Genetic foundation of unrivaled survival strategies

- Of water bears and carnivorous plants -

Genetische Grundlagen einzigartiger Überlebensstrategien

- Über Bärtierchen und fleischfressende Pflanzen -

Felix Mathias Bemm

from Weimar

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

Universität Würzburg Würzburg, March 2017

BY NC SA

Submitted on:

Members of the thesis committee

Chairperson:

Primary Supervisor: Prof. Jörg Schultz

Dept. of Bioinformatics
Biocentre, Am Hubland
University Würzburg
D-97074 Würzburg

Second Supervisor: Dr. Dirk Becker

Dept. of Botany I
Julius-von-Sachs-Institute of Biosciences
University Würzburg
D-97074 Würzburg

Third Supervisor: Prof. Roy Gross

Department of Microbiology
Biocentre, Am Hubland
University Würzburg
D-97074 Würzburg

Date of Public Defence:

Date of Receipt of Certificates:

I would like to dedicate this thesis to my whole family which has been a tireless critic, an outstanding example and a faithful friend throughout my whole life . . .

Declaration

I hereby confirm that my thesis entitled "Genetic foundation of unrivaled survival strategies - Of waterbears and carnivorous plants" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and / or materials applied are listed and specified in the thesis. Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Felix Mathias Bemm
Würzburg, March 2017

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Dissertation "Genetic foundation of unrivaled survival strategies - Of waterbears and carnivorous plants" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben. Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Felix Mathias Bemm
Würzburg, March 2017

Acknowledgements

I would first like to thank my thesis advisory board, namely Prof. Jörg Schultz, Prof. Dirk Becker and Prof. Roy Gross (all University of Würzburg) for constant support throughout the past years. Additionally, I would like to thank the Graduate School of Life Sciences (University of Würzburg) for the initial and the extended grant support. Many thanks go to all former colleagues of the Department of Bioinformatics (University of Würzburg) especially Dr. Frank Förster, Dr. Alexander Keller and Dr. Thomas Hackl. I am especially grateful for the students that accompanied me during the past years, namely Anna-Lena Van de Weyer, Clemens Weiss, Markus Ankenbrand and Lorenz Weber. Finally, I would like to thank Detlef Weigel, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen for patiently waiting on my final graduation.

Dionaea muscipula **Transcriptomics** I would further like to thank Prof. Rainer Hedrich (University of Würzburg) for allowing me to participate and center my PhD thesis around the *D. muscipula* transcriptome project. Within this context, I would like to thank Prof. Dirk Becker, who patiently trained me to think like a biologists but on the other side allowed me to constantly overwhelm him with bioinformatical chitchat. Special thanks to Dr. Ines Kreuzer (University of Würzburg) for constant experimental support.

Tardigrade Genomics Similarly, I would like thank to Prof Jörg Schultz for allowing me to work in the field of tardigrade genomics and for patient, non-authoritarian support of my whole scientific career. Special thanks to Dr. Frank Förster for help with computing and programming matters, to Prof. Christian Janzen for FACS training and support, to Dr. Ralph Oliver Schill for provision of tardigrade specimen, Dr. Alexander Keller for lab support and Prof. Ulrich Certa, F. Hoffmann-La Roche AG for initial sequencing of *Milnesium tardigradum*.

Finally, I would like to thank my family, my fiancée and the Flöten-Franz (*Canis lupus familiaris*, FCI Group 3, Section 2 Small Terriers #345) for unquestioned backing, positive diversion and all-embracing love.

Summary

All living organisms leverage mechanisms and response systems to optimize reproduction, defense, survival, and competitiveness within their natural habitat. Evolutionary theories such as the universal adaptive strategy theory (UAST) developed by John Philip Grime (1979) attempt to describe how these systems are limited by the trade-off between growth, maintenance and regeneration; known as the universal three-way trade-off. Grime introduced three adaptive strategies that enable organisms to coop with either high or low intensities of stress (e.g., nutrient deficiency) and environmental disturbance (e.g., seasons). The competitor is able to outcompete other organisms by efficiently tapping available resources in environments of low intensity stress and disturbance (e.g., rapid growers). A ruderal specism is able to rapidly complete the life cycle especially during high intensity disturbance and low intensity stress (e.g., annual colonizers). The stress tolerator is able to respond to high intensity stress with physiological variability but is limited to low intensity disturbance environments. Carnivorous plants like *D. muscipula* and tardigrades like *M. tardigradum* are two extreme examples for such stress tolerators. *D. muscipula* traps insects in its native habitat (green swamps in North and South Carolina) with specialized leaves and thereby is able to tolerate nutrient deficient soils. *M. tardigradum* on the other side, is able to escape desiccation of its terrestrial habitat like mosses and lichens which are usually covered by a water film but regularly fall completely dry. The stress tolerance of the two species is the central study object of this thesis. In both cases, high througput sequencing data and methods were used to test for transcriptomic (*D. muscipula*) or genomic adaptations (*M. tardigradum*) which underly the stress tolerance. A new hardware resource including computing cluster and high availability storage system was implemented in the first months of the thesis work to effectively analyze the vast amounts of data generated for both projects. Side-by-side, the data management resource TBro [14] was established together with students to intuitively approach complex biological questions and enhance collaboration between researchers of several different disciplines. Thereafter, the unique trapping abilities of *D. muscipula* were studied using a whole transcriptome approach. Prey-dependent changes of the transcriptional landscape as well as individual tissue-specific aspects of the whole plant were studied. The analysis revealed that non-stimulated traps of *D. muscipula* exhibit the expected hallmarks of any typical leaf but operates evolutionary conserved stress-related pathways including defense-associated responses when digesting prey. An integrative approach, combining proteome and transcriptome data further enabled the detailed description of the digestive cocktail and the potential nutrient uptake machinery of the plant. The published work [25] as well as a accompanying video material (https://www.eurekalert.org/pub_releases/2016-05/cshl-fgr042816.php; Video credit: Sönke Scherzer) gained global press coverage

and successfully underlined the advantages of *D. muscipula* as experimental system to understand the carnivorous syndrome. The analysis of the peculiar stress tolerance of *M. tardigradum* during cryptobiosis was carried out using a genomic approach. First, the genome size of *M. tardigradum* was estimated, the genome sequenced, assembled and annotated. The first draft of *M. tardigradum* and the workflow used to established its genome draft helped scrutinizing the first ever released tardigrade genome (*Hypsibius dujardini*) and demonstrated how (bacterial) contamination can influence whole genome analysis efforts [27]. Finally, the *M. tardigradum* genome was compared to two other tardigrades and all species present in the current release of the Ensembl Metazoa database. The analysis revealed that tardigrade genomes are not that different from those of other Ecdysozoa. The availability of the three genomes allowed the delineation of their phylogenetic position within the Ecdysozoa and placed them as sister taxa to the nematodes. Thereby, the comparative analysis helped to identify evolutionary trends within this metazoan lineage. Surprisingly, the analysis did not reveal general mechanisms (shared by all available tardigrade genomes) behind the arguably most peculiar feature of tardigrades; their enormous stress tolerance. The lack of molecular evidence for individual tardigrade species (e.g., gene expression data for *M. tardigradum*) and the non-existence of a universal experimental framework which enables hypothesis testing withing the whole phylum Tardigrada, made it nearly impossible to link footprints of genomic adaptations to the unusual physiological capabilities. Nevertheless, the (comparative) genomic framework established during this project will help to understand how evolution tinkered, rewired and modified existing molecular systems to shape the remarkable phenotypic features of tardigrades.

Zusammenfassung

Alle lebenden Organismen verwenden Mechanismen und Rückkopplungssysteme um Reproduktion, Überlebenswahrscheinlichkeit, Abwehreffizienz und Konkurrenzfähigkeit in ihrem natürlichen Habitat zu optimieren. Evolutionäre Theorien, wie die von John Philip Grime (1979) entwickelte „universal adaptive strategy theory“ (UAST), versuchen zu beschreiben wie diese Systeme durch eine Balance zwischen Wachstum, Erhaltung und Regeneration, auch gemeinhin bekannt als universeller Dreiwege-Ausgleich, des jeweiligen Organismus limitiert sind. Grime führte dazu drei adaptive Strategien ein, die es Organismen ermöglicht sich an hohe oder niedrige Stress-Intensitäten (z.B. Nahrungsknappheit) oder umweltbedingte Beeinträchtigung (z.B. Jahreszeiten) anzupassen. Der Wettkämpfer ist in der Lage seine Konkurrenz durch eine effiziente Ressourcengewinnung zu überflügeln und ist vor allem bei niedrigem Stresslevel und minimalen umweltbedingten Beeinträchtigungen effizient (z. B. schnelles Wachstum). Ruderale Organismen hingegen durchlaufen den Lebenszyklus in kurzer Zeit und sind damit perfekt an starke umweltbedingte Beeinträchtigungen, wie zum Beispiel Jahreszeiten, angepasst. Allerdings können auch sie nur bei niedrigen Stresslevel effizient wachsen. Die letzte Gruppe von Organismen, die Stresstoleranten sind in der Lage sich an hohen Stressintensitäten mithilfe extremer physiologischer Variabilität anzupassen, können das allerdings nur in Umgebungen mit niedrigen umweltbedingten Beeinträchtigungen. Fleischfressende Pflanzen wie die Venusfliegenfalle (*D. muscipula*) oder Bärtierchen (*M. tardigradum*) sind zwei herausragende Beispiele für stresstolerante Organismen. Die Venusfliegenfalle ist in der Lage Insekten mit spezialisierten Blätter, welche eine einzigartige Falle bilden, zu fangen. Die Pflanze kompensiert so die stark verminderte Mengen an wichtigen Makronährstoffen (z.B. Stickstoff) in den Sümpfen von Nord- und Süd-Carolina. Bärtierchen dagegen sind in der Lage in schnell austrocknenden Habitaten wie Moosen oder Flechten, die normalerweise mit einem Wasserfilm überzogen sind, durch eine gesteuerte Entwässerung ihres Körpers zu überleben. Die Stresstoleranz beider Spezies ist zentraler Forschungsschwerpunkt dieser Dissertation. In beiden Fällen werden Hochdurchsatz-Methoden zur Sequenzierung verwendet um genomische (Bärtierchen) sowie transkriptomische (Venusfliegenfalle) Anpassungen zu identifizieren, die der enorem Stresstoleranz zugrunde liegen. Um den erhöhten technischen Anforderungen der Datenanalysen beider Projekte Rechnung zu tragen wurde in den ersten Monaten der Dissertation eine neue zentrale Rechenumgebung und ein dazugehöriges Speichersystem etabliert. Parallel wurde die Datenmanagementplattform TBro [14] zusammen mit Studenten aufgesetzt, um komplexe biologische Fragestellung mit einem fachübergreifendem Kollegium zu bearbeiten. Danach wurden die einzigartigen Fangfähigkeiten der Venusfliegenfalle mittels einem transkriptomischen Ansatz untersucht. Vor allem wurden transkriptionelle Änderungen infolge

eines Beutefangs sowie gewebespezifische Aspekte der ruhenden Pflanzen untersucht. Die Analyse zeigte deutlich, dass die Fallen der fleischfressenden Pflanze immer noch Merkmale von typischen „grünen“ Blättern aufweisen. Während des Beutefangs und -verdauens jedoch wird eine Vielzahl an evolutionär konservierten Systemen aktiviert, die bisher nur mit Stressantworten und zellulärer Verteidigung in Verbindung gebracht worden sind. Die Integration von proteomischen und transkriptomischen Hochdurchsatzdaten ermöglichte es zudem den Verdauungssaft der Venusfliegenfalle genaustens zu beschreiben und wichtige Komponenten der Aufnahmemaschinerie zu identifizieren. Die wissenschaftliche Arbeit [25] und das begleitende Videomaterial (https://www.eurekalert.org/pub_releases/2016-05/cshl-fgr042816.php; Video credit: Sönke Scherzer) erfreute sich einer breiten Berichterstattung in den Medien und unterstreicht die Vorteile der Venusfliegenfalle als experimentelles System um fleischfressende Pflanzen besser zu verstehen. Die genomische Analyse des Bärtierchen (*M. tardigradum*) zielte auf die außerordentliche Stresstoleranz, vor allem auf die Kryptobiose, einen Zustand in dem Stoffwechselvorgänge extrem reduziert sind, ab. Dazu wurden das komplette genetische Erbgut (Genom) entschlüsselt. Die Größe des Genomes wurde bestimmt und das Erbgut mittels Sequenzierung entschlüsselt. Die gewonnenen Daten wurden zu einer kontinuierlichen Sequenz zusammengesetzt und Gene identifiziert. Der dabei etablierte Arbeitsablauf wurde verwendet um ein weiteres Bärtierchengenom genau zu überprüfen. Im Rahmen dieser Analyse stellte sich heraus, dass eine große Anzahl an Kontaminationen im Genom von *H. dujardini* vorhanden sind [27]. Das neu etablierte Genom von *M. tardigradum* wurde im folgenden verwendet um einen speziesübergreifenden Vergleich dreier Bärtierchen und aller Spezies aus der Metazoadatenbank von Ensembl durchzuführen. Die Analyse zeigte, dass Bärtierchengenome sehr viel Ähnlichkeit zu den bereits veröffentlichten Genomen aus dem Überstamm der Urmünder (Protostomia) aufweisen. Die erstmalige Verfügbarkeit aller Bärtierchengenome ermöglichte es zudem, das Phylum der Bärtierchen als Schwester der Nematoden mittels einer phylogenomische Analyse zu platzieren. Die vergleichende Analyse identifizierte außerdem zentrale evolutionäre Trends, vor allem einen enormen Verlust an Genen in dieser Linie der Metazoa. Die Analyse ermöglichte es aber nicht, generelle Mechanismen, die zur enormen Stresstoleranz in Bärtierchen führen, artübergreifend zu identifizieren. Vor allem das Fehlen von weiteren molekularen Daten für einzelne Bärtierchenspezies (z.B. transkriptionelle Daten für *M. tardigradum*) machten es unmöglich die wenigen genomische Adaptionen mit den physiologischen Besonderheiten der Bärtierchen in Deckung zu bringen. Nichtsdestotrotz konnten die vergleichenden Analysen zeigen, dass Evolution auch innerhalb der Bärtierchen verschiedenste Systeme neu zusammensetzt, neue Funktionen erschafft oder bestehenden Systeme modifiziert und damit die außerordentliche phänotypische Variabilität ermöglicht.

Table of contents

List of figures xvii

List of tables xix

Nomenclature xxi

1 *D. muscipula* Transcriptomics 1

1.1 Abstract 1
1.2 Introduction 2
1.2.1 Carnivory in the plant kingdom 2
1.2.2 The Venus flytrap 3
1.2.3 Project objectives 7
1.3 Material and Methods 11
1.3.1 Computing environment 11
1.3.2 Experimental procedures 11
1.3.3 Data preparation and transcriptome assembly 11
1.3.4 Feature annotation 12
1.3.5 Transcript filtering 13
1.3.6 Transcript abundance estimation and differential expression tests . . 13
1.3.7 Enrichment analysis and comparison 13
1.3.8 Definition of the putative secretome, transportome and kinome . . . 14
1.3.9 Metagenomic profiling and comparison 14
1.4 Results 15
1.4.1 Data inventory 15
1.4.2 Data quality control and trimming 18
1.4.3 The reference transcriptome 24
1.4.4 Transcriptome quantification 40
1.4.5 The molecular make-up of a non-stimulated trap 46

1.4.6 The trigger hairs 52
1.4.7 Secretion-response to insect feeding 63
1.4.8 Inventory of the glandular nutrient uptake machinery 67
1.4.9 Trap-wide response to insect feeding 71
1.4.10 Regulatory capacities of the *D. muscipula* 77
1.4.11 Sensory capacities of the *D. muscipula* 79
1.4.12 Metagenomics changes during insect feeding 83
1.4.13 Conserved stress signals 86
1.4.14 Gene family expansions in *D. muscipula* 91
1.5 Discussion 94
1.6 Epilogue 101
1.7 Published Elements 102

2 TBro: Visualization and management of denovo transcriptomes 103
2.1 Preface 103
2.2 Supplementary Material 112

3 Tardigrade Genomics 115
3.1 Abstract 115
3.2 Introduction 116
3.2.1 Project objectives 120
3.3 Material and Methods 121
3.3.1 Animal culture 121
3.3.2 Sequencing 121
3.3.3 Genome size estimation 122
3.3.4 Genome and transcriptome assembly 122
3.3.5 Genome feature annotation 123
3.3.6 Protein Domain Expansions and Contractions 123
3.3.7 Phylogenetic Reconstruction and Hypothesis Testing 124
3.3.8 Protein Domain Gain and Loss Estimation 124
3.3.9 Functional Hypothesis Testing 124
3.4 Results 126
3.4.1 Genomic Features of *M. tardigradum* 126
3.4.2 Comparative Framework 128
3.4.3 Protein Domain Expansions and Contractions 133
3.4.4 Phylogenetic Reconstruction 140
3.4.5 Protein Domain Gain and Loss Estimation 142

3.4.6 Functional Hypotheses Testing 144
3.5 Discussion 147
3.6 Epilogue 150
3.7 Published elements 150
3.8 Submitted elements 150

References **151**

Appendix A GT34 Protein Phylogenetics **177**
A.1 Introduction 177
A.2 Material and Methods 178
A.2.1 Pinus taeda EST database mining 178
A.2.2 Multiple sequence alignments and phylogenetic trees 178
A.3 Results and Discussion 179
A.4 Summary 180
A.5 Published elements 181

Appendix B Curriculum vitae **183**

List of figures

1.1 Per sequence quality scores 19
1.2 Read duplication levels 20
1.3 Per sequence GC content 21
1.4 Read trimming results 22
1.5 Contamination screen 23
1.6 Post-assembly intron correction 26
1.7 Post-assembly fusion correction 27
1.8 Isoform per unigene distribution 28
1.9 Identity and coverage distribution 30
1.10 Post-assembly filtering process 31
1.11 RNA-seq re-mapping rates 40
1.12 Observed quality vs. phred quality score 41
1.13 Distances between experiments 43
1.14 Standard deviation of expression values 44
1.15 Principal component analysis — non-stimulated plant 46
1.16 Intersection analysis — non-stimulated plant 48
1.17 Tissues-by-tissue correlation — non-stimulated plant 50
1.18 Amplification benchmark I (trap) 53
1.19 Amplification benchmark II (gland & hairs) 54
1.20 Tissues-by-tissue correlation — non-stimulated plant & hairs 56
1.21 Principal component analysis — non-stimulated plant & hairs 58
1.22 Intersection analysis — non-stimulated plant & hairs 59
1.23 GO tree map — putative transcriptome of hairs 61
1.24 Hydrolytic cocktail inventory 65
1.25 Transporter inventory 68
1.26 AHA1-like P-Type H^+ ATPase expression 69
1.27 GO tree map — insect-stimulated traps 73

1.28 Insect-stimulated JA biosynthesis 74
1.29 Transcription factor inventory 77
1.30 Kinase inventory 79
1.31 RLK expression overview 81
1.32 FLS2-like RLK expression 82
1.33 Metagenomic PCA — non-stimulated plant 84
1.34 Metagenomic PCA — COR-stimulated traps 84
1.35 Metagenomic PCA — insect-stimulated traps 85
1.36 Metagenomic analysis — insect-stimulated traps 85
1.37 QC - *Genlisea nigrocaulis* & *Genlisea hispidula* 87
1.38 Semantic similarity — insect-stimulated traps 90
1.39 Semantic similarity — insect-stimulated & non-stimulated traps 91
1.40 eggNOG assignment levels 92
1.41 GO tree map — euNOG expansions 93

2.1 Supplementary figure 1 112

3.1 Genome size estimation 127
3.2 Ensembl Metazoa genome selection 131
3.3 Genome completeness 132
3.4 Expansions intersection 134
3.5 Phylogenetic reconstruction 141
3.6 Hypothesis testing 141
3.7 Domain loss in three Ecdysozoan lineages 142
3.8 Trehalose biosynthesis & metabolism 144
3.9 Heat shock proteins expression 145

A.1 Phylogenetic tree of family GT34 proteins. 180

List of tables

1.1 Sequencing experiments 16
1.3 Proteome screenning experiments 17
1.5 BUSCO benchmarking results 29
1.6 Repeat annotation results 32
1.7 Open reading frame annotation results 33
1.8 Interpro annotation results 33
1.9 Gene ontology annotation results 34
1.10 MapMan annotation results 35
1.11 TF classification results 36
1.13 Kinase classification results 38
1.14 CRB assignment results — *Arabidopsis thaliana* 39
1.15 EggNOG assignment results 39
1.17 GO enrichment — 1st component 47
1.18 GO enrichment — 2nd component 47
1.19 GO enrichment — 3rd component 47
1.20 GO enrichment — flower-specific DEGs 48
1.21 GO enrichment — gland & root shared DEGs 50
1.22 MapMan bin enrichment — gland & root shared DEGs 51
1.23 GO Enrichment — putative transcriptome of hairs 60
1.24 GO enrichment — insect-stimulated glands 63
1.25 MapMan bin enrichment — insect-stimulated glands 64
1.26 Exocytosis component inventory 66
1.28 Endocytosis component inventory 70
1.29 GO enrichment — insect-stimulated traps 71
1.30 MapMan bin enrichment — insect-stimulated traps 72
1.31 MapMan bin counts — insect-stimulated traps 75
1.32 GO term PCD — insect-stimulated traps 76

1.33 Transcription factor family enrichment 78
1.34 Kinome family enrichment 80
1.35 GO enrichment — DEGs *G. nigrocaulis* traps 87
1.36 GO enrichment — PEGs *G. hispidula* & *G. nigrocaulis* traps 88
1.37 GEM2Net enrichment 89
1.38 euNOG functional categories 93

2.1 TBro responsibilities & contributions 104
2.2 TBro responsibilities & contributions to figures 104
2.3 Supplementary table 1 113

3.1 Assembly metrics 129
3.2 Protein domain expansions Class I 135
3.4 Protein domain expansions Class II 137
3.6 Protein domain contractions Class I 139
3.8 Protein domain contractions Class II 139
3.10 Protein domain gains within the Tardigrada 143
3.12 Heat shock proteins 145
3.13 Tardigrada-unique genes 146

Nomenclature

Acronyms / Abbreviations

AOX	alternative oxidase
AP	action potential
ATPases	adenosine triphosphatases
BMA	best match strategy
Ca	calcium
CAHS	cytoplasmic abundant heat soluble
cDNA	complementary DNA
CDS	coding sequence
Cl	chloride
CLI	command line interface
CO_2	carbon dioxide
COR	Coronatine
CP	carnivorous plants
CPU	central processing unit
CRB	conditional reciprocal best BLAST
DE	differential expression
DEG	differentially expressed gene

DNA	deoxyribonucleic acid
Dsup	damage suppressor
ER	endoplasmic reticulum
exp	expression experiment
FDR	false discovery rate
GC	guanine/cytosin
GO	gene ontology
GT	glycosyltransferases
HD/HDUJ	*H. dujardini*
HMM	hidden Markov model
HSD	honest significance difference
HSP	heat shock proteins
HTS	high throughput screening
JA	jasmonic acid
K	potassium
LCA	lowest common ancestor
LEA	late embryo abundant
MAHS	mitochondria-targeted heat-soluble
Mg	magnesium
MT/MTAR	*M. tardigradum*
N	nitrogen
NCBI	National Center for Biotechnology Information
NDP	nucleoside diphosphate
NOG	non-supervised orthologous groups

ORF	open reading frame
P	phosphorus
PCA	principal component analysis
PCD	programmed cell death
PCR	polymerase chain reaction
PEG	preferentially expression genes
PM	plasma membrane
pro	proteomic experiment
QC	quality control
rbcL	Ribulose-1,5-bisphosphate carboxylase/oxygenase
RLK	receptor-like-kinase
RNA	ribonucleic acid
$RNA-seq$	RNA sequencing
ROS	reactive oxygen species
$RV/RVAR$	*Ramazzottius varieornatus*
S	sulfur
SAHS	secretory abundant heat soluble
SOD	superoxide dismutase
SRA	short read archive
TC	transporter classification
TE	transposable element
TF	transcription factor
WGD	whole genome duplication

Chapter 1

D. muscipula Transcriptomics

1.1 Abstract

Since Darwin's time the concept of plant carnivory is known. Nevertheless, limited molecular evidence was available until now to understand the mechanisms underlying the carnivorous trait. The following chapter presents the first robust transcriptomic study of a carnivorous plant, namely *D. muscipula*. The comprehensive data analysis characterizes the molecular transition that occurs when *D. muscipula* activates its trapping organs. Results indicate that resting traps still operate in a leaf-like manner but rapidly alter the transcriptomic landscape upon insect feeding. *D. muscipula*, specifically the traps activate defense response pathways during prey digestion alongside with a highly productive secretory system and a complex nutrient uptake machinery. A first comparative analysis with non-carnivorous plants during wounding stress revealed several conserved pathways and expression patterns. Thus, the data presented provides rich insight how generic stress response pathways might have been rewired during evolution to suite prey capture, digestion, and nutrient acquisition.

1.2 Introduction

1.2.1 Carnivory in the plant kingdom

Carnivorous plants are one of the most spectacular curiosities that can be found in the plant kingdom. Presently, the carnivorous syndrome is recognized in over 600 species, across 11 families and 19 genera. Its existence was first announced in 1769 by Carl von Linné. Starting with the first experimental evidence presented in Charles Darwin's book "Insectivorous Plants" [70], carnivorous plants have attracted thousands of researchers to study their capturing and digesting mechanisms. The first phylogenetic analysis using nucleotide diversity of the plastid Ribulose-1,5-bisphosphate carboxylase/oxygenase (rbcL) genes [8] and numerous recent studies (e.g., [240, 259, 133] provided evidence that carnivorous plants are a polyphyletic group and that the carnivorous syndrome has evolved at least six times within angiosperms [113, 114]. The most complex capturing and digestion machineries can be found within the Caryophyllales, scattered over four different families, namely Droseraceae, (Drosera, Dionaea, Aldrovanda), Drosophyllaceae (Drosophyllum), Nepenthaceae (Nepenthes), and partially within the Dioncophyllaceae (Triphyophyllum). Phylogenetic analysis [49] suggested all of the four being part of a monophyletic group within the non-core Caryophyllales and hypothesised that the common ancestor had adhesive flypaper traps. In modern species the trapping mechanisms greatly vary, from passive flypaper, pitfall and suction traps to active snap traps. They all serve the same purpose; capturing of small organisms for nutrition. Prey is usually attracted by pigments, reflection patterns or chemical mimicry and captured by either surface immobilization (e.g., *Drosophyllum tbd*), pitfall trapping (e.g., *Nepenthes pervillei*) or formation of a hermetically-sealed "green stomach" (e.g., *D. muscipula*). Multicellular glands which are either sessile, stalked, or pitted, allow the plants to secrete digestive enzymes and absorb the digested material. Carnivorous plants can benefit from the acquired nutrients in multiple ways [108]. Nutrient acquisition is commonly followed by plant growth [4], nutrient storage [47] and an increase in reproduction [276, 300]. In some aquatic carnivorous plants (e.g., *Aldrovanda vesiculosa*) even carbon is directly taken up and might compensate for the CO_2 deficit of the aquatic environment that otherwise limits the photosynthetic rate [2, 3]. On the downside, the costs for the construction of specialized morphological trapping structures, the generation of digestive enzymes and the activity of the uptake machinery can be substantial [84, 83, 159, 219].

1.2.2 The Venus flytrap

Carnivorous plants commonly inhabit bright and wet areas that are very low in nutrients [108]. Normally, nutrient deficiency caused by low concentration of nutrients or their constrained availability dramatically effects plant growth. Especially, macronutrients such as nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), sulfur (S) and magnesium (Mg) are vital for plant growth. If limited, plants are unable to generate essential metabolites necessary for a normal life cycle. Extreme nutrient deficiency affects plant growth and development so strongly that plants experience stunting, deformity, discoloration, distress, or death. Natural deficiencies often involve multiple nutrients especially in swamps. Still, nitrogen is the nutrient most often in shortest supply. The Venus flytrap (*D. muscipula*), native to the nutrient-low Green Swamps in North and South Carolina [235, 2, 248] copes with nutrient deficiency by feeding on nitrogen-rich insects. The plant develops snap traps at the end of its petioles to catch larger prey [105]. Insects are attracted by various stimuli like the release of volatile organic compounds [175] and captured by the rapidly closing bi-lobed snap traps. Closure is triggered by touch stimulation of mechano-sensitive trigger hairs at the inner surface of the snap traps [119, 99, 88]. Structurally, trigger hairs can be partitioned into four zones [119, 296]. The indentation zone (see [119], zone III) contains the sensory cells. Both poles of the sensory cells show high amounts of endoplasmic reticulum (ER) but low number of ribosomes [41]. Sensory cells are connected to podium cells (see [119], zone IV) with numerous plasmodesmata. Due to the high number of plasmodesmata it has been hypothesized that signal transduction favours the symplastic pathway ([296, 123]). The center of the sensory cells comprises the nucleus, organelles (mostly mitochondria), lipid droplets and smaller vacuoles [41]. The cell walls of the sensory cells are generally thickened with the exception of the central parts, probably to form a predetermined soft spot that guides bending. The larger vacuoles contain polyphenolic compounds which might be involved in storage, binding and release of ions needed for the transformation of the mechanical stimulus into an electrochemical signal [296, 41]. The process of perception and transformation of the mechanical stimulus into a physiological signal remains unclear. Already Darwin realized that the prey capturing is a process driven by electrical excitability and very fast biomechanical movements, but it was Burdon-Sanderson in 1872 who discovered the responsible action potential (AP) [44]. 130 years later, it was suggested that the trigger hairs perceive the mechanical stimulus [119]. At first, the mechanical stimulus locally creates a receptor potential that, if strong enough, develops into an AP. The AP travels across the electrically coupled trap and prepares it for an upcoming closure. An additional AP within the next 10s causes the fast closure by releasing the elastic energy stored in the traps. Further stimulation of the trigger hairs and additional propagated APs lead to hermetical sealing of the trap. At least five sequential APs are necessary to seal the capture organ. Interestingly, the plant is able to memorize the number of generated APs to prevent false alarms [288, 32]. Recent studies demonstrated that the action potentials at the plasma membrane of *D. muscipula* depend on Cl^{-} and K^{+} for de- and re-polarisation [87]. Nevertheless, the genetic mark-up necessary to describe this model is still lacking molecular evidence. After recognition of the prey, the trap closes within a fraction of a second. While small insects might get a chance to escape the closed traps through the teeth-like structure at the outer rim of the trap [70], larger ones are inevitable stuck; a process that saves the plant respiratory energy to start digestion of unworthy prey. Nevertheless, recent results point towards a less selective capturing when it comes to prey size [141]. The traps show the typical morphological and physiological properties of a green leaf although it serves as highly specialized organ to hunt for prey [290]. The morphological features between the trap-forming leaf tip (trap) are very much different from the leaf base (petiole). Still, physiological experiments demonstrate the expected photosynthetic activity of the trap even though at a lower rate [218]. The most striking difference between the two organ parts are the approximately 37,000 glands on the inner surface of the trap [155, 87] that are directly linked to prey digestion and nutrient uptake activity during insect feeding. After sealing is completed, the plant starts secretion of its mucous acidic fluid which contains enzymes to digest the prey [271, 249]). Other carnivorous plants directly release the enzymes into water containing traps (e.g., *A. vesiculosa*) upon activation of the feeding process. The only exception is *Utricularia gibba* which seem to constantly release its digestive enzymes into the traps. The secretion process of *D. muscipula* can also be triggered by the known touch hormone jasmonic acid (JA) or Coronatine (COR), a bacterial toxin that mimics the plant hormone [87]. This suggests a direct link between electrical signal and internal chemical response. Acidification of the digestive fluid probably aides the breakdown process and optimizes the enzyme activity. Early studies suggested that proteases and phosphatase are the major enzyme components in a number of different species [277, 236, 102, 122] and that the secretion mechanism is highly regulated by signaling processes to maximize gain of the extracted nutrients in a prey dependent manner. Nevertheless, a comprehensive and deep analysis of such cocktails is still missing due to poor experimental design (e.g., PCR-based single gene analysis) or lack of resolution (e.g., proteomics only studies). After breakdown of the prey [124], nutrients are taken up. While non-carnivorous plants mostly rely on their complex root system - especially the root hairs - to selectively absorbe and transport nutrients, carnivorous plants such as *D. muscipula* are able to aquire nutrients not only with their root system but also through the trap surface during the feeding process. Several studies with organic nitrogen and carbon suggested the glands as responsible for nutrient resorption [248, 176], underlining its previously speculated dimorphic character. But, the molecular landscape of the uptake machinery remained unknown. The nutrients gained provide *D. muscipula* with an extreme competitive advantage and help it to overcome the nitrogen deficiency caused by its natural low-nutrient soil habitats [2]. Although several carnivorous plant species were targeted by high-throughput sequencing projects [142, 143, 185, 94, 20, 52–54, 264, 279] none of them was able to improve our understanding of the molecular basis of the carnivorous syndrome dramatically. Efforts focusing on the comparison of two trumpet pitcher species (*Sarracenia psittacina* and *Sarracenia purpurea*) only revealed that genes under positive selection are associated with molecular binding activity [262]. Unfortunately, the work did not shed light on the unique combination of lures (e.g., scent, drugged nectar, waxy deposits to clog insect feet) that most Sarracenia species use in combination with their one-way traps to capture prey. The first genome study of a carnivorous plant targeted the floating bladderwort *U. gibba* [143]. The study reported a genome size contraction, mainly due to compression of intronic and intergenic regions as well as gene loss in highly multigenic gene families, after several rounds of whole genome duplication (WGD). A similar contraction was observed in *Genlisea aurea*, the largest carnivorous species in the genus Genlisea [185]. Still, the *U. gibba* genome encoded a diverse gene landscape. The sequencing and analysis of the *U. gibba* transcriptome revealed vegetative shoots and traps most akin to each other [142]. Traps expressed various hydrolytic enzymes that are potentially involved in prey digestion and previously were thought to be encoded by bacterial genomes only. Additionally, the plant showed an accelerated respiration activity coupled with a high activity of DNA repair and reactive oxygen species (ROS) detoxification enzymes. By analyzing preferentially expressed genes (gene specifically but not uniquely expressed in specific tissues or conditions), authors also argued that traps are mainly responsible for phosphate uptake (through high affinity inorganic phosphate transporters) while nitrate uptake happens more likely in vegetative parts of the plants (through action of low- and high-affinity nitrate transporters). Nevertheless, the transportome analysis remained superficial due to the absence of biological replicates and accompanying feeding experiments. Efforts focusing on other Utricularia species (e.g., *Utricularia vulgaris*) did not reveal further insights into the feeding process but mainly confirmed previous presence and absence patterns of root-associated genes [20]. The first transcriptomic studies in the order of the Caryophyllales were carried out for *Nepenthes ampullaria* and *D. muscipula*. Transcript data for *N. ampullaria* was only deposited but not further analyzed [292]. The transcriptome analysis of *D. muscipula* included generation of a transcriptome assembly from flowers and traps as well as its functional annotation [147]. Authors reported an abundant representation of processes related to catalytic, antioxidant and electron carrier activities but did not link their results to known morphological or functional traits associated with the carnivorous capabilities of *D. muscipula*. In summary, present high-throughput sequencing studies barely shed light on the molecular mechanism underlying plant carnivory. Most processes remain poorly understood due to missing functional evidence or week experimental designs. All projects heavily suffer from missing biological replicates (especially those that leverage transcriptome sequencing), shallow sequencing depth, suboptimal tissues sampling schemes and uncontrolled feeding states of the trapping organs. Here, *D. muscipula* provides a perfect model to understand prey-dependent changes of the transcriptional landscape as well as to study individual tissue-specific aspects of the whole plant since its trapping and feeding mechanisms can be actively triggered [87]. Even better, the process can be tightly controlled by application of mechanical stimuli, insect feeding or trough the JA mimicry COR. Thereby, *D. muscipula* seems the most suitable model to understand prey-induced transcriptional changes that underlie the carnivorous syndrome. The here presented thesis work thus focuses on the analysis and interpretation of the transcriptomic landscape of *D. muscipula*, tries to overcome the above mentioned previous shortcomings and aims to provide a blueprint to study other carnivorous plant species.

1.2.3 Project objectives

The reference transcriptome First objective is to generate a reliable transcriptome resource. Since genome-wide sequencing efforts are still ongoing the transcriptome needs to be established *de novo*. Thorough quality control and cross checks need to be carried out to i) assure that a maximum number of transcripts are present, ii) the contamination level (e.g., metagenomic and prey-derived contamination) is low and iii) the structural integrity of the transcriptome is consistent.

The transcriptomic landscape of a snap trap Based on the reference transcriptome, global expression patterns should be compared using RNA sequencing (RNA-seq) data from the four major organs (petiole, trap, roots and flowers). It should be tested whether individual organs share expression profiles globally and what the underlying transcripts are biologically associated with. Additionally, it should be tested whether tissue-specific expression patterns exists and to what extend they explain the morphological and physiological make-up of the organs. Since RNA-seq data from traps integrates the expression profile of various typical leaf cell types as well as glandular tissues and trigger hair cells, it could be hypothesized that traps display a expression patchwork that can be linked to different other organs. The objective is to use additional RNA-seq samples from the trap rim, which is free of digesting glands, and glands to compare their expression profiles to non-trap tissues like petiole, flower and root. It should also be tested to what other organ glands are most akin to and to what extend this fits with the assumption that glandular tissue evolved from flower nectaries [58].

Targeting low input transcriptome profiling Trigger hairs consist of a small number of highly specialized cells. Since overall tissue mass is extremely small, it is not possible to submit them to sequencing-based transcriptome profiling without pooling hairs from several traps or even from several hundred plants. The only alternative to massive pooling is PCR-based amplification of extracted RNA and subsequent transcriptome profiling. The objective is to test whether RNA-seq results are concordant when amplified and non-amplified experiments are compared. Tests should be carried out against a non-amplified reference (e.g., traps and COR-treated traps) to assess the results from a qualitative and quantitative perspective. Additionally, it should be tested whether non-amplified and amplified trigger hair samples already exhibit expression patterns that can be linked to PCR-based amplification or other batch effects. Final results should provide insights whether PCR-based RNA amplification can be used throughout transcript quantification in *D. muscipula*.

The transcriptome of the trigger hair Sensory cells of the trigger hairs are shaped by apical and basal ER cisternae, numerous mitochondria, as well as vacuoles and lipid droplets. The objective is to describe the transcriptome profile of the trigger hair in respective to its specialized structures. The analysis needs to correct for eventual gland signatures as consequence of tissue contamination. Additionally, the trigger hair specific kinome should be described as well as anion and cation transporters that might play a role during the generation of the AP.

The secreted hydrolytic enzyme cocktail of *D. muscipula* Understanding the composition of the hydrolytic cocktail is one of the most important objectives of this thesis. The primary goal is to identify cocktail components by overlaying differential expression testing data from insect-stimulated traps, profile-based signal peptide annotations (secreted enzymes carry these signal peptides) and High Throughput Screening (HTS) proteomics data from three different stimulation experiments (insect-feeding, COR treatment and mechanical stimulation). Transcription of most secreted enzymes should respond to insect-feeding. A sub-selection of those identified in HTS proteomic data and with a signal peptide annotation should yield a human feasible set of transcripts for manual inspection. In parallel, it should be tested whether secreted enzymes exist (those detected in HTS proteomic data) that are produced in advance in non-stimulated glands and thus might be early secreted. Hydrolytic enzymes are likely to be secreted via exocytosis. The exocytosis machinery, specifically the components of the exocyst complex should be identified and expression differences between insect-stimulated and resting plants described.

The transportome Non-carnivorous plants mostly rely on soil reservoirs as supply for macronutrients. The complex root system especially the root hairs selectively absorb and transport the nutrients from the soil. On the contrary, carnivorous plants such as *D. muscipula* are able to absrob nutrients not only with their root system but also through the trap surface during the feeding process. Several studies with organic nitrogen and carbon suggested the dimorphic glands as responsible for the nutrient resorption [248, 176]. Nevertheless, the molecular landscape of the uptake machinery remained unknown in larger parts. The objective is to combine profile-based transmembrane annotations, homology-based transporter classification and differential expression testing to describe potential transport in both; the resting and the insect-feeding traps of *D. muscipula*. Transporters should be classified according to their target substrate and their site of action. The electrochemical gradient across the plasma membrane, maintained mostly by H^+ adenosine triphosphatases (ATPases), drives most of the transport processes. Highly abundant H^+ ATPases should be cataloged and checked for expression patterns dependent on insect-feeding. Additionally, components of the endocytotic machinery should be identified and monitored in a similar manner, since they provide an enzyme saving mechanisms to complement nutrient resorption through transporters and channels [6].

Trap-wide response to insect-feeding Multiple attempts have been made to describe "carnivory" related genes using transcriptome data. Most studies suffered from a weak experimental design (e.g., no biological replicates), hard to control conditions (e.g., feeding state of *U. gibba* traps) or monitored only a small number of genes through quantitative PCR techniques [40]. The objective is to fully characterize the transcriptomic landscape of an insect-feeding *D. muscipula* by RNA-seq using crickets as prey. Results should present the global expression patterns of carnivorous plant and broadly describe biological processes and molecular functions associated with the feeding process. Pathways associated with wound signaling should be characterized in-depth since previous results strongly suggest that JA plays a central role during insect-feeding of *D. muscipula* and other carnivorous species [203]. The objective is to identify core components of the alpha-linolenic acid cascade including associated transcriptional regulators and analyse their expression patterns. Additionally, trap-wide transcriptional regulators should be identified and broadly compared against those active in non-stimulated tissues.

Microbial aspects of the feeding process Plants exhibit a unique microbial community on their surface that often has a beneficial effect even though surfaces like leaves are extremely hostile. On the contrary, plants can also be target of organisms that cause infectious diseases including fungi, oomycetes, bacteria, viruses, and others. Interestingly, *D. muscipula* is rarely infected by microbes [278] even when the plant meets a completely new prey-associated microbiome during insect-feeding. The molecular basis for this immunity is still unknown as well as the fate of the microbial community after activation of the feeding process. Since microbial communities can be profiled easily from whole genome or whole transcriptome data from non-sterile host targets, resting and insect-feeding traps as well as glands should be metagenomically profiled and compared.

The sensory and regulatory capacities of traps Fast and precise transcriptional regulation of the secretion process is vital for *D. muscipula* to assess chemistry and quality of the prey. The objective aims to i) establish a classified kinome with special focus on receptor-like-kinases (RLKs) [255] that might be involved in chemical sensing and ii) analyse the transcription factor potential of both non-stimulated and insect-stimulated tissue.

Conserved response signals during insect-feeding Early molecular studies of carnivorous plants suggested several defense related processes to take part in the insect-feeding process. The transcriptomic profile of resting and insect-feeding *D. muscipula* traps should be compared to other carnivorous transcriptomes by directly looking at orthologous mappings (e.g., preferentially expressed genes (PEGs) from *U. gibba* traps) and against existing insect-treatments of non-carnivorous species using measurements of semantic similarity. Additionally, expanded protein families should be identified and insect-feeding specific expression patterns examined.

1.3 Material and Methods

1.3.1 Computing environment

All computational analysis were carried out on a compute cluster with 200 CPUs and a total of 1280 Gb of main memory distributed across 4 compute nodes. Data was stored on the submission node providing further 24 CPUs, 64 Gb of main memory and 137 Tb of storage capacities. Compute nodes and submission node were interconnected by a 10 GbE network. Cluster and submission node, storage and network system were assembled, setup and administered by the author of this thesis and Frank Förster during the first half of the thesis due to limited existing resources at the Department for Bioinformatics, University Würzburg. Compute and submission nodes were running Ubuntu 12.04.5 LTS (Precise Pangolin) at the time of the thesis. See section 1.4.1 for basic paths to raw data, important intermediate and final results.

1.3.2 Experimental procedures

Sequencing data presented in this thesis was generated by LGC Genomics GmbH, Ostendstraße 25, 12459 Berlin or GATC Biotech AG, Jakob-Stadler-Platz 7, 78467 Konstanz. Experimental methods as well as results concerning plant cultivation, RNA extraction, protein extraction as well as quantitative PCRs, kinetic and electro-physiological experiments that are cited or displayed in the thesis were implemented and produced by various members of the Department for Botany I, University of Würzburg, Julius-von-Sachs-Platz 2, 97082 Würzburg. See the final publication of the transcriptome data [25] for a detailed overview of the experimental methods (section Methods, subsections Plant growth, RNA extraction, sequencing, and qPCR, Proteomics, Electron microscopy, Electrophysiology) as well as for remarks on contributions (section Acknowledgments).

1.3.3 Data preparation and transcriptome assembly

Raw RNA-seq read data sets were screened for quality issues with FastQC (version 0.11.4) [13]. Passed data sets were quality trimmed using skewer (version 0.1.67; -Q 30; -q 30; -l 75; -m pe) [148]. The transcriptome was assembled using Trinity (release 2013-02-16; –jaccard_clip; –min_kmer_cov 2; –path_reinforcement_distance 75) [111]. The assembly was screened for artificial fusion events caused by low-complexity regions or highly similar UTRs. For transcripts with more than one potential coding region (see Feature annotation) a linkage map was constructed. Evidence from homology based database searches and mapped paired end reads were used to link the potential coding regions. Paired end mappings were produced with Bowtie (version 0.12.7) [179], while homology evidence was generated by searching transcripts against a plant-comprising subset of the Uniprot database (release 2014-06-19) [275] using blast (version 2.2.29) [48]. In case no linkage evidence was found, the transcript was considered to be fused. Fusion sites were detected by searching for the lowermost covered site between two coding regions. Transcripts were cut apart and the region with the lowermost coverage was trimmed from both resulting transcripts. Whenever a transcript was defused the original isoform-gene relation returned by Trinity was disregarded. To re-create a reliable isoform-gene relation, all defused isoforms derived from the same gene were partitioned using transitivity clustering (version 1.0) [298] into new genes. If sufficient evidence for a linkage was found the transcripts were considered intron containing. Introns were removed by aligning high scoring templates from the homology search to the transcript using genewise (version 2.4.1) [30]. Only proper alignments were used to cut introns from transcripts. Otherwise, transcripts were left untouched.

1.3.4 Feature annotation

Corrected transcripts were annotated using homology and profile based methods. Coding regions were identified using TransDecoder (release 2014-01-16) [118]. Putative peptides were preferred when they had a significant match to a Pfam domain. Only peptides with a length of at least 90 bp were considered. Protein families and domains were classified using InterProScan (release 44.0) [149, 92]. Phobius (version 1.01) [158], SignalP (version 4.0) [221] and TMHMM (version 2.0) [260] were integrated into the default signature recognition methods. Gene ontology (GO) terms where predicted using Blast2GO (version 2.5.2) [62]. GO terms where augmented with Interpro annotations using ANNEX [210]. Interspersed repeats and low complexity regions were identified using RepeatMasker (version 4.0.3) [258]. Putative orthologues between the Venus flytrap transcriptome and *A. thaliana* (TAIR10) [178] were assigned using a conservative conditional reciprocal best blast (CRB-BLAST; release 2015-05-19) [16] were assigned with a two-step approach. First, a CRB-BLAST against the TCDB [239] sequence set was used to identify conditional reciprocal best hits. If no distinct one-to-one hit was found, the sum of the best hits (e-value 1e-5) was evaluated. If all hits were annotated with the same TCDB family, the annotation was mapped to the Venus flytrap sequence. Putative peptides were further assigned to clusters of orthologous groups defined by the eggNOG databases and its command-line assignment tool eggnog-mapper (release 0.12.7) [137]. MapMan bins [227] were assigned using Mercator [284].

1.3.5 Transcript filtering

Prior to differential expression analyses and enrichment studies, the following filtering steps were applied to the refined assembly. (1) Non-coding RNAs were excluded by searching for isoforms without a potential coding region. (2) Ambiguous low abundance genes were excluded when no sequenced sample produced an expected count higher than 5. (3) In addition, ambiguous isoforms were removed by only considering isoforms with an abundance higher than 1 % of the abundance of its parental gene. (4) Possible contaminations were filtered with a two-step approach. Isoforms were split into overlapping k-Mers (k-Mer size 19bp). Each k-Mer was then searched in a database of trusted k-Mers created from genomic sequencing data (unpublished). Isoforms without a single k-Mer being present in the trusted k-Mer databases were disregarded. Remaining isoforms were searched against the complete non-redundant database (release 2015-01-13) [241] using blastn. Resulting hits were taken to calculate the lowest common ancestor (LCA) with MEGAN4 [140], using default settings. Isoforms with an LCA in bacteria, fungi or metazoa were disregarded in the following. (5) Transposon-like isoforms detected by RepeatMasker were excluded if they only contained the transposable element or a protein domain associated with interspersed repeats in the current RepBase databases (release 2014-04-2) [156].

1.3.6 Transcript abundance estimation and differential expression tests

Isoform and gene abundances were quantified with RSEM (version 1.2.5) [189] using unfiltered read data sets. The resulting count matrix for each experiment was normalized using the trimmed mean of M-values normalization method implemented in the DESeq package (version 1.22.0) [12]. Using a variance stabilized transformation of the normalized count data set possible outliers were detected using arrayQualityMetrics (version 3.26.0) [161]. The same expression data set was used for the principal component analysis. Differentially expressed genes were detected using DESeq. Only genes exhibiting an adjusted p-value equal or smaller than 0.01 were considered as significant. DEGs for *A. thaliana* microarray experiments (GSE48676, GSE49981, GSE5520, and GSE50526) were detected using GEO2R [19].

1.3.7 Enrichment analysis and comparison

GO enrichment studies were carried out using topGO (version 2.22.0) [9] whenever a raw or adjusted p-value was available for term eighting. The statistical measures were included using the weighted algorithm while fisher exact test was used as test statistic. Terms were considered significant with an FDR of 0.01. In case no p-values were available Ontologizer (version 2.1) [22] was used. Enrichments were calculated using the Parent-Child-Intersection of the terms and considered significant at an FDR of 0.01 after Bonferroni mutiple testing correction. Semantic similarity of GO enrichments was measured with the GNU R Bioconductor package GOSemSim [303] using Wang's measurement method and the best-match average (BMA) strategy to combine scores for individual terms. Gene set enrichments (GSEAs) were carried out with GAGE [192] using MapMap bins, TCs, GEM2Net [304] clusters or manually assigned gene classifications. DESeq derived adjusted p-values were used as per gene score, and gene sets were considered significant at a q-value equal or lower than 0.1.

1.3.8 Definition of the putative secretome, transportome and kinome

The secretome of *D. muscipula* was defined with the following rule set. Each putative member had to be a differentially expressed unigene when comparing resting with insect-stimulated glands (both down- and up-regulated DEGs were considered). The underlying unigene needed to have at least one differentially expressed isoform with an annotated signal peptide and evidence of a detectable peptide in the same open reading frame. The transportome was defined similar to the secretome. Putative members had to be a differentially expressed gene in resting or insect-stimulated glands. The unigene needed to have at least one differentially expressed isoform annotated with two or more transmembrane domains. Furthermore, the same sequence needed to have a proper TCDB classification. Transportome annotation and classification were manual refined by Dirk Becker, Department for Botany I, University of Würzburg. The putative kinome was defined as all unigenes with a properly annotated protein kinase domain (Pfam entry PF00069).

1.3.9 Metagenomic profiling and comparison

Raw RNA-seq read data sets were aligned to NCBI-nr [241] as a protein reference database using DIAMOND [42]. Reads were assigned to taxonomic bins using MEGAN-CE [139] and its implemented lowest common ancestor (LCA) algorithm. Reads were considered assigned at a minimum score of 50 and a maximum e-Value of 0.001. Normalized metagenomic profiles were exported and analyzed using STAMP [216]. A principal component analysis (PCA) was used to inspect sample clustering and the variance explained by each component. Post-hoc plots were used to inspect results of the multiple group statistic test carried out using ANOVA as test framework and the Tukey-Kramer method as post-hoc HSD (honest significant difference) test. The effect size was set to Eta-squared and p-values adjusted with Bonferroni correction.

1.4 Results

Note: The results presented in this chapter heavily build on experiments, data sets and analysis described in [25]. The following sections are based on the computational analysis carried out by the author of this thesis.

1.4.1 Data inventory

The following computational analysis builds on two raw data types, namely RNA-seq reads from high throughput sequencing experiments and protein quantificatios from high throughput proteome screenings. Experimental designs underlying each sequencing or screening experiments are given in each result section. See table 1.1 and 1.3 for basic output numbers and experiment identifiers used throughout this thesis. RNA-seq data was produced for various tissues as well as treatments over a period of four years. In total 8.3 billion reads (see table 1.1) were produced. All read data sets beside exp005, exp006 and exp007 were deposited in the sequence read archive (SRA) under BioProject PRJNA203407. Proteome data was generated for secreted mucilage and two membrane extracts from petioles and traps. Primary analysis was carried out by Prof. Dr. Waltraud Schulze (Department of Plant Systems Biology, University of Hohenheim, 70593 Stuttgart, Germany). Raw data was deposited in the EBI PRIDE archive under project PXD003480. A summary of each proteome quantification experiment is given in the respective result sections. RNA-seq data sets were used to generate the reference transcriptome and for transcript quantification while proteome data was used to detect translated transcript in a binary fashion. See sections 1.4.3, 1.4.4, 1.4.7 and 1.4.8 for results generated with the two data types. See data box 1.4.1 for paths to RNA-seq and proteomics data on the storage system (currently `wrzh089`).

Data: RNA-seq and HTS Proteomics

```
BASE=/storage/genomics/projects/dmuscipula/transcriptome
$BASE/data/illumina/exp00* # RNA-seq data
$BASE/proteomics/qt1.03/data/current # HTS Proteomics data
```

Table 1.1 Overview of sequencing experiments used during the analysis of the *D. muscipula* transcriptomic landscape. Experimental procedure are described in [25]. Single experiments are referenced using the experiment id (e.g., exp001) throughout this thesis.

ID	Tissue	Treatment	Replicate	SRA ID	Read Pairs
exp001	Petiole	none	1	SRR2807633	48,249,836
			2	SRR2807634	43,846,727
			3	SRR2807644	109,174,125
	Flower	none	1	SRR2807648	82,344,072
			2	SRR2807650	94,576,872
			3	SRR2807649	82,183,197
	Roots	none	1	SRR2807642	60,765,302
			2	SRR2807641	42,496,016
			3	SRR2807643	71,275,544
	Rim	none	1	SRR2807654	59,582,735
			2	SRR2807655	79,822,192
			3	SRR2807656	94,468,755
	Traps	none	1	SRR2807638	65,538,410
			2	SRR2807639	85,082,118
			3	SRR2807640	48,264,648
		COR	1	SRR2807651	68,174,440
			2	SRR2807652	76,033,562
			3	SRR2807653	51,132,019
	Glands	none	1	SRR2807635	61,104,656
			2	SRR2807636	46,414,468
			3	SRR2807637	67,264,172

ID	Tissue	Treatment	Replicate	SRA ID	Read Pairs
exp002	Pooled Tissues	Pooled Conditions	1	SRR2795277	91,624,980
exp003	Glands	none	1	SRR2807627	56,579,279
			2	SRR2807628	57,727,793
			3	SRR2807630	56,602,986
		COR	1	SRR2807629	51,821,988
			2	SRR2807632	53,260,935
			3	SRR2807631	60,636,940
exp004	Traps	none	1	SRR2807621	53,162,749
			2	SRR2807622	48,202,899
			3	SRR2807623	61,188,245
		Insect	1	SRR2807624	40,138,272
			2	SRR2807625	57,466,411
			3	SRR2807626	42,863,945
exp005	Traps	none	1	none	124,238,574
		COR	1	none	80,966,642
exp006	Glands	none	1	none	40,571,750
			2	none	42,518,329
			3	none	52,026,420
	Hair	none	1	none	49,708,826
			2	none	54,423,368
			3	none	45,625,187
exp007	Glands	none	1	none	72,139,439
			2	none	82,771,089
			3	none	43,308,712
	Hair	none	1	none	60,852,154
			2	none	62,450,736
			3	none	74,677,776
exp008	Glands	none	1	SRS1131959	179,188,714
			2	SRR2807657	176,247,961
			3	SRR2807659	176,206,340
		Insect	1	SRR2807660	198,188,982
			2	SRR2807662	195,144,623
			3	SRR2807662	164,505,395
				Sum	8,289,664,610

Table 1.3 Overview of proteome screening experiments used during the analysis of the *D. muscipula* transcriptomic landscape. Experimental procedure are described in [25]. Single experiments are referenced using the experiment id (e.g., pro001) throughout this thesis.

ID	Tissue / Sample	Condition / Treatment	PRIDE ID	Peptides
pro001	Mucilage	Mechanics, Hormone, Insect	PXD003480	1,392
pro002	Membranes	Trap, Petiole	none	4,381

1.4.2 Data quality control and trimming

Raw RNA-seq data was submitted to a thorough quality control. First quality metrics where generated and visually inspected (see figures 1.1, 1.2 and 1.3). None of the data sets showed serious issues that would have led to a total loss of the data set. Data quality varied according to data production time point (and most likely according to improvements of the sequencing chemistry and base calling performance; see figure 1.1). Duplication levels only showed expected issues for amplified sequencing experiments (e.g., exp005; see figure 1.2). Same was true for GC content levels (see figure 1.3). GC content was lower for amplified samples (e.g., exp005) and higher for Oligo(dT) primed samples (e.g., exp002). Quality-based trimming consequently displayed the same effects (see figure 1.4). Younger experiments (e.g., exp004) were less trimmed than older experiments (e.g., exp001). High and low GC content samples (e.g., exp008) showed less (reverse) reads surviving than experiments with an approximate Gaussian GC distribution probably pointing towards re-calibration issues for the 2nd read caused by the skimmed GC content. A subsequent contamination screen showed no dramatic levels of non-host reads. Experiments with an additional cDNA amplification step as well as insect-treatment experiments showed an elevated contamination rate up to 25 %. While the latter might point towards a prey-derived metagenomic signal the first is likely an effect of the amplification bias towards lower GC content (bacterial genomes are usually low in GC and are therefore more likely to be amplified).

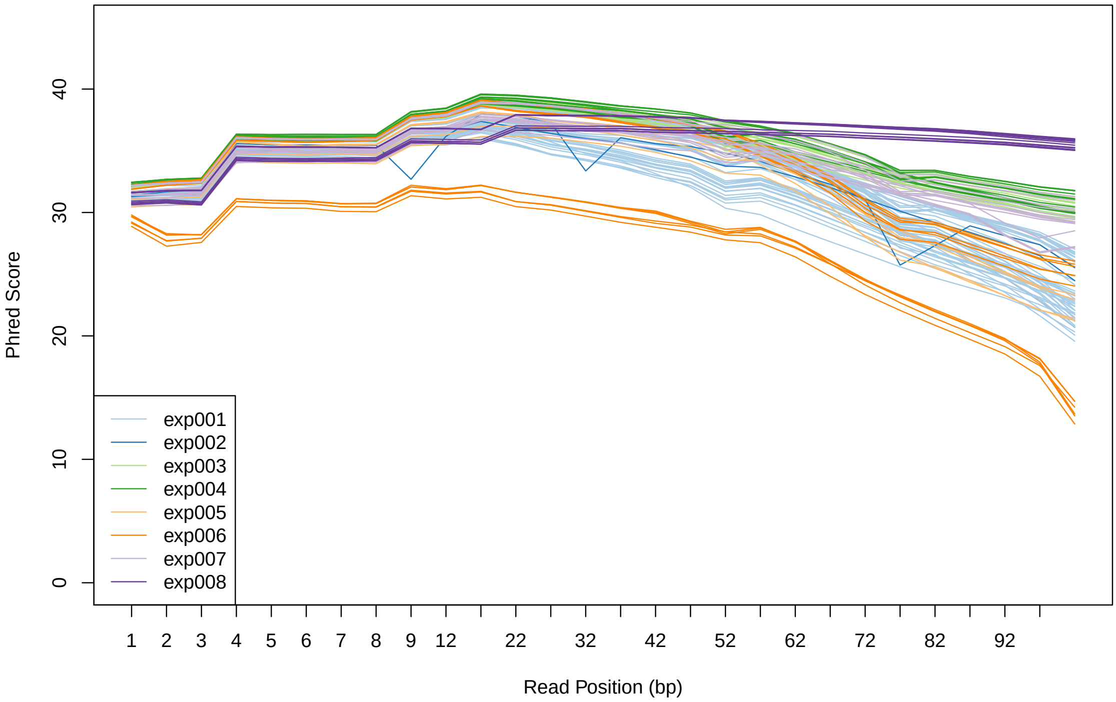


Fig. 1.1 Per sequence quality scores for all RNA-seq experiments used throughout this thesis. Quality scores are plotted per read position. Two obvious trends are visible. Experiments which involved cDNA amplification steps (e.g., exp005, exp006) show overall low quality which indicates a systematic problem. Second, quality scores behave according to generation date (exp001 < exp002 < exp003 < exp004 < exp005 < exp005 < exp006 < exp007 < exp008), respectively according to sequence chemistry and base calling software version used. Quality drops in exp002 are likely to occur due to re-calibration issues during sequencing.

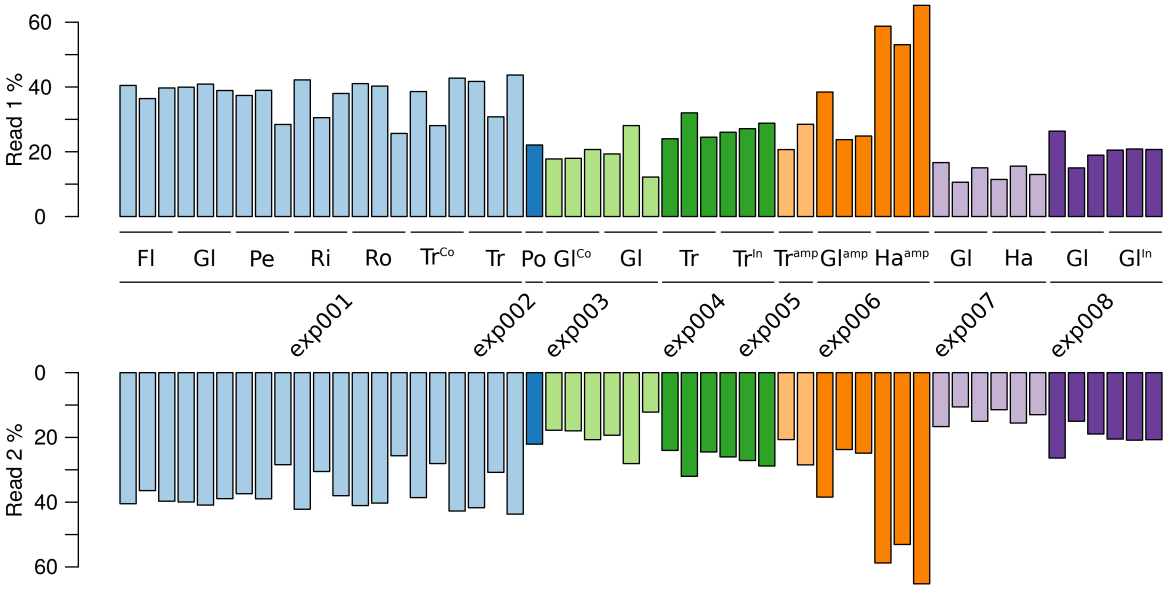


Fig. 1.2 Read duplication levels (in %) for all RNA-seq experiments used throughout this thesis. Most experiments show a level of read duplication typical for RNA-seq libraries. Since all RNA-seq libraries were over-sequenced to observe lowly expressed transcripts, high expressed transcripts potentially created the large set of duplicates displayed. Only samples from exp006, especially hair samples, show a dramatically higher duplication level likely caused by an amplification bias towards low GC content and thus resulting in a less complex sampled cDNA pool.

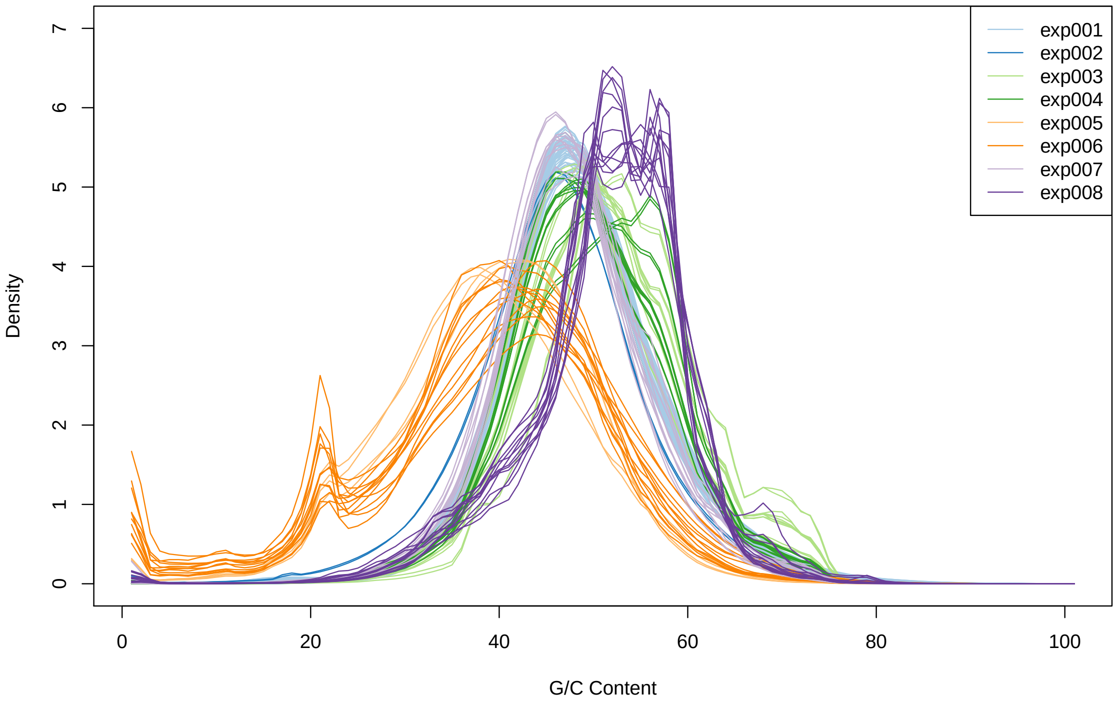


Fig. 1.3 Per sequence GC content across the whole length of each sequence for all RNA-seq experiments used throughout this thesis. Libraries exp001-004 show a normal distribution of GC content while experiments involving cDNA amplification steps are massively shifted towards a lower GC content and show further peaks at 20 % and 0 % GC content. Both, peaks at very low GC levels and the general left shift of the distribution point towards an amplification-related systematic bias. See section 1.4.6 for further details. The GC content of exp008 is shifted to the right showing a double peak. Mostly likely this indicates a contamination with substantial amounts of cDNA or DNA from microorganisms.

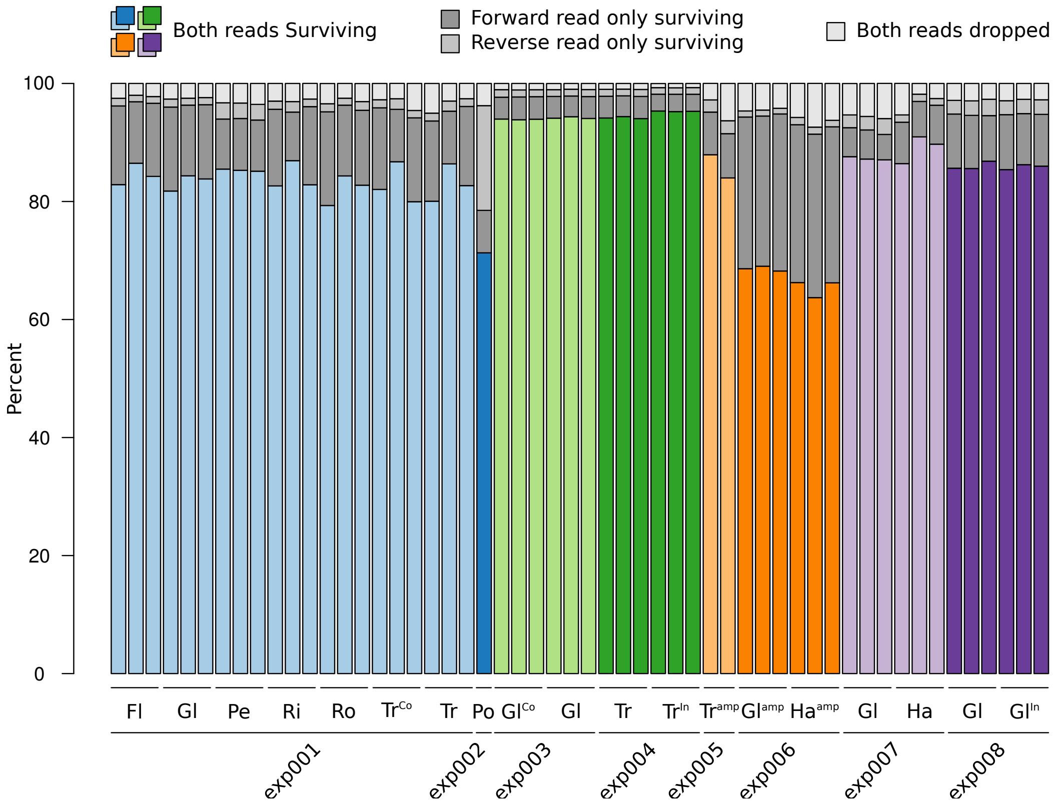


Fig. 1.4 Read trimming results for all RNA-seq experiments used throughout this thesis. Trimming results correspond roughly to per sequence quality scores in figure 1.1. Only exp002 and exp006 showed systematic problems with the 2nd read leading to a loss of about 30 % of sequencing data. Since base-calling can be influenced heavily by library complexity it is likely that the skimmed nucleotide distribution and complexity of the amplified (see figure 1.3) libraries are related to these quality issues.

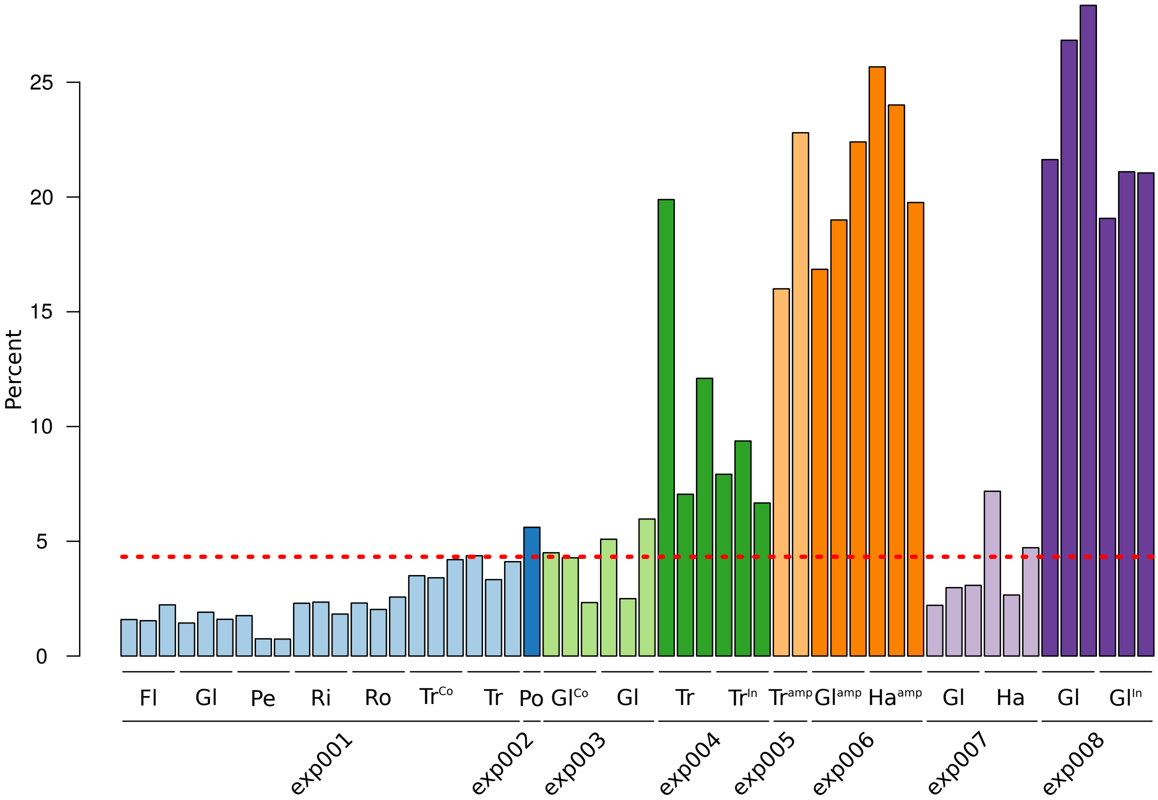


Fig. 1.5 Contamination screen for all RNA-seq experiments used throughout this thesis. Reads were compared against a subset of the RefSeq database containing complete bacterial/archaeal genomes using kraken [299]. Experiments with an additional cDNA amplification step as well as insect-treatment experiments show an elevated contamination rate up to 25 %. See section 1.4.12 for a detailed analysis.

1.4.3 The reference transcriptome

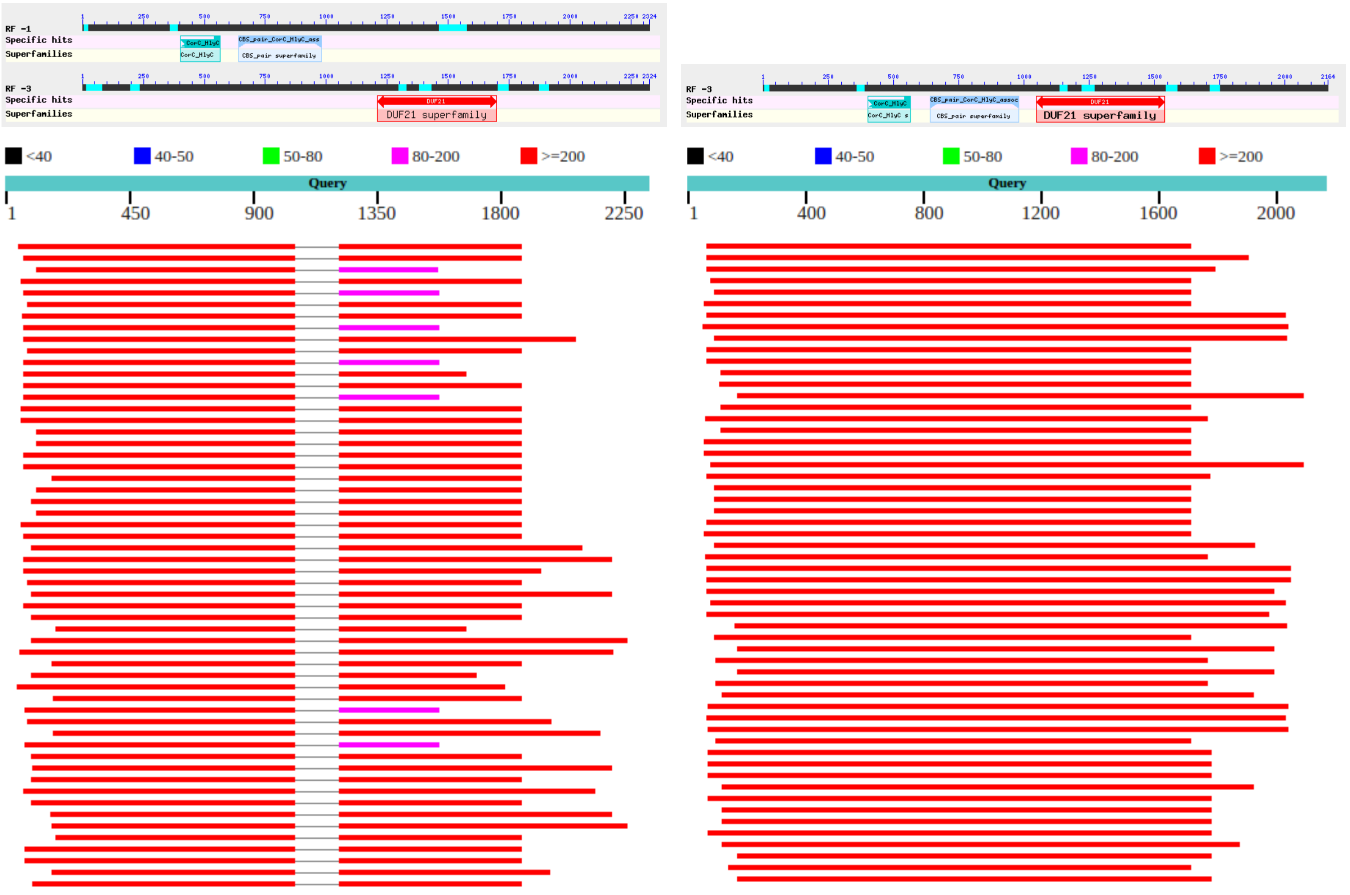
The primary object of this project was to generate a reliable reference transcriptome that consists of all detectable primary and alternative transcripts using RNA-seq data. Since different experimental data sets were produced over a longer time course of the project, often in a step-wise manner with later experiments building on results of previous ones, not all data sets were initially available to generate the reference. Data from exp001 was used to assembly the reference transcriptome including all major tissues without any treatment. Additionally, COR-treated traps were deeply sequenced, since the phytohormone triggers a) the slow trap closure and b) effectively mimics the hormonal stimulus to start the digestion process and c) maintain the activity of the digestion machinery thus representing a typical insect-feeding traps. The latter experiment was complemented with exp002 which was generated by pooling various RNA samples collected across all major tissues and target conditions (insect-treatment, COR-treatment). The reverse transcribed cDNA pool was experimentally normalized before sequencing to enrich for low abundance transcripts (see method box "Experimental cDNA Normalization" for a short experimental description). The data was used to test for the completeness of the reference transcriptome in a qualitative manner by simply asking if all produced sequencing reads effectively mapped back to the reference. See details below. Several transcriptome assembler were tested prior to the final transcriptome assembly, namely Trans-ABySS, Oasis (Velvet) and Trinity. Only Trinity was technically able to a) deal with the amount sequencing data from exp001 and b) handle heavily out-crossing species like *D. muscipula* with a high rate of heterozygosity. Oasis and Trans-AByss always terminated throughout the course of the assembly process due to memory overflow although all assemblers were run on a compute node providing 512 Gb of main memory.

Method: Experimental cDNA normalization

The procedure iteratively allows denaturation and hybridization of double-stranded cDNA molecules and the enzymatic removal of the double-stranded fraction at higher temperatures. Since the hybridization rate of a cDNA is proportional to its square concentration [252], highly abundant cDNAs will re-associate faster than low abundant cDNAs. High abundant cDNAs will be hydrolyzed more often than low abundant cDNAs and thus removed at a higher rate. The final pool contains cDNAs with a more similar abundance. The resulting cDNA library needs to be sequenced at a lower depth to reach the same complexity or vice versa achieves more complexity at the same depth.

The Assembly: The final reference assembly was started using 1,437,793,866 trimmed reads from exp001 covering 6 different tissues and COR-treated traps. The initial greedy assembly of Trinity with Inchworm resulted in a collection of 74,667,918 linear contigs that were drawn from the k-Mer graph. Chrysalis bundled the linear contigs into 212,382 pools roughly representing genes or highly similar paralogous gene families. Contigs were drawn together if they shared at least one (k-1)-Mer (24 bp) and a sufficient numbers of reads spanned the join. Resulting pools were used to build individual de Bruijn graphs. In the last step, Butterfly trimmed spurious edges and compacted linear paths. The final graph was reconciled with reads. Finally, Butterfly returned one linear sequence for each splice form and/or paralogous transcript. The final assembly comprised 345,803 isoforms with a total length of 376,689,236 bp. The transcript assembly had a N50 of 2133 bp and a N90 of 403 bp (NXX is defined as the shortest sequence length at XX% of the total length). The shortest sequence had a length of 201 bp while longest was 15,692 bp long. The full sequence set was clustered into 179,888 unigenes with a unigene consisting of two isoforms on average (see figure 1.8).

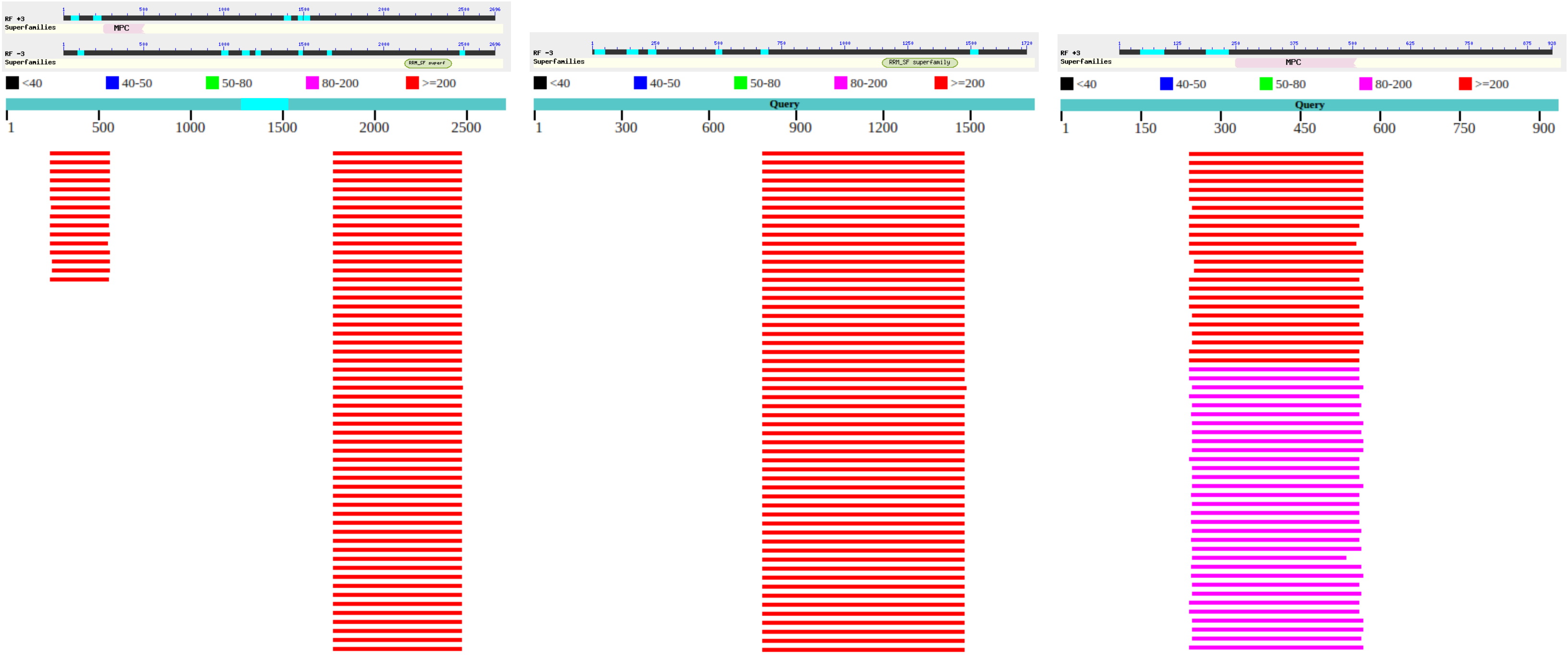
Post-assembly correction, filtering and validation: The final assembly was assessed for structural integrity by searching for artificial transcript fusions and retained intron sequences based on homologous sequences. Overall, 6,683 unigenes (respectively their corresponding isoforms) were identified to contain artificial fusions or retained intron sequences. Ambiguous unigenes contained 1,433 unigenes with one or multiple fusions, 4,100 unigenes with at least one intron-containing isoform and 1,150 unigenes showing both, fusions and retained introns. The fusion correction process was able to successfully split and rebuild 5,222 unigenes from 2,583 artificial fusions. For 3,584 isoforms (distributed over 508 unigenes) no coverage valley was found and the fused sequences were split in the middle of the fusion event. For 155 unigenes the cluster rebuilding process did not converge and the unigene was removed from the transcriptome. The intron correction process removed introns from 22,121 isoforms scattered over 3,383 unigenes. For 17,104 isoforms insufficient homology evidence was found and the sequences were not altered. The corrected the transcriptome sets comprised 315,584 isoforms and 183,578 unigenes with 2 isoforms per unigene on average (see figure 1.8) and a total length of 283,676,975 bp (N50 of 1,595 bp).



Raw Isoform

Corrected Isoform

Fig. 1.6 Example of an post-assembly intron correction. Left side: The raw isoform contains a intron and the two "exons" are in different reading frames. Right side: The intron and the frame shift were successfully removed and a complete isoform restored. Using this procedure 22,121 out of 39,305 detected introns were removed. The ratio correlates with the number of homology supported isoforms. Retained introns might be explained by immature transcript present in the final sequencing pool and likely be a consequence of the missing poly(A) enrichment during the library preparation of exp001.



Raw Isoform

Corrected Isoform 1

Corrected Isoform 2

Fig. 1.7 Example of a post-assembly fusion correction. Left side: The raw isoform contains two open reading frames not connected by homology evidence. Right side: The isoforms was successfully split based on the available homology and read mapping evidence. Using this procedure 5,222 unigenes were rebuild from 2,583 likely fused unigenes. In most cases highly repetitive elements such as the ribosomal binding domain (reading frame -3) likely caused the isoforms/unigene to be falsely assembled.

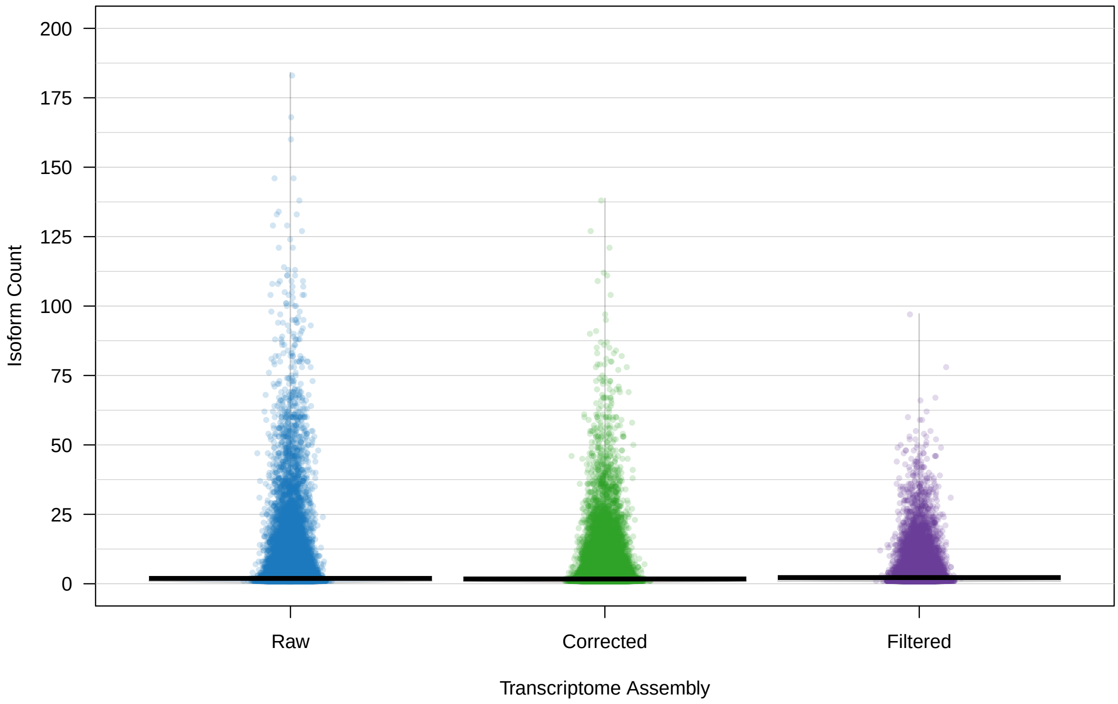


Fig. 1.8 Distribution of the number of isoforms per unigene after correction and filtering. Correction and filtering both reduced the number of isoforms. The correction process was able to remove isoforms during rebuilding of the defused unigenes when not all given isoforms contained the same fusion event. The filtering process reduced the number of isoforms based on their overall expression contribution and in case their were classified as contamination or TE-only isoform.

The corrected transcriptome was filtered with the 5 different approaches described in section 1.3.5. The following numbers represent redundant, overlapping counts of isoforms and unigenes flagged by each filter individually for exclusion. The ORF filter flagged 104,625 isoforms and 79,542 unigenes. The count-based isoform and unigene filter flagged 1179 isoforms and 1169 unigenes. The LCA-based contamination filter flagged 15,949 isoforms and 14,814 unigenes. The k-Mer based contamination filter flagged 20,186 isoforms and 19,277 unigenes. The TE-only filter removed 49,387 isoforms and 19,921 unigenes. The final RefSeq transcriptome contained 114,103 isoforms and 51,436 unigenes. Assessing the completeness of the final filtered transcriptome revealed that 89 % of all genes of BUSCOs plantae benchmark were present (see table 1.5). The correction process removed 1 % of the BUSCOs mainly due to fusion events. The filtering process removed 1 % of the BUSCOs mainly due to the count-based isoform and unigene filters applied. The final filtered transcriptome was further assessed using a homology-based method to determine isoforms sequence coverage and global identity in comparison to possible database homologs. Both sequence coverage (as base positions of the isoform covered by the alignment to a homologous sequence) and global identity (as the number of base positions identical between the query and the homolgous reference) increased during the correction and filtering process.

Table 1.5 Tabular overview of BUSCOs benchmarking results. BUSCO provides a quantitative measure of completeness for the three different transcriptomes. Completeness measurements are inferred based on evolutionary-informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB

Transcriptome	Completeness	Singletons	Duplicates	Fragments	Missing
Raw	91%	244	632 (66%)	30 (3.1%)	50 (5.2%)
Corrected	90%	286	583 (60%)	32 (3.3%)	55 (5.7%)
Filtered	89%	300	555 (58%)	26 (2.7%)	75 (7.8%)

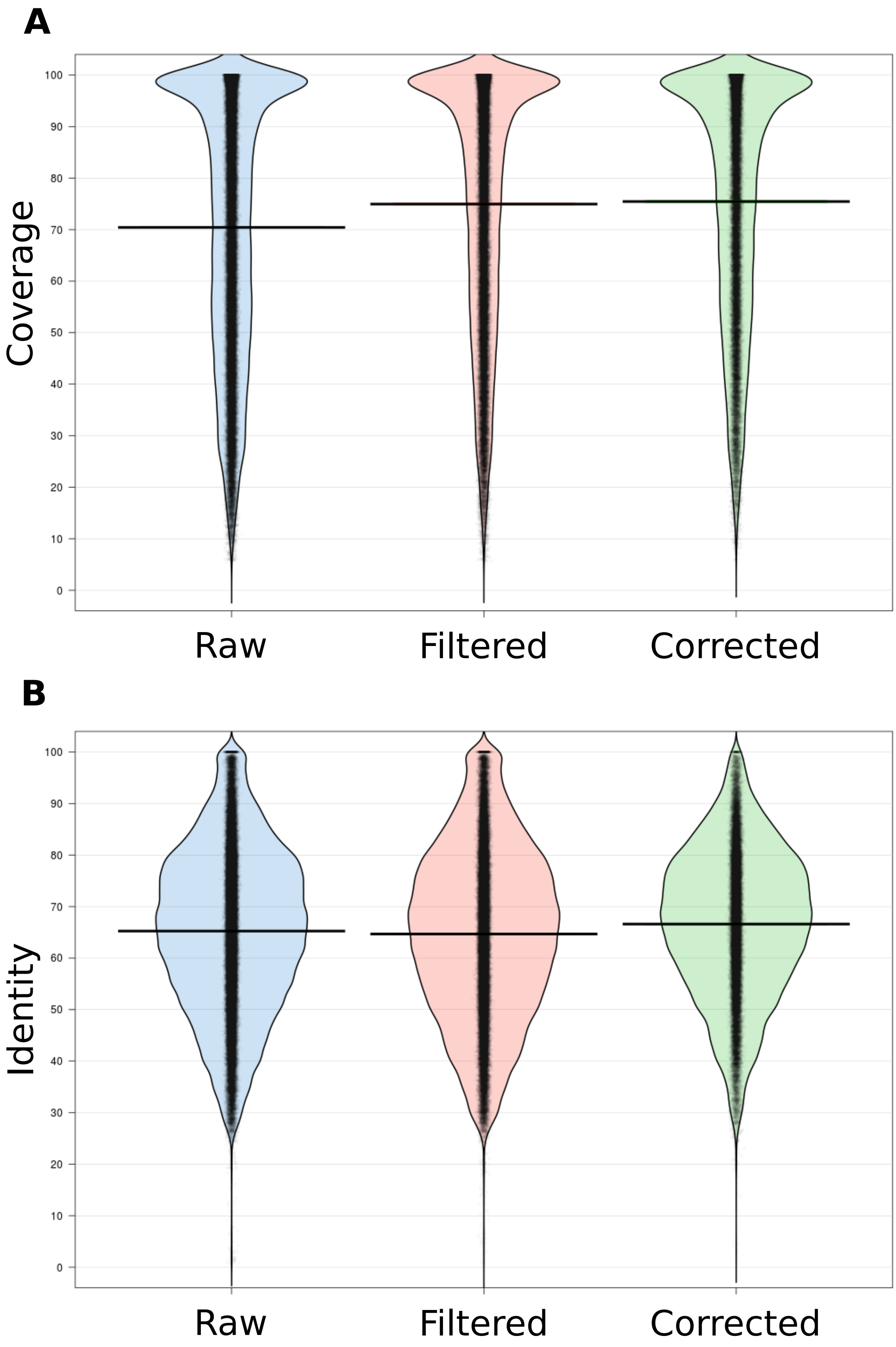


Fig. 1.9 Distribution of per sequence coverage and best-hit identity for the three different assembly stages. A) The per sequence coverage (as base positions of the isoform covered by the alignment to the best-hit homologous sequence) increased during the correction and filtering process. B) The global identity (as the number of base positions identical between the query and the best-hit homologous sequence in the alignment) drops slightly during the correction process probably due to introduction of new isoforms. The filtering process reverses this process and outperforms both the raw and the corrected transcriptome.

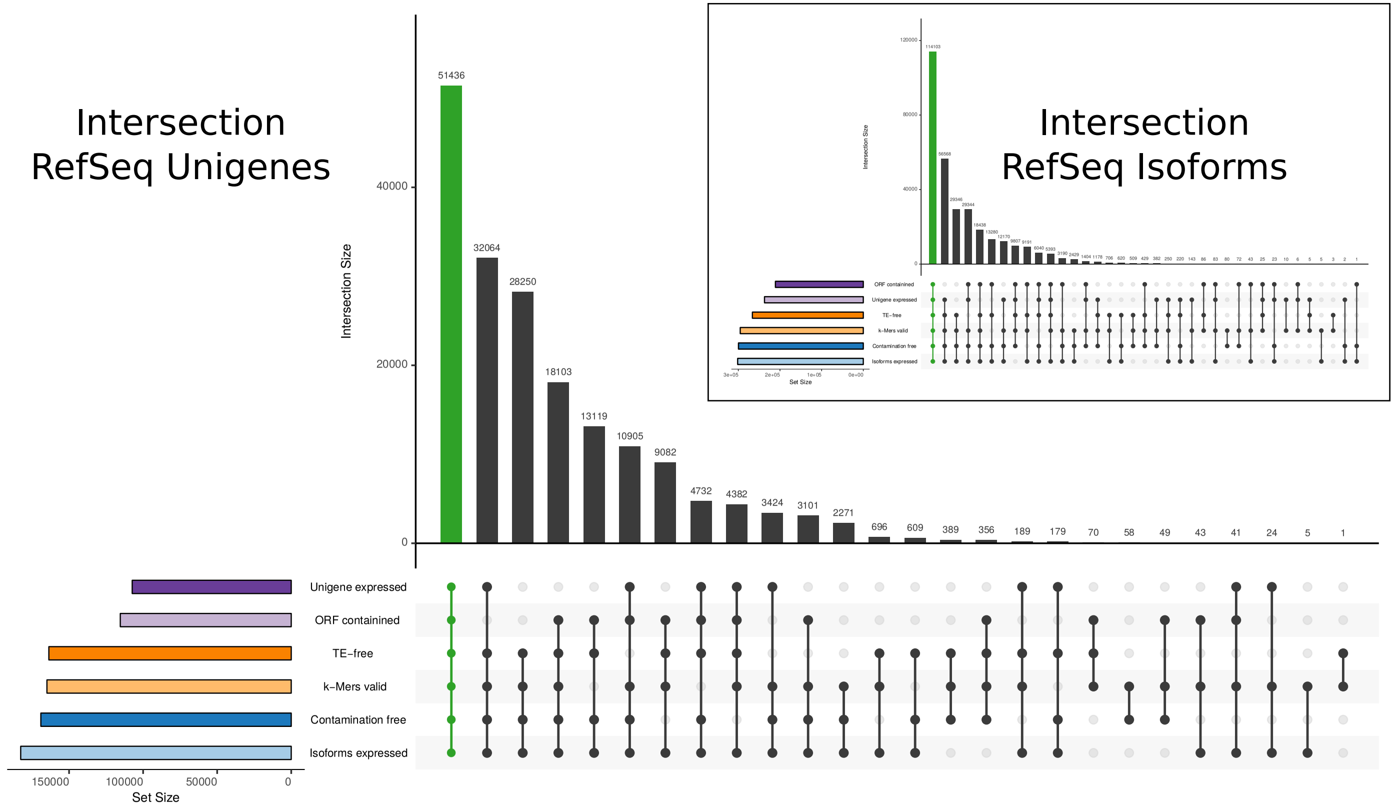


Fig. 1.10 Post-assembly filtering results based on multiple evidence tracks. Two classes of filters were applied. Sequence-driven filtering (e.g., ORF, k-Mer and Contamination filter) was carried out on isoform level while expression driven filtering was either carried out on isoform or unigene level. Full unigenes, including all isoforms were only removed when the unigenes had less than 5 counts. All other filters removed isoforms. If all isoforms of a unigene were removed the unigene was excluded automatically. The bar-plot on the lower left shows the size of individual inclusion lists while the plot on the right defines different logical subsets similar to the areas of a Venn diagram. The central histogram depicts the size if each interaction set. Green bars mark the final unigene and isoform reference size used throughout this thesis.

Transcriptome annotation: The corrected transcriptome was annotated using homology and profile based methods to predict conserved elements such as protein domains and motifs, assign functional classifications and predict evolutionary relationships. All annotation procedures were carried out on isoform sequences. Unigenes were assigned Gene Ontology terms, MapMan bins as well as functional classification (e.g., transcription factor class) by using a consensus decision across all individual isoform annotations for a given unigene.

Structural Annotation: Repetitive elements of individual isoforms were annotated with RepeatMasker. Overall 18,277,850 bp (6.44%) of the corrected transcriptome were masked as repetitive. Over 80% were removed during the filtering process. See table 1.6 for further details. Isoforms were further searched for possible open reading frames (ORFs). A reading frame had to be at least 90 bp long and protein domain containing ORFs were preferred over those without any domain annotation. The final transcriptome contained 29,638 unigenes with a complete ORF, 36,752 unigenes with only partial ORFs and 5290 unigenes that were nested into larger ones. The amount of unigenes dropped substantially during the filtering process after a slight increase due to the defusions happening during the correction. See table 1.7 for more details. The complete set of ORFs was used to search for conserved elements such as protein domains and motifs defined in Interpro. See results of InterProScan in table 1.8.

Table 1.6 Repeat annotation of the corrected and the filtered transcriptome. Numbers indicate isoform counts and covered base pairs (e.g., 41,025 isoforms / 13,430,716 bp covered). Dominating repetitive elements were retrotransposons and DNA transposons. 49,387 isoforms and 19,921 unigenes were purely annotated with repetitive elements and removed from the final transcriptome.

Type	Corrected	Filtered
Retroelements	41,025 / 13,430,716	5033 / 783,449
DNA transposons	8490 / 1,688,122	2904 / 400,302
Unclassified	868 / 187,046	360 / 56,235
Small RNA	919 / 130,241	90 / 6662
Satellites	268 / 21,419	86 / 5469
Simple repeats	27,873 / 1,138,420	53,627 / 1982.687
Low complexity	34,513 / 1,724,695	8923 / 444,617
Sum	113,956 / 18,320,659	71,023 / 3,679,421

Table 1.7 Open reading frame annotation results of the corrected and filtered transcriptome. Numbers indicate isoform and unigene counts (e.g., 108,414 isforms / 42,218 unigenes). The ORF counts increased during correction (not shown) as expected since new unigenes and isoforms are introduced. The filtering process removed 12,580 ORF containing unigenes.

Type	Corrected	Filtered
Complete ORFs	108,414 / 42,218	75,170 / 29,638
5-prime fragmented ORFs	66,798 / 42,033	35,601 / 19,680
3-prime fragmented ORFs	52,321 / 31,819	30,080 / 17,072
Internal ORFs	37,602 / 34,728	6332 / 5290

Table 1.8 Interpro annotation results based on the ORF complement of the corrected and filtered transcriptome. Numbers indicate isoforms and unigene counts (e.g., 15,314 isforms / 6545 unigenes). The amount of unigenes with at least one annoated isoform was substantially reduced during the filtering process mostly due to the removal of contamination and repetitive elements.

Source	Corrected	Filtered
Coils	15,314 / 6545	9620 / 3366
Gene3D	62,158 / 30,355	34,538 / 11,686
Hamap	1814 / 944	1223 / 548
PANTHER	83,106 / 38,887	45,509 / 146,961
Pfam	86,770 / 41,291	47,347 / 16,163
Phobius	57,763 / 28,605	37,446 / 17,035
PIRSF	2645 / 1320	1838 / 8128
PRINTS	10,042 / 5128	6121 / 2440
SignalP	18,337 / 9560	11,783 / 5685
SMART	21,539 / 8223	14,769 / 4971
SUPERFAMILY	66,585 / 32,152	35,960 / 12,085
TIGRFAM	6883 / 3010	5198 / 1956
TMHMM	38,464 / 19,676	24,881 / 11,670

Functional classification results Functional classes were initially assigned to isoforms and then lifted to parental unigenes in case the isoform complement of an unigene did not disagree. Gene ontology terms were assigned by combining blast-based homology hits and protein domain annotations using Blast2GO. The lifting process allowed terms to be different between isoforms when they had a common parent at the next parental node. Overall, 15,244 unigenes were assigned at least one gene ontology term. See table 1.9 for further details.

Table 1.9 Gene ontology annotation results. Numbers indicate isoforms and unigene counts (e.g., 65,076 isforms / 21,335 unigenes). The filtering process removed two thirds of the unigenes mostly due to the removal of contamination and repetitive elements.

Source	Corrected	Filtered
Blast	141,517 / 68,896	65,076 / 21,335
Generic GO	95,672 / 47,994	47,384 / 15,344
Plant GOSlim	95,672 / 47,994	47,384 / 15,344

MapMan ontology germs were mapped to isoforms using Mercator. The lifting process assigned 15,649 unigenes of the filtered transcriptome to 940 different MapMan ontology terms at 4 different levels. See table 1.10 for details. Transcription factor and transporter classes were mapped to isoforms using 1-to-1 orthologous relationships between isforms and the target databases PlantTFDB (transcription factors) and TCDB (transporter). The filtered transcriptome contained 1052 unigenes annotated as transcription factor and 1313 unigenes as transporter. See tables 1.11 for further details on the different transcription factor classes. The kinase complement was annotated using protein domain annotations and the sub-classification system provided by kinomer.

Table 1.10 MapMan annotation results. Numbers indicate isoforms and unigene counts (e.g., 95,669 isforms / 55,788 unigenes). The filtering process removed two thirds of the annotated unigenes mostly due to low expression and contamination flags. A crosscheck with known bacterial proteins confirmed that Mercators mapping procedure also assigns non-eukaryotic sequences especially to plastid-associated terms. The latter might explain why 2426 unigenes where assigned to a MapMan term but removed by the contamination filter. See table 1.13 for class assignments.

Level	Terms	Corrected	Filtered
1	23	95,669 / 55,788	43,334 / 15,649
2	233	95,616 / 55,762	43,296 / 15,638 /
3	450	14,677 / 5078	10,429 / 2452 /
4	223	4690 / 1995	3226 / 803 /

Table 1.11 Transcription factor classification results. Numbers indicate isoforms and unigene counts (e.g., 181 isforms / 85 unigenes). The most annotated classes were WRKYs, NACs and C3H transcription factors. WRKYs are involved in responses to biotic and abiotic stresses, seed dormancy, seed germination and senescence. NACs are involved in regulation of biotic and abiotic stress responses while C3H transcription factors usually encode Zinc finger (Znf) domains and have a large structural and functional diversity.

Class	Corrected	Filtered
AP2	181 / 85	150 / 79
ARF	302 / 73	251 / 67
ARR-B	621 / 214	524 / 172
B3	296 / 55	243 / 52
BBR-BPC	16 / 8	16 / 8
BES1	52 / 23	47 / 20
bHLH	414 / 109	318 / 91
bZIP	293 / 92	216 / 48
C2H2	507 / 249	340 / 160
C3H	2051 / 760	1443 / 428
CAMTA	500 / 135	394 / 104
CO-like	108 / 50	97 / 41
CPP	28 / 4	22 / 4
DBB	46 / 23	40 / 18
Dof	38 / 24	35 / 23
$E2F_{DP}$	16 / 5	12 / 5
EIL	16 / 3	11 / 3
ERF	181 / 85	150 / 79
FAR1	565 / 121	105 / 32
G2-like	540 / 168	468 / 150
GATA	145 / 65	126 / 51
GeBP	42 / 21	34 / 13
GRAS	109 / 30	93 / 28
GRF	55 / 14	43 / 13
HB-other	188 / 54	157 / 49
HB-PHD	299 / 104	239 / 87
HD-ZIP	196 / 69	166 / 58
HSF	28 / 18	23 / 14

Class	Corrected	Filtered
LBD	105 / 36	74 / 31
LFY	4 / 3	2 / 1
LSD	18 / 4	18 / 4
MIKC	139 / 55	127 / 48
M-type	127 / 57	117 / 50
MYB	951 / 333	754 / 254
$MYB_{related}$	730 / 230	604 / 183
NAC	1607 / 483	1510 / 473
NF-X1	5 / 2	5 / 2
NF-YA	16 / 3	16 / 3
NF-YB	46 / 24	36 / 15
NF-YC	46 / 24	36 / 15
Nin-like	148 / 37	106 / 24
RAV	396 / 129	327 / 120
S1Fa-like	2 / 2	1 / 1
SBP	78 / 19	75 / 19
SRS	13 / 7	12 / 6
STAT	7 / 4	5 / 3
TALE	61 / 23	56 / 19
TCP	41 / 19	35 / 16
Trihelix	152 / 57	121 / 43
Whirly	10 / 2	9 / 2
WOX	140 / 47	121 / 43
WRKY	2225 / 895	1732 / 668
YABBY	30 / 11	21 / 8
ZF-HD	12 / 10	12 / 10

Table 1.13 Kinase classification results. Numbers indicate isoforms and unigene counts (e.g., 242 isforms / 98 unigenes). Kinomer assigned 455 out of 1010 annotated kinases to 12 sub classes. The most frequent classes were tyrosine kinases (TKs), Ca^{2+}/calmodulin-dependent protein kinase (CAMKs) and CMGC kinases. The latter encodes MAPK growth- and stress-response kinases, cell cycle CDKs (cyclin dependent kinases) and other kinases involved in splicing and metabolic control. CAMKs and TKs can be further sub classified in other related families and cover a broad structural and functional spectrum.

Class	Corrected	Filtered
AGC	242 / 98	206 / 88
Alpha	101 / 41	93 / 39
CAMK	463 / 179	386 / 162
CK1	32 / 15	31 / 15
CMGC	353 / 133	292 / 126
PDHK	13 / 8	13 / 8
PIKK	44 / 14	33 / 14
RGC	83 / 24	61 / 21
RIO	28 / 10	25 / 10
STE	142 / 57	118 / 54
TK	656 / 230	526 / 217
TKL	234 / 96	196 / 90

Ortholog Mapping: Orthologous relationships to individual species or evolutionary conserved groups of proteins were detected with a conditional reciprocal Best BLAST against individual species (e.g., *A. thaliana*) or the eggnog-mapper. 13,907 unigenes were assigned an orthologous relationship in comparison to *A. thaliana* while only 1149 where assigned to one of the evolutionary conserved clusters defined by the eggNOG database (release 4.5). See table 1.14 and 1.15 for details.

Table 1.14 Ortholog mapping between *D. muscipula* and *A. thaliana*. Numbers indicate isoforms and unigene counts (e.g., 6252 isforms / 6110 unigenes). *D. muscipula* proteins with a 1:1 relation to *A. thaliana* are usually represented by unigenes with a single isoform. The reduction of 1:1 relationships is less substantial than the reduction of 1:n relations during the filtering process. The latter might be related to technical challenges during the ortholog detection. The most influencing negative factors might be the presence of unigenes with a high number of isoforms as consequence of the genetic heterogeneity in the RNA pool underlying the assembly and unresolved paralogous relationships rather than a biological signal caused by alternative splicing events.

Type (*D. muscipula* : *A. thaliana*)	Corrected	Filtered
1:1	6252 / 6110	5489 / 5368
1:n	40421 / 13747	28955 / 8539

Table 1.15 Assignments of *D. muscipula* proteins to evolutionary conserved clusters defined by the eggNOG database. The number of unigenes assigned to an individual NOG almost correspond to a 1:1 relationship suggesting only a small number of possible expansions in evolutionary conserved protein groups. The assignment was restricted to the filtered transcriptome and might be heavily biased due to removal of real *D. muscipula* transcripts during the filtering process. Especially the application of the low isoform usage or the low unigene expression filters might have introduced the unwanted bias.

	Covered Data sets	Covered NOGs	Isoforms	Unigenes
Assignments	47	1067	3038	1149
Expansions[1]	31	191	764	227

---

[1] Expansion are nested (e.g., comp234559_c0.0_seq28 is part of 0347W-artNOG is part of 0IK0C-euNOG)

1.4.4 Transcriptome quantification

Read mapping and transcript quantification Using the corrected transcriptome as reference, all read data sets were re-mapped. Re-mapping rates differed substantially between experiments and samples. Mapping rates (see figure 1.11) and contamination rates (see figure 1.5) were literally reversed indicating a sample-specific contamination especially in insect-treated and amplified samples. The high remapping rate for exp002 suggests a near complete mapping target and further demonstrates the high complexity of the corrected transcriptome since the data sets represents a pool of various tissues and treatments. The fragment length distribution for all experiments peaked around 150 bp and the mapping quality correlated with the Phred Quality Distribution of the reads (see figure 1.12). Transcripts were quantified based on the raw read mappings without noticeable problems.

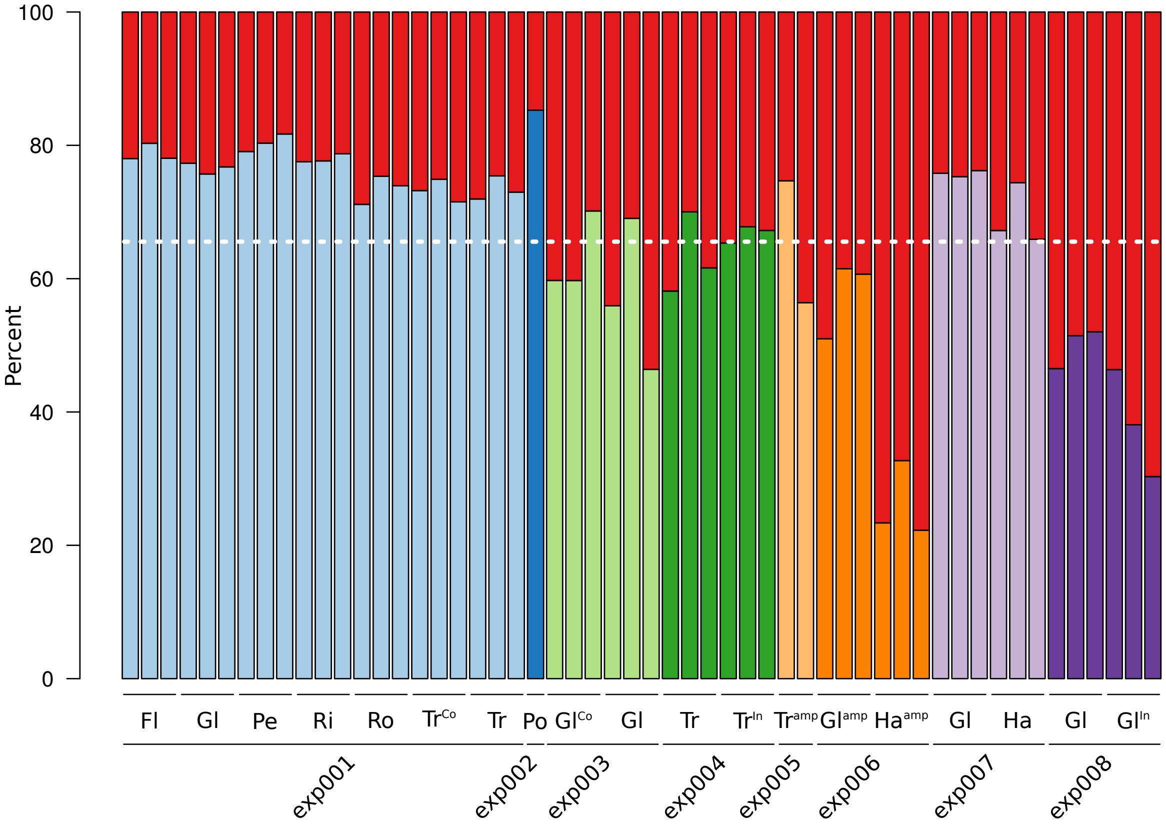


Fig. 1.11 Short read re-mapping rates of all RNA-seq experiments used throughout this thesis. Mappings rates efficiency develops independent of the tissues but according to the respective RNA-seq experiment. Data sets that have been produced from amplified RNA-seq pools tend to have lower re-mapping rates (e.g., exp006). Mapping rates roughly anti-correlate with the contamination rates (see figure 1.5).

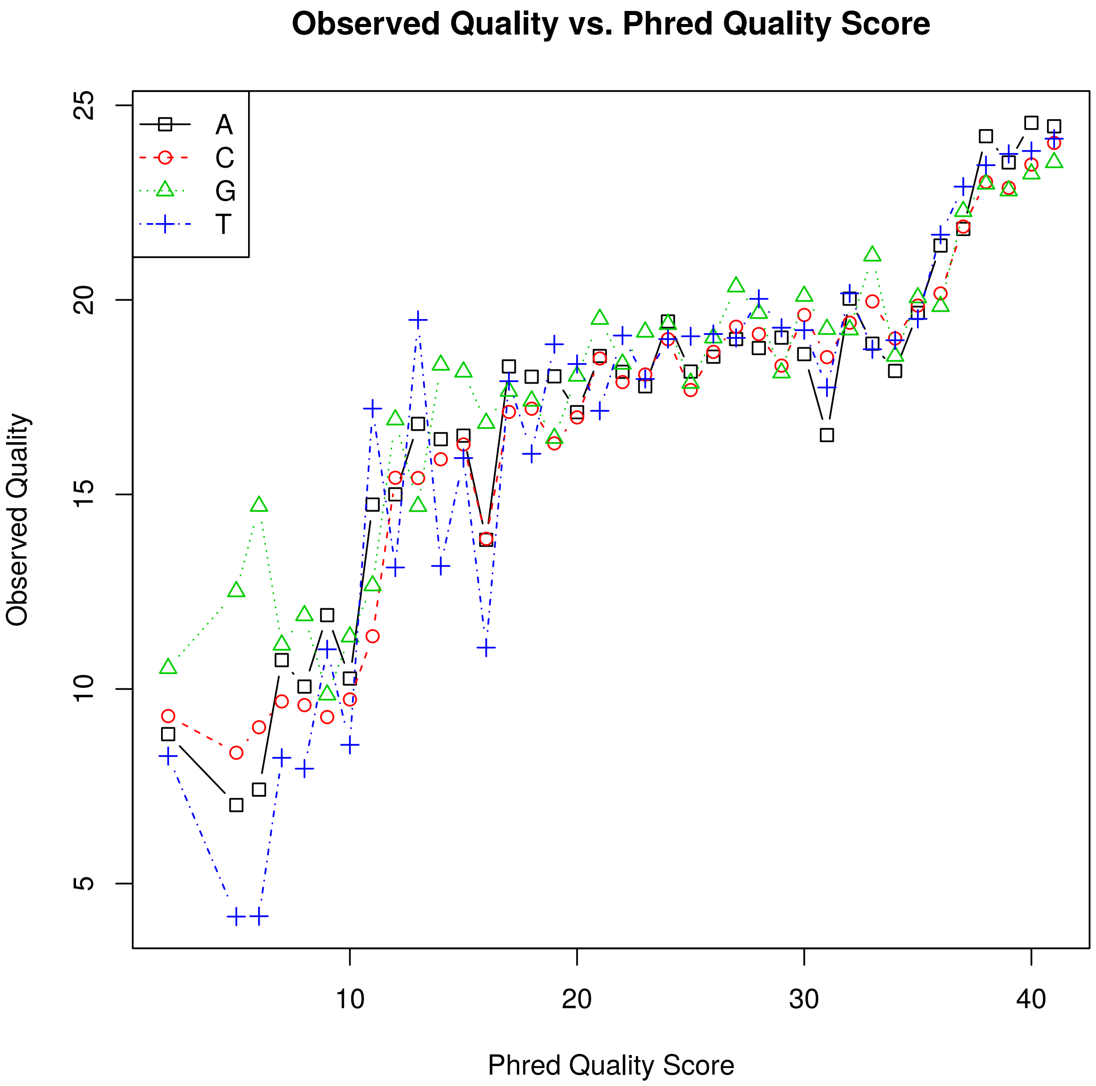


Fig. 1.12 Exemplary plot of the observed mapping quality vs. the Phred quality scores of the underlying short reads (exp001 Trap Replicate 1). Read qualities and mapping qualities are developing almost linear.

Normalization Quantification results were integrated and normalized per experiment. Quality control was carried out on blindly normalized expression data sets that were variance stabilized (see section 1.3.6). The actual transcriptome analysis was carried out using the conditional normalization procedure implemented in DESeq. All experiments (exp005 and exp006 are described in section 1.4.6) were tested for outliers and only two samples (exp001 Gland L2 and exp004 $Trap^{Insect}$ L1) were marked (during 1 out of 3 outlier tests). Conditions and tissues always clustered according to the intended experimental factor (e.g., tissues or treatment) and never show unintended causes such as batch effects. Experiments showed no noticeable problems in the standard deviation of the expression values (see figure 1.14). The running median of the standard deviation developed approximately horizontal, independent of the mean expression values and shows no substantial trend. Most experiments showed a hump on the right, normally indicating a saturation of the expression intensities in microarray data. In case of RNA-seq data the hump probably indicates incomplete saturation caused by a lower sequencing depth.

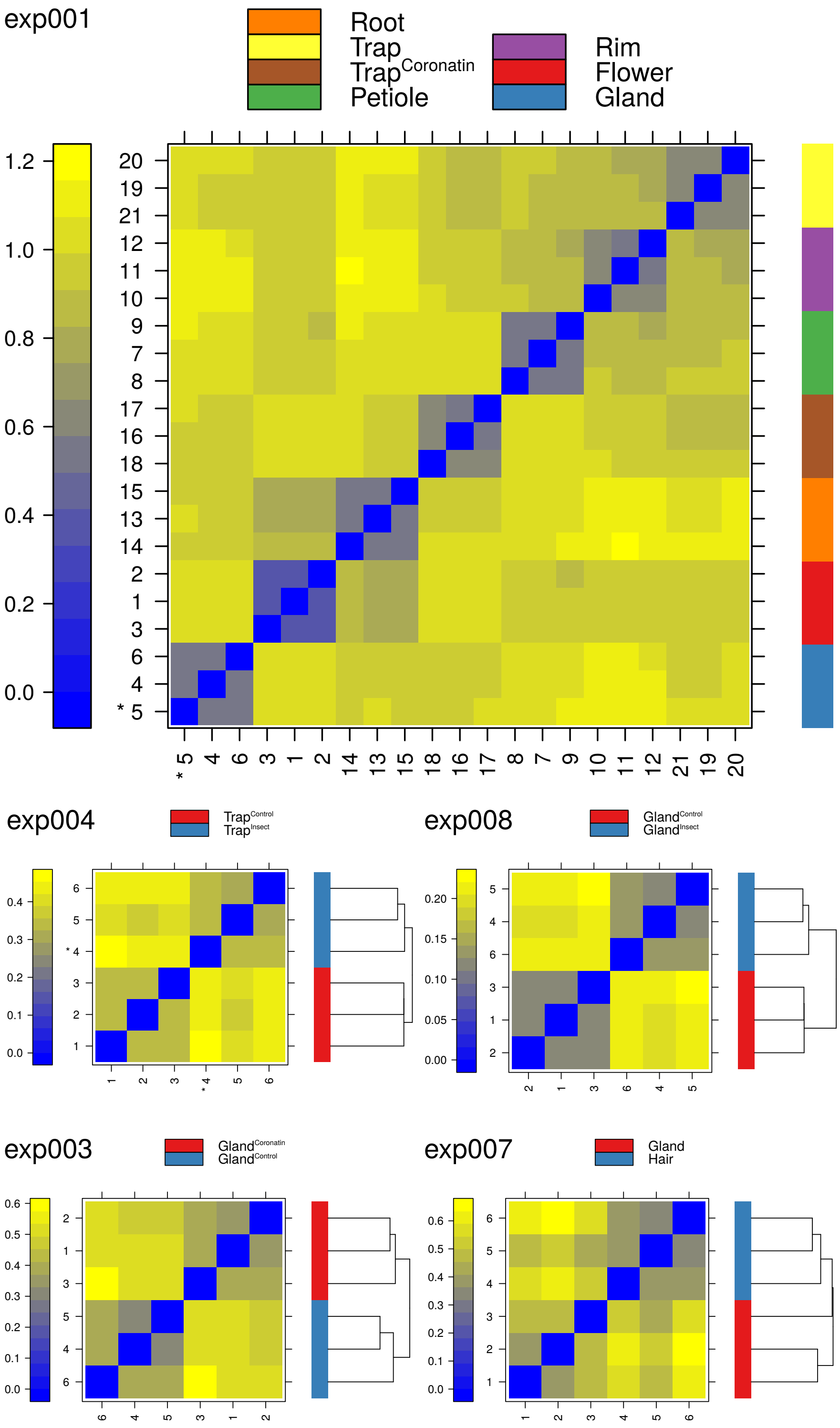


Fig. 1.13 False color heatmap of the distances between samples of all experiments. The scale is adjusted to the distance range observed in each experiment individually. The distance d_{ab} between two samples a and b was defined as $\delta ab = mean|M_{ai} - M_{bi}|$, where M_{ai} is the value of the i-th unigene of the a-th experiment. Samples were marked as outlier with asterisk, when their sum of the distances to all other samples, $S_a = \sum_b d_{ab}\, d_{ab}$ was exceptionally large. Two such arrays were detected (exp001 Gland L2 and exp004 $Trap^{Insect}$ L1). Nevertheless, all experiments showed a valid clustering of their samples in conditional groups or tissues.

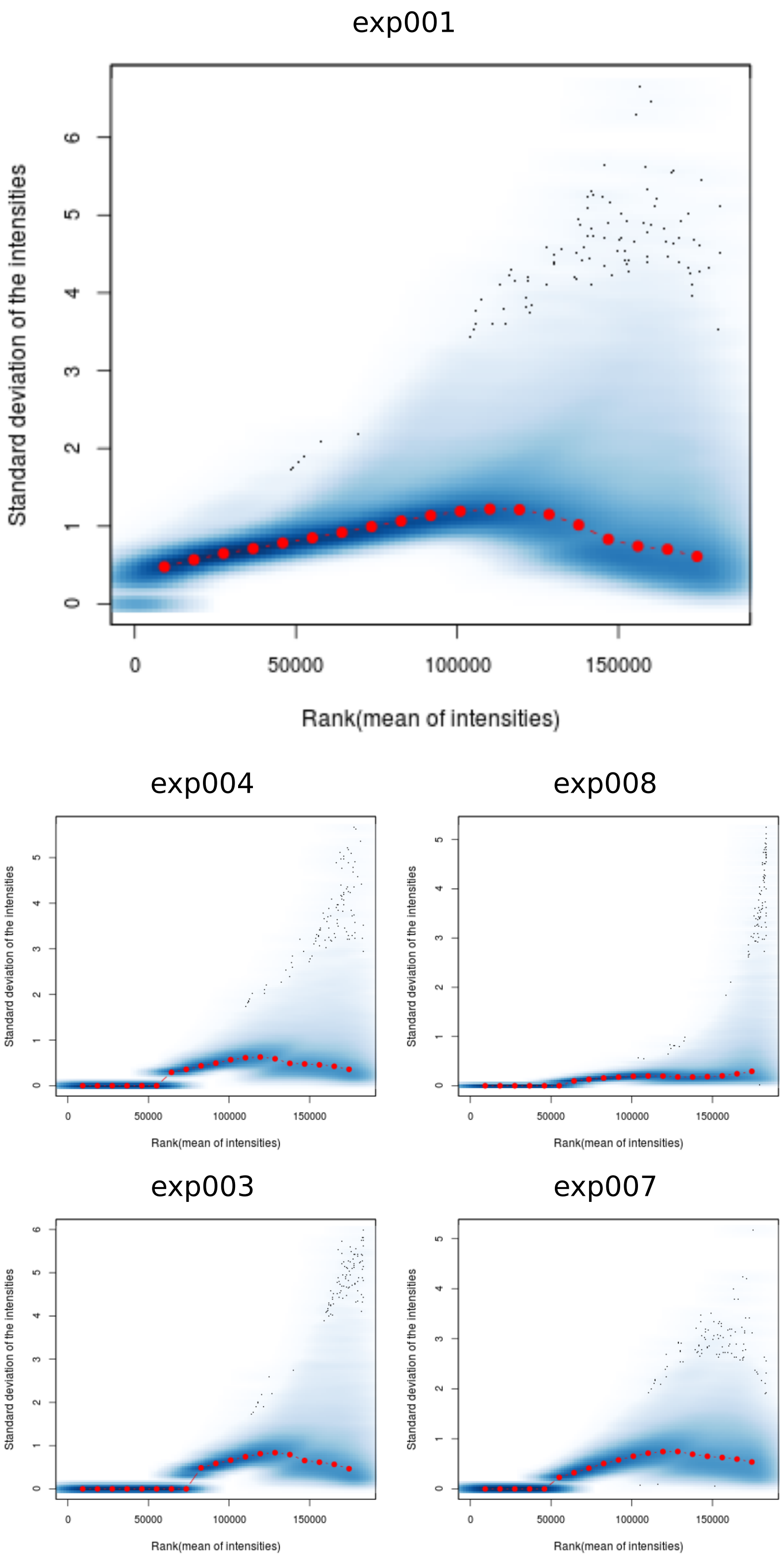


Fig. 1.14 Standard deviation of the expression counts across all RNA-seq experiments. All samples behave accordingly. The approximately horizontal lines indicate no substantial trend as typically expected. Humps on the right are caused by a slightly higher variability in the higher expression range. The trend might indicate a normalization bias at higher expression ranges and mimics the different sequencing depths.

Data: Annotation and Expression Results

Sequence annotations, expression quantifications and differential testing results (described in the previous sections) of the *D. muscipula* reference transcriptome are available through the Carnivorome transcriptome browser (http://tbro.carnivorome.org). Raw sequence data has been submitted to the NCBI BioProject (http://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA203407. The data is also accessible on the storage system (currently `wrzh089`).

```
BASE=/storage/genomics/projects/dmuscipula/transcriptome
$BASE/mapping # Read mappings and quantification
$BASE/annotation # Transcriptome annotation
```

1.4.5 The molecular make-up of a non-stimulated trap

The reference transcriptome was used to identify linearly uncorrelated variables (principal components) with a Principal Component Analysis (PCA) on variance-stabilized expression data. The first three principal components accounted for 72.46 % of variability in the reference transcriptome (see figure 1.15A and B). The first component explained 34.87 % of the variability and separated roots and flowers from petioles and traps. Genes positively correlated with roots and flowers showed cell cycle and protein modification processes enriched while petioles and traps displayed higher transcriptional activity in photosynthesis and translation (see table 1.17). The second component accounted for 22.08 % of variability and separated flowers from the rest of the tissues. Genes positively correlated with flowers showed a number of development and differentiation-related biological processes enriched while petioles, traps and roots showed signaling-related enrichments (see table 1.18). The third component represented 15.50 % of the overall variability and separated traps from the rest of the tissues. Traps were associated with responses to external stimuli and the generation of precursor metabolites and energy while the other tissues were not enriched for any particular biological processes (see table 1.19).

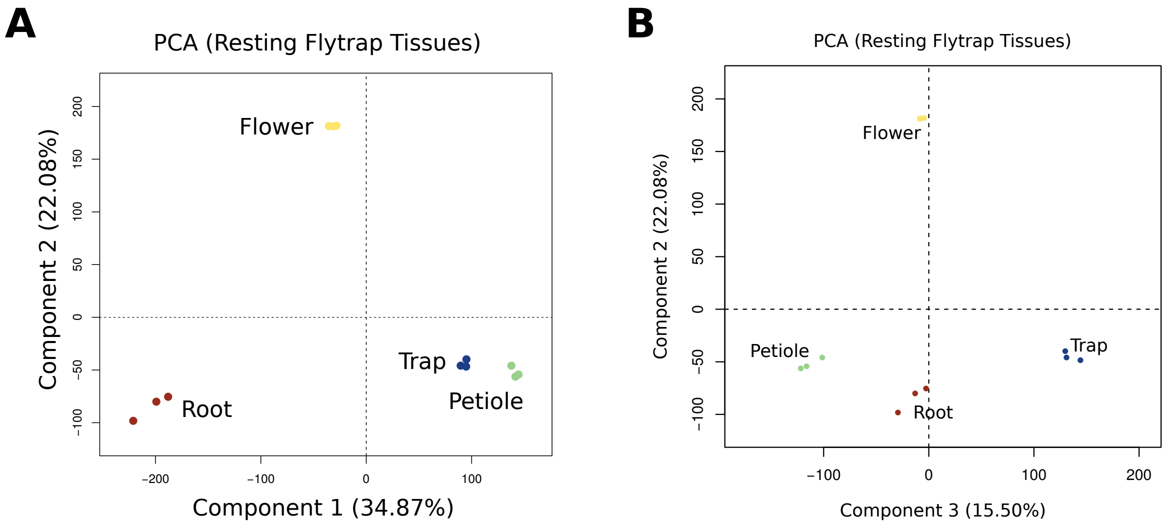


Fig. 1.15 (A) Principal component analysis of all biological replicates (n=3) from petiole, trap, root and flower. The first two dimensions account for 57 % of all the variance in the resting *D. muscipula* data. (B) Third component of the PCA accounting for 15.50 %

Table 1.17 Gene ontology enrichment result of genes significantly contributing to the 1st principal component of non-stimulated tissue. The analysis was run separately for genes positively and negatively correlated with the respective principal component.

Group	Term	Genes	DEG	adj. P-Value
Root & Flower	cell cycle	435	184	5.60×10^{-11}
	cellular protein modification process	1205	397	4.30×10^{-5}
Trap & Petiole	photosynthesis	370	138	1.16×10^{-22}
	translation	623	157	3.88×10^{-8}

Table 1.18 Gene ontology enrichment result of genes significantly contributing to the 2nd principal component of non stimulated tissue. The analysis was run separately for genes positively and negatively correlated with the respective principal component.

Group	Term	Genes	DEG	adj. P-Value
Flower	regulation of gene expression, epigeneti...	153	72	8.73×10^{-9}
	cell cycle	279	107	8.54×10^{-7}
	DNA metabolic process	303	113	2.04×10^{-6}
	cell differentiation	278	103	1.36×10^{-5}
	anatomical structure morphogenesis	489	161	3.88×10^{-5}
	cellular component organization	911	272	5.14×10^{-5}
	multicellular organismal development	962	294	2.72×10^{-3}
	flower development	281	95	3.88×10^{-3}
Others	response to extracellular stimulus	80	33	5.72×10^{-4}
	cell communication	567	127	2.62×10^{-3}

Table 1.19 Gene ontology enrichment result of genes significantly contributing to the 3rd principal component of non-stimulated tissue. The analysis was run separately for genes positively and negatively correlated with the respective principal component.

Group	Term	Genes	DEG	adj. P-Value
Traps	response to external stimulus	232	76	7.80×10^{-6}
	generation of precursor metabolites and ...	171	62	1.80×10^{-5}

The principal component analysis was complemented with a global all-versus-all pairwise differential expression test to further shed light on the commonalities underlying first threes observed variables. The DE tests between roots, traps, flowers and petioles revealed 17,715 out of the 51,196 genes as differentially expressed in at least one pairwise comparison (1.16). 14,089 DEGs were detected DE in more than one comparison. Intersecting the four sets of DEGs from each tissues showed 3626 organ-specific DEGs, with the highest number for the roots (1470) followed by the flowers (1113), petioles (655), and traps (388). Gene ontology enrichments for all four sets of organ-specific DEGs (those genes DE in only a single tissues/organ) showed only significant enrichments for flower-specific DEGs (see table 1.20). All others tissues lacked organ-specific enrichment signatures.

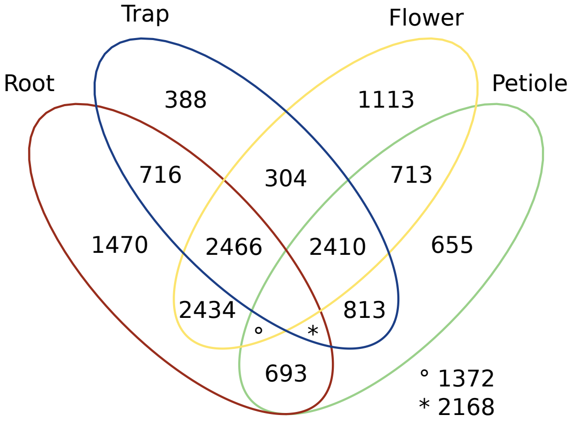


Fig. 1.16 Intersection analysis of all-vs-all differential expression testing results. Overall, 14,088 DEGs were shared by at least two tissues while 3626 DEGs are most likely expressed in a tissue-specific manner.

Table 1.20 Gene ontology enrichment result of DEGs only detected in the flower using the 4-way Venn analysis. Only genes with a base mean expression value of equal or larger than 100 were taken into account.

Term	Genes	DEG	adj. P-Value
anatomical structure morphogenesis	1,286	88	1.96×10^{-6}
cell growth	537	40	3.40×10^{-3}
cell cycle	676	46	8.24×10^{-3}

Since the majority of DEGs were frequently detected in more than one pairwise comparison the global relationship between non-stimulated tissues was further investigated with a correlation analysis on complete expression profiles. Traps correlated the most with petioles (see figure 1.17A; Pearson correlation coefficient 0.77, p-value 0), followed by flowers (Pearson correlation coefficient 0.6, p-value 0) and roots ((Pearson correlation coefficient 0.54, p-value 0)). Thus, the correlation analysis revealed textbook knowledge, but for the first time with molecular evidence, that the traps of *D. muscipula* are indeed modified leaves. Because the medium-ranged correlation coefficient between traps and leaves pointed to the existence of carnivorous signatures, a second correlation analysis was carried out. The expression profile of traps was replaced by two profiles from glands and rims (free of digesting glands but containing nectaries) to represent the heterogeneous nature of the different trap tissues better. The glands on the inner surface of the trap were previously identified as being responsable for prey digestion and the nutrient uptake (Juniper et al. 1989; Escalante-Perez et al. 2011) and thus, are a good candidate tissues for carnivorous signatures. The correlation analysis revealed rims and petioles as being the most strongly correlated expression profiles (see figure 1.17B; Pearson correlation coefficient 0.93, p-value 0) supporting the signals from the first correlation analysis. Interestingly, the 2nd most strongest correlation was found for roots and glands (see figure 1.17B; Pearson correlation coefficient 0.55, p-value 0), two sink tissues that have secretory capabilities [291]. Since strong functional signals are often encoded in DEGs, both tissues were tested against petioles. Glands revealed 7,427 DEGs when while roots revealed 6,897 DEGs when compared to petioles. Both DEG sets overlapped in 5,013 DEGs. Shared DEGs (and the associated p-values from the comparison petiole versus gland) were used as input for an gene ontology enrichment which revealed strong signals of transport, stress response and protein metabolic processes (see table 1.21).

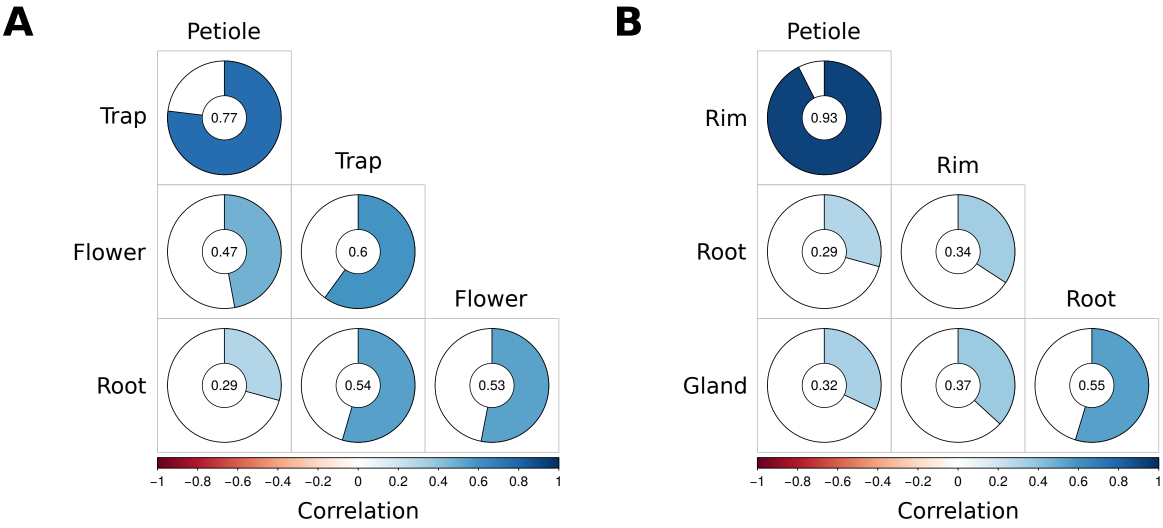


Fig. 1.17 (A) Hierarchically clustered visualization of the global Pearson correlation between all major tissues/organs. All individual pairwise correlations are significant according to multiple testing adjusted probabilities. (B) Visualization of the global Pearson correlation between all major tissues with traps being represented by non-glandular rims and glands. Again all correlations tested are significant.

Table 1.21 Gene ontology enrichment result of DEGs detected in the roots and glands alike. Only genes with an base mean expression value of equal or larger than 100 were taken into account.

Term	Genes	DEG	adj. P-Value
protein metabolic process	3361	1282	2.50×10^{-11}
transport	2416	929	3.40×10^{-7}
response to stress	2567	970	8.50×10^{-6}

A second enrichment was carried out to complement biological process aspects with molecular functions. To obtain the most fine-grained result the enrichment was conducted using lower level MapMan bins (level 3; Thimm et al. 2004) coupled with a generally applicable gene-set enrichment (GAGE; Luo et al. 2009a). The enrichment test revealed that DE gene-set shared between glands and roots accumulated genes involved transcriptional regulation via AP2, C2H2, WRKY and bHLH-like transcription factors, protein synthesis (ribosome biogenesis), protein modification and degradation, as well as protein targeting along the secretory pathway (see table 1.22).

Table 1.22 MapMan bin enrichment result of DEGs detected in glands and roots alike. Enrichment tests were carried out on level three assignments. Only genes with an base mean expression value of equal or larger than 100 were taken into account.

Term	Annotated	Significant	P-Value	Q-Value
stress / abiotic / heat	120	27	4.65×10^{-3}	2.01×10^{-2}
stress / abiotic / drought/salt	43	12	6.09×10^{-2}	8.97×10^{-2}
RNA / regulation of transcription / WRKY family	23	12	6.21×10^{-2}	8.97×10^{-2}
RNA / regulation of transcription / unclassified	115	56	2.29×10^{-6}	2.97×10^{-5}
RNA / regulation of transcription / putative transcription regulator	78	29	2.97×10^{-3}	1.54×10^{-2}
RNA / regulation of transcription / C2H2 zinc finger family	35	19	3.81×10^{-2}	7.07×10^{-2}
RNA / regulation of transcription / bHLH family	25	11	1.78×10^{-2}	4.23×10^{-2}
RNA / regulation of transcription / AP2/EREBP	22	13	6.91×10^{-3}	2.49×10^{-2}
protein / targeting / secretory pathway	95	42	1.74×10^{-5}	1.51×10^{-4}
protein / synthesis / ribosome biogenesis	79	28	7.67×10^{-3}	2.49×10^{-2}
protein / synthesis / ribosomal protein	646	92	5.01×10^{-5}	3.25×10^{-4}
protein / synthesis / initiation	87	20	1.96×10^{-2}	4.23×10^{-2}
protein / postranslational modification / kinase	62	23	1.93×10^{-2}	4.23×10^{-2}
protein / degradation / ubiquitin	345	145	1.75×10^{-12}	4.54×10^{-11}
not assigned / no ontology / pentatricopeptide (PPR) repeat	133	67	1.87×10^{-2}	4.23×10^{-2}

1.4.6 The trigger hairs

The trigger hairs of *D. muscipula* play a pivotal role during insect capturing [32]. It detects the insect movement and converts the mechanical stimulus into an electrical signal which in the following travels across the whole trap surface triggering the fast closure of the trap. Compared to the "rest" of the trap, the trigger hair comprises only a few hundred cells. To investigate the transcriptomic landscape of trigger hairs, two different RNA-Seq strategies were tested. The first strategy involved collection of small amounts of hairs from an individual plant as well as a control tissues (gland), amplification of extracted RNA (respectively cDNA) and subsequent transcriptome profiling (exp006). The procedure was also carried out with RNA isolated from whole traps and COR-stimulated traps using a single replicate (exp005). The second strategy involved collection of large amounts of trigger hairs from multiple isolates (which might be genetically different) as well as a control tissue (gland) followed by transcriptome profiling (exp007). Amplification and RNA-seq was carried out by LGC Genomics GmbH, Ostendstraße 25, 12459 Berlin. Each individual RNA-seq study was assessed standalone and compared to similar samples from the resting transcriptome profile (exp001; gland, trap and COR-stimulated trap). Quality of exp005 was not assessed due to missing replicates. Instead, it was directly compared with selected samples from exp001.

Amplification benchmark I The comparison of non-stimulated traps and glands as well as COR-stimulated traps (all three exp001) with amplified samples from trap and COR-stimulated (exp005) showed a significant batch effect (see figure 1.18 C). The first component of the PCA accounted for 31.15% of the variance and separated amplified from non-amplified data sets pointing again towards a dramatic unintended experimental effect. The second component separated COR-stimulated from non-stimulated samples and accounted for 21,6% of the overall variance. Sample distances of the two amplified samples to the rest were exceptionally larger than for the non-amplified samples (see figure 1.18 A,D). Additionally, the variance mean dependence showed a slight increase in the standard deviation of the variance for highly expressed genes (see figure 1.18 B).

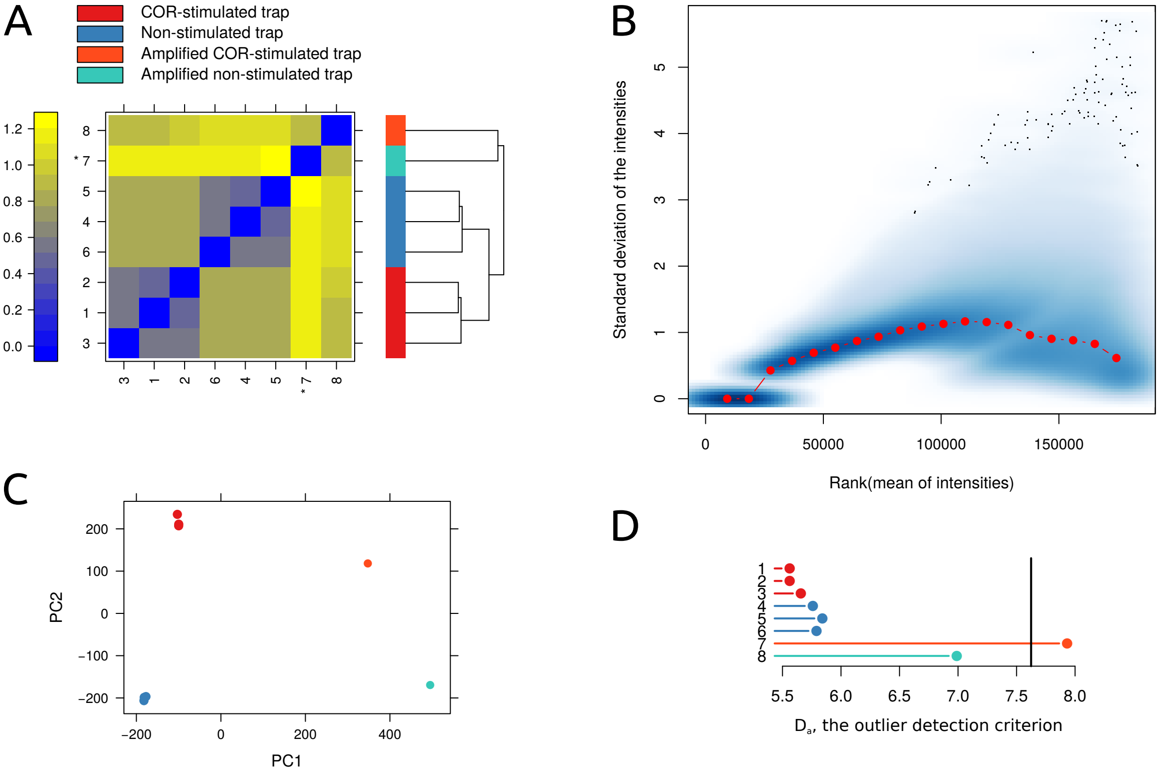


Fig. 1.18 Benchmark results from amplified trap samples. (A) The between sample distance heatmap showed an unintended clustering of amplified trap samples. Axis labels indicate simple sample numbering. (B) The variance mean dependency showed the typical hump on the right but with a non-horizontal trend indicating an amplification bias for highly expressed genes. (C) The Principal Component Analysis splits amplified and non-amplified samples apart and supports the trend already seen in the distance heatmap. (D) The outlier detection for distances between samples clearly showed the amplified samples to be more problematic and marked at least two of them as outlier.

Amplification benchmark II Comparison of amplified samples from glands and hairs with a) gland samples from the resting transcriptome profiling and b) with samples from pooled hairs and glands resulted in a similar picture as in traps. Amplified glands and hairs show a greater sample distance to the other samples and are marked as outlier (see figure 1.19 A,D). The PCA again showed a strong batch effect separating amplified hairs from the rest accounting for $27,84\%$ of the variance. The second component did not separate the samples according to condition/tissues but roughly according to experiments. Samples from pooled tissue samples (exp007) are much more scattered than samples from the resting transcriptome profiling (exp001, see figure 1.19 C). The variance mean dependence showed a dramatic increase in the standard deviation of the variance for highly expressed genes (see figure 1.19 B).

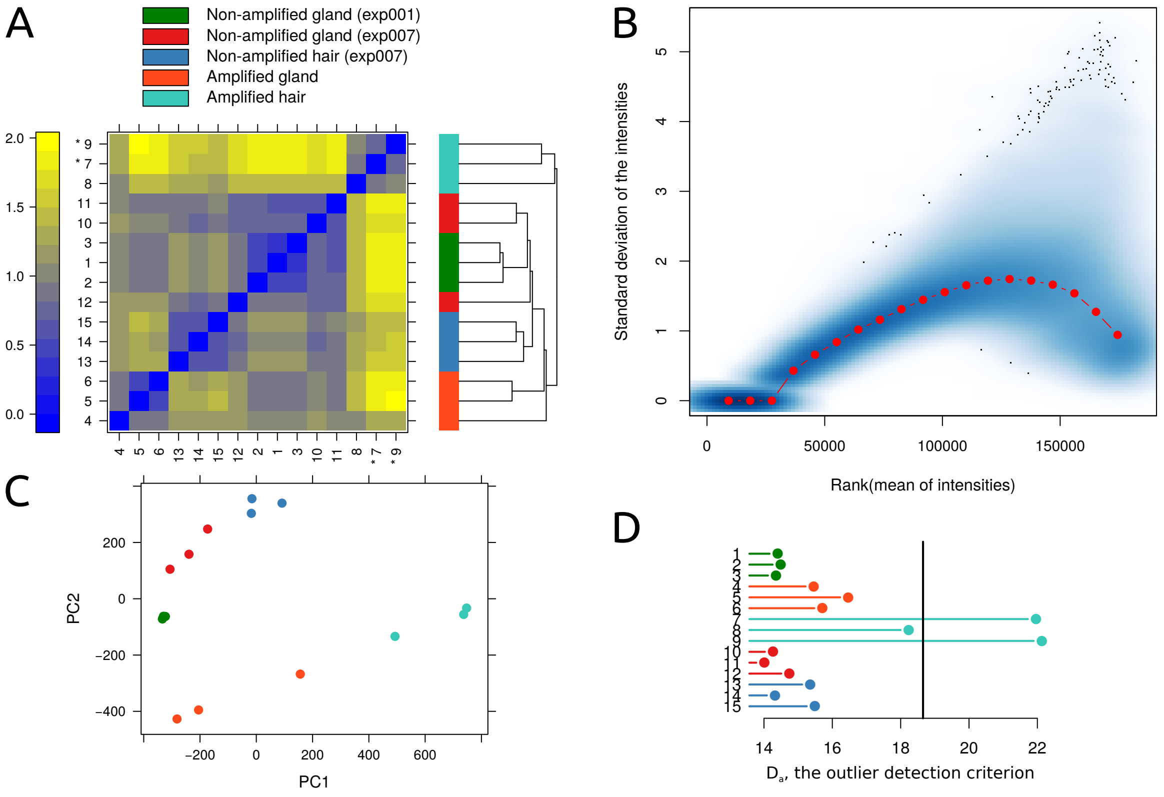


Fig. 1.19 Benchmark results from amplified gland and hair samples — (A) The between sample distance heatmap showed an unintended clustering of amplified hair and gland samples. (B) The variance mean dependency showed the typical hump on the right but with a non-horizontal trend indicating an amplification bias for highly expressed genes. (C) The Principal Component Analysis splits amplified and non-amplified samples apart and supports the trend already seen in the distance heatmap. (D) The outlier detection for distances between samples clearly showed the amplified samples to be more problematic and marked at least two of them as outlier (see sub figure A and D; samples 7 and 8).

Bias analysis All above mentioned expression data sets were used to assess a) skims in the nucleotide distribution and b) the ratios of 5 prime as well as 3 prime biases [250]. The GC content (see figure 1.3) was dramatically shifted towards lower GC content in all amplified data sets. CollectRnaSeqMetrics from the picard command line tools was used to produce a metrics describing the distribution of the bases within the transcripts and identify the median coverage depth as well as the ratios of 5 prime and 3 prime biases [36]. Especially, amplified data sets from exp005 and exp006 experienced a strong median 3 and 5 prime bias in the 1000 most highly expressed transcripts (data not shown).

Summary PCR amplification prior to RNA sequencing introduced strong batch effects when compared to non-amplified samples even when the RNA was extracted from a pool tissues/individuals. The strength of the effect is tissue-depended since glands are much weaker affected than hairs. On the contrary, samples from pooled tissues or individuals (exp007) are much more variant than samples from more homogeneous experiments (exp001). Since the amplifications bias is very strong, the amplification strategy was not considered suitable. All subsequent analysis of the trigger hairs are therefore carried out on the samples from pooled tissues (exp007) keeping the heterogeneity of the samples in mind.

Transcriptomic landscape of the trigger hairs The trigger hairs of *D. muscipula* are part of the sensory system that enables the plant to quickly close its traps. Nevertheless, the trigger hairs are not the only tissue that can be mechanically stimulated (e.g., glands) or is capable of generating action potentials. To better understand the transcriptomic markup of the trigger hair it was thus compared to trap rims, a tissue thought to be incapable to generate electrical impulses. Thus, rim might not express the components of the sensory system and is likely the most suitable tissue to be used as a testing background. RNA-seq profiles of hairs, traps, glands and petioles were tested against rim as background. The overall expression profile of trigger hairs correlated most with the expression profile of flowers (see figure 1.20).

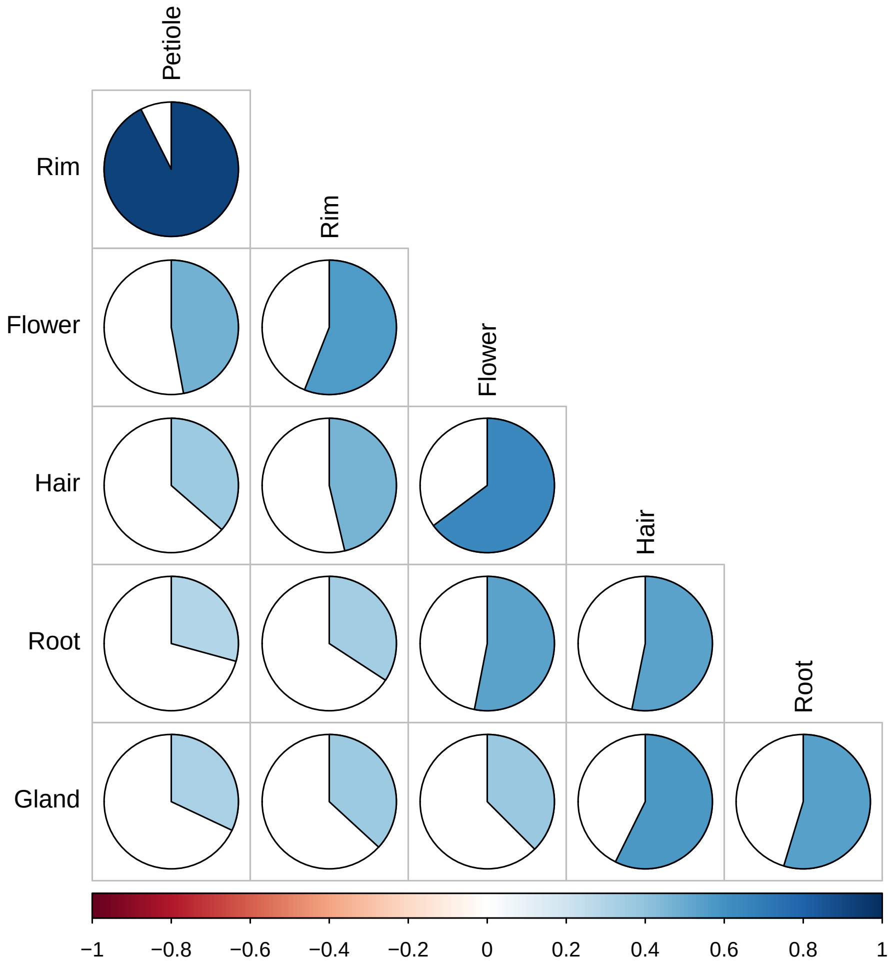


Fig. 1.20 Tissues-by-tissue correlation of non-stimulated tissues including hairs. Hierarchically clustered visualization of the global Pearson correlation between all major tissues or organs including hairs. All individual pairwise correlations are significant according to multiple testing adjusted probabilities.

Similar to the analysis of the resting reference transcriptome a Principal Component Analysis (PCA) on variance-stabilized expression data was carried out. The first three principal components accounted for 95.50 % compared to 72.46 % of variability in the reference transcriptome (see figure 1.15 A and B as well as 1.23). The first component explained 28.52 % of the variability and separated hairs from flowers, traps, rims and petioles which are separated from glands and roots. Genes positively correlated with roots, glands and flowers showed translation and transport processes enriched while petioles, rims, hairs and traps displayed processes involved in tropism, reproduction and development enriched. The second component separated photosynthetically active tissues from the rest while the third component separated trap tissues (trap, gland an hair) from non-trap organs with trap tissues being enriched for biological processes involved in translation, response to stress and transport.

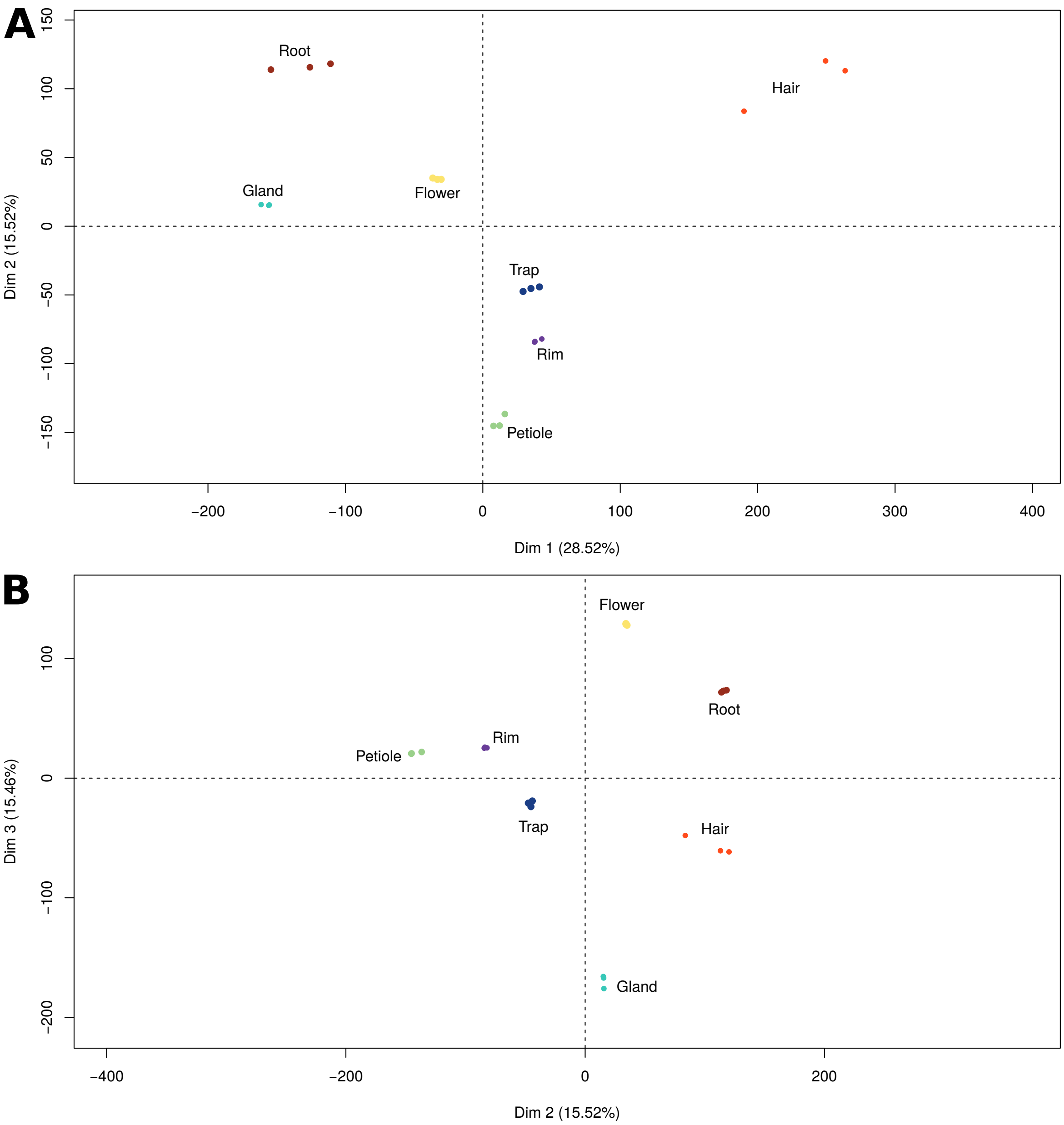


Fig. 1.21 (A) Principal component analysis of all biological replicates (n=3) from petiole, trap, root, flower, gland and hair. The first two dimensions accounted for 48.33 % of all the variance in the resting *D. muscipula* data. (B) The third component of the PCA accounted for 15.46 %

Trigger hairs showed 5,527 unigenes differentially expressed compared to rim. Of these, 1,363 overlapped with traps, 2,075 overlapped with glands and 673 overlapped with petioles. Enrichments were carried out on subsets defined by a Venn analysis (see figure 1.23). Unigenes differentially expressed only in hairs were mostly involved in developmental processes and cell differentiation (see table 1.23. Unigenes shared with traps and glands were involved in transport processes while unigenes shared with petioles were enriched for DNA metabolic processes (see table 1.23.

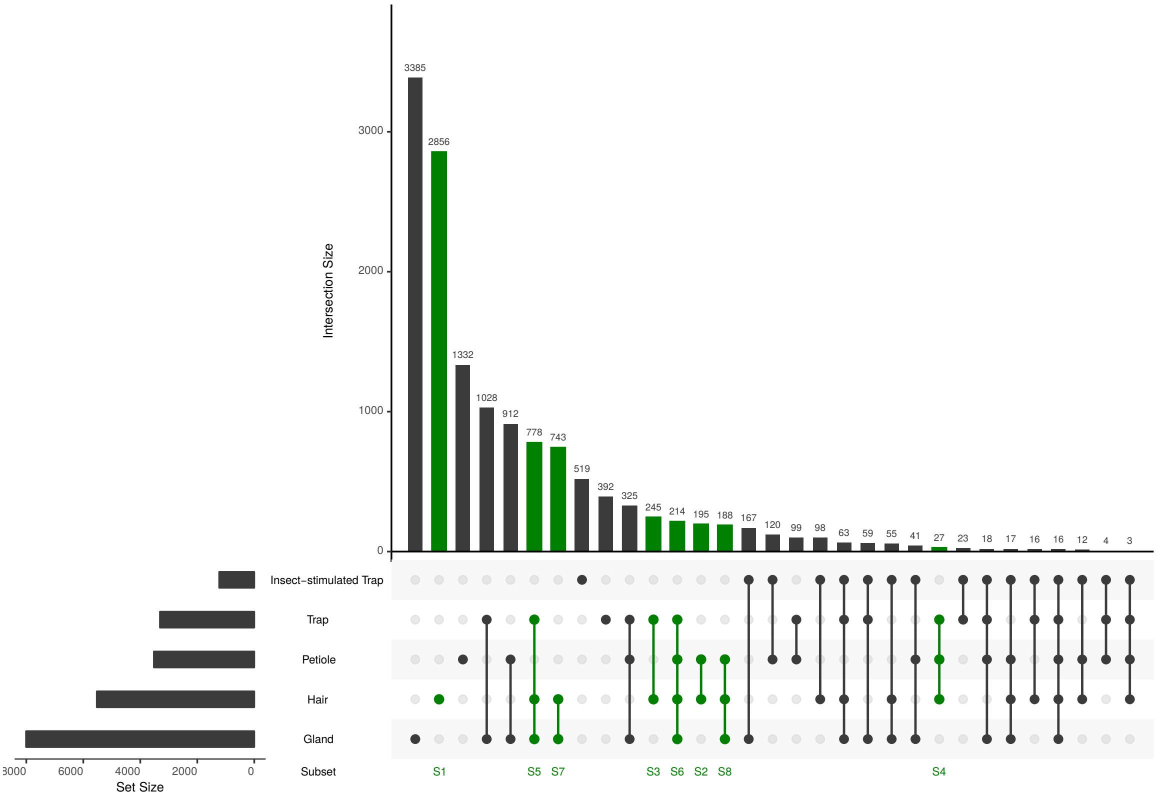


Fig. 1.22 Intersection analysis of DEGs from rim, hair and gland compared to petiole. Green connectors indicate subsets that include DEGs detected in hairs.

Table 1.23 Gene ontology enrichment results of DEG subsets derived from the intersection analysis. DEG results from a comparison between rim and petiole were removed prior to the enrichment. Only genes with a base mean expression value of equal or larger than 100 were taken into account. Subset ids correspond to ids in figure 1.22. S7 is not shown due to unavailability of the data at the time of submission.

Subset	Term	Genes	DEG	adj. P-Value
S1	tropism	140	41	1.3e-16
	regulation of gene expression, epigenetic	363	64	2.9e-13
	cell cycle	675	96	3.0e-13
	reproduction	1115	134	2.0e-12
	cellular protein modification process	1899	187	2.6e-09
	cellular component organization	2670	242	1.4e-08
	anatomical structure morphogenesis	1301	135	3.5e-08
	multicellular organismal development	2583	271	4.0e-08
	post-embryonic development	1588	169	1.4e-07
	nucleobase-containing compound metabolic process	3438	308	2.4e-07
	cell-cell signaling	79	19	6.2e-07
	signal transduction	1227	118	1.3e-05
	DNA metabolic process	793	83	1.5e-05
	embryo development	525	60	2.0e-05
	flower development	686	72	5.3e-05
	cell differentiation	830	83	7.8e-05
	growth	713	70	0.00055
S2	nucleobase-containing compound metabolic process	3438	47	6.6e-05
	cell cycle	675	15	0.00011
	DNA metabolic process	793	14	0.00188
S3	anatomical structure morphogenesis	1301	27	0.0012
	reproduction	1115	23	0.0031
	tropism	140	6	0.0050
S4	signal transduction	1227	6	0.0035
S5	transport	2487	145	5.1e-05
S6	cellular protein modification process	1899	37	3e-04
S8	DNA metabolic process	793	22	4.7e-06
	reproduction	1115	25	3.9e-05
	nucleobase-containing compound metabolic process	3438	59	0.00031
	cell cycle	675	15	0.00182
	post-embryonic development	1588	22	0.00570

In order to describe the nature of trigger hair DEGs (from a comparison to rim) and their enrichment signals more meaningful, a clustering relying on semantic similarity measures was applied (see section 1.4.9). A inspection of the treemap revealed that the terms "trichome morphogenesis", "gravitropism" and "protein desumoylation" successfully represented most of the raw GO terms. See section 1.4.9 for more detail on the tree mapping procedure.

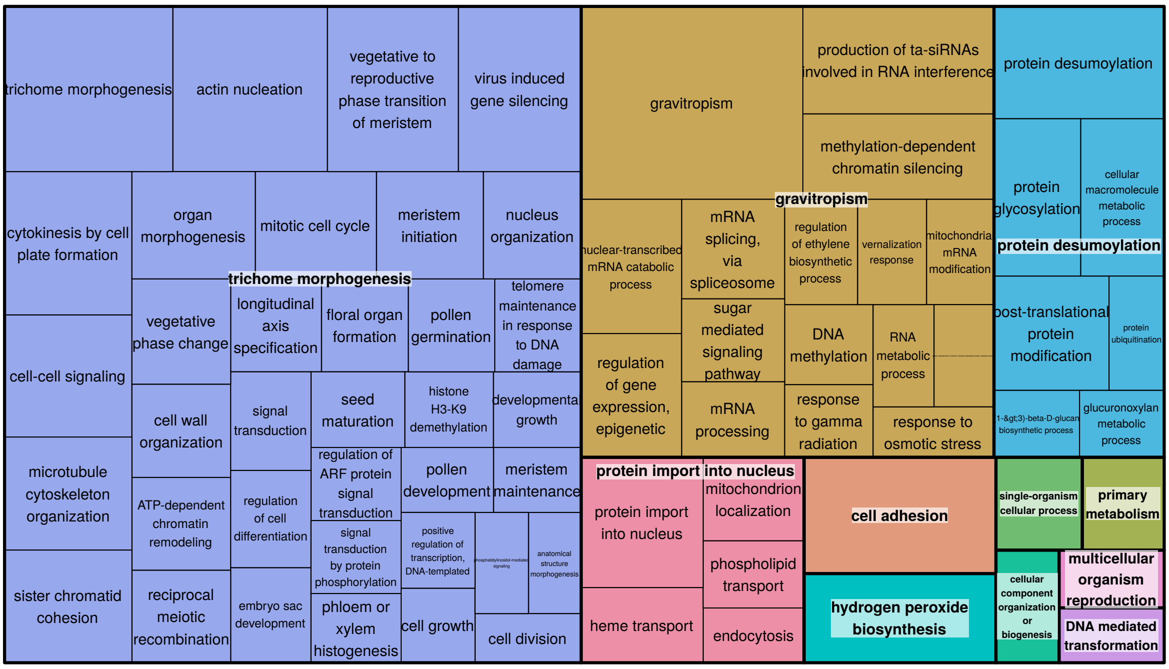


Fig. 1.23 Gene Ontology treemap of terms enriched in trigger hairs DEGs from a comparison to rims. The plot was drawn with Revigo [269] using Resniks normalized semantic similarity measures and a low similarity score of 0.4 to optimally group similar terms.

Mechano-sensitive potential Plants can survive mechanical stimulation and even use it to change their own architecture [222]. *D. muscipula* is using mechanical stimulation to to drive its trap closure. Especially the trigger hairs are able to couple a mechanical stimulus to an electrical signal that activates the trapping mechanism. Three groups of proteins are able to fulfill this biological functions, namely linkage proteins, mechanosensitive ion channels and some structural elements. A homology-based search for linkage proteins in particular NDR1-like integrins [171] resulted in a well annotated *D. muscipula* homolog (comp226242_c2_seq11) expressed mainly in trap, down-regulated by insect stimulation. All other hits contained no transmembrane domain and were not considered further. Expression in hairs was similar to glands but not as high as in whole traps. The second group of proteins linked to mechanosensitivity are mechanosensitive ion channels. A profile-based search for MscS-like channels, Mid1-complementing activity channels (MCA) [211], two-pore potassium (TPK) channels and PIEZO-like channels was conducted. Five TPKs (PF07885) were identified in the reference transcriptome, but none was differentially expressed in hairs and only one in glands (TPK1-like, comp220323_c1_seq4). Using the PLAC8 domain 14 MCA-like unigenes were identified. Three were differentially expressed in hairs (comp234198_c1.0_seq1, comp226866_c0.1_seq1, comp211686_c0_seq1) compared to rim but were also differentially expressed in glands and in parts during insect stimulation. Due to the collection procedure of the hairs gland contamination can not be excluded. Unigenes where hairs and glands both show differentially expression whereas gland exhibit a higher expression mean are considered contamination in hair samples. Using the MSL domain (PF00924) 6 MSL-like unigenes were identified. Three were differentially expressed in hairs. A MSL6-like unigene (comp215746_c0.0_seq1) was differentially expressed only in flowers, roots and hair with nearly no expression in the other tissues. A MSL10-like unigene (comp224798_c0_seq1) was exclusively expressed in hairs experiencing a 46 fold change compared to rim. The third unigenes was misjoint with a transcription factor sequence and discarded. Lastly, homology-based search for PIEZO1-like channels was conducted. A single significant hit identified a PIEZO1-like unigene (comp234339_c0.0_seq1) with Uniprot entry Q92508 as query. The channel was slightly up-regulated in hairs compared to rim but was in general ubiquitously expressed in all tissues and organs.

Signaling-related Kinases Recent studies suggested that plant mechanosensors could also continuously survey the mechanical status of the cell walls [56, 207, 138]. So far no cell-wall integrity sensors, as those existing in yeast [146, 186] were identified. Instead, plants might leverage their very large family of membrane-localized receptor-like kinases (RLKs) for this task. Potential candidate RLK cell-wall integrity sensors contain a cytosolic kinase domain, a single membrane-spanning domain, and an extracellular domain, putatively to bind carbohydrates [107, 135]. Using a homology-based search, the most promising candiates were identified in the reference transcriptome and expression evaluated. Five potential candidate RLK were found, two wall-associated kinases (WAK2 - comp232235_c0_seq2, WAK5 - comp230140_c0.0_seq1; [172]) with significantly higher expression in hairs and flowers, a THE1-like kinase (comp226143_c0.0_seq5; [126]) with higher expression in roots and flowers, a FER-like kinase (comp234198_c0.0_seq4; [77]) with high expression in roots, flowers and hairs as well as a HERK1-like kinase (comp234459_c0.1_seq1, [117]) with high expression in all tissues and slight up-regulation in hairs. An ANX1 or ANX2 homolog was not identified. None of the kinases showed a hair-exclusive expression pattern or a complete absence in rims.

1.4.7 Secretion-response to insect feeding

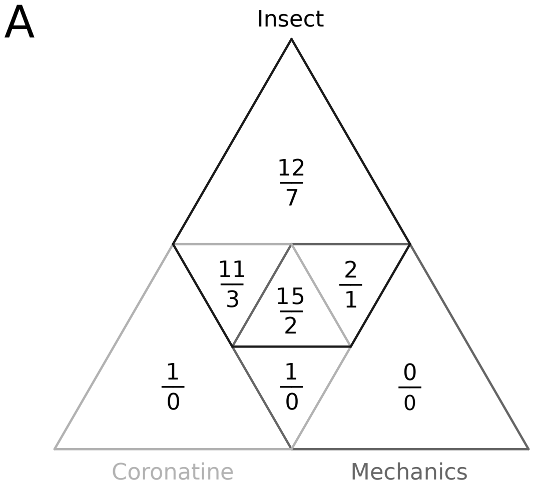
The molecular make-up of a non-stimulated *D. muscipula* revealed a strong similarity between the transcriptional landscape of glands and roots. Both showed a typical expression profile of a sink tissue. A subsequent morphological analysis via transmission electron microscopy (TEM) confirmed the enrichment results from the initial expression analysis (see table 1.22, [25]). The analysis also revealed that the glands undergo massive structural changes during insect-feeding or COR treatment. In order to associate these ultra-structural changes with molecular activities, resting and active glands were compared. Glands were stimulated with insects to activate secretion. Mechanically separated glands from the inner trap surface were subjected to transcriptome profiling after 24h of the onset of the stimulus. Stimulated glands show 3,447 genes up- and 2,826 genes down-regulated. The majority of the up-regulated genes were indicative for highly active transport, signal transduction and stress responses (see table 1.24 and 1.25).

Table 1.24 Gene ontology enrichment result of DEGs detected in insect-stimulated glands. Only genes with a base mean expression value of equal or larger than 100 were taken into account.

Term	Genes	DEG	adj. P-Value
transport	2286	460	2.2e-10
signal transduction	1147	233	8.7e-06
response to endogenous stimulus	1024	210	1.3e-05
response to biotic stimulus	831	164	0.00085
response to stress	2439	430	0.00250
response to external stimulus	1198	228	0.00355
cell death	257	57	0.00387
response to abiotic stimulus	1877	334	0.00486
catabolic process	1670	297	0.00829

Table 1.25 MapMan Bin enrichment result of DEGs detected in insect-stimulated glands. Only genes with a base mean expression value of equal or larger than 100 were taken into account.

Term	Annotated	Significant	P-Value	Q-Value
RNA.regulation of transcription.C2H2 zinc finger family	35	13	2.61E-04	5.23E-03
protein.targeting.secretory pathway	95	29	7.06E-04	7.06E-03
stress.abiotic.heat	120	23	4.35E-03	2.10E-02
RNA.regulation of transcription.AP2/EREBP	22	10	4.70E-03	2.10E-02
protein.degradation.ubiquitin	345	93	6.34E-03	2.10E-02
RNA.regulation of transcription.bZIP transcription factor family	24	10	6.43E-03	2.10E-02
protein.degradation.serine protease	49	11	7.84E-03	2.10E-02
RNA.regulation of transcription.WRKY domain transcription factor family	23	15	8.41E-03	2.10E-02
protein.degradation.cysteine protease	44	11	1.87E-02	4.15E-02
stress.biotic.PR-proteins	24	11	3.38E-02	6.75E-02



B

Enzyme	FC
Purple Acid Phosphatase 27	971
S1/P1 Nuclease 1	802
Cystein Peptidase C1A (SAG12)	667
Beta-Glucanase (BGL2)	602
Serine Carboxypeptidase 49 (SCPL49)	496
Ribonuclease T2 (RNS1)	472
Chitinase Class I (VF CHITINASE I)	401
Plant Peroxidase	362
S1/P1 Nuclease 2	340
Plant Lipid Transfer Protein	328
Peptide-N4-Asparagine Amidase A	26
Pathogenesis-related Protein	13
LysM-containing Protein	9
Aspartic peptidase 1	7
Aspartic peptidase 2	5

Fig. 1.24 The hydrolytic cocktail of *D. muscipula* adopted from [25]. (A left) Venn diagram of potential secretome members from an overlay of RNA-seq data (resting and insect-stimulated glands) and HTS proteomic measurements of secretions from chemically (COR), mechanically and insect-stimulated traps. Numbers indicate up- (top) and down- (bottom) regulated transcripts. Potential candidates are limited to transcripts being differentially regulated, containing a proper signal peptide and being detected at least once using HTS proteomics. (B) List of potential secretome members that were differentially up-regulated after insect-stimulation and detected in all three HTS proteomic measurements.

Results of the GO and MapMan bin enrichments point towards a elevated transcription of components which are potentially part of the hydrolytic cocktail (e.g., MapMan bins protein.degradation.serine protease and stress.biotic.PR-proteins). Three layers of evidence were combined to understand the composition of the hydrolytic cocktail in detail. DE results from exp008 for both, non-stimulated and insect-stimulated glands were overlay with profile-based signal peptide annotations (10,562 unigenes had an annotated signal peptide) and High Throughput Screening (HTS) proteomics data. The proteomic data comprised peptide measurements from the secreted mucilage collected after mechanical (140 unigenes with at least one peptide matching isoform), insect (380 unigenes with at least one peptide matching isoform) and COR (247 unigenes with at least one peptide matching isoform) stimulation. Within the group of up-regulated, signal peptide containing transcripts, 42 were differentially expressed. Using the HTS proteomic data set of the secretion fluid, all 42 candidate genes were confirmed to be actively secreted (see left figure 1.4.7). Among them, fifteen proteins were secreted irrespective of the nature of the stimulus (COR, trigger hair stimulation or insect). They encoded secretome-related proteins with hydrolase activity such as proteases, phosphatases and chitinases, as well as defensin-like (DEFL) cysteine-rich proteins (see right figure 1.4.7). On the contrary, only 13 down-regulated DEGs had proteomic evidence together with an annotated signal peptide (see left figure 1.4.7, denominators). The most abundant transcripts that were secreted irrespective of the nature of the stimulus (COR, trigger hair stimulation or insect) encoded the aspartic protease nepenthesin (Nep) [40] and a Lipid Transfer Protein (LTP). In summary, some components of the hydrolytic cocktail already showed a high expression in resting glands but the majority of components showed a massive induction ranging from 5 to 900-fold after stimulation. Abovementioned enzymes are likely to be secreted into the traps through exocytosis [6]. Potential components of the exocytosis machinery were identified using annotated *A. thaliana* components ("Exocyst complex component") and the CRB ortholog map (see table 1.28). Using 15 *A. thaliana* components, 19 putative *D. muscipula* components of the exocytosis machinery were identified. Additional components were identified using Pfam protein domain annotations (PF15469, PF15278, PF16528, PF09763, PF07393, PF04091, PF04048, PF15277, PF06046, PF03081). Most components were not up-regulated after insect stimulation but experienced an higher expression in resting Glands than in resting Petioles (exp001).

Table 1.26 Summary of putative members of the exocytosis machinery and their transcriptional regulation. Most components experienced an higher expression in resting Glands than compared to resting Petioles (exp001).

Function	Component Count	DE(Insect)	DE(Gland)[2]
Sec5 (PF15469)	2	0	0
Sec3_C_2 (PF15278)	0	0	0
Exo84_C (PF16528)	0	0	0
Sec3_C (PF09763)	5	0	5
Sec10 (PF07393)	3	0	1
Sec15 (PF04091)	2	0	2
Sec8_exocyst (PF04048)	1	0	1
Sec3-PIP2_bind (PF15277)	2	0	0
Sec6 (PF06046)	1	0	0
Exo70 (PF03081)	15	0	8

---

[2] compared to Petiole

1.4.8 Inventory of the glandular nutrient uptake machinery

During the digestion of the prey, metabolites and minerals containing essential plant nutrients like nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), sulphur (S), and magnesium (Mg) are released (Adamec, 1997). Several studies with organic nitrogen and carbon suggest that the glands are responsible for the nutrient resorption (Schulze et al., 2001; Kruse et al., 2014), but the molecular landscape of the uptake machinery remained unknown in larger parts. Again, a combination of profile-based transmembrane annotations (see TMHMM annotations described in section 1.8), homology-based transporter classification (see TCDB annotations described in section 1.3.8) and differential expression testing (exp008) was used to mine for potential transporter in both, the resting and the insect-feeding glands of *D. muscipula*. Transporter classification as well as substrate targets and likely transporter localizations were manually curated by Dirk Becker, Department of Botany I, University of Würzburg. Out of 2950 transcripts classified according to the transporter classification (TC) system, 145 were down-regulated and 148 up-regulated after insect-stimulation of glands. Secondary active transporters represented the most frequent and strongly regulated class (see figure 1.25 A). The amount of transport proteins involved in plasma membrane transport was significant increased in insect-stimulated glands. On the contrary, plastid localized transporters appeared strongly down-regulated, with the exception of the energy-supplying, plastidic ATP/ADP antiporter NTT1, that was strongly induced (see figure 1.25 B; Flugge et al., 2011). Several transporter substrate classes showed transcripts with a drastic up-regulation (see figure 1.25 C). Especially phosphate, metal, nitrogen and nucleotide transporter showed larger expression changes. Notably, highly induced plasma membrane phosphate transporters PT1 and PT2 (Shin et al., 2004) and a high-affinity molybdate transporter MOT1 (Tomatsu et al., 2007) were up-regulated in insect-stimulated glands. Likewise, plasma membrane transporters for sulphate, as well as nitrogen-containing solutes, were induced by insect stimulation. See supplementary table 15 in [25] for a detailed overview of the transportome or explore the transportome under http://tbro.carnivorom.com/tbro/graphs/S15_Transportome. Transporters, especially transporters up-regulated in non-stimulated glands were cross-validated using High Throughput Screening (HTS) proteomics data produced from whole trap membrane extracts (unpublished material from Prof. Dr. Waltraud Schulze, Plant Systems Biology, University of Hohenheim). None of the 293 differentially regulated transporters were found with the HTS screen most likely due to sub optimal experimental resolution or an insufficient expression level (e.g., compared to expression levels of secreted proteins).

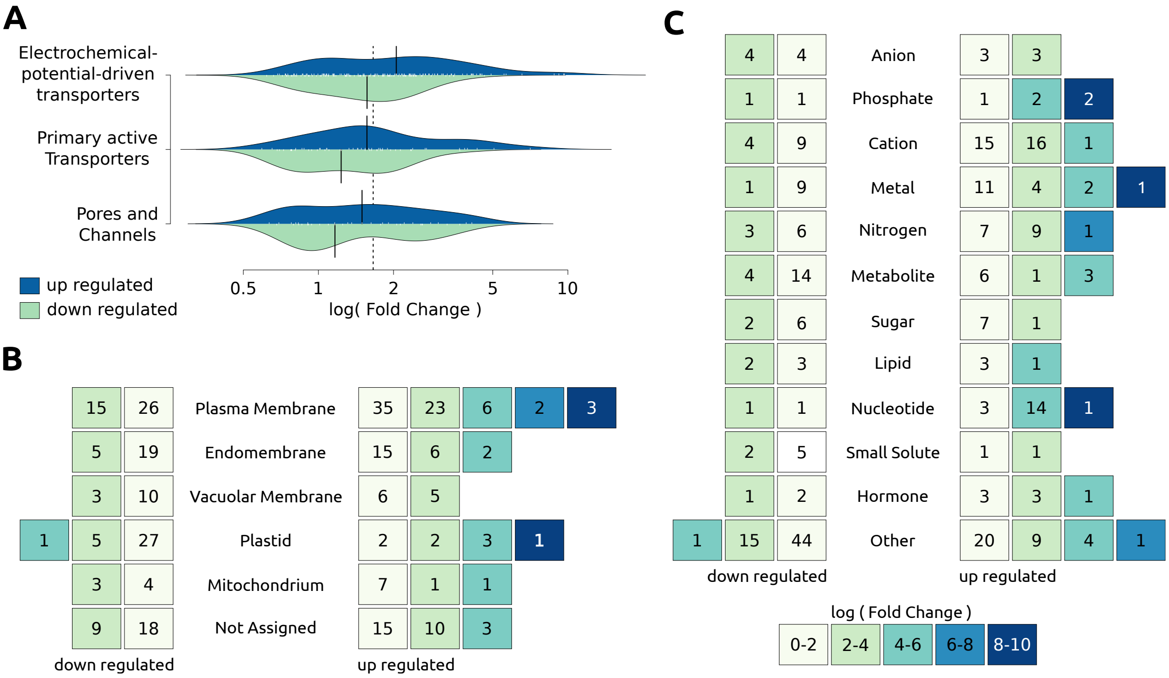


Fig. 1.25 The *D. muscipula* transporter inventory and potential components of the glandular nutrient machinery. (A) Bean plot of logarithmic fold changes (LFC) for all transporters grouped into major TCDB classes. Upper and lower distribution for each bean show up- and down-regulated transporters respectively. Black bars indicate the median LFC of each class per condition. White ticks indicate individual transporters. The class “Electrochemical potential-driven transporters” contains the most up regulated members. (B) Numbers of differentially regulated transporters grouped by their potential sub cellular location. Transporters are binned according to their LFC. Bin colours indicate the range of the underlying LFC. (C) Numbers of differentially regulated transporters grouped by their potential substrate class.

Most transport processes are driven by electrochemical gradients across the plasma membrane. Especially P-Type H^+ ATPases contribute to the maintenance of the gradient. An additional P-Type H^+ ATPase search using the Interpro profile IPR008250 was conducted. From 93 unigenes (respectively 187 isoforms) annotated as P-Type H^+ ATPase, 10 were up-regulated in insect-stimulated glands while 12 were down-regulated. The most strongest up-regulated transcript was a AHA1-like P-Type H^+ ATPase thought to be involved in jasmonate-induced ion fluxes and stomatal closure (see figure 1.26) [198].

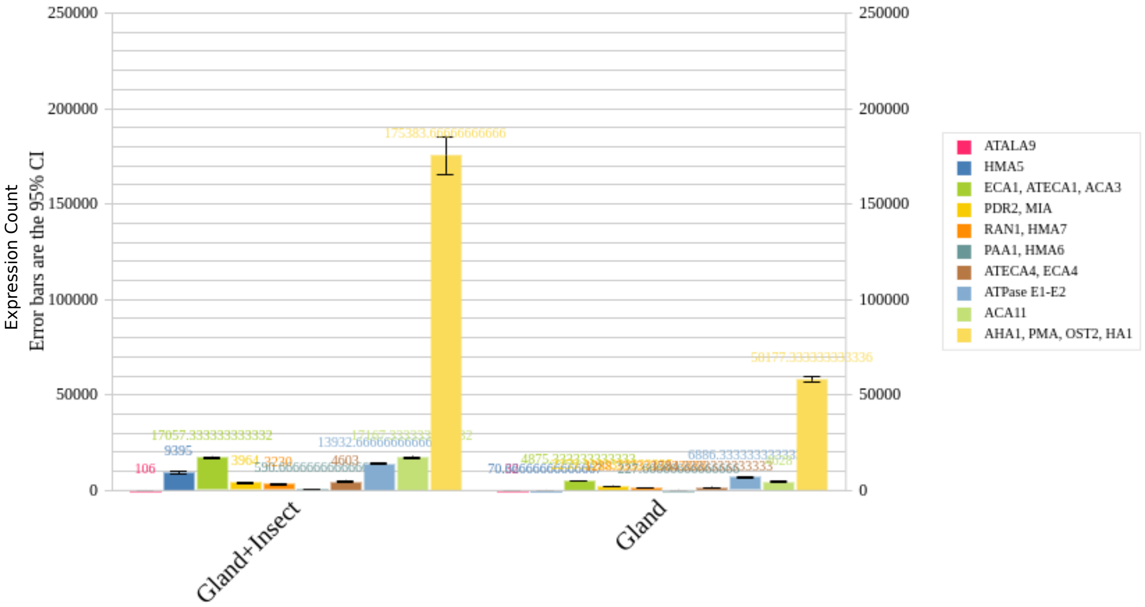


Fig. 1.26 Expression changes of a AHA1-like P-Type H^+ ATPase after insect stimulation in glands. The already highly expressed ATPase gets further up-regulated after insect stimulation. Expression data was taken from the current *D. muscipula* transcriptome browser (http://tbro.carnivorom.com)

Endocytosis provides another potential nutrient resorption mechanisms next to transporters and channels (Adlassnig et al., 2012) but is less understood in plants. Potential components of the endocytosis machinery were identified using published *A. thaliana* components and the CRB ortholog map (see table 1.28). Most components were not up-regulated after insect stimulation but experienced a higher expression in resting glands than in resting petioles (exp001).

Table 1.28 [Summary of putative members of the endocytosis machinery and their transcriptional regulation

Function	*A. thaliana* Proteins	CRBs	DE(Insect)	DE(Gland)[3]
Adaptors	14	18	4	9
Accessory proteins	6	8	1	4
Clathrin coat	5	13	1	2
Scission	6	10	1	5

---

[3] compared to Petiole

1.4.9 Trap-wide response to insect feeding

Functional enrichment of insect-stimulated DEGs in glands showed a stress-like response. The expression of stress-related proteins with hydrolytic, proteinase-inhibitory and membrane-permeabilizing ability (e.g., MapMan bin stress.biotic.PR-proteins, protein.degradation.serine protease) indicates that molecular pathways resembling defence and or wounding responses might be active during the feeding process. In order to capture the trap-wide response to insect stimulation transcriptomic profiles were generated. Testing for differential expression between insect-stimulated and resting traps (exp004) identified 2,137 genes as up-regulated while 852 where suppressed. Similar to the stimulated glands, insect-stimulated traps displayed a massive stress response and a high transport activity (see table 1.29 and 1.30).

Table 1.29 Gene ontology enrichment result of DEGs detected in insect-stimulated traps. Only genes with a base mean expression value of equal or larger than 100 were taken into account.

Term	Genes	DEG	adj. P-Value
response to stress	2520	276	1.2×10^{-10}
transmembrane transport	643	84	1.5×10^{-6}
cell death	282	43	1.5×10^{-5}
immune system process	506	66	2.1×10^{-5}
signal transduction	1175	126	0.0001
transport	2373	237	0.0011

Table 1.30 MapMan Bin enrichment result of DEGs detected in insect-stimulated traps. Only genes with a base mean expression value of equal or larger than 100 were taken into account.

Term	Annotated	Significant	Q-Value
protein.synthesis.ribosomal protein	646	77	0.008,451,343
protein.degradation.cysteine protease	44	11	0.014338123
stress.abiotic.drought/salt	43	11	0.023722335
secondary metabolism.flavonoids.dihydroflavonols	23	11	0.046274575
RNA.regulation of transcription.bZIP transcription factor family	24	12	0.054495196
stress.biotic.PR-proteins	24	15	0.054495196
RNA.regulation of transcription.WRKY domain transcription factor family	23	17	0.064214115
protein.degradation.serine protease	49	17	0.064214115

In order to describe the nature of the GO enrichment set more meaningful a clustering relying on semantic similarity measures was applied [269]. GO terms where summarized by removing redundant GO terms when they contained "is a" relations (e.g., "car" is similar to "bus", but is also related to "road" and "driving"). The resulting representative subset was visualized with tree mapping, a method that uses nested rectangles to display hierarchical data (see 1.27).

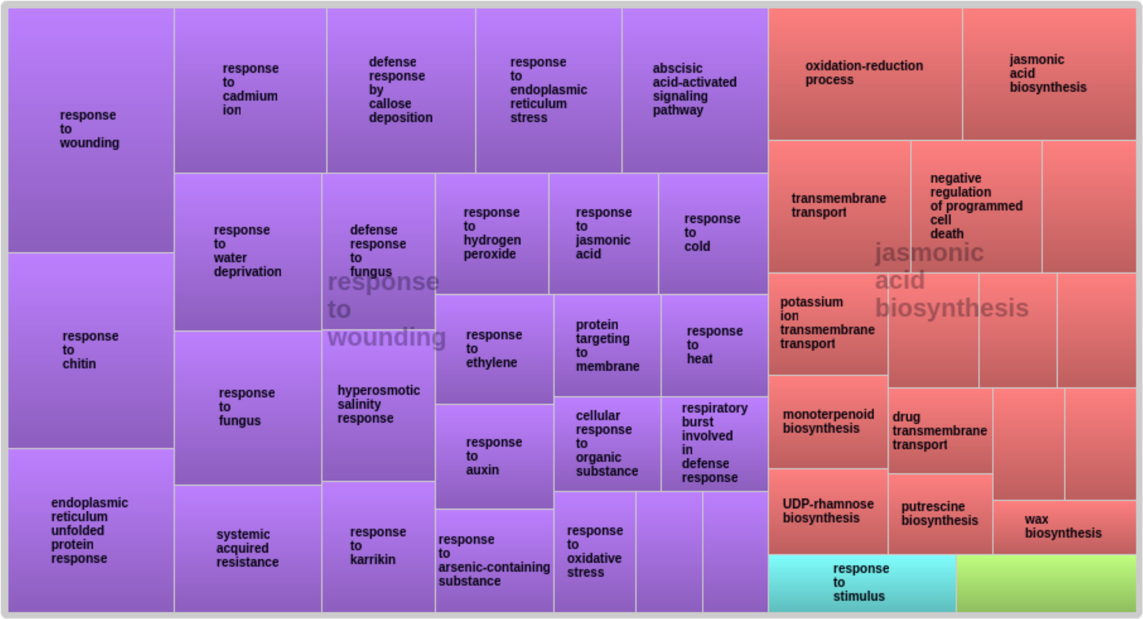


Fig. 1.27 Gene ontology treemap of insect-stimulated traps. The plot was drawn with Revigo [269] using Resniks normalized semantic similarity measures and a low similarity score of 0.4 to optimally group similar terms. The two terms "response to wounding" and "jasmonic acid biosynthesis" almost entirely represented the raw gene ontology terms.

A visual inspection of the treemap revealed that the terms "response to wounding" and "jasmonic acid biosynthesis" successfully represented most of the raw GO terms. An individual analysis of JA core components in glands and traps (see figure 1.28) further showed that insect stimulation accelerates the expression of genes related to jasmonic acid biosynthesis. A highly similar transcriptional response was triggered when the JA mimic COR was applied to glands and traps (see figure1.28). Both COR and insects seem to activate the oxylipin pathway leading to the formation of 12-oxo phytodienoic acid (OPDA). Insects, however, seem to further promote biosynthesis of JA-Ile, the true COI1 ligand [253] by an elevated production of downstream biosynthesis components.

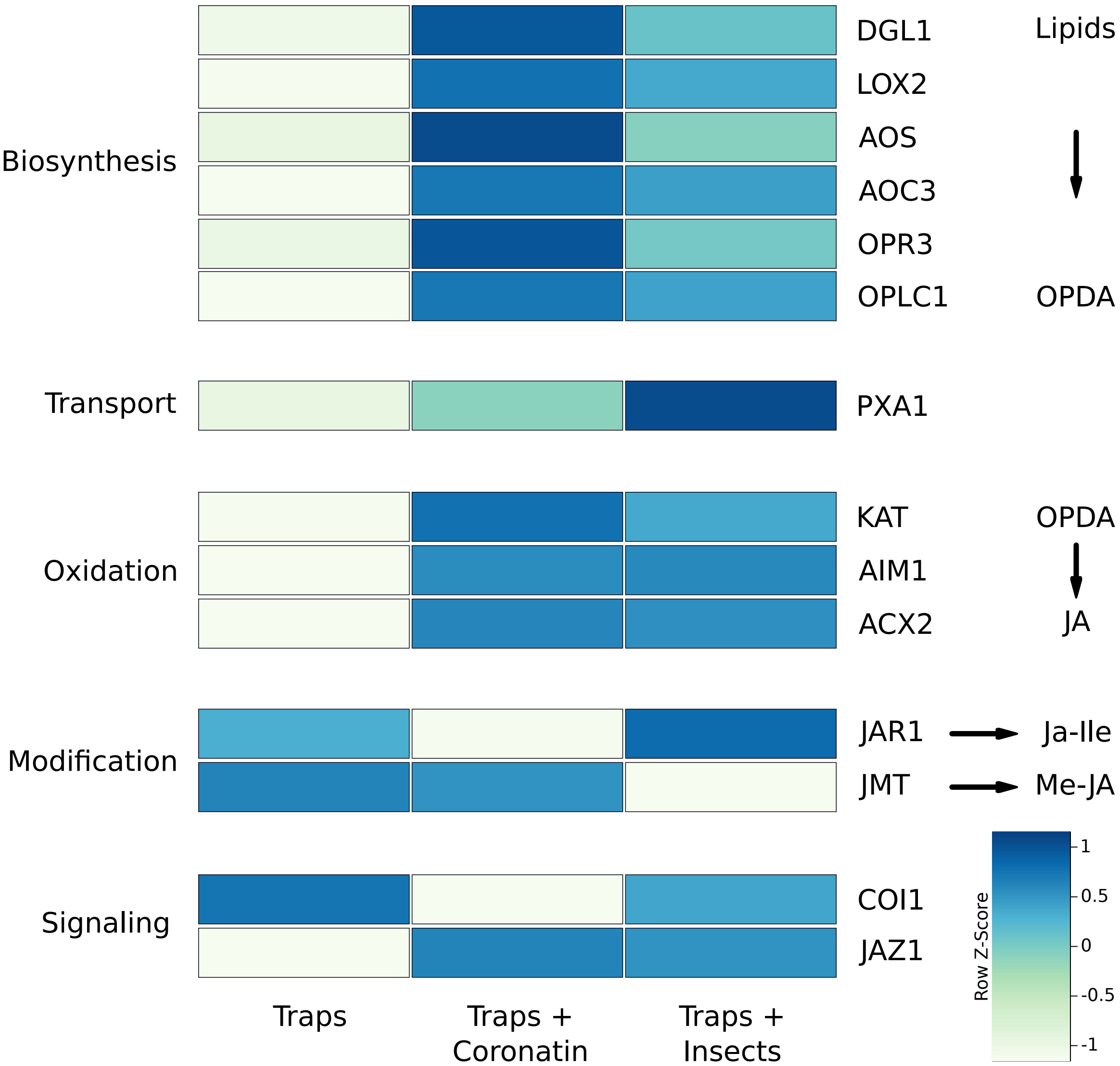


Fig. 1.28 Expression patterns of jasmonic acid biosynthesis core components in resting traps as well as traps after COR and insect stimulation adopted from [25]. Expression values are scaled by rows (Z-scoring). Most components are up regulated after insect and COR stimulation. Only insect stimulation shows a up regulation of JAR1, a component that promotes biosynthesis of JA-Ile. Abbreviations are DAD1 (DEFECTIVE ANTHER DEHISCENCE 1): phospholipase A1/ triacylglycerol lipase; LOX2 (LIPOXYGENASE 2); AOS (ALLENE OXIDE SYNTHASE); AOC3 (ALLENE OXIDE CYCLASE 3); OPR3 (OPDA-REDUCTASE 3): 12-oxophytodienoate reductase; OPCL1 (OPC-8:0 COA LIGASE1); PXA1 (PEROXISOMAL ABC TRANSPORTER 1. CTS (COMATOSE); KAT2/PED1/PKT3 (PEROXISOMAL 3-KETOACYL-COA THIOLASE 3); AIM1 (ABNORMAL INFLORESCENCE MERISTEM); ACX2 (ACYL-COA OXIDASE 2); JAR1 (JASMONATE RESISTANT 1); JMT (JASMONIC ACID CARBOXYL METHYLTRANSFERASE).

Table 1.31 MapMan Bin counts of terms related to translation and stress in insect-stimulated traps. Only genes with a base mean expression value of equal or larger than 100 were taken into account. Abbreviations are SCPL (SERINE CARBOXYPEPTIDASE-LIKE); SBT (SUBTILISIN-TYPE SERINE PROTEASE); APM (AMINOPEPTIDASE); HSP (HEAT SHOCK PROTEIN); CPN (CHAPERONIN); DGL (DOLICHYL-DIPHOSPHOOLIGOSACCHARIDE); HEXO (HEXOSAMINIDASE); GAMMA-VPE (GAMMA-VACUOLAR-PROCESSING ENZYME), PEX (PEROXISOME BIOGENESIS FACTOR); DIL (DILUTES; MYOSIN HEAVY CHAIN); GPX (GLUTATHIONE PEROXIDASE); SIR (SULFITE REDUCTASE); CB5 (CYTOCHROME B5); CHI (ENDOCHITINASE), ARL (ARGOS-LIKE), ERD (EARLY RESPONSIVE TO DEHYDRATION)

Class	Member	Families
Protein degradation	67	SCPLs, SBTs, APMs
Protein folding	6	HSPs, CPNs
Protein glycolysation	7	DGLs, HEXOs
Protein targeting	15	GAMMA-VPEs, PEXs
Redox response	50	DILs, GPXs, SIRs, CB5s
Stress response	69	HSPs, CHIs, ARLs, ERDs

A sub classification of DEGs into MapMan bins associated with translation and stress resulted in 214 group-able DEGs (see table 1.30). The set comprised transcripts related to the production of reactive oxygen species (ROS) scavengers [204] and components of the ER-quality control (ER-QC) machinery [190, 295] as well as members of the putative hydrolytic cocktail. Additionally, 38 transcripts were predicted to negatively regulate programmed cell death (PCD, see table 1.32).

Table 1.32 Members of the GO term "programmed cell death" (PCD) differentially expressed in insect-stimulated traps (IsTr) and glands (IsGl). Only genes with a base mean expression value of equal or larger than 100 were taken into account. Four genes without any annotation were excluded from the table.

Synonym	Description	IsTrap	IsGland
DPL1	dihydrosphingosine phosphate lyase	yes	yes
NTF2B	nuclear transport factor 2B	yes	yes
NA	Eukaryotic aspartyl protease family protein	yes	yes
PLDP1	phospholipase D P1	yes	yes
NA	alpha/beta-Hydrolases superfamily protein	yes	yes
MLO5	Seven transmembrane MLO family protein	yes	yes
PPI1	proton pump interactor 1	yes	yes
PLL4	poltergeist like 4	yes	yes
NA	Major facilitator superfamily protein	yes	yes
PDIL1-1	PDI-like 1-1	yes	yes
BAH1	SPX (SYG1/Pho81/XPR1) domain-containing protein	yes	yes
HXK1	hexokinase 1	yes	yes
NA	Glycosyl hydrolase superfamily protein	yes	yes
NA	Leucine-rich repeat protein kinase family protein	yes	yes
NA	Chloroplast-targeted copper chaperone protein	yes	yes
S6K2	serine/threonine protein kinase 2	yes	yes
PCS1	Eukaryotic aspartyl protease family protein	yes	yes
AAP3	amino acid permease 3	yes	yes
ELI3-1	elicitor-activated gene 3-1	yes	yes
CCR3	CRINKLY4 related 3	yes	yes
NA	DCD (Development and Cell Death) domain protein	yes	yes
DND1	Cyclic nucleotide-regulated ion channel family protein	yes	yes
CBL1	calcineurin B-like protein 1	yes	yes
FC1	FUS3-complementing gene 1	yes	yes
RLK7	Leucine-rich receptor-like protein kinase family protein	yes	yes
NA	SBP (S-ribonuclease binding protein) family protein	yes	no
NA	Protein kinase superfamily protein	yes	no
NA	Target of Myb protein 1	yes	no
CPK9	calmodulin-domain protein kinase 9	yes	no
LHT1	lysine histidine transporter 1	yes	no
SOBIR1	Leucine-rich repeat protein kinase family protein	yes	no
CDF1	cycling DOF factor 1	yes	no
NSL1	MAC/Perforin domain-containing protein	yes	no

1.4.10 Regulatory capacities of the *D. muscipula*

Plant cells are competent for the activation of defence responses [184]. They evolved a complex regulatory network able to activate defense genes after perception of different primary signals like wounding. An important part of the regulatory complex are transcription factors, proteins that can control transcription by binding to specific DNA sequences [160, 182]. Putative transcription factors of *D. muscipula* were identified by applying the family assignment rules defined at PlantTFDB [116]. The databases established a reliably assignment scheme (see website http://planttfdb.cbi.pku.edu.cn/help_famschema.php) that uses profile-based domain annotations to classify transcription factors. Within the complete transcriptome, 3.184 isoforms and 1.088 unigenes respectively were classified into 36 different transcription factor families. Inspection of the classification resulted in 1,088 distinct assignments. Six unigenes showed ambiguous multi-annotations (e.g., isoform 1 B3, isoform 2 AP2-ERF) but where resolved manually by inspection of the underlying isoforms. Out of 1.088 unigenes 735 were at least once differentially expressed in exp001, exp004 or exp008.

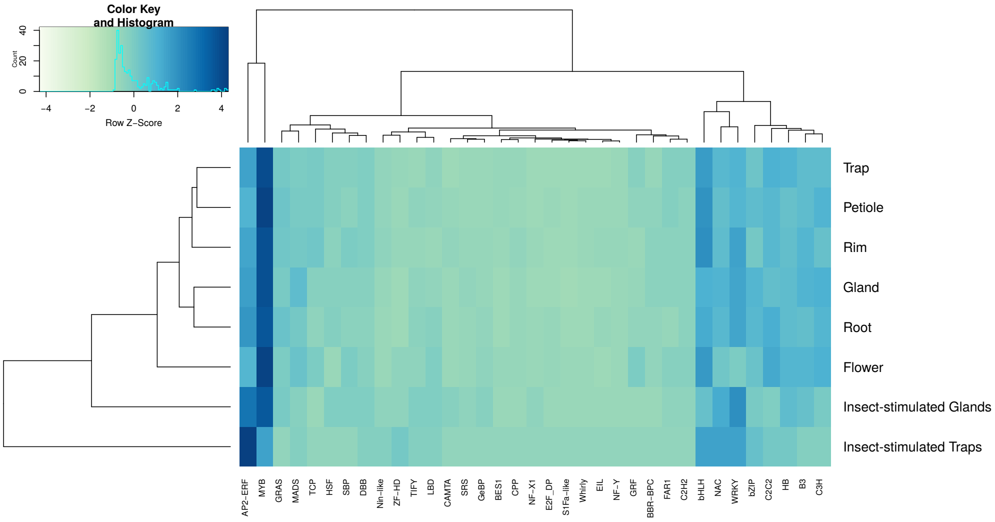


Fig. 1.29 Differentially regulated transcription factors of *D. muscipula*. The color gradient corresponds to the number of differentially regulated members of each transcription factor family. Differentially expressed transcription factors were defined based on a) the all-versus-all comparison from exp001 (see figure 1.4.5) and b) for conditional experiments exp004 (insect-stimulated traps) and exp008 (insect-stimulated Glands). Absolute numbers of the conditional experiments differ from the resting tissues since they were only compared to their control while the resting tissues were compared all-versus-all and differential expression results from each comparison was taken into account.

AP2-ERF and MYB transcriptions factors were the most strongly regulated groups (see figure 1.29). A gene family enrichment further supported these results (see table 1.33).

From 38 tested groups 12 were enriched in resting traps while nine were enriched in petiole using DEGs from esting traps and petioles (exp001). Differentially expressed unigenes from insect-stimulated traps were enriched for six families that were already highly active in resting traps (exp004). Glands showed only an enrichment in four families.

Table 1.33 Transcription factor families enriched in exp001, exp004 and exp008. The enrichment was carried out using GAGE and the PlantTFDB classification. Individual unigenes were considered significant with a adjusted P-value of ≤ 0.01 while a transcription factor family was considered enriched with a Q-value of ≤ 0.1 (corresponds to a false positive rate of ≤ 0.1). Conditional experiments (exp004 and exp008) were compared with their respective control which might differ from the resting tissues shown here.

Class	Petiole	Trap	Insect-stimulated Trap	Insect-stimulated Gland
AP2-ERF	1.22E-02	6.55E-04	2.88E-03	1.23E-02
WRKY	2.51E-03	3.66E-03	7.75E-04	3.50E-03
bZIP	6.00E-03	7.69E-03	3.33E-02	-
C2C2	-	7.69E-03	-	-
MYB	1.47E-04	1.27E-02	2.88E-03	9.06E-02
NAC	4.86E-02	1.66E-02	2.44E-02	-
C3H	1.73E-03	3.20E-02	-	-
MADS	-	3.22E-02	2.30E-02	-
HB	6.10E-02	3.22E-02	-	-
B3	2.43E-02	6.86E-02	-	-
GRAS	-	7.21E-02	-	-
bHLH	1.88E-03	9.24E-02	-	9.06E-02

Detailed inspection of the 53 highly regulated transcription factors (families) expressed in insect-stimulated traps revealed several transcription factors involved in stress response like CBF1, ABR1, JAZ1, ATAF2 and AIB [257]. Additionally, several transcription factors (e.g., CDF1, CDF4, DORNROSCHEN-like as well as RHL1-like) were associated with developmental processes [170], like flowering timing [97] or ploidy-dependent cell growth [268]. Most differentially expressed members of the WRKY transcription factor family (11 out of 17) remained without any further functional annotation next to the detectable protein domains. Only a WRKY20-like (regulation of ABA signalling, [193]), two WRKY22-like (dark-induced leaf senescence, [306]) , a WRKY33-like, a WRKY40-like (both regulating defense against pathogens, [305]) and a WRKY42-like (regulating of phosphate homeostasis, [266]) transcription factor could be further described.

1.4.11 Sensory capacities of the *D. muscipula*

Assessing chemistry and quality of the prey is vital for *D. muscipula*. Kinases, especially receptor-like-kinases (RLKs) are often involved in chemical sensing (Antolin-Llovera et al., 2012) and play a key role in most cellular activities. The *D. muscipula* kinome was characterized by searching the profile-based annotation for isoforms with protein kinase domains (PF00069). Isoforms were further classified using Kinomer [202]. From 872 identified unigenes with at least one kinase-containing isoform, 590 were at least once differentially expressed in exp001, exp004 or exp008. Sub classification resulted in 467 distinct classifications spread across 12 different kinase groups. 123 unigenes showed ambiguous multi-annotations (e.g., isoform 1 TK, isoform 2 CAMK) and where not considered downstream.

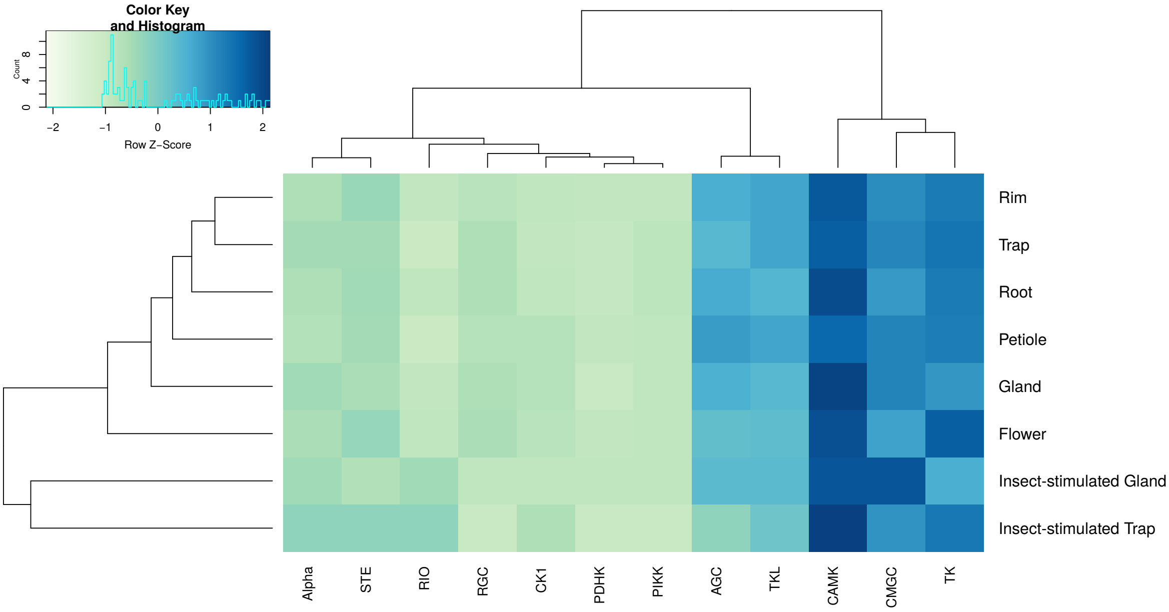


Fig. 1.30 Differentially regulated kinome of *D. muscipula*. The color gradient corresponds to the number of differentially regulated members of each kinase group. Differentially expressed kinases were defined based on a) the all-versus-all comparison from exp001 (see figure 1.4.5) and b) for conditional experiments exp004 (insect-stimulated traps) and exp008 (insect-stimulated Glands). Absolute numbers of the conditional experiments differ from the resting tissues since they were only compared to their control while the resting tissues were compared all-versus-all and differential expression results from each comparison were taken into account.

Tyrosine kinases, Ca^{2+}/Calmodulin-dependent protein kinases and CGMC kinases were the most strongly regulated groups (see figure 1.30). A gene family enrichment further supported these results (see table 1.34). From 12 tested groups, seven were enriched in resting traps while six were enriched in petioles using DEGs from resting traps and petioles (exp001). In insect-stimulated traps, members of the CAMK, TKL, CMGC and AGC group were differentially expressed compared to resting traps (exp004). A similar picture could be found in insect-stimulated glands were all kinase groups apart from TKLs showed a high number of differentially regulated members.

Table 1.34 Kinase groups enriched in exp001, exp004 and exp008. The enrichment was carried out using GAGE and the Kinomer group classification. Individual unigenes were considered significant with an adjusted P-value of ≤ 0.01 while a kinome group was considered enriched with a Q-value of ≤ 0.1 (corresponds to a false positive rate of ≤ 0.1). Conditional experiments (exp004 and exp008) were compared with their respective control which might differ from the resting tissues shown here.

Class	Petiole	Trap	Insect-stimulated Trap	Insect-stimulated Gland
CAMK	1.47×10^{-3}	8.88×10^{-7}	1.07×10^{-2}	5.13×10^{-2}
TK	1.40×10^{-4}	8.88×10^{-7}	-	5.13×10^{-2}
TKL	1.47×10^{-3}	2.30×10^{-3}	3.33×10^{-2}	-
CMGC	5.45×10^{-3}	2.35×10^{-3}	1.07×10^{-2}	5.13×10^{-2}
AGC	5.45×10^{-3}	1.74×10^{-2}	1.07×10^{-2}	9.82×10^{-2}
STE	2.01×10^{-2}	1.94×10^{-2}	-	5.13×10^{-2}
Alpha	-	4.78×10^{-2}	-	9.82×10^{-2}

Inspection of the differentially regulated kinome of resting and insect-stimulated traps showed that most kinases belong to the class of receptor-like kinases (RLK; see figure 1.31). A kinase was defined as RLK when the isoform showed at least one transmembrane domain in the same ORF as the protein kinase domain. A simple intersection analysis revealed that 27 kinases were already highly expressed in resting traps (see figure 1.31) when compared to petioles but get a further boost after insect stimulation. A subset of 43 kinases are purely regulated in an insect-specific manner with 22 kinases being gland-specific.

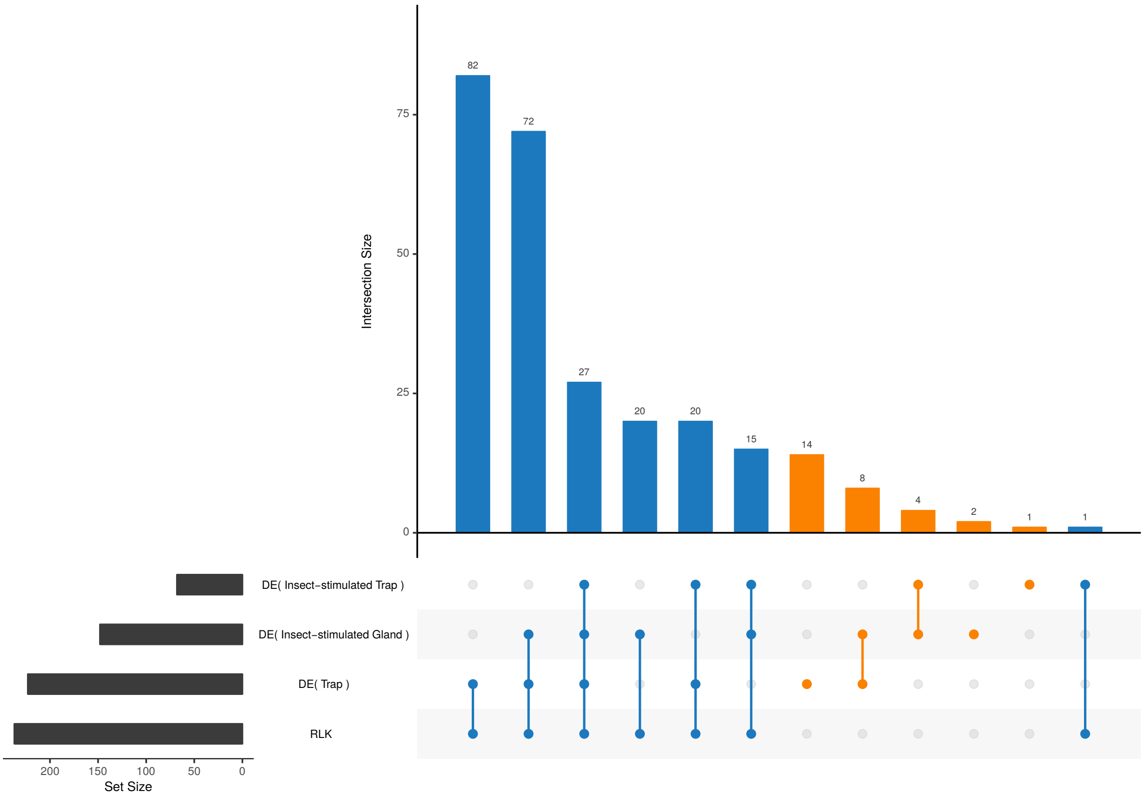
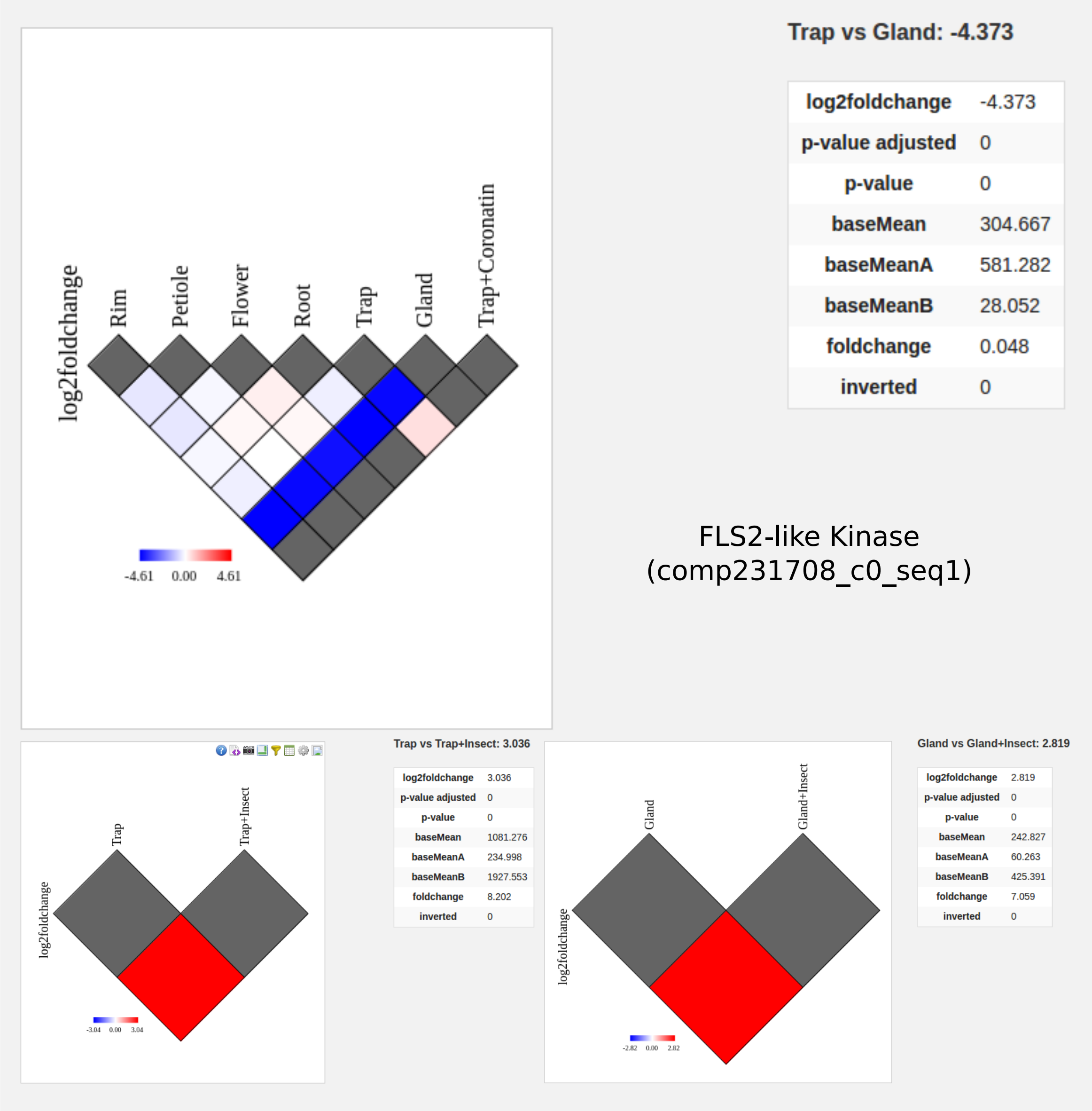


Fig. 1.31 Overlay of the kinases differentially expressed in traps (compared to petioles) as well as insect-stimulated kinases from glands and traps. RLKs are defined as isoform with a protein kinase domain and a transmembrane domain within the same open reading frame. The RLK subsets are colored in blue while the non-RLKs are colored in orange.

From the kinases regulated in a insect-specific manner, 15 were classified as RLK. Among them several RLKs were found involved in wounding response (PRB1-like - comp230357_c2) as well as MAP kinase signaling relay involved in innate immunity (FLS2-like - comp231708_c0; see figure 1.32). The latter is involved in the perception of bacterial flagellin, a potent elicitor of the defence response. Interestingly, also receptors involved in control of stomatal movements in response to CO_2 (HT1 - comp226371_c1) and negative regulation of root growth (PERK10 - comp215021_c0) were found.



Trap vs Gland: -4.373

log2foldchange	-4.373
p-value adjusted	0
p-value	0
baseMean	304.667
baseMeanA	581.282
baseMeanB	28.052
foldchange	0.048
inverted	0

Trap vs Trap+Insect: 3.036

log2foldchange	3.036
p-value adjusted	0
p-value	0
baseMean	1081.276
baseMeanA	234.998
baseMeanB	1927.553
foldchange	8.202
inverted	0

Gland vs Gland+Insect: 2.819

log2foldchange	2.819
p-value adjusted	0
p-value	0
baseMean	242.827
baseMeanA	60.263
baseMeanB	425.391
foldchange	7.059
inverted	0

Fig. 1.32 Expression patterns of a FLS2-like RLK of *D. muscipula* across resting and insect-stimulated tissues. The unigene showed a low expression in resting glands and a weak up-regulation in COR-stimulated traps (upper part) while it showed a strong induction in glands and traps after insect stimulation (lower part).

1.4.12 Metagenomics changes during insect feeding

The initial RNA-seq contamination screen and the detection of RLKs associated with the perception of bacterial flagellin in several transcriptome samples suggested a tailored response of *D. muscipula* to insect-stimulation depending on the quality of the prey and potentially towards the associated microbiome. Since microbial communities can be profiled easily from whole transcriptome data (at least those that remain on traps or petioles after the wash procedure and are also accessible through the RNA extraction), stimulated and no-stimulated samples from exp001 and exp004 where characterized. Results from the initial contamination screen where reused to assign the most likely taxonomic label to individual reads. Sample assignment rates varied from 0.75 to 28.35% (see figure 1.5) at the domain level (only counting assignments to Bacteria and Archaea). At lower levels less reads were assigned on average (e.g. only 84% of the reads assigned to the bacterial domain could also be assigned to a phylum). The taxonomic level order was chosen for all subsequent analysis since most samples had still assignment rates of more than 64% at a moderate standard deviation of 4%. The phylum was not considered since it is usually less informative. All other levels experienced lower assignments rates and where as well not considered (e.g., class with 35%, family 49%). Scaled assignment counts (z-scores normalization) where inspected with STAMP. Principal component analysis for exp001 showed no clear clustering of samples according to biological conditions (see figure 1.33). The same was true for the non-stimulated and COR-stimulated trap samples from exp001 (see figure 1.34). Samples from exp004 were roughly separated in a non-stimulated and a insect-stimulated group with two samples from each condition being very similar to each other (see figure 1.35). A test comparing the two groups revealed several bacterial orders with significant differences under non-stimulated and insect-stimulated conditions (see figure 1.36). The most prominent differences where detected in the order Myxococcales. Since most of the myxobacteria ("slime bacteria") predominantly live in the soil and feed on insoluble organic substances it is unlikely that the differences are of biological significance but rather experimental contamination. Overall, the metagenomic profiling was able to assign a substantial amount of reads to various, frequently low taxonomic levels but no clear sample grouping nor group differences were detectable at the order level. Thus, the current data set is not sufficient to describe metagenomic changes during the insect-feeding process.

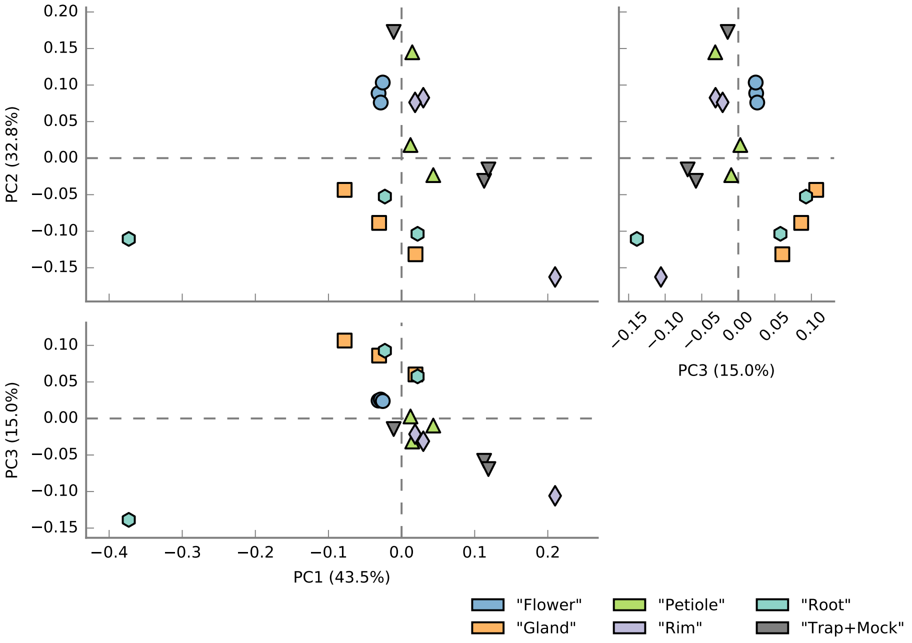


Fig. 1.33 PCA of metagenomic profiles from all resting samples. No scaling other than centering the data was applied before determining the PCA transformation. Tissue replicates only show a rough clustering. No clear separation between tissues is detectable on the three displayed components.

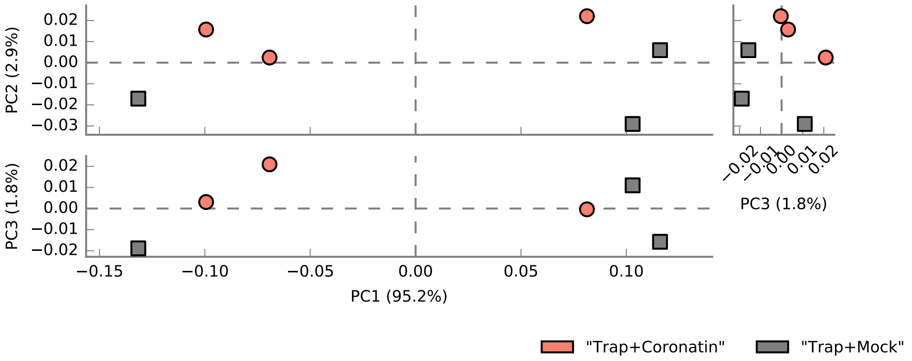


Fig. 1.34 PCA of metagenomic profiles from COR- and non-stimulated traps (exp001). No scaling other than centering the data was applied before determining the PCA transformation. Tissue replicates show no obvious clustering according to their source. No clear separation between tissues is detectable on the three displayed components.

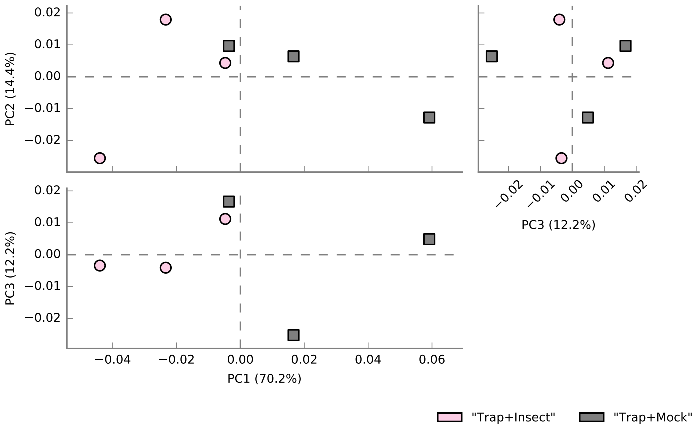


Fig. 1.35 PCA of metagenomic profiles from insect- and non-stimulated traps (exp004). No scaling other than centering the data was applied before determining the PCA transformation. Tissue replicates show no obvious clustering according to their source. No clear separation between tissues is detectable on the three displayed components.

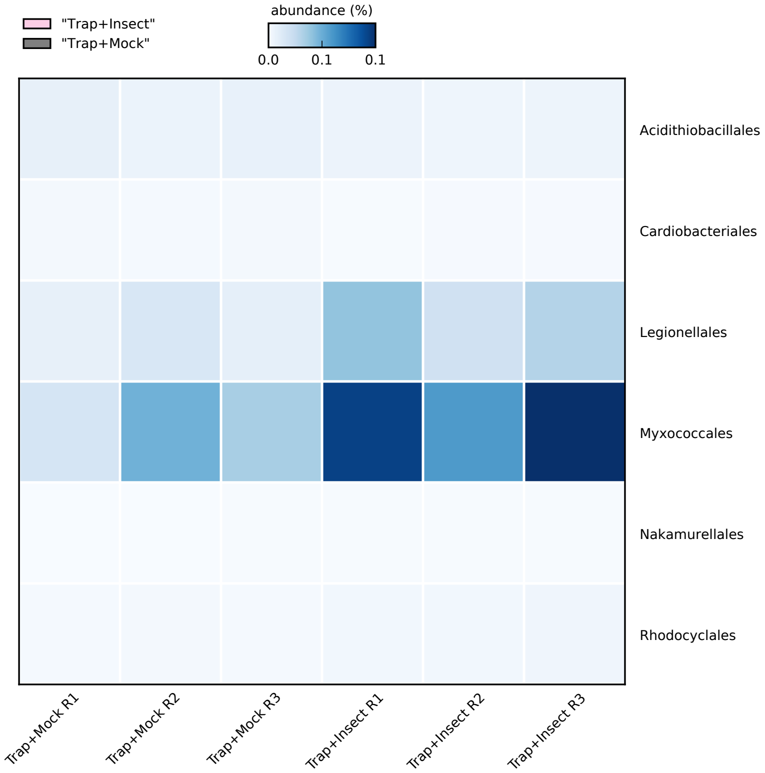


Fig. 1.36 Abundance plot indicating the proportion of sequences assigned to each of the orders detected to be different under non-stimulated and insect-stimulated conditions. Replicates show roughly the same abundance levels in most of the detected orders. The most prominent differences where detected in the order Myxococcales.

1.4.13 Conserved stress signals

Gene ontology and MapMan-based enrichments carried out on unigenes differentially up-regulated after insect stimulation showed terms and functional categories predominant that can be observed in typical stress response studies in plants [209, 307, 169]. To test whether the stress signals are conserved in other carnivorous plants to a higher extend or the signals can be sub-classified using stress response studies in non-carnivorous plants four different analysis were carried out. First, transcriptomic data from two Genlisea specimen was analyzed with the same approach used for *D. muscipula* (methods applied as in section if not stated otherwise). *G. nigrocaulis* and *G. hispidula* are two species from the genus Genlisea which use a lobster pot trapping mechanism to supplement their nutrition in nutrient-poor habitats. In contrast to *D. muscipula* both species are using a passive trapping system and the digestive enzymes are thought to be constantly present. The transcriptome assemblies of both species where comparable to *D. muscipula* (see table). The *G. hispidula* assembly comprised 374,255 isoforms with a total length of 232,855,968 bp. The transcript assembly had a N50 of 998 bp and a N90 of 255 bp (NXX is defined as the shortest sequence length at XX% of the total length). The shortest sequence had a length of 201 bp while longest was 17,173 bp long. The full sequence set was clustered into 285,800 unigenes with a unigenes consisting of less than two isoforms on average. The *G. nigrocaulis* assembly comprised 364,334 isoforms with a total length of 255,510,630 bp. The transcript assembly had a N50 of 1335 bp and a N90 of 265 bp (NXX is defined as the shortest sequence length at XX% of the total length). The shortest sequence had a length of 201 bp the longest was 17,092 bp. The full sequence set was clustered into 326,214 unigenes with a unigene consisting of one isoform on average. The two transcriptomes were annotated using Pfam as profile based method to predict conserved elements such as protein domains and to assign functional classifications. All annotation procedures were carried out on isoform sequences. The *G. hispidula* transcriptome contained 74,120 isoforms represented by 52,228 unigenes with at least one detectable protein domain annotation while the *G. nigrocaulis* transcriptome contained 78,676 isoforms represented by 65,279 unigenes with a proper annotation. Database-derived relations between Pfam domains and GO terms were used to further annotate the two data sets resulting in 42,170 *G. nigrocaulis* unigenes and 32,598 *G. hispidula* unigenes feasible for GO enrichment tests. 5,503 isoforms (represented by 4,191 unigenes) of *G. hispidula* and 4,824 isoforms (represented by 3,881 unigenes) of *G. nigrocaulis* were assigned a high confidence ortholog in *D. muscipula* using the previously described CRB-BLAST procedure. RNA-seq reads were used to quantify transcript expression and resulted in remapping rates between 75 and 87% (see figure 1.13 for a comparison to *D. muscipula*). Expression levels were used to test for differentially expressed genes after a thorough quality control on variance stabilized expression counts. Quality control revealed a unusual clustering of samples in the *G. hispidula* data set (see figure 1.37).

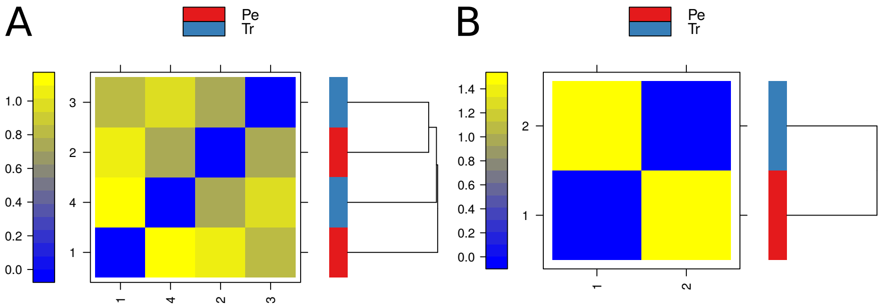


Fig. 1.37 False color heatmap of the distances between samples of RNA-seq experiments from *G. hispidula* (A) and *G. nigrocaulis* (B). The scale is adjusted to the distance range observed in each experiment individually. The distance d_{ab} between two samples a and b was defined as $\delta ab = mean|M_{ai} - M_{bi}|$, where M_{ai} is the value of the i-th unigene of the a-th experiment. RNA-seq sampels from the two biological replicates of *G. hispidula* show an usual clustering. Only samples one and four were considered for further analysis. The removal of biological replicates drastically lowered the statistical power of the differential expression test.

Only samples one and four of the *G. hispidula* data set were further used for the analysis to maximize the conditional differences. As consequence of the missing biological replicates the statistical power of the DE testing procedure was low. Only 261 unigenes were differentially expressed in *G. hispidula* when traps were compared to petioles while 499 were differentially expressed in *G. nigrocaulis* in the same comparison. The subsequent GO enrichment did not show any stress related signals (see table 1.35).

Table 1.35 Gene ontology enrichment result of DEGs detected in *G. nigrocaulis* traps compared to petioles. Similar to the approach on *D. muscipula* only genes with expression count higher than 100 were considered for the enrichment.

Species	Term	Genes	DEG	adj. P-Value
G. nigrocaulis	translation	535	53	1.2×10^{-15}
	photosynthesis	134	14	4.2×10^{-5}
	proton transport	83	9	0.002

To circumvent the low statistical power of the DE testing procedure a second enrichment was carried out on preferentially expressed unigenes as defined by a minimum count of 10 in traps and a fold changes of at least two compared to petioles for both species. *G. hispidula* show 5,302 unigenes preferentially expressed in traps while *G. hispidula* showed 8,505 unigenes preferentially expressed in traps. The subsequent GO enrichment again showed no similar stress signals as in *D. muscipula*.

Table 1.36 Gene ontology enrichment result of PEGs detected in *G. nigrocaulis* and *G. hispidula* traps compared to petioles. Only genes with expression count higher than 100 were considered for the enrichment.

Species	Term	Genes	DEG	adj. P-Value
G. nigrocaulis	translation	535	209	1.4×10^{-5}
G. nigrocaulis	intracellular protein transport	161	75	0.000,78
G. nigrocaulis	ammonium transport	11	9	0.000,88
G. nigrocaulis	gluconeogenesis	17	12	0.001,10
G. nigrocaulis	vesicle-mediated transport	171	62	0.003,30
G. nigrocaulis	SRP-dependent cotranslational protein ta...	22	13	0.007,05
G. hispidula	protein phosphorylation	635	199	3.7×10^{-8}
G. hispidula	microtubule-based movement	43	25	5.3×10^{-7}
G. hispidula	fatty acid biosynthetic process	36	20	1.9×10^{-5}
G. hispidula	cell wall modification	41	20	0.000,21
G. hispidula	carbohydrate metabolic process	406	114	0.000,84
G. hispidula	response to oxidative stress	51	22	0.000,91
G. hispidula	multicellular organism development	17	10	0.001,41
G. hispidula	DNA replication	40	21	0.001,44

A third species, *U. gibba* was tested for similar stress signals as in *D. muscipula*. *U. gibba*, an aquatic carnivorous plant found on all continents except Antarctica, uses ovoid traps attached to its leaf-like structures to capture prey. Similar to the trigger hairs of *D. muscipula*, the plant leverages several setiform branched appendages to detect mechanical stimuli. Upon touch the traps are set off and vacuum the prey into the bladder [272]. A previous transcriptomics study detected a number of non-stress and stress-induced transcripts preferentially expressed in traps. Transcript sequences were used to assign putative orthologous relationships to *D. muscipula* isoforms. Out of 112,424 *U. gibba* transcripts, 1,528 were assigned a one-to-one orthologous relationship to 1,146 *D. muscipula* isoforms and 1060 unigenes respectively. Only 11 of the *D. muscipula* orthologs were stress-induced in *U. gibba* traps while 18 were preferentially expressed in resting *U. gibba* traps. Orthologs to transcripts active in stressed *U. gibba* traps were annotated as ribosomal proteins, heat shock proteins and ubiquitins. Orthologs to transcripts preferentially expressed in resting *U. gibba* traps were partially annotated as phosphatases (PAP10), 14-3-3 proteins, cystein proteases and oxidoreductase. Generally, no clear stress signal was observed, although some of the orthologs were putative members of the *D. muscipula* secretome.

Comparing stress signals of carnivorous and non-carnivorous plants Term-based or gene family-based enrichment studies heavily rely on the experimental annotation of typical model species like *A. thaliana*. The pure presence of stress signals in *D. muscipula* during insect feeding suggest that gene complements are in operation that share a certain amount of sequence conservation or structural similarity in terms of protein domain architectures (which in most cases determine the functional capacities of its underlying sequence). Two tests were carried out to test conservation of the stress signals and sub classify gene complements or functional networks in detail. The first test leveraged gene co-expression clusters thought to be involved in biotic and abiotic stresses in *A. thaliana* as defined by the GEM2Net platform. Based on 13,021 high confidence orthologs between *D. muscipula* and *A. thaliana* (see methods 1.3.7), 5,943 *A. thaliana* genes were selected with an *D. muscipula* ortholog that was differentially expressed after insect stimulation. The selected genes, and the corresponding adjusted P-value of its *D. muscipula* ortholog was used as input to a gene family enrichment against the GEM2Net co-expression cluster assignments. The ortholog gene set was enriched for 9 different clusters associated with various stress responses ranging from biotic to abiotic stresses (see table 1.37) corroborating the hypothesis that insect stimulation-associated stress signals of *D. muscipula* partially rely on gene expression networks that are also active in non-carnivorous species like *A. thaliana*.

Table 1.37 GEM2Net enrichment results. The enrichment was carried out on *A. thaliana* genes with a one-to-one *D. muscipula* ortholog. Only *D. muscipula* orthologs were considered which were differentially up-regulated after insect stimulation. The enrichment was weighted by mapping adjusted P-values of the differential expression test to the respective *A. thaliana* gene.

Cluster ID	Stresses (Project ID)	P-Value	Q-Value
1425	Biotrophic Bacteria	1.21E-05	7.84E-03
127	Fungi	7.90E-05	2.56E-02
881	Oxidative Stress	1.45E-04	2.81E-02
1560	Necrotrophic Bacteria	1.74E-04	2.81E-02
554	Nitrogen	3.82E-04	4.59E-02
1135	Temperature	4.25E-04	4.59E-02
313	Heavy Metal	5.22E-04	4.83E-02
1414	Biotrophic Bacteria	6.92E-04	5.60E-02
1334	Virus	9.59E-04	6.90E-02

The second test leveraged microarray expression data sets from *A. thaliana* where different stimuli were applied to study the plants stress response. The approach followed the same idea underlying the *D. muscipula* transcriptome study. First, DE tests were carried out for each individual data set. The resulting DE gene list and associated adjusted P-values were used as input to a gene ontology based term enrichment. All significantly enriched terms were then used to sub-classify stress signals in *D. muscipula* against various stress experiments from *A. thaliana*. GO terms were compared based on their semantic similarity and a global distance value (as the inverse of the similarity) between a given *A. thaliana* experiment and non-stimulated and insect-stimulated *D. muscipula* traps were calculated. The resulting distance matrix was used to determine the most similar stress experiments from both species. The transcriptomic profile of *D. muscipula* after insect stimulation closely resembles that of *A. thaliana* plants facing herbivore attack or wounding rather than fungal or bacterial infections (see figure 1.38). Non-stimulated experiments of both species cluster accordingly (see figure 1.39).

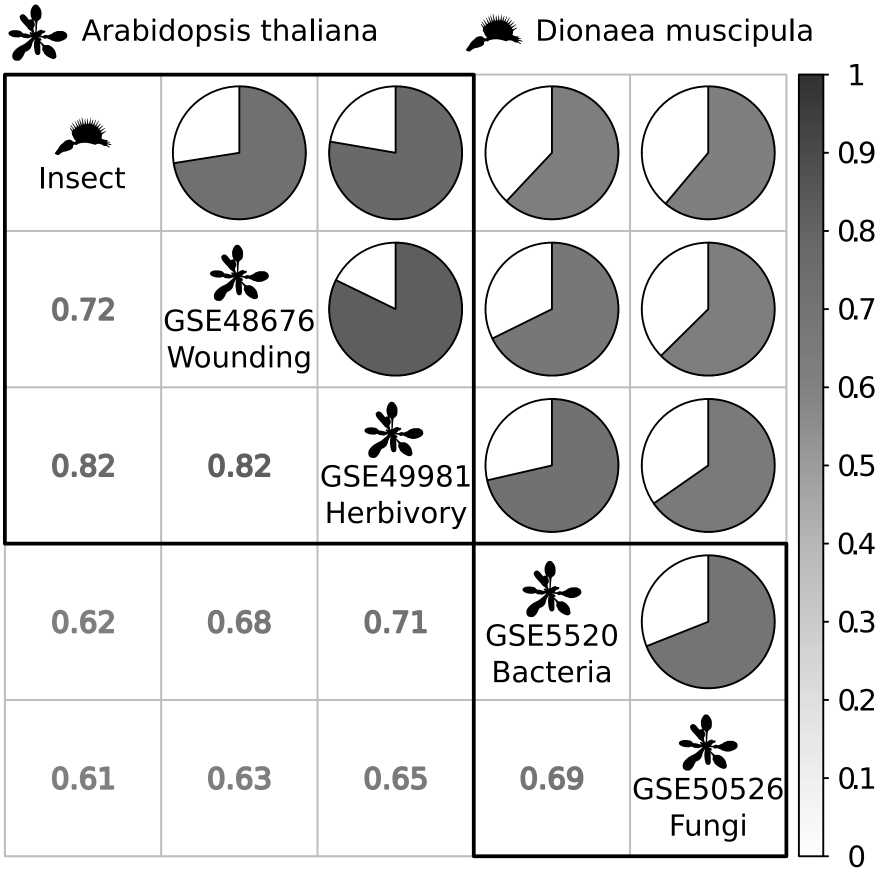


Fig. 1.38 Semantic similarity between different *A. thaliana* microarray experiments and insect-stimulated traps. The semantic similarity is based on the quantitative comparison of all sets of significantly enriched gene ontology terms for each individual expression experiment.

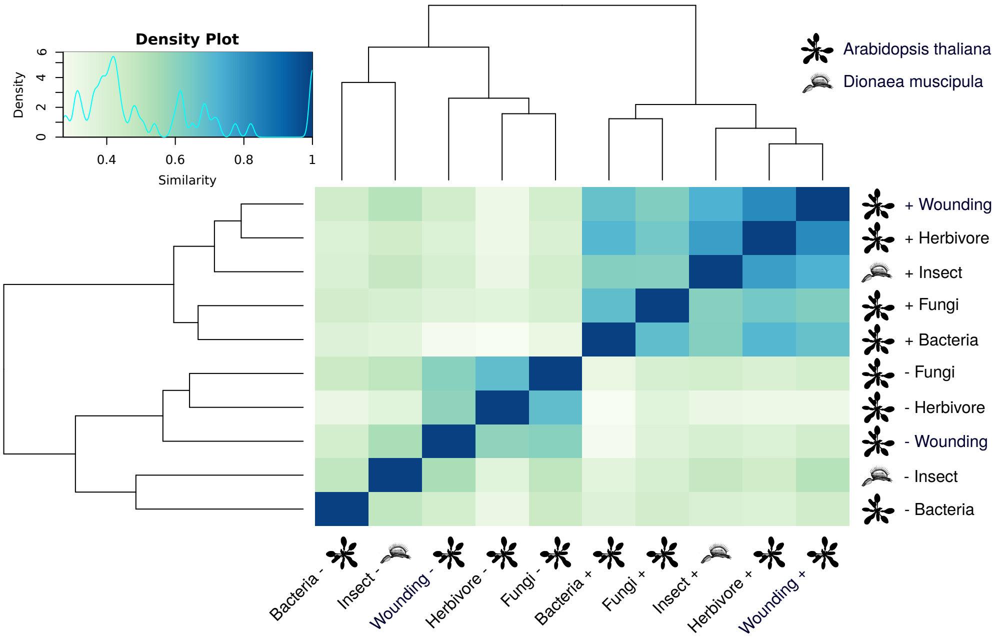


Fig. 1.39 Semantic similarity between different *A. thaliana* microarray experiments and insect-stimulated traps including non-treated controls. The figure is based on the same semantic similarity measurements that have been used in figure 1.38. Controls showed generally less similarity than the different treatments.

1.4.14 Gene family expansions in *D. muscipula*

Manual secretome and transportome annotation revealed several protein families with considerable amount of paralogous unigenes partially expressed after insect stimulation. A common source for such paralogous gene families can be unequal crossing over (or ectopic recombination; [46]), replication slippage (or erroneous DNA replication; [209]), retrotransposition [197, 301], aneuploidy or whole genome duplication. In any case the newly generated genetic material can lead to evolutionary innovation by freeing one or both copies from selective pressure [157]. Most commonly subsequent mutation accumulation causes loss of function or pseudogenization. In rare cases the expansion of certain genetic elements can develop a new or different function (neofunctionalization). In other cases neutral "subfunctionalization" can occur where the original function of the single gene is now performed non-redundantly by the different copies of the duplicated element [265, 96]. To test whether such duplicated genes exist in *D. muscipula* and whether they are heavily expanded in comparison to other eukaryotic species, *D. muscipula* isoforms were assigned to clusters of orthologous groups using the eggNOG databases and its associated mapper. 1,395 isoforms represented by 570 unigenes from *D. muscipula* were assigned to an existing eggNOG (see figure 1.40 for a simple distribution).

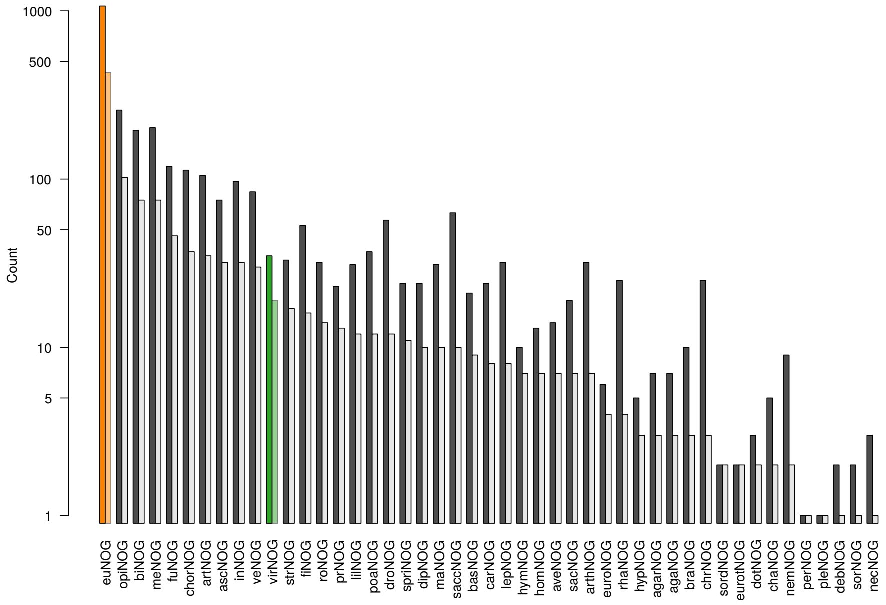


Fig. 1.40 eggNOG assignment level histogram for *D. muscipula*. Counts represent isoforms (dark) and unigenes (light) assigned to eggNOGs. Most isoforms were assigned to euNOGs (eukaryotic NOGs; orange). Interestingly, the second highest NOG count is not virNOG (Viridiplantae NOG, green) but opiNOG (Ophistokonta NOG) probably indicating presence of a eukaryotic contamination in the filtered transcriptome.

In 191 eggNOGs *D. muscipula* unigenes counts were the highest (and therefore considered expanded without statistically testing the expansion). Since several assignments levels exist within eggNOG only the 83 euNOGs (eukaryotic NOGs) were considered. Expansions were distributed across 13 different functional categories with most expansion in euNOGs with unknown function. Only 29 euNOGs showed *D. muscipula* expansions and were properly annotated (see table 1.38). All *D. muscipula* unigenes responsible for the expansions were submitted to a gene ontology enrichment (see section 1.3.7 for methods) to crosscheck the functional categorization of the eggNOG mapping procedure. Results of the gene ontology enrichment compared similar to the functional categories of the expanded euNOGs (see figure 3.4). Several enriched gene ontology terms were associated with response to stress or stimulus response. The responsible unigenes were members of 51 out of 83 euNOGs. A manual inspection of the eggNOGs which comprised 1,516 isoforms represented by 153 unigenes revealed a diffuse annotation pattern. The expanded eggNOGs often contained leucine-rich repeats or were annotated as regulatory proteins such as transcription factors or DNA binding proteins and it is likely that the assignments (eggNOG mapper) and the individual GO annotations (blast2go) are heavily biased by their repeat content.

Table 1.38 euNOG functional categories assignment for *D. muscipula*. Member counts represent the number of euNOGs where *D. muscipula* showed the highest number paralogs.

Members	Symbol	Category
54	S	Function unknown
5	U	Intracellular trafficking, secretion, and vesicular transport
4	D	Cell cycle control, cell division, chromosome partitioning
4	T	Signal transduction mechanisms
3	O	Posttranslational modification, protein turnover, chaperones
3	Z	Cytoskeleton
2	A	RNA processing and modification
2	K	Transcription
2	L	Replication, recombination and repair
1	B	Chromatin structure and dynamics
1	I	Lipid transport and metabolism
1	J	Translation, ribosomal structure and biogenesis
1	W	Extracellular structures

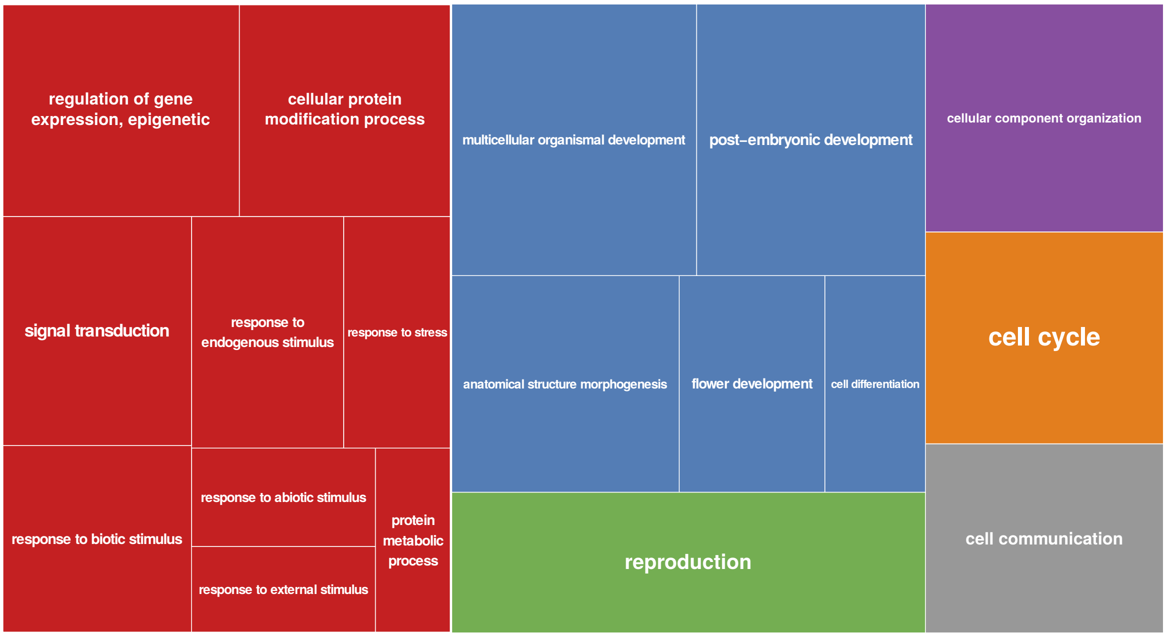


Fig. 1.41 Gene Ontology treemap of terms enriched in expanded *D. muscipula* euNOGs. The plot was drawn with Revigo [269] using Resniks normalized semantic similarity measures and a low similarity score of 0.4 to optimally group similar terms.

1.5 Discussion

D. muscipula is one of the most fascinating plants on earth. It's carnivorous life style interests researchers, gardeners and kids alike since centuries. Already Darwin recognized that the snap traps resemble a sensory-motor system operating at high speed and precision without nerves or muscles [39]. Since then numerous papers have been published describing the morphological aspects of the plant (e.g., the trigger hairs), its lifestyle, the kinetics of trapping, secretion and digestion as well as the electrical properties of the plant, important for its electrical memory and the fast trap closure. Nevertheless, the genetic mark-up necessary to describe each of the proposed models are still unclear in larger parts due to the lack of molecular evidence. The main goal of this project, respectively this chapter of the thesis was to establish a reliable genetic resource that can be used to test several hypothesis regarding the molecular mechanisms driving the plants carnivorous lifestyle. Since a genomic resource was still missing a RNA-seq based transcriptome approach was selected to assemble as much as possible of the coding complement. An initial set of 1.4 billion Illumina reads were assembled into 345,803 isoforms with a total length of 376,689,236 bp. Subsequent correction and filtering reduced the sequence set to a reference transcriptome containing 114,103 isoforms and 51,436 unigenes. Although the pure number of unigenes roughly fits gene count expectations (the *Beta vulgaris ssp. vulgaris* genome encodes 27,421 protein-coding genes) the reference transcriptome still experiences several issues challenging all downstream analysis. Approximately 33,487 isoforms respectively 18,338 unigenes (out of 51,436) only contain partial open reading frames indicating incompletely assembled transcripts. At least 3,172 partial unigenes contain protein domain annotations and are thus unlikely to be non-coding "debris" but real assembly artifacts. Profile-based annotation of the reference transcriptome resulted in 47,347 isoforms respectively 16,163 unigenes having a protein domain (Pfam) while homology-based functional classifications (Gene Ontology) resulted in similar number of classified sequences (47,384 isoforms and 15,344 unigenes). Conservatively calculated the overall annotation rate is less than 50 %. A value that is most likely caused by the small number of evolutionary closely related and well annotated species available for the annotation process. Next to partially assembled isoforms several unigenes are likely to contain sets of highly similar paralogous gene families. Roughly 1,092 unigenes contain multiple isoforms which are in-homogeneously assigned to multiple different *A. thaliana* transcripts as orthologs und thus point towards paralogous multigene-containing unigenes being an artifact of the assembly process. Nevertheless, most transcripts respectively unigenes show a decent completeness (see figure 1.9) if homology evidence is available. Additionally, the reference transcriptome shows a high degree of complete single-copy orthologs specific to plants as estimated by BUSCO. Thus, the current reference transcriptome (version QT1.3) can be considered the by far best genetic resource to understand the carnivorous lifestyle of *D. muscipula* at the moment.

Comparing and understanding global expression patterns of tissues or organs heavily relies on the correct quantification of individual transcripts. Usually, RNA-seq reads are mapped back to a transcriptome reference and counted to calculate the transcript abundance [280, 234, 233]. Ambiguously-mapping reads are usually not considered. Since the reference transcriptome of *D. muscipula* contains numerous isoforms for individual unigenes (most likely as an artifact of the assembler not being able to handle the amount of heterozygocity present in the species) non-ambiguous assignment of reads to their transcript of origin is challenging. Here an approach was chosen that took advantage of the isoform-unigene relation provided by the transcriptome assembler Trinity and effectively combined with the ability of the transcript quantification tool RSEM to work with ambiguously-mapped reads. The strategy allowed to accurately estimate unigene-level and bypassed issues associated with the calculation of isoform-level abundances. Although this process leads to mis-quantification in cases were members of paralogous gene families are clustered into a single unigene it usually provides more stable and accurate results for most of the assembled transcripts respectively unigenes. Quantification results for all experimental data sets showed acceptable remapping rates (see figure 1.11) especially when the different contamination levels were taken into account (see figure 1.5) that often reversely mimicked the mapping efficiency. A stringent quality control and multi-approach outlier detection showed that all experiments showed a sample clustering as desired and unintended batch effects were absent in non-amplified data sets. Furthermore, the quantification process lead to abundance estimates that showed no noticeable problems in the standard deviation of the expression values at different expression strengths (see figure 1.4.4).

Taking advantage of the robust unigene abundance quantification global expression patterns of non-stimulated tissues of *D. muscipula* were characterized. The analysis in cooperating PCA, correlation and differential testing approaches coupled with a Gene Ontology driven description of associated biological processes and molecular functions revealed that traps of *D. muscipula* exhibit the hallmarks of a typical leaf (see figure 1.17). A thorough intersection analysis of unigenes differentially expressed in each tissues showed that the transcriptomic landscape of the traps is a patchwork consisting of unigenes also differentially expressed in other organs. The same trend was true for other tissues or organs. Only flowers exhibited organ-specific expression patterns related to flower physiology, namely anatomical structure morphogenesis, cell growth and the cell cycle. The analysis of principal components describing most of the variability in the non-stimulated data set showed that tissues can be assigned into a photosynthetic active (petiole and trap) and inactive group (roots and flowers) or into reproductive (flower) and non-reproductive (trap, petiole and root) organs thus displaying the typical properties of a green plant. A dissection of traps into glandular (glands), non-glandular (rim) tissue as well as trigger hairs and subsequent transcriptome profiling further revealed that the transcriptomic landscape of the trap is heterogeneous. While the expression profile of the rim strongly correlated with petioles, the expression profile of glands strongly correlated with roots and flowers, both typical sink tissues with secretory capabilities and a highly active molecular transport system. A morphological analysis published side-by-side with the transcriptome data confirmed the dimorphic nature of the glands with a cell layer presumably being responsible for energy storage, and two layers involved in production and secretion of hydrolytic enzymes as well as metabolite shuttling through a highly brush-border like folded, increase membrane surface [25]. Surprisingly, the expression profile of the trigger hairs showed higher similarity to flowers than to glands and roots. Although the sensory cells of the trigger hairs are shaped by apical and basal ER cisternae, numerous mitochondria, as well as vacuoles and lipid droplets, what makes them literally similar to parts of the glands, they mainly exhibited signals related to developmental, especially differentiation processes. Gene ontology enrichment revealed unigenes being active that are linked to anatomical structure morphogenesis mostly to the establishment and maintenance of cell polarity. The analysis of the non-stimulated tissues provides the first transcriptome-wide molecular support for the common assumption that the traps of *D. muscipula* are indeed modified leaves. The analysis of individual trap tissues further showed that only parts of the traps are still photo-synthetically active and cause the strong expression correlation with petioles. On the contrary, glands showed expression profiles related to their secretory and nutrient uptake capabilities and thus provide first insights into carnivorous signatures. Interestingly, even gland specific expression profiles were shaped by a patchwork of genes also active in other tissues (e.g., roots and flowers) thus one can hypothesize that the glands sitting on the photo-synthetically active traps originated from an existing heterotrophic, secreting tissue like roots or floral nectaries.

The second most important goal of this project was to relate (differentially) expressed genes from insect-feeding *D. muscipula* to the plants carnivorous lifestyle. Since *D. muscipula* can be easily switched from a resting state into an insect processing one on demand (using either COR or real prey like crickets) again RNA-seq was used to fully characterize the transcriptomic landscape and provide unprecedented insights into the molecular and physiological processes active. The analysis was carried out on two different tissues/organs, namely insect-stimulated glands and whole traps. While the first was used to describe the secretory and the nutrient uptake system as fine grained as possible, the second was used to assess the trap-wide response generally and with special focus on regulatory and sensory capacities of the traps. The analysis of the secretome revealed the hydrolytic cocktail consists of at least 17 highly abundant enzymes with proteolytic, oxidative or hydrolytic activity. Most of them were stimulated by Insect treatment while a smaller number where transcribed highly in non-stimulated glands. The latter indicates that some components of the digestive cocktail might act as first strike against the prey. A detailed kinetic study of selected hydrolases (e.g., VF CHITINASE I) further showed that expression levels reach peak level after 24 - 48h and that expression can be boosted depending on the nutrients nature thus pointing towards an active chemo-sensing to assess prey presence and quality [25].

The assessment of the nutrient uptake system uncovered a diverse set of transporters active in both non-stimulated and insect-stimulated glands. Especially insect-stimulated glands showed an overall expression profile geared towards the acquisition of prey-derived nutrients such as nitrogen-containing solutes but also inorganic cations like potassium that are require for tugor formation or growth [294, 215]. The latter is particular important during the reopening of the traps as well as during development of new ones [289]. The overall transporter profile suggests that *D. muscipula* uses metabolites as well as macro molecules like amino acids, peptides or nucleotides released during prey digestion to counteract the male nutrition it experiences in its native habitat. The transport system is likely to be driven by electrochemical gradients maintained by multitude of different P-Type H^+ ATPases of which some seem specifically activated through insect stimulation. Additionally, the glands shows signs of an accelerated endocytosis system, indicating that nutrient resorption through transporters and channels might be complemented with a system engulfing prey-derived macro molecules such as proteins for selective update.

The analysis of the trap-wide response after insect stimulation confirmed that conserved stress-related pathways indeed play a central role during prey digestion. Transcriptomic profiles of insect-stimulated traps showed typical signs of a wounding response mostly caused by a strong up regulation of components of the JA biosynthesis pathway. The wounding response was complemented by a high activity of ROS scavenging systems as well as a regulatory systems controlling stress- and pathogen-induced cell death. While the first signal likely originated in glands as consequence of the highly active translation machinery, the second might indicate an effective strategy that shields *D. muscipula* traps from detrimental ROS effects. Gene family enrichment (see table 1.30) further displayed signals of elevated transcription factor activity and a pronounced signaling transduction network. A detailed analysis of transcription factors families active after insect-stimulation indicated a strong up regulation of AP2/ERF-like transcription factors. While most of them were involved in stress responses like JAZ1, there were a number related to flowering time control (CDFs) and developmental growth processes (e.g., DORNRÖSCHEN). The latter is especially interesting since reopening of the traps is thought to involve growth processes in certain trap lobe regions [37, 297, 224]. The TF analysis suggests a crucial role for several transcription factor families (e.g., AP2/ERFs, WRKYs or MYBs) to orchestrate the insect-feeding response, a result similar to other stress response studies in plants [209, 307, 169]. Nevertheless, the role of individual transcription factors remains unresolved due to further experimental evidence and the complete lack of knowledge about the TF-miRNA co-regulatory networks operating in *D. muscipula*. Analysis of the active signal transduction network reveled a high transcriptional activity of Tyrosine kinases, Ca^{2+}/Calmodulin-dependent protein kinases and CGMC kinases, most of them putatively cell-surface receptor kinases (RLKs). Among them LysM-type chitin oligosaccharide-responsive CERK1-like kinase [205] was significantly upregulated. It's high activity might indicate the digestion process builds on the ability to assess the chemistry and quality of the prey progressively and adjust the secretion process accordingly. Interestingly, a homolog of the defense response associated receptor-like kinase FLS2, involved in perception of bacterial flagellin, was highly expressed after insect stimulation but less active after COR treatment. Similar to the CERK1-like kinase, the expression patterns of FLS2 suggests a prey specific response not only tailored to the prey itself but maybe also to the microbial load associated with it.

Since it has been previously shown that *D. muscipula* is rarely infected by microbes (Tokunaga, 2004) even though the plant meets a completely new prey-associated microbiome during insect-feeding the microbial load and possible changes were assessed again making use of the "contaminated" RNA-seq data from the transcriptomic study. Although a substantial amount of sequencing reads were assigned to the bacterial kingdom, less was assigned to meaningful taxonomic levels like phylum or order. A principal component analysis as well as conditional tests only suggested a differences between resting plant organs (exp001) and the insect stimulation experiment (exp008) probably indicating unintended experimental batch effects rather than a biological signal. The comparison of a prey associated microbial profile (derived from transcriptomic data generated from pure crickets) showed no higher similarity to the insect-stimulated RNA-seq samples than to others either indicating that most of the prey derived microbiome already vanished in the progress of the digestion process (*D. muscipula* insect-stimulated samples where extracted after 24h) or that the experimental methods and conditions where not comparable. Since recent studies in other carnivorous plant species suggested (beneficial) effects of the microbial community on the host system future studies probably should leverage microbial profiling through 16S sequencing or similar approaches.

Early molecular studies of carnivorous plants suggested several conserved defense related processes take part in the insect-feeding process. Using the quantified *D. muscipula* transcriptome, its annotation and enriched terms representing biological functions this hypothesis was tested in a comparative manner with two different approaches respectively comparisons. A direct comparison of expression profiles from other carnivorous plants (e.g., *U. gibba*, *G. hispidula*, *G. nigrocaulis*) with the expression profile of non-stimulated and insect-stimulated *D. muscipula* traps revealed no greater similarity between them. The hypothesis that the defense response is conserved across carnivorous plants needs to be rejected on the obtained results. But, it is highly likely that the data quality, depth and nature of the experiments are simply not comparable and a likely signal is hidden. In case of *G. hispidula* and *G. nigrocaulis* missing biological replicates between conditions as well as the unexpected clustering of the present samples make it nearly impossible to detect differentially expressed genes and run a proper functional enrichment or semantic similarity analysis. Same is true for the *U. gibba* data set where differential expression testing results are completely absent and the assignment of stress responsiveness of a genes is purely based on the preferential expression scheme that a previous study introduced [142]. Again, results would reject the hypothesis of a conserved stress response but the input data does not qualify to test hypothesis after all. The second hypothesis tested, was that *D. muscipula*s "stress" response after insect stimulation is at least partially driven by expression networks also active in non-carnivorous plants. The test revealed that several clusters of co-expressed genes are a) conserved between species and b) are both stress triggered. Due to the incompleteness of the GEM2net database a precise classification of *D. muscipula*s stress response was not possible. Especially the lack of wounding or "feeding" experiments (e.g., aphid feeding, [191]) limited this approach. A third test, designed to overcome the limitations of the latter, was used to test whether stress response signals present in *D. muscipula* can be sub classified using semantic similarity of gene ontology enrichments. A comparison of *A. thaliana* stress signals triggered by various different stimuli revealed that the stress response of *D. muscipula* resembles that of *A. thaliana* plants facing herbivore attack or wounding rather than fungal or bacterial infections. Especially, the presence of terms related to jasmonic acid biosynthesis and action drove herbivore attack, wounding and feeding experiments together. Noteworthy, each individual stimulus was only represented by a single experiment and thus the analysis has a limited statistical power although individual enrichments are based on differential testing results that incorporated biological replicates. In summary, tests showed conservation of stress response systems across *D. muscipula* and non-carnivorous species like *A. thaliana* but not between *D. muscipula* and other carnivorous plants most likely due to the poor resolution or design of the underlying data set. A recent genome study of the heterophyllous pitcher plant *Cephalotus follicularis* revealed expression differences between carnivorous and non-carnivorous leaves that encode biological processes related to prey attraction, capture, digestion and nutrient absorption [100]. Especially the digestive fluid exhibited a similar composition in comparison to *D. muscipula*. Again, results suggested that pathways involved in the carnivorous syndrome are at least partially conserved and that the numerous species within the "non-core" Caryophyllales are a perfect framework to understand evolutionary paths that led angiosperms to become carnivorous.

1.6 Epilogue

The transcriptome studies presented in this thesis provide first insights into the molecular basis of the carnivorous plant *D. muscipula*. It closes the gap between observations that already Darwin had made and our physiological understanding of the carnivorous syndrome in plants. The work systematically cataloged biological processes and molecular activities which shape the morphological and anatomical appearance of this astonishing plant. It presents the first molecular data that separates resting from active trapping organs, confirms that resting traps still operate in a leaf-like manner and for the first time provides evidence that the carnivorous syndrome builds on conserved pathways and expression patterns during capturing and digestion. Nevertheless, the work also demonstrates how limited our methodological capabilities are even in the advent of high throughput screening methods such as RNA-seq or proteomics. A substantial amount of the quantitative and qualitative transcriptome that is active while the plant captures and digest its prey lacks any functional assignment. Thus it is left out during enrichment attempts or comparative analyses and negatively biases our view on the plants nature. Future studies need to carefully complement quantitative technologies such as RNA-seq with qualitative, functional screens at the DNA (e.g., genetic interaction mappings), the RNA (e.g., ribosome profiling) and the protein (e.g., affinity purification and mass spectrometry) level to fully understand functional elements of a species like *D. muscipula*, their regulation, site of action and origin.

1.7 Published Elements

Bemm, F., Becker, D., Larisch, C., Kreuzer, I., Escalante-Perez, M., Schulze, W. X., Ankenbrand, M., Weyer, A.-l. V. D., Krol, E., Al-rasheid, K. A., Mithöfer, A., Weber, A. P., Schultz, J., and Hedrich, R. (2016a). Venus flytrap carnivorous lifestyle builds on herbivore defense strategies. *Genome Res.*, 26(6):1–14 Gao, P., Loeffler, T. S., Honsel, A., Kruse, J., Krol, E.,

Scherzer, S., Kreuzer, I., Bemm, F., Buegger, F., Burzlaff, T., Hedrich, R., and Rennenberg, H. (2015). Integration of trap- and root-derived nitrogen nutrition of carnivorous Dionaea muscipula. *The New phytologist*, 205(3):1320–9 Scherzer, S., Böhm, J., Krol, E., Shabala,

L., Kreuzer, I., Larisch, C., Bemm, F., Al-Rasheid, K. A. S., Shabala, S., Rennenberg, H., Neher, E., and Hedrich, R. (2015). Calcium sensor kinase activates potassium uptake systems in gland cells of Venus flytraps. *Proc Natl Acad Sci U S A*, 112(23):7309–7314 Schulze,

W. X., Sanggaard, K. W., Kreuzer, I., Knudsen, A. D., Bemm, F., Thogersen, I. B., Brautigam, A., Thomsen, L. R., Schliesky, S., Dyrlund, T. F., Escalante-Perez, M., Becker, D., Schultz, J. J., Karring, H., Weber, A., Hojrup, P., Hedrich, R., and Enghild, J. J. (2012). The Protein Composition of the Digestive Fluid from the Venus Flytrap Sheds Light on Prey Digestion Mechanisms. *Molecular & Cellular Proteomics*, 11(11):1306–1319

Chapter 2

TBro: Visualization and management of denovo transcriptomes

2.1 Preface

The following chapter presents TBro, a published software package developed as central hub for transcriptome data sets and curated annotations from the Carnivorom project. See table 2.1 and 2.2 for project responsibilities and contributions.

Publication: Ankenbrand, M. J., Weber, L., Becker, D., Förster, F., and Bemm, F. (2016). TBro: visualization and management of de novo transcriptomes. *Database*, 2016(0):baw146

Table 2.1 Statement of individual author contributions and of legal second publication rights. Responsibilities and contributions in decreasing order from left to right. Abbreviation: MJA = Ankenbrand, M. J.; LW = Weber, L.; FF = Förster, F.; DB = Becker, D.; FB = Bemm, F.

Participated in				
Study Design	FB	FF, MJA	DB	LW, MJA
Methods Development	LW	MJA	FB	
Data Collection	FB			
Data Analysis and Interpretation	FB	NJA	FF, DB	
Writing of Introduction	FB	MJA	All others	
Writing of Materials & Methods	MJA	FB	All others	
Writing of Discussion	FB	MJA	All others	
Writing of First Draft	FB,MJA			

Table 2.2 Statement of individual author contributions to figures/tables/chapters included in the manuscripts. Responsibilities and contributions in decreasing order from left to right. Abbreviation: MJA = Ankenbrand, M. J.; FB = Bemm, F.

Participated in		
Figure 1	FB	MJA
Figure 2	FB	MJA
Supplementary Table S1	MJA	
Supplementary Figure S2	FB	

The doctoral researcher confirms that she/he has obtained permission from both the publishers and the co-authors for legal second publication.

Felix Bemm January 22, 2018 Würzburg

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment.

Prof. Jörg Schultz January 22, 2018 Würzburg

Database, 2016, 1–7
doi: 10.1093/database/baw146
Original article

Original article

TBro: visualization and management of *de novo* transcriptomes

Markus J. Ankenbrand[1,†], Lorenz Weber[2,3,†], Dirk Becker[4], Frank Förster[2,3] and Felix Bemm[2,5,*]

[1]Department of Animal Ecology and Tropical Biology, Biocenter, Am Hubland, 97074 Würzburg, Germany, [2]Department of Bioinformatics, Biocenter, Am Hubland, 97074 Würzburg, Germany, [3]Center for Computational and Theoretical Biology, University of Würzburg, 97074 Würzburg, Germany, [4]Institute for Molecular Plant Physiology and Biophysics, University of Würzburg, 97082 Würzburg, Germany and [5]Department Molecular Biology (Detlef Weigel), Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany

***Corresponding author:** E-mail: felix.bemm@tuebingen.mpg.de

Present address: Felix Bemm, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany

[†]These authors contributed equally to this work.

Citation details: Ankenbrand,M. J., Weber,L. et al. TBro: visualization and management of de novo transcriptomes. *Database* (2016) Vol. 2016: article ID baw146; doi:10.1093/database/baw146

Received 18 April 2016; Revised 12 October 2016; Accepted 18 October 2016

Abstract

RNA sequencing (RNA-seq) has become a powerful tool to understand molecular mechanisms and/or developmental programs. It provides a fast, reliable and cost-effective method to access sets of expressed elements in a qualitative and quantitative manner. Especially for non-model organisms and in absence of a reference genome, RNA-seq data is used to reconstruct and quantify transcriptomes at the same time. Even SNPs, InDels, and alternative splicing events are predicted directly from the data without having a reference genome at hand. A key challenge, especially for non-computational personnal, is the management of the resulting datasets, consisting of different data types and formats. Here, we present TBro, a flexible *de novo* transcriptome browser, tackling this challenge. TBro aggregates sequences, their annotation, expression levels as well as differential testing results. It provides an easy-to-use interface to mine the aggregated data and generate publication-ready visualizations. Additionally, it supports users with an intuitive cart system, that helps collecting and analysing biological meaningful sets of transcripts. TBro's modular architecture allows easy extension of its functionalities in the future. Especially, the integration of new data types such as proteomic quantifications or array-based gene expression data is straightforward. Thus, TBro is a fully featured yet flexible transcriptome browser that supports approaching complex biological questions and enhances collaboration of numerous researchers.

Database URL: tbro.carnivorom.com

© The Author(s) 2016. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Background

RNA sequencing (RNA-seq) provides a fast and cost-effective method to access transcribed genes in a qualitative and quantitative manner (1, 2). Without prior knowledge this technology enables transcript discovery and quantification at the same time (3). In particular, for non-model organisms and in absence of a reference genome, RNA-seq has been proven a successful strategy to elucidate the role of candidate genes in physiological pathways or developmental programs as well as the underlying molecular mechanisms (4–7).

Nowadays, transcriptome assemblers such as Velvet/Oases (8) and Trinity (9, 10) are capable to accurately reconstruct full length transcripts, even for recently duplicated genes or alternative splice isoforms from RNA-seq data. Most assemblers operate over a broad range of expression levels. The assembled sequences are usually organized into hypothetical genes (unigenes) represented by multiple isoforms. Those isoforms are usually searched for candidate coding regions. Their deduced proteins are annotated by employing homology as well as profile based methods such as InterproScan (11) and Mercator (12). Furthermore, reusing the generated RNA-seq data for tools like RSEM (13) or Salmon (14) provide quantification of isoforms and their subordinate unigenes. Quantification results serve as input for differential expression (DE) testing, one of the major applications of RNA-seq. Both DE testing results as well as isoform annotation are subject to Gene Ontology or gene family enrichment analysis with tools like topGO (15) or GAGE (16) on either whole transcriptomes or curated subsets.

In the end, most *de novo* RNA-seq studies result in a multitude of different datasets, including sequences, their annotation, expression levels and DE as well as co-expression testing results. Since most of the datasets contain thousands of entries they remain hard to handle. The vast amount of different data types necessitates the usage of a simple interface, optimally through a web browser, to allow uniform data access also for non-IT personal. Researchers need to refine functional annotations (e.g. unigene/isoform synonyms or descriptions) or flag individual unigenes or isoforms with personal metadata. Additionally, classification of biologically related unigenes or isoforms into functional groups or protein families is often pivotal to help understanding their specific roles and interplay in given pathways and networks. Currently, only a small number of tools and platforms are available that provides these basic functions. Most tools are tailored for genome reference based RNA-seq studies [e.g. Tripal (17, 18), Intermine (19), TraV (20), RNASeqExpressionBrowser (21)] or aim for a specific species [e.g. dbWFA (22)] with Intermine and Tripal the most feature rich and best maintained tools available. Intermine is specifically designed for the integration and analysis of complex biological data sets on top of genome annotations but comes with a higher hardware footprint and a complex backend not ideal for smaller lab environments. Tripal on the other hand, serves as online biological knowledgement system displaying predefined queries and thus making it inflexible for large amounts of different user requests. Only TrinotateWeb (23) provides a unified way to create, organize, and visualize results from de novo transcriptome studies. However, it allows no multi user access, lacks the ability to store user-defined unigene or isoform collections, offers only a very sparse search interface and is not capable to provide pathway information. Beyond that, it is hard to extend since the back-end does neither rely on a documented database schema, such as Chado nor does the front-end make use of a modular web service system necessary for new visualizations or analyses. Here we present TBro, a flexible *de novo* transcriptome browser, written to overcome the above-mentioned constraints thereby enabling researchers to analyse and share their data in a collaborative and standardized manner.

Features

TBro represents an easy to use multi-user *de novo* transcriptome data mining platform. It is developed as web application, works across platforms, and is browser independent. The TBro interface provides structured access to a given transcriptome and its annotation by modelling unigene → isoform relations. Unigene subpages (e.g. http://tbro.carnivorom.com/tbro/details/byId/439690) offer a tabular list of all available isoforms including high level visualization functions for expression profiles and DE testing results. Similarly, isoforms are presented on individual comprehensive subpages allowing users to inspect annotations and metadata (e.g. synonyms and descriptions) as well as enabling visualization of analysis results (e.g. quantifications or DE testing results) dynamically in one place (e.g. http://tbro.carnivorom.com/details/byId/439692). Isoforms and annotated peptides are sent directly to NCBI's blast suite (24). Annotated features like repeats, predicted peptides and interpro hits are displayed in an overview graph and listed as separate tables. If available, a link to the underlying external database entry is provided. Simple annotations like Gene Ontology terms, MapMan bins and Enzyme commision numbers are displayed underneath. All coordinate-based annotations (e.g. open reading frames, protein domains) as well as expression profiles and differential expression results are visualized by CanvasXpress (25).

The visualization itself as well as the underlying data tables are modified dynamically using the context menu of the CanvasXpress library. Users can simply change graphical parameters, scaling and limits of the plots as well as transform or correlate them in different ways. In addition, users can add a custom alias and description for each unigene or isoform at the top of each subpage. Advanced users can use TBro's web services as an application programming interface (API) to access and integrate data into other applications.

One of TBro's major achievements is the implemented cart system to comfortably organize and analyse user-specified collections of unigenes and isoforms. They are compiled from the underlying transcriptome database by different exploration methods. Users can select unigenes or isoforms of interest by homology searches (e.g. BLAST), annotated protein signatures (e.g. Interpro) or pathway assignments (e.g. KEGG) as well as through fine grained filtering of expression and differential expression results. Furthermore, users can search for unigenes and isoforms by their id or alias or enter complete paragraphs of a paper to mine them for potential hits. The search for an id or alias is carried out in a strict mode to perfectly match a database entry or in non-strict mode to expand the results to related entries. The latter is used to easily retrieve all isoforms for a unigene. Resulting hits are further refined by simple string or data type specific filters. Results are usually displayed as tables and selected rows can easily be added to a cart via the table menu or simply by drag and drop onto the desired cart. Carts are rapidly synchronized between tabs within a browser session and user can share them in a collaborative manner using TBro's controlled import and export functions.

Whole carts are visualized similar to individual unigenes or isoforms. Expression results are displayed as heat map for multiple selected conditions or tissues. Results from DE tests are graphed in a Bland–Altman plot [MA plot; (26)]. The latter is especially useful to localize selected unigenes or isoforms within the context of an entire expression experiment. Users can annotate Carts with an alias as well as a detailed description and store the cart itself and its corresponding annotation within TBro's database. The OpenID-based user authentication system enables hundreds of users to store personal annotations generically however eliminating the need for its own centralized login system.

Implementation

TBro is divided into three environments (Figure 1A). The user environment (Figure 1A, light grey) consists of a client interface and an admin interface, which is used to control TBro. The admin tools are implemented in PHP with a command line interface (CLI) using multiple pear packages (Log, Console_CommandLine, Console_Table and Console_ProgressBar), propel for database abstraction (object-relational mapping), and phing for setting up databases and web interfaces. The client interface is structured using PHP and javascript with the Foundation Front-end framework. User interface interactions such as drag and drop capabilities, effects, widgets are built with the jQueryUI library. Displayed tables are created using the DataTables plugin for jQuery to make tables searchable and add multi-column ordering functionalities. Experimental as well as sequence annotation data are visualized using the CanvasXpress (25) graphing library. The Front-end is developed under the convention of the Document Object Model (DOM). DOM traversals, modifications and event binding are handled with jQuery. Ajax (Asynchronous JavaScript and XML) is used to update the parts of the frontend without reloading it completely. Data collections, arrays, and objects are manipulated using Underscore.js. Dynamic content is directly injected into the front-end using Underscore.js client-side templating.

TBro's core environment (Figure 1a, black) consists of an Apache web server, delivering the web interface and providing core functionalities as atomic web services as well as a PostgreSQL server hosting the modified Generic Model Organism Database (GMOD) Chado database (27). Caching capability is provided by a memcached (28) server. The separated provision of each component provides high-availability and allows for resource optimizations (e.g. load balancing). REST Web services are written in PHP and return results formatted as JavaScript Object Notation (JSON). Database queries are logged and optimized using loggedPDO. Users are authenticated with lightOpenID. User session data is stored with webStorage on the client side to optimize server requests. Sequences and sequence annotations are stored using the Chado sequence module. Relationships between features such as unigenes and isoforms or proteins and protein domains are modelled using the feature relationship table. Quantification and DE testing results are stored in two newly introduced tables. Both tables complement the Chado Mage module to easily store non-microarray expression data. Future releases will store tabular data (e.g. quantification and DE testing results) using PostgreSQL NoSQL capabilities to speed up requests. User annotation data from carts and individual annotations are kept in a specifically created table (webuser_data). User data received from the front-end is inserted as decomposed binary format (JSONB).

The analysis environment (Figure 1A, dark grey) is used to perform computations like BLAST searches. Jobs are triggered by users via the web browser and tracked in a separate database. An arbitrary number of workers on

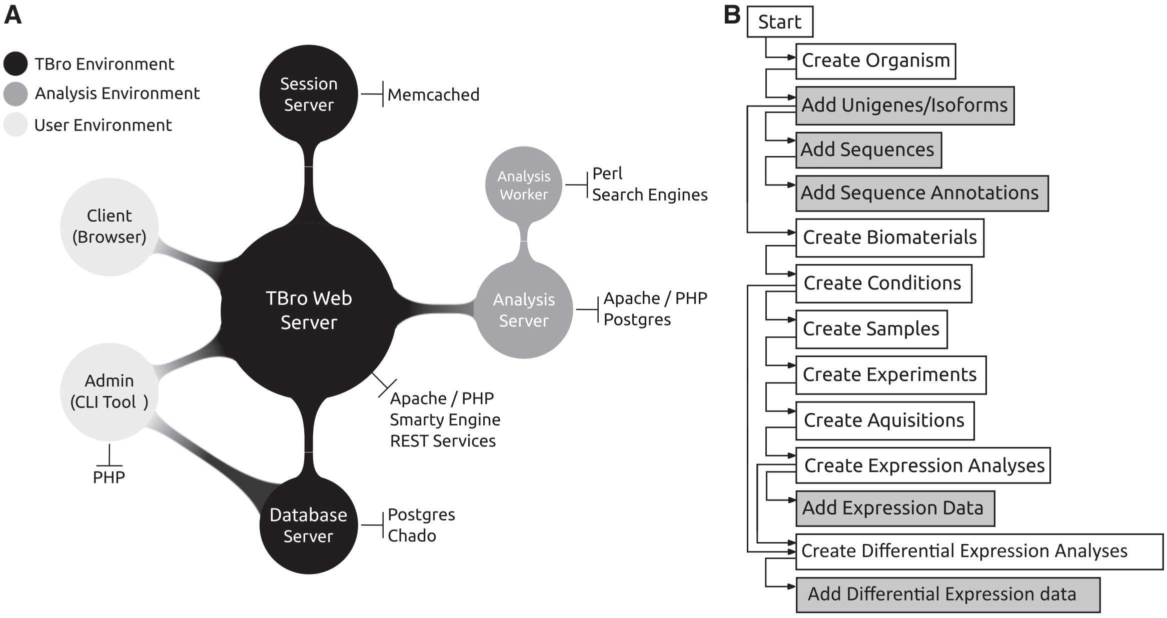


Figure 1. (A) TBro's architecture is divided into three sections. The TBro environment builds the backbone with the central web server. The web server is connected to the database server and the session server for caching. The analysis environment is used to perform computationally intensive tasks. It is divided into a server and an arbitrary number of workers that can run on heterogeneous systems. The user environment consists of the client (a web browser) which is used to interact with a running instance of TBro and the command line tools which are used to import and manage data by a qualified administrator. (B) A typical data import hierarchically prepares and adds all transcriptomic data sets. Tasks performed by TBro-db are coloured in grey while tasks performed with TBro-import are coloured in white. The complete workflow tightly builds on the reference Chado schema to ease maintenance and usability.

heterogeneous host systems (currently Linux and Windows are supported) is utilized to run the job. Workers query the database for unallocated jobs, run them and report the results back to the database. The status of the job and eventually the results are accessible by the user via a unique URL. The analysis environment builds on a modular structure to easily extend it to other tools (e.g. HMMER for profile based searches).

Usage

TBro knows two principal roles: administrator and user. The administrator imports and manages data using a CLI while the user accesses and searches the data with a web browser. The CLI is divided into three subcommands, TBro-db for managing data values (list, insert, edit and delete of e.g. contacts, organisms), TBro-import for importing multiple data values from files (e.g. ids, sequences) and TBro-tool which provides helper scripts (e.g. format converter). All tools come with support for auto completion in Linux environments. The CLI tools hierarchically prepare and import all data sets but can also be used to retrieve data from the database. An exemplary import workflow is available in the TBro documentation (http://tbro-tutorial.readthedocs.org). Sequence information and relations are imported by supplying relation maps (Unigene → Isoforms and Isoform → Open Reading Frame) and simple fasta files. The same is done for generic pathway associations (EC → KEGG Map). Annotation results are imported using a two-column tab-separated file (Sequence ID → GO/EC/Synonym) or source-defined multi-column files (Interpro, RepeatMasker, MapMan). Expression counts and DE results are imported after deep modelling the sample relations with TBro's database control tool (TBro-db, Figure 2B). Each expression dataset is associated with a biomaterial (e.g. tissue), a condition (e.g. treatment) and a sample name (e.g. replicate-1) according to the Chado database schema. The combination of biomaterial, condition and sample name is connected with an experiment. Each experiment is assigned to one or multiple acquisitions corresponding to a sequencing runs or array hybridization. Acquisitions are associated with a corresponding analysis e.g. quantification and normalization of unigene and isoform counts or DE test results. Finally, the datasets are imported by simply supplying a quantification and analysis id.

The online demo (http://tbro.carnivorom.com) hosts data from the recently published Venus flytrap *(Dionaea muscipula)* deep transcriptome sequencing project (Bemm et al., 2016, in press). The unfiltered data sets contain 315 584 isoforms for 183 578 subordinate unigenes. A total of 3 221 001 annotation entries of various types are stored within TBro's database backend. Expression data

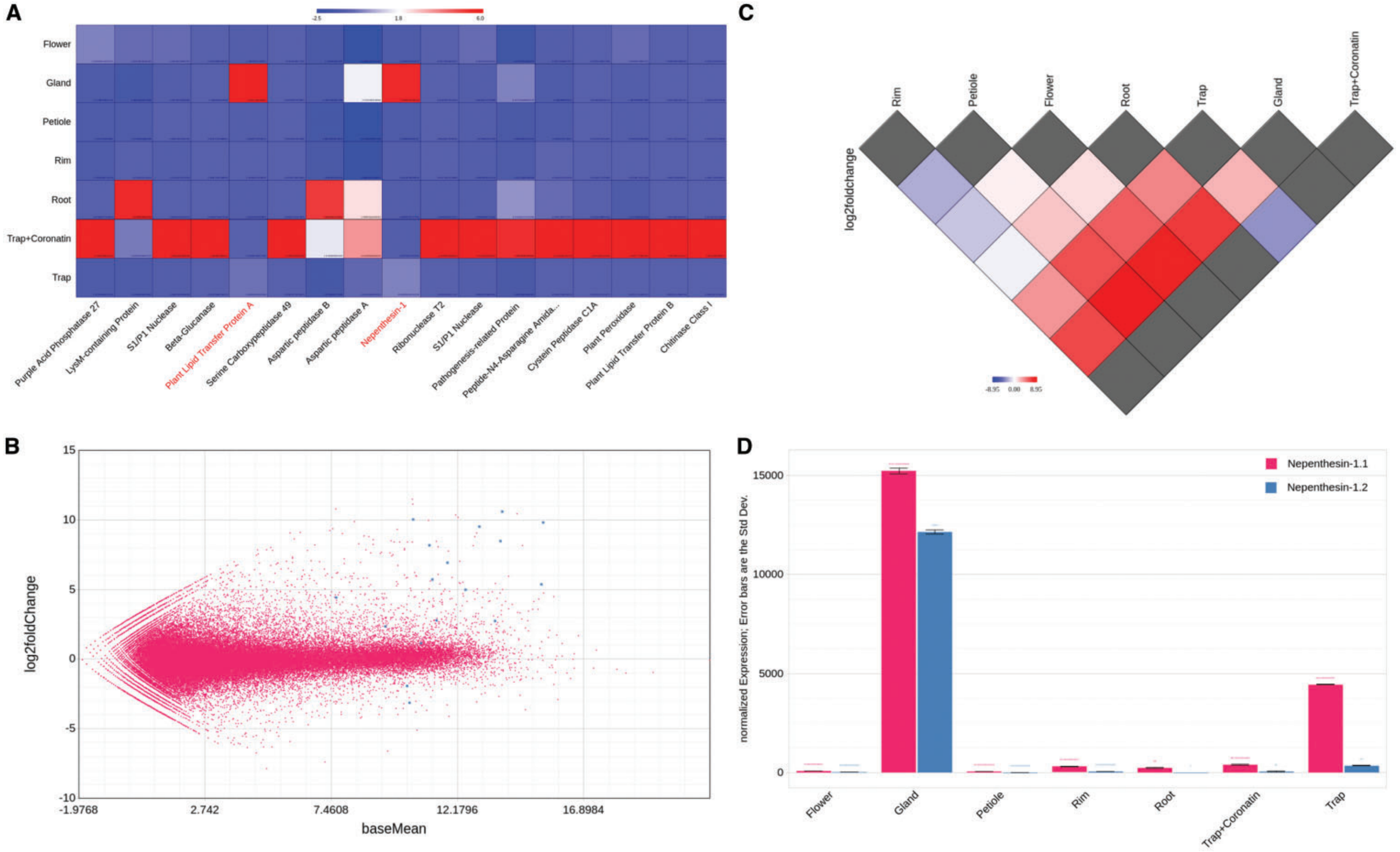


Figure 2. (A) Z-transformed expression heatmap of a cart containing putative members of the hydrolytic cocktails secreted by Venus flytrap during its hunting cycle. Two unigenes are being expressed in a non-stimulated gland specific manner. (B) MA plot for the same cart based on DE testing results from DESeq. The plot indicates that most members of the hydrolytic cocktail are being highly expressed compared to the majority of the unigenes. (C) Triangular visualization of the DE testing results for an individual gene (Nepenthesis-1). (D) Simple expression barplot of the Nepenthesin-1 gene with two isoforms showing different expression patterns. All plots were generated directly in TBro. Z-transformation, scaling and layouts were adjusted using functions from CanvasXpress context menu directly in the browser.

from four experiments with a total of 39 samples contain 19 467 318 distinct expression values and results for 2 744 423 DE comparisons are aggregated. The total size of the PostgreSQL database on disk is approximately 14 GB. All components of the Venus Flytrap TBro instance are running on a single virtual machine [Intel(R) Xeon(R) CPU E5-2640 v3, 2 cores, 8 GB RAM, Ubuntu 12.04, 64 bit].

One of the major questions during the deep transcriptome sequencing project of the Venus flytrap was about the nature and abundance of the hydrolytic enzymes which are secreted by specialized glands on the inner trap surface to digest animal prey. Several high-throughput proteomics experiments using different stimuli (insect and hormone treatment as well as mechanical stimulation) were conducted to stimulate secretion and detect hydrolytic enzymes in *Dionaea's* digestive fluid. Following sampling of the secretion fluid, peptides were identified by mass spectrometry and mapped onto the reference transcriptome. Thereby 368 isoforms, respectively their deduced proteins, were identified as secreted independent of the nature of the stimulus. The resulting isoforms were searched within TBro and stored using its cart system. This initial 'secretome' cart was searched for entries exhibiting an annotated signal peptide (indicative for secreted proteins) employing the cart annotation search. Eligible isoforms were added to a 'filtered secretome' cart. Subordinate unigenes were added to the new 'filtered secretome' cart via the table menu and DE results from insect-stimulated glands (exp008) were visualized using a MA plot (Figure 2B). It became immediately obvious that the hydrolytic cocktail consists of enzymes being already expressed in non-stimulated glands (Figure 2B, blue dots with $\log_2$ fold change < 0, 2 unigenes) and those triggered upon insect stimulation (Figure 2B, blue dots with $\log_2$ fold change > 0, 15 unigenes). The two differentially expressed unigenes in non-stimulated tissues were further analysed with TBro's triangular DE plot using an expression experiment comprising different non-stimulated tissues (exp001, Figure 2C). This plot revealed that the two unigenes, encoding Nepenthesin-1 and a Lipid Transfer Protein (LTP), are indeed excessively transcribed in a gland specific manner. The refined cart was directly used as supplementary data for the publication and to ease the review process.

Altogether, TBro successfully enhanced collaboration of numerous researchers working in the Venus flytrap transcriptome project team. It was particularly helpful to visualize expression strength or expression variability in publication-associated carts (Figure 2A). It was further intensively used to identify representative isoforms for individual unigenes using its adjustable expression bar plots (Figure 2D). Researchers frequently visualized DE test results using TBro's triangular DE plot (Figure 2C) to identify DE patterns over a large set of different tissues. Finally, TBro's pathway module was used to provide functional associations (e.g. Jasmonic acid biosynthesis, Supplementary Figure S2).

Conclusion

TBro provides simple-to-use interfaces to (i) inspect and refine functional annotations, (ii) analyse and visualize expression as well as (iii) DE testing data. It handles user derived sets of unigenes/isoforms as well as entire experimental datasets and thus outperforms competing packages in terms of functionality, user-friendliness and flexibility. The cart system helps collecting, organizing and sharing biological meaningful sets of unigenes/isoforms and thus offers an effective way to export meta-data for external review. Building on the Chado database schema empowers TBro to handle complex representations of biological knowledge and a multitude of different data types. Although TBro was developed with RNA-seq experiments in mind, it can easily be adopted to host proteomic or other quantification data sets. Furthermore, it provides interoperability between different biological databases and applications of the GMOD toolkit. The modular backend, organized into different environments and the heavy use of highly flexible atomic services allow an easy extension of TBro's functionalities in the future. It also provides a fast prototyping platform to test and develop functionalities for genome-centred data warehouse systems such as Intermine and Tripal. Upcoming releases will introduce cart operations such as union or intersection as well as transformations (e.g. unigene ↔ isoform) to further ease TBro's usage. Finally, we aim to develop new features that enable users to switch between organisms or data releases in context of their personal carts again using Chado's built-in relationship model.

Availability

TBro is available as docker images (https://hub.docker.com/u/tbroteam) as well as source code (https://github.com/tbroteam). It is easily set-up using preconfigured docker images. Core applications, databases and job handlers are distributed in separate images. Functional tests are continuously performed with Travis-CI (https://travis-ci.org/TBroTeam/TBro) while code review is automatically performed by codeclimate (https://codeclimate.com/github/TBroTeam/TBro). A tutorial leads user through the installation as well as analysis process (https://tbro-tutorial.readthedocs.org). TBro is distributed under the MIT license. All included modules have compatible licenses (see Supplementary Table S1). The CanvasXpress (http://canvasxpress.org) release distributed with TBro is an earlier version available under the LGPL. Nevertheless, its version easily updated during the setup procedure.

Funding

FB and MJA were supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. This publication was supported by the Open Access Publication Fund of the University of Würzburg. We would like to thank Jörg Schultz for fruitful discussions and valuable suggestions throughout the project.

Conflict of interest. None declared.

References

1. Mortazavi,A., Williams,B.A., McCue,K. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5, 621–628.
2. Wang,Z., Gerstein,M., and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 57–63.
3. Trapnell,C., Williams,B.A., Pertea,G. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28, 511–515.
4. Garg,R., Patel,R.K., Tyagi,A.K. *et al.* (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.*, 18, 53–63.
5. Wenping,H., Yuan,Z., Jie,S. *et al.* (2011) De novo transcriptome sequencing in Salvia miltiorrhiza to identify genes involved in the biosynthesis of active ingredients. *Genomics*, 98, 272–279.
6. Xia,Z., Xu,H., Zhai,J. *et al.* (2011) RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*. *Plant Mol. Biol.*, 77, 299–308.
7. Wang,X.W., Luan,J.B., Li,J.M. *et al.* (2010) De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, 11, 400.
8. Schulz,M.H., Zerbino,D.R., Vingron,M. *et al.* (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28, 1086–1092.
9. Grabherr,M.G., Haas,B.J., Yassour,M. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29, 644–652.
10. Haas,B.J., Papanicolaou,A., Yassour,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, 8, 1494–1512.
11. Jones,P., Binns,D., Chang,H.Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30, 1236–1240.
12. Lohse,M., Nagel,A., Herter,T. *et al.* (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. *Plant Cell Environ.*, 37, 1250–1258.
13. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.
14. Patro,R., Duggal,G. and Kingsford,C. (2015) Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. *bioRxiv*.
15. Alexa,A. and Rahnenfuhrer,J. (2010) R package version 2, topGO: enrichment analysis for gene ontology. http://www.bioconductor.org/packages/release/bioc/html/topGO.html.
16. Luo,W., Friedman,M.S., Shedden,K. *et al.* (2009) GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10, 161.
17. Sanderson,L.A., Ficklin,S.P., Cheng,C.H. *et al.* (2013) Tripal v1.1: a standards-based toolkit for construction of online genetic and genomic databases. *Database*, 2013, bat075.
18. Ficklin,S.P., Sanderson,L.A., Cheng,C.H. *et al.* (2011) Tripal: a construction toolkit for online genome databases. *Database*, 2011, bar044.
19. Smith,R.N., Aleksic,J., Butano,D. *et al.* (2012) InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, 28, 3163–3165.
20. Dietrich,S., Wiegand,S. and Liesegang,H. (2014) TraV: a genome context sensitive transcriptome browser. *PLoS One*, 9, e93677.
21. Nussbaumer,T., Kugler,K.G., Bader,K.C. *et al.* (2014) RNASeqExpressionBrowser—a web interface to browse and visualize high-throughput expression data. *Bioinformatics*, 30, 2519–2520.
22. Vincent,J., Dai,Z., Ravel,C. *et al.* (2013) dbWFA: a web-based database for functional annotation of *Triticum aestivum* transcripts. *Database*, 2013, bat014.
23. TrinotateWeb: Graphical Interface for Navigating Trinotate Annotations and Expression Analyses. https://trinotate.github.io/TrinotateWeb.html (29 March 2016, date last accessed).
24. Camacho,C., Coulouris,G., Avagyan,V. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, 10, 421.
25. Neuhaus,I. CanvasXpress http://canvasxpress.org (29 March 2016, date last accessed).
26. Martin Bland,J. and Altman,D. (1986) Originally published as Volume 1, Issue 8476 Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 327, 307–310.
27. Mungall,C.J., Emmert,D.B. and FlyBase Consortium. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23, i337–i346.
28. Fitzpatrick,B. (2004) Distributed caching with Memcached. *Linux J.*, 2004, 5.

2.2 Supplementary Material

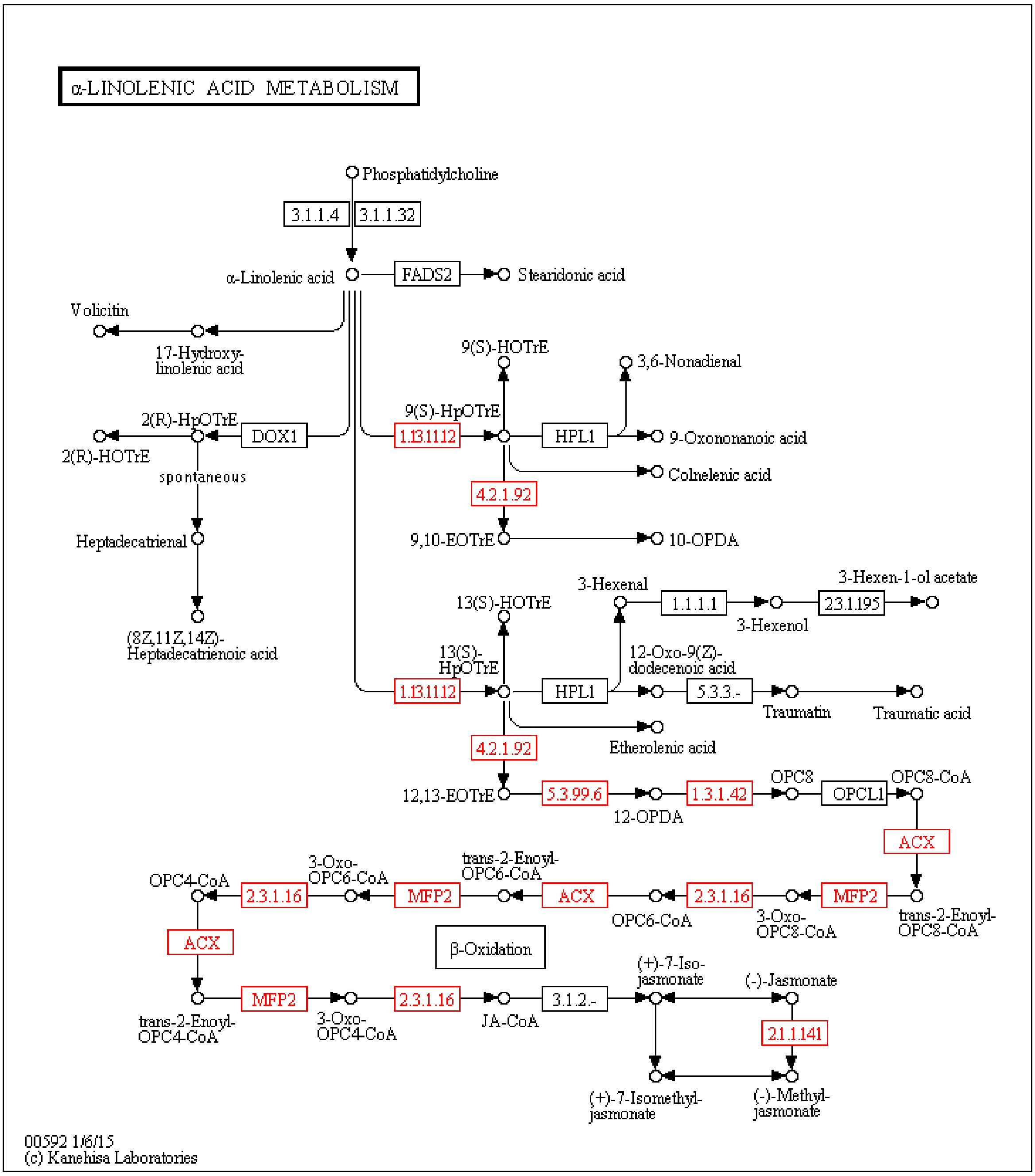


Fig. 2.1 Supplementary Figure 1 — KEGG map of the alpha-Linolenic acid metabolism with highlighted components present in a published cart (S1_JA_Pathway). Future releases will color the components dependent on their transcriptional regulation.

Table 2.3 Supplementary Table 1 — External libraries included in TBro.

Library	Version	Link	Licence	Note
smarty	3.1.13	http://www.smarty.net	LGPL	Server side templating
lightOpenID		http://code.google.com/p/lightopenid	MIT	User authentication
loggedPDO		http://github.com/phryneas/loggedPDO	MIT	Logged Database Connection
Foundation	"4.0.8 (js) & 4.1.6 (css)"	http://foundation.zurb.com/	MIT	Web Framework (css)
jQuery	1.9.1	http://jquery.com	MIT	DOM traversal, event binding, AJAX calls, etc.
jQueryUI	1.10.2	http://jqueryui.com	MIT	autocomplete, accordion, etc.
underscore.js	1.4.4	http://underscorejs.org	MIT	client side templating, helper functions
DateTables	1.9.4	http://www.datatables.net	BSD 3-clause	tables
TableTools	1.0.4	http://datatables.net/extras/tabletools	BSD 3-clause	tables
canvasXpress	7.1	http://canvasxpress.org/	LGPL	plots
sprintf.js		http://github.com/alexei/sprintf.js	BSD 3-clause	
webStorage		https://github.com/ryanttb/webStorage	MIT	local storage (synchronization)
alphanum.js		http://www.davekoelle.com/alphanum.html	LGPL	sorting
PEAR		http://pear.php.net/package	MIT / BSD 2-clause	Log, Console_CommandLine, Console_Table, Console_ProgressBar
Propel		http://propelorm.org	MIT	db abstraction layer
Phing		http://www.phing.info	LGPL	build

Chapter 3

Tardigrade Genomics

3.1 Abstract

Tardigrades are among the most stress tolerant animals and survived even unassisted exposure to space in low earth orbit. Still, the adaptations leading to these unusual physiological features are a mystery. Even the phylogenetic position of this phylum within the Ecdysozoa is enigmatic. Complete or draft genome sequence might help to address these questions as genomic adaptations can be revealed and phylogenetic reconstructions can be based on new markers. The following chapter presents the genome of a new Eutardigrada member, namely *M. tardigradum* and integrates it into a comparative framework with two previously published Eutardigrada species, namely *R. varieornatus* and *H. dujardini*. Using the comparative framework, the phylum of the tardigrades was placed as sister group of the nematodes and the arthropods as outgroup. A phase of massive gene loss thus far attributed to the nematodes could be pre-dated to the split from the tardigrades. A comprehensive catalog of protein domain expansions and contractions further revealed that the coding complements of the three tardigrades might have been shaped in much more species-specific manner than previously thought. Taken together, the comparative framework established in this chapter provides novel directions for further research on stress tolerance in tardigrades and has as a direct impact for the understanding of the Ecdysozoa evolution including prominent model organisms.

3.2 Introduction

There is no life without water. Antony van Leeuwenhoek must have been well aware of this fact when in 1702 he collected some dry dust from a roof gutter. He was up to a surprise when he viewed the sample with one of his self-built microscopes. Soon after mixing with some clean water, he found tiny animals, which he called ‘animalcules’ [183]. Thus, seemingly dead animals came fully alive again after rehydration. In 1959, D. Keilin coined the term ‘cryptobiosis’, which can be triggered by low oxygen (anoxybiosis), low temperature (cryobiosis), high salt concentrations (osmobiosis) or desiccation (anhydrobiosis) [163]. To date, this ‘peculiar state of biological organization’ between dead and alive [60] has been described in various species all over the tree of life. This includes prokaryotes [29], plants [17] with the highlight example of the ‘resurrection plants’ [208, 21, 200] and a wide range of invertebrate animals. The latter include species of arthropods [63, 61], nematodes [251], rotifers [281] and tardigrades [206]. Tardigrades (from latin tardus = slow and gradi = walk [261]) are small animals of about 0.1 to 1.2 mm with a peculiar shape reminiscent of bears. Accordingly, they have also been called ‘kleine Wasserbärchen’ (little water bear) in German. They were first identified at the end of the 18th century [33]. Today, more than 1,000 species are known [85]. Together, they form a phylum of their own belonging to the ecdysozoa [7]. The detailed phylogenetic position of this phylum is still under discussion. There are mainly two hypotheses, which place the tardigrades as sister taxon to the nematodes [125, 181, 199, 229, 237, 196, 34] or the arthropods [196, 50, 106, 238, 273] respectively, but so far neither molecular nor morphological investigations have come to an unambiguous conclusion [78, 80].

As their German name already suggests, tardigrades are an aquatic life form and can only survive as long as they are covered by a water film. Still, most species inhabit terrestrial habitats like mosses and lichens which regularly fall completely dry. At these times, adults, juveniles and also embryos can only survive until the next rain period by changing from the active state into the anhydrobiotic tun state [244]. As metabolic conversion of nutrients requires water, tardigrades in the tun state suspend life and do not age [127]. In this form, they survived being frozen [129, 130], heated [131] and exposed to enormous levels of UV [10] or ionizing radiation [151].

Although evolved as a mechanism to survive anhydrobiosis, tardigrades in the tun state are also resistant against other environmental stressors not typical for their habitat. The tardigrade species *M. tardigradum* even survived the exposure to space in low earth orbit [152]. Accordingly, tardigrades have been suggested as model organisms for space research [150].

The molecular mechanisms enabling anhydrobiosis are a field of active research. Starting with a search for a single causative agent, nowadays four lines of defense, each consisting of different options, are considered: (i) stabilization of proteins and membranes, (ii) avoiding damage caused by reactive oxygen species (ROS) and other toxins, (iii) restructuring of cellular components to evade stress on structures caused by drying (iv) and detection of desiccation and coordination of reactions by regulatory mechanisms and signaling pathways [64].

More than 40 years ago, water replacement and vitrification were suggested as core mechanisms for stabilization [65]. Here, water is replaced by other biomolecules resulting in a glass state of the cell. Mainly two types of molecules enabling this transition, sugars [67, 66] and late embryo abundant (LEA) proteins [109, 282] have been described. But, even within closely related organisms the relevance of these molecules differs [180, 128]. As further candidates, heat shock proteins have been suggested. As they assist in protein folding and are able to refold denatured proteins, they could provide a self-evident mechanism to repair damage arising in anhydrobiosis. Still, the relevance of heat shock proteins for tardigrades is discussed controversially. HSP70 expression is increased at rehydration [153] but not increased in desiccated animals [153, 232]. Directly comparing different variants of HSPs revealed complex patterns [245, 230].

The emergence of ROS is of considerable danger for a cell, as it can damage all cellular components. Already a challenge for a ‘standard’ cell, this problem increases dramatically when a cell desiccates. Accordingly, genes involved in the reduction of ROS are upregulated at the entrance of anhydrobiosis [228]. Still, ROS can damage cellular components in anhydrobiotic animals. In the case of Deoxyribonucleic acid (DNA), these can be recovered by effective repair systems [212]. Evidence for restructuring of the cytoskeleton in anhydrobiosis was found in the nematode *Panagrolaimus superbus* [283]. Similarly, the expression of some cytoskeletal proteins changed in anhydrobiosis of the tardigrade *M. tardigradum* [293]. The signaling pathways and regulatory mechanisms involved in anhydrobiosis are thus far only poorly understood. Only for the *Caenorhabditis elegans* Dauer larvae, notch signalling in the head neurons was suggested [86].

The first studies addressing the unique physiological peculiarities of tardigrades revealed different hypothesis regarding their underlying genomic basis. A genome wide analysis of the gene coding complement of the tardigrade *H. dujardini* by Boothby et al. [274]. found that horizontal gene transfer might have shaped the functional capacity of the animal much more than previously suspected. The analysis identified several thousand genes likely to be derived from non-metazoan sources mostly from bacteria. Especially proteins containing domains involved in classic stress response, including heat shock proteins, chaperones, DNA damage repair enzymes and antioxidant pathway members were expanded in numbers compared to *C. elegans* and *D. melanogaster*. Furthermore, the study suggested that some of the expansions build completely on foreign sources that replaced the original host genes. Based on their results, authors argued that stress tolerant organisms might show a predisposition to acquiring foreign genes presumably driven by membrane leakiness and DNA breakages during harsh conditions like desiccation, ultimately correlating horizontal gene transfer with the rate of survival during desiccation. A second independent genome study of *H. dujardini* reported strong conflicts between the two assembled and annotated genomes although the biomaterial for both studies was taken from the same original stock culture [174]. Analysis of the second genome reference for *H. dujardini* suggested a very low level of horizontal gene transfer. Authors demonstrated that the high rate of horizontal gene transfer was rather an artifact of non-eliminated contaminants than biological signal. They further prove that most candidates for foreign horizontal gene transfer could neither been confirmed using long read or short read sequencing data nor by assessing their expression status. They conclude that the previous genome assembly and accompanied analysis are heavily compromised by the almost ten thousand genes derived from bacterial contaminants and that most conclusion are rather artifactual. The conclusions of the second genome release were further supported by to additional independent studies reusing both previously generated genome sequencing data sets [73, 27]. Both studies showed that the data published by Boothby et al. can be used to assemble full bacterial genomes. Surprisingly, one of the genomes could be assigned to the bacterial family of the Chitinophagaceae, a family that is known to harbor genes coding for chitin degradation and utilization. Furthermore, a detailed analysis of the genome showed that it harbors genes associated with biosynthesis of proteorhodopsin, host invasion and intracellular resistance, dormancy, sporulation and oxidative stress. Delmont et al. further concluded that the detected genome might belong to a microbial inhabitant of *H. dujardini* since both, the data set published by Boothby et. al and Koutsovoulos et. al showed traces of it but could not rule out the possibility that it may be associated with the food source.

Studies in *H. dujardini* were complemented by a similar analysis in a second tardigrade species, namely *R. varieornatus* [121]. The authors leveraged their high-quality genome sequence of *R. varieornatus* and were able to show that only a small proportion of the gene coding complement might represent putative foreign genes. Their study further showed that the species (selectively) lost several members of pathways that promote stress damage (e.g., peroxisomal oxidative pathway, stress responsive pathway) during hypoxia, genotoxic or oxidative stress but simultaneously display expansion of gene families related to ameliorating damage (e.g., superoxide dismutases (SOD)). A close examination of gene expression profiles during dehydration and rehydration by the authors revealed only minor differences between the two states. Additionally, the study identified a tardigrade-unique DNA-associated protein that when transferred to human cell culture suppresses DNA damage and shows high irradiation viability. In summary, the study suggested that a) tardigrades can enter a dehydrated state without a massive transcriptional turnover and b) that the genome provides mechanisms that prevents, extenuates or protects against damage caused by harsh environmental conditions.

All molecular studies so far revealed a first but very different glimpse into the mechanisms underlying anhydrobiosis and other peculiarities specific to taridgrades. Even 300 years after van Leeuwenhoek's discovery, there is no general understanding of the genetic basis that encodes the strong stress tolerance of species from the phylum Tardigrada. The following chapter adds a third tardigrade species, namely *M. tardigradum* to the realm. *M. tardigradum* does not only represent a second class of tardigrades but also is arguably one of the most stress resistant tardigrades [152] with a wealth set of transcriptomic [293, 98, 195], proteomic [246, 247] and metabolomic [24] studies reusable for hypothesis testing.

3.2.1 Project objectives

Comparative Framework The first objective of this chapter is to establish a proper comparative genomics framework to test several hypotheses related to the unusual physiological features of tardigrades. A quality assessment of the different published *H. dujardini* genomes should pinpoint the most reliable one. A completeness screen of all available metazoan genomes from Ensembl and their annotation should select the most trustworthy ones. The *M. tardigradum* genome should be assembled, annotated and validated. The resulting set of genomes and their respective proteomes should be tested for ortholog relationships for later downstream analysis. Finally, the set of selected species should be tested for expanded, contracted as well as lost protein domains (see objective 3.2.1) and cross compared to existing expansion studies in tardigrades.

Phylogenetic Reconstruction Since the position of the phylum Tardigrada is is still under discussion an attempt should be made to reconstruct a whole genome based phylogeny in-cooperating a set of high quality metazoan genomes from Ensembl as well as the three available tardigrade genomes. Additionally, the two main hypotheses which place the tardigrades as sister taxon to the nematodes or the arthropods should be tested whether or not the whole genome based phylogeny results in a clear placement. The tests should include domain absence/presence patterns, domain architecture patterns and ortholog group memberships.

Gain and Loss Patterns Since previous studies reported a substantial loss in domains and domain architectures from the Ur-ecdysozoan to current day nematodes [308] it should be tested whether or not tardigrades undergo the same process independently of the outcome of the phylogeny reconstruction. The test should involve the estimation of gain and loss rates alongside the different hypothesis tested in the previous objectives. Gains as well as losses should be summarized using a gene ontology enrichment. Additionally, it should be discussed whether or not these patterns can be linked to the lifestyle of tardigrades or *M. tardigradum* in specific.

3.3 Material and Methods

The following subsections concentrate on the generation of the *M. tardigradum* genome, its annotation and the comparative analysis with other metazoan genomes . Methods for the assessment of different *H. dujardini* genomes can be found [27, 73]. Animal culturing was carried out by Laura Burleigh, F. Hoffmann-La Roche AG, Konzern-Hauptsitz, Grenzacherstrasse 124, 4070 Basel, Switzerland and Frank Förster, Department for Bioinformatics, Genomics Group, Biocentre / Am Hubland, 97074 Würzburg, Germany. 454 sequencing was carried out by F. Hoffmann-La Roche AG. Illumina sequencing was carried out by GATC Biotech AG (Hauptsitz), European Genome and Diagnostics Centre, Jakob-Stadler-Platz 7, 78467 Konstanz. Genome size estimation was carried out by the author of the thesis under supervision by Christian Janzen, Department of Cell and Developmental Biology, Am Hubland, 97074 Würzburg, Germany.

3.3.1 Animal culture

Tardigrade specimens of *M. tardigradum* (described in [75]; Eutardigrada, Apochela), cultured in the laboratory for a decade, were used to study the genome. Originally they were collected from dry moss in Tübingen, Germany. Animals were provided by Dr. Ralph Oliver Schill. The carnivorous tardigrade species was reared in plastic culture dishes on a small layer of 3% agar, covered with Volvic™ water (Danone Waters Deutschland, Wiesbaden, Germany). Rotifers of the species Philodina citrina were provided as food twice a week (Roche culture). The cultures were maintained in an environmental chamber at 20 °C using an artificial light source with a 12 h light, 12 h dark cycle. For the DNA/RNA extraction exuvia with eggs and embryos were collected and cleaned by five washing steps with Volvic™ water. Subsequently, they were placed separately in a 24-well plate until they hatched. Juveniles have been transferred in a reaction tube, frozen in liquid nitrogen and stored at -80°C.

3.3.2 Sequencing

DNA was extracted from approximately 1000 freshly hatched animals (to avoid bacterial contamination) using the Qiagen DNeasy kit according to the manufacturer's instructions for animal tissues (spin column protocol). Animals were washed five times in RNAse/DNAse-free water, resuspended in Qiagen buffer ATL and disrupted with a FastPrep-24 homogenizer (MP Biomedicals for 2 x 30 seconds at 4 m/sec). Following overnight incubation in buffer ATL and proteinase K at 56ºC, samples were treated with RNAse A and purified on a spin column. RNA was extracted from a similar sized animal culture as for the DNA. Animals were disrupted as described above, resuspended in buffer RLT, and RNA was extracted using the Qiagen RNeasy kit. Ribosomal RNA was depleted using the RiboMinus Kit for RNA-seq (Invitrogen) and reverse transcribed using random hexamers (Promega Im-Prom II Reverse Transcription System). cDNA was amplified using the GenomePlex Complete Whole Genome Amplification Kit (Sigma). The 95ºC fragmentation step was omitted from the whole genome amplification, as RNA had been fragmented during homogenization. Bead libraries were prepared from DNA (1.8 ug) and cDNA (2 ug) using the GS FLX Titanium general library preparation kit (454 Life Sciences), followed by amplification using emulsion PCR with the LV emPCR kit (Lib-L) (454 life Sciences). Sequencing was performed on a 454 FLX instrument (454 Life Sciences). A second sequencing data set was produced from an additional batch of animals. DNA was extracted as above and subjected to whole genome amplification with Qiagen REPLI-g prior to sequencing. TruSeq DNA library prep and Illumina sequencing was carried out by GATC.

3.3.3 Genome size estimation

The genome size of *M. tardigradum* was estimated using flow cell cytometry. *Drosophila melanogaster* was used as standard[112]. A culture of M. tardigradum was washed (4 times, M9 buffer) and placed into modified Galbraith's buffer. Nuclei were released with a tissue grinder (Kontes Dounce tissue grinder, "A" pestle) and filtered to a 30-µm Nylon mesh. The same procedure was carried out with a single head from *Drosophila melanogaster* female. The nuclear suspension was stained with propidium iodid (PI) for 2 hours and measured immediately with a FACScalibur flow cytometer (Becton Dickinson, USA) and analyzed with CellQuest Pro version 6.0. PI-positive cells were gated and fluorescence intensity was analyzed in FL2-H channel and displayed on a linear scale. Non-stained cells served as a negative control. The whole procedure was benchmarked by comparing *Drosophila melanogaster* and *Apis mellifera* [15]. Results indicated an error of about 0.025 (data not shown).

3.3.4 Genome and transcriptome assembly

Genomic and transcriptomic reads were prepared by masking vector contamination's and adapters using SMALT [223]. Both read read sets were compared against NCBI-nr using diamond [42]. The resulting alignments were prepared for MEGAN using daa-meganizer [140, 139]. MEGAN was used to compute the lowest common ancestor for each read individually. Reads assigned to the superkingdom Bacteria or Archaea were removed from the data set. Remaining genomic reads were assembled with Canu (release 1.3; error-

Rate=0.035, genomeSize=75000000, minReadLength=50, corMinCoverage=0, corMaxEvidenceErate=0.15, minOverlapLength=50, trimReadsCoverage=2) [173]. Transcriptomic reads were assembled using MIRA4 with the accurate settings [57]. The resulting EST library was further used to assess the genome assembly completeness. The genome completeness was validate with CEGMA [217] and BUSCO [256].

3.3.5 Genome feature annotation

Known repetitive elements were annotated with RepeatMasker (v.4.0.4, species=metazoa) [258]. Coding genes were annotated with Braker1 (version 1.9; default parameters) by combining de novo gene predictions and evidence alignments from ESTs [134]. Evidence alignments were generated by aligning all ESTs against the genome using BLAT [165]. Resulting alignments were converted into intron boundaries and passed to Braker1. The resulting proteins were functionally classified using homology and profile based methods. Protein families, domains and important sites were assigned using InterproScan5 (release 5.20; default parameters) [149] and the Interpro database (release 59.0) [92]. Gene ontology terms and basic functional descriptions were assigned by lifting protein domain gene ontology annotations to their respective gene/protein.

3.3.6 Protein Domain Expansions and Contractions

Significantly expanded and contracted Interpro terms (restricted to those predicted by the Pfam sub module) were identified by comparing their occurrence in the three tardigrades to all species present in the Ensembl Metazoa database (Release 34) using a chi square test [166]. The occurrence of a specific Interpro term in the three tardigrades was compared to the occurrence of the same term in each of the reference species individually. The number of all genes associated with at least one Interpro term was used as background for each species. The resulting p-values for each Interpro term were combined into a weighted consensus p-value since they addressed the same null hypothesis, that an Interpro term is not expanded or contracted significantly. For that, all p-values were z-transformed and a weighted consensus test was applied. The final weighted consensus p-value was adjusted using the Bonferroni method and considered significant at a level of 5%. Expansion and contractions were used to test for enriched gene ontology terms with dcGOR [90]. Enrichments were statistically verified with the hypergeometric test. P-values were adjusted using Bonferroni's method. The significance of a term was not only required when using the whole domains as background but also using domains annotated to all its direct parents/ancestors as background (Parent-Child algorithm). All Interpro terms (restricted to those that were Pfam-derived) found in the 56 species were used as background.

3.3.7 Phylogenetic Reconstruction and Hypothesis Testing

The phylogenetic reconstruction was carried out using 56 selected species present in release 34 of the Ensembl Metazoa database [166]. Potential inparalogs, orthologs and co-ortholog pairs were identified using orthAgauge [82]. The species phylogeny was reconstructed using ete3 (build 3.0.0b36; -w mcoffee_ensembl-trimal01-prottest_default-treebest_ensembl -m cog_25-alg_concat_default-raxml_default) based on the ortholog groups generated with MCL [285]. Alternative phylogenetic hypothesis for the placement of the tardigrades were tested using RAxML [263] and CONSEL [254]. Testing was done using binary representations of the absence-presence matrices for protein domain families, domain architectures and orthologous groups. Domain architectures were defined on annotated Pfam domains families. Repetitive stretches of domains were collapsed into a single representation. RAxML was used to calculated per-site log likelihoods for each of the alternative hypothesis. Test statistics for alternative hypothesis were calculated using the approximately unbiased test implemented in CONSEL.

3.3.8 Protein Domain Gain and Loss Estimation

Gain and loss events for protein domains were detected using the most likely tree topology and the corresponding absence-presence matrices. Ancestral nodes were reconstructed using RAxML [263]. To account for uncertainties during the reconstruction an expected value for each gain and loss event was calculated by multiplying the probability of the parent and the child state. Gain states were further assessed by cross comparison to the out-groups and marked ambiguous if one of the species within this group already encoded the tested protein domain. Subsequently, ambiguous gains were excluded from all downstream analysis.

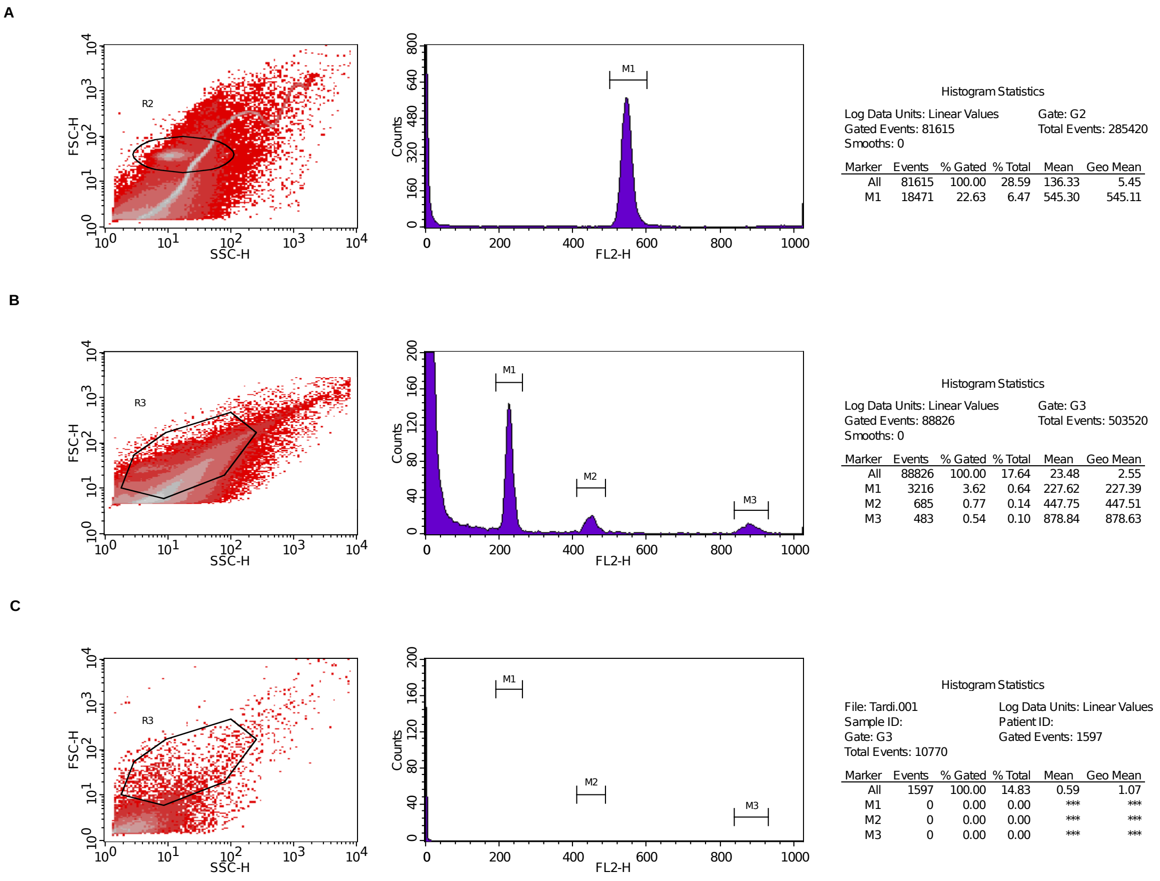
3.3.9 Functional Hypothesis Testing

SAHS/CAHS/MAHS containing proteins previously identified in *R. varieornatus* were detected using a profile based approach. Template from *R. varieornatus* were aligned, the alignment manually curated and used to build a hidden Markov model (HMM) [79]. A reverse search of the model against *R. varieornatus* proteins was conducted and the results used to define an optimal inclusion e-value (CAHS = 1.2×10^{-22}; SAHS = 5.1×10^{-40}). The final model was used to screen proteins from all species. Dsup and MAHS homologs were identified by a simple protein blast (BLASTP) [48] against the complete set of Ensembl Metazoa Release 34 species and the three tardigrades. HSPs were identified by using predefined Pfam protein domains (PF00011, PF00012, PF00118, PF00166, PF00183, PF00226). Trehalose biosynthesis and metablosim components were identified by using predefined Pfam protein domains (PF00128, PF00358, PF00534, PF00982, PF01204, PF02056, PF02358, PF02922, PF03632, PF03633, PF03636, PF09071, PF11941, PF11975, PF16657). Late Embryogenesis Abundant proteins (LEA proteins) were identified using predefined Pfam protein domains (PF00477, PF02987, PF03168, PF03242, PF03760, PF10714). Gene expression values for *R. varieornatus* were directly taken from the supplementary material published by Hashimoto et al. [121].

3.4 Results

3.4.1 Genomic Features of *M. tardigradum*

The assembled genome of *M. tardigradum* comprised 75.1 Mb, which is in close agreement to the results of a flow cytometry based determination of 73.3±1.8 Mb (see figure 3.1). Overall, 6654 contigs were assembled with a contig N50 size of 50 kb. The assembly was validated with two approaches. First, a prediction of 248 core eukaryotic genes with CEGMA revealed a 96 % completeness of the genome (see figure 3.3 A). Second, a prediction of near-universal single-copy orthologs with BUSCO was used to benchmark the *M. tardigradum* genome against three different lineages. Benchmarking against a nematode specific BUSCO set revealed a completeness of 41 % while benchmarks against an arthropods specific and a Metazoa specific set revealed a completeness of 35 % respectively 56 % (see figure 3.3 A and table 3.1 for details). Only 1.23 % of the assembly was classified as repetitive or low-complexity. Based on a metazoan repeat library, 1271 DNA transposons (mostly hobo-Activator and Tc1-IS630-Pogo) and 2033 LTR elements (mostly BEL/Pao and Gypsy/DIRS1) were identified. The integrative gene annotation approach predicted 19,401 protein coding genes. 1684 genes were putatively derived through tandem duplication while 43 genes probably originated through segmental duplication when compared to *H. dujardini* and *R. varieornatus*. The subsequent functional annotation found homologs for 12,518 genes (65 %) while 7534 had an ortholog within the reference species from Ensembl Metazoan Release 34, *R. varieornatus* or *H. dujardini*. Based on the curated gene set, 10,966 genes were functionally assigned to either a protein family, protein domain or an important site, excluding low-complexity, transmembrane and coiled-coil assignments. 7357 genes had at least one associated gene ontology term.



Histogram Statistics

Log Data Units: Linear Values   Gate: G2
Gated Events: 81615   Total Events: 285420
Smooths: 0

Marker	Events	% Gated	% Total	Mean	Geo Mean
All	81615	100.00	28.59	136.33	5.45
M1	18471	22.63	6.47	545.30	545.11

Histogram Statistics

Log Data Units: Linear Values   Gate: G3
Gated Events: 88826   Total Events: 503520
Smooths: 0

Marker	Events	% Gated	% Total	Mean	Geo Mean
All	88826	100.00	17.64	23.48	2.55
M1	3216	3.62	0.64	227.62	227.39
M2	685	0.77	0.14	447.75	447.51
M3	483	0.54	0.10	878.84	878.63

Histogram Statistics

File: Tardi.001   Log Data Units: Linear Values
Sample ID: Patient ID:
Gate: G3   Gated Events: 1597
Total Events: 10770

Marker	Events	% Gated	% Total	Mean	Geo Mean
All	1597	100.00	14.83	0.59	1.07
M1	0	0.00	0.00	***	***
M2	0	0.00	0.00	***	***
M3	0	0.00	0.00	***	***

Fig. 3.1 Genome size estimation for *M. tardigradum*. The histograms of relative DNA content were obtained after flow cytometric analysis of propidium iodide-stained nuclei. Sub-figures: A = Stained nuclei from whole *M. tardigradum* animals; B = Stained nuclei from a single *D. melanogaster* head; C = Unstained nuclei from whole *M. tardigradum* animals. Marker M1 corresponds to the diploid genomes size in all samples. The ratio of M1 peak means (*M. tardigradum* : *D. melanogaster*) was equal to 0,41 and hence the 2C DNA amount of *M. tardigradum* was estimated to about 0,75 pg corresponding to a genome size of 73.3±1.8 Mb (SD was calculated from a cross comparison of *D. melanogaster* and *A. mellifera*, data not shown).

3.4.2 Comparative Framework

***H. dujardini* and *R. varieornatus*:** The three publically available tardigrade genomes (two for *H. dujardini* and one for *R. varieornatus*) were benchmarked using CEGMA and BUSCO. The *R. varieornatus* and the *H. dujardini* genome published by Koutsovoulos et al. showed similar results compared to *M. tardigradum* (see table 3.1). Only the *H. dujardini* genome published by Boothby et al. showed a dramatic increase in total assembly size and a high duplication rate in both CEGMA and BUSCO. A detailed assessment of Boothby et al. genome showed that the assembly is heavily contaminated with non-host sequences [174, 73, 27]. Additionally, the assembly shows a great number of highly similar contigs, probably representing haplotypes of the same allele which where not resolved into a single haploid representation during the assembly process. The Boothby et al. assembly was not considered for any downstream analysis to minimize the influence of the assembly quality on the expansion, contraction and loss analysis.

Table 3.1 Tabular overview of key assembly metrics for *M. tardigradum*, *R. varieornatus* and two different *H. dujardini* assemblies. BUSCO and CEGMA results are given in percent. Duplicated genes can be part of partial and complete counts.

		M. tardigradum	*H. dujardini*		*R. varieornatus*
			[174]	[274]	[121]
Assembly	Sequences	6,654	22,497	13,202	199
	Total Length	75 Mb	252 Mb	135 Mb	56 Mb
	Longest	470 kb	1.5 Mb	594 kb	9 Mb
	Shortest	0.3 kb	2 kb	0.5 kb	1 kb
	N50	50 kb	15 kb	50 kb	4 Mb
	N90	4 kb	5 kb	6 kb	1 Mb
	Avg. GC	42 %	47 %	45 %	47 %
	N's	0	36 kb	3 Mb	0
BUSCO Arthropoda	Complete	35	40	43	48
	Partial	12	11	13	10
	Duplicated	2	35	3	3
	Missing	51	48	43	40
BUSCO Nematoda	Complete	41	38	51	47
	Partial	6	5	4	5
	Duplicated	2	20	2	2
	Missing	52	57	45	47
BUSCO Metazoa	Complete	56	64	65	70
	Partial	10	9	10	6
	Duplicated	3	42	3	4
	Missing	33	27	24	23
CEGMA	Complete	96	88	89	96
	Partial	98	97	95	97
	Duplicated	1	3	1	1
	Missing	2	3	5	3

Ensembl Metazoa genomes: 65 metazoan genomes (Ensembl Metazoan Release 34) were benchmarked with CEGMA and BUSCO (lineage Metazoa odb9) to assess their quality and remove incomplete genomes that could dramatically influence the expansion, contraction and loss analysis. Results from the CEGMA benchmark showed relatively stable and high completeness values. Only non-ecdysozoa genomes generally displayed low completeness. Results of the BUSCO benchmark varied strongly (see figure 3.3) and probably indicate BUSCO lineage specific effects already seen for the three tardigrade genomes (see table 3.1, differences between BUSCO Arthropoda, Nematoda and Metazoa). Due to BUSCOs lineage specific effects only CEGMA results were considered for the final species selection. The empirical distribution of CEGMA completeness values were used to compute descriptive parameters which where then visualized in a skewness-kurtosis plot (see figure 3.2 A). Completeness values fitted best to a theoretical exponential distribution and the summary statistics suggested a minimum completeness of 75% still fit to the theoretical distribution. An additional inspection of the empirical density and the cumulative distribution further supported the assumption that CEGMA completeness values lower then 75% can be considered outliers (see figure 3.2 B). Nine species with less than 75% completeness were removed (see figure 3.2 B right). The final Ensembl Metazoa Genomes selection contained 56 species.

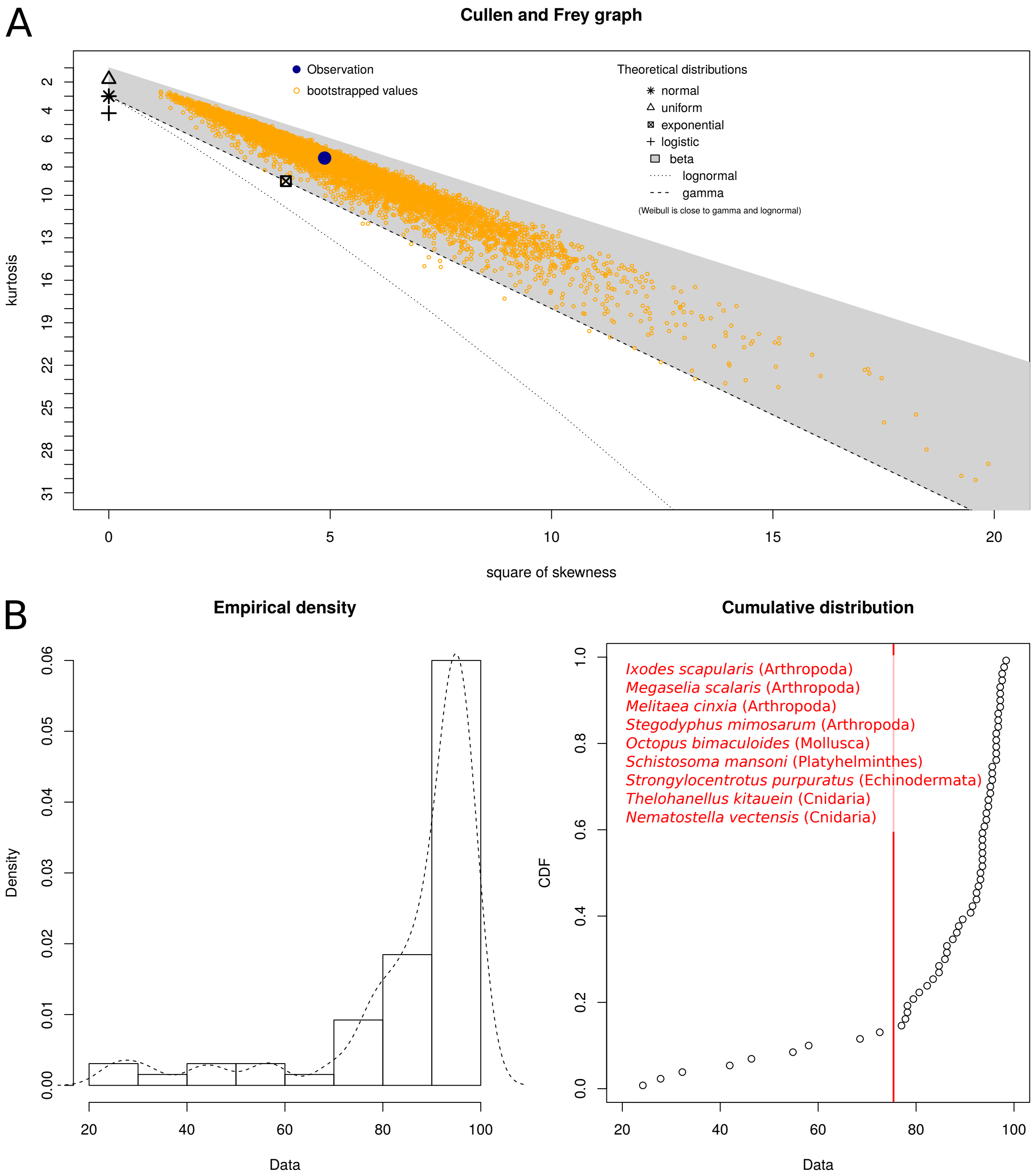


Fig. 3.2 Ensembl Metazoa genome selection based on visual inspection of their completeness values. A) Skewness-kurtosis plot that visualizes the empirical distribution of CEGMA completeness values and computed descriptive parameters. Completeness values fitted best to a theoretical exponential distribution and the summary statistics suggested a minimum completeness of 75% still fit to the theoretical distribution. B) Empirical density and the cumulative distribution of CEGMA completeness values (75% was chosen as the final threshold; species indicated in red were excluded based on the threshold). The analysis was carried using the fitdistrplus package [72].

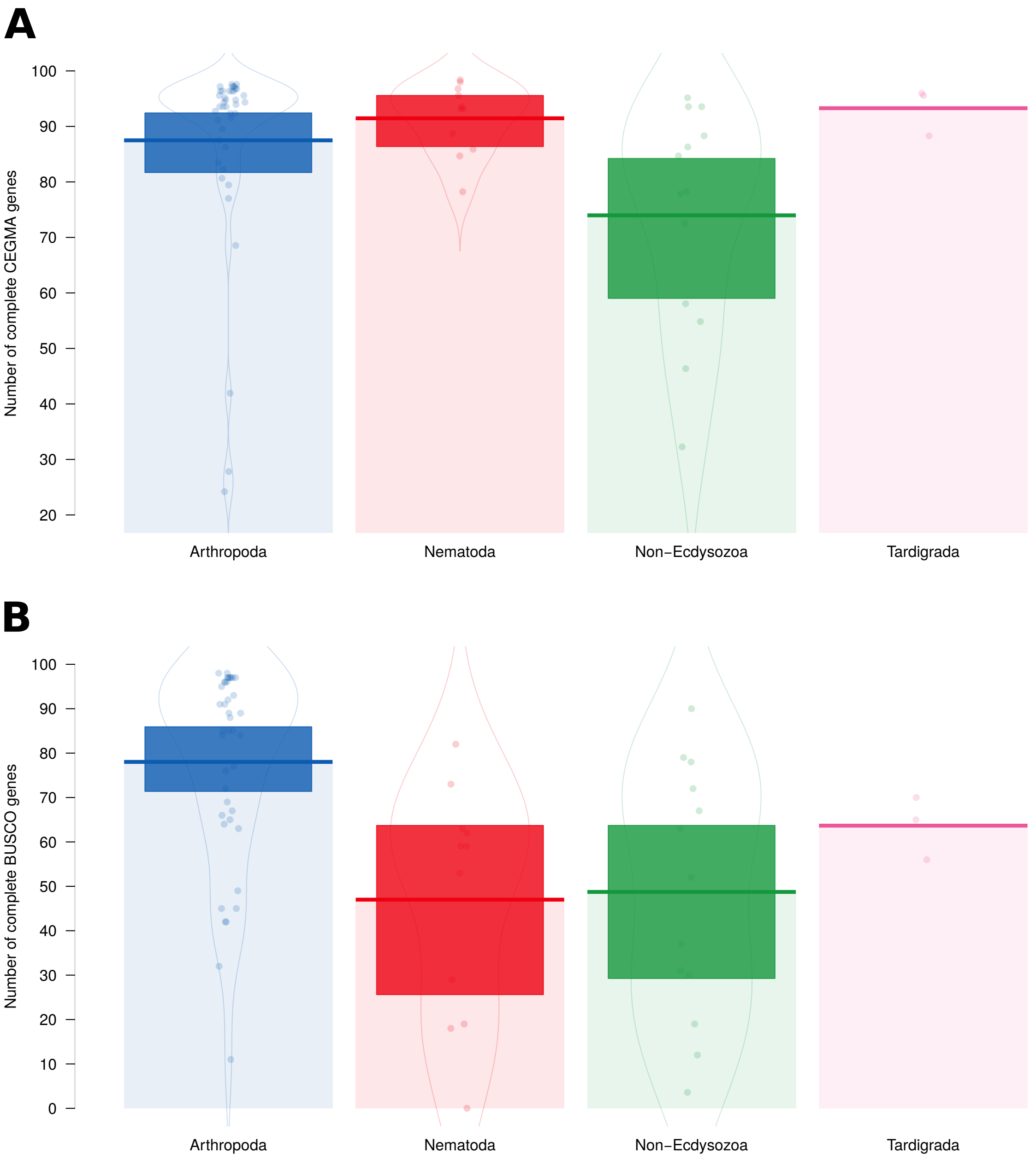


Fig. 3.3 Genome completeness values of BUSCO and CEGMA across all Ensemble Metazoa Release 34 species and the three tardigrades. CEGMA completeness values show less spread within and between phyla. BUSCO completeness (using a set of metazoa BUSCOs) shows lower values for Nematoda, Non-Ecdysozoa and Tardigrada.

3.4.3 Protein Domain Expansions and Contractions

Protein domains are evolutionary conserved units usually with independent structural and functional properties [145, 136, 43]. They are widely distributed over all existing organisms [187] with some of them being universal others being clade-specific. Only a limited set of protein domains was established during evolution but their combination into different protein architectures provides a huge functional capacity. Especially in eukaryotes the re-usage of protein domains can be high and up to 90% of all domains can be found in multiple proteins within a single specimen. Duplication and integration events are driven by whole genome duplication or genetic mechanisms like exon-shuffling, retrotranspositions, recombination and horizontal gene transfer. Duplicated domains are often found in new functional arrangements, they play an important role during sub-functionalization and they can positively influence the adaptability of an organisms [18, 167, 226]. Reversibly, the detection of expanded protein domains could provide hints on which biological processes and molecular functions are most likely associated with certain (adaptive) traits. The three proteomes from *M. tardigradum*, *R. varieornatus* and *H. dujardini* were subjected to an expansion and contraction search. Protein domains were classified into Class I and II expansions as well as Class I and II contractions. Class I expansions and contractions are those where the query species experienced the highest or lowest protein domain occurrence count whereas Class II expansions and contractions indicated protein domains where the query species belongs to the group with the 5% highest or lowest occurrence. While expansions are generally easy to detect and score, contractions are much harder to assess in terms of their significance. Especially protein domains with low counts and a very narrow species distribution are unsuitable for chi-square testing. Overall, 7939 protein domains where tested for contractions and expansions. 130 protein domains tested significant for an expansion while 12 tested significant for a contraction (see tables 3.2, 3.4, 3.6 and 3.8). *M. tardigradum* showed 27 Class I, 23 Class II expansions, 2 Class I contractions and 6 Class II contractions. *R. varieornatus* showed 38 Class I, 26 Class II expansions, 0 Class I contractions and 4 Class II contractions while *H. dujardini* showed 41 Class I, 35 Class II expansions, 2 Class I contractions and 2 Class II contractions. Expanded and contracted proteins domain sets were subjected to a gene ontology enrichment separately for each species. Class I and II for expansions and contractions were combined but none of the subsets showed enrichments for a specific term. A subsequent intersection analysis showed that only small amount of expansions and contractions are shared across the three tardigrade species (see figure 3.4). Surprisingly, all species showed a substantial amount of undetectable protein domains (*M. tardigradum*: 2666; *R. varieornatus*: 2576; *H. dujardini*: 2551).

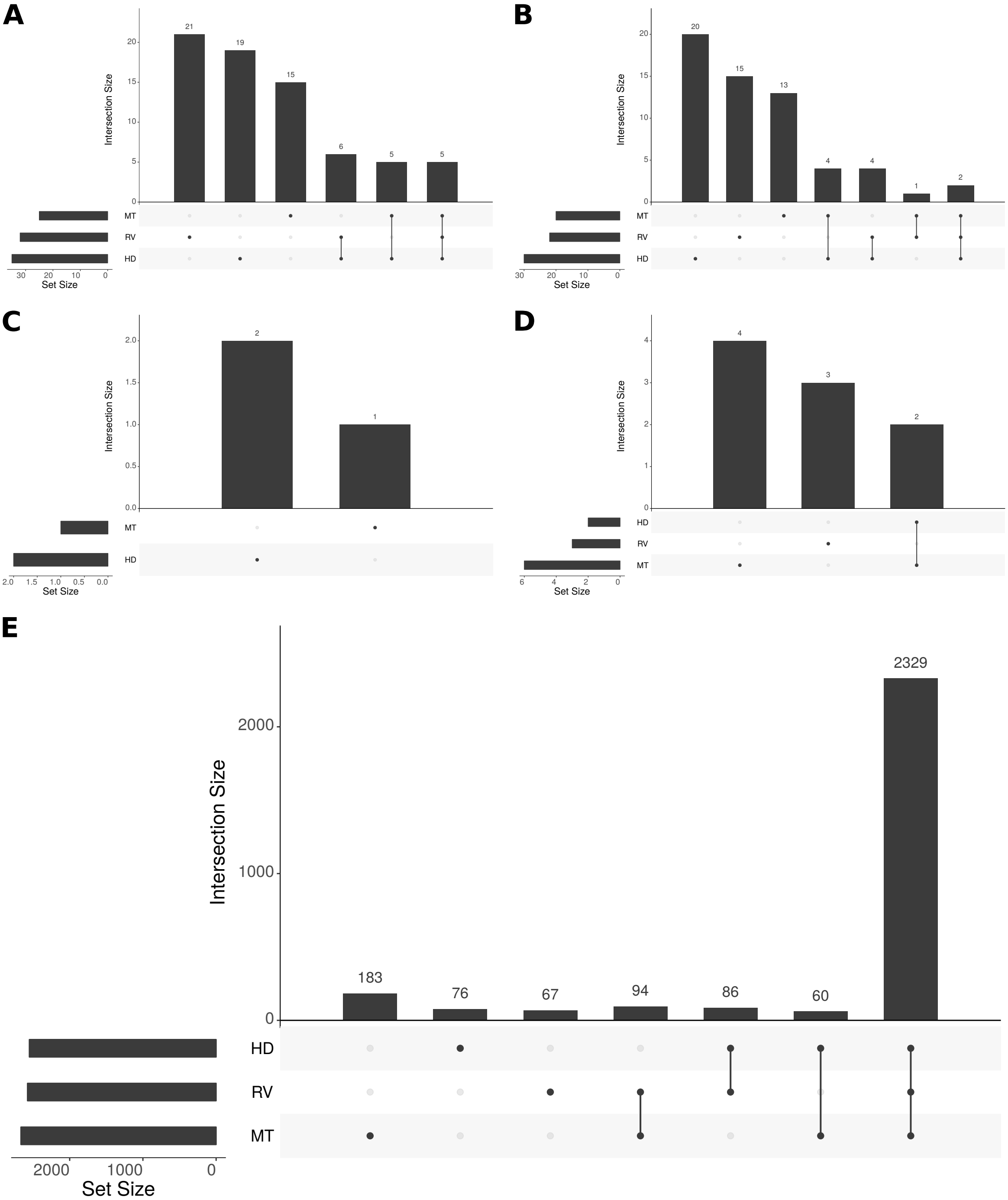


Fig. 3.4 A) Intersection analysis of expanded and contracted protein domains in *M. tardigradum*, *R. varieornatus* and *H. dujardini*. Expansions were generally species-specific and only a small number was shared across the three species. Lost protein domains were largely shared across the three species. Several of the expansions as well as lost protein domains were likely contamination in either the tardigrades themselves or the respective outgroups used during the detection. The analysis was carried using the UpSetR package [188]. A) Class I Expansions B) Class II expansions C) Class I contractions D) Class II contractions E) Undetectable protein domains ; Abbreviation: HD = *H. dujardini*, MT = *M. tardigradum*, RV = *R. varieornatus*

Table 3.2 Protein domain expansions Class I. Abbreviation: HD = *H. dujardini*, MT = *M. tardigradum*, RV = *R. varieornatus*. All expansions were tested significantly with an adjusted p-value smaller than 0.05.

Interpro ID	HD	MT	RV	Description
IPR032406	1	1	1	Cyclic nucleotide-gated channel
IPR018392	1	1	1	LysM domain
IPR015938	1	1	1	Glycine N-acyltransferase, N-terminal
IPR014851	1	1	1	BCS1, N-terminal
IPR001548	1	1	1	Peptidase M2, peptidyl-dipeptidase A
IPR026767	1	1	0	Transmembrane protein 151
IPR008197	1	1	0	WAP-type ’four-disulfide core’ domain
IPR001447	1	1	0	Arylamine N-acetyltransferase
IPR000726	1	1	0	Glycoside hydrolase, family 19, catalytic
IPR000627	1	1	0	Intradiol ring-cleavage dioxygenase, C-terminal
IPR027353	1	0	1	NET domain
IPR013871	1	0	1	Cysteine-rich secretory protein
IPR011547	1	0	1	SLC26A/SulP transporter domain
IPR007858	1	0	1	Dpy-30 motif
IPR002645	1	0	1	STAS domain
IPR001429	1	0	1	P2X purinoreceptor
IPR025667	1	0	0	SprB repeat
IPR025533	1	0	0	Protein of unknown function DUF4419
IPR024989	1	0	0	Major facilitator superfamily associated domain
IPR024370	1	0	0	PBP domain
IPR022234	1	0	0	Protein of unknown function DUF3759
IPR021255	1	0	0	Putative auto-transporter adhesin, head GIN domain
IPR018999	1	0	0	RNA helicase UPF1, UPF2-interacting domain
IPR015399	1	0	0	Domain of unknown function DUF1977, DnaJ-like
IPR009688	1	0	0	Domain of unknown function DUF1279
IPR009283	1	0	0	Apyrase
IPR005554	1	0	0	Nrap protein
IPR005378	1	0	0	Vacuolar protein sorting-associated protein 35
IPR004352	1	0	0	Glycoside-hydrolase family GH114, TIM-barrel domain
IPR003661	1	0	0	Signal transduction histidine kinase
IPR003594	1	0	0	Histidine kinase-like ATPase, C-terminal domain

Interpro ID	HD	MT	RV	Description
IPR001480	1	0	0	Bulb-type lectin domain
IPR000716	1	0	0	Thyroglobulin type-1
IPR000519	1	0	0	P-type trefoil domain
IPR000407	1	0	0	Nucleoside phosphatase GDA1/CD39
IPR025340	0	1	0	Protein of unknown function DUF4246
IPR023796	0	1	0	Serpin domain
IPR018629	0	1	0	XK-related protein
IPR018473	0	1	0	Hermes trasposase, DNA-binding domain
IPR013424	0	1	0	PEP-CTERM protein-sorting domain
IPR010513	0	1	0	KEN domain
IPR009492	0	1	0	TniQ
IPR008514	0	1	0	Type VI secretion system effector, Hcp
IPR007365	0	1	0	Transferrin receptor-like, dimerisation domain
IPR007016	0	1	0	O-antigen ligase-related
IPR006214	0	1	0	Bax inhibitor 1-related
IPR005105	0	1	0	Protein-PII uridylyltransferase, N-terminal
IPR004993	0	1	0	GH3 family
IPR004308	0	1	0	Glutamate-cysteine ligase catalytic subunit
IPR002557	0	1	0	Chitin binding domain
IPR032776	0	0	1	CECR6/TMEM121 family
IPR032405	0	0	1	Kinesin-associated
IPR025959	0	0	1	Winged helix-turn helix domain
IPR019774	0	0	1	Aromatic amino acid hydroxylase, C-terminal
IPR012938	0	0	1	Glucose/Sorbosone dehydrogenase
IPR012887	0	0	1	L-fucokinase
IPR012308	0	0	1	DNA ligase, ATP-dependent, N-terminal
IPR011607	0	0	1	Methylglyoxal synthase-like domain
IPR008893	0	0	1	WGR domain
IPR007644	0	0	1	RNA polymerase, beta subunit, protrusion
IPR007642	0	0	1	RNA polymerase Rpb2, domain 2
IPR007350	0	0	1	Transposase, Tc5, C-terminal
IPR007225	0	0	1	Exocyst complex component EXOC6/Sec15
IPR006710	0	0	1	Glycoside hydrolase, family 43
IPR006204	0	0	1	GHMP kinase N-terminal domain
IPR004854	0	0	1	Ubiquitin fusion degradation protein UFD1

Interpro ID	HD	MT	RV	Description
IPR004030	0	0	1	Nitric oxide synthase, N-terminal
IPR003929	0	0	1	Potassium channel, calcium-activated, BK, alpha subunit
IPR002759	0	0	1	Ribonuclease P/MRP protein subunit
IPR001736	0	0	1	Phospholipase D/Transphosphatidylase
IPR001424	0	0	1	Superoxide dismutase, copper/zinc binding domain

Table 3.4 Protein domain expansions Class II. Abbreviation: HD = *H. dujardini*, MT = *M. tardigradum*, RV = *R. varieornatus*. All expansions were tested significantly with an adjusted p-value smaller than 0.05.

Interpro ID	HD	MT	RV	Description
IPR018487	1	1	1	Hemopexin-like repeats
IPR006068	1	1	1	Cation-transporting P-type ATPase, C-terminal
IPR007743	1	1	0	Immunity-related GTPases-like
IPR001507	1	1	0	Zona pellucida domain
IPR001424	1	1	0	Superoxide dismutase, copper/zinc binding domain
IPR000276	1	1	0	G protein-coupled receptor, rhodopsin-like
IPR032751	1	0	1	Protein fuseless
IPR010796	1	0	1	B9 domain
IPR005018	1	0	1	DOMON domain
IPR004156	1	0	1	Organic anion transporter polypeptide OATP
IPR031569	1	0	0	Apextrin, C-terminal domain
IPR019545	1	0	0	DM13 domain
IPR019344	1	0	0	Mitochondrial F1-F0 ATP synthase subunit F, predicted
IPR013126	1	0	0	Heat shock protein 70 family
IPR012462	1	0	0	Peptidase C78, ubiquitin fold modifier-specific peptidase 1/ 2
IPR011735	1	0	0	HtrL protein
IPR011234	1	0	0	Fumarylacetoacetase, C-terminal-related
IPR009112	1	0	0	GTP cyclohydrolase I, feedback regulatory protein
IPR009009	1	0	0	RlpA-like protein, double-psi beta-barrel domain
IPR007305	1	0	0	Vesicle transport protein, Got1/SFT2-like
IPR007053	1	0	0	LRAT-like domain
IPR004993	1	0	0	GH3 family
IPR004878	1	0	0	Otopetrin
IPR004871	1	0	0	Cleavage/polyadenylation specificity factor

Interpro ID	HD	MT	RV	Description
IPR004181	1	0	0	Zinc finger, MIZ-type
IPR004014	1	0	0	Cation-transporting P-type ATPase, N-terminal
IPR003929	1	0	0	Potassium channel, calcium-activated, BK, alpha subunit
IPR003719	1	0	0	Phenazine biosynthesis PhzF protein
IPR001279	1	0	0	Metallo-beta-lactamase
IPR000375	1	0	0	Dynamin central domain
IPR007645	0	1	1	RNA polymerase Rpb2, domain 3
IPR026854	0	1	0	Vacuolar protein sorting-associated protein 13
IPR025799	0	1	0	Protein arginine N-methyltransferase
IPR024989	0	1	0	Major facilitator superfamily associated domain
IPR018164	0	1	0	Alanyl-tRNA synthetase, class IIc, N-terminal
IPR007120	0	1	0	DNA-directed RNA polymerase, subunit 2, domain 6
IPR007080	0	1	0	RNA polymerase Rpb1, domain 1
IPR005515	0	1	0	Vitelline membrane outer layer protein I (VOMI)
IPR002645	0	1	0	STAS domain
IPR001736	0	1	0	Phospholipase D/Transphosphatidylase
IPR001372	0	1	0	Dynein light chain, type 1/2
IPR001098	0	1	0	DNA-directed DNA polymerase, family A, palm domain
IPR000814	0	1	0	TATA-box binding protein
IPR000547	0	1	0	Clathrin, heavy chain/VPS, 7-fold repeat
IPR024970	0	0	1	Maelstrom domain
IPR022140	0	0	1	Kinesin-like KIF1-type
IPR011645	0	0	1	Haem NO binding associated
IPR008250	0	0	1	P-type ATPase, A domain
IPR007325	0	0	1	Kynurenine formamidase
IPR007281	0	0	1	Mre11, DNA-binding
IPR007021	0	0	1	Domain of unknown function DUF659
IPR004875	0	0	1	DDE superfamily endonuclease domain
IPR002772	0	0	1	Glycoside hydrolase family 3 C-terminal domain
IPR001932	0	0	1	PPM-type phosphatase domain
IPR001568	0	0	1	Ribonuclease T2-like
IPR001180	0	0	1	Citron homology (CNH) domain
IPR000731	0	0	1	Sterol-sensing domain
IPR000648	0	0	1	Oxysterol-binding protein
IPR000407	0	0	1	Nucleoside phosphatase GDA1/CD39

Interpro ID	HD	MT	RV	Description

Table 3.6 Protein domain contractions Class I. Abbreviation: HD = *H. dujardini*, MT = *M. tardigradum*, RV = *R. varieornatus*. All contractions were tested significantly with an adjusted p-value smaller than 0.05.

Interpro ID	HD	MT	RV	Description
IPR006671	1	0	0	Cyclin, N-terminal
IPR000195	1	0	0	Rab-GTPase-TBC domain
IPR001752	0	1	0	Kinesin motor domain

Table 3.8 Protein domain contractions Class II. Abbreviation: HD = *H. dujardini*, MT = *M. tardigradum*, RV = *R. varieornatus*. All contractions were tested significantly with an adjusted p-value smaller than 0.05.

Interpro ID	HD	MT	RV	Description
IPR001251	1	1	0	CRAL-TRIO lipid binding domain
IPR000408	1	1	0	Regulator of chromosome condensation, RCC1
IPR025110	0	1	0	AMP-binding enzyme C-terminal domain
IPR020683	0	1	0	Ankyrin repeat-containing domain
IPR006652	0	1	0	Kelch repeat type 1
IPR001810	0	1	0	F-box domain
IPR011765	0	0	1	Peptidase M16, N-terminal
IPR007863	0	0	1	Peptidase M16, C-terminal
IPR003008	0	0	1	Tubulin/FtsZ, GTPase domain

3.4.4 Phylogenetic Reconstruction

Undetectable protein domains can be either a consequence of loss events or simply be an artifact of clade- or even species-specific gains. To test the loss likelihood of a given protein domain the phylogenetic relationship of the testing framework needs to be know. Since the position of the phylum Tardigrada is is still under discussion, a phylogenomic analysis was performed to reconstruct the phylogeny underlying the 56 Ensembl Metazoa species and the three tardigrades. Single copy clusters of orthologs with at least one tardigrade member were identified and used for phylogenomic reconstruction. Using all target sequences (filtered Ensembl Metazoa Release 34 and the three tardigrades), 2245 of these cluster were detected (containin at least 15 species and at least one tardigrade). A Maximum Likelihood-based phylogenetic reconstruction based on this supermatrix placed *M. tardigradum*, *R. varieornatus* and *H. dujardini* as the sister group of the Nematodes (see figure 3.5). Additionally, all previous hypotheses placing the tardigrades, namely as the sister group of the Arthropods in a pan-Arthropoda cluster (i), as the sister group of the nematodes grouping the tardigrades into the cycloneuralia (ii) and as the outgroup to both, arthropods and nematodes (iii) (see figure 3.6) were tested [125, 181, 199, 229, 237, 196, 34, 196, 50, 106, 238, 273]. The super-matrix was used to extract the per site log-likelihood calculated by RaxML for each hypothesis and the "approximately unbiased test" as implemented in CONSEL was performed. The same approach was carried out on domain repertoire, domain architectures and shared orthologous groups of all thee tardigrades and the 56 Ensembl Metazoa species. In each case, the presence and absence of the feature was encoded in a binary matrix. For all data sets, the placement as sister group to the nematodes had the highest rank and the lowest p-value (see figure 3.6). Only using domain architectures, the hypothesis placing the tardigrades as the sister group of the Arthropods in a pan-Arthropoda cluster showed an equally good p-value. Taken together, the three different tardigrade genomes strongly support the tardigrades as members of a Cycloneuralia cluster and rejects their placement within the pan-Arthropoda.

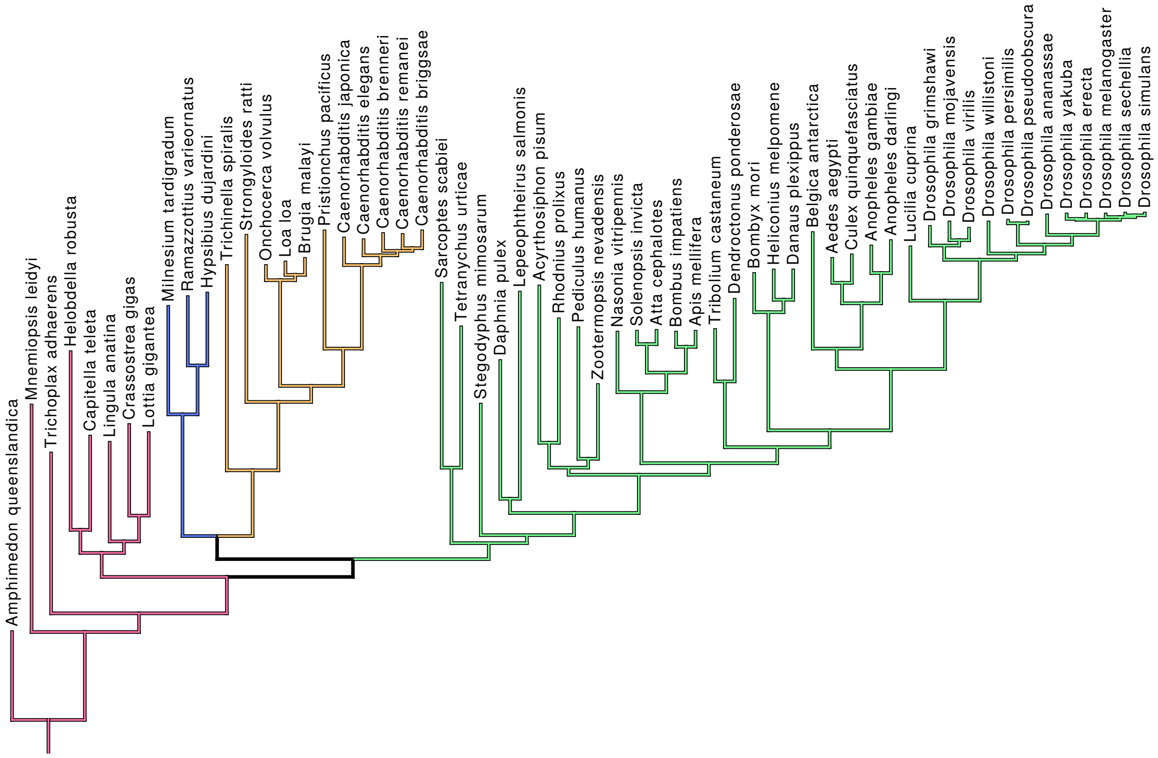
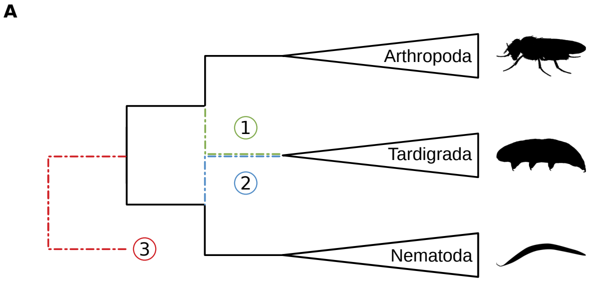


Fig. 3.5 Phylogenetic position of tardigrades. Maximum Likelihood based reconstruction of ecdysozoan phylogeny by a supermatrix approach using 2245 single copy clusters of orthologs. The three tardigrade species are placed as sister taxa to the nematodes. The tree was drawn and annotated with Mesquite [194]. Color code: Blue = Tardigrada; Yellow = Nematoda; Green = Arthropoda; Red = Outgroup



B

p-Value of the approximately Unbiased Test

Hypothesis	①	②	③
Sequence Supermatrix	0.014	**0.001**	0.986
Orthologous Groups	$9e^{-39}$	$\mathbf{2e^{-79}}$	1.000
Domain Occurences	$7e^{-5}$	$\mathbf{6e^{-5}}$	0.608
Domain Architectures	$1e^{-6}$	$\mathbf{2e^{-8}}$	1.000

Fig. 3.6 Phylogenetic position of tardigrades. A) According to three different hypotheses, tardigrades are either the sister taxon to the arthropods (1), the nematodes (2) or the outgroup to both (3). B) The probability of acceptance for each of these hypotheses was tested on different data sets. Hypothesis 2 ranked highest in all tests.

3.4.5 Protein Domain Gain and Loss Estimation

A substantial domain loss has been reported on route from the Ur-ecdysozoan to current day nematodes [308]. The phylogenetic placement of tardigrades was used to assess this loss in detail. A maximum likelihood approach was used to reconstruct the domain repertoire of the extinct ancestors. A greater part of the losses happened before the split of tardigrades and nematodes followed by further nematode and tardigrade specific losses (see figure 3.7). Interestingly, no gene ontology terms were enriched in protein domains which were lost before the split of tardigrades and nematodes (Cycloneuralia losses) or in proteins specifically lost in tardigrades. The maximum likelihood approach also suggested several protein domain gains (see table 3.10), again without any enriched gene ontology terms. Manual inspection of the protein domains suggested 12 of them either being present due to contamination issues or horizontal gene transfer.

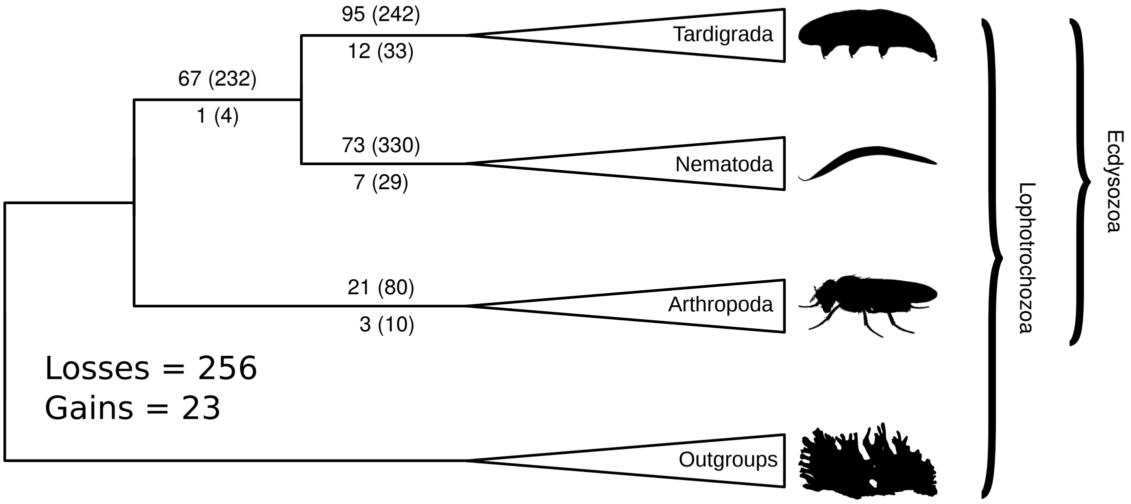


Fig. 3.7 Domain loss in three Ecdysozoan lineages (see figure 3.6 , hypothesis 2). The domain repertoire of ancestral species was reconstructed using Maximum Likelihood. Numbers above branches indicate losses and number underneath branches indicate gains. Bracketed numbers represent raw counts while none bracketed numbers indicated likelihood-corrected counts. Heavy losses were detected tardigrades and nematodes before and after their split.

Table 3.10 Protein domain gains after the split from the nematodes. Pfam IDs colored in red represent contamination or horizontal transferals. Bold Pfam IDs denote gains shared between all tardigrade species. Abbreviation: HD = *H. dujardini*, MT = *M. tardigradum*, RV = *R. varieornatus*

Pfam ID	MT	RV	HD	Description
PF06527	1	1	0	TniQ
PF04616	1	0	1	Glycosyl hydrolases family 43
PF07705	1	1	0	CARDB
PF08444	0	1	1	Aralkyl acyl-CoA:amino acid N-acyltransferase
PF09469	1	0	1	Cordon-bleu ubiquitin-like domain
PF13668	1	0	1	Ferritin-like domain
PF05621	1	1	0	Bacterial TniB protein
PF04236	1	1	1	Tc5 transposase C-terminal domain
PF10091	1	0	1	Putative glucoamylase
PF02015	1	1	0	Glycosyl hydrolase family 45
PF12585	1	0	1	Protein of unknown function (DUF3759)
PF01345	1	1	0	Domain of unknown function DUF11
PF17017	1	0	1	Aberrant zinc-finger
PF02987	1	1	0	Late embryogenesis abundant protein
PF12019	1	1	0	Type II transport protein GspH
PF03330	1	1	1	Lytic transglycolase
PF10988	1	0	1	Putative auto-transporter adhesin, head GIN domain
PF07484	1	1	0	Phage Tail Collar Domain
PF14539	1	0	1	Domain of unknown function (DUF4442)
PF04397	1	1	0	LytTr DNA-binding domain
PF03169	1	1	0	OPT oligopeptide transporter protein
PF11443	1	1	1	Domain of unknown function (DUF2828)
PF03269	1	1	0	Caenorhabditis protein of unknown function, DUF268
PF03851	1	0	1	UV-endonuclease UvdE
PF02954	1	0	1	Bacterial regulatory protein, Fis family
PF08448	1	1	0	PAS fold
PF13394	1	1	0	4Fe-4S single cluster domain
PF16732	1	1	0	Type IV minor pilin ComP, DNA uptake sequence receptor
PF15099	1	1	1	Phosphoinositide-interacting protein family
PF13441	0	1	1	YMGG-like Gly-zipper
PF13930	0	1	1	DNA/RNA non-specific endonuclease
PF17124	1	0	1	ThiJ/PfpI family-like

3.4.6 Functional Hypotheses Testing

Based on previous molecular studies and the recent release of the *R. varieornatus* and *H. dujardini* genome several functional hypotheses were checked. See section 3.3.9 for methodical details.

Trehalose & LEA The first line of defense against damage caused by anhydrobiosis is the stabilization of cellular structures. In many species studied so far, one of two biomolecules, trehalose [68, 59, 23] and LEA proteins [110, 282, 168], are involved. A protein domain based annotation of key components of the trehalose biosynthesis and metabolism pathway suggested a possible route from D-Glucose to Trehalose being present in all three tardigrades (see figure 3.8). The *R. varieornatus* genome seems to encode a second route to convert UDP-Glucose to Trehalose. Additionally, all three species encode enzymes to reconvert alpha,alpha-trehalose back to D-glucose and beta-D-glucose 1-phosphate using phosphate as a substrate. LEA proteins were present in only two of the tardigrades (*M. tardigradum*: 1, *H. dujardini*: 1) although Hashimoto et al. reported 10 proteins in *R. varieornatus*. A detailed inspection revealed that none of the LEA-classified proteins contained any of the typical Pfam protein domains but were mostly classified as "Bacterial protein of unknown function".

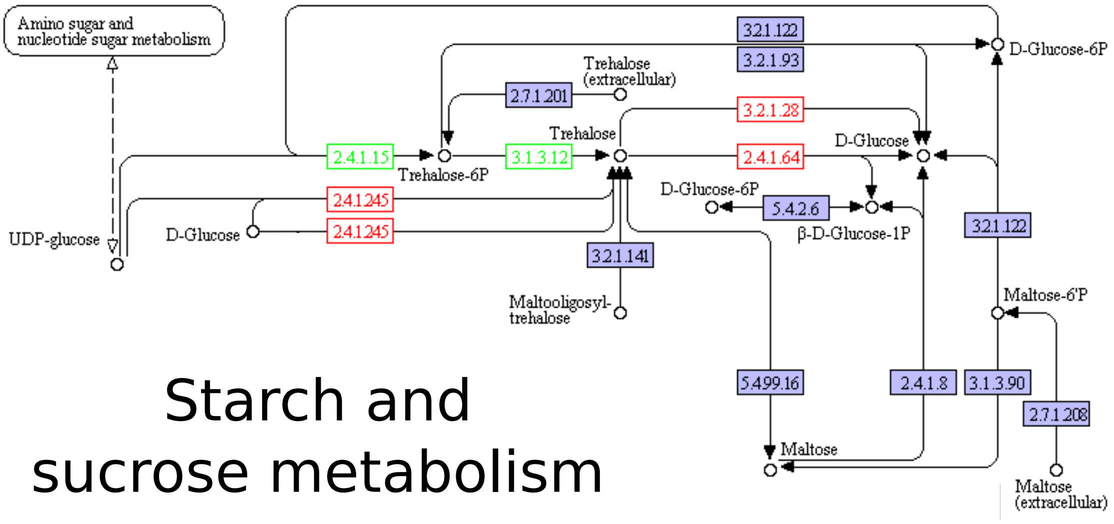


Fig. 3.8 Components of the trehalose biosynthesis & metabolism pathway in tardigrades. Numbers represent enzyme commission numbers. Color code: Red = Enzymes present in all three tardigrades; Green = Enzymes present in *R. varieornatus*; 3.2.1.141, 5.499.16 and 3.2.1.93 are likely present as well since all tardigrades genomes encode at least one alpha-amylase (PF00128). 2.7.1.201 and 3.2.1.122 are missing in all three tardigrades.

HSPs Although via a different mechanism, also heat shock proteins (HSPs) can stabilize proteins under stress conditions. All typical protein families are encoded in the three tardigrade genomes (see 3.12). None of the families is significantly expanded in comparison to the other metazoan species. Only two HSPs show stronger expression in *R. varieornatus* (see figure 3.9). This is in congruence with further experimental evidence, which did not find a consistent importance for these families [153, 230]. Taken together, the three genomes as well as experimental studies indicate only a minor role for the stabilization of proteins.

Table 3.12 Heat Shock Proteins present in the three tardigrade genomes.

Family	Alternative Name	Pfam ID	InterPro	HD	MT	RV
HSP70	HSP70/HSP110	PF00012	IPR013126	69	12	13
DNAJ	HSP40	PF00226	IPR001623	43	42	33
HSPB	small HSPs	PF00011	IPR008978	11	9	7
HSPC	HSP90	PF00183	IPR001404	2	3	2
HSPD	GroEL	PF00118	IPR002423	13	17	10
HSPE	GroES	PF00166	IPR020818	1	1	1

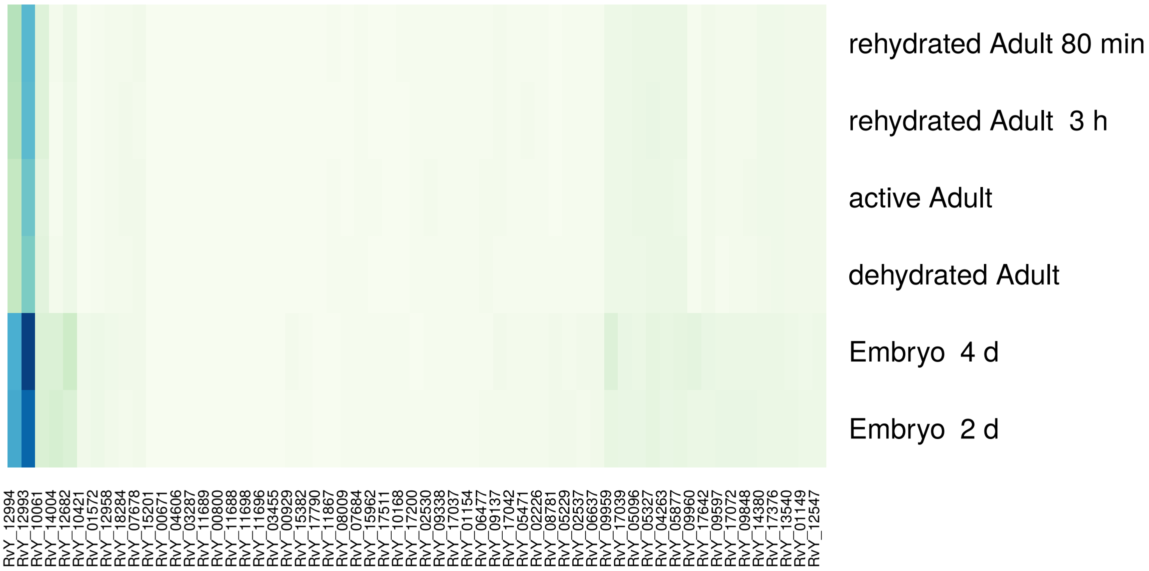


Fig. 3.9 Heat shock proteins expression in *R. varieornatus*

Tardigrada-unique genes Hashimoto et al. suggested the presence of several tardigrada-unique genes in the genome of *R. varieornatus* which are constitutively expressed and associated with stress tolerance. The most abundant expressed proteins included Cytoplasmic Abundant Heat Soluble (CAHS) and Secretory Abundant Heat Soluble (SAHS) [302] proteins that were previously identified as well as a newly characterized DNA Damage suppressor (Dsup) [121]. Furthermore, a Mitochondrial Abundant Heat Soluble (MAHS) was previously identified. Using the annotated *R. varieornatus* templates the genome of *H. dujardini* and *M. tardigradum* were screened using a profile and a homology based strategy. Both species encoded several CAHS (see table 3.13) but only *H. dujardini* encoded also 5 SAHS proteins. Neither *M. tardigradum* nor *H. dujardini* showed a homolog to Dsup. MAHS-like proteins were only found in *R. varieornatus* and *H. dujardini*. The erratic distribution of tardigrade-unique proteins underlines their variable relevance for stress tolerance but again suggests the phyla Tardigrada as a rich source of new protection genes and mechanisms.

Table 3.13 Tardigrada-unique genes associated with stress tolerance.

Species	CAHS	SAHS	MAHS	Dsup
R. varieornatus	16	13	2	1
M. tardigradum	5	0	0	0
H. dujardini	9	5	1	0

3.5 Discussion

The draft genomes of *M. tardigradum*, *R. varieornatus* and *H. dujardini* offer a first chance to understand the evolutionary adaptations required for the remarkable stress tolerance of tardigrades. So far, the genomes of only two other species capable of anhydrobiosis, the bdelloid rotifer *Adineta vaga* [95], and the antarctic midge *Belgica antarctica* [164] have been sequenced. In the case of *A. vaga*, the genome revealed different peculiarities in its structure and the inclusion of about 8% horizontally transferred genes [164]. In contrast, no such features are present in the genome of *M. tardigradum* and *R. varieornatus*. Horizontal gene transfer was suggested for *H. dujardini* but three independent studies indicated that the signal is likely due to heavy sequencing contamination [174, 27, 73].

The comparative framework established in this chapter provided new data to address the phylogenetic placement of the tardigrades, which still is discussed controversially. Mainly two hypotheses exist, either as the outgroup of the arthropods building the pan-Arthropoda taxon, or together with the nematodes as a member of the Cycloneuralia. A Maximum Likelihood approach was used to analyze the domain repertoire, the domain architectures and groups of orthologous genes. In all cases the placement of the tardigrades as sister group to the nematodes was the most highest ranking hypothesis. A phylogenetic reconstruction based on single copy cluster of orthologs found in 59 species including arthropods, nematodes, tardigrades and several basal eukaryotes likewise placed the tardigrades together with the nematodes.

Additionally, the comparative framework was used to detect expanded and contracted protein domains to provide hints on which biological processes and molecular functions are most likely associated with tardigrade specific traits. The analysis revealed numbers of several protein domains being significantly altered in each of the three tardigrade species and recovered previously published expansions in *R. varieornatus* [121]. Nevertheless, little overlap was detected when comparing the expanded or contracted protein domain complement of all three tardigrades (see figure 3.4). Some of the expanded protein domains encoded stress related gene families. *R. varieornatus* displayed a significant expansion of superoxide dismutases (SOD, IPR024134), a class of proteins necessary for the inactivation of ROS. Especially when desiccating, tardigrades are exposed to endogenous and exogenous stresses and inactivation of ROS might be an important line of defense for anhydrobiosis. An additional screen for protein domains related to ROS production and scavenging unexpectedly revealed that all three tardigrade genomes encode for an alternative oxidase (AOX, IPR002680). The protein is able to lower the internal production of ROS at the mitochondria [144, 286]. While common in bacteria, plants and fungi, AOX was thus far only found in few metazoan species mostly living in salt water. This includes the bdelloid rotifer *Adineta vaga* which is capable of anhydrobiosis [95] respectively cryptobiosis. In addition to the inactivation of ROS, avoiding its emergence would be a complementary strategy. The presence of AOX proteins might indicate an overlooked mechanisms for anhydrobiotic metazoa and tardigrades, the so far first terrestrial animal utilizing this mechanism for ROS defense. In addition to the chemical stresses, desiccation also leads to a physical stress by changing the volume of cells. In the case of plants, it has been suggested that this stress implies a third line of defense; the rearrangements of the cytoskeleton. Likewise, tardigrades might obviate structural stress by restructuring the cell wall, membranes and the cytoskeleton. The expansion analysis revealed three protein domains associated with catabolic cell wall processes are significantly over-represented in all of the three tardigrade species (Glycoside hydrolase, IPR000726 [93]; LysM domain, IPR018392 [154]; Chitin binding domain, IPR002557 [267]). On the contrary, the latter protein domains might also be involved in digestive processes since carnivorous tardigrades like *M. tardigradum* often consume their prey as a whole.

Protein domain losses (as extreme case of contractions) and gains (as extreme case of expansions) could arguably be an effect of contamination present in the 56 Ensembl Metazoa reference species. The reliability was tested reusing the phylogenetic positioning of the tardigrades alongside with the presence-absence information of protein domains. Ancestors states were imputed using Maximum likelihood estimate. The analysis revealed a substantial amount of domain losses that pre-dated the split of nematodes and tardigrades but also indicated further loss in both phyla independently. However, no comparable loss was found at the base of the arthropod lineage. In the resulting scenario, a major trend in the evolution of the last common ancestor of nematodes and tardigrades was the reduction of the domain repertoire starting from a complex Ur-Ecdysozoan. Thus, the evolution of nematodes and tardigrades recalls the general trend of reduction already observed at the base of the bilateria [55].

At last, the comparative framework was used to test several functional hypotheses that were devised on previous molecular experiments. It was suggested that the evolution of anhydrobiosis was as simple as finding the right water substitute [60, 101]. Nowadays, more complex models are discussed [64], but still the water replacement and vitrification of the cell plays a central role. Here, one possible adaptation is the accumulation of the sugar trehalose. An analysis of the trehalose biosynthesis and metabolism pathway showed that necessary components to generate and reconvert trehalose are present in all three tardigrades although to different extends. Small amounts of trehalose were previously found in Eutartigrades like *R. varieornatus* and *H. dujardini* but not in *M. tardigradum* [128, 131]. This could indicate independent adaptations of the pathway in different branches of the tardigrades. Thus, contrasting the three species on a transcriptional, translational or functional level in the future will provide insights into the different mechanisms underlying anhydrobiosis and the involvement of biomolecules such as trehalose. Another type of biomolecules which stabilize proteins in the case of desiccation are LEA proteins [287, 38]. Additionally, HSPs or the recently discovered Cytoplasmic Abundant Heat Soluble (CAHS), Secretory Abundant Heat Soluble (SAHS) and Mitochondrial Abundant Heat Soluble (MAHS) proteins as well as the newly characterized DNA Damage suppressor (Dsup) present in *R. varieornatus* might play an important role. A profile and homology based detection of the above-mentioned protein families revealed a scattered distribution of the protein similar to the observation of expanded and contracted protein domains. While HSPs and CAHS-like proteins were generally conserved in all three tardigrades, SAHS-like proteins were only discovered in the two Eutardigrades. LEA-like proteins were absent in *R. varieornatus* although otherwise reported [121], while Dsup was absent in *H. dujardini* and *M. tardigradum*. HSPs were generally weekly expressed in *R. varieornatus* and probably play only a minor role during desiccation. The scattered distribution of LEA, SAHS-like and tardigrade-unique proteins like Dsup suggests species or at least class specific adaptations towards typical tardigrade traits.

3.6 Epilogue

From a human's perspective, the stress tolerance of tardigrades is indeed remarkable. Still, the genomes of the three tardigrades are not that different from that of other Ecdysozoa. The available genomes of *R. varieornatus* and *H. dujardini* and the newly introduced genome of *M. tardigradum* allowed the delineation of their phylogenetic position within the Ecdysozoa and thereby helped to identify evolutionary trends within this metazoan lineage. Nevertheless, the comparative framework established in this chapter did not reveal general mechanisms behind the arguably most peculiar feature of tardigrades; their enormous stress tolerance. On the contrary, it revealed that the coding complements of the three tardigrades might have been shaped in much more species-specific manner than previously thought. The lack of further evidence from transcriptional, translational or generally experimental evidence made it nearly impossible to link these genomic species-specific footprints to the unusual physiological capabilities. Still, tardigrades are yet another example on how evolution tinkers even within the same phyla, rewires and modifies existing systems by smaller changes or generates completely new paths to remarkable phenotypic features.

3.7 Published elements

Bemm, F., Weiß, C. L., Schultz, J., and Förster, F. (2016b). Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? *Proceedings of the National Academy of Sciences of the United States of America*, 113(22):E3054—-6

3.8 Submitted elements

Bemm, F., Burleigh, L., Förster, F., Schmucki, R., Ebeling, M., Janzen, C., Dandekar, T., Schill, R., Certa, U., and Schultz, J. (2017). Draft genome of the Eutardigrade Milnesium tardigradum sheds light on ecdysozoan evolution. *bioRxiv* [The current version of the manuscript is accessible at bioRxiv using the working title.]

References

[1] Access, D. N. A. R. A. (2008). Genome Structure of the Legume, Lotus japonicus. *DNA Research*, pages 1–13.

[2] Adamec, L. (1997). Mineral nutrition of carnivorous plants: A review. *The Botanical Review*, 63(3):273–299.

[3] Adamec, L. (2006). Respiration and photosynthesis of bladders and leaves of aquatic Utricularia species. *Plant Biology*, 8(6):765–769.

[4] Adamec, L., Botany, S., and Republic, C. (2008). Soil fertilization enhances growth of the carnivorous plant Genlisea violacea. *Biologia*, 63(2):201–203.

[5] Ade, C. P., Bemm, F., Dickson, J. M. J., Walter, C., and Harris, P. J. (2014). Family 34 glycosyltransferase (GT34) genes and proteins in Pinus radiata (radiata pine) and Pinus taeda (loblolly pine). *Plant Journal*, 78(2):305–318.

[6] Adlassnig, W., Koller-Peroutka, M., Bauer, S., Koshkin, E., Lendl, T., and Lichtscheidl, I. K. (2012). Endocytotic uptake of nutrients in carnivorous plants. *The Plant journal : for cell and molecular biology*, 71(2):303–313.

[7] Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632):489–493.

[8] Albert, V. A., Williams, S. E., and Chase, M. W. (1992). Carnivorous plants: phylogeny and structural evolution. *Science*, 257(5076):1491–1495.

[9] Alexa, A., Rahnenfuhrer, J., and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607.

[10] Altiero, T., Guidetti, R., Caselli, V., Cesari, M., and Rebecchi, L. (2011). Ultraviolet radiation tolerance in hydrated and desiccated eutardigrades. *Journal of Zoological Systematics and Evolutionary Research*, 49(s1):104–110.

[11] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped {BLAST} and {PSI-BLAST}: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402.

[12] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

[13] Andrews, S. (2010). FastQC - A quality control tool for high throughput sequence data.

[14] Ankenbrand, M. J., Weber, L., Becker, D., Förster, F., and Bemm, F. (2016). TBro: visualization and management of de novo transcriptomes. *Database*, 2016(0):baw146.

[15] Ardila-Garcia, A. M., Umphrey, G. J., and Gregory, T. R. (2010). An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints. *Insect Mol. Biol.*, 19(3):337–346.

[16] Aubry, S., Kelly, S., Kümpers, B. M. C., Smith-Unna, R. D., and Hibberd, J. M. (2014). Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis. *PLoS Genet*, 10(6):e1004365.

[17] Author, P., Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S.-i., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Klitgaard-Kristensen, D., Sejrup, H. P., Haflidason, H., Johnsen, S., Spurk, M., and Sei, Q. (2008). The Physcomitrella Genome Reveals Evolutionary Insights into the Conquest of Land by. *Sciences, New Series*, 319(5859):64–69.

[18] Bagowski, C. P., Bruins, W., and Te Velthuis, A. J. W. (2010). The nature of protein domain evolution: shaping the interaction network. *Current genomics*, 11(5):368–76.

[19] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5.

[20] Bárta, J., Stone, J. D., Pech, J., Sirová, D., Adamec, L., Campbell, M. A., and Štorchová, H. (2015). The transcriptome of Utricularia vulgaris, a rootless plant with minimalist genome, reveals extreme alternative splicing and only moderate sequence similarity with Utricularia gibba. *BMC Plant Biology*, 15(1):78.

[21] Bartels, D. and Salamini, F. (2001). Desiccation tolerance in the resurrection plant Craterostigma plantagineum. A contribution to the study of drought tolerance at the molecular level. *Plant physiology*, 127(4):1346–53.

[22] Bauer, S., Grossmann, S., Vingron, M., and Robinson, P. N. (2008). Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651.

[23] Behm, C. A. (1997). The role of trehalose in the physiology of nematodes. *Int. J. Parasitol.*, 27(2):215–229.

[24] Beisser, D., Grohme, M. a., Kopka, J., Frohme, M., Schill, R. O., Hengherr, S., Dandekar, T., Klau, G. W., Dittrich, M., and Müller, T. (2012). Integrated pathway modules using time-course metabolic profiles and EST data from Milnesium tardigradum. *BMC systems biology*, 6(1):72.

[25] Bemm, F., Becker, D., Larisch, C., Kreuzer, I., Escalante-Perez, M., Schulze, W. X., Ankenbrand, M., Weyer, A.-l. V. D., Krol, E., Al-rasheid, K. A., Mithöfer, A., Weber, A. P., Schultz, J., and Hedrich, R. (2016a). Venus flytrap carnivorous lifestyle builds on herbivore defense strategies. *Genome Res.*, 26(6):1–14.

[26] Bemm, F., Burleigh, L., Förster, F., Schmucki, R., Ebeling, M., Janzen, C., Dandekar, T., Schill, R., Certa, U., and Schultz, J. (2017). Draft genome of the Eutardigrade Milnesium tardigradum sheds light on ecdysozoan evolution. *bioRxiv*.

[27] Bemm, F., Weiß, C. L., Schultz, J., and Förster, F. (2016b). Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? *Proceedings of the National Academy of Sciences of the United States of America*, 113(22):E3054—-6.

[28] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41(D1):D36–42.

[29] Billi, D. and Potts, M. (2002). Life and death of dried prokaryotes. *Research in microbiology*, 153(1):7–12.

[30] Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res*, 14(5):988–995.

[31] Birol, I., Raymond, A., Jackman, S. D., Pleasance, S., Coope, R., Taylor, G. A., Yuen, M. M. S., Keeling, C. I., Brand, D., Vandervalk, B. P., Kirk, H., Pandoh, P., Moore, R. A., Zhao, Y., Mungall, A. J., Jaquish, B., Yanchuk, A., Ritland, C., Boyle, B., Bousquet, J., Ritland, K., MacKay, J., Bohlmann, J., and Jones, S. J. M. (2013). Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. *Bioinformatics*, 29(12):1492–1497.

[32] Böhm, J., Scherzer, S., Krol, E., Kreuzer, I., von Meyer, K., Lorey, C., Mueller, T. D. D., Shabala, L., Monte, I., Solano, R., Al-Rasheid, K. A. A. S., Rennenberg, H., Shabala, S., Neher, E., and Hedrich, R. (2016). The Venus Flytrap Dionaea muscipula Counts Prey-Induced Action Potentials to Induce Sodium Uptake. *Curr Biol*, 26(3):286–295.

[33] Bonnet, C. and Goeze, J. A. E. (1773). *Herrn Karl Bonnets Abhandlungen aus der Insektologie / aus dem Französischen übersetzt und mit einigen Zusätzen herausgegeben von Joh. August Ephraim Goeze*. Bey J.J. Gebauers Wittwe und Joh. Jac. Gebauer, Halle :.

[34] Borner, J., Rehm, P., Schill, R. O., Ebersberger, I., and Burmester, T. (2014). A transcriptome approach to ecdysozoan phylogeny. *Mol. Phylogenet. Evol.*, 80:79–87.

[35] Breton, C., Snajdrová, L., Jeanneau, C., Koca, J., and Imberty, A. (2005). Structures and mechanisms of glycosyltransferases. *Glycobiology*, 16(2):29R–37R.

[36] Broad Institute (2014). Picard - A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

[37] Brown, W. H. (1916). The Mechanism of Movement and the Duration of the Effect of Stimulation in the Leaves of Dionaea. *American Journal of Botany*, 3(2):68.

[38] Browne, J., Tunnacliffe, A., and Burnell, A. (2002). Anhydrobiosis: Plant desiccation gene found in a nematode. *Nature*, 416(6876):38–38.

[39] Brownlee, C. (2013). Carnivorous plants: trapping, digesting and absorbing all in one. *Curr Biol*, 23(17):R714–6.

[40] Buch, F., Kaman, W. E., Bikker, F. J., Yilamujiang, A., Mithofer, A., Mithöfer, A., and Hause, B. (2015). Nepenthesin protease activity indicates digestive fluid dynamics in carnivorous nepenthes plants. *PLoS One*, 10(3):e0118853.

[41] Buchen, B., Hensel, D., and Sievers, A. (1983). Polarity in mechanoreceptor cells of trigger hairs of Dionaea muscipula Ellis. *Planta*, 158(5):458–468.

[42] Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1):59–60.

[43] Buljan, M. and Bateman, A. (2009). The evolution of protein domain families. *Biochemical Society Transactions*, 37(4).

[44] Burdon Sanderson, J. (1872). Note on the electrical phenomena which accompany irritation of the leaf of Dionaea muscipula. *Proceedings of the Royal Society of London*, vol. 21(no. 139-147):495–496.

[45] Burleigh, J. G., Barbazuk, W. B., Davis, J. M., Morse, A. M., and Soltis, P. S. (2012). Exploring Diversification and Genome Size Evolution in Extant Gymnosperms through Phylogenetic Synthesis. *Journal of Botany*, 2012:1–6.

[46] Bush, G. L., Case, S. M., Wilson, A. C., and Patton, J. L. (1977). Rapid speciation and chromosomal evolution in mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 74(9):3942–6.

[47] Butler, J. L., Gotelli, N. J., and Ellison, A. M. (2008). Linking the brown and green: Nutrient transformation and fate in the Sarracenia microecosystem. *Ecology*, 89(4):898–904.

[48] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421.

[49] 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):1503–1509.

[50] Campbell, L. I., Rota-Stabelli, O., Edgecombe, G. D., Marchioro, T., Longhorn, S. J., Telford, M. J., Philippe, H., Rebecchi, L., Peterson, K. J., and Pisani, D. (2011). MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proceedings of the National Academy of Sciences of the United States of America*, 108(38):15920–4.

[51] Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic acids research*, 37(Database issue):D233–8.

[52] Cao, H. X., Schmutzer, T., Scholz, U., Pecinka, A., Schubert, I., and Vu, G. T. (2015). Metatranscriptome analysis reveals host-microbiome interactions in traps of carnivorous Genlisea species. *Front Microbiol*, 6:526.

[53] Carretero-Paulet, L., Chang, T. H., Librado, P., Ibarra-Laclette, E., Herrera-Estrella, L., Rozas, J., and Albert, V. A. (2015a). Genome-wide analysis of adaptive molecular evolution in the carnivorous plant Utricularia gibba. *Genome Biol Evol*, 7(2):444–456.

[54] Carretero-Paulet, L., Librado, P., Chang, T. H., Ibarra-Laclette, E., Herrera-Estrella, L., Rozas, J., and Albert, V. A. (2015b). High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant Utricularia gibba. *Mol Biol Evol*, 32(5):1284–1295.

[55] Chakrabortee, S., Boschetti, C., Walton, L. J., Sarkar, S., Rubinsztein, D. C., and Tunnacliffe, A. (2007). Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46):18073–8.

[56] Cheung, A. Y. and Wu, H.-M. (2011). THESEUS 1, FERONIA and relatives: a family of cell wall-sensing receptor kinases? *Current Opinion in Plant Biology*, 14(6):632–641.

[57] Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. In *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, volume 99, pages 45–56.

[58] Clarke, C. (2001). Nepenthes of Sumatra and Peninsular Malaysia. *Natural History Publications (Borneo), Kota Kinabalu.*

[59] CLEGG, J. S. (1965). THE ORIGIN OF TREHALOSE AND ITS SIGNIFICANCE DURING THE FORMATION OF ENCYSTED DORMANT EMBRYOS OF ARTEMIA SALINA. *Comparative biochemistry and physiology*, 14:135–43.

[60] Clegg, J. S. (2001). Cryptobiosis - a peculiar state of biological organization. *Comparative biochemistry and physiology. Part B*, 128(4):613–624.

[61] Clegg, J. S. (2005). Desiccation tolerance in encysted embryos of the animal extremophile, artemia. *Integrative and comparative biology*, 45(5):715–24.

[62] Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18):3674–3676.

[63] Cornette, R. and Kikawada, T. (2011). The induction of anhydrobiosis in the sleeping chironomid: current status of our knowledge. *IUBMB life*, 63(6):419–29.

[64] Crowe, J. H. (2014). Anhydrobiosis: an unsolved problem. *Plant Cell Environ.*, 37(7):1491–1493.

[65] Crowe, J. H. and Clegg, J. S. (1973). *Anhydrobiosis*. Dowden, Hutchinson {&} Ross, Inc., Stroudsburg, Pennsylvania,.

[66] Crowe, J. H., Hoekstra, F. A., and Crowe, L. M. (1992). Anhydrobiosis. *Annual Review of Physiology*, 54(1):579–599.

[67] Crowe, L. M. (2002). Lessons from nature: the role of sugars in anhydrobiosis. *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology*, 131(3):505–13.

[68] Crowe, L. M. and Crowe, J. H. (2007). Trehalose as a "chemical chaperone": fact and fantasy. *Adv. Exp. Med. Biol.*, 594:143–158.

[69] Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)*, 27(8):1164–5.

[70] Darwin, C. (1875). *Insectivorous Plants*. Murray, London, UK.

[71] Davies, G. J. (2001). Sweet secrets of synthesis. *Nature Structural Biology*, 8(2):98–100.

[72] Delignette-Muller, M. L. and Dutang, C. (2015). {fitdistrplus}: An {R} Package for Fitting Distributions. *Journal of Statistical Software*, 64(4):1–34.

[73] Delmont, T. O. and Eren, A. M. (2016). Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, 4:e1839.

[74] Doblin, M. S., Pettolino, F., and Bacic, A. (2010). <i>Evans Review</i> : Plant cell walls: the skeleton of the plant world. *Functional Plant Biology*, 37(5):357.

[75] Doyere, L.-M.-F. (1842). Mémoire sur les Tardigrades. *Annales des Sciences Naturelles*, Seconde Sé(Tome XVIII).

[76] Driouich, A., Follet-Gueye, M.-L., Bernard, S., Kousar, S., Chevalier, L., Vicré-Gibouin, M., and Lerouxel, O. (2012). Golgi-mediated synthesis and secretion of matrix polysaccharides of the primary cell wall of higher plants. *Frontiers in plant science*, 3:79.

[77] Duan, Q., Kita, D., Li, C., Cheung, A. Y., and Wu, H.-M. (2010). FERONIA receptor-like kinase regulates RHO GTPase signaling of root hair development. *Proceedings of the National Academy of Sciences*, 107(41):17821–17826.

[78] Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. a., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., and Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–9.

[79] Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10):e1002195.

[80] Edgecombe, G. D. (2009). Palaeontological and Molecular Evidence Linking Arthropods, Onychophorans, and other Ecdysozoa. *Evolution: Education and Outreach*, 2(2):178–190.

[81] Edwards, M. E., Dickson, C. A., Chengappa, S., Sidebottom, C., Gidley, M. J., and Reid, J. S. G. (1999). Molecular characterisation of a membrane-bound galactosyltransferase of plant cell wall matrix polysaccharide biosynthesis. *Plant Journal*, 19(6):691–697.

[82] Ekseth, O. K., Kuiper, M., and Mironov, V. (2014). orthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics (Oxford, England)*, 30(5):734–6.

[83] Ellison, A. M. (2006). Nutrient limitation and stoichiometry of carnivorous plants. *Plant Biology*, 8(6):740–747.

[84] Ellison, A. M. and Farnsworth, E. J. (2005). The cost of carnivory for Darlingtonia californica (Sarraceniaceae): Evidence from relationships among leaf traits. *American Journal of Botany*, 92(7):1085–1093.

[85] Emilia, R., Charles, L., and Survey, A. (2016). Actual checklist of Tardigrada species (2009-2016, 31. Technical Report 1, Department of Life Sciences, University of Modena and Reggio Emilia, Modena.

[86] Erkut, C., Vasilj, A., Boland, S., Habermann, B., Shevchenko, A., and Kurzchalia, T. V. (2013). Molecular strategies of the Caenorhabditis elegans dauer larva to survive extreme desiccation. *PLoS One*, 8(12):e82473.

[87] Escalante-Pérez, M., Krol, E., Stange, A., Geiger, D., Al-Rasheid, K. A. S., Hause, B., Neher, E., and Hedrich, R. (2011). A special pair of phytohormones controls excitability, slow closure, and external stomach formation in the Venus flytrap. *Proceedings of the National Academy of Sciences of the United States of America*, 108(37):15492–15497.

[88] Escalante-Perez, M., Scherzer, S., Al-Rasheid, K. A., Dottinger, C., Neher, E., and Hedrich, R. (2014). Mechano-stimulation triggers turgor changes associated with trap closure in the Darwin plant Dionaea muscipula. *Mol Plant*, 7(4):744–746.

[89] Faik, A., Price, N. J., Raikhel, N. V., and Keegstra, K. (2002). An Arabidopsis gene encoding an -xylosyltransferase involved in xyloglucan biosynthesis. *Proceedings of the National Academy of Sciences*, 99(11):7797–7802.

[90] Fang, H. (2014). dcGOR: An R Package for Analysing Ontologies and Protein Domain Annotations. *PLoS Computational Biology*, 10(10):e1003929.

[91] Felsenstein, J. (1989). PHYLIP - Phylogeny inference package - v3.2. *Cladistics*, 5:164–166.

[92] Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S., Wu, C. H., Xenarios, I., Yeh, L.-S., Young, S.-Y., and Mitchell, A. L. (2017). InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Research*, 45(D1):D190–D199.

[93] Flach, J., Pilet, P. E., and Jollès, P. (1992). What's new in chitinase research? *Experientia*, 48(8):701–16.

[94] Fleischmann, A., Michael, T. P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E. M., Greilhuber, J., Muller, K. F., and Heubl, G. (2014). Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann Bot*, 114(8):1651–1663.

[95] Flot, J.-F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E. G. J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthélémy, R.-M., Bast, J., Bazykin, G. A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J. F., Vakhrusheva, O. A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A. S., Mark Welch, D. B., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O., and Van Doninck, K. (2013). Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. *Nature*, 500(7463):453–457.

[96] Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545.

[97] Fornara, F., Panigrahi, K. C., Gissot, L., Sauerbrunn, N., Rühl, M., Jarillo, J. A., and Coupland, G. (2009). Arabidopsis DOF Transcription Factors Act Redundantly to Reduce CONSTANS Expression and Are Essential for a Photoperiodic Flowering Response. *Developmental Cell*, 17(1):75–86.

[98] Förster, F., Beisser, D., Grohme, M. A., Liang, C., Mali, B., Siegl, A. M., Engelmann, J. C., Shkumatov, A. V., Schokraie, E., Müller, T., Schnölzer, M., Schill, R. O., Frohme, M., and Dandekar, T. (2012). Transcriptome analysis in tardigrade species reveals specific molecular pathways for stress adaptations. *Bioinform. Biol. Insights*, 6:69–96.

[99] Forterre, Y., Skotheim, J. M., Dumais, J., and Mahadevan, L. (2005). How the Venus flytrap snaps. *Nature*, 433(7024):421–425.

[100] Fukushima, K., Fang, X., Alvarez-Ponce, D., Cai, H., Carretero-Paulet, L., Chen, C., Chang, T.-H., Farr, K. M., Fujita, T., Hiwatashi, Y., Hoshi, Y., Imai, T., Kasahara, M., Librado, P., Mao, L., Mori, H., Nishiyama, T., Nozawa, M., Pálfalvi, G., Pollard, S. T., Rozas, J., Sánchez-Gracia, A., Sankoff, D., Shibata, T. F., Shigenobu, S., Sumikawa, N., Uzawa, T., Xie, M., Zheng, C., Pollock, D. D., Albert, V. A., Li, S., and Hasebe, M. (2017). Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. *Nature Ecology & Evolution*, 1(3):0059.

[101] Galau, G. A., Bijaisoradat, N., and Hughes, D. W. (1987). Accumulation kinetics of cotton late embryogenesis-abundant mRNAs and storage protein mRNAs: coordinate regulation during embryogenesis and the role of abscisic acid. *Developmental biology*, 123(1):198–212.

[102] Gallie, D. R. and Chang, S. C. (1997). Signal transduction in the carnivorous plant Sarracenia purpurea. Regulation of secretory hydrolase expression during development and in response to resources. *Plant physiology*, 115(4):1461–71.

[103] Gao, P., Loeffler, T. S., Honsel, A., Kruse, J., Krol, E., Scherzer, S., Kreuzer, I., Bemm, F., Buegger, F., Burzlaff, T., Hedrich, R., and Rennenberg, H. (2015). Integration of trap- and root-derived nitrogen nutrition of carnivorous Dionaea muscipula. *The New phytologist*, 205(3):1320–9.

[104] Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695.

[105] Gibson, T. C. and Waller, D. M. (2009). Evolving Darwin's 'most wonderful' plant: Ecological steps to a snap-trap. *New Phytologist*, 183(3):575–587.

[106] Giribet, G. and Edgecombe, G. D. (2012). Reevaluating the arthropod tree of life. *Annu. Rev. Entomol.*, 57:167–186.

[107] Gish, L. A. and Clark, S. E. (2011). The RLK/Pelle family of kinases. *The Plant Journal*, 66(1):117–127.

[108] Givnish, T. J., Burkhardt, E. L., Happel, R. E., and Weintraub, J. D. (1984). Carnivory in the Bromeliad Brocchinia reducta, with a Cost/Benefit Model for the General Restriction of Carnivorous Plants to Sunny, Moist, Nutrient-Poor Habitats. *The American Naturalist*, 124(4):479–497.

[109] Goyal, K., Walton, L. J., Browne, J. A., Burnell, A. M., and Tunnacliffe, A. (2005a). Molecular anhydrobiology: identifying molecules implicated in invertebrate anhydrobiosis. *Integrative and comparative biology*, 45(5):702–9.

[110] Goyal, K., Walton, L. J., and Tunnacliffe, A. (2005b). LEA proteins prevent protein aggregation due to water stress. *The Biochemical journal*, 388(Pt 1):151–7.

[111] Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–52.

[112] Gregory, T. R. and Johnston, J. S. (2008). Genome size diversity in the family Drosophilidae. *Heredity*, 101(3):228–238.

[113] Group, A. P. (1998). An Ordinal Classification for the Families of Flowering Plants. *Annals of the Missouri Botanical Garden*, 85(4):531.

[114] Group, A. P. (2003). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society*, 141(4):399–436.

[115] Guindon, S., Dufayard, J.-F. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3):307–21.

[116] Guo, A.-Y., Chen, X., Gao, G., Zhang, H., Zhu, Q.-H., Liu, X.-C., Zhong, Y.-F., Gu, X., He, K., and Luo, J. (2007). PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Research*, 36(Database):D966–D969.

[117] Guo, H., Li, L., Ye, H., Yu, X., Algreen, A., and Yin, Y. (2009). Three related receptor-like kinases are required for optimal cell elongation in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, 106(18):7648–7653.

[118] Haas, B. J. (2014). TransDecoder (Find Coding Regions Within Transcripts).

[119] Haberlandt, G. (1906). *Sinnesorgane im Pflanzenreich zur Perzeption mechanischer Reize*. Engelmann.

[120] Harris, P. J. and Stone, B. A. (2005). Chemistry and Molecular Organization of Plant Cell Walls. In *Biomass Recalcitrance*, pages 61–93. Blackwell Publishing Ltd., Oxford, UK.

[121] Hashimoto, T., Horikawa, D. D., Saito, Y., Kuwahara, H., Kozuka-Hata, H., Shin-I, T., Minakuchi, Y., Ohishi, K., Motoyama, A., Aizu, T., Enomoto, A., Kondo, K., Tanaka, S., Hara, Y., Koshikawa, S., Sagara, H., Miura, T., Yokobori, S.-i., Miyagawa, K., Suzuki, Y., Kubo, T., Oyama, M., Kohara, Y., Fujiyama, A., Arakawa, K., Katayama, T., Toyoda, A., and Kunieda, T. (2016). Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nature communications*, 7:12808.

[122] Hatano, N. and Hamada, T. (2008). Proteome analysis of pitcher fluid of the carnivorous plant Nepenthes alata. *Journal of Proteome Research*, 7(2):809–816.

[123] Haupt, W. (1982). Physiology of Movement. In *Progress in Botany / Fortschritte der Botanik*, pages 222–230. Springer Berlin Heidelberg, Berlin, Heidelberg.

[124] Hedrich, R. (2015). Carnivorous plants. *Current biology : CB*, 25(3):R99—-R100.

[125] Hejnol, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., Martinez, P., Baguñà, J., Bailly, X., Jondelius, U., Wiens, M., Müller, W. E. G., Seaver, E., Wheeler, W. C., Martindale, M. Q., Giribet, G., and Dunn, C. W. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. Biol. Sci.*, 276(1677):4261–4270.

[126] Hématy, K., Sado, P.-E., Van Tuinen, A., Rochange, S., Desnos, T., Balzergue, S., Pelletier, S., Renou, J.-P., and Höfte, H. (2007). A Receptor-like Kinase Mediates the Response of Arabidopsis Cells to the Inhibition of Cellulose Synthesis. *Current Biology*, 17(11):922–931.

[127] Hengherr, S., Brümmer, F., and Schill, R. O. (2008a). Anhydrobiosis in tardigrades and its effects on longevity traits. *Journal of Zoology*, 275(3):216–220.

[128] Hengherr, S., Heyer, A. G., Köhler, H.-R., and Schill, R. O. (2008b). Trehalose and anhydrobiosis in tardigrades–evidence for divergence in responses to dehydration. *The FEBS journal*, 275(2):281–8.

[129] Hengherr, S., Reuner, A., Brümmer, F., and Schill, R. O. (2010). Ice crystallization and freeze tolerance in embryonic stages of the tardigrade Milnesium tardigradum. *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology*, 156(1):151–5.

[130] Hengherr, S., Worland, M. R., Reuner, A., Brümmer, F., and Schill, R. O. (2009a). Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades. *The Journal of experimental biology*, 212(Pt 6):802–7.

[131] Hengherr, S., Worland, M. R., Reuner, A., Brümmer, F., and Schill, R. O. (2009b). High-temperature tolerance in anhydrobiotic tardigrades is limited by glass transition. *Physiological and biochemical zoology*, 82(6):749–55.

[132] Henry, R. J., editor (2005). *Plant diversity and evolution: genotypic and phenotypic variation in higher plants*. CABI, Wallingford.

[133] Hilu, K. W., Borsch, T., Muller, K., Soltis, D. E., Soltis, P. S., Savolainen, V., Chase, M. W., Powell, M. P., Alice, L. A., Evans, R., Sauquet, H., Neinhuis, C., Slotta, T. A. B., Rohwer, J. G., Campbell, C. S., and Chatrou, L. W. (2003). Angiosperm phylogeny based on matK sequence information. *American Journal of Botany*, 90(12):1758–1776.

[134] Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. *Bioinformatics*, 32(5):767–769.

[135] HOK, S., DANCHIN, E. G. J., ALLASIA, V., PANABIÈRES, F., ATTARD, A., and KELLER, H. (2011). An Arabidopsis (malectin-like) leucine-rich repeat receptor-like kinase contributes to downy mildew disease. *Plant, Cell & Environment*, 34(11):1944–1957.

[136] Holm, L. and Sander, C. (1994). Parser for protein folding units. *Proteins: Structure, Function, and Genetics*, 19(3):256–268.

[137] Huerta-Cepas, J., Forslund, K., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2016). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *bioRxiv*.

[138] Humphrey, T. V., Bonetta, D. T., and Goring, D. R. (2007). Sentinels at the wall: cell wall receptors and sensors. *New Phytologist*, 176(1):7–21.

[139] Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Computational Biology*, 12(6):e1004957.

[140] Huson, D. H., Mitra, S., Ruscheweyh, H.-J. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res*, 21(9):1552–1560.

[141] Hutchens, J. J. and Luken, J. O. (2009). Prey capture in the Venus flytrap: collection or selection? *Botany*, 87(Luken 2005):1007–1010.

[142] Ibarra-Laclette, E., Albert, V. A., Perez-Torres, C. A., Zamudio-Hernandez, F., Ortega-Estrada Mde, J., Herrera-Estrella, A., Herrera-Estrella, L., Pérez-Torres, C. A., Zamudio-Hernández, F., Ortega-Estrada, M. d. J., Herrera-Estrella, A., Herrera-Estrella, L., Perez-Torres, C. A., Zamudio-Hernandez, F., Ortega-Estrada Mde, J., Herrera-Estrella, A.,

Herrera-Estrella, L., Pérez-Torres, C. A., Zamudio-Hernández, F., Ortega-Estrada, M. d. J., Herrera-Estrella, A., and Herrera-Estrella, L. (2011). Transcriptomics and molecular evolutionary rate analysis of the bladderwort (Utricularia), a carnivorous plant with a minimal genome. *BMC Plant Biol*, 11(1):101.

[143] Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A. J., Juárez, M. J. A., Simpson, J., Fernández-Cortés, A., Arteaga-Vázquez, M., Góngora-Castillo, E., Acevedo-Hernández, G., Schuster, S. C., Himmelbauer, H., Minoche, A. E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Pérez, S. A., de Jesús Ortega-Estrada, M., Cervantes-Luevano, J. I., Michael, T. P., Mockler, T., Bryant, D., Herrera-Estrella, A., Albert, V. A., and Herrera-Estrella, L. (2013). Architecture and evolution of a minute plant genome. *Nature*, 498(7452):94–8.

[144] Ito, Y., Saisho, D., Nakazono, M., Tsutsumi, N., and Hirai, A. (1997). Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature. *Gene*, 203(2):121–129.

[145] Janin, J. and Chothia, C. (1985). Domains in proteins: Definitions, location, and structural principles. *Methods in Enzymology*, 115:420–430.

[146] Jendretzki, A., Wittland, J., Wilk, S., Straede, A., and Heinisch, J. J. (2011). How do I begin? Sensing extracellular stress to maintain yeast cell wall integrity. *European Journal of Cell Biology*, 90(9):740–744.

[147] Jensen, M. K., Vogt, J. K., Bressendorff, S., Seguin-Orlando, A., Petersen, M., Sicheritz-Pontén, T., and Mundy, J. (2015). Transcriptome and Genome Size Analysis of the Venus Flytrap. *PLOS ONE*, 10(4):e0123887.

[148] Jiang, H., Lei, R., Ding, S.-W. W., and Zhu, S. (2014). Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15(1):182.

[149] Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., and Hunter, S. (2014). {InterProScan} 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

[150] Jönsson, K. I. (2007). Tardigrades as a potential model organism in space research. *Astrobiology*, 7(5):757–66.

[151] Jönsson, K. I., Harms-Ringdahl, M., and Torudd, J. (2005). Radiation tolerance in the eutardigrade Richtersius coronifer. *International journal of radiation biology*, 81(9):649–56.

[152] Jönsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M., and Rettberg, P. (2008). Tardigrades survive exposure to space in low Earth orbit. *Current biology : CB*, 18(17):R729–R731.

[153] Jönsson, K. I. and Schill, R. O. (2007). Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade Richtersius coronifer. *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology*, 146(4):456–60.

[154] Joris, B., Englebert, S., Chu, C. P., Kariyama, R., Daneo-Moore, L., Shockman, G. D., and Ghuysen, J. M. (1992). Modular design of the Enterococcus hirae muramidase-2 and Streptococcus faecalis autolysin. *FEMS microbiology letters*, 70(3):257–64.

[155] Juniper, B. E., Robins, R. J., and Joel, D. M. (1989). *The Carnivorous Plants*. AcademicPress: Harcourt, Brace, Jovanovich, London, San Diego, New York, Berkeley, Boston, Sydney, Tokyo, Toronto, academic p edition.

[156] Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4):462–467.

[157] Kaessmann, H. (2009). Genetics. More than just a copy. *Science (New York, N.Y.)*, 325(5943):958–9.

[158] Käll, L., Krogh, A., Sonnhammer, E. L., Kall, L., Krogh, A., and Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 338(5):1027–1036.

[159] Karagatzides, J. D. and Ellison, A. M. (2009). Construction costs, payback times, and the leaf economics of carnivorous plants. *American Journal of Botany*, 96(9):1612–1619.

[160] Karin, M. (1990). Too many transcription factors: positive and negative interactions. *The New biologist*, 2(2):126–31.

[161] Kauffmann, A., Gentleman, R., and Huber, W. (2009). arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, 25(3):415–416.

[162] Keegstra, K. and Cavalier, D. (2010). Glycosyltransferases of the GT34 and GT37 Families. In *Annual Plant Reviews*, pages 235–249. Wiley-Blackwell, Oxford, UK.

[163] Keilin, D. (1959). The problem of anabiosis or latent life: history and current concept. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, 150(939):149–91.

[164] Kelley, J. L., Peyton, J. T., Fiston-Lavier, A.-S., Teets, N. M., Yee, M.-C., Johnston, J. S., Bustamante, C. D., Lee, R. E., and Denlinger, D. L. (2014). Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nature Communications*, 5:1–8.

[165] Kent, W. J. (2002). {BLAT–the} {BLAST-like} alignment tool. *Genome Res.*, 12(4):656–664.

[166] Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Humphrey, J., Kerhornou, A., Khobova, J., Aranganathan, N. K., Langridge, N., Lowy, E., McDowall, M. D., Maheswari, U., Nuhn, M., Ong, C. K., Overduin, B., Paulini, M., Pedro, H., Perry, E., Spudich, G., Tapanari, E., Walts, B., Williams, G., Tello–Ruiz, M., Stein, J., Wei, S., Ware, D., Bolser, D. M., Howe, K. L., Kulesha, E., Lawson, D., Maslen, G., and Staines, D. M. (2016). Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research*, 44(D1):D574–D580.

[167] Kersting, A. R., Bornberg-Bauer, E., Moore, A. D., and Grath, S. (2012). Dynamics and Adaptive Benefits of Protein Domain Emergence and Arrangements during Plant Genome Evolution. *Genome Biology and Evolution*, 4(3):316–329.

[168] Kikawada, T., Nakahara, Y., Kanamori, Y., Iwata, K.-i., Watanabe, M., McGee, B., Tunnacliffe, A., and Okuda, T. (2006). Dehydration-induced expression of LEA proteins in an anhydrobiotic chironomid. *Biochemical and biophysical research communications*, 348(1):56–61.

[169] Kim, Y., Tsuda, K., Igarashi, D., Hillmer, R., Sakakibara, H., Myers, C., and Katagiri, F. (2014). Mechanisms Underlying Robustness and Tunability in a Plant Immune Signaling Network. *Cell Host & Microbe*, 15(1):84–94.

[170] Kirch, T., Simon, R., Grünewald, M., and Werr, W. (2003). The DORNROSCHEN/ENHANCER OF SHOOT REGENERATION1 gene of Arabidopsis acts in the control of meristem cell fate and lateral organ development. *The Plant cell*, 15(3):694–705.

[171] Knepper, C., Savory, E. A., and Day, B. (2011). Arabidopsis NDR1 is an integrin-like protein with a role in fluid loss and plasma membrane-cell wall adhesion. *Plant physiology*, 156(1):286–300.

[172] Kohorn, B. D. and Kohorn, S. L. (2012). The cell wall-associated kinases, WAKs, as pectin receptors. *Frontiers in Plant Science*, 3.

[173] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv*.

[174] Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., and Conlon, C. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. *Pnas*, 113(23):1–6.

[175] Kreuzwieser, J., Scheerer, U., Kruse, J., Burzlaff, T., Honsel, A., Alfarraj, S., Georgiev, P., Schnitzler, J. P., Ghirardo, A., Kreuzer, I., Hedrich, R., and Rennenberg, H. (2014). The Venus flytrap attracts insects by the release of volatile organic compounds. *Journal of Experimental Botany*, 65(2):755–766.

[176] Kruse, J. J., Gao, P., Honsel, A., Kreuzwieser, J. J., Burzlaff, T., Alfarraj, S., Hedrich, R., and Rennenberg, H. (2014). Strategy of nitrogen acquisition and utilization by carnivorous Dionaea muscipula. *Oecologia*, 174(3):839–851.

[177] Lairson, L., Henrissat, B., Davies, G., and Withers, S. (2008). Glycosyltransferases: Structures, Functions, and Mechanisms. *Annual Review of Biochemistry*, 77(1):521–555.

[178] Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40(D1):D1202–10.

[179] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25.

[180] Lapinski, J. and Tunnacliffe, A. (2003). Anhydrobiosis without trehalose in bdelloid rotifers. *FEBS Letters*, 553(3):387–390.

[181] Lartillot, N. and Philippe, H. (2008). Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 363(1496):1463–1472.

[182] Latchman, D. S. (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–12.

[183] Leeuwenhoek, V. and Antony (1816). *The Select Works of Antony Van Leeuwenhoek: Containing His Microscopical Discoveries in Many of the Works of Nature, Volume 1*. Whittingham and Arliss, London.

[184] León, J., Rojo, E., and Sánchez-Serran, J. J. (2001). Wound signalling in plants. *Journal of Experimental Botany*, 52(354):1–9.

[185] Leushkin, E. V., Sutormin, R. a., Nabieva, E. R., Penin, A. a., Kondrashov, A. S., and Logacheva, M. D. (2013). The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. *BMC Genomics*, 14(1):476.

[186] Levin, D. E. (2011). Regulation of Cell Wall Biogenesis in Saccharomyces cerevisiae: The Cell Wall Integrity Signaling Pathway. *Genetics*, 189(4):1145–1175.

[187] Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, 106(27):11079–84.

[188] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics (IEEE InfoVis {\textquoteright}14).*

[189] Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12:323.

[190] Liu, J. X. and Howell, S. H. (2010). Endoplasmic reticulum protein quality control and its relationship to environmental stress responses in plants. *Plant Cell*, 22(9):2930–2942.

[191] Louis, J., Singh, V., and Shah, J. (2012). <i>Arabidopsis thaliana</i> —Aphid Interaction. *The Arabidopsis Book*, 10:e0159.

[192] Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):1–17.

[193] Luo, X., Bai, X., Sun, X., Zhu, D., Liu, B., Ji, W., Cai, H., Cao, L., Wu, J., Hu, M., Liu, X., Tang, L., and Zhu, Y. (2013). Expression of wild soybean WRKY20 in Arabidopsis enhances drought tolerance and regulates ABA signalling. *Journal of Experimental Botany*, 64(8):2155–2169.

[194] Maddison, W. P. and D.R. (2017). Mesquite: a modular system for evolutionary analysis.

[195] Mali, B., Grohme, M. a., Förster, F., Dandekar, T., Schnölzer, M., Reuter, D., Wełnicz, W., Schill, R. O., and Frohme, M. (2010). Transcriptome survey of the anhydrobiotic tardigrade Milnesium tardigradum in comparison with Hypsibius dujardini and Richtersius coronifer. *BMC genomics*, 11:168.

[196] Mayer, G., Martin, C., Rüdiger, J., Kauschke, S., Stevenson, P. A., Poprawa, I., Hohberg, K., Schill, R. O., Pflüger, H.-J., and Schlegel, M. (2013). Selective neuronal staining in tardigrades and onychophorans provides insights into the evolution of segmental ganglia in panarthropods. *BMC Evol. Biol.*, 13:230.

[197] McCLINTOCK, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6):344–55.

[198] Merlot, S., Leonhardt, N., Fenzi, F., Valon, C., Costa, M., Piette, L., Vavasseur, A., Genty, B., Boivin, K., Müller, A., Giraudat, J., and Leung, J. (2007). Constitutive activation of a plasma membrane H+-ATPase prevents abscisic acid-mediated stomatal closure. *The EMBO Journal*, 26(13):3216–3226.

[199] Meusemann, K., von Reumont, B. M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., von Haeseler, A., Burmester, T., Hadrys, H., Wägele, J. W., and Misof, B. (2010). A phylogenomic approach to resolve the arthropod tree of life. *Mol. Biol. Evol.*, 27(11):2451–2464.

[200] Mihailova, G., Petkova, S., Büchel, C., and Georgieva, K. (2011). Desiccation of the resurrection plant Haberlea rhodopensis at high temperature. *Photosynthesis research*, 108(1):5–13.

[201] Miller, C., Gurd, J., Brass, A., Wu, H.-B., Hwang, P.-I., Guo, L., Ashlock, D., Schnable, P., Almeida, N., Felipe, M., Kerlavage, A., McCombie, W., and Venter, J. (1999). A RAPID algorithm for sequence database comparisons: application to the identification of vector contamination in the EMBL databases. *Bioinformatics*, 15(2):111–121.

[202] Miranda-Saavedra, D. and Barton, G. J. (2007). Classification and functional annotation of eukaryotic protein kinases. *Proteins: Structure, Function, and Bioinformatics*, 68(4):893–914.

[203] Mithöfer, A., Reichelt, M., and Nakamura, Y. (2014). Wound and insect-induced jasmonate accumulation in carnivorous Drosera capensis: two sides of the same coin. *Plant Biology*, 16(5):982–987.

[204] Mittler, R., Vanderauwera, S., Gollery, M., and Van Breusegem, F. (2004). Reactive oxygen gene network of plants. *Trends Plant Sci*, 9(10):490–498.

[205] Miya, A., Albert, P., Shinya, T., Desaki, Y., Ichimura, K., Shirasu, K., Narusaka, Y., Kawakami, N., Kaku, H., and Shibuya, N. (2007). CERK1, a LysM receptor kinase, is essential for chitin elicitor signaling in Arabidopsis. *Proceedings of the National Academy of Sciences*, 104(49):19613–19618.

[206] Møbjerg, N., Halberg, K. a., Jørgensen, A., Persson, D., Bjørn, M., Ramløv, H., and Kristensen, R. M. (2011). Survival in extreme environments - on the current knowledge of adaptations in tardigrades. *Acta physiologica (Oxford, England)*, 202(3):409–20.

[207] Monshausen, G. B. and Gilroy, S. (2009). Feeling green: mechanosensing in plants. *Trends in Cell Biology*, 19(5):228–235.

[208] Moore, J. P., Lindsey, G. G., Farrant, J. M., and Brandt, W. F. (2007). An overview of the biology of the desiccation-tolerant resurrection plant Myrothamnus flabellifolia. *Annals of botany*, 99(2):211–7.

[209] Mor, A., Koh, E., Weiner, L., Rosenwasser, S., Sibony-Benyamini, H., and Fluhr, R. (2014). Singlet Oxygen Signatures Are Detected Independent of Light or Chloroplasts in Response to Multiple Stresses. *PLANT PHYSIOLOGY*, 165(1):249–261.

[210] Myhre, S., Tveit, H., Mollestad, T., and Laegreid, A. (2006). Additional gene ontology structure for improved biological reasoning. *Bioinformatics*, 22(16):2020–2027.

[211] Nakagawa, Y., Katagiri, T., Shinozaki, K., Qi, Z., Tatsumi, H., Furuichi, T., Kishigami, A., Sokabe, M., Kojima, I., Sato, S., Kato, T., Tabata, S., Iida, K., Terashima, A., Nakano, M., Ikeda, M., Yamanaka, T., and Iida, H. (2007). Arabidopsis plasma membrane protein crucial for Ca2+ influx and touch sensing in roots. *Proceedings of the National Academy of Sciences of the United States of America*, 104(9):3639–44.

[212] Neumann, S., Reuner, A., Brümmer, F., and Schill, R. O. (2009). DNA damage in storage cells of anhydrobiotic tardigrades. *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology*, 153(4):425–9.

[213] Neves, L. G., Davis, J. M., Barbazuk, W. B., and Kirst, M. (2013). Whole-exome targeted sequencing of the uncharacterized pine genome. *Plant Journal*, 75(1):146–156.

[214] Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hällman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Käller, M., Luthman, J., Lysholm, F., Niittylä, T., Olson, Å., Rilakovic, N., Ritland, C., Rosselló, J. A., Sena, J., Svensson, T., Talavera-López, C., Theißen, G., Tuominen, H., Vanneste, K., Wu, Z.-Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T. R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Lee Thompson, S., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P. K., Lundeberg, J., and Jansson, S. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497(7451):579–584.

[215] Osakabe, Y., Arinaga, N., Umezawa, T., Katsura, S., Nagamachi, K., Tanaka, H., Ohiraki, H., Yamada, K., Seo, S.-U., Abo, M., Yoshimura, E., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2013). Osmotic Stress Responses and Plant Growth Controlled by Potassium Transporters in Arabidopsis. *The Plant Cell*, 25(2):609–624.

[216] Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30(21):3123–3124.

[217] Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067.

[218] Pavlovic, A., Demko, V., and Hudák, J. (2010). Trap closure and prey retention in Venus flytrap (Dionaea muscipula) temporarily reduces photosynthesis and stimulates respiration. *Annals of botany*, 105(1):37–44.

[219] Pavlovič, A. and Saganová, M. (2015). A novel insight into the cost–benefit model for the evolution of botanical carnivory. *Annals of Botany*, 115(7):1075–1092.

[220] Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, 19(5):651–652.

[221] Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*, 8(10):785–786.

[222] Peyronnet, R., Tran, D., Girault, T., and Frachisse, J.-M. (2014). Mechanosensitive channels: feeling tension in a world under pressure. *Frontiers in Plant Science*, 5:558.

[223] Ponstingl, H. (2014). SMALT - A mapper for DNA sequencing reads.

[224] Poppinga, S., Kampowski, T., Metzger, A., Speck, O., and Speck, T. (2016). Comparative kinematical analyses of Venus flytrap (Dionaea muscipula) snap traps. *Beilstein Journal of Nanotechnology*, 7(1):664–674.

[225] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The Pfam protein families database. *Nucleic acids research*, 40(Database issue):D290—-301.

[226] Qian, W. and Zhang, J. (2014). Genomic evidence for adaptation by gene duplication. *Genome research*, 24(8):1356–62.

[227] Ramšak, Ž., Baebler, Š., Rotter, A., Korbar, M., Mozetič, I., Usadel, B., and Gruden, K. (2014). GoMapMan: integration, consolidation and visualization of plant gene annotations within the MapMan ontology. *Nucleic Acids Research*, 42(D1):D1167–D1175.

[228] Rebecchi, L. (2013). Dry up and survive: the role of antioxidant defences in anhydrobiotic organisms. *J. Limnol.*, 72(1s):62–72.

[229] Rehm, P., Borner, J., Meusemann, K., von Reumont, B. M., Simon, S., Hadrys, H., Misof, B., and Burmester, T. (2011). Dating the arthropod tree based on large-scale transcriptome data. *Mol. Phylogenet. Evol.*, 61(3):880–887.

[230] Reuner, A., Hengherr, S., Mali, B., Förster, F., Arndt, D., Reinhardt, R., Dandekar, T., Frohme, M., Brümmer, F., and Schill, R. O. (2010). Stress response in tardigrades: differential gene expression of molecular chaperones. *Cell stress & chaperones*, 15(4):423–30.

[231] Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in genetics : TIG*, 16(6):276–7.

[232] Rizzo, A. M., Negroni, M., Altiero, T., Montorfano, G., Corsetto, P., Berselli, P., Berra, B., Guidetti, R., and Rebecchi, L. (2010). Antioxidant defences in hydrated and desiccated states of the tardigrade Paramacrobiotus richtersi. *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology*, 156(2):115–21.

[233] Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011a). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329.

[234] Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., and Pachter, L. (2011b). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3):R22.

[235] Roberts, P. R. and Oosting, H. J. (1958). Responses of Venus Fly Trap (Dionaea muscipula) to Factors Involved in Its Endemism. *Ecological Monographs*, 28(2):193–218.

[236] Robins, R. J. (1976). The nature of the stimuli causing digestive juice secretion in Dionaea muscipula Ellis (venus's flytrap). *Planta*, 128(3):263–265.

[237] Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., von Haeseler, A., Kube, M., Reinhardt, R., and Burmester, T. (2007). {EST} sequencing of Onychophora and phylogenomic analysis of Metazoa. *Mol. Phylogenet. Evol.*, 45(3):942–951.

[238] Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J. L., Telford, M. J., Pisani, D., Blaxter, M., and Lavrov, D. V. (2010). Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biol. Evol.*, 2:425–440.

[239] Saier, M. H., Reddy, V. S., Tsu, B. V., Ahmed, M. S., Li, C., and Moreno-Hagelsieb, G. (2016). The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Research*, 44(D1):D372–D379.

[240] Savolainen, V., Chase, M. W., Hoot, S. B., Morton, C. M., Soltis, D. E., Bayer, C., Fay, M. F., de Bruijn, A. Y., Sullivan, S., and Qiu, Y.-L. L. (2000). Phylogenetics of Flowering Plants Based on Combined Analysis of Plastid atpB and rbcL Gene Sequences. *Systematic Biology*, 49(2).

[241] Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., and Ye, J. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic acids research*, 39(Database issue):D38–51.

[242] Scheller, H. V. and Ulvskov, P. (2010). Hemicelluloses. *Annual Review of Plant Biology*, 61(1):263–289.

[243] Scherzer, S., Böhm, J., Krol, E., Shabala, L., Kreuzer, I., Larisch, C., Bemm, F., Al-Rasheid, K. A. S., Shabala, S., Rennenberg, H., Neher, E., and Hedrich, R. (2015). Calcium sensor kinase activates potassium uptake systems in gland cells of Venus flytraps. *Proc Natl Acad Sci U S A*, 112(23):7309–7314.

[244] Schill, R. O. and Fritz, G. B. (2008). Desiccation tolerance in embryonic stages of the tardigrade. *Journal of Zoology*, 276(1):103–107.

[245] Schill, R. O., Steinbrück, G. H. B., and Köhler, H.-R. (2004). Stress gene (hsp70) sequences and quantitative expression in Milnesium tardigradum (Tardigrada) during active and cryptobiotic stages. *J. Exp. Biol.*, 207(Pt 10):1607–1613.

[246] Schokraie, E., Hotz-Wagenblatt, A., Warnken, U., Mali, B., Frohme, M., Förster, F., Dandekar, T., Hengherr, S., Schill, R. O., and Schnölzer, M. (2010). Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms. *PloS one*, 5(3):e9502.

[247] Schokraie, E., Warnken, U., Hotz-Wagenblatt, A., Grohme, M. A., Hengherr, S., Förster, F., Schill, R. O., Frohme, M., Dandekar, T., and Schnölzer, M. (2012). Comparative proteome analysis of Milnesium tardigradum in early embryonic state versus adults in active and anhydrobiotic state. *PLoS One*, 7(9):e45682.

[248] Schulze, W., Schulze, E. D., Schulze, I., and Oren, R. (2001). Quantification of insect nitrogen utilization by the venus fly trap Dionaea muscipula catching prey with highly variable isotope signatures. *Journal of experimental botany*, 52(358):1041–9.

[249] Schulze, W. X., Sanggaard, K. W., Kreuzer, I., Knudsen, A. D., Bemm, F., Thogersen, I. B., Brautigam, A., Thomsen, L. R., Schliesky, S., Dyrlund, T. F., Escalante-Perez, M., Becker, D., Schultz, J. J., Karring, H., Weber, A., Hojrup, P., Hedrich, R., and Enghild, J. J. (2012). The Protein Composition of the Digestive Fluid from the Venus Flytrap Sheds Light on Prey Digestion Mechanisms. *Molecular & Cellular Proteomics*, 11(11):1306–1319.

[250] Shanker, S., Paulson, A., Edenberg, H. J., Peak, A., Perera, A., Alekseyev, Y. O., Beckloff, N., Bivens, N. J., Donnelly, R., Gillaspy, A. F., Grove, D., Gu, W., Jafari, N., Kerley-Hamilton, J. S., Lyons, R. H., Tepper, C., and Nicolet, C. M. (2015). Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *Journal of biomolecular techniques : JBT*, 26(1):4–18.

[251] Shannon, A. J., Browne, J. A., Boyd, J., Fitzpatrick, D. A., and Burnell, A. M. (2005). The anhydrobiotic potential and molecular phylogenetics of species and strains of Panagrolaimus (Nematoda, Panagrolaimidae). *The Journal of experimental biology*, 208(Pt 12):2433–45.

[252] Shcheglov, A. S., Zhulidov, P. A., Bogdanova, E. A., and Shagin, D. A. (2007). *NORMALIZATION OF cDNA LIBRARIES*. Springer Netherlands.

[253] Sheard, L. B., Tan, X., Mao, H., Withers, J., Ben-Nissan, G., Hinds, T. R., Kobayashi, Y., Hsu, F. F., Sharon, M., Browse, J., He, S. Y., Rizo, J., Howe, G. A., and Zheng, N. (2010). Jasmonate perception by inositol-phosphate-potentiated COI1-JAZ co-receptor. *Nature*, 468(7322):400–405.

[254] Shimodaira, H. and Hasegawa, M. (2001). {CONSEL}: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246–1247.

[255] Shiu, S.-H. and Bleecker, A. B. (2001). Plant Receptor-Like Kinase Gene Family: Diversity, Function, and Signaling. *Science Signaling*, 2001(113):re22–re22.

[256] Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212.

[257] Singh, K. (2002). Transcription factors in plant defense and stress responses. *Current Opinion in Plant Biology*, 5(5):430–436.

[258] Smit, AFA, Hubley, R & Green, P. (2015). RepeatMasker Open-4.0.

[259] SOLTIS, D. E., SOLTIS, P. S., CHASE, M. W., MORT, M. E., ALBACH, D. C., ZANIS, M., SAVOLAINEN, V., HAHN, W. H., HOOT, S. B., FAY, M. F., AXTELL, M., SWENSEN, S. M., PRINCE, L. M., KRESS, W. J., NIXON, K. C., and FARRIS, J. S. (2000). Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. *Botanical Journal of the Linnean Society*, 133(4):381–461.

[260] Sonnhammer, E. L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*, 6:175–82.

[261] Spallanzani, L. (1776). *Opuscoli di Fisica animale e vegetabile*. Societa tipografica.

[262] Srivastava, A., Rogers, W. L., Breton, C. M., Cai, L., and Malmberg, R. L. (2011). Transcriptome analysis of sarracenia, an insectivorous plant. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 18(4):253–61.

[263] Stamatakis, A. (2014). {RAxML} version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

[264] Stephens, J. D., Rogers, W. L., Heyduk, K., Cruse-Sanders, J. M., Determann, R. O., Glenn, T. C., and Malmberg, R. L. (2015). Resolving phylogenetic relationships of the recently radiated carnivorous plant genus Sarracenia using target enrichment. *Mol Phylogenet Evol*, 85:76–87.

[265] Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, 49(2):169–181.

[266] Su, T., Xu, Q., Zhang, F.-C., Chen, Y., Li, L.-Q., Wu, W.-H., and Chen, Y.-F. (2015). WRKY42 Modulates Phosphate Homeostasis through Regulating Phosphate Translocation and Acquisition in Arabidopsis. *Plant Physiology*, 167(4):1579–1591.

[267] Suetake, T., Tsuda, S., Kawabata, S., Miura, K., Iwanaga, S., Hikichi, K., Nitta, K., and Kawano, K. (2000). Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. *The Journal of biological chemistry*, 275(24):17929–32.

[268] Sugimoto-Shirasu, K., Roberts, G. R., Stacey, N. J., McCann, M. C., Maxwell, A., and Roberts, K. (2005). RHL1 is an essential component of the plant DNA topoisomerase VI complex and is required for ploidy-dependent cell growth. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18736–41.

[269] Supek, F., Bošnjak, M., Škunca, N., Šmuc, T., and O'Donovan, C. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLoS ONE*, 6(7):e21800.

[270] Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic acids research*, 36(Database issue):D1009—-14.

[271] TAKAHASHI, K., MATSUMOTO, K., NISHII, W., MURAMATSU, M., and KUBOTA, K. (2009). COMPARATIVE STUDIES ON THE ACID PROTEINASE ACTIVITIES IN THE DIGESTIVE FLUIDS OF. *Carnivorous Plant Newsletter*, 38.

[272] Taylor, P. and Royal Botanic Gardens, K. (1994). *The genus Utricularia : a taxonomic monograph*. Royal Botanic Gardens, London.

[273] Telford, M. J., Bourlat, S. J., Economou, A., Papillon, D., and Rota-Stabelli, O. (2008). The evolution of the Ecdysozoa. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1496).

[274] Tenlen, J. R., Smith, F. W., Wang, J. R., Kiera, A., Nishimura, E. O., Tintori, S. C., Li, Q., Jones, C. D., Yandell, M., Messina, D. N., and Goldstein, B. (2016). Correction for Boothby et al., Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences*, 113(36):E5364—-E5364.

[275] The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169.

[276] Thoren, L. M. and Karlsson, P. S. (1998). Effects of supplementary feeding on growth and reproduction of three carnivorous plant species in a subarctic environment. *Journal of Ecology*, 86(3):501–510.

[277] Tökés, Z. A., Woon, W. C., and Chambers, S. M. (1974). Digestive enzymes secreted by the carnivorous plant Nepenthes macferlanei L. *Planta*, 119(1):39–46.

[278] Tokunaga, T., Takada, N., and Ueda, M. (2004). Mechanism of antifeedant activity of plumbagin, a compound concerning the chemical defense in carnivorous plant. *Tetrahedron Letters*, 45(38):7115–7119.

[279] Tran, T. D., Cao, H. X., Jovtchev, G., Neumann, P., Novak, P., Fojtova, M., Vu, G. T., Macas, J., Fajkus, J., Schubert, I., and Fuchs, J. (2015). Centromere and telomere sequence alterations reflect the rapid genome evolution within the carnivorous plant genus Genlisea. *Plant J*, 84(6):1087–1099.

[280] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515.

[281] Tunnacliffe, A. and Lapinski, J. (2003). Resurrecting Van Leeuwenhoek's rotifers: a reappraisal of the role of disaccharides in anhydrobiosis. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 358(1438):1755–71.

[282] Tunnacliffe, A. and Wise, M. J. (2007). The continuing conundrum of the LEA proteins. *Die Naturwissenschaften*, 94(10):791–812.

[283] Tyson, T., O'Mahony Zamora, G., Wong, S., Skelton, M., Daly, B., Jones, J. T., Mulvihill, E. D., Elsworth, B., Phillips, M., Blaxter, M., and Burnell, A. M. (2012). A molecular analysis of desiccation tolerance mechanisms in the anhydrobiotic nematode Panagrolaimus superbus using expressed sequenced tags. *BMC research notes*, 5(1):68.

[284] USADEL, B., POREE, F., NAGEL, A., LOHSE, M., CZEDIK-EYSENBERG, A., and STITT, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. *Plant, Cell & Environment*, 32(9):1211–1229.

[285] Van Dongen, S. (2008). Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141.

[286] Vanlerberghe, G. and McIntosh, L. (1997). ALTERNATIVE OXIDASE: from gene to function. *Annu Rev Plant Physiol Plant Mol Biol*, 48:703–734.

[287] Vesely, M. D. and Vesely, D. L. (1999). Environmental upregulation of the atrial natriuretic peptide gene in the living fossil, Limulus polyphemus. *Biochemical and biophysical research communications*, 254(3):751–6.

[288] Volkov, A. G., Adesina, T., and Jovanov, E. (2008a). Charge induced closing of Dionaea muscipula Ellis trap. *Bioelectrochemistry*, 74(1):16–21.

[289] Volkov, A. G., Adesina, T., Markin, V. S., and Jovanov, E. (2008b). Kinetics and mechanism of Dionaea muscipula trap closing. *Plant physiology*, 146(2):694–702.

[290] Volkov, A. G., Harris, S. L., Vilfranc, C. L., Murphy, V. A., Wooten, J. D., Paulicin, H., Volkova, M. I., and Markin, V. S. (2013). Venus flytrap biomechanics: Forces in the Dionaea muscipula trap. *Journal of Plant Physiology*, 170(1):25–32.

[291] Walker, T. S., Bais, H. P., Grotewold, E., and Vivanco, J. M. (2003). Root exudation and rhizosphere biology. *Plant physiology*, 132(1):44–51.

[292] Wan Zakaria, W.-N.-A., Loke, K.-K., Zulkapli, M.-M., Mohd Salleh, F.-I., Goh, H.-H., and Mohd Noor, N. (2015). RNA-seq Analysis of Nepenthes ampullaria. *Frontiers in plant science*, 6:1229.

[293] Wang, C., Grohme, M. A., Mali, B., Schill, R. O., and Frohme, M. (2014). Towards Decrypting Cryptobiosis-Analyzing Anhydrobiosis in the Tardigrade Milnesium tardigradum Using Transcriptome Sequencing. *PloS one*, 9(3):e92663.

[294] Wang, M., Zheng, Q., Shen, Q., and Guo, S. (2013). The critical role of potassium in plant stress response. *International journal of molecular sciences*, 14(4):7370–90.

[295] Williams, B., Verchot, J., and Dickman, M. B. (2014). When supply does not meet demand-ER stress and plant programmed cell death. *Front Plant Sci*, 5:211.

[296] Williams, M. E. and Mozingo, H. N. (1971). The Fine Structure of the Trigger Hair in Venus's Flytrap. *American Journal of Botany*, 58(6):532.

[297] WILLIAMS, S. E. and BENNETT, A. B. (1982). Leaf Closure in the Venus Flytrap: An Acid Growth Response. *Science*, 218(4577):1120–1122.

[298] Wittkop, T., Emig, D., Lange, S., Rahmann, S., Albrecht, M., Morris, J. H., Bocker, S., Stoye, J., Baumbach, J., Böcker, S., Stoye, J., and Baumbach, J. (2010). Partitioning biological data with transitivity clustering. *Nat Methods*, 7(6):419–420.

[299] Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46.

[300] Worley, A. C. and Harder, L. D. (1999). Consequences of preformation for dynamic resource allocation by a carnivorous herb, Pinguicula vulgaris (Lentibulariaceae). *American Journal of Botany*, 86(8):1136–1145.

[301] Xiong, Y. and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO journal*, 9(10):3353–62.

[302] Yamaguchi, A., Tanaka, S., Yamaguchi, S., Kuwahara, H., Takamura, C., Imajoh-Ohmi, S., Horikawa, D. D., Toyoda, A., Katayama, T., Arakawa, K., Fujiyama, A., Kubo, T., and Kunieda, T. (2012). Two Novel Heat-Soluble Protein Families Abundantly Expressed in an Anhydrobiotic Tardigrade. *PLoS ONE*, 7(8):e44209.

[303] Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978.

[304] Zaag, R., Tamby, J. P., Guichard, C., Tariq, Z., Rigaill, G., Delannoy, E., Renou, J.-P. P., Balzergue, S., Mary-Huard, T., Aubourg, S., Martin-Magniette, M.-L. L., and Brunaud, V. (2015). GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response. *Nucleic Acids Res*, 43(Database issue):D1010–7.

[305] Zheng, Z., Qamar, S. A., Chen, Z., and Mengiste, T. (2006). Arabidopsis WRKY33 transcription factor is required for resistance to necrotrophic fungal pathogens. *The Plant Journal*, 48(4):592–605.

[306] Zhou, X., Jiang, Y., and Yu, D. (2011). WRKY22 transcription factor mediates dark-induced leaf senescence in Arabidopsis. *Molecules and cells*, 31(4):303–13.

[307] Zhurov, V., Navarro, M., Bruinsma, K. A., Arbona, V., Santamaria, M. E., Cazaux, M., Wybouw, N., Osborne, E. J., Ens, C., Rioja, C., Vermeirssen, V., Rubio-Somoza, I., Krishna, P., Diaz, I., Schmid, M., Gomez-Cadenas, A., Van de Peer, Y., Grbic, M., Clark, R. M., Van Leeuwen, T., and Grbic, V. (2014). Reciprocal Responses in the Interaction

between Arabidopsis and the Cell-Content-Feeding Chelicerate Herbivore Spider Mite. *PLANT PHYSIOLOGY*, 164(1):384–399.

[308] Zmasek, C. M. and Godzik, A. (2011). Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biology*, 12(1):R4.

Appendix A

GT34 Protein Phylogenetics

A.1 Introduction

Membrane-bound glycosyltransferases (GTs) catalyze the transfer of glycosyl residues from donor nucleotide sugars to acceptors during biosynthesis of plant cell-wall polysaccharides. Those located in the plasma membrane (PM) form cellulose and callose while Golgi-associated GTs form other cell wall polysaccharides [76, 74, 242]. GTs are grouped into a variety of families based on sequence similarity [51]. Family 34 is a nucleoside diphosphate (NDP) sugar-dependent super-family with a typical GT-A fold present in all glycosyltransferases [35, 177]. Characterized members of this super-family transfer an a-linked monosaccharide in the NDP-sugar donor to the acceptor to forming an a-glycosidic linkage [71]. They are often involved in synthesis of xyloglucans and galactomannans [162], acting as xylosyltransferase or galactosyltransferase. Enzymes encoding the two actives differ by their sequence and form two distinct phylogenetic sub clades [162, 89]. Due to missing genomes sequences little is known about the GT34 family in gymnosperms, a group of seed-producing plants with many extremely large genomes [214, 31, 45]. The following work analyses the GT34 members of *Pinus radiata*, *Pinus taeda*, and (characterized) GT34 members available in the CaZy database using a phylogenetic approach.

A.2 Material and Methods

A.2.1 Pinus taeda EST database mining

Full length CDS of *P. taeda* were identified using an EST mining approach while *P. radiata* CDS were isolated experimentally [5] using assembled *P. taeda* sequences as template. The following sections describe the EST mining approach for *P. taeda* and the phylogenetic analysis. The downloaded *P. taeda* NCBI GenBank EST dataset (328,662 sequences, January 2013) [28] was edited using SeqClean [201]. The resulting high-quality EST dataset was clustered and assembled using GICL [220]. The clustered sequences and the remaining singletons were translated using EMBOSS TranSeq [231]. The predicted peptides were searched for a galactosyltransferase GMA12/MNN10 domain (Pfam ID PF05637) using a customized hidden Markov model. The customized hidden Markov model was constructed using HMMER3 [79] using the original seed alignment of the Pfam profile [225] but retaining only seeds from species in Chloroplastida (Plantae) to ensure an improved specificity for green plants. All identified ESTs were manually re-assembled into the final contigs using Lasergene SeqMan 5.01 (DNASTAR, http://dnastar.com).

A.2.2 Multiple sequence alignments and phylogenetic trees

Full-length *A. thaliana* [270], *Lotus japonicus* [1] and *Trigonella foenum-graecum* [81] proteins containing a PF05637 domain were retrieved, domain annotations were recalculated using the customized hidden Markov model, and domains covering at least 75% of the hidden Markov model profile were excised and aligned using hmmalign [79]. Protest3 was used to calculate the best-fit model for amino acid substitution for the final curated alignment [69]. Using Protest3, the LG model was chosen based on the Akaike information criterion. The phylogenetic reconstruction was performed using PhyML 3.0 [115], protpars from PHYLIP [91] and BIONJ [104]. Support for each clade was analyzed using 100 bootstrap calculations. The final tree was visualized and illustrated using Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

A.3 Results and Discussion

P. radiata is a species with considerable commercial importance and currently the subject of a whole-genome sequencing project (http://dendrome.ucdavis.edu/NealeLab/lpgp). Still genome data is missing. A close relative, *P. taeda* (loblolly pine) was used as a template during the experimental isolation and characterization of *P. radiata* GT34 CDS. Publicly accessible EST databases contained over 300,000 cDNA sequences for *P. taeda*, suggesting close to complete coverage of its transcriptome. The EST dataset was mined using a customized hidden Markov model profile of the conserved GT34 domain (PF05637) and complemented with an extensive BLAST search [11] using all sequences of GT34 members annotated in the CaZy database as template. 119 single *P. taeda* ESTs were identified and assembled into seven contigs of which four comprised complete CDS that translate into the proteins PtGT34A–D, respectively (Figure A.1). A 146bp gap in PtGT34D was closed using the sequence of a single *Pinus contorta* cDNA clone (GenBank accession number GT249984). Remaining 3 fragments were discarded as mis-annotation due to missing homology to other GT34s or seed-plant proteins in general. Additionally, a published whole exome dataset [213] was screened, but no other GT34 genes were found. The phylogenetic analysis of the selected GT34's from *A. thaliana*, *L. japonicus*, *T. foenum-graecum*, *P. taeda* and *P. radiata* resulted in a tree having three sub-clades : 34-1, 34-2 and 34-3 (see figure A.1). *P. taeda* GT34A and D cluster together with *L. japonicus* protein LjGMGT and the *T. foenum-graecum* protein TfGMGT, both having galactomannan (1-6)-α-galactosyltransferase activity, suggesting that these are galactosyltransferases. *P. taeda* GT34B cluster together with *A. thaliana* XXT1 and XXT2 into sub-family 34-2 while GT34B clusters together with XXT3, XXT4 and XXT5 into sub family 34-3. Both sub families show xyloglucan (1-6)-α-xylosyltransferase activity. Experimental analysis of enzyme activities for *P. radiata* showed that PrGT34A and PrGT34C were not enzymatically active [5] which is also likely the case for their *P. taeda* counterparts.

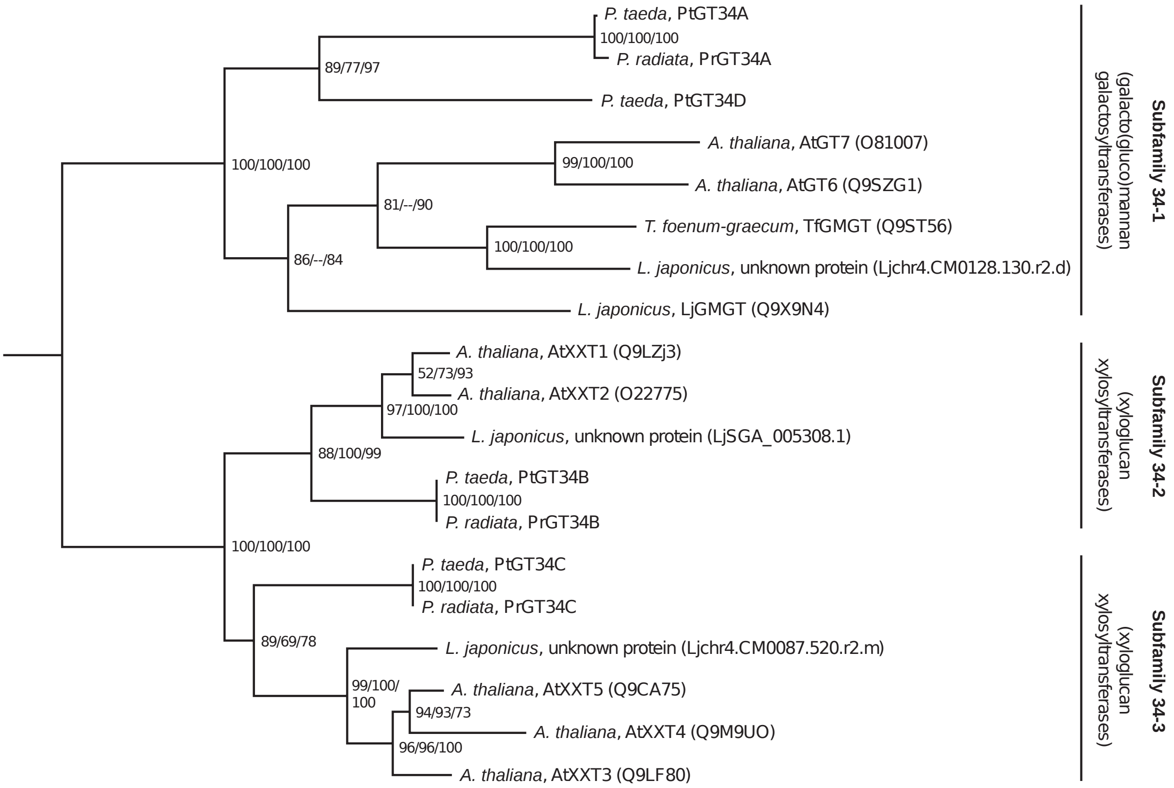

P. taeda, PtGT34A
100/100/100
P. radiata, PrGT34A
89/77/97
P. taeda, PtGT34D
A. thaliana, AtGT7 (O81007)
99/100/100
A. thaliana, AtGT6 (Q9SZG1)
100/100/100
81/--/90
T. foenum-graecum, TfGMGT (Q9ST56)
100/100/100
L. japonicus, unknown protein (Ljchr4.CM0128.130.r2.d)
86/--/84
L. japonicus, LjGMGT (Q9X9N4)
A. thaliana, AtXXT1 (Q9LZj3)
52/73/93
A. thaliana, AtXXT2 (O22775)
97/100/100
L. japonicus, unknown protein (LjSGA_005308.1)
88/100/99
P. taeda, PtGT34B
100/100/100
P. radiata, PrGT34B
100/100/100
P. taeda, PtGT34C
100/100/100
P. radiata, PrGT34C
89/69/78
L. japonicus, unknown protein (Ljchr4.CM0087.520.r2.m)
99/100/100
A. thaliana, AtXXT5 (Q9CA75)
94/93/73
A. thaliana, AtXXT4 (Q9M9UO)
96/96/100
A. thaliana, AtXXT3 (Q9LF80)
Subfamily 34-1
(galacto(gluco)mannan galactosyltransferases)
Subfamily 34-2
(xyloglucan xylosyltransferases)
Subfamily 34-3
(xyloglucan xylosyltransferases)

Fig. A.1 Phylogenetic tree of family GT34 proteins from radiata and loblolly pines (*P. radiata* and *P. taeda*), *A. thaliana*, *T. foenum-graecum* and *L. japonicus*. The phylogenetic reconstruction was performed using PhyML 3.0 [115], protpars from PHYLIP [91] and BIONJ [104], and support for each clade was determined using 100 bootstrap calculations. Node values indicate bootstrap support for tree reconstruction using maximum likelihood/maximum parsimony/pairwise distance methods. UniProt accession numbers are shown for annotated proteins.

A.4 Summary

In conclusion, the study identified four expressed GT34 genes in *P. taeda*. Three of them having a heterologously expressed orthologue in *P. radiata* of which one (PrGT34B) was found to principally have xylosyltransferase activity using UDP-xylose as the donor and cello-oligosaccharides as the acceptor substrate [5]. Remaining proteins may be involved in the synthesis of this polysaccharide as well as heteromannans [132, 120].

A.5 Published elements

Ade, C. P., Bemm, F., Dickson, J. M. J., Walter, C., and Harris, P. J. (2014). Family 34 glycosyltransferase (GT34) genes and proteins in Pinus radiata (radiata pine) and Pinus taeda (loblolly pine). *Plant Journal*, 78(2):305–318

Appendix B

Curriculum vitae

FELIX BEMM

CURRICULUM VITAE

PERSONAL INFORMATION

NAME	**Felix Mathias Bemm**
BIRTHDAY	**27th of June, 1984**
BIRTHPLACE	**Weimar, Deutschland**
FAMILY STATUS	**unmarried**

EDUCATION

PERIOD	**August 2014 — present**
INSTITUTION	Max Planck Institute für Developmental Biology, Tübingen
POSITION	Postdoctoral Researcher, Genome Informatics
GROUP	Functional and Structural Genomics
ADVISOR	Detlef Weigel (Department Molecular Biology)

PERIOD	**March 2011 — July 2014**
INSTITUTION	University Würzburg
GRADE	Dr. rer. nat. (expected July 2017)
THESIS	Genetic foundation of unrivaled survival strategies
ADVISOR	Jörg Schultz (Bioinformatics), Dirk Becker (Botany), Roy Gross (Microbiology)

PERIOD	**October 2005 — February 2011**
INSTITUTION	University Würzburg
GRADE	Dipl. Biol. univ (1.3)
PROFESSION(S)	Bioinformatics, Biotechnology, Cell Biology
THESIS	Searching novel domains in genomic data of *Paramecium tetraurelia*
ADVISOR	Jörg Schultz

PERIOD	**June 2003 — September 2005**
INSTITUTION	Polit. Voluntary Service, Thuringian Parliament, Erfurt

PERIOD	**August 1995 — June 2003**
INSTITUTION	Grammar School, Rudolstadt
GRADE	University-Entrance Diploma (2.2)

Academic Experience

Period	**September 2013 — January 2014**
Institution	Centre for Genomic Regulation (CRG), Barcelona, Spain
Project	Pitcher Plant Genome Project (Group Heinz Himmelbauer)

Period	**February 2009 — April 2009**
Institution	Centre for Integrative Bioinformatics, Vienna, Austria
Project	Deep Metazoan Phylogeny Project, HaMSTR (Group Arndt von Haeseler)

Fellowships

Period	**February 2011 — July 2014**
Institution	Graduate School of Life Science, Würzburg
Funding	Excellence Initiative of the German Federal and State Government

Special Trainings

March 2014	GMOD Workshop, Kuala Lumpur, Malaysia
October 2013	NGS Workshop, CRG, Barcelona, Spain
August 2012	Evolutionary Genomics, Otto-Warburg Summer School, Berlin, Germany

Computational Skills

Programming	Perl, Python, Bash, JavaScript and PHP (both basic)
Databases	PostgreSQL, MySQL, Oracle
Statistics	GNU R, Statistica

Felix Bemm, Würzburg, Friday 31st March, 2017