Making the most of phylogeny: Unique adaptations in tardigrades and 216374 internal transcribed spacer 2 structures

Universität Würzburg

Kumulative Dissertation
zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

Making the most of phylogeny: Unique adaptations in tardigrades and 216374 internal transcribed spacer 2 structures

vorgelegt von
Frank Förster
aus Hohenmölsen

Würzburg, 2010

Eingereicht am: . . . .

Mitglieder der Promotionskommission:
            Vorsitzender: Prof. Dr. Thomas Dandekar
            1. Gutachter: Prof. Dr. Thomas Dandekar
            2. Gutachter: Prof. Dr. Markus Engstler

Tag des Promotionskolloquiums: . . . .
Doktorurkunde ausgehändigt am: . . . .

Acknowledgements

Contents

List of Figures

Acronyms

15-LOX-DICE   15-lipoxygenase differentiation control element

AQP   Aquaporin

BLAST   basic local alignment search tool

CBC   compensatory base change
cDNA   complementary desoxyribonucleic acid
CLANS   cluster analysis of sequences
COG   cluster of orthologous group
COI   Cytochrome c oxidase subunit I
CPU   central processing unit

DBMS   database management system
DDBJ   DNA Data Bank of Japan
DNA   desoxyribonucleic acid

EMBL   European Molecular Biology Laboratory
EST   expressed sequence tag
ETS   external transcribed spacer

FABP   fatty acid binding protein

gDNA   genomic desoxyribonucleic acid
GO   Gene Ontology
GST   Gluthathione S-transferase
GTR   generalised time reversible

HMM   hidden markov model
HPC   high performance computing
HSP   heat shock protein

ITS   internal transcribed spacer

ITS2	internal transcribed spacer 2
LEA	late embryogenesis abundant
LSU	large subunit
mRNA	messenger ribonucleic acid
mtDNA	non-coding regions of animal mitochondrial DNA
NAS	network attached storage
NCBI	National Center for Biotechnology Information
NRdb	non-redundant protein database
NTS	non transcribed spacer
ORF	open reading frame
PCR	polymerase chain reaction
PFAM	Protein Families
rbcL	the large subunit of the chloroplast gene encoding ribulose bisphosphate carboxylase
rDNA	ribosomal desoxyribonucleic acid
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
SCOS	shared candidate orthologous sequence
SMART	Simple Modular Architecture Research Tool
SOD	Superoxide dismutase
SSU	small subunit
TSP	tardigrade-specific protein
UniProt	Universal Protein Resource
UTR	untranslated region

Part I.

Introduction

Thesis outline

The members of the phylum Tardigrada are outstanding animals with the capability to resist environmental stress like low and high temperatures or pressure, the lack of water and radiation. In this thesis I describe bioinformatical approaches to investigate such capabilities and their impact on tardigrade physiology. Further the taxonomic relationship between several tardigrade species are of interest in this thesis which are currently unknown. Thus the second part of the thesis gives attention to phylogenetic methods used to distinguish species and concerning their systematics.

The first part of the thesis is an introduction into the phylum Tardigrada and the internal transcribed spacer 2 (ITS2) as a phylogenetic marker.

The second part describes the methods I used, which cover bioinformatical methods such as sequence alignment, database searches as well as database generation and laboratory work.

The third part, as the main part and results section of this thesis, is composed of the scientific articles as individual chapters we published or submitted to several journals: Chapter 3 (pages 23–34) describes the creation of a web-based workbench for our FUNCRYPTA project. It covers the presentation of the sequencing results and their annotation and additionally offers some tools like motif search or BLAST in our sequence database. The next article (chapter 4 (pages 34–47)) gives an overview about the still ongoing nucleotide sequencing project for *Milnesium tardigradum*. I was involved in the cleaning, assembly and annotation of the sequences and contributed to further steps of data analysis. The third publication (chapter 5 (pages 47–98)) covers the survey and translation of the complete available transcriptome sequence information from different tardigrades. We compare this information to other organisms including *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. The chapter 6 (pages 98–137) shows the first results of the proteomic approach of the FUNCRYPTA-Project for which I prepared the hypothetical proteins to allow the identification of the picked protein spots. The next publication (chapter 7 (pages 137–146)) shows an investigation of several heat shock proteins (HSPs) in *M. tardigradum*. Chapter 8 (pages 146–160) shows several examples and techniques of bioinformatical analyses applied to the available tardigrade data, e.g. the identification of ribonucleic acid (RNA) stability motifs and clustering of protein families. We showed the diversity of DNA-j like proteins which maybe a reason for the adaptation capability of *M. tardigradum*.

The second major topic of this thesis–and the resulting publications—focuses on the ITS2 as a genetic marker to answer phylogenetic questions. Currently about 200,000 sequences of ITS2s are available in common databases like Genbank. I was involved in a complete redesign of the ITS2 database. As a public database, it offers sequence and structure information for more then 160,000 complete ITS2s (chapter 9

(pages 160–167)). In a simulation study, we showed for the first time the benefit of additional structure information for phylogenetic reconstructions with the ITS2 (chapter 10 (pages 167–180)). The next publication (chapter 11 (pages 180–194)) corroborates a monophyletic chlorophycean DO-group (Sphaeropleales). I obtained the sequence information in our laboratory for this study by own experiments. Cryptic species in the phylum Tardigrada are often found. So chapter 12 (pages 194–222) shows how we distinguish between several species of the genus *Paramacrobiotus* and describe three new species based on an analysis of compensatory base changes (CBCs) of the ITS2s

The phylum Tardigrada

The phylum Tardigrada is formed by the tardigrades which are commonly known as 'water bears' and is part of the superphylum Ecdysozoa.

The first description of a tardigrade was published by Goeze (1773). He named the animal 'kleiner Wasserbär' (=little water bear). He depicted a small animal with a head with two eyes and four pairs of clawed legs. (Figure 1).

The name Tardigrada was given by Spallanzani (1776) and means 'slow walker' (Latin: tardis-slow, gradi-walk). Spallanzani (1776) also described the capability to desiccate and to become alive again if the desiccation was slow. This reversible suspension of metabolism of this multi-cellular animals is known as cryptobiosis (Keilin 1959; Ramazzotti and Maucci 1983).

The size of tardigrades ranges from 50–1500 μm. They are globally distributed, however their habitats may be very diverse and characterised by extreme environmental conditions. Usually, they live in marine, freshwater and terrestrial environments, especially in lichens and mosses (Marcus 1928; Marcus and Dahl

Figure 1.: **First drawing of a 'waterbear' by Goeze (1773)** with head and mouth (a), eyes (b) and pairs of legs (c) with claws (d).

1928; Nelson 2002). For example, the cosmopolitan *M. tardigradum* was found in altitudes up to 6000 m in the Himalaya, in central Europe, Java and Canada (Marcus 1928). In addition, *Coronarctus tenellus* was found in the deep sea at a depth of 3700 m (Kinchin 1994, p. 94).

In most cases it is their dehydrated form, also known as tun-stage (Baumann 1922, 1927), that allows them to resist such extreme environments. Some tardigrades can survive low temperatures down to $-273\,°C$ (Becquerel 1950) and high temperatures up to $151\,°C$ (Rahm 1921). They tolerate high (7.5 GPa) (Ono et al. 2008) and low pressure down to the space vacuum (Jønsson et al. 2008) and different types of radiation like X-rays up to 10,000 Gy (May et al. 1964), γ-rays up to 6000 Gy (Horikawa et al. 2006; Jønsson et al. 2005), and Helium ion beam up to 7000 Gy (Horikawa et al. 2006).

From a systematic point of view, the phylum Tardigrada is divided into three classes: Eutardigrada, Heterotardigrada and Mesotardigrada. These classes are subdivided into four orders with 21 families. Recent checklists by Degma and Guidetti (2007); Degma et al. (2010); Guidetti and Bertolani (2005) summarise more than 1,000 described tardigrade species.

The class Mesotardigrada is under discussion. It was established on base of the species *Thermozodium esakii* from a hot spring near Nagasaki (Rahm 1937). No other mesotardigrade has been found and neither type material nor type locality have survived due to an earthquake. The members of the class Heterotardigrada mainly exist in terrestrial and maritime environments. Members of Eutardigrada mainly live in terrestrial and limnic habitats.

The phylogenetic position of the phylum Tardigrada within the animal kingdom is still unclear. Recent publications suggest a position next to the Nematoda (Roeding et al. 2007; Telford et al. 2005). Former publications suggest a close relationship to the Arthopoda (Aguinaldo et al. 1997; Marcus 1928; Ramazzotti and Maucci 1983). Tardigrada as a sister group to Onychophora was discussed by Mallatt and Giribet (2006).

In this thesis I present data of three tardigrade species (*M. tardigradum*, *Hypsibius dujardini* and *Richtersius coronifer*) and the genus *Paramacrobiotus*, described in the following.

Milnesium tardigradum

Milnesium tardigradum is a well-known cosmopolitan species tardigrade with body length between 500 μm and more then 1000 μm (Kinchin 1994). In contrast to most other tardigrades, *M. tardigradum* is carnivorous, feeding on rotifers and nematodes.

The reproduction of *M. tardigradum* seems to be parthenogenetic as reported by Baumann (1964). Also Schuetz (1987) could not observe males during a cultivation

time of one year. Suzuki (2008) reported a low frequency of males in the culture of *M. tardigradum*.

The eggs are colourless or brown and the diameter ranges from 70–110 μm (Kinchin 1994). Suzuki (2003) found that the most rapidly developing animal laid eggs 12 days after hatching.

Hypsibius dujardini

Hypsibius dujardini is an obligatory parthenogenetic species (Ammermann 1967). *H. dujardini* can be cryopreserved and cultured continuously for long time. The genome size of about 78 Mbp (Gregory et al. 2007) is one of the smallest genomes of animals (Gabriel et al. 2007). This is about three-quarters of *C. elegans* and less than half the size of the *D. melanogaster* genome. The generation time is about 14 days at room temperature, the embryos have a stereotyped cleavage pattern with asymmetric cell divisions, nuclear migrations, and cell migrations occurring in reproducible patterns.

The rate of protein evolution in *H. dujardini* is similar to the rate in *D. melanogaster* and, thus similar to most other known genomes in metazoans (Gabriel et al. 2007). Only sparse molecular data are available, but Bavan et al. (2009) described a purinergic receptor expressed in *H. dujardini* and Gabriel and Goldstein (2007) examined the expression of Pax3/7 during embryogenesis.

Richtersius coronifer

Richtersius coronifer, formerly *Adorybiotus* or *Macrobiotus*, is a large tardigrade with body sizes up to 1 mm (Kinchin 1994). *R. coronifer* is a bryophilic (moss-living) tardigrade, which can be found all over the world mainly in alpine and arctic environments up to 5600 m (Ramløv and Westh 2001).

R. coronifer can survive severe dehydration for years (Ramløv and Westh 2001) and is capable of tolerating temperatures down to −196 °C independent from the actual state (tun or normal living) (Ramløv and Westh 1992; Westh and Ramløv 1991). The disaccharide trehalose is accumulated up to approximately 2.3 % dry weight during desiccation (Westh and Ramløv 1991).

Genus *Paramacrobiotus*

Recently the genus *Paramacrobiotus* was erected within the family of Macrobiotidae using morphological characters and gene sequences. Within the genus several cryptic species from different places around the world could be found, which can not be clearly separated by morphological or common molecular markers alone (Guidetti et al. 2009). This cryptic species have been already used for biochemical

and physiological studies and formally described as *Paramacrobiotus 'richtersi* group', numbered consecutively (Hengherr et al. 2009*a,b*; Hengherr et al. 2008).

Internal transcribed spacer 2

Traditionally breeding tests and similarity of morphology were used to analyse genetic relatedness at the genus and species level. Molecular methods have played a minor role so far, but the usage of molecular analyses should shed light on any question of genetic relatedness between two organisms (Coleman and Mai 1997).

A lot of different markers have been used depending on the relatedness of the organisms to be analysed (Figure 2). The large subunit (LSU) and the small subunit (SSU) of the ribosome are highly conserved phylogenetic markers which can be used for higher taxonomic levels (Hershkovitz and Lewis 1996).

The demand for a marker usable for distinguishing species as well as genera or even higher taxonomic levels is satisfied with the ITS2. This spacer is located between the 5.8S and the 28S ribosomal ribonucleic acid (rRNA) which is part of the rRNA cistron Figure 3 (Coleman 2003, 2007). The highly conserved flanking regions of the 5.8S and the 28S rRNA can be used to annotate ITS2 sequences by application of hidden markov models (HMMs) (Keller et al. 2009).

The evolutionary rate of the ITS2 sequence is relatively high, but due to its importance in the ribogenesis the secondary structure of the ITS2 is highly conserved throughout the eukaryotes (Coleman 2007; Mai and Coleman 1997; Schultz et al. 2005).

The structure consists of helices which are necessary recognition and binding sites for the processing of the primary ribosomal desoxyribonucleic acid (rDNA) transcripts. Therefore, the ITS2 may not only be considered as a spacer but further as a non-coding gene (Hoshina et al. 2006).

The common structure core shows four helices (Figure 4). The first and the fourth helix have a faster evolutionary rate, than the other two helices. The second helix shows a characteristic `U-U` mismatch (Coleman 2003). The third helix owns the most conserved sequence and has a `UGGU` motif (Schultz et al. 2005).

Further, the predominance of CBCs is proof for a common secondary structure (Coleman 2003). CBC maintain pairings which are essential for the secondary structure helices by changing the two nucleotides of a pairing to keep this pairing (for example `A-U` to `C-G`). In a hemi- or half-CBC only one base of a pairing (`A-U` to `G-U`) was changed (Coleman 2003). The appearance of a full CBC can be used to distinguish between species with a accuracy of 93 % (Müller et al. 2007).

The existence of a common core within the Eukaryota allows homology modelling to transfer a known structure from a template sequence-structure pair of one ITS2 to another one with unknown structure (Wolf et al. 2005). At the moment more then
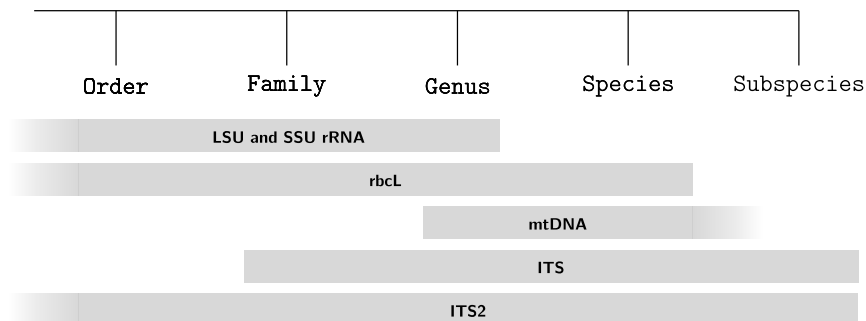
Figure 2.: **Ranges for often used phylogenetic markers.**
The bars indicate the range for which the corresponding DNA sequences are commonly utilised for phylogenetic comparisons. Abbreviations: ITS internal transcribed spacer; mtDNA non-coding regions of animal mitochondrial DNA; rbcL the large subunit of the chloroplast gene encoding ribulose bisphosphate carboxylase. The ITS2 bar represents the range possible by use of ITS2 secondary structure. Redrawn from Coleman (2003).
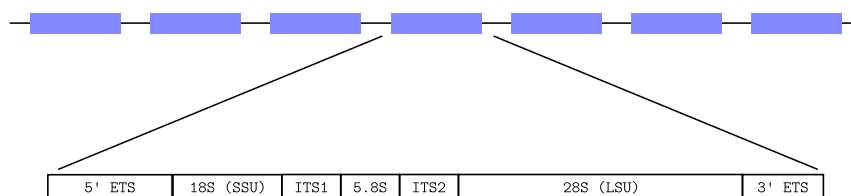
Figure 3.: **Organisation of a typical eukaryotic rRNA cistron.**
The drawing illustrates the organisation of the nuclear ribosomal cistrons (blue boxes) of a typical eukaryote composed of the 5′ ETS, the 18S-rRNA, internal transcribed spacer (ITS) 1, the 5.8S-rRNA, the ITS2, the 28S-rRNA and the 3′ ETS. Between all ribosomal cistrons are NTSs. Redrawn from Coleman (2003).
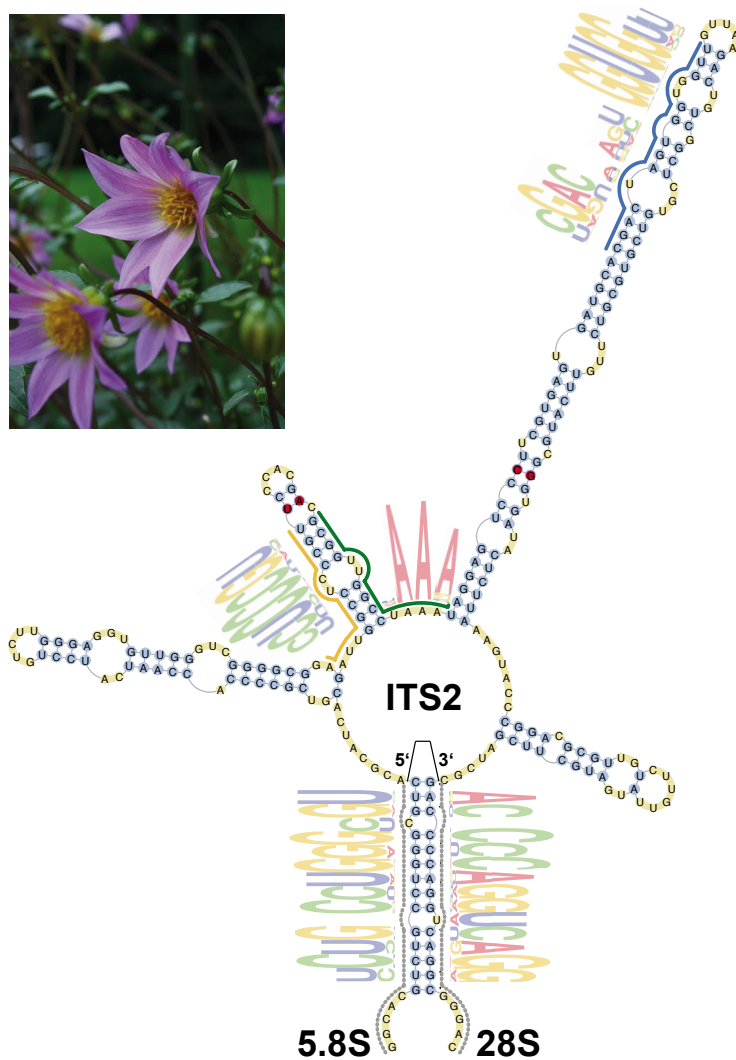
Figure 4.: **Secondary structure of the ITS2 of *Dahlia brevis*.** General ITS2 topology and visualisation of plant HMM motifs for the secondary structure of *D. brevis* (gi:31281745). Annotation from HMMs of 5.8S and 28S are displayed as dotted lines tracing the outline of their position, whereas the ITS2 motif HMMs are represented by coloured lines. In parts of these motifs, nucleotide frequencies are presented (Byun and Han 2009; Gorodkin et al. 1997). Nucleotides are coloured yellow in unpaired regions, whereas paired nucleotides are blue. CBCs between secondary structures of *D. brevis* and *Dahlia iscapigeroides* (gi:31281755) are shown in red. Image taken from Koetschan et al. (2010)

200,000 sequence-structure pairs are present at the ITS2 database (Koetschan et al. 2010; Schultz et al. 2006; Selig et al. 2008) – a web server which offers structures derived from a free energy minimisation algorithm (Mathews et al. 1999) and homology modelled sequences (Wolf et al. 2005). Furthermore, ITS2 sequences can be annotated (Keller et al. 2009) and structures can be predicted. For the database the taxonomic information from National Center for Biotechnology Information (NCBI) is used (Benson et al. 2009; Sayers et al. 2009).

We have recently shown that the addition of the structure information improves the quality of phylogenetic reconstructions (Keller et al. 2010).

This thesis focuses to the amazing phylum of the Tardigrada. These animals are capable to resist extreme environmental conditions, such as temperature, pressure and radiation. Within this work—in framework of the Funcrypta project, we are seeking to answer the central question what reasons are behind the tardigrades' robustness. Our approach is based on the computational analysis of various sequence information.

Another important issue addressed in this thesis is the still debatable interrelatedness of species within the Tardigrada. Therefore, the second part of this thesis concentrates on the usability of the ITS2 as phylogenetic marker to distinguish between cryptic species in the tardigrade genus *Paramacrobiotus*.

Part II.

Material and Methods

Chapter 1.

Material

1.1. Hardware requirements

The main part of the calculations were carried out on a computer with SuSE 10.3 or SuSE 11.0 and sparse Microsoft Windows XP™ as operating system, Intel® Core™ 2 6600 central processing unit (CPU) with 2.40 GHz and 2 GB DDR2-SDRAM memory. Furthermore, I had access to a high performance computing (HPC) cluster system and used it for longer calculations. The HPC cluster consists of 40 nodes each having two dual core Intel® Xeon® 5140 (2.33 GHz) CPUs. Each node offers 8–16 GB of memory and a local 20 GB hard drive. The entire cluster additionally holds 777 GB of network attached storage (NAS). This gives a total of 160 CPU cores, 448 GB memory and more then 1.5 TB storage.

1.2. Databases

These included the major public databases, starting from primary sequence information: Genbank (Benson et al. 2000; Benson et al. 2009) at the NCBI, the EMBL nucleotide sequence database at the European Molecular Biology Laboratory (EMBL) (Baker et al. 2000; Leinonen et al. 2010), DNA Data Bank of Japan (DDBJ) (Kaminuma et al. 2010; Tateno et al. 2000) and Universal Protein Resource (UniProt) (Jain et al. 2009; UniProt Consortium 2010). Furthermore, specific secondary databases such as Simple Modular Architecture Research Tool (SMART) (Letunic et al. 2009; Schultz et al. 1998) or Protein Families (PFAM) (Finn et al. 2010) were used.

Additionally we host some local installed databases using POSTGRES or MYSQL as database management system (DBMS) (e.g. tardigrade workbench (Förster et al. 2009), the BLAST2GO database (Götz et al. 2008) or the ITS2 database (Koetschan et al. 2010)).

Chapter 2.

Methods

2.1. Bioinformatical methods

2.1.1. Sequence alignment

For the calculation of an optimal global alignment between two sequences I used NEEDLE from the EMBOSS-program suite (Rice et al. 2000). It is a implementation of the Needleman-Wunsch algorithm (Needleman and Wunsch 1970).

All NEEDLE versions before 6.2.0 NEEDLE allowed only semi-global alignments, which means that the gaps at the $3'$ and $5'$ ends are not penalised. For a global alignment the program STRETCHER from the EMBOSS-program suite was used. It implements the Myers-Miller algorithm (Myers and Miller 1988) to find the optimal global alignment in an amount of computer memory that is proportional to the size of the smaller sequence $O(N)$ but needs about twice the time to calculate the alignment $O(2N)$.

The program WATER was used to obtain optimal local alignments computed with the Smith-Waterman algorithm (Smith and Waterman 1981). This software is also a part of the EMBOSS-program suite. For scanning large databases I applied basic local alignment search tool (BLAST) (Altschul et al. 1990, 1997)—a heuristic algorithm—instead of using time-consuming optimal local alignment methods. I used the current NCBI-BLAST implementation (version 2.2.20) to search protein and nucleotide databases, e.g. Genbank or UniProt.

Another tool for database searches is the HMMER software (Durbin et al. 1998). The software is based on profile-HMMs (Krogh et al. 1994) and is used for sensitive database searches.

2.1.2. Expressed sequence tags

Pre-processing   The base calling for the expressed sequence tags (ESTs) was performed with PHRED (Ewing et al. 1998; Ewing and Green 1998) using a score threshold of 20. ESTs could contain contaminations from cloning vectors, low complexity regions or low quality bases. I used the software SeqClean for removal

of these contaminations (TIGR 2010). The database containing information of the vector, adaptor and primer sequences for SEQCLEAN was created by myself. I also scanned the sequences against the NCBI UniVec database (Kitts et al. 2009). The minimum sequence length was set to 100 bp. Another problem during assembly could be repetitive elements. These elements were removed with REPEATMASKER (Smit et al. 1996–2004).

Assembly The assembly of the cleaned ESTs was done by the software CAP3 (Huang and Madan 1999) and TGICL (Pertea et al. 2003). Therefore the standard parameters were used. Finally the obtained sequences were checked again for low complexity regions.

Annotation The annotation pipeline is depicted in Figure 5. The cleaned and assembled sequences were used for translation. To determine nucleotide sequences coding for known proteins, we used a BLASTX search against the UniProtKB/ SwissProt-, UniProtKB/TrEMBL- and non-redundant protein database (NRdb) (Figure 5, step 1). In parallel a six frame translation using VIRTUAL RIBOSOME (Wernersson 2006) in version 1.1 was done followed by a HMMER search against the PFAM-database (Figure 5, step 2). For all sequences resulting in a hit either in the BLASTX or the PFAM search the corresponding ORFs were extracted (Figure 5, step 4). Sequences which were not identified via BLASTX were searched against the next more comprehensive database. Sequences with no significant result either using BLASTX or using HMMER against PFAM were translated into six frames and all ORFs consisting of 100 or more amino acids were extracted. If no ORF had a length of 100 or more amino acids, we took the longest ORF (Figure 5, step 3). All sequences which seemed to be rRNA were identified using a BLASTN against a database of eukaryotic rRNAs. These sequences were subtracted from translated sequences (Figure 5, step 5). Finally sequences resulting in a hit in BLASTX and PFAM search (Figure 5, steps 1 and 2) were merged (Figure 5, step 6).

2.1.3. CLANS clustering

The program CLANS performed an all-against-all pairwise sequence comparison using TBLASTX. Afterwards, all similar proteins were clustered in 3D applying the cluster analysis of sequences (CLANS) method (Frickey and Lupas 2004) with a cut-off E-value of 0.001. The program calculates pairwise attraction values based on the HSP P-values obtained from the BLAST-run. Finally, the clusters were identified using convex clustering and NJ-trees with standard parameters.
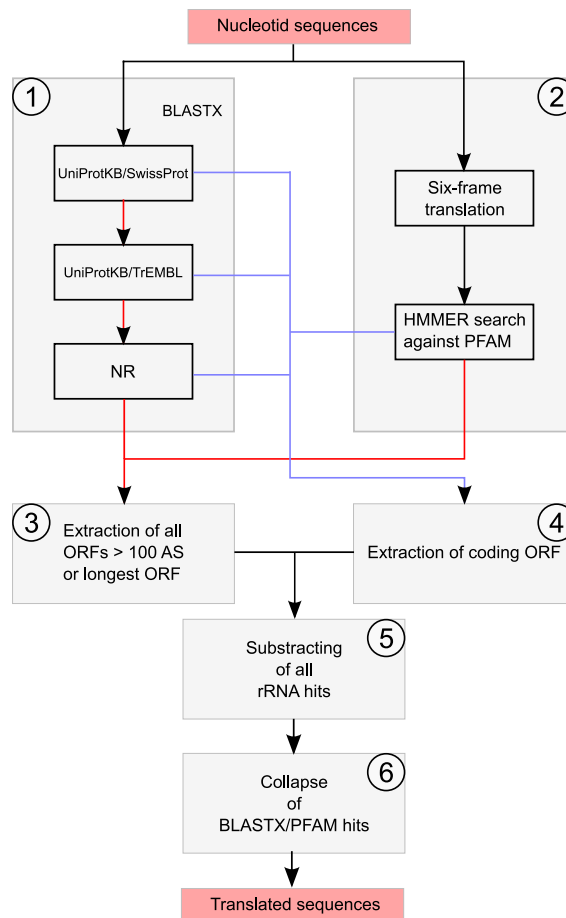
Figure 5.: **Annotation pipeline.** The flowchart depicts the steps of the annotation method. The cleaned EST sequences were scanned for known proteins using BLASTX against UniProtKB/SwissProt-, UniProtKB/TrEMBL- and non-redundant protein database (NRdb) from NCBI (step 1). In parallel they were translated into all six reading frames using Virtual Ribosome followed by a HMMer search against PFAM (step 2). The ORFs of hits against the databases were extracted (step 4). All remaining sequences which had no result in the database search were translated into all ORFs with a minimum length of 100 aa or the longest ORF if no ORF extends the 100 aa. Afterwards, the nucleotide sequences were scanned for rRNA and sequences which gave hits were subtracted from the sequence set (step 5). The last step was the merging of all ORFs (step 6). Image taken from Förster et al. (2009).

2.1.4. InParanoid clustering

After all-against-all pairwise sequence comparisons between the EST sets of *M. tardigradum* and *H. dujardini* using TBLASTX, the results of this comparison were imported into INPARANOID (Remm et al. 2001) for prediction of orthologs. As the EST sets cover a substantial fraction of the encoded proteins but do not yet represent the whole protein set of the compared tardigrades, we called the orthologs given by INPARANOID shared candidate orthologous sequences (SCOSs) and the remaining sequences candidate single sequence ESTs.

2.1.5. Identification of regulatory elements

The ESTs of *H. dujardini* and *M. tardigradum* were systematically screened using the software UTRSCAN (Grillo et al. 2010). This software screens 30 regulatory elements for RNA regulation with a focus on the 3′ untranslated region (UTR) elements and stability of messenger ribonucleic acid (mRNA). The default settings for batch mode were used and all reported elements were collected.

2.1.6. ITS2 work-flow description

The ITS2 work-flow is described by Schultz and Wolf (2009) and consists of the following steps: HMM annotation, secondary-structure prediction, homology modelling, sequence-structure alignments, analysis of CBCs and finally the distinction of species.

HMM annotation of the ITS2   The HMM annotation was performed according to Keller et al. (2009). I used the HMMs for metazoan 5.8S and 28S. The suggested e-value threshold of 0.001 was used.

Secondary-structure prediction   For the secondary-structure prediction I calculated the secondary structure for all tardigrade ITS2s using the software RNASTRUCTURE 3.46 (Mathews et al. 2004). RNASTRUCTURE implements a free energy minimisation algorithm and therefore calculate the secondary structure with the lowest free energy. The structure of *P. 'richtersi* group 2' was chosen as template for the homology modelling.

Homology modelling   The homology modelling (Wolf et al. 2005) was used to transfer the structure of *P. 'richtersi* group 2' to all other tardigrade ITS2s. It is based on semi-global alignments with the software NEEDLE and was completely rewritten by myself during our redesign of the ITS2 database. If not stated otherwise in the

publication a threshold of 75 % was used for the helix transfer. For Schill et al. (2010) we used 66 %.

Sequence-structure alignments   All obtained sequence-structure pairs were multiple aligned with 4SALE (Seibel et al. 2006). The algorithm is based on the translation of sequence-structure pairs into pseudoproteins following the multiple alignment of the pseudoproteins with CLUSTALW (Thompson et al. 1994) and the retranslation into sequence-structure pairs.

Tree reconstruction and CBC-analysis   The CBC analysis was done with 4SALE using the obtained multiple sequence-structure alignment as input. Therefore, all possible sequence-structure pairs were scanned against each other for CBCs. The resulting CBC matrix was used to distinguish between species on appearance of a CBC with an accuracy of 93 % (Müller et al. 2007). Also a CBC tree was reconstructed with CBCANALYZER (Wolf et al. 2005). Additionally, I reconstructed a neighbour-joining tree with PROFDISTS (Wolf et al. 2008).

2.1.7. ITS2 database generation

The ITS2 database was completely redesigned and new features were added to the pipeline process, e.g. HMM annotation. Therefore, it was necessary to create new generation scripts which also allow automated updates at regular intervals.

     The pipeline consists of the following steps: Retrieval of all NCBI sequences which match to a search string which is specific for the ITS2, a rescan of Genbank using the HMM annotation and a merge of all obtained sequences. Afterwards, the sequences which are annotated with a start and end are folded by UNAFOLD (Markham and Zuker 2005, 2008). The following iterative step tries to obtain homology modelled structures. All sequence-structure pair which were generated by the last step or the last iteration are used as input. The loop iterates as long as new sequence-structure pairs can be obtained.

     The last step searches through the remaining sequences using BLASTN, followed by a homology modelling step to get additional partial sequence-structure pairs. All sequence-structure pairs are stored in a POSTGRES database in our department and are publicly available via the web interface of the ITS2 database.

2.1.8. Simulation of ITS2 evolution

We simulated $600,000$ sequence-structure pairs using SISSI v0.98 (Gesell and Haeseler 2006). The secondary structures were included in the simulation process of coevolution by application of two separate substitution models. Therefore it was necessary to estimate substitution models for the non-paired and the paired

nucleotides of the ITS2. For the unpaired nucleotides we used a 4×4 generalised time reversible (GTR) substitution model. For the paired nucleotides a 16×16 GTR was used. Both GTRs were estimated by a manually verified alignment based on 500 individual ITS2 sequences and structures with a variant of the method described by Müller and Vingron (2000).

2.1.9. Reconstruction of LEA and DnaJ protein trees

The proteins for the reconstruction of the trees for late embryogenesis abundant (LEA) and DnaJ proteins was done using the PHYLIP version 3.68 package (Felsenstein 2005). First, the hits for the corresponding PFAM domains were extracted from the protein sequences, next they were multiple aligned with CLUSTALW and finally the tree was calculated using the programs SEQBOOT, PROTDIST and NEIGHBOR from the PHYLIP package.

2.2. Biochemical methods

The extraction of the genomic desoxyribonucleic acid (gDNA) from cultured cells of *Ankyra judayi*, *Atractomorpha porcata* and *Sphaeroplea annulina* was done using Dynabeads® (DNA DIRECT Universal, Dynal Biotech, Oslo, Norway) according to the manufacturer's protocol. Polymerase chain reaction (PCR) reactions were performed in a 50 µL reaction volume containing 25 µL FastStart PCR Master (Roche Applied Science), 5 µL gDNA and 300 nM of each primer ITS3 (5′-GCA TCG ATG AAG AAC GCA GC-3′) and ITS4 (5′-TCC TCC GCT TAT TGA TAT GC-3′) designed by White et al. (1990).

Cycling conditions for amplification consisted of 94 °C for 10 min, 30 cycles of 94 °C for 30 s, 50 °C for 30 s and 72 °C for 45 s, followed by a final extension step of 10 min at 72 °C. PCR products were analysed by 3 % TBE agarose gel electrophoresis and ethidium bromide staining.

PCR probes were purified with the PCR Purificaton Kit (Qiagen) according to the manufacturer's protocol. Afterwards, the probes were quantified by spectrometry. Each sequencing probe was prepared in an 8 µL volume containing 20 ng desoxyribonucleic acid (DNA) and 1.25 µM primer. Sequencing was carried out using an annealing temperature of 50 °C with the sequencer Applied Biosystems QST 3130 Genetic Analyzer by the Institute of Hygiene and Microbiology (Würzburg, Germany).

Part III.

Results

Chapter 3.

Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades

BMC Genomics

Research article

Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades

Frank Förster[†1], Chunguang Liang[†1], Alexander Shkumatov[†2], Daniela Beisser[1], Julia C Engelmann[1], Martina Schnölzer[3], Marcus Frohme[4], Tobias Müller[1], Ralph O Schill[5] and Thomas Dandekar*[1]

Address: [1]Dept of Bioinformatics, Biocenter University of Würzburg, 97074 Würzburg, Germany, [2]EMBL, Hamburg Outstation, Notkestrasse 85, 22603 Hamburg, Germany, [3]Functional Proteome Analysis, German Cancer Research Center, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany, [4]University of Applied Sciences, Bahnhofstraße 1, 15745 Wildau, Germany and [5]Dept of Zoology, Institute for Biology, Universität Stuttgart, 70569 Stuttgart, Germany

Email: Frank Förster - frank.foerster@biozentrum.uni-wuerzburg.de; Chunguang Liang - liang@biozentrum.uni-wuerzburg.de; Alexander Shkumatov - ashkumat@embl-hamburg.de; Daniela Beisser - daniela.beisser@biozentrum.uni-wuerzburg.de; Julia C Engelmann - julia.engelmann@klinik.uni-regensburg.de; Martina Schnölzer - m.schnoelzer@dkfz.de; Marcus Frohme - mfrohme@tfh-wildau.de; Tobias Müller - tobias.mueller@biozentrum.uni-wuerzburg.de; Ralph O Schill - ralph.schill@bio.uni-stuttgart.de; Thomas Dandekar* - dandekar@biozentrum.uni-wuerzburg.de

* Corresponding author   †Equal contributors

Abstract

Background: Tardigrades represent an animal phylum with extraordinary resistance to environmental stress.

Results: To gain insights into their stress-specific adaptation potential, major clusters of related and similar proteins are identified, as well as specific functional clusters delineated comparing all tardigrades and individual species (*Milnesium tardigradum*, *Hypsibius dujardini*, *Echiniscus testudo*, *Tulinus stephaniae*, *Richtersius coronifer*) and functional elements in tardigrade mRNAs are analysed. We find that 39.3% of the total sequences clustered in 58 clusters of more than 20 proteins. Among these are ten tardigrade specific as well as a number of stress-specific protein clusters. Tardigrade-specific functional adaptations include strong protein, DNA- and redox protection, maintenance and protein recycling. Specific regulatory elements regulate tardigrade mRNA stability such as lox P DICE elements whereas 14 other RNA elements of higher eukaryotes are not found. Further features of tardigrade specific adaption are rapidly identified by sequence and/or pattern search on the web-tool tardigrade analyzer http://waterbear.bioapps.biozentrum.uni-wuerzburg.de. The work-bench offers nucleotide pattern analysis for promotor and regulatory element detection (tardigrade specific; nrdb) as well as rapid COG search for function assignments including species-specific repositories of all analysed data.

Conclusion: Different protein clusters and regulatory elements implicated in tardigrade stress adaptations are analysed including unpublished tardigrade sequences.

Background

Tardigrades are small metazoans resembling microscopic bears ("water-bears", 0.05 mm to 1.5 mm in size) and live in marine, freshwater and terrestrial environments, especially in lichens and mosses [1-3]. They are a phylum of multi-cellular animals capable of reversible suspension of their metabolism and entering a state of cryptobiosis [4,5]. A dehydrated tardigrade, known as anhydrobiotic tun-stage [6,7], can survive for years without water. Moreover, the tun is resistant to extreme pressures and temperatures (low/high), as well as radiation and vaccuum [8-13].

Well known species include *Hypsibius dujardini* which is an obligatory parthenogenetic species [14]. The tardigrade *H. dujardini* can be cultured continuously for decades and can be cryopreserved. It has a compact genome, a little smaller than that of *Caenorhabditis elegans* or *Drosophila melanogaster*, and the rate of protein evolution in *H. dujardini* is similar to that of other metazoan taxa [15]. *H. dujardini* has a short generation time, 13-14 days at room temperature. Embryos of *H. dujardini* have a stereotyped cleavage pattern with asymmetric cell divisions, nuclear migrations, and cell migrations occurring in reproducible patterns [15]. Molecular data are sparse but include the purinergic receptor occuring in *H. dujardini* [16].

Milnesium tardigradum is an abundant and ubiquitous terrestrial tardigrade species in Europe and possibly worldwide [17]. It has unique anatomy and motion characteristics compared to other water bears. Most water bears prefer vegetarian food, *M. tardigradum* is more carnivorous, feeding on rotifers and nematodes. The animals are really tough and long-living, one of the reasons why *M. tardigradum* is one of the best-studied species so far.

Questions of general interest are: How related are tardigrade proteins to each other? Which protein families provide tardigrade-specific adaptations? Which regulatory elements influence the mRNA stability? Starting from all published tardigrade sequences as well as 607 unpublished new sequences from *Milnesium tardigradum*, we analyse tardigrade specific clusters of related proteins, functional protein clusters and conserved regulatory elements in mRNA mainly involved in mRNA stability. The different clusters and identified motifs are analysed and discussed, all data are also available as a first anchor to study specific adaptations of tardigrades in more detail (Tardigrade workbench). Furthermore, the tardigrade analyzer, a sequence server to analyse individual tardigrade specific sequences, is made available. It will be regularly updated to include new tardigrade sequences. It has a number of new features for tardigrade analysis not available from standard servers such as the NIH Entrez system [18]: several new species-specific searches (*Echiniscus tes-*

tudo, *Tulinus stephaniae*), additional new sequence information (*M. tardigradum*) and pattern-searches for nucleotide sequences (including pattern search on non-redundant protein database, NRDB). An easy search for clusters of orthologous groups (COG, [19]) different from the COGnitor tool [20] allowing tardigrade specific COG and eukaryotic COG (KOG) searches is also available.

Furthermore, a batch mode allows a rapid analysis of up to 100 sequences simultaneously when uploaded in a file in FASTA format (for tardigrade species or NRDB).

Two fifths of the tardigrade sequences cluster in longer protein families, and we hypothesise for a number of these that they are implicated in the unique stress adaptation potential of tardigrades. We find also ten tardigrade specific clusters. The unique tardigrade adaptions are furthermore indicated by a number of functional COGs and KOGs identified here, showing a particular emphasis on the protection of proteins and DNA. RNA read out is specifically regulated by several motifs for mRNA stability clearly overrepresented in tardigrades.

Results and Discussion

We analysed all publicly available tardigrade sequences (status 9th of April 2009) as well as 607 unpublished *M. tardigradum* sequences from our ongoing transcriptome analysis.

Major tardigrade protein clusters of related sequence-similar proteins

All available tardigrade sequences were clustered by the CLANS algorithm [21]. Interestingly, 39.3 % of the predicted proteins (mainly EST-based predictions) cluster in just 58 major families, each with at least 20 sequences [see additional file 1: Table S1]. These include 4,242 EST sequences from a total of 10,787.

Using these clusters, a number of tardigrade-specific adaptations become apparent (Table 1 [and additional file 1: Table S1]): the clusters include elongation factors (cluster 12), ribosomal RNAs and proteins (cluster 1, 4, 32 and 56) which are part of the transcriptional or translational machinery. Cluster 5 (chitinase binding domain [22]) could provide membrane and structural reorganization or immune protection (e.g. fungi) according to homologous protein sequences characterized in other organisms. Other clusters show protein families related to the tardigrade stress adaptation potential, e.g. ubiquitin-related proteins (cluster 14; maybe stress-induced protein degradation) and cytochrome oxidase-related proteins (cluster 2, suggested to be involved in respiratory chain).

Moreover, proteins responsible for protein degradation (cluster 15) were found as well as proteins regulating

Table 1: CLANS clusters of sequence similar proteins in published tardigrade sequences[1]

Number/color	Cluster description	Sequences/percentage
2	Cytochrome c oxidase like (subunit 1, EC 1.9.3.1)	425 (3.94%)
3	Uncharacterized protein U88/Glycosyltransferase 8 family	302 (2.80%)
5	Proteins containing a Chitin binding domain	191 (1.77%)
6	Proteins containing an IBR/Neuroparsin/DUF1096 domain	189 (1.75%)
7	Fatty-acid binding protein (FABP) family	127 (1.18%)
8	TSP[2] (remote homology to Sericin 1)	126 (1.17%)
9	Proteins containing a RNA polymerase Rpb3/Rpb11 dimerisation domain	92 (0.85%)
10	Metallothionein superfamily (Type 15 family./Thioredoxin like)	84 (0.78%)
12	GTP-binding elongation factor family. EF-Tu/EF-1A sub- family	79 (0.72%)
13	GST superfamily. Sigma family	78 (0.70%)
14	Ubiquitin family	75 (0.69%)
15	Cathepsin family (EC 3.4.22.-)	74 (0.67%)
16	Carboxypeptidase A inhibitor like	72 (0.64%)
17	Trichohyalin/Translation initiation factor like	69 (0.60%)
18	TSP[2]	65 (0.57%)
19	TSP[2]	61 (0.56%)
20	RNA/DNA-binding proteins	60 (0.55%)
	...	
23	Small Heat Shock Protein (HSP20) family	53 (0.47%)
24	Diapause-specific proteins	51 (0.44%)
	...	
38	LEA type 1 family proteins	31 (0.28%)
	...	

[1] Shown are the number of proteins found for the specified cluster, their percentages and the corresponding cluster number in Figure 1. The full Table with all clusters and their color code matching to Figure 1 is given in [additional file 1: Table S1]. Clusters 1, 4 and 11 contain rRNA and are given in [additional file 1].
[2] Tardigrade specific protein cluster

peptidases (cluster 16). Cluster 23 consists of 53 heat shock proteins which are involved in many stress response reactions [23]. Few diapause specific proteins (cluster 24) are known from other animals. Diapause is a reversible state of developmental suspension. It is observed in diverse taxa, from plants to animals, including marsupials and some other mammals [24] as well as insects (associated molecular function varies but involves calcium channel inhibition [25]) and should here support the tun formation or regulate other (e.g. developmental) metabolic inactive states. Furthermore, proteins involved in storage or transportation of fatty acids also seem to be important (cluster 31, [26]). Late embryogenesis abundant (LEA) protein expression seems to be linked to desiccation stress and the acquisition of desiccation tolerance in organisms [27] e.g. nematodes [28,29] and rotifers [30]. Thirty-one LEA type 1 family proteins were found in cluster 38.

LEA proteins are wide-spread among plants and synthesized in response to certain stresses [31,32]. The LEA type 1 family is well known in higher plants (rice, maize, carrots) to be synthesized during late embryogenesis and in ABA stress response. It includes desiccation-related protein PCC3-06 of *Cratersostigma plantagineum*. LEA type 1 family occurs in bacteria (e.g. *Haemophilus influenzae*, *Dei-*

nococcus radiodurans), but is atypical for animals. However, this is an animal example where LEA family type 1 is well represented and forms a full cluster.

Moreover, ten clusters (8, 18, 19, 30, 33, 35, 37, 42, 51, 55) consist of proteins which seem to be specific for tardigrades. These show no significant homology to known proteins.

Functional clusters of stress-specific adaptations present in tardigrades

To gain a systematic overview of involved tardigrade functions, all available tardigrade sequences were classified species-specific according to COG functional category [19,20] as well as according to COG number and molecular function encoded. Note that in this section "protein" implies one type of protein. A COG or KOG comprises often several sequences from different tardigrades. Prokaryotic (COG) and eukaryotic (KOG) gene clusters were compared (Table 2; details on the WEB http://water bear.bioapps.biozentrum.uni-wuerzburg.de/). Again, several tardigrade-specific adaptations stand out, e.g. highly represented COGs regulate translation elongation factor and sulfate adenylate transferase and a strong ubiquitin system. There are many cysteine proteases (21 proteins). For redox protection there are 14 thioredoxin-domain

Table 2: Highly represented protein functions in Tardigrades (COGs and KOGs).

Information from COG clusters[1]:

Information storage and processing
    75 Translation elongation factor EF-1 (COG5256)
    64 GTPases - translation elongation (COG0050)
    58 Peptide chain release factor RF-3 (COG4108)

Cellular processes and signaling
    31 Ubiquitin (COG5272)
    25 Membrane GTPase LepA (COG0481)
    21 Cysteine protease (COG4870)

Metabolism
    75 Heme/copper-type cytochrome/quinol oxidases (COG0843)
    67 GTPases - Sulfate adenylate transferase (COG2895)

Poorly characterized
    11 Dehydrogenases with different specificities (COG1028)
    11 Uncharacterized homolog of Blt101 (COG0401)

Information from KOG clusters[1]:

Information storage and processing
    77 Translation elongation factor EF-1 (KOG0052)
    71 Polypeptide release factor 3 (KOG0459)
    70 Elongation factor 1 alpha (KOG0458)
    53 Mitochondrial translation elongation factor Tu (KOG0460)

Cellular processes and signaling
    52 Glutathione S-transferase (KOG1695)
    46 Cysteine proteinase Cathepsin L (KOG1543)
    34 Apolipoprotein D/Lipocalin (KOG4824)
    31 Cysteine proteinase Cathepsin F (KOG1542)

Metabolism
    78 Cytochrome c oxidase (KOG4769)
    74 Fatty acid-binding protein FABP (KOG4015)

Poorly characterized
    31 Ubiquitin and ubiquitin-like proteins (KOG0001)
    16 GTPase Rab18, small G protein superfamily (KOG0080)
    15 Ras-related GTPase (KOG0394)
    15 GTPase Rab21, small G protein superfamily (KOG0088)

[1] Detailed data and all COG/KOG numbers are given on the WEB page http://waterbear.bioapps.biozentrum.uni-wuerzburg.de/cgi-bin/cog_stat.pl. Summarized here are the functions of those clusters of orthologous groups (COGs) occurring particularly often or suggesting tardigrade specific adaptations.

containing proteins and 75 Heme/copper-type cytochrome/quinol-like proteins as well as ubiquinone oxidoreductase subunits (15 proteins). There are ten proteins involved in seleno-cysteine specific translation [33,34]. In eukaryotes, selenoproteins show a mosaic occurrence, with some organisms, such as vertebrates and algae, but notably also tardigrades, having dozens of these proteins, while other organisms, such as higher plants and fungi, having lost all selenoproteins during evolution [34]. Membrane GTPases (25 proteins) are often of Lep A (leader peptidase [35]) type in tardigrades. In general, members of the GTPase superfamily regulate membrane signaling pathways in all cells. However, LepA, as well as NodO, are prokaryotic-type GTPases very similar to protein synthesis elongation factors but apparently have

membrane-related functions [35]. It is interesting to observe this prokaryotic-type GTPase in tardigrades. We suggest that it will have similar function as known in other organisms and thus ensure protein translation (elongation factor) coupled to membrane integrity and possibly cytoskeletal rearrangement which would again boost the tardigrade resistance to stress.

The KOGs show similar highly represented families and adaptations. Abberant proteins are rapidly recognized by ubiquitination-like proteins (220 proteins) and ubiquitin-ligase related enzymes (71 proteins) as well as proteasome regulatory subunits (85 proteins). For protein protection and refolding disulfide isomerases (26 proteins) and cyclophilin type peptidyl-prolyl cis-trans isomerases (43 proteins; KOG 0879-0885) are available. Connected to redox protection are also thirty AAA+type ATPases and three peroxisome assembly factor 2 containing proteins (KOG0736). This broad effort in protein protection is further supported by molecular chaperones (HSP70, mortalins and other; total of 50 proteins) and chaperonin complex components (32 proteins; KOG0356-0364). There are six superoxide dismutases and six copper chaperons for thioredoxins (37 proteins), glutaredoxin-like proteins (nine) and ten thiodisulfide isomerases as well as 52 glutathione-S-transferases. We found 22 hits to helicases. Tardigrade DNA protection is represented by 52 proteins of the molecular chaperone DNA J family: proteins of the DNA J family are classified into 3 types according to their structural domain decomposition. Type I J proteins compose of the J domain, a gly-rich region connecting the J domain and a zinc finger domain, and possibly a C-terminal domain. Type II lacks the Zn-finger domain and type III only contains the J domain [36,37]. The latter two are referred to as DnaJ-like proteins. Analysis of the domains present in tardigrade proteins by SMART [38] and Pfam [39] searches reveals only the J domain and in some cases a transmembrane region, identifying them as type III DnaJ-like proteins. For further information on these COGs/KOGs see Table 3.

Moreover, undesired proteins can be rapidly degraded by cathepsin F-like proteins (31 proteins) or L-like proteins (46 proteins). There are several calcium-dependent protein kinases (25 proteins; KOG0032-0034) and actin-bundling proteins. According to this observation calcium signaling should be implicated in adaptive rearrangement of the cytoskeleton during tardigrade rehydration. The cytoskeleton is a key element in the organisation of eukaryotic cells. It has been described in the literature that the properties of actin are modulated by small heat-shock proteins including a direct actin-small heat-shock protein interaction to inhibit actin polymerization to protect the cytoskeleton [40,41] (compare with the CLANS cluster 24 (Diapause proteins) found in the above analysis).

Table 3: Identified DnaJ-family COGs/KOGs in Tardigrades and *Milnesium tardigradum*[1].

KOG/COG number	COG distribution	COG name	present in	
			Tardigrades	*M. tardigradum*[*]
COG0484		DnaJ-class molecular chaperone with C-terminal Zn finger domain	5	
COG2214		DnaJ-class molecular chaperone	8	
KOG0550	A-DH-P-	Molecular chaperone (DnaJ superfamily)	3	2
KOG0691	ACDHYP-	Molecular chaperone (DnaJ su perfamily)	7	2
KOG0712	ACDHYPE	Molecular chaperone (DnaJ su perfamily)	8	2
KOG0713	ACDH---	Molecular chaperone (DnaJ su perfamily)	5	1

[1]Shown are the number of proteins found for the specified COG/KOG number, the KOG distribution of the KOG in different eukaryotic species (see abbreviations) and the COG/KOG annotation.
Abbreviations: A *Arabidopsis thaliana*, C *Caenorhabditis elegans*, D *Drosophila melanogaster*, H *Homo sapiens*, Y *Saccharomyces cerevisiae*, P *Schizosaccharomyces pombe*, E *Encephalitozoon cuniculi*. *These include specific unpublished data from ongoing work on *M. tardigradum*

Translation in tardigrades includes polypeptide release factors (71 proteins) and proteins for translation elongation (77 proteins). There are about 80 GTP-binding ADP-ribosylation factors. The secretion system and Rab/Ras GTPases are fully represented (183 proteins). Seventeen tubulin anchor proteins show that the cytoskeleton is well maintained. Finally, we find 14 TNF-associated factors and 34 apolipoprotein D/lipocalin proteins.

Typical motifs in tardigrade mRNAs

The regulatory motif search showed a number of known regulatory RNA elements involved in tardigrade mRNA regulation (Table 4 for *H. dujardini* and *M. tardigradum*). Certainly it can not be formally ruled out that some of these elements work in a tardigrade modified way. Similarly, there are probably further patterns which are tardigrade specific, but not detected with the UTRscan software [42] applied for analysis.

Table 4: Regulatory elements in *Hypsibius dujardini*[1] and *Milnesium tardigradum*[2] mRNA sequences.

Motiv	*Hypsibius dujardini*	*Milnesium tardigradum*
15-LOX-DICE	1528 (1269)[3]	46 (45)[3]
ADH DRE	60 (58)[3]	1 (1)[3]
BRE	1(1)[3]	--
Brd-Box	152 (149)[3]	28 (22)[3]
CPE	21 (21)[3]	15 (15)[3]
Elastin G3A	1 (1)[3]	--
GLUT1	1 (1)[3]	--
GY-Box	406 (372)[3]	21 (21)[3]
IRE	1 (1)[3]	--
IRES	1353 (1353)[3]	90 (90)[3]
K-Box	469 (447)[3]	35 (33)[3]
SECIS-1	1 (1)[3]	--
SECIS-2	6 (6)[3]	--
TGE	5 (5)[3]	1 (1)[3]
TOP	50 (50)[3]	1 (1)[3]

[1] We considered 5,378 ESTs in *H. dujardini*. [2] We considered 607 ESTs in *M. tardigradum*. [3] The number of hits is followed by the number of mRNAs with this hit in brackets to indicate multiple hits.

The RNA elements found include the lox-P DICE element [43] in *H. dujardini* as top hit with as many as 1,269 ESTs (23.6% of all *H. dujardini* EST sequences). The cytidine-rich 15-lipoxygenase differentiation control element (15-LOX DICE, [44]) binds KH domain proteins of the type hnRNP E and K (stronger in multiple copies), mediating mRNA stabilization and translational control [43].

Furthermore, a high number of mRNAs contains K-Boxes (cUGUGAUa, [45]) and brd Boxes (AGCUUUA, [46]). All these elements are involved in mRNA storage and mRNA stability. These two elements are potential targets for miR-NAs as shown in *Drosophila melanogaster* [47].

However, in the two tardigrade species compared, only 16 of 30 well known RNA elements are found, suggesting a clear bias in tardigrade mRNA regulation. For example, the widely used AU rich elements in higher organisms [42] such as vertebrates are absent in tardigrades [see additional file 1].

Regulatory elements in tardigrade mRNA are probably important for their adaptation, in particular to support transformation to tun stage and back to active stage again. The list of RNA elements found can be compared for instance to our data on regulatory elements in human anucleate platelets [48] where mRNAs have to be stockpiled for the whole life of the platelet. Due to this comparatively long life, a long mRNA untranslated region is important in these cells. The same should apply to tardigrade mRNAs, since their average UTR is predicted to be long. A different stock-piling scenario occurs in unfertilized eggs, but due to developmental constraints, here localization signals are often in addition important for developmental gradients. We tested for these in tardigrades but did not find a high representation of localization motifs.

Web-tool tardigrade analyzer

We created a convenient platform to allow rapid sequence comparisons of new protein sequences, in particular from new sequencing efforts in tardigrades, to our database by applying rapid heuristic local alignment using BLAST [49] and allowing to search in selected species.

A batch mode allows the analysis of up to 100 sequences simultaneously when uploaded in a file in FASTA format. Output data are displayed according to an enhanced BLAST output format with graphical illustrations. Low expected E-values result for searches using the option of our tardigrade specific databases: a more specific smaller database reduces the probability of false positives. As an alternative for general sequence analysis, a search against the non-redundant database of GenBank can be performed. This takes more computational power and yields higher E-values, however, it identifies functions for most sequences. An additional useful feature is to scan all available data for peptide motifs or PROSITE signatures using a "pattern" module [additional file 1: Fig. S1] or assign potential functions by COGs [19]. The first is helpful to recognize tardigrade proteins in cases where the tardigrade sequence has diverged far, and only critical residues for function are still conserved as motif signatures. It can also be applied to search for regulatory RNA motifs such as polyadenylation sites (e.g. AAUAAA or AAUUAA) or recognize promotor modules such as the glucocorticoid receptor element (GRE; palindromic pattern: AGAA CAnnnTGTTCT). For this purpose, both, the tardigrade sequences and the non redundant database can be searched (e.g. to look for stress-specific regulatory RNA elements; [additional file 1, Fig. S2]).

Interestingly, this nucleotide (RNA or DNA) specific option is not available on some common servers, e.g. the PHI-BLAST [50] server at NIH. Further options include a user-defined database [additional file 1: Fig. S3] and interactively animated stress clusters (Figure 1).

The tool http://waterbear.bioapps.biozentrum.uni-wuerzburg.de/ allows rapid searches for tardigrade specific sequences, e.g. molecular adaptations against stress [see additional file 1 for screenshots and a tutorial]. For instance, a search for trehalase sequences shows no trehalase mRNA in the *H. dujardini* sequences. In contrast, there are several heat shock proteins in tardigrades, an example is HSP90 proteins (identified by sequence similarity as well as by a pattern hit based approach using the PROSITE entry PS00298 with the signature Y-x- [NQHD]- [KHR]- [DE]- [IVA]-F- [LM]-R- [ED]; Table 5). Specific COGs are also rapidly assigned for any desired sequence. This includes the option to map the query sequence of interest to any of the known tardigrade specific COGs. Furthermore, nucleotide patterns such as
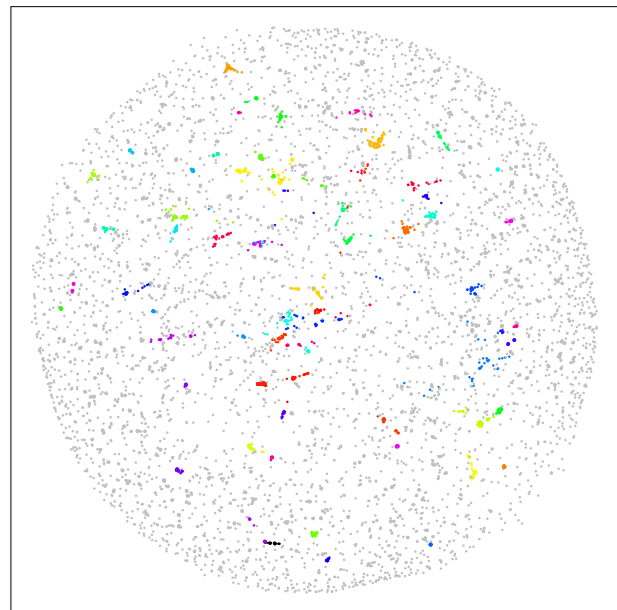
Figure 1
Functional clusters by CLANS of sequence related proteins in tardigrades. All available [see additional file 1: Figure S5] tardigrade protein sequences were clustered in a 3D sphere according to their sequence distance and were projected to the paper plane. Individual protein functions are colored [for color code see additional file 1: Table S1] and all listed in Table 1. Functional clusters appear as patches of an individual color. Color code and clusters can be interactively examined at the Tardigrade workbench http://waterbear.bio-apps.biozentrum.uni-wuerzburg.de and are given in [additional file 1 Table S1]. figure1.pdf

mRNA polyadenylation sites are rapidly identified e.g. in *H. dujardini* mRNAs [additional file 1: Fig. S4]. Similarly, other mRNA 3'UTR elements can be identified, e.g. AU rich sequences mediating mRNA instability or regulatory K-boxes (motif cUGUGAUa, [45]) in tardigrades.

Implications

Tardigrades show a surprising large amount of related sequences. Certainly, one has to correct for a few genes sequenced from many lineages for phylogenetic studies in tardigrades (cytochrome c, rRNA etc.) However, despite this, a number of tardigrade-specific clusters still remain. Furthermore, Table 1 shows that most of the annotated clusters are stress-related.

Looking at specific protein functions, both COG and KOG proteins show that tardigrades spend an extraordinary effort in protein protection, turnover and recycling as well as redox protection. Some other specific adaptations become apparent also from Table 2, but the complete extent of these adaptations is unclear given the limited

Table 5: HSP90 proteins identified in *Hypsibius dujardini* using the Tardigrade analyzer[1].

Hit	Predicted function/name (Tardigrade analyzer)	Pattern matched	Start	End position
gi:37213462	hsp90[2]	YSNKEIFLRE	68	77
gi:37213713	hsp90[2]	YSNKEIFLRE	70	79

[1]These are hits using the pattern hit option and the heat shock protein PROSITE entry PS00298 for pattern generation and recognition. The pattern has the signature of `Y-x- [NQHD]- [KHR]- [DE]- [IVA]-F- [LM]-R- [ED]`. [2]Predicted similarity to Q7PT10 (HSP83 ANOGA) from Swissprot

sampling of available tardigrade sequences. Furthermore, protection of DNA is critical as it has been shown that tardigrade tuns accumulate DNA damage which first has to be repaired before resurrection occurs [51,52]. Taking this into consideration, DNA J proteins were investigated in more detail since proteins of this family are well represented in tardigrades, including several COGs and KOGs. Several data underline the extremely high resistance of tardigrades to temperature, pressure and radiation as well as a high repair potential regarding DNA [11,51]. Thus, we suggest that the high repair potential is also mediated by this well represented protein family. Phylogenetic analysis (Table 3) shows that these proteins are represented by several KOGs as well as the classic COGs in tardigrades. In particular, the first three KOG families are also used in *M. tardigradum*, where extreme stress tolerance requires strong repair mechanisms [17]. Furthermore, all these tardigrade proteins in Table 3 are small, having neither zinc-finger domains nor low complexity regions, but instead consisting of single DNA J domains which would always place them in type I (subfamily A) of DNA-J like proteins. This suggests that the direct interaction with DNA-J like proteins is the key molecular function.

Finally, we could show that there are 16 regulatory elements used in tardigrade mRNA, while a number of other elements known from higher eukaryotic organisms and vertebrates is not used. It is interesting to note that the elements often used in tardigrades are all involved in regulation of mRNA stability. Thus, they may be implicated in stage switching, as presumably in the initial phases of the tun awakening or tun formation, new supply of mRNA is turned off and instead regulation of synthesized mRNA becomes important.

In addition, and for further research we supply the web tool tardigrade analyzer. There are a number of alternative tools available, e.g. from NCBI http://www.ncbi.nlm.nih.gov/. However, we offer some species-specific searches not available from these sources as well as RNA and promotor pattern search (not only for tardigrades but also for NRDB; not available from NIH). Furthermore, there are functional COG prediction as well as new, unpublished tardigrade sequences from *M. tardigradum*,

all above reported data including the reported sequences and detailed functional clusterings as well as regular server updates. A better understanding of the survival mechanisms in these organisms will lead to the development of new methods in several areas of biotechnology. For example, preservation of biological materials *in situ*, macromolecules and cells from non-adapted organisms [53]. This is, of course, only a first and very general overview on potential tardigrade specific adaptations, more species-specific data will be considered as more information becomes available.

Conclusion

Tardigrade genomes invest in stress-specific adaptations, this includes major sequence related protein clusters, functional clusters for stress as well as specific regulatory elements in mRNA. For further tardigrade genome analysis we offer the tardigrade workbench as a flexible tool for rapid and efficient analysis of sequence similarity, protein function and clusters, COG membership and regulatory elements.

Methods
Tardigrade-sequences

The cosmopolitan eutardigrade species *M. tardigradum* Doyére 1849 (Apochela, Milnesidae) was cultured. Tardigrades were kept and reared on petri dishes (diameter: 9.4 cm) filled with a small layer of agarose (3 %) (peqGOLD Universal Agarose, peqLAB, Erlangen, Germany) and covered with spring water (Volvic™ water, Danone Waters Deutschland, Wiesbaden, Germany) at 20 ± 2°C and a light/dark cycle of 12 h. Rotifers *Philodina citrina* and nematodes *Panagrellus* sp. were provided as food source, juvenile tardigrades were also fed with green algae *Chlorogonium elongatum*. For all experiments adult animals in good physical condition were taken directly from the culture and starved for three days to avoid preparation of additional RNA originating from not completely digested food in the intestinal system. For an overview of RNAs present both in active and tun stage we used a mixture of the same number of animals.

Total RNA extraction was performed using the QIAGEN RNeasy®Mini kit (Qiagen, Hilden, Germany). The cDNA

synthesis was reversed transcribed using 1 μg total RNA by the Creator™ SMART™ cDNA Library Construction Kit (Clontech-Takara Bio Europe, France). The resulting cDNA was amplified following the manufacturers protocol and cloned into pDNR-Lib cloning vector. The resulting plasmids were used to transform *Escherichia coli* by electroporation. Sequencing of the cDNA-library was done by ABI 3730XL capillary sequencer (GATC Biotech AG, Konstanz, Germany). All obtained EST sequences were deposited with Genbank including dbEST databank.

Nucleotide sequences from other tardigrades were collected from Genbank. For *H. dujardini*, the best represented species, we composed 5,235 ESTs. We stored *H. dujardini* as well as all published sequences of other tardigrade species (e.g. *T. stephaniae*, *E. testudo*, *M. tardigradum*, *R. coronifer*) in a database (10,787 sequences including translated sequences, details in [additional file 1], status on April, 2009).

CLANS clustering
For a systematic overview on tardigrade specific adaptations we first clustered all published tardigrade nucleotide sequences into functional clusters (Figure 1) using the Cluster analysis of sequences (CLANS) algorithm [21]. All sequences were clustered in 3D space using 0.001 as an E-value cut-off for TBLASTX all-against-all searches. [additional file 1: Fig. S4].

Identification of regulatory elements
For this the ESTs of *H. dujardini* and *M. tardigradum* were systematically screened using the software UTRscan [42]. This software screens 30 regulatory elements for RNA regulation with a focus on 3' UTR elements and stability of mRNA. The default settings for batch mode were used and all reported elements were collected.

COG clustering and identification
In order to acquire a systematic overview of the functionalities, we used the latest version of COG/KOG databases ftp://ftp.ncbi.nih.gov/pub/COG and the BLAST hits from both nucleotide search and protein search were clustered according to their COG ID. Searches were carried out in parallel on all the tardigrade species including *M. tardigradum*, *H. dujardini*, *E. testudo*, *T. stephaniae* and *R. coronifer*. The results are summarized in a table shown in the tardigrade analyzer, the background color from cold to warm (blue to red) indicates the cluster size, which enables an easy comparison. Moreover, users are allowed to click the COG ID and the hit number. The server then reports the corresponding sequence ID, description, conservation and the homologous entries recorded in the database. The server with its data is automatically updated bi-monthly according to the latest tardigrade databases.

Tardigrade workbench
The tardigrade workbench is implemented in Perl using the Bioperl modules [54]. NCBI BLAST program of 2.2.17 is involved in the software package. A database of Post-gresql 8.1.9 is applied to manage the tardigrade entries so as to accelerate the searching queried by investigators. The COG cluster information is automatically updated each week and warehoused on the server. In addition, the run of tardigrade workbench requires an Apache server, a linux system of at least 2 GB memory is highly recommended.

Authors' contributions
FF did tardigrade protein data analysis including CLANS clustering and RNA motif analysis. CL established the current version of the tardigrade workbench including programming new routines, data management and nucleotide motif analysis. AS did the initial setup of the server, of the virtual ribosome and the CLANS clustering. DB, JE, MS and MF participated in tardigrade data analysis. TM gave expert advice and input on statistics, RS gave expert advice on tardigrade physiology and zoology. TD led and guided the study including analysis of data and program, supervision, and manuscript writing. All authors participated in the writing of the manuscript and approved the final version.

Additional material

Additional file 1
Additional Tables and Figures. The file contains seven additional figures and two additional tables. One of these tables summarizes annotation and different identifiers for 607 new EST sequences from Milne-sium tardigradum.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-469-S1.PDF]

Acknowledgements

References
1. Marcus E, Dahl F: *Spinnentiere oder Arachnoidea IV. Bärtierchen (Tardi-grada)* Urban & Fischer Bei Elsevier; 1928.
2. Marcus E: **Zur Ökologie und Physiologie der Tardigraden.** *Zool Jahrb Abt Phys* 1928, **44:**323-370.
3. Nelson DR: **Current Status of the Tardigrada: Evolution and Ecology.** *Integr Comp Biol* 2002, **42:**652-659.
4. Keilin D: **The Leeuwenhoek Lecture: The problem of anabio-sis or latent life: History and current concept.** *Proc R Soc Lond B Biol Sci* 1959, **150:**149-191.
5. Ramazzotti G, Maucci W: **The Phylum Tardigrada.** *Memorie dell'Istituto Italiano di Idrobiologia, Pallanza* 1983, **41:**309-314.

6. Baumann H: **Die Anabiose der Tardigraden.** *Zool Jahrb* 1922, **45**:501-556.
7. Baumann H: **Bemerkungen zur Anabiose von Tardigraden.** *Zool Anz* 1927, **72**:175-179.
8. Horikawa DD, Sakashita T, Katagiri C, Watanabe M, Kikawada T, Nakahara Y, Hamada N, Wada S, Funayama T, Higashi S, Kobayashi Y, Okuda T, Kuwabara M: **Radiation tolerance in the tardigrade Milnesium tardigradum.** *Int J Radiat Biol* 2006, **82**:843-848.
9. Hengherr S, Worland MR, Reuner A, Brümmer F, Schill RO: **Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades.** *J Exp Biol* 2009, **212**:802-807.
10. Hengherr S, Worland MR, Reuner A, Brümmer F, Schill RO: **High-Temperature Tolerance in Anhydrobiotic Tardigrades Is Limited by Glass Transition.** *Physiol Biochem Zool* 2009, **82(6)**:749-755.
11. Jönsson KI, Rabbow E, Schill RO, Harms-Ringdahl M, Rettberg P: **Tardigrades survive exposure to space in low Earth orbit.** *Curr Biol* 2008, **18**:R729-R731.
12. Jönsson KI, Schill RO: **Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade Richtersius coronifer.** *Comp Biochem Physiol B Biochem Mol Biol* 2007, **146**:456-460.
13. Wright JC: **Cryptobiosis 300 Years on from van Leuwenhoek: What Have We Learned about Tardigrades?** *Zoologischer Anzeiger - A Journal of Comparative Zoology* 2001, **240**:563-582.
14. Ammermann D: **The cytology of parthenogenesis in the tardigrade Hypsibius dujardini.** *Chromosoma* 1967, **23(2)**:203-213.
15. Gabriel WN, McNuff R, Patel SK, Gregory TR, Jeck WR, Jones CD, Goldstein B: **The tardigrade Hypsibius dujardini, a new model for studying the evolution of development.** *Dev Biol* 2007, **312**:545-559.
16. Bavan S, Straub VA, Blaxter ML, Ennion SJ: **A P2X receptor from the tardigrade species Hypsibius dujardini with fast kinetics and sensitivity to zinc and copper.** *BMC Evol Biol* 2009, **9**:17.
17. Kinchin I, Dennis R: *The biology of tardigrades* Portland Press London; 1994.
18. Baxevanis AD: **Searching the NCBI databases using Entrez.** *Curr Protoc Hum Genet* 2006, **Chapter 6:**. Unit 6.10
19. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28**:33-36.
20. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4**:41.
21. Frickey T, Lupas A: **CLANS: a Java application for visualizing protein families based on pairwise similarity.** *Bioinformatics* 2004, **20**:3702-3704.
22. Tjoelker LW, Gosting L, Frey S, Hunter CL, Trong HL, Steiner B, Brammer H, Gray PW: **Structural and functional definition of the human chitinase chitin-binding domain.** *J Biol Chem* 2000, **275**:514-520.
23. Qian SB, McDonough H, Boellmann F, Cyr DM, Patterson C: **CHIP-mediated stress recovery by sequential ubiquitination of substrates and Hsp70.** *Nature* 2006, **440**:551-555.
24. Chen WH, Ge X, Wang W, Yu J, Hu S: **A gene catalogue for post-diapause development of an anhydrobiotic arthropod Artemia franciscana.** *BMC Genomics* 2009, **10**:52.
25. Kim YJ, Nachman RJ, Aimanova K, Gill S, Adams ME: **The pheromone biosynthesis activating neuropeptide (PBAN) receptor of Heliothis virescens: identification, functional expression, and structure-activity relationships of ligand analogs.** *Peptides* 2008, **29**:268-275.
26. Alvarez-Ordóñez A, Fernández A, López M, Bernardo A: **Relationship between membrane fatty acid composition and heat resistance of acid and cold stressed Salmonella senftenberg CECT 4384.** *Food Microbiol* 2009, **26**:347-353.
27. Tunnacliffe A, Wise MJ: **The continuing conundrum of the LEA proteins.** *Naturwissenschaften* 2007, **94**:791-812.
28. Browne JA, Dolan KM, Tyson T, Goyal K, Tunnacliffe A, Burnell AM: **Dehydration-specific induction of hydrophilic protein genes in the anhydrobiotic nematode Aphelenchus avenae.** *Eukaryot Cell* 2004, **3**:966-975.
29. Goyal K, Tisi L, Basran A, Browne J, Burnell A, Zurdo J, Tunnacliffe A: **Transition from natively unfolded to folded state induced by desiccation in an anhydrobiotic nematode protein.** *J Biol Chem* 2003, **278**:12977-12984.
30. Tunnacliffe A, Lapinski J, McGee B: **A putative LEA protein, but no trehalose, is present in anhydrobiotic bdelloid rotifers.** *Hydrobiologia* 2005, **546**:315-321.
31. Kobayashi Y, Maeta E, Terashima A, Takumi S: **Positive role of a wheat HvABI5 ortholog in abiotic stress response of seedlings.** *Physiol Plant* 2008, **134**:74-86.
32. Hong-Bo S, Zong-Suo L, Ming-An S: **LEA proteins in higher plants: structure, function, gene expression and regulation.** *Colloids Surf B Biointerfaces* 2005, **45**:131-135.
33. Fagegaltier D, Lescure A, Walczak R, Carbon P, Krol A: **Structural analysis of new local features in SECIS RNA hairpins.** *Nucleic Acids Res* 2000, **28**:2679-2689.
34. Lobanov AV, Hatfield DL, Gladyshev VN: **Eukaryotic selenoproteins and selenoproteomes.** *Biochim Biophys Acta* 2009 in press.
35. March PE: **Membrane-associated GTPases in bacteria.** *Mol Microbiol* 1992, **6**:1253-1257.
36. Walsh P, Bursać D, Law YC, Cyr D, Lithgow T: **The J-protein family: modulating protein assembly, disassembly and translocation.** *EMBO Rep* 2004, **5**:567-571.
37. Cheetham ME, Caplan AJ: **Structure, function and evolution of DnaJ: conservation and adaptation of chaperone function.** *Cell Stress Chaperones* 1998, **3**:28-36.
38. Letunic I, Doerks T, Bork P: **SMART 6: recent updates and new developments.** *Nucleic Acids Res* 2009, **37**:D229-D232.
39. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36**:D281-D288.
40. Mounier N, Arrigo AP: **Actin cytoskeleton and small heat shock proteins: how do they interact?** *Cell Stress Chaperones* 2002, **7**:167-176.
41. Sun Y, MacRae TH: **Small heat shock proteins: molecular structure and chaperone function.** *Cell Mol Life Sci* 2005, **62**:2460-2476.
42. Pesole G, Liuni S: **Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs.** *Trends Genet* 1999, **15**:378.
43. Ostareck-Lederer A, Ostareck DH, Hentze MW: **Cytoplasmic regulatory functions of the KH-domain proteins hnRNPs K and E1/E2.** *Trends Biochem Sci* 1998, **23**:409-411.
44. Ostareck-Lederer A, Ostareck DH, Standart N, Thiele BJ: **Translation of 15-lipoxygenase mRNA is inhibited by a protein that binds to a repeated sequence in the 3' untranslated region.** *EMBO J* 1994, **13**:1476-1481.
45. Lai EC, Burks C, Posakony JW: **The K box, a conserved 3' UTR sequence motif, negatively regulates accumulation of enhancer of split complex transcripts.** *Development* 1998, **125**:4077-4088.
46. Lai E: **Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation.** *Nat Genet* 2002, **30**:363-364.
47. Lai EC, Tam B, Rubin GM: **Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs.** *Genes Dev* 2005, **19**:1067-1080.
48. Dittrich M, Birschmann I, Pfrang J, Herterich S, Smolenski A, Walter U, Dandekar T: **Analysis of SAGE data in human platelets: features of the transcriptome in an anucleate cell.** *Thromb Haemost* 2006, **95**:643-651.
49. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
50. Zhang Z, Schäffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV, Altschul SF: **Protein sequence similarity searches using patterns as seeds.** *Nucleic Acids Res* 1998, **26**:3986-3990.
51. Neumann S, Reuner A, Brümmer F, Schill RO: **DNA damage in storage cells of anhydrobiotic tardigrades.** *Comp Biochem Physiol A Mol Integr Physiol* 2009, **153**:425-429.
52. Schill R, Neumann S, Reuner A, Brümmer F: **Detection of DNA damage with single-cell gel electrophoresis in anhydrobiotic tardigrades.** *Comp Biochem Physiol A Mol Integr Physiol* 2008, **151**:32-32.

53. Schill RO, Mali B, Dandekar T, Schnölzer M, Reuter D, Frohme M:
     Molecular mechanisms of tolerance in tardigrades: new per-
     spectives for preservation and stabilization of biological
     material. *Biotechnol Adv* 2009, **27:**348-352.
54. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C,
     Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C,
     Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD,
     Stupka E, Wilkinson MD, Birney E: **The Bioperl toolkit: Perl mod-**
     ules for the life sciences. *Genome Res* 2002, **12:**1611-1618.

Chapter 4.

Transcriptome survey of the anhydrobiotic tardigrade *Milnesium tardigradum* in comparison with *Hypsibius dujardini* and *Richtersius coronifer*

BMC
Genomics

RESEARCH ARTICLE

Open Access

Transcriptome survey of the anhydrobiotic tardigrade *Milnesium tardigradum* in comparison with *Hypsibius dujardini* and *Richtersius coronifer*

Brahim Mali[1*], Markus A Grohme[1], Frank Förster[2], Thomas Dandekar[2], Martina Schnölzer[3], Dirk Reuter[4], Weronika Wełnicz[1], Ralph O Schill[5], Marcus Frohme[1]

Abstract

Background: The phenomenon of desiccation tolerance, also called anhydrobiosis, involves the ability of an organism to survive the loss of almost all cellular water without sustaining irreversible damage. Although there are several physiological, morphological and ecological studies on tardigrades, only limited DNA sequence information is available. Therefore, we explored the transcriptome in the active and anhydrobiotic state of the tardigrade *Milnesium tardigradum* which has extraordinary tolerance to desiccation and freezing. In this study, we present the first overview of the transcriptome of *M. tardigradum* and its response to desiccation and discuss potential parallels to stress responses in other organisms.

Results: We sequenced a total of 9984 expressed sequence tags (ESTs) from two cDNA libraries from the eutardigrade *M. tardigradum* in its active and inactive, anhydrobiotic (tun) stage. Assembly of these ESTs resulted in 3283 putative unique transcripts, whereof ~50% showed significant sequence similarity to known genes. The resulting unigenes were functionally annotated using the Gene Ontology (GO) vocabulary. A GO term enrichment analysis revealed several GOs that were significantly underrepresented in the inactive stage. Furthermore we compared the putative unigenes of *M. tardigradum* with ESTs from two other eutardigrade species that are available from public sequence databases, namely *Richtersius coronifer* and *Hypsibius dujardini*. The processed sequences of the three tardigrade species revealed similar functional content and the *M. tardigradum* dataset contained additional sequences from tardigrades not present in the other two.

Conclusions: This study describes novel sequence data from the tardigrade *M. tardigradum*, which significantly contributes to the available tardigrade sequence data and will help to establish this extraordinary tardigrade as a model for studying anhydrobiosis. Functional comparison of active and anhydrobiotic tardigrades revealed a differential distribution of Gene Ontology terms associated with chromatin structure and the translation machinery, which are underrepresented in the inactive animals. These findings imply a widespread metabolic response of the animals on dehydration. The collective tardigrade transcriptome data will serve as a reference for further studies and support the identification and characterization of genes involved in the anhydrobiotic response.

Background

Desiccation tolerance or anhydrobiosis is the ability of an organism to survive almost complete drying without sustaining damage. Anhydrobiosis is observed in certain micro-organisms, plants and animals such as rotifers, brine shrimp cysts, tardigrades and insect larvae, as for example those of the *Polypedilum vanderplanki* [1-3]. Studying the mechanisms of tolerance against desiccation may lead to development of new methods for preserving biological materials, which is of enormous practical importance in industrial as well as in medical fields [4]. In the dry state, the metabolism is suspended and the duration that anhydrobiotic organisms can survive ranges from years to centuries. Tardigrades are able to survive long periods of desiccation [5-8]. The hitherto longest known observation of an

* Correspondence: bmali@tfh-wildau.de
[1]Molecular Biology and Functional Genomics, University of Applied Sciences Wildau, Bahnhofstraße 1, 15745 Wildau, Germany

extended lifespan of 20 years has been demonstrated in the anhydrobiotic state of the species *Echiniscus testudo* Doyère 1840 [9]. Anhydrobiosis probably depends on a series of complex morphological, physiological and genetic adaptations that involve the stabilization of macromolecular complexes. As a consequence, a number of components have been identified and appear to be important for protecting these organisms from desiccation damage. Among them are the highly hydrophilic LEA proteins, which have been initially described in plants but have been identified in several invertebrates [10,11,2,12], as well as non-reducing disaccharides like trehalose [13-16]. We are studying anhydrobiosis in the limno-terrestrial tardigrade *Milnesium tardigradum* Doyère 1840 which shows remarkable resistance to adverse environmental conditions in all stages of life [17] - even to extreme levels of ionizing radiation [18] and the vacuum of space in low earth orbit [19]. *M. tardigradum* outperforms several other tardigrade species in tolerance e.g. survival of extreme temperatures above 100°C [20] as well as freezing [21,22]. Similar anhydrobiotic resistance to extreme environmental stress has been observed in other animals such as bdelloid rotifers or chironomid larvae [23,24] suggesting common mechanisms that allow anhydrobiotic survival and conferring radiation tolerance. The tardigrade phylum currently includes more than 1000 species living in the sea, in fresh water and on land. These last, needing at least a film of water to be active, are called limno-terrestrial and include most of the anhydrobiotic species [25]. They have been studied for their fascinating ability to perform anhydrobiosis and consequently serve as a potential model for studying tolerance and survival of multicellular organisms to a variety of extreme environmental conditions. Although there are several physiological, morphological and ecological studies on anhydrobiotic tardigrades [26-30], only limited DNA sequence information from molecular phylogenetic studies is available [30-34]. However, some sequence resources are only available from the species *Hypsibius dujardini* Doyère 1840 [Daub et al. Unpublished data 2003] and *Richtersius coronifer* (Richters 1903) [35]. Studies of *H. dujardini* have been focused mainly on developmental and evolutionary biology [36-38]. In this study we generated 9984 ESTs of *M. tardigradum* from active and inactive (anhydrobiotic/tun) stages, thereby establishing *M. tardigradum* as a model for anhydrobiosis research. These ESTs and the resulting unigenes were functionally annotated using Gene Ontology vocabulary. Furthermore, a cross-species comparison between *M. tardigradum*, *H. dujardini* and *R. coronifer* has been performed.

Results and discussion
cDNA libraries and sequence datasets
We have generated two directionally cloned cDNA libraries from active and inactive stages of *M. tardigradum* and subjected them to single pass Sanger sequencing. Furthermore we retrieved two EST datasets from public sequence databases (see Table 1 and 2). The datasets used in this study consisted of EST sequences from, *M. tardigradum* active and inactive stages, *H. dujardini* [Daub et al. Unpublished data 2003] and *R. coronifer* [35] of which the latter two were retrieved from NCBI (National Center for Biotechnology Information) dbEST and the NCBI Trace Archive. The source of all tardigrade samples consisted of whole adult animals except for the *H. dujardini* sample where adults and juveniles had been pooled.

Analysis of the *M. tardigradum* cDNA library
As summarized in Table 1, a total of 9979 clones were sequenced from the *M. tardigradum* library generated from two stages, active and inactive, in order to obtain various transcripts and to extract putative anhydrobiotic candidate genes. Assembly of the ESTs allowed the identification of 1997 and 1858 non-redundant sequences for active and inactive stages, respectively. The average unigene length was 579 nucleotides. Homology search (BLASTX) using *M. tardigradum* unigenes against the NCBI database showed that nearly 50% of the ESTs had no corresponding entry in GenBank. All ESTs were deposited in GenBank (See accession numbers in the additional file 1). The three available tardigade datasets were processed and compared (Table 2, Figure 1) in order to get an overview of the similarity and redundancy between our library and the other two EST resources.

GO enrichment analysis of *M. tardigradum* ESTs
To study the functional differences between active and inactive stages of *M. tardigradum*, we performed a GO

Table 1 Summary of the expressed sequence tag (EST) analysis of the *M. tardigradum* stages (active and inactive).

Description	Active	Inactive
Total number of raw sequences	4992	4987
Total number of quality ESTs	3617	3498
Number of contigs	466	431
Number of ESTs in contigs	2103	2106
Average clone per contig	4.5	4.8
Number of singletons	1540	1437
Total non-redundant sequences	1997	1858
Blast hits (%)	52.83	51.18
No blast hits (%)	47.17	48.82

Figure 1 Clustered heatmap of *M. tardigradum* TBLASTX relative bit score similarities against *H. dujardini* & *R. coronifer*. Top BLAST bit scores are shown as color gradient corresponding to their value. The red trace was added for better visualization of similarity values. The asterisk (*) marks sequences that could not be found in *H. dujardini* but are highly conserved between *M. tardigradum* and *R. coronifer*.

enrichment analysis between the two datasets (Figure 2; additional file 2). Studying functional differences give insight into global mechanisms that are at work in the desiccating animals. Comparing the datasets revealed that 24 GO terms were significantly underrepresented in the inactive stage. The underrepresented GO-terms which were mapped to "nucleosome", "nucleosome assembly", "chromatin assembly or disassembly" and "chromatin assembly" (GO:0000786, GO:0006334, GO:0006333, GO:0031497) consist exclusively of transcripts coding for histones. The cellular component (CC) subset of differential terms is also solely associated with structural components of the genome, such as "nucleosome" (GO:0000786), "chromatin" (GO:0005694), "chromosome" (GO:0000785), and "chromosomal part" (GO:0044427). Finding only underrepresented terms is consistent with the global metabolic arrest of animals

undergoing cryptobiosis. Histone mRNA expression is tightly linked to DNA replication and regulated by the cell cycle [39]. A study in *Caenorhabditis elegans* under anoxia showed similar adaptations such as cell cycle arrest, dephosphorylation of the histone H3 and morphological changes in the chromatin distribution [40]. A metabolic suppression could limit cellular and genomic damage by reducing the energy turnover to a minimum making the organism less susceptible to stress and therefore ensuring cell survival e.g. by decreased production of free radicals. Also GO-terms involved in translation regulation seem to be affected e.g. "regulation of translation" (GO:0006417), "translation regulator activity" (GO:0045182) and "translation factor activity, nucleic acid binding" (GO:0008135), implying modulation of translational activity as a response to desiccation.

The most abundant ESTs in active and inactive libraries of *M. tardigradum*

The total EST count obtained by comparing the active against the inactive dataset of *M. tardigradum* is summarized in the Table 3. The relative abundance of some transcripts in the inactive stage may indicate that they have been transcribed during the desiccation process or have been stored to be translated on rehydration. Biologically, high survival rates in *M. tardigradum* are accomplished only when drying slowly at high relative

Table 2 Summary of number of EST sequences, contigs, and singletons in tree tardigrade cDNA libraries.

	M. tardigradum	*H. dujardini*	*R. coronifer*
# of raw sequences	9984	5235	3360
# of quality ESTs	7209	5221	2819
Singleton	2419	1640	1083
Contigs	864	707	373
unigene	**3283**	**2347**	**1456**

Figure 2 Differentially represented Gene Ontology terms between active and inactive EST libraries of *M. tardigradum*. GO enrichment analysis between active and inactive EST libraries using Fisher's exact test with a false discovery rate (FDR) cutoff of p ≤ 0.05 (*) and p ≤ 0.01 (**). The number of transcripts associated with a specific GO term are represented as percentage of all functionally annotated EST in their respective libraries (active: N = 1207 and inactive: N = 1055).

humidity [16,41], suggesting that anhydrobiotic tardigrades like rotifers [42], need time to activate certain mechanisms for optimal anhydrobiosis. Probably this is because the transcription of RNAs coding for protection components has to take place. Among the genes that have a higher relative representation in the inactive stage are as follows:

Lipid-related transcripts
Lipid-related transcripts are represented mainly by intracellular fatty acid binding protein (FABP). FABPs have a low molecular mass and bind with high affinity to hydrophobic ligands such as saturated and

unsaturated long-chain fatty acids. Various functions have been proposed for FABPs such as the uptake, transport, and delivery of fatty acids to beta-oxidation [43,44]. FABPs are also thought to be active fatty acid chaperones by protecting and shuttling fatty acids within the cell [45,46]. However the biological role and mechanisms of action of FABPs remain poorly understood. The transcript level of FABP was identified by cDNA array and Northern blot analysis as being up-regulated during hibernation of ground squirrels [43,47]. Members of the FABPs family have recently been identified and reported to increase in the monogonont rotifer *Brachionus plicatilis* during dormancy

Table 3 The most abundantly represented transcripts in the *M. tardigradum* active and inactive libraries.

Gene family	EST count		E-value
	Active	Inactive	
unknown	65	113	
cytochrome b	43	37	1E-080
intracellular fatty acid binding protein	30	43	7E-015
kazal-type serine proteinase inhibitor	24	38	7E-06
unknown	38	32	
ATP synthase F0 subunit 6	28	28	7E-017
unknown	30	21	
unknown	21	28	
unknown	19	23	
cytochrome c oxidase subunit III	21	20	4E-036
cytochrome c oxidase subunit I	8	29	5E-144
40S ribosomal protein S27	13	15	2E-038
40S ribosomal protein S25	19	13	4E-023
40S ribosomal protein S21	10	20	7E-024
NADH dehydrogenase subunit 4	13	15	2E-021
unknown	16	10	
vitellogenin	9	16	2E-013
similar to Actin-5C isoform 2	3	16	1E-015
cytochrome oxidase subunit II	8	10	2E-034
unknown	5	12	
cystatin B	5	10	5E-014
NADH dehydrogenase subunit 5	4	10	3E-021
elongation factor 1 alpha	10	2	4E-124

[48]. The presence of FAPB in inactive stage of *M. tardigradum* may imply conserved mechanisms shared between rotifer dormancy and anhydrobiosis in tardigrades and presumably other organisms as well. FAPB may protect membranes and ensure fatty acids as energy saving storage during anhydrobiosis.

Protease inhibitors
To date, little is known about the possible mechanisms of proteolytic inhibition or suppression in anhydrobiotic organisms. Protease inhibitors are candidate genes which would offer protection against protein degradation during anhydrobiosis. Among the abundant protease inhibitors transcripts in inactive stages of *M. tardigradum* are Kazal-type serine proteinase inhibitor and Cystatin B. Overexpression of Cystatin B (an intracellular cysteine proteinase inhibitor) in transgenic yeast and *Arabidopsis* showed an increase in the resistance to high salt, drought, oxidative, and cold stresses [49]. Elevated levels of transcripts coding for protease inhibitors such as Cystatin B have also been found in brine shrimp cysts [50]. The abundance of protease inhibitors may inhibit proteolytic reactions of proteases that could damage tissues during the desiccation process or as a response to induction of proteases as a result of

aggregated proteins. Also a protection against microbial degradation could be possible as this can occur at humidity levels at which tardigrades can't rehydrate and actively mobilize any cellular defence mechanisms.

Cytochrome c oxidase subunit I
Cytochrome c oxidase subunit I (COXI) is a mitochondrial gene that encodes the cytochrome c oxidase subunit I, a crucial enzyme involved in oxidative phosphorylation and thus energy production. COXI was over threefold more represented in the inactive state. Transcripts encoding COXI were also abundantly expressed during dehydration stress in the antarctic nematode *Plectus murrayi* [51] and up regulated by temperature increase in the yeast-like fungus *Cryptococcus neoformans* [52]. The mitochondrial COXI upregulation may serve to prevent the damage to the electron transport chain caused by desiccation and to keep an increased energy production for the survival of the tardigrades.

The sequences, which could not be assigned any function based on homology search in NCBI, were searched for conserved domains in ProDom [53] and Swiss-Prot databases [54] but did not show any hits. Since these are not all full-length sequences, it is possible that they may have missed characteristic motifs or domains for classification. A detailed investigation of their function as well as other identified transcripts presented in Table 3 will be a task in the future.

Transcripts with putative functions in desiccation resistance identified in all three tardigrade species datasets
In a cross-search over the four tardigrade EST resources (active and inactive libraries of *M. tardigradum*, *R. coronifer* and *H. dujardini*), transcripts which are potentially associated with desiccation tolerance during anhydrobiosis in other organisms were identified (see additional files 3, 4 and 5).

Detoxification-related genes
Oxidative stress proteins have been shown to be an important component in many biological processes [55]. They mediate detoxification and have putative roles as antioxidants such as glutathione S-transferase (GST), thioredoxin, superoxide dismutase (SOD), glutathione peroxidases and peroxiredoxin. It was shown that overexpression of GST/glutathione peroxidase increased the resistance to oxidative and water stress in transgenic tobacco plants [56]. GSTs are a diverse superfamily of multifunctional proteins that are reported to play a prominent role in the detoxification metabolism in nematodes [57]. In particular the up-regulation of detoxifying enzymes GST and SOD in *Plectus murrayi* [51] suggests

an efficient role of reactive oxygen species (ROS) scavenging mechanisms under desiccation stress. These observations led us to postulate that the tardigrade GST and SOD are likely to deal with oxidatively damaged cellular components during desiccation. These enzymes that help in the removal of these compounds contribute to cellular survival after oxidative damage.

Aquaporins
Many organisms adapt to desiccation stress by the activation of various water-channel proteins, called aquaporins (AQP) [58,59]. Data from *Polypedilum vanderplanki* indicates that of the two aquaporins isolated from this organism, one is involved in anhydrobiosis, whereas the other controls water homeostasis of the fat body during normal conditions [60]. Similarly, the aquaporins in larvae of the goldenrod gall fly, *Eurosta solidaginis* were either upregulated (AQP3) or downregulated (AQP2 and AQP4) following desiccation [61]. The upregulated AQP3 is especially intriguing because it is permeable to water and glycerol across the cell membrane as larvae prepare for the osmotic stress associated with desiccation. In our study aquaporin transcripts have been identified in all tardigrade datasets. These AQPs may act in concert with other transmembrane proteins to mediate the rapid transport of water across the plasma membrane during anhydrobiosis when its diffusion through the phospholipid layer of the membrane is limited.

Molecular chaperones
In the four tardigrade datasets we have identified some putative heat shock protein (HSP) encoding genes. HSPs are highly conserved throughout evolution and they function as molecular chaperones and play primary roles in protein biosynthesis and folding [62]. In tardigrades, there is considerable debate concerning the role of HSPs under desiccation stress. In the *R. coronifer*, a lower level of Hsp70 protein was found in desiccated animals when compared with active ones [63]. In *M. tardigradum*, one isoform of the *hsp70* tarnscripts showed up-regulation during the transition from active to the inactive state [64,65], while the other *hsp70* isoforms are downregulated and seem not to be directly involved in anhydrobiosis. Using the same model *M. tardigradum*, Reuner et al. [65] found an upregualtion of hsp90 in the inactive state. Certainly Hsp70 isoforms and hsp90 are involved in tardigrade desiccation, but further studies are necessary to understand how these proteins work to protect anhydrobiotic organisms.

Much attention was recently paid on the chaperone-like LEA (late embryogenesis abundant) proteins in anhydrobiotic animals [66,67]. LEA proteins are mainly low molecular weight (10-30 kDa) proteins associated with tolerance to water stress resulting from desiccation and cold shock [68,69]. Genes encoding LEA-like proteins have been identified in the nematode *Aphelenchus avenae* under dehydration condition [70-72]. A similar gene was identified and upregulated in the larvae of *P. vanderplanki* by water stress imposed by either desiccation or hypersalinity [73]. Recently, LEA have also been identified and shown to be induced under dehydration in the springtail *Megaphorura arctica* [74]. In the tardigrade EST libraries, LEA transcripts have been found in the *H. dujardini* library (the less tolerant tardigrade) and also in the proteome map of *M. tardigardum* [75]. These data suggest that LEA-like proteins could be widespread in anhydrobiotic organisms and serve important functions during desiccation.

The translationally controlled tumor protein (TCTP) found in all tardigrada datasets is often designated as a stress-related protein because of its highly regulated expression in stress conditions and its close relation to a family of small chaperone proteins [76]. Importantly, TCTP can bind to native proteins and protect them from thermal denaturation [77].

Trehalose synthesis-related gene
Trehalose, which accumulates in many anhydrobiotic organisms during desiccation is proposed to act as a common water replacement molecules and stabilizer of biological structures [78-80]. The accumulation of trehalose has been reported in the cysts of the crustacean *Artemia franciscana* [81], in the nematode *Aphelenchus avenae* [82] and in the insect larvae of the *P. vanderplanki* [83]. However, anhydrobiotic Bdelloid rotifers are unable to produce trehalose [41,84]. In addition, the trehalose-6-phosphate synthase genes (*tps*) have not been found in rotifer genomes [41]. Although trehalose accumulates substantially in the eutardigrade *Adorybiotus coronifer* [85], it was surprisingly immeasurable in *M. tardigardum* [16] and we could not find transcripts of *tps* in *M. tardigradum* ESTs. Nevertheless, transcripts coding for trehalases have been described in *M. tardigradum* [86] but we propose that its function is probably limited to the catabolism of trehalose taken up from food sources. The hypothesis of trehalose as a protective agent during desiccation may not be applicable to all anhydrobiotic organisms and in *M. tardigradum* other strategies are probably employed.

Comparative ESTs analysis between the three tardigrade species
The datasets analysed in this study represent most of the available transcriptome data from tardigrades, and until now there is little information on tardigrade genome and transcriptome structure. The genome sizes range from very compact genomes, ~75 Mb for *H.*

dujardini, considered as one of the smallest tardigrade genomes [36], up to 800 Mb for other species http://www.genomesize.com. Our dataset adds a substantial part towards the complete gene content in the tardigrada species.

To investigate the complementation of the three tardigrade datasets (*M. tardigradum, H. dujardini* and *R. coronifer*) we searched for putative orthologous sequences across all three datasets. Using a TBLASTX search with an e-value threshold of 10^{-5} we compared the *M. tardigradum* unigenes against the other two datasets. The BLAST bit-score of each top-scoring hit was extracted and *M. tardigradum* sequences that exhibited sequence similarity against at least one other tardigrade species are presented as a clustered heatmap in Figure 1 (see also additional file 6). *M. tardigradum* unigenes show similarities against both other species with some hits only present in either one of them (N = 785). A higher coverage of *M. tardigradum* transcripts can be seen in the *H. dujardini* dataset compared to *R. coronifer* which is likely due to the smaller *R. coronifer* dataset. This cross species comparison implies that the remaining 2498 unigenes contained in the *M. tardigradum* dataset represent further yet unknown tardigrade transcripts and expands the known tardigrade sequence data. These might be very interesting for studying the evolutionary relationships of protein families.

To calculate the average relative transcriptome sequence similarity between *M. tardigradum* and the other two tardigrade species we included only sequences that were common to all three tardigrade species (N = 368). These contained mainly abundant transcripts e.g. ribosomal proteins, ADP-ribosylation factor, ubiquitin, glyceraldehyde-3-phosphate dehydrogenase and heat shock proteins. The resulting average transcript similarity for *M. tardigradum* against *H. dujardini* was 147.66 +/- 88.27 and *M. tardigradum* against *R. coronifer* 150.95 +/- 93.74. This is reflected in the phylogenetic distance calculated using 18S rRNA sequences (see additional file 7), which positions *R. coronifer* closer to *M. tardigradum*.

Conclusions

This study describes novel sequence data from the tardigrade *M. tardigradum* that identified a set of 3283 unigenes, which significantly contributes to the available tardigrade sequence data and will help to establish this tardigrade as a model for studying desiccation tolerance. The comparison of active and inactive stage EST libraries by performing an exploratory GO enrichment analysis suggests a metabolic suppression in terms of replication and translation during desiccation. The tardigrade-EST resource generated from this study will serve as a reference for future global gene expression

experiments, aiming at the identification of key regulators of desiccation resistance during anhydrobosis. Furthermore the datasets of *H. dujardini* and *R. coronifer* will serve as additional resources that could give clues about the evolutionary conservation of these regulators between tardigrade species of different anhydrobiotic capabilities.

Methods
Animal culture and sampling

M. tardigradum was reared in a laboratory culture on 3% agar plates covered with Volvic® mineral water at 20 ± 2°C and a light/dark cycle of 12 h as previously published [8]. For all experiments, adult animals (eight weeks after hatching) in good condition were collected directly from the agar plate using a pipette and a stereomicroscope. Tardigrades were starved for 3 days, and washed for several times with Volvic® mineral water before being processed to avoid contaminations. A total of 1000 animals were collected into 1.5 ml Eppendorf tubes in aliquots of 200 animals each. Animals representing the active state were frozen directly in liquid nitrogen. Anhydrobiotic stages of *M. tardigradum* were generated by a previously published protocol [8]. Briefly, *M. tardigradum* (200 animals) were placed in 1.5 ml Eppendorf tube and desiccated at room-temperature at 85% relative humidity (RH) for 12 to 16 hours (till they have completed the tun formation) and then at 35% RH for further 48 hours. The animals were frozen at -80°C until their experimental use.

Library construction

Total RNA extraction was performed by following the instructions of QIAGEN RNeasy R Mini kit (Qiagen, Hilden, Germany). Complete lysis of the tardigrades and especially disruption of their harsh cuticle was achieved by sonication on ice for 1 min (duty cycle 0.5s) by using a microsonicator (Probe 73, Sonopuls; Bandelin). For cDNA synthesis 1 μg total RNA was reverse transcribed using the Creator™ SMART™ cDNA Library Construction Kit following the manufacturers recommendations (Clontech-TaKaRa Bio Europe, France). The resulting first strand cDNA was amplified by LD-PCR for 18 cycles according to the manufacturers protocol using the 5' PCR primer (5'-AAGCAGTGGTATCAACGCAGAGT-3') as the forward and the CDSIII/3'PCR Primer (5'-ATTCTAGAGGCCGAGGCGGCCGACATG-d(T)30N-1N-3') as reverse primer. The amplified PCR products were then analyzed by agarose gel electrophoresis. After digestion of the amplified cDNA with the *Sfi*I restriction enzyme, products smaller than 300 bp were removed using the Chroma Spin-400 column as described in the Creator SMART™ protocol and cloned into pDNR-Lib cloning vector. This procedure was

chosen because of the low amount of starting material. Plasmids were transferred via electroporation to *Escherichia coli* (strain DH10B, Invitrogen, Karlsruhe, Germany).

cDNA sequencing

In total, 9984 cDNA clones were either picked by hand or automatically using a QPix robot (Genetix, UK) into 384 well LB-agar culture plates containing chloramphenicol. Sequencing was mostly from the 5' end using standard M13 forward sequencing primer. The sequencing of the cDNA library was sequenced on a ABI 3730XL capillary sequencer by GATC Biotech AG (Konstanz, Germany).

Sequence analysis and annotation

The EST analysis pipeline (Figure 3) includes typically, EST pre-processing, EST assembly and annotation of the resulting unigenes. The result is the generation of a clean, high-quality EST sequence set. Both chromatogram (*M. tardigradum*) and FASTA sequences (*H. dujardini* and *R. coronifer*) files are accepted as entry point to the analysis. Base calling was performed with phred [87,88] using a score threshold of 20. Low quality sequences, cloning vector, poly A or T tails, adaptors, and short sequences (<100 base pairs) are removed from the sequences with SeqClean [89]. Repetitive elements are masked with RepeatMasker [90]. Sequences that can be considered contaminants and unexpected vector sequences are also removed with SeqClean, using NCBI's UniVec database (v5.1) [91].

TIGR Gene Indices clustering tools (TGICL) with standard parameters [92] and CAP3 [93] have been used for the assembly step. For functional annotation, processed putative unique transcripts were loaded into the Blast2GO software [94]. Blasting was done with BLASTX algorithm using Blast2GO (v2.3.5) standard parameters. Unigenes were annotated with GO terms using standard evidence GO weight parameters. The 'Augment Annotation by ANNEX' function was used to refine annotations. Subsequently, Inter-ProScan [95] was performed to find conserved functional domains. GO terms derived from domains were merged into the existing GO annotation of the respective unigenes.

GO enrichment analysis

Identification of GO terms differentially enriched between the active and inactive *M. tardigradum* datasets

Figure 3 Overview of the tardigrade EST processing pipeline. Software used by each process in this flowchart are shown in brackets.
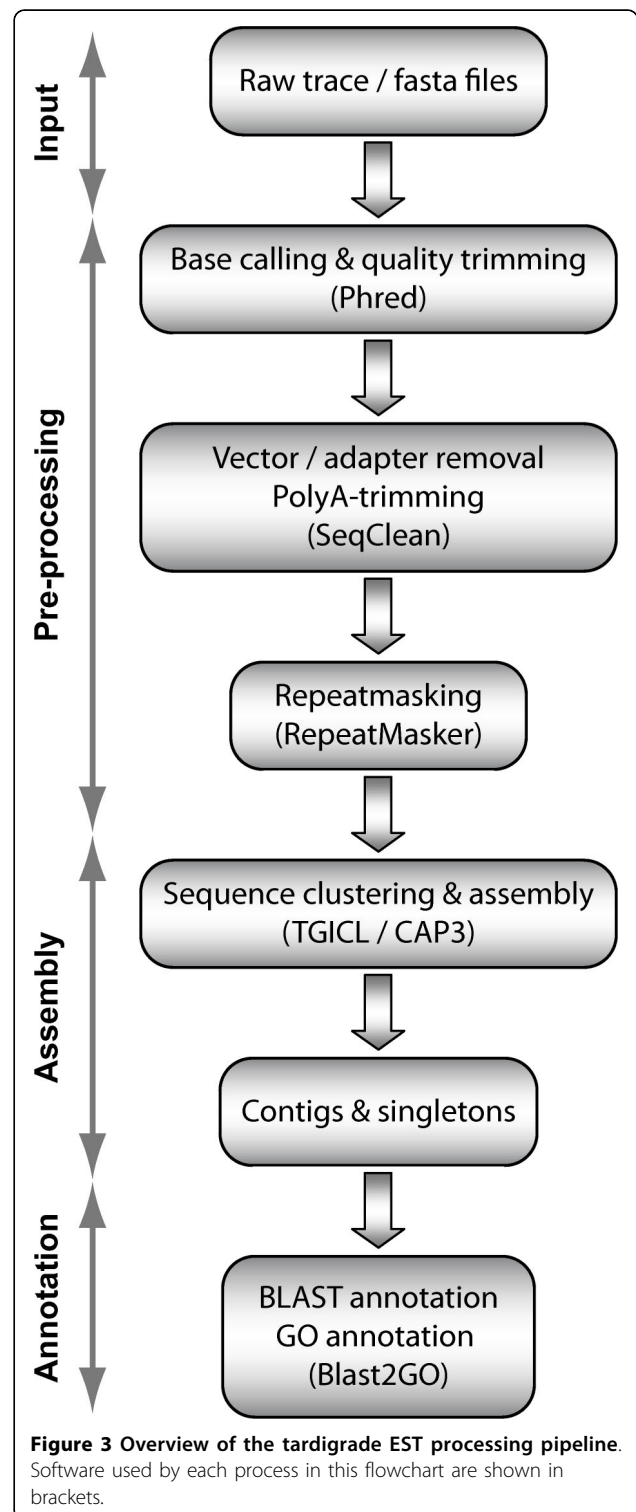
was performed using the GOSSIP statistical framework [96] webservice via the BLAST2GO software. GOSSIP employs 2 × 2 contingency tables of annotation frequencies for each GO term and computes p-values using Fisher's exact test. The statistical framework accounts for false positives (type-I-errors) that arise from multiple testing by calculating adjusted p-values. We screened for significantly enriched GO-terms by controlling the false discovery rate (FDR), setting a cut-off threshold of pFDR(p) ≤ 0.05. GO terms fulfilling this criterion were considered differentially enriched between the two *M. tardigradum* datasets.

Additional file 1: List of *M. tardigradum* ESTs and their GenBank accession numbers. This file provides a list of dbEST ID, User ID and GenBank accession numbers of all *M. tardigradum* ESTs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-168-S1.XLS]

Additional file 2: GO-enrichment analysis statistics of *M. tardigradum* datasets. This file contains details about the GO-enrichment analysis between the active and inactive stage of *M. tardigradum* using GOSSIP. A list of enriched ESTs is provided.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-168-S2.XLS]

Additional file 3: Putative anhydrobiotic transcripts identified in *M. tardigradum* dataset. This file provides a list of ESTs in active and inactive stages of *M. tardigradum* that are potentially associated with desiccation tolerance.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-168-S3.XLS]

Additional file 4: Putative anhydrobiotic transcripts identified in *H. dujardini* dataset. This file provides a list of *H. dujardini* sequences that are potentially associated with desiccation tolerance.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-168-S4.XLS]

Additional file 5: Putative anhydrobiotic transcripts identified in *R. coronifer* dataset. This file provides a list of *R. coronifer* sequences that are potentially associated with desiccation tolerance.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-168-S5.XLS]

Additional file 6: Putative orthologous sequences of *M. tardigradum* against *H. dujardini* and *R. coronifer*. This file provides a list of the putative orthologues shared by tardigrade species investigated in this study.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-168-S6.XLS]

Additional file 7: Phylogenetic tree based on tardigrade 18S rRNA sequences. Displays a phylogenetic tree constructed from *E. testudo, M. tardigradum, R. coronifer* and *H. dujardini* 18S rRNA sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-11-168-S7.PDF]

Author details
[1]Molecular Biology and Functional Genomics, University of Applied Sciences Wildau, Bahnhofstraße 1, 15745 Wildau, Germany. [2]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany. [3]German Cancer Research Center (DKFZ), Protein Analysis Facility, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany. [4]Oncoscience AG, Wedel, Germany. [5]Biological Institute, Zoology, Universität Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany.

Authors' contributions
BM established and optimized the tardigrade RNA extraction protocol and constructed and managed the cDNA clone libraries, MG performed functional annotation and enrichment analysis, putative orthologue prediction and gave useful comments on sequence analysis, MF was responsible for oversight, budget, obtaining the funding for the project, and contributing advice at each step of the research. FF performed quality control, processing and assembly of ESTs and was involved in data analysis, TD contributed to the bioinformatic analysis. WW performed the phylogenetic analysis, RS provided the animals and coordinated the project and contributed comments on candidate anhydrobiotic genes, MS and DR supported the identification of anhydrobiotic genes. BM and MG wrote the main part of the manuscript. All authors read and approved the final manuscript.

References
1. Crowe JH, Hoekstra FA, Crowe LM: **Anhydrobiosis**. *Annu Rev Physiol* 1992, **54**:579-599.
2. Tunnacliffe A, Lapinski J: **Resurrecting van Leeuwenhoek's rotifers: a reappraisal of the role of disaccharides in anhydrobiosis**. *Philos Trans R Soc Lond B Biol Sci* 2003, **358**:1755-1771.
3. Watanabe M: **Anhydrobiosis in invertebrates**. *Appl Entomol Zool* 2006, **41**:15-31.
4. Schill RO, Mali B, Dandekar T, Schnölzer M, Reuter D, Frohme M: **Molecular mechanisms of tolerance in tardigrades: New perspectives for preservation and stabilization of biological material**. *Biotech Adv* 2009, **27**:348-352.
5. Baumann H: **Bemerkungen zur Anabiose der Tardigraden**. *Zool Anz* 1927, **72**:1-4.
6. Guidetti R, Jönsson KI: **Long-term anhydrobiotic survival in semi-terrestrial micrometazoans**. *J Zool* 2002, **257**:181-187.
7. Bertolani R, Guidetti R, Jönsson KI, Altiero T, Boschini D, Rebecchi L: **Experiences with dormancy in tardigrades**. *J Limnol* 2004, **63**:16-25.
8. Hengherr S, Brümmer F, Schill RO: **Anhydrobiosis in tardigrades and its effects on longevity traits**. *J Zool* 2008, **275**:216-220.
9. Jørgensen A, Møbjerg N, Kristensen RM: **A molecular study of the tardigrade *Echiniscus testudo* (Echiniscidae) reveals low DNA sequence diversity over a large geographical area**. *J Limnol* 2007, **66**(Suppl 1):77-83.
10. Browne J, Tunnacliffe A, Burnell A: **Anhydrobiosis: plant desiccation gene found in a nematode**. *Nature* 2002, **416**:38.
11. Gal TZ, Glazer I, Koltai H: **An LEA group 3 family member is involved in survival of *C. elegans* during exposure to stress**. *FEBS Lett* 2004, **577**:21-26.
12. Kikawada T, Nakahara Y, Kanamori Y, Iwata K, Watanabe M, McGee B, Tunnacliffe A, Okuda T: **Dehydration-induced expression of LEA proteins in an anhydrobiotic chironomid**. *Biochem Biophys Res Commun* 2006, **348**:56-61.
13. Oliver AE, Leprince O, Wolkers WF, Hincha DK, Heyer AG, Crowe JH: **Non-disaccharidebased mechanisms of protection during drying**. *Cryobiology* 2001, **43**:151-167.
14. Crowe LM: **Lessons from nature: the role of sugars in anhydrobiosis**. *Comp Biochem Physiol Part A Mol Integr Physiol* 2002, **131**:505-513.
15. Crowe JH: **Trehalose as a "chemical chaperone": fact and fantasy**. *Adv Exp Med Biol* 2007, **594**:143-158.
16. Hengherr S, Heyer AG, Köhler HR, Schill RO: **Trehalose and anhydrobiosis in tardigrades-evidence for divergence in responses to dehydration**. *FEBS J* 2008, **275**:281-288.
17. Schill RO, Fritz GB: **Desiccation tolerance in embryonic stages of the tardigrade *Milnesium tardigradum***. *J Zool* 2008, **276**:103-107.

18. Horikawa DD, Sakashita T, Katagiri C, Watanabe M, Kikawada T, Nakahara Y, Hamada N, Wada S, Funayama T, Higashi S, Kobayashi Y, Okuda T, Kuwabara M: Radiation tolerance in the tardigrade *Milnesium tardigradum*. *Int J Radiat Biol* 2006, **82**:843-848.

19. Jönsson KI, Rabbow E, Schill RO, Harms-Ringdahl M, Rettberg P: Tardigrades survive exposure to space in low Earth orbit. *Curr Biol* 2008, **18**:R729-R731.

20. Hengherr S, Worland MR, Reuner A, Brümmer F, Schill RO: High-temperature tolerance in anhydrobiotic tardigrades is limited by glass transition. *Physiol Biochem Zool* 2009, **82**:749-755.

21. Hengherr S, Reuner A, Worland R, Brümmer F, Schill RO: Freeze tolerance, super cooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades. *J Exp Biol* 2009, **212**:802-807.

22. Hengherr S, Reuner A, Brümmer F, Schill R: Ice crystallization and freeze tolerance in embryonic stages of tardigrades. *Comp Biochem Physiol Part A Mol Integr Physiol* 2010.

23. Gladyshev E, Meselson M: Extreme resistance of bdelloid rotifers to ionizing radiation. *Proc Natl Acad Sci USA* 2008, **105**:5139-5144.

24. Watanabe M, Nakahara Y, Sakashita T, Kikawada T, Fujita A, Hamada N, Horikawa DD, Wada S, Kobayashi Y, Okuda T: Physiological changes leading to anhydrobiosis improve radiation tolerance in *Polypedilum vanderplanki* larvae. *J Insect Physiol* 2007, **53**:573-579.

25. Guidetti R, Bertolani R: Tardigrade taxonomy: an updated check list of the taxa and a list of characters for their identification. *Zootaxa* 2005, **845**:1-46.

26. Ramløv H, Westh P: Survival of the cryptobiotic eutardigrade *Adorybiotus coronifer* during cooling to -196℃ : effect of cooling rate, trehalose level, and short-term acclimation. *Cryobiology* 1992, **29**:125-130.

27. Ramløv H, Westh P: Cryptobiosis in the eutardigrade *Adorybiotus coronifer* : tolerance to alcohols, temperature and de novo protein synthesis. *Zool Anz* 2001, **240**:517-523.

28. Wright JC: The significance of four xeric parameters in the ecology of terrestrial Tardigrada. *J Zool* 1991, **224**:59-77.

29. Jönsson KI, Borsari S, Rebecchi L: Anhydrobiotic Survival in Populations of the Tardigrades *Richtersius coronifer* and *Ramazzottius oberhaeuseri* from Italy and Sweden. *Zool Anz* 2001, **240**:419-423.

30. Guil N, Giribet G: Fine scale population structure in the *Echiniscus blumi-canadensis* series (Heterotardigrada, Tardigrada) in an Iberian mountain range–When morphology fails to explain genetic structure. *Mol Phylogenet Evol* 2009, **51**:606-613.

31. Nichols PB, Nelson DR, Garey1 JR: A family level analysis of tardigrade phylogeny. *Hydrobiologia* 2006, **558**:53-60.

32. Sands CJ, Convey P, Linse K, McInnes SJ: Assessing meiofaunal variation among individuals utilising morphological and molecular approaches: an example using the Tardigrada. *BMC Ecol* 2008, **8**:7-18.

33. Schill RO, Steinbrück G: Identification and differentiation of heterotardigrada and eutardigrada species by riboprinting. *J Zool Syst Evol Res* 2007, **45**:184-190.

34. Schill RO, Förster F, Dandekar T, Wolf M: Distinguishing species in Paramacrobiotus (Tardigrada) via compensatory base change analysis of internal transcribed spacer 2 secondary structures, with the description of three new species. *Org Divers Evol* .

35. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD, Sørensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G: Broad hylogenomic sampling improves resolution of the animal tree of life. *Nature* 2008, **452**:745-749.

36. Gabriel WN, McNuff R, Patel SK, Gregory TR, Jeck WR, Jones CD, Goldstein B: The tardigrade *Hypsibius dujardini*, a new model for studying the evolution of development. *Dev Biol* 2007, **312**:545-559.

37. Gabriel WN, Goldstein B: Segmental expression of Pax3/7 and engrailed homologs in tardigrade development. *Dev Genes Evol* 2007, **217**:421-433.

38. Bavan S, Straub VA, Blaxter ML, Ennion SJ: A P2X receptor from the tardigrade species *Hypsibius dujardini* with fast kinetics and sensitivity to zinc and copper. *BMC Evol Biol* 2009, **9**:17.

39. Marzluff WF, Duronio RJ: Histone mRNA expression: multiple levels of cell cycle regulation and important developmental consequences. *Curr Opin Cell Biol* 2004, **14**:692-699.

40. Padilla PA, Nystul TG, Zager RA, Johnson ACM, Roth MB: Dephosphorylation of cell cycle-regulated proteins correlates with anoxia-induced suspended animation in *Caenorhabditis elegans*. *Mol Biol Cell* 2002, **13**:1473-1483.

41. Horikawa DD, Higashi S: Desiccation tolerance of the tardigrade *Milnesium tardigradum* collected in Sapporo, Japan, and Bogor, Indonesia. *Zoolog Sci* 2004, **21**:813-816.

42. Lapinski J, Tunnacliffe A: Anhydrobiosis without trehalose in bdelloid rotifers. *FEBS Letters* 2003, **553**:387-390.

43. Hittel D, Storey KB: Differential expression of adipose - and heart-type fatty acid binding proteins in hibernating ground squirrels. *Biochim Biophys Acta* 2001, **1522**:238-243.

44. Haunerland NH, Spener F: Fatty acid-binding proteins–insights from genetic manipulations. *Prog Lipid Res* 2004, **43**:328-349.

45. Coe NR, Bernlohr DA: Physiological properties and functions of intracellular fatty acid-binding proteins. *Biochim Biophys Acta* 1998, **1391**:287-306.

46. Makowski L, Hotamisligil GS: Fatty Acid Binding Proteins - The Evolutionary Crossroads of Inflammatory and Metabolic Responses. *J Nutr* 2004, **134**:2464S-2468S.

47. Hittel D, Storey KB: The translation state of differentially expressed mRNAs in the hibernating 13-lined ground squirrel (*Spermophilus tridecemlineatus*). *Arch Biochem Biophys* 2002, **401**:244-254.

48. Denekamp NY, Thorne MAS, Clark MS, Kube M, Reinhardt R, Lubzens E: Discovering genes associated with dormancy in the monogonont rotifer *Brachionus plicatilis*. *BMC Genomics* 2009, **10**:108.

49. Zhang X, Liu S, Takano T: Two cysteine proteinaseinhibitors from *Arabidopsis thaliana*, AtCYSa and AtCYSb, increasing the salt, drought, oxidation and cold tolerance. *Plant Mol Biol* 2008, **68**:131-143.

50. Chen WH, Ge X, Wang W, Yu J, Hu S: A gene catalogue for post-diapause development of an anhydrobiotic arthropod *Artemia franciscana*. *BMC Genomics* 2009, **10**:52.

51. Adhikari BN, Wall DH, Adams BJ: Desiccation survival in an Antarctic nematode: molecular analysis using expressed sequenced tags. *BMC Genomics* 2009, **10**:69.

52. Toffaletti DL, Del Poeta M, Rude TH, Dietrich F, Perfect JR: Regulation of cytochrome c oxidase subunit 1 (COX1) expression in *Cryptococcus neoformans* by temperature and host environment. *Microbiology* 2003, **149**:1041-1049.

53. Corpet F, Servant F, Gouzy J, Kahn D: ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* 2000, **28**:267-269.

54. Bairoch A, Apweiler R: The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucleic Acids Res* 1998, **26**:38-42.

55. Françaa MB, Paneka AD, Eleutherio ECA: Oxidative stress and its effects during dehydration. *Comp Biochem Physiol* 2007, **146**:621-631.

56. Roxas VP, Smith RK, Allen ER, Allen RD: Overexpression of glutathione S-transferase/glutathione peroxidase enhances the growth of transgenic tobacco seedlings during stress. *Nat Biotechnol* 1997, **15**:988-991.

57. Lindblom TH, Dodd AK: Xenobiotic detoxification in thenematode *Caenorhabditis elegans*. *J Exp Zool* 2006, **305**:720-730.

58. Izumi Y, Sonoda S, Yoshida H, Danks HV, Tsumuki H: Roleof membrane transport of water and glycerol in the freeze toleranceof the rice stem borer, *Chilo suppressalis* Walker (Lepidoptera: Pyralidae). *J Insect Physiol* 2006, **52**:215-220.

59. Philip BN, Yi SX, Elnitsky MA, Lee RE Jr: Aquaporins play a role in desiccation and freeze tolerance in larvae of the goldenrod gall fly, *Eurosta solidaginis*. *J Exp Biol* 2008, **211**:1114-1119.

60. Kikawada T, Saito A, Kanamori Y, Fujita M, Snigórska K, Watanabe M, Okuda T: Dehydration-inducible changes in expression of twoaquaporins in the sleeping chironomid, *Polypedilum vanderplanki*. *Biochim Biophys Acta* 2008, **1778**:514-520.

61. Philip BN, Yi SX, Elnitsky MA, Lee RE Jr: Aquaporins play a role in desiccation and freeze tolerance in larvae of the goldenrod gall fly, *Eurosta solidaginis*. *J Exp Biol* 2008, **211**:1114-1119.

62. Feder ME, Hofmann GE: Heat shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annu Rev Phys* 1999, **61**:243-282.

63. Jönsson KI, Schill RO: Induction of Hsp70 by desiccation, ionising radiation and heatshock in the eutardigrade *Richtersius coronifer*. *Comp Biochem Physiol B Comp Physiol* 2007, **146**:456-460.

64. Schill RO, Steinbrück GH, Köhler H-R: **Stress gene (hsp70) sequences and quantitative expression in** *Milnesium tardigradum* **(Tardigrada) during active and cryptobiotic stages.** *J Exp Biol* 2004, **207**:1607-1613.

65. Reuner A, Hengherr S, Mali B, Förster F, Arndt D, Reinhardt R, Dandekar T, Frohme M, Brümmer F, Schill RO: **Stress-response in tardigrades: Differential gene expression of molecular chaperones.** *Cell Stress & Chaperones* 2009.

66. Tunnacliffe A, Wise MJ: **The continuing conundrum of the LEA proteins.** *Naturwissenschaften* 2007, **94**:791-812.

67. McGee B, Schill RO, Tunnacliffe A: **Hydrophilic proteins in invertebrate anhydrobiosis.** *Integrative and Comparative Biology* 2004, **44**:679-699.

68. Ingram J, Bartels D: **The molecular basis of dehydration tolerance in plants.** *Annu Rev Plant Physiol Plant Mol Biol* 1994, **47**:377-403.

69. Hoekstra FA, Golovina EA, Buitink J: **Mechanisms of plant desiccation tolerance.** *Trends Plant Sci* 2001, **6**:431-438.

70. Browne JA, Tunnacliffe A, Burnell A: **Anhydrobiosis - plant desiccation gene found in a nematode.** *Nature* 2002, **416**:38.

71. Browne JA, Dolan KM, Tyson T, Goyal K, Tunnacliffe A, Burnell AM: **Dehydration-specific induction of hydrophilic protein genes in the anhydrobiotic nematode** *Aphelenchus avenae. Eukaryot Cell* 2004, **3**:966-975.

72. Goyal K, Walton LJ, Browne JA, Burnell AM, Tunnacliffe A: **Molecular anhydrobiosis: identifying molecules implicated in invertebrate anhydrobiosis.** *Integr Comp Biol* 2005, **45**:702-709.

73. Kikawada T, Nakahara Y, Kanamori Y, Iwata K, Watanabe M, McGee B, Tunnacliffe A, Okuda T: **Dehydration-induced expression of LEA proteins in an anhydrobiotic chironomid.** *Biochem Biophys Res Commun* 2006, **348**:56-61.

74. Bahrndorff S, Tunnacliffe A, Wise MJ, McGee B, Holmstrup M, Loeschcke V: **Bioinformatics and protein expression analyses implicate LEA proteins in the drought response of Collembola.** *J Insect Physiol* 2009, **55**:210-217.

75. Schokraie E, Hotz-Wagenblatt A, Warnken U, Mali B, Frohme M, Förster F, Dandekar D, Hengherr S, Schill RO, Schnoelzer M: **Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms.** *PLoS ONE* 2010.

76. Bommer UA, Thiele BJ: **The translationally controlled tumour protein (TCTP).** *Int J Biochem Cell Biol* 2004, **36**:379-385.

77. Gnanasekar M, Dakshinamoorthy G, Ramaswamy K: **Translationally controlled tumor protein is a novel heat shock protein with chaperone-like activity.** *Biochem Biophys Res Commun* 2009, **386**:333-337.

78. Wolkers WF, Tablin F, JH Crowe: **From anhydrobiosis to freeze-drying of eukaryotic cells.** *Comp Biochem Physiol* 2002, **131**:535-543.

79. Shirkey B, McMaster NJ, Smith SC, Wright DJ, Rodriguez H, Jaruga P, Birincioglu M, Helm R, Potts M: **Genomic DNA of** *Nostoc commune* **(Cyanobacteria) becomes covalently modified during long-term (decades) desiccation but is protected from oxidative damage and degradation.** *Nucleic Acids Res* 2003, **31**:2995-3005.

80. Crowe JH, Crowea LM, Wolkers WF, Oliver AE, Ma X, Auh J-H, Tang M, Zhu S, Norris J, Tablin F: **Stabilization of Dry Mammalian Cells: Lessons from Nature.** *Integr Comp Biol* 2005, **45**:810-820.

81. Clegg JS: **The origin of trehalose and its significance during emergence of encysted dormant embryos of** *Artemia salina. Comp Biochem Physiol* 1965, **14**:135-143.

82. Crowe JH, Madin KAC: **Anhydrobiosis in nematodes: evaporative water loss survival.** *J Exp Zool* 1975, **193**:323-334.

83. Watanabe M, Kikawada T, Minagawa N, Yukuhiro F, Okuda T: **Mechanism allowing an insect to survive complete dehydration and extreme temperatures.** *J Exp Biol* 2002, **205**:2799-22802.

84. Caprioli M, Kathiolm AK, Melno G, Ramløv H, Ricci C, Santo N: **Trehalose in desiccated rotifers: a comparison between a bdelloid and a monogonont species.** *Comp Biochem Physiol* 2004, **139**:527-532.

85. Westh P, Ramløv H: **Trehalose accumulation in the tardigrade** *Adorybiotus coronifer* **during anhydrobiosis.** *J Exp Biol* 1991, **258**:303-311.

86. Takekazu K, Takeo K: **Identification and analysis of trehalase from tardigrades,** *Milnesium tardigradum. Zool Sci* 2005, **22**:1493.

87. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.

88. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.

89. DFCI Gene Indices Software Tools. [http://compbio.dfci.harvard.edu/tgi/software/].

90. RepeatMasker homepage. [http://www.repeatmasker.org].

91. UniVec database at National Center for Biotechnology Information. [http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html].

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

93. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.

94. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36**:3420-3435.

95. Zdobnov EM, Apweiler R: **InterProScan-an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**:847-848.

96. Blüthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D: **Biological profiling of gene groups utilizing Gene Ontology.** *Genome Inform* 2005, **16**:106-15.

Chapter 5.

Transcriptome analyzed at different levels in *Hypsibius dujardini* and *Milnesium tardigradum*: Specific adaptations, motifs and clusters as well as general protective pathways

Transcriptome analyzed at different levels in *Hypsibius dujardini* and *Milnesium tardigradum:* Specific adaptations, motifs and clusters as well as general protective pathways

Frank Förster[1,§], Daniela Beisser[1,§], Markus Grohme[2,§], Chunguang Liang[1], Brahim Mali[2], Andy Reuner[3], Alexander Matthias Siegl[4], Julia Engelmann[1], Alexander Shkumatov[5], Elham Schokraie[6], Tobias Müller[1], Mark Blaxter[7], Martina Schnölzer[6], Ralph O. Schill[3], Marcus Frohme[2,*] and Thomas Dandekar[1,*]

[1]Dept of Bioinformatics, Biocenter University of Würzburg, 97074 Würzburg, Germany

[2]University of Applied Sciences, Bahnhofstraße 1, 15745 Wildau, Germany

[3]Dept. of Zoology, Institute for Biology, Universität Stuttgart, 70569 Stuttgart, Germany

4 Lehrstuhl für Botanik II, Julius-von-Sachs-Institut University of Würzburg, 97082 Würzburg, Germany

[5]EMBL, Hamburg Outstation, Notkestrasse 85, 22603 Hamburg, Germany

[6]Functional Proteome Analysis, German Cancer Research Center, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany

[7]Institute of Evolutionary Biology, University of Edingburgh, EH9 3JT,UK

[*]Corresponding authors, phone: 0049-931-888-4551. fax -4552;

e-mail: dandekar@biozentrum.uni-wuerzburg.de

[§]Contributed equally

Running title: *Comparing tardigrade transcriptomes*

Abstract

Tardigrades are unique metazoans with dormant stages ("tuns") resistant to extremes of cold, heat, radiation, and vacuum. Here we survey and translate all currently available transcriptome sequence information from the tardigrades *Hypsibius dujardini* and *Milnesium tardigradum* and systematically compare it on all levels to other organisms including *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*: ESTs from an ongoing large-scale transcriptome effort are collected and made available on the tardigrade workbench. mRNA stability and translational motifs influence protein expression and show a number of mRNA stability motifs for *M. tardigradum* while avoiding several typical motifs known from vertebrates. Encoded clusters of sequence similar, orthologous proteins shared in *M. tardigradum* and *H. dujardini* as well as more general protein families are identified by different clustering methods. Resulting pathways include stress-protective metabolic storage pathways for glycogen, glycolipids and specific secondary metabolism. Redox and protein protection pathways are present in both species, but are more diverse in *M. tardigradum*. This tardigrade also expresses specific pathways implicated in DNA protection. Stress pathways include late embryogenesis abundant (LEA) proteins, heat shock proteins and Bmh2 pathway. Finally, stress pathways have partly conserved regulation (man, nematodes) and a number of key tardigrade specific adaptations. [196 words]

Introduction

Tardigrades represent one of the three main invertebrate taxa where anhydrobiotic populations are wide-spread, the other two groups being nematodes and rotifers[1]. In these taxa, the anhydrobiotic state may be induced over the whole life cycle, from the egg to the adult stage (holo-anhydrobiosis,[2]). The mechanisms by which these so-called anhydrobiotic organisms preserve their cells under extreme desiccation have remained a challenge to biologists. In previous research on the biochemistry of anhydrobiotic organisms, a strong focus has been on the role of polyhydroxy compounds, mainly carbohydrates, as membrane stabilizers in the dry state[3; 4]. Thus the eutardigrade *Richtersius coronifer* was shown to accumulate the disaccharide trehalose at about 2.3% dry weight at the entrance of the anhydrobiotic state[5]. Physiological and molecular studies showed the role of several stress proteins (heat-shock proteins and late embryogenesis proteins) in the protection against desiccation damage[6; 7]. Stress and LEA (late embryogenesis abundant) proteins act as molecular chaperones and bind to other proteins, thereby preventing aggregation or unfolding of the protein or promoting protein folding[8; 9; 10; 11]. Stress proteins may also protect cells from oxidative damage both *in vivo* and *in vitro*[12; 13; 14].

Schill et al.[15] documented three heat-shock protein genes (Hsp70 family) in the tardigrade species *M. tardigradum* and the inductions of their different expression levels[16] in the cycle of dehydration, cryptobiosis and rehydration.

However, a more diversified approach is needed to fully capture the different facets of the superior stress adaptation in these organisms. Advances in transcriptomics allow us now diversified large-scale comparisons exploiting different EST resources for tardigrades. The present work is a comparison of new large-scale ESTs generated from the cosmopolitan terrestrial tardigrade *M. tardigradum* with its well known remarkable resistance to heat,

radiation and cold[17; 18; 19] and *H. dujardini,* another tardigrade studied for instance regarding development[20]. The transcriptome is analyzed on all hierachical levels. Sequence reads for ESTs are compared and assembled and added to the Tardigrade Workbench[21], a unique resource for tardigrade sequences. Next a transcriptome-based overview of encoded proteins, protein clusters and protein relatedness is achieved. This allows for the first time a systematic overview and comparison of the pathways in these tardigrades. We test the generality of the identified tardigrade specific stress adaptations by comparison to other organisms including *C. elegans*, *D. melanogaster* and *H. sapiens*. Details on the numerous sequences and comparisons are given in the supplementary material on our tardigrade website:

`http://tardigrade.bioapps.biozentrum.uni-wuerzburg.de.`

Key results are presented in the paper:

An ongoing transcriptome effort generates *M. tardigradum* ESTs., which are collected and provided on the Tardigrade Workbench[21]. The EST data allows us to identify RNA motifs. For *M. tardigradum* we can show that mRNA regulation by specific stability signals could be a further adaptation. We then examine and compare different higher levels of the transcriptome. Starting with the encoded proteins we find common proteins for both species. These have identical domain compositions according to Inparanoid clustering[22]. Furthermore, we compare clusters of similar proteins within *H. dujardini*, *M. tardigradum* or both. The general sequence similarity of encoded proteins is examined by CLANS[23], clustering proteins sufficiently similar (Blast e-value < 0.001). These two comparison types reveal a number of proteins with unknown function occurring only in tardigrades partly involved also in their unique physiology as well as clusters of proteins enhancing adaptive responses to stress.

To systematically compare encoded cellular functions and potential stress adaptations to other organisms we next used the well established clusters of orthologous groups (COGs) and identify the encoded protein repertoire in both species which is shared with other organisms. This includes LEA (late embryonic abundant) proteins implicated in stress tolerance and we

examine which subgroups occur in these two tardigrade species. Western blot analysis and PCR confirm experimentally LEA proteins and heat shock proteins, respectively, in *M. tardigradum*.

Building on these findings we systematically investigate tardigrade pathways: different enzymes indicate several metabolic pathways by their presence. For *H. dujardini* or *M. tardigradum* these include pathways for basic primary and secondary metabolism, but also storage and protective stress. Key adaptations were found in e.g. bmh2, redox and DNA protection pathways as well as heat-shock protein regulated paths. Pathways for DNA protections are especially strong in *M. tardigradum*, additionally both species feature a number of redox and stress protective pathways.

Finally we extend the comparison of stress proteins and pathways to pro- and eukaryotes to get a general comparison of the pathway inventory compared to other organisms. The data suggests a strong protection of proteins including redox protection and specific protection of DNA, but also points to specific pathways found in other organisms including *H. sapiens*.

Results

Transcriptome generation and analysis platform

H. dujardini: Expressed sequence tags were generated for *H. dujardini* in Edinburgh as part of a research program to study early development in this species. The ESTs were generated from a non-directional cDNA library. The library was constructed from adults and juveniles from an established laboratory culture derived from a single parthenogentic female (culture ED52). As *H. dujardini* releases eggs in the shed cuticle at each post-larval moult, the sample also includes egg and embryo transcripts. A total of 5235 sequences were deposited in dbEST.

M. tardigradum: Expressed sequence tags were generated for *M. tardigradum* as part of a research program to study stress adaptations for this species. The ESTs were generated from a

non-directional cDNA library. For these experiments adult animals in a good physical condition were taken directly from the culture and starved for three days to avoid preparation of additional RNA originating from not completely digested food in the intestinal system. The derived *M. tardigradum* ESTs are a mixture of tun stage and active motile animals. A total of 9982 *M. tardigradum* sequences were deposited in dbEST.

Tardigrade workbench: The tardigrade workbench server[21] was updated accordingly and includes now all the above mentioned new and all previous public EST data. Furthermore, adding systematic translation, all encoded proteins are available as well as standard protein databanks for comparison at the workbench. An online tutorial was added at

`http://tardigrade.bioapps.biozentrum.uni-wuerzburg.de/cgi-bin/about.pl`.

Furthermore, the data structure and the management of the database were improved. The implementation of the database was changed. This includes more options for the user, e.g. new options for COG statistics (see also below). The new server is available at

`http://tardigrade.bioapps.biozentrum.uni-wuerzburg.de` and serves as an analysis platform for further studies in tardigrade molecular makeup (for refereeing: username and password are "confidential" and "secretary" respectively and will be removed after acception of the manuscript; for anonymous access use e.g. http://www.the-cloak.com/anonymous-surfing-home.html).

Regulatory RNA motifs involved in post transcriptional regulation in *M. tardigradum*

As we here analyze EST data from many *M. tardigradum* sequences we systematically looked for regulatory RNA motifs in all available *M. tardigradum* sequences (Table I). From a screen of 30 regulatory motifs used in higher eukaryotes and in particular vertebrates (see Materials and methods) only twelve are used in *M. tardigradum* sequences. The regulatory motifs used include mRNA stability signals known from vertebrates such as the LOX 15-DICE element[24], brd box[25] and the alcohol dehydrogenase 3'UTR downregulation control element[26], in

contrast, other elements such as the widely used AU elements in vertebrates[27] are absent in

the *M. tardigradum* ESTs. Translational elements such as the IRE are found but used rarely.

Encoded proteins shared in both species

Building on these results, higher information levels of the transcriptome were analyzed. This

includes large-scale comparisons of encoded tardigrade proteins from the assembled data.

Candidate orthologue sequences occur in both species and have the same domain composition

and hence, very likely the same function. We used Inparanoid[22] to identify such shared

candidate orthologue sequences (sCOS), present in the *H. dujardini* and *M. tardigradum* EST

data (Fig. 1, top). Furthermore, if two sequences with the same domain composition occur in

the same species this is also identified by Inparanoid. Thus, we identified 178 sCOS,

corresponding to 214 *M. tardigradum* and 234 *H. dujardini* ESTs. The set of sCOS are a

conservative estimate as only the full genome sequence of both organisms would allow us to

identify all orthologues. However, these candidate orthologue sequences already show that a

substantial fraction of closely shared proteins exist in both species and a major fraction of

these include adaptations against environmental stress (Fig. 1, top).

Clusters of sequence similar encoded proteins occurring within
***M. tardigradum*, *H. dujardini* or both tardigrades**

CLANS[23] identifies *de novo* protein families just by sequence similarity (cut-off see Materials

and methods) in any collection of sequences and does not demand the sequence to have an

orthologue in another species. This allowed us to estimate and obtain the number of clusters

of sequence similar proteins either for both tardigrades by pooling all available ESTs or only

the ones present in the *H. dujardini* or only the ones in *M. tardigradum* EST data (Fig. 1,

bottom).

There are 13 CLANS clusters consisting of 203 proteins for *M. tardigradum*, 16 CLANS

clusters consisting of 246 proteins in *H. dujardini*, and 37 CLANS clusters pooling all

available ESTs from both species consisting of 684 proteins. These always include tardigrade-specific protein clusters (a total of 314 proteins in different clusters). Within the

M. tardigradum clusters, there are genes related to stress detoxification (cytochrome p450), protein turnover (cathepsin, metalloproteinase-like cluster and three ubiqutin-like clusters), protein secretion (Rab-like), redox protection (glutathione S-transferase, cystatin-like) and regulation (ras, calmodulin-like). The 2nd largest cluster is *M. tardigradum* specific.

H. dujardini also features species-specific clusters (four with no homologies including the top cluster) and again clusters of ras-like, of cathepsins, of glutathion S-transferase-like and cystatin-like proteins. Furthermore, there are dehydrogenases and a number of diapause-specific proteins. Diapause is a reversible state of developmental suspension and found from plants to animals, including marsupials and some other mammals[28] as well as insects and here, its presence supports either the tun form or the regulation of other (e.g. developmental) metabolic inactive states. Cell wall protection is probably supported by chitin binding proteins. Both Inparanoid and CLANS analyses and comparisons are presented in detail in the supplementary information on our tardigrade website.

The shared CLANS clusters suggest a tardigrade-specific optimization of protein protection by clusters of related proteins (details in supplementary table I on tardigrade website): Twelve CLANS clusters have no sequence homology to any functionally described sequence. Thus, they are only known from tardigrades but shared between both species (S-Table Ic). Furthermore, there are CLANS clusters with a superior functional description: clusters forming a ras subfamily[29], cathepsin-like proteins[30], cystatin-like[31], and chitin-binding proteins[32; 33], dehydrogenases and cytochrome p450-like proteins[34], as well as ubiquitin- and ubiquitin-conjugating enzyme-like[35; 36], thioredoxin-like and cysteine-rich proteins[37]. However, there are several additional families, for instance sulfotransferase-like[38], serin/threonine protein kinases[39], serin protease inhibitor-like[40] and signal recognition protein like[41] clusters.

Specific encoded protein families and those shared with other organisms that occur in *M. tardigradum* and *H. dujardini* ESTs

To compare the repertoire of functions in more detail, in this study we use for the first time the full transcriptome available from the sequencing effort on *M. tardigradum* ESTs and compare it to the available data regarding clusters of orthologous groups (COGs) and eukaryotic clusters of orthologous groups (KOGs). COGs were originally introduced by Tatusov et al.[42]. The COG clusters permit a genomic perspective on protein families from the rapidly accumulating genome sequences, as in this system all conserved genes are classified according to their homologous relationships into different COG clusters (or eukaryotic clusters, KOGs, see above). By looking at the COG/KOG repertoire the *M. tardigradum* inventory is directly compared to the protein inventory shared by all currently available genome sequences (several thousand clusters of genes encoding defined protein functions). The same analysis was done for *H. dujardini* as well as for clusters shared by both species (Table IIa-c; complete list in supplementary material, S-Table II on our tardigrade website). Table II summarizes specific molecular categories according to the COG/KOG classification in this comparison. This includes (Table IIa) molecular chaperons in *H. dujardini* [category O], uncharacterized proteins [category R] including unusual protein kinases and membrane ATPases [category C]. Ribosomal proteins occur in any organism [category J], however, there is COG/KOG evidence that there are specific proteins used here and that there are even more such specific ribosomal proteins in *M. tardigradum*. Furthermore, *M. tardigradum* has even more *Milnesium*-specific not yet well characterized proteins [categories S and R], metabolic enzymes [category C] and secondary metabolism involved in protein protection and turnover [category O]. Furthermore, there is evidence for specific DNA turnover and protection [category K] and transport processes including for instances Guanine nucleotide exchange factors and SNARE proteins[43] [category U]. Of course, many other functions are shared between both tardigrades.

Individual COG/KOGs: Among the COGs and KOGs, we find presumable tardigrade-specific reinforcements for protein protection such as some ubiquitin-like proteins (KOG0001) including ubiquitin-ligases and the ribosomal protein L40 fusion (KOG0003; remaining COGs/KOGs described here are given in Table II) , many ATP-dependent RNA helicases, several AAA+-type ATPases ([44]; including peptidase M41 domain) and mRNA splicing factor ATP-dependent RNA helicase shared by both tardigrade species. Furthermore, for protein protection, we find cyclophilin type peptidyl-prolyl cis-trans isomerase[45], proteasome components, cytochrome P450-like, GTPases and GTP-binding proteins (for the cytoskeleton). Regulation includes Ser/Thr kinases and phosphatases as well as glycogen synthase kinase-3[46]. There are only four (KOG2733-membrane, KOG3098, KOG4604, KOG4431) shared uncharacterized proteins, KOG4431 induced by hypoxia could be involved in tardigrade-specific adaptations. Apart from this, there are tardigrade versions of the basic eukaryotic proteome present such as ribosomal proteins, metabolic enzymes (e.g. triose phosphate isomerase, central carbohydrate metabolism, compare with above, nucleotide metabolism). For *M. tardigradum*, a number of species-specific adaptations compared to *H. dujardini* become apparent: Mitogen activated protein kinase, cyclin C-dependent kinase CDK8, Ser/Thr protein kinases, an adenylyl cyclase as well as germ-line stem cell division protein Hiwi/Piwi[47] and specific splicing factors, DNA/RNA helicases[48], SNF2-family DNA-dependent ATPase[49], probably involved in DNA protection (see also [50]), DNA polymerase epsilon, nucleotide excision repair protein RAD16 as well as additional DNA-directed RNA polymerases and a site-specific recombinase. There are several large families of uncharacterized conserved proteins, leucine-rich repeats and, quite interesting, a stress-induced morphogen[51] as well as the ribosomal biogenesis protein Nop58pa[52]. Different metabolic enzymes include several ubiquitin-specific enzymes and glutathione peroxidase for protein protection, general metabolic enzymes such as succinyl CoA synthetase and

dihydrolipoamide acetyltransferase as well as enzymes of secondary metabolism such as polypeptide N-acetylgalactosaminyltransferase.

For *H. dujardini* there are again several specific adaptations, these include molecular chaperons, archaeal/vacuolar H-ATPases, WD40 repeat proteins, an unusual protein kinase, some metabolic enzymes and for protein protection for instance the ubiquinol reductase (Table IIc). Once more, there seem to be specifically adapted ribosomal proteins.

Experimental validation of heat shock proteins and LEA proteins.

We directly tested the presence of stress proteins in *M. tardigradum* by PCR regarding hsps (Fig. 2) and by Western blot analysis using antibodies against heat shock protein Hsp70 and LEA protein (Fig. 3). A strong band was detected for Hsp70 slightly below 75 kDa as expected. When using the antibody against LEA proteins, one prominent band appeared at approx. 60 kDa accompanied by several faint bands in the lower and higher molecular weight region. Thus, the presence of Hsp70 and LEA proteins could be experimentally confirmed.

Phylogenetic distribution of LEA stress proteins occurring in these two tardigrades compared to other species

Preliminary data suggested as a first hint to stress-pathways that several LEA(late Embryogenesis abundant)-like proteins[53] also occur in tardigrade species. They are known to improve stress resistance in other organisms in particular in plants. As currently there is for the *M. tardigradum* LEA proteins no complete sequence available, we compared (Fig. 4) all identified LEA proteins from *H. dujardini* (13 sequences, red) and *Richtersius coronifer* (3 sequences, orange). The identified sequences were compared to a number of other *bona fide* LEA proteins from other organisms (PFAM seed alignment). We use the new classification by Tunnacliffe and Wise[53], and compare to their LEA protein groups 1, 2 and 3. Remaining higher LEA groups (4-6 in reference[53]) have no similarity to the tardigrade proteins. Groups 1, 2 and 3 correspond to the PFAM domains Lea_1, Lea_2, and Lea_4 and are shown (full seed alignment) in Fig. 4. Apart from one protein, all *H. dujardini* proteins cluster in two

branches, but have only Lea 4 proteins from other organisms as relatives. Thus all tardigrade sequences including all three *R. coronifer* sequences are found in the area of Lea 4-like proteins, a potential independent Lea protein family is represented by gi50297072 from *H. dujardini* (Fig. 4).

Predicted metabolic enzymes and resulting pathways from *M. tardigradum* and *H. dujardini* EST analysis

The sequencing effort on *M. tardigradum* as well as the existing EST sequences on *H. dujardini* allow us to give first estimates on central metabolic pathways involved in metabolic protection and stress resistance (Table III, details in S-Table IV on the tardigrade website). Thus *H. dujardini* and *M. tardigradum* have key glycolytic enzymes (the list in *M. tardigradum* is even more complete), aldolase is however up to now only found in *H. dujardini*. Starch and sucrose metabolism is represented by a number of enzymes. On the other hand, basic fatty acid biosynthesis has only one (*H. dujardini*) or two (*M. tardigradum*) general enzymes but clear evidence for fatty acid elongation in mitochondria and fatty acid conversion into different CoA-enzyme intermediates. The latter is again more completely demonstrated in *M. tardigradum* and here is also clear evidence for ketone body synthesis and their conversion (which should allow long term adaptation against carbohydrate shortage). Furthermore, linoleic acids are synthesized and there are several ESTs found encoding enzymes of the glycerophospholipid and sphingolipid metabolism. In the latter, we find ceramidase in both organisms, in *M. tardigradum* in addition sphingomyelin phosphodiesterase and several others. For prostaglandin mediated stress responses[54] enzyme evidence is again found in both species (Phospholipase A2, prostaglandin-D synthase), in *M. tardigradum* this includes furthermore the thromboxane A-synthase. As further signalling compounds, there is a detailed carotenoid metabolism. Furthermore, steroids are synthesized in both organisms (e.g. isopentenyl-diphosphate delta-isomerase), in *M. tardigradum* there are many enzymes found already in our EST library of the steroid metabolism. Also the ability to

branch off to terpenoid biosynthesis[55] is present in both organisms. Regarding intermediate metabolism, both organisms have aminophosphonate metabolism and a detailed ether lipid metabolism.

There are several metabolite pathways implicated in stress defence. Trehalose pathway is not clear in *M. tardigradum* or *H. dujardini* but there are clear pathways for glycogen storage and glycolipids (found in both organisms), or specific protein and membrane modifications (several such enzymatic pathways available for both organisms).

Stress pathways predicted from *M. tardigradum* and *H. dujardini* ESTs

In the first place, we examined the hypothesis whether a regulatory stress pathway around Bmh2 protein, which is known to exist in yeast[56], is also conserved in tardigrades. This stress pathway is shown here to be present in tardigrades. 14-3-3 related protein Bmh2 regulates several metabolic enzymes including production of trehalose and lipid synthesis. All these proteins form a complex in yeast[56]. The protein Bmh2 is found to be present in tardigrades according to the EST data *H. dujardini* (Fig. 5) as well as in *M. tardigradum* (suppl.Tables S-V, S-VI).

Furthermore, the regulatory input protein Sck1[57] is found in *H. dujardini* as well as in *M. tardigradum.* Several further metabolic proteins are conserved in man and yeast. Interestingly, this includes Trehalase and its metabolite Trehalose, but it has not yet been detected in *H. dujardini or M. tardigradum*[58]. However, the sequences of the master regulator Bmh2 and the dependent kinase Sck1 were found in *H. dujardini* as well as *M. tardigradum* and further verified by additional sequence analysis such as back-searches, Prosite Motifs and SMART domain composition[59]. Other bmh2-dependent metabolic enzymes are LCB1 or LCB2[56]. LCB1 and LCB2-like proteins were in fact found for *H. dujardini* but not for *M. tardigradum.* This includes six predicted acyltransferases in *H. dujardini* (COG0596, see suppl. material). However, as we are not sure which of these is the exact LCB1 or LCB2 homologue and have no experimental data which are directly regulated by Bmh2 there

remains a question mark in Fig. 5. However, the data suggest the complete stress pathway including dependent metabolic enzymes in *H. dujardini* but only key switches in *M. tardigradum,* their effector proteins still remain to be identified.

We next investigated the extent of standard desiccation tolerance pathways in the available *H. dujardini* and *M. tardigradum* sequence data. Alpert[60] compiled a number of pathways in desiccation tolerance which were all investigated. We found no genes indicating stress protection by photo pigments[61; 62]. This is in accordance with morphological data. Using systematic sequence to sequence comparisons starting from verified orthologues (see Material and Methods) of a number of organisms, we obtained the following results regarding stress pathway proteins using the available EST data from *H. dujardini* and *M. tardigradum* (Table IV, details in S-Table V a,b on the tardigrade website):

There are no pigment protection pathways present. As members of **protective pathways** there are RNA helicases[63] as well as peroxiredoxin, peroxidase and superoxide dismutase[64]. There is strong redox protection in both tardigrades, cold-shock like proteins and specific membrane protection pathways, e.g. regulated by MIP (major intrinsic protein,[65]) and by aquaporins to boost dessication tolerance[66; 67]. Several **LEA protein**[68; 69] candidates were found including very good matches to known LEA proteins (validation see Fig. 4). Different larger and smaller **heat shock proteins**[70] could be identified including hsp70 in *H. dujardini* as well as in *M. tardigradum.*

DNA repair includes the MutS pathway[71]. Furthermore, *M. tardigradum* contains specific DNA repair (RAD51, DnaJ family) as well as other DNA protection (DNA helicases) pathway proteins. In addition, further mechanisms such as anti-freezing proteins[72] were considered. Here, however, extensive searches including structure domain search, sequence

and domain analysis did not identify related sequences in *H. dujardini* ESTs or

M. tardigradum ESTs[1].

Moreover, we next checked how complete the different indicated pathways are in

M. tardigradum according to our data (Table IVc): Besides Bmh2 pathway, the homologous

DNA recombination, as well as other standard DNA repair pathways are present and also

regulated by a number of different Rad proteins. Different heat shock proteins and their

regulators as well as dependent enzymes become apparent. LEA proteins occur in plants and

animals. According to the EST data in tardigrades a specific animal signalling pathway is

present, the HOG signalling including PKA-mediated regulation. Also for the protective

pathways a number of regulatory proteins and dependent enzymes are predicted to be present.

Tardigrade-specific stress adaptations versus general adaptations shared with other animals

There are two other desiccation resistant phyla, rotifera and nematodes. To get an insight into

the general and tardigrade specific adaptations we compared our inventory of stress pathways

found in *M. tardigradum* and/or *H. dujardini* with stress pathways known from rotifera or

from nematodes. Conservation of key pathways in man is shown in Fig. 6. We considered all

proteins found either in *M. tardigradum* or *H. dujardini* or both to be involved in stress

protection and whether these are also found in nematodes, rotifers or man, as well as specific

genes involved in nematode or rotifer stress response and whether these are found in

tardigrades. A first overview on this is shown by clusters of orthologous groups (COGs or

eukaryotic KOGs;[42]) and using available genome information regarding protein clusters in

Nematodes (*C. elegans*) and *H. sapiens*[73]. We considered three major COG/KOG categories[73]

involved in protective functions: We compared our tardigrade data and protein clusters to the

other two organisms regarding the COG/KOG categories V (defence mechanisms), O

---

[1] We have a strong indications from mass spectrometry for anti-freezing proteins in *M. tardigradum*.

(posttranslational modification) and L (replication, recombination and repair). The overview (Venn diagram) in Fig 6 shows that there are many shared functions (172), but there are 81 tardigrade-"specific" protein clusters (i.e. they were neither found in man nor in nematodes). Overlap between organism groups is largest between nematodes and man. To better identify unique tardigrade adaptations we subtracted clusters common for tardigrades and yeast (Fig. 7) and end up with a list of 46 specific adaptations and COG/KOG clusters genuine for tardigrades within the context of this comparison. These are adaptations and clusters of orthologous groups identified from our EST analysis which do not occur in these other groups of desiccation tolerant animals and are in this sense and specific comparison tardigrade-specific. However, in any annotated tardigrade protein the sequence homology from which the annotation was derived, shows already that the protein occurs also in some other species. Only the tardigrade-specific protein clusters (supplementary table I on tardigrade website) are strictly tardigrade-specific but have not yet an annotation.

In the above sense and in comparison to rotifers, nematodes, human and yeast there are many tardigrade specific proteins (46 protein clusters, Fig. 6) and they include COGs for glutaredoxin-related proteins and molecular chaperons as well as KOGs involved in molecular chaperoning. This includes the DnaJ superfamily with a large number of different clusters (Milnesium-specific, Table IVc; supplementary table V and II on tardigrade website; see above). The COG clusters 4973, 4974 support this by supplying site-specific recombinases[74]. Collectively, these should enhance DNA protection and this helps to explain the extraordinary resistance of *M. tardigradum* against DNA damage documented recently[50; 75]. Details on all involved functions and protein clusters are given in the supplementary material on our tardigrade website including basic COGs/KOGs shared with yeast in tardigrades (Table S-VII). Fig. 8 summarizes all EST predicted adaptations and compares their distribution in other organisms, focussing on the predicted encoded protein families. Tardigrades exploit adaptations from nematodes, rotifers and vertebrates as well as individual adaptations. Note

that these are only those which can be predicted from the transcriptome. In addition, there are a number of tardigrade specific EST clusters where no sequence similarity allows predictions on any specific function. Interestingly, some protein functions (11) are shared only between tardigrades and *H. sapiens* in this comparison (Fig. 8 and supplementary table VIId on tardigrade website): These are pathways including DNA repair protein RHP57[76] and ubiquitin protein ligases[77] as well as proteasome maturation factor[78]. Finally information on rotifers is sparse. Supplementary material (S-Table VIIf,g) on our tardigrade website shows at least that a mitochondrial chaperonin is shared with man and *C. elegans* and that an hsp90 family-type molecular chaperone[79; 80] is instead shared with tardigrades.

Discussion

Unique stress adaptation capabilities are found after by tardigrades transcriptome sequence analysis. We present here a detailed comparative analysis including predicted protein clusters and resulting pathways of *H. dujardini* and *M. tardigradum*. Our study considers all available transcriptome data from two current EST sequencing projects in the tardigrades *M. tardigradum* and *H. dujardini*. The present study focuses on the transcriptome data and predicted protein families and pathways, however, for two protein families (LEA proteins, heat shock proteins), there is also complementary experimental evidence from PCR and Western blotting. By direct comparison of the different clusters, the predicted proteins shared by both species are identified. Note that both EST libraries are independent efforts. The EST library of *H. dujardini* was generated in the laboratory of M. Blaxter and is available at Genbank. The *M. tardigradum* transcriptome is newly generated as part of a general effort (www.funcrypta.de) to understand the adaptation and physiology of *M. tardigradum*. The aim of this study is to analyze the different levels of the transcriptome, starting from RNA motifs but focussing on different types of sequence clusters and pathways as well as their verification by experiments and by further comparisons. Further studies will be or are already looking into

other aspects of the adaptation process such as differential EST libraries between tun and active state in *M. tardigradum* or changes in specific metabolites during adaptation as well as comparison to further tardigrade adaptations.

Our study is a conservative estimate of the different encoded protein clusters and pathways analyzed, with a full *M. tardigradum* genome sequence and full transcriptome data the description will become more complete. However, already the available data (9982 EST sequences from *M. tardigradum*, 5235 EST sequences from *H. dujardini*) show a number of interesting features in these two different tardigrades. Thus our study shows for the available data in a systematic sequence to sequence comparison that a substantial fraction of the sequences (12.2% for all sequences no matter from which of the two compared tardigrades) comes in sequence-similar families involved in protein protection, redox protection, protein turnover, signalling as well as tardigrade-specific unique clusters (Fig. 1 bottom). Interestingly, CLANS clustering of all sequence-similar related proteins shows this holds for both *M. tardigradum* (6.2%, 13 clusters) and *H. dujardini* (10.5%, 16 clusters). However, there is stronger stress adaptation potential in *M. tardigradum*. This general trend is supported by the further analyses reported in this paper. Note that each of the following different analyses reveals different specific sequence families and pathways involved in tardigrade-specific adaptations:

First, we compare the predicted metabolic inventory of both tardigrades as well as specific regulatory motifs in *M. tardigradum* mRNA. A number of specific motifs are used to regulate mRNA in *M. tardigradum*. The key motifs (lox P DICE[24] k-box[81], brd-box[25]) are all regulatory for stability but avoid standard motifs known from vertebrates such as the AU-rich element for mRNA instability[27].

Specific metabolic pathways include central primary metabolism and adaptions in lipid and carbohydrate metabolism. Furthermore, there are metabolite connections to stress defence, for

instance regarding storage carbohydrates such as glycogen and lipid metabolism to foster membrane protection.

Moreover, we identify a number of specific pathways involved in stress adaptation including some that also occur in vertebrates. We can point out several major protein families and assemble them to pathways in *M. tardigradum* which are implicated in its excessive stress resistance to high temperatures or vacuum and cold in its tun stage[16; 17; 18; 19; 58; 82]. We see several stress-related clusters of sequence similar proteins, clusters of orthologous groups in eukaryotes (KOGs) or prokaryotes as well as metabolic pathways. While there is generally high overlap in the latter, the other two comparisons show high protection for proteins and against redox stress in both organisms but a number of species-specific families not found in the other. In accordance with observations from physiology, *M. tardigradum* has even more specific protection pathways and hence, involved protein families than *H. dujardini*. This includes specific protein families involved in DNA protection[83]. A recent study highlighted the high repair potential in *M. tardigradum* in comet assays[50] but the machinery behind the high repair potential and the numerous DNA breaks endured during tun stage was not clear. The different DNA repair protein pathways analyzed here and compared with the situation in *H. dujardini* better explain involved families, in particular of the DnaJ type. The high repair potential is supported by a comparison of protective protein clusters of tardigrades according to the COG classification system[42] to nematodes, man, and yeast as a control. We identify 46 tardigrade-specific adaptations including the just mentioned chaperones also involved in DNA protection[84] as well as 11 protein clusters shared with pathways and protein clusters in man. These include bmh2 (Fig. 5, Fig. 8), proteasome maturation factors and others. The maturation factor could in principle also be manipulated in man to improve stress tolerance, as proteasome inhibitory drugs are known from cancer therapy[85; 86]. However, to improve stress tolerance a suitable activation level is critical. DNA repair protein RHP57 (*RAD57* homolog of *Schizosaccharomyces pombe*) is another conserved and attractive target to improve stress

tolerance in man. It is a recombination repair gene[76] and shown here to be present both in tardigrades and man. DNA stability and repair is considered to be a key factor to lower cancer risk and reach high including very high age in man[87].

Conclusion

This study systematically analyzes sequence families and clusters apparent from tardigrade transcriptome data from ongoing EST efforts and their resulting proteins and pathways. The results show molecular functions involved in the unique stress-adaptation of tardigrades, species-specific differences and general features shared even with man, thereby providing a good basis for detailed functional studies.

Material and Methods

Tardigrade culture

Cultures of the cosmopolitan eutardigrade species *M. tardigradum* DOYÈRE 1849 (Apochela, Milnesidae) were kept and reared on petri dishes (ø 9.4 cm) filled with a small layer of agarose (3%) (peqGOLD Universal Agarose, peqLAB, Erlangen, Germany) and covered with spring water (Volvic™ water, Danone Waters Deutschland, Wiesbaden, Germany) at 20±2 °C and a light/dark cycle of 12 h. Rotifers *Philodina citrina* and nematodes *Panagrellus* sp. were provided as food source to adults, while, juvenile tardigrades were additionally fed with green algae *Chlorogonium elongatum*. For all experiments adult animals in good condition were taken from the culture and starved for three days. This prevents the preparation of additional RNA originating from incompletely digested food in the intestinal system.

Sequence resources

We obtained all ESTs of *M. tardigradum* from our ongoing sequencing project of *M. tardigradum.* For *H. dujardini* sequences we obtained all available EST sequences from GenBank[88]. Proteins were predicted using a BLASTX search against UniProtKB/SwissProt-,

UniProtKB/TrEMBL- and NR-database. The ORFs for nucleotide sequences, showing significant results (Evalue < 0.001), were extracted. To be more efficient only sequences without results were searched against a more extensive databases (UniProtKB/TrEMBL- and finally NR-database, respectively). For sequences without homology the longest ORF was extracted. All available new and existing sequence information (RNA, encoded proteins) was integrated on the tardigrade workbench. For this a new server was hosted at

`http://tardigrade.bioapps.biozentrum.uni-wuerzburg.de`. It includes now enhanced management capabilities, all new databases and rewritten source code.

Identification of regulatory elements in *M. tardigradum* mRNAs

Here, all known ESTs of *M. tardigradum* were systematically screened using the software UTRscan[89]. This software screens 30 regulatory elements for RNA regulation (translational elements, stability elements, other well known regulatory elements). The default settings for batch mode were used and all reported elements were collected.

PCR methods

Primers for different HSPs (HSP10: forward CCACTTCGCTACAAACAG, reverse GTGATGCCAATAGGAGTG, product length 171; HSP17.2: forward GTCGATGCAGTCAAACAG, reverse GCGTGCTCAGTTACTCTC, product length 99; HSP40: forward CTCTGGGAAGCCATTACT, reverse TGAGACTTGCTTCTCTCC, product length 119; HSP90: forward GTCGATGCAGTCAAACAG, reverse GCGTGCTCAGTTACTCTC, product length 99; HSP70: forward CCAAGTCTTCGAGTGATC, reverse TCTGCTCAGACAACAGTC, product length 84;) were used to validate EST data on *M. tardigradum* heat shock proteins by PCR. The used PCR protocol consists of a 8 minute denaturation step at 94 °C, 35 cycles of 30 seconds denatuaration (94 °C), 30 seconds annealing (56 °C) and 30 seconds elongation at 72 °C and a final elongation at 72 °C for 10 minutes in a reaction volume of 20 µl 1 µl of cDNA was added.

We used primer3 [90] for the design of the oligonucleotides.

Sequence and domain analysis.

A number of different iterative sequence alignment procedures and sequence analysis methods was used as described previously (Gaudermann et al., 2006) including gene context, interaction predictions, domain analysis and phylogenetic tests. Similar sequences were clustered by different methods, below CLANS clustering and InParanoid clustering are decribed in more detail, however other clusters considered included COG/KOG clustering, as well as containing specific domains evident by Pfam membership and different domain families according to Interpro and SMART.

Clustering using the CLANS method.

CLANS performed an all-against-all pairwise sequence comparisons using TBLASTX. Afterwards the similar proteins were clustered in 3D applying the CLANS method[23] with a cut-off E-value of 0.001. The program calculates pairwise attraction values based on the HSP P-values obtained from the BLAST-run. Finally the clusters were identified using convex clustering and NJ-trees with standard parameters.

Clustering using the Inparanoid method.

After all-against-all pairwise sequence comparisons using TBLASTX, the results of the comparison were imported into Inparanoid[22] for prediction of orthologs within the ESTs. As the EST sets cover a substantial fraction of the encoded proteins but do not yet represent the whole protein set of the compared tardigrades, we called the orthologs given by Inparanoid "shared candidate ortholog sequences" (scos) and the remaining sequences "candidate single sequence ESTs". The scos represent the lower bound of identified orthologs, as further identified proteins will certainly increase the number of orthologs between the two tardigrades.

Sequence comparisons to identify tardigrade proteins and stress pathways

We conducted systematic sequence to sequence comparisons starting from verified orthologous of a number of organisms (*Drosophila melanogaster*, *Caenorhabditis elegans*; different plants; fungi; as well as vertebrates) and used these to screen tardigrade proteins for similarity.

Good hits to stress proteins and best hits to LEA proteins were further examined for functionality by a number of sequence analysis methods including sensitive back-searches[91], PROSITE motifs and domain composition[59].

Furthermore, we include a large-scale mapping of all available sequence data from *M. tardigradum* and *H. dujardini* to clusters of orthologous groups found in complete prokaryotic genomes (COGs,[42]) or complete eukaryotic genomes (KOGs). This allows to directly predict individual protein functions and protein families by comparing the sequence to well established families from known genome sequences. The latest version (last major update in 2008) of the COG/KOG database was used[73].

Phylogenetic comparison of LEA proteins

LEA proteins from *H. dujardini* and *M. tardigradum* were collected after the annotation of ESTs. LEA sequences from other organisms were collected from NCBI. They were aligned and distances calculated applying ClustalW. Bootstrap support for the different branches were calculated. The final unrooted tree (Fig. 4) was drawn using the software FigTree (version 1.2.3; `http://tree.bio.ed.ac.uk/software/figtree/`).

Western Blotting

Total protein lysate from *M. tardigradum* (10 µg) was separated on a NuPAGE^{TM} 4–12% Bis-Tris mini gel (Invitrogen) using MES running buffer. A voltage of 200 was applied until the bromophenol blue front had reached the bottom of the gel (approx. 40 cm). Separated proteins were electro- transferred onto PVDF membrane for 1.5 h at maximum 50 mA (0.8/cm^2) in a semi-dry transfer unit (Hoefer^{TM} TE 77) using the following transfer solution: concentrated anode buffer (300 mM Tris-HCl pH 10.4), anode buffer (25 mM Tris-HCl pH 10.4) and

cathode buffer (25 mM Tris-HCl pH 9.4, 40 mM aminohexanoic acid). The PVDF membrane

was incubated in blocking buffer containing 5% non-fat milk, 0.1% Tween20 in PBS.

Detection of Hsp70 was carried out using anti Hsp70 (BD Biosciences Pharmingen) as

primary antibody and a horseradish peroxidase conjugated anti mouse IgG (GE-Healthcare) as

secondary antibody. For detection of LEA proteins we used Ari LEA Ab as primary antibody

and a horseradish peroxidase conjugated anti-rabbit IgG (GE-Healthcare) as secondary

antibody. For molecular mass determination of the target proteins on film ECL DualVue

marker (GE-Healthcare) was used. Immunoreaction was detected using the ECL Western

Blotting Detection kit (GE Healthcare). Images were acquired using an Image Scanner Model

UTA-1100 (Amersham Biosciences).

Supplementary data

Supplementary data is available at `http://tardigrade.bioapps.biozentrum.uni-wuerzburg.de`

Acknowledgements

References

1. **Wright, J. C. (1989). Desiccation Tolerance and Water-Retentive Mechanisms in Tardigrades. *J Exp Biol* 142, 267-292.**
2. **Jonsson, K. I., Harms-Ringdahl, M. & Torudd, J. (2005). Radiation tolerance in the eutardigrade Richtersius coronifer. *Int J Radiat Biol* 81, 649-656.**
3. **Crowe, L. M. (2002). Lessons from nature: the role of sugars in anhydrobiosis. *Comp Biochem Phys A* 131, 505-513.**
4. **Hoekstra, F. A., Wolkers, W. F., Buitink, J., Golovina, E. A., Crowe, J. H. & Crowe, L. M. (1997). Membrane stabilization in the dry state. *Comp Biochem Phys A* 117, 335-341.**
5. **Westh, P. & Ramlov, H. (1991). Trehalose Accumulation in the Tardigrade Adorybiotus-Coronifer during Anhydrobiosis. *J Exp Zool* 258, 303-311.**

6.   Clegg, J. S. (2005). Desiccation tolerance in encysted embryos of the animal extremophile, Artemia. *Integr Comp Biol* 45, 715-724.

7.   Goyal, K., Browne, J. A., Walton, L. J., Pinelli, C., Rastogi, R. K., Burnell, A. M. & Tunnacliffe, A. (2004). Molecular anhydrobiology: Identifying molecules implicated in invertebrate anhydrobiosis. *Integr Comp Biol* 44, 560-560.

8.   Ellis, R. J. & Vandervies, S. M. (1991). Molecular Chaperones. *Annu Rev Biochem* 60, 321-347.

9.   Feder, M. E. & Hofmann, G. E. (1999). Heat-shock proteins, molecular chaperones, and the stress response: Evolutionary and ecological physiology. *Annu Rev Physiol* 61, 243-282.

10.   Gething, M. J. & Sambrook, J. (1992). Protein Folding in the Cell. *Nature* 355, 33-45.

11.   Kiang, J. G. & Tsokos, G. C. (1998). Heat shock protein 70 kDa: Molecular biology, biochemistry, and physiology. *Pharmacol Therapeut* 80, 183-201.

12.   Browne, J., Tunnacliffe, A. & Burnell, A. (2002). Anhydrobiosis: plant desiccation gene found in a nematode. *Nature* 416, 38.

13.   Jacquier-Sarlin, M. R., Fuller, K., Dinh-Xuan, A. T., Richard, M. J. & Polla, B. S. (1994). Protective effects of hsp70 in inflammation. *Experientia* 50, 1031-8.

14.   Plumier, J. C. L., Ross, B. M., Currie, R. W., Angelidis, C. E., Kazlaris, H., Kollias, G. & Pagoulatos, G. N. (1995). Transgenic Mice Expressing the Human Heat-Shock Protein-70 Have Improved Postischemic Myocardial Recovery. *J Clin Invest* 95, 1854-1860.

15.   Schill, R. O., Steinbruck, G. H. & Kohler, H. R. (2004). Stress gene (hsp70) sequences and quantitative expression in Milnesium tardigradum (Tardigrada) during active and cryptobiotic stages. *J Exp Biol* 207, 1607-13.

16.   Jonsson, K. I. & Schill, R. O. (2007). Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade Richtersius coronifer. *Comp Biochem Physiol B Biochem Mol Biol* 146, 456-60.

17.   Hengherr, S., Worland, M. R., Reuner, A., Brummer, F. & Schill, R. O. (2009). Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades. *J Exp Biol* 212, 802-807.

18.   Hengherr, S., Worland, M., Reuner, A., Brümmer, F. & Schill, R. (2009). Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades. *Journal of Experimental Biology* 212, 802.

19.   Horikawa, D. D., Sakashita, T., Katagiri, C., Watanabe, M., Kikawada, T., Nakahara, Y., Hamada, N., Wada, S., Funayama, T., Higashi, S., Kobayashi, Y., Okuda, T. & Kuwabara, M. (2006). Radiation tolerance in the tardigrade Milnesium tardigradum. *Int J Radiat Biol* 82, 843-848.

20.   Gabriel, W. N. & Goldstein, B. (2007). Segmental expression of Pax3/7 and engrailed homologs in tardigrade development. *Dev Genes Evol* 217, 421-33.

21.   Förster, F., Liang, C., Shkumatov, A., Beisser, D., Engelmann, J. C., Schnölzer, M., Frohme, M., Müller, T., Schill, R. O. & Dandekar, T. (2009). Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades. *BMC Genomics* 10, 469.

22.   Remm, M., Storm, C. E. V. & Sonnhammer, E. L. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314, 1041-1052.

23.   Frickey, T. & Lupas, A. (2004). CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20, 3702-3704.

24. Ostareck-Lederer, A., Ostareck, D. H., Standart, N. & Thiele, B. J. (1994). Translation of 15-lipoxygenase mRNA is inhibited by a protein that binds to a repeated sequence in the 3' untranslated region. *Embo J* 13, 1476-81.

25. Lai, E. C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* 30, 363-4.

26. Parsch, J., Russell, J. A., Beerman, I., Hartl, D. L. & Stephan, W. (2000). Deletion of a conserved regulatory element in the Drosophila Adh gene leads to increased alcohol dehydrogenase activity but also delays development. *Genetics* 156, 219-27.

27. Chen, C. Y. & Shyu, A. B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci* 20, 465-70.

28. Chen, W. H., Ge, X., Wang, W., Yu, J. & Hu, S. (2009). A gene catalogue for post-diapause development of an anhydrobiotic arthropod Artemia franciscana. *BMC Genomics* 10, 52.

29. Fan, S., Meng, Q., Laterra, J. J. & Rosen, E. M. (2007). Ras effector pathways modulate scatter factor-stimulated NF-kappaB signaling and protection against DNA damage. *Oncogene* 26, 4774-96.

30. Xiao, K., Jehle, F., Peters, C., Reinheckel, T., Schirmer, R. H. & Dandekar, T. (in press). The set of CA/C1 peptidases of the malaria parasites Plasmodium falciparum and P. berghei and their mammalian hosts - a bioinformatical analysis. *Biol. Chem.*

31. Nishiyama, K., Konishi, A., Nishio, C., Araki-Yoshida, K., Hatanaka, H., Kojima, M., Ohmiya, Y., Yamada, M. & Koshimizu, H. (2005). Expression of cystatin C prevents oxidative stress-induced death in PC12 cells. *Brain Res Bull* 67, 94-9.

32. Raikhel, N. V., Lee, H. I. & Broekaert, W. F. (1993). Structure and Function of Chitin-Binding Proteins. *Annu Rev Plant Phys* 44, 591-615.

33. Tjoelker, L. W., Gosting, L., Frey, S., Hunter, C. L., Trong, H. L., Steiner, B., Brammer, H. & Gray, P. W. (2000). Structural and functional definition of the human chitinase chitin-binding domain. *J Biol Chem* 275, 514-20.

34. Guo, Y., Breeden, L. L., Zarbl, H., Preston, B. D. & Eaton, D. L. (2005). Expression of a human cytochrome p450 in yeast permits analysis of pathways for response to and repair of aflatoxin-induced DNA damage. *Mol Cell Biol* 25, 5823-33.

35. Furuchi, T., Hwang, G. W. & Naganuma, A. (2002). Overexpression of the ubiquitin-conjugating enzyme Cdc34 confers resistance to methylmercury in Saccharomyces cerevisiae. *Mol Pharmacol* 61, 738-41.

36. Madura, K., Prakash, S. & Prakash, L. (1990). Expression of the Saccharomyces cerevisiae DNA repair gene RAD6 that encodes a ubiquitin conjugating enzyme, increases in response to DNA damage and in meiosis but remains constant during the mitotic cell cycle. *Nucleic Acids Res* 18, 771-8.

37. Inoue, K., Takano, H., Shimada, A. & Satoh, M. (2009). Metallothionein as an anti-inflammatory mediator. *Mediators Inflamm* 2009, 101659.

38. Gidda, S. K. & Varin, L. (2006). Biochemical and molecular characterization of flavonoid 7-sulfotransferase from Arabidopsis thaliana. *Plant Physiol Biochem* 44, 628-36.

39. Barnett, M. E., Madgwick, D. K. & Takemoto, D. J. (2007). Protein kinase C as a stress sensor. *Cell Signal* 19, 1820-9.

40. Huang, Y., Xiao, B. & Xiong, L. (2007). Characterization of a stress responsive proteinase inhibitor gene with positive effect in improving drought resistance in rice. *Planta* 226, 73-85.

41. Hasona, A., Crowley, P. J., Levesque, C. M., Mair, R. W., Cvitkovitch, D. G., Bleiweis, A. S. & Brady, L. J. (2005). Streptococcal viability and diminished

stress tolerance in mutants lacking the signal recognition particle pathway or YidC2. *Proc Natl Acad Sci U S A* 102, 17466-71.

42. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278, 631-637.

43. Sollner, T. H. (2003). Regulated exocytosis and SNARE function (Review). *Mol Membr Biol* 20, 209-20.

44. Jou, Y., Chiang, C. P., Jauh, G. Y. & Yen, H. E. (2006). Functional characterization of ice plant SKD1, an AAA-type ATPase associated with the endoplasmic reticulum-Golgi network, and its role in adaptation to salt stress. *Plant Physiol* 141, 135-46.

45. Doyle, V., Virji, S. & Crompton, M. (1999). Evidence that cyclophilin-A protects cells against oxidative stress. *Biochem J* 341 (Pt 1), 127-32.

46. Dombrowski, J. E., Baldwin, J. C. & Martin, R. C. (2008). Cloning and characterization of a salt stress-inducible small GTPase gene from the model grass species Lolium temulentum. *J Plant Physiol* 165, 651-61.

47. Sharma, A. K., Nelson, M. C., Brandt, J. E., Wessman, M., Mahmud, N., Weller, K. P. & Hoffman, R. (2001). Human CD34(+) stem cells express the hiwi gene, a human homologue of the Drosophila gene piwi. *Blood* 97, 426-34.

48. Laursen, L. V., Ampatzidou, E., Andersen, A. H. & Murray, J. M. (2003). Role for the fission yeast RecQ helicase in DNA repair in G2. *Mol Cell Biol* 23, 3692-705.

49. Shaked, H., Avivi-Ragolsky, N. & Levy, A. A. (2006). Involvement of the Arabidopsis SWI2/SNF2 chromatin remodeling gene family in DNA damage response and recombination. *Genetics* 173, 985-94.

50. Neumann, S., Reuner, A. & Brümmer, F. (2009). DNA damage in storage cells of anhydrobiotic tardigrades. *Comp Biochem Phys A* 153, 425-429.

51. Li, B., Li, F., Puskar, K. M. & Wang, J. H. (2009). Spatial patterning of cell proliferation and differentiation depends on mechanical stress magnitude. *J Biomech* 42, 1622-7.

52. Lafontaine, D. L. & Tollervey, D. (1999). Nop58p is a common component of the box C+D snoRNPs that is required for snoRNA stability. *Rna* 5, 455-67.

53. Tunnacliffe, A. & Wise, M. J. (2007). The continuing conundrum of the LEA proteins. *Naturwissenschaften* 94, 791-812.

54. Furuyashiki, T. & Narumiya, S. (2009). Roles of prostaglandin E receptors in stress responses. *Curr Opin Pharmacol* 9, 31-8.

55. Nakamura, A., Shimada, H., Masuda, T., Ohta, H. & Takamiya, K. (2001). Two distinct isopentenyl diphosphate isomerases in cytosol and plastid are differentially induced by environmental stresses in tobacco. *FEBS Lett* 506, 61-4.

56. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B. & Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631-636.

57. Soto, T., Fernandez, J., Cansado, J., VicenteSoler, J. & Gacto, M. (1997). Protein kinase Sck1 is involved in trehalase activation by glucose and nitrogen source in the fission yeast Schizosaccharomyces pombe. *Microbiol-Uk* 143, 2457-2463.

58. Hengherr, S., Heyer, A. G., Kohler, H. R. & Schill, R. O. (2008). Trehalose and anhydrobiosis in tardigrades - evidence for divergence in responses to dehydration. *Febs J* 275, 281-288.

59. Gaudermann, P., Vogl, I., Zientz, E., Silva, F. J., Moya, A., Gross, R. & Dandekar, T. (2006). Analysis of and function predictions for previously conserved hypothetical or putative proteins in Blochmannia floridanus. *Bmc Microbiol* 6, -.

60. Alpert, P. (2006). Constraints of tolerance: why are desiccation-tolerant organisms so small or rare? *J Exp Biol* 209, 1575-84.

61. Melgar, J. C., Guidi, L., Remorini, D., Agati, G., Degl'innocenti, E., Castelli, S., Camilla Baratto, M., Faraloni, C. & Tattini, M. (2009). Antioxidant defences and oxidative damage in salt-treated olive plants under contrasting sunlight irradiance. *Tree Physiol*.

62. Sanchez-Estudillo, L., Freile-Pelegrin, Y., Rivera-Madrid, R., Robledo, D. & Narvaez-Zapata, J. A. (2006). Regulation of two photosynthetic pigment-related genes during stress-induced pigment formation in the green alga, Dunaliella salina. *Biotechnol Lett* 28, 787-91.

63. Ursic, D., Chinchilla, K., Finkel, J. S. & Culbertson, M. R. (2004). Multiple protein/protein and protein/RNA interactions suggest roles for yeast DNA/RNA helicase Sen1p in transcription, transcription-coupled DNA repair and RNA processing. *Nucleic Acids Res* 32, 2441-52.

64. Loprasert, S., Vattanaviboon, P., Praituan, W., Chamnongpol, S. & Mongkolsuk, S. (1996). Regulation of the oxidative stress protective enzymes, catalase and superoxide dismutase in Xanthomonas--a review. *Gene* 179, 33-7.

65. Chepelinsky, A. B. (2003). The ocular lens fiber membrane specific protein MIP/Aquaporin 0. *J Exp Zoolog A Comp Exp Biol* 300, 41-6.

66. Maurel, C. (2007). Plant aquaporins: novel functions and regulation properties. *FEBS Lett* 581, 2227-36.

67. Yang, S. & Cui, L. (2009). The action of aquaporins in cell elongation, salt stress and photosynthesis. *Sheng Wu Gong Cheng Xue Bao* 25, 321-7.

68. Hong-Bo, S., Zong-Suo, L. & Ming-An, S. (2005). LEA proteins in higher plants: structure, function, gene expression and regulation. *Colloids Surf B Biointerfaces* 45, 131-5.

69. Liu, Y. & Zheng, Y. (2005). PM2, a group 3 LEA protein from soybean, and its 22-mer repeating region confer salt tolerance in Escherichia coli. *Biochem Biophys Res Commun* 331, 325-32.

70. Morimoto, R. I., Kline, M. P., Bimston, D. N. & Cotto, J. J. (1997). The heat-shock response: regulation and function of heat-shock proteins and molecular chaperones. *Essays Biochem* 32, 17-29.

71. Tessmer, I., Yang, Y., Zhai, J., Du, C., Hsieh, P., Hingorani, M. M. & Erie, D. A. (2008). Mechanism of MutS searching for DNA mismatches and signaling repair. *J Biol Chem* 283, 36646-54.

72. Dandekar, T. & Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J Mol Biol* 256, 645-60.

73. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L. & Yaschenko, E. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 35, D5-D12.

74. Orr-Weaver, T. L., Szostak, J. W. & Rothstein, R. J. (1981). Yeast transformation: a model system for the study of recombination. *Proc Natl Acad Sci U S A* **78**, 6354-8.

75. Schill, R. O., Neumann, S., Reuner, A. & Brümmer, F. (2008). Detection of DNA damage with single-cell gel electrophoresis in anhydrobiotic tardigrades. *Comparative Biochemistry and Physiology - Part A: Molecular & Integrative Physiology* **151**, S32-S32.

76. Tsutsui, Y., Morishita, T., Iwasaki, H., Toh, H. & Shinagawa, H. (2000). A recombination repair gene of Schizosaccharomyces pombe, rhp57, is a functional homolog of the Saccharomyces cerevisiae RAD57 gene and is phylogenetically related to the human XRCC3 gene. *Genetics* **154**, 1451-61.

77. Robinson, P. A. & Ardley, H. C. (2004). Ubiquitin-protein ligases. *J Cell Sci* **117**, 5191-4.

78. Fricke, B., Heink, S., Steffen, J., Kloetzel, P. M. & Kruger, E. (2007). The proteasome maturation protein POMP facilitates major steps of 20S proteasome formation at the endoplasmic reticulum. *EMBO Rep* **8**, 1170-5.

79. Nathan, D. F., Vos, M. H. & Lindquist, S. (1997). In vivo functions of the Saccharomyces cerevisiae Hsp90 chaperone. *Proc Natl Acad Sci U S A* **94**, 12949-56.

80. Zhao, R., Davey, M., Hsu, Y. C., Kaplanek, P., Tong, A., Parsons, A. B., Krogan, N., Cagney, G., Mai, D., Greenblatt, J., Boone, C., Emili, A. & Houry, W. A. (2005). Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone. *Cell* **120**, 715-27.

81. Lai, E. C., Tam, B. & Rubin, G. M. (2005). Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes Dev* **19**, 1067-80.

82. Jonsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M. & Rettberg, P. (2008). Tardigrades survive exposure to space in low Earth orbit. *Curr Biol* **18**, R729-R731.

83. Setlow, P. (2007). I will survive: DNA protection in bacterial spores. *Trends Microbiol* **15**, 172-80.

84. Yamamoto, T., Mori, Y., Ishibashi, T., Uchiyama, Y., Ueda, T., Ando, T., Hashimoto, J., Kimura, S. & Sakaguchi, K. (2005). Interaction between proliferating cell nuclear antigen (PCNA) and a DnaJ induced by DNA damage. *J Plant Res* **118**, 91-7.

85. McConkey, D. J. & Zhu, K. (2008). Mechanisms of proteasome inhibitor action and resistance in cancer. *Drug Resist Updat* **11**, 164-79.

86. Yang, H., Zonder, J. A. & Dou, Q. P. (2009). Clinical development of novel proteasome inhibitors for cancer treatment. *Expert Opin Investig Drugs* **18**, 957-71.

87. Arking, R. (2006). *The biology of aging observations and principles*. 3rd edit, Oxford University Press, Oxford ; New York.

88. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2009). GenBank. *Nucleic Acids Res* **37**, D26-31.

89. Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P. J., Duarte, J., Saccone, C. & Pesole, G. (2005). UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res* **33**, D141-6.

90. Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386.

91. Bork, P. & Gibson, T. J. (1996). Applying motif and profile searches. *Methods Enzymol* 266, 162-84.

92. Reuner, A., Hengherr, S., Mali, B., Förster, F., Arndt, D., Reinhardt, R., Dandekar, T., Frohme, M., Brummer, F. & Schill, R. O. (2009). Stress response in tardigrades: differential gene expression of molecular chaperones. *Cell Stress Chaperones*.

93. Mtwisha, L., Brandt, W., McCready, S. & Lindsey, G. G. (1998). HSP 12 is a LEA-like protein in Saccharomyces cerevisiae. *Plant Mol Biol* 37, 513-521.

94. Schuller, C., Brewster, J. L., Alexander, M. R., Gustin, M. C. & Ruis, H. (1994). The Hog Pathway Controls Osmotic Regulation of Transcription Via the Stress-Response Element (Stre) of the Saccharomyces-Cerevisiae Ctt1 Gene. *Embo J* 13, 4382-4389.

95. White, C. N. & Rivin, C. J. (1995). Sequence and Regulation of a Late Embryogenesis Abundant Group-3 Protein of Maize. *Plant Physiol* 108, 1337-1338.

96. Espelund, M., Debedout, J. A., Outlaw, W. H. & Jakobsen, K. S. (1995). Environmental and Hormonal-Regulation of Barley Late-Embryogenesis-Abundant (Lea) Messenger-Rnas Is Via Different Signal-Transduction Pathways. *Plant Cell Environ* 18, 943-949.

97. Finkelstein, R. R. (1993). Abscisic Acid-Insensitive Mutations Provide Evidence for Stage-Specific Signal Pathways Regulating Expression of an Arabidopsis Late Embryogenesis-Abundant (Lea) Gene. *Mol Gen Genet* 238, 401-408.

98. Suzuki, M., Ketterling, M. G., Li, Q. B. & McCarty, D. R. (2003). Viviparous1 alters global gene expression patterns through regulation of abscisic acid signaling. *Plant Physiol* 132, 1664-1677.

99. Immenschuh, S. & Baumgart-Vogt, E. (2005). Peroxiredoxins, oxidative stress, and cell proliferation. *Antioxid Redox Sign* 7, 768-777.

Figure Captions

Fig. 1. Sequence similar protein clusters in *Milnesium tardigradum* and *Hypsibius dujardini.*

Top: *Shared candidate orthologue sequences (sCOS), present in the* H. dujardini *and* M. tardigradum *EST.* The Venn diagram illustrates the results given by Inparanoid. The left and the right sequence sets show the candidate single sequence ESTs for *Milnesium tardigradum* and *Hypsibius dujardini*. Between the two sCOS-sets are the shared candidate ortholog sequences (sCOS) from *Milnesium tardigradum* (left) and *Hypsibius dujardini* (right). The number of shared orthologs is given in brackets, above are given the number of sCOS for both species. For a complete listing see supplementary table III on tardigrade website.

Bottom: Number of clusters containing no less than ten sequence similar proteins according to CLANS clustering. The top 5 clusters for *Milnesium tardigradum* are: 1- Cytochrome p450 like (29 proteins), 2- no homology (26 proteins), 3- Cathepsin like (18 proteins), 4- Rab subfamily like (15 proteins), 5- Ras subfamily like (16 proteins). Top 5 clusters for *H. dujardini* are: 1- no homology (28 proteins), 2- Chitin binding protein like (23 proteins), 3- Ras related proteins (23 proteins), 4- Cystatin like (22 proteins), 5-Glutathione S-transferase like (20 proteins). Top 5 clusters shared by both tardigrades are: 1- no homology (52 proteins), 2- Ras subfamily like (46 proteins), 3- Cathepsin like (38 proteins), 4- Chitin binding protein like (36 proteins), 5-Cystatin like (35 proteins). Details and all found clusters are given in supplementary material (S-Table I a-c) on the tardigrade website.

Fig. 2. PCR validation of heat shock protein transcripts in *Milnesium tardigradum* and *H. dujardini.* Shown are detected PCR bands using heat shock protein encoding EST primers as described in M&M. Lanes: (1) Hyperladder IV (Bioline); (2) hsp10 (expected: 171 bp); (3) hsp10 Non template control (NTC); (4) hsp20/30 (99 bp); (5) hsp20/30 NTC; (6) hsp40 (119 bp); (7) hsp40 NTC; (8)hsp70 (84 bp); (9) hsp70 NTC; (10) hsp90 (99 bp); (11) hsp90 NTC; (12) Hyperladder IV (Bioline).

Fig. 3. Western blot of heat shock and LEA proteins in *Milnesium tardigradum*. The presence of Hsp70 (middle panel) and LEA proteins (right panel) in *Milnesium tardigradum* was confirmed by Western blot analysis using the corresponding antibodies. The ECL DualVue size marker is shown on the left.

Fig. 4. Phylogenetic tree of different tardigrade LEA-like proteins. Besides several LEA protein tardigrade sequences from *Hypsibius dujardini* (13 proteins, red) and *Richtersius coniferi* (3 sequences, orange) all sequences from the seed alignments for LEA1 (blue), LEA2 (grey) and LEA4 (green) profile from PFAM are shown. LEA groups 3, 5, 6[53] are even less related. Bootstrap values were tested but are not shown here to prevent cluttering. The tree is generated by Figtree (version 1.2.3).

Fig. 5. Bmh2 stress pathway is conserved in tardigrades. The Bmh2 pathway was shown to exist not only in vertebrates but also in yeast[56]. It is also conserved in *Hypsibius dujardini* and *Milnesium tardigradum.* Key regulators (bmh2, Sck1) as well as depending metabolic enzymes are indicated together with genbank identifiers (top) comparing man, yeast and *Hypsibius dujardini*. There are sequence similar candidate sequences for Lcb1 and Lcb2 in *Hypsibius dujardini*, but it is not known which is regulated by bmh2 or Sck1. Trehalase has not been found (question mark). In *Milnesium tardigradum* only the two regulators, bmh2 and Sck1 have been identified.

Fig. 6. Comparison of COGs/KOGs for *M. tardigradum, H. dujardini* (Tardigrades), *H. sapiens* and *C. elegans* (Nematodes). The Venn diagram depicts intersections of COGs/KOGs between the two desiccation resistant phyla tardigrades and nematodes (*C. elegans*) versus *H. sapiens.* The numbers in brackets indicate COGs/KOGs that were shown to be also present in rotifers. For the comparison only COGs/KOGs were used that belong to the functional groups [L]: Replication, recombination and repair, [V]: Defense mechanisms or [O]: Posttranslational modification, protein turnover, chaperones. The COGs/KOGs with descriptions can be found in supplementary Table VII on our tardigrade website.

Fig. 7. Comparison of COGs/KOGs specific for *M. tardigradum, H. dujardini* **(Tardigrades) and** *S. cerevisiae and S. pombe* **(Yeast).** The Venn diagram compares the specific COGs/KOGs only found in tardigrades to unicellular clusters of the yeasts *S. cerevisiae* and *S. pombe*. For the comparison only COGs/KOGs were used that belong to the functional groups [L]: Replication, recombination and repair, [V]: Defense mechanisms or [O]: Posttranslational modification, protein turnover, chaperones. The COGs/KOGs with descriptions can be found in supplementary Table VII on our tardigrade website.

Fig. 8. Tardigrade stress adaptation: Which are specific, which are shared? Depicted are predicted tardigrade specific protein families in the center of the diagram and protein families shared with other organisms (Top: Man and nematodes; left: yeast (*S. pombe*, *S. cerevisiae*; right: rotifers; bottom: man, not in nematodes (NB: 14-3-3-like proteins occur in *C. elegans*, but their relation to the bmh2 pathway is not clear).

Tables

Table I. Regulatory Motifs found by UTR-Scan in 3264 *Milnesium tardigradum* unique genes:

Motif	Number of hits (Number of unique genes)
15-LOX-DICE	224 (218)
ADH_DRE	53 (52)
Brd-Box	107 (106)
CPE	37 (37)
GY-Box	98 (96)
IRE	4 (4)
IRES	1029 (1029)
K-Box	315 (295)
SECIS-1	4 (4)
SECIS-2	18 (18)
TGE	4 (4)
TOP	6 (6)

Table IIa. Subset of important COGs/KOGs identified uniquely in *Hybsibius dujardini*.
Overview: The entire set contains 44 COGs and 134 KOGs with the following number of COG-classes: 2[A] 1[AD] 26[C] 1[D] 11[E] 4[F] 3[G] 1[GM] 1[GO] 4[H] 7[I] 1[IG] 1[IO] 1[IR] 24[J] 1[K] 4[M] 1[N] 1[NU] 26[O] 2[P] 1[Q] 18[R] 1[RP] 8[S] 11[T] 1[TUZ] 1[TZ] 7[U] 1[V] 2[W] 4[Z]. COGs/KOGs with two or more letters are assigned to several functional classifications. The complete table can be found in the supplementary table IIa on our tardigrade website.

KOG4582:[R]Uncharacterized conserved protein, contains ZZ-type Zn-finger (9)
KOG0712:[O]Molecular chaperone (DnaJ superfamily) (8)
KOG0714:[O]Molecular chaperone (DnaJ superfamily) (8)
KOG0715:[O]Molecular chaperone (DnaJ superfamily) (8)
KOG0102:[O]Molecular chaperones mortalin/PBP74/GRP75, HSP70 superfamily (7)
KOG0691:[O]Molecular chaperone (DnaJ superfamily) (7)
KOG2835:[F]Phosphoribosylamidoimidazole-succinocarboxamide synthase (6)
COG0152:[F]Phosphoribosylaminoimidazolesuccinocarboxamide (SAICAR) synthase (5)
KOG0187:[J]40S ribosomal protein S17 (5)
COG0089:[J]Ribosomal protein L23 (4)
COG2007:[J]Ribosomal protein S8E (4)
COG0092:[J]Ribosomal protein S3 (3)
KOG1714:[J]60s ribosomal protein L18 (3)
KOG1768:[J]40s ribosomal protein S26 (3)
COG0355:[C]F0F1-type ATP synthase, epsilon subunit (mitochondrial delta subunit) (2)
COG1156:[C]Archaeal/vacuolar-type H+-ATPase subunit B (2)
COG1997:[J]Ribosomal protein L37AE/L43A (2)
KOG0300:[S]WD40 repeat-containing protein (2)
KOG0310:[S]Conserved WD40 repeat-containing protein (2)
KOG1235:[R]Predicted unusual protein kinase (2)
KOG3458:[C]NADH:ubiquinone oxidoreductase, NDUFA8/PGIV/19 kDa subunit (2)

Table IIb. Important COGs/KOGs identified uniquely in *Milnesium tardigradum*.

Overview: The entire set contains 123 COGs and 497 KOGs with the following number of COG-classes: 13[A] 3[AJ] 6[B] 3[BK] 50[C] 1[CIQ] 14[D] 1[DL] 1[DN] 1[DO] 1[DR] 1[DZ] 14[E] 2[EI] 1[ET] 10[F] 34[G] 1[GMO] 10[H] 1[HE] 24[I] 1[IE] 1[IOVE] 1[IT] 62[J] 30[K] 2[KL] 1[KLB] 1[KT] 16[L] 1[LR] 7[M] 1[MJ] 1[N] 46[O] 1[ODR] 1[OK] 1[OKT] 1[OUT] 4[P] 3[PET] 1[PQ] 5[Q] 2[QR] 67[R] 1RD 1[RTU] 63[S] 43[T] 1[TDK] 1[TR] 1[TU] 2[TV] 1[TZ] 38[U] 1[UO] 1[UR] 7[V] 1[WT] 9[Z]. COGs with two or more letters are assigned to several functional classifications. The complete list of all Milnesium specific COGs/KOGs can be found in the supplementary table IIb on our tardigrade website.

KOG0660:[T]Mitogen-activated protein kinase (12)
KOG0666:[K]Cyclin C-dependent kinase CDK8 (10)
KOG3381:[S]Uncharacterized conserved protein (6)
KOG0472:[S]Leucine-rich repeat protein (5)
KOG0473:[S]Leucine-rich repeat protein (5)
COG0271:[T]Stress-induced morphogen (activity unknown) (4)
KOG1042:[D]Germ-line stem cell division protein Hiwi/Piwi; negative developmental regulator (4)
KOG1685:[S]Uncharacterized conserved protein (4)
KOG2443:[S]Uncharacterized conserved protein (4)
KOG2989:[S]Uncharacterized conserved protein (4)
KOG3618:[R]Adenylyl cyclase (4)
KOG3737:[O]Predicted polypeptide N-acetylgalactosaminyltransferase (4)
KOG3738:[O]Predicted polypeptide N-acetylgalactosaminyltransferase (4)
COG0045:[C]Succinyl-CoA synthetase, beta subunit (2)
COG0074:[C]Succinyl-CoA synthetase, alpha subunit (2)
COG0386:[O]Glutathione peroxidase (2)
COG0553:[KL]Superfamily II DNA/RNA helicases, SNF2 family (2)
COG1278:[K]Cold shock proteins (2)
COG1758:[K]DNA-directed RNA polymerase, subunit K/omega (2)
COG1761:[K]DNA-directed RNA polymerase, subunit L (2)
COG2012:[K]DNA-directed RNA polymerase, subunit H, RpoH/RPB5 (2)
COG4973:[L]Site-specific recombinase XerC (2)
COG4974:[L]Site-specific recombinase XerD (2)
KOG0119:[A]Splicing factor 1/branch point binding protein (RRM superfamily) (2)
KOG0272:[A]U4/U6 small nuclear ribonucleoprotein Prp4 (contains WD40 repeats) (2)
KOG0388:[L]SNF2 family DNA-dependent ATPase (2)
KOG0389:[B]SNF2 family DNA-dependent ATPase (2)
KOG0391:[R]SNF2 family DNA-dependent ATPase (2)
KOG0392:[K]SNF2 family DNA-dependent ATPase domain-containing protein (2)
KOG0557:[C]Dihydrolipoamide acetyltransferase (2)
KOG0558:[C]Dihydrolipoamide transacylase (alpha-keto acid dehydrogenase E2 subunit) (2)
KOG0559:[C]Dihydrolipoamide succinyltransferase (2-oxoglutarate dehydrogenase, E2 subunit) (2)
KOG0607:[T]MAP kinase-interacting kinase and related serine/threonine protein kinases (2)
KOG0653:[D]Cyclin B and related kinase-activating proteins (2)
KOG0654:[D]G2/Mitotic-specific cyclin A (2)
KOG0655:[D]G1/S-specific cyclin E (2)
KOG0656:[D]G1/S-specific cyclin D (2)
KOG0698:[T]Serine/threonine protein phosphatase (2)
KOG0699:[T]Serine/threonine protein phosphatase (2)
KOG0870:[K]DNA polymerase epsilon, subunit D (2)
KOG1002:[L]Nucleotide excision repair protein RAD16 (2)
KOG1088:[S]Uncharacterized conserved protein (2)
KOG1435:[IT]Sterol reductase/lamin B receptor (2)
KOG1679:[I]Enoyl-CoA hydratase (2)
KOG1680:[I]Enoyl-CoA hydratase (2)
KOG1681:[I]Enoyl-CoA isomerase (2)
KOG1716:[V]Dual specificity phosphatase (2)
KOG1717:[V]Dual specificity phosphatase (2)
KOG1718:[V]Dual specificity phosphatase (2)
KOG1766:[R]Enhancer of rudimentary (2)
KOG1863:[O]Ubiquitin carboxyl-terminal hydrolase (2)
KOG1864:[O]Ubiquitin-specific protease (2)
KOG1868:[O]Ubiquitin C-terminal hydrolase (2)
KOG1873:[O]Ubiquitin-specific protease (2)

Table IIc. **Subset of important COGs/KOGs identified both in *Hybsibius dujardini* and *Milnesium tardigradum*.** The entire set contains 129 COGs and 606 KOGs with the following number of COG-classes: 42[A] 2[AJ] 1[AR] 6[B] 1[BK] 1[BL] 51[C] 1[CD] 1[CP] 1[CR] 9[D] 1[DKL] 2[DO] 1[DR] 2[DT] 1[DZ] 9[E] 1[EG] 1[EM] 5[F] 1[FGR] 10[G] 1[GT] 6[H] 1[HC] 19[I] 1[IE] 1[IQR] 1[IT] 1[IU] 127[J] 1[JD] 21[K] 4[L] 1[LKJ] 2[M] 2[MG] 129[O] 3[OC] 1[OE] 1[OR] 1[OT] 1[OUT] 12[P] 1[PT] 15[Q] 1[QI] 72[R] 1[RT] 1[RTKL] 1[RV] 17[S] 57[T] 2[TR] 2[TU] 1[TUZ] 2[TZ] 1[TZR] 38[U] 1[UR] 1[UT] 1[UZ] 6[V] 26[Z] 1[ZD]. COGs/KOGs with two or more letters are assigned to several functional classifications. The complete table can be found in the supplementary table IIc on our tardigrade website.

COG5272:[O]Ubiquitin (31/16)
KOG0001:[OR]Ubiquitin and ubiquitin-like proteins (31/16)
KOG0003:[J]Ubiquitin/60s ribosomal protein L40 fusion (31/16)
KOG0004:[J]Ubiquitin/40S ribosomal protein S27a fusion (31/16)
KOG0005:[DO]Ubiquitin-like protein (31/16)
KOG0881:[O]Cyclophilin type peptidyl-prolyl cis-trans isomerase (6/12)
COG0149:[G]Triosephosphate isomerase (3/2)
KOG0266:[R]WD40 repeat-containing protein (3/5)
KOG0329:[A]ATP-dependent RNA helicase (3/8)
COG0451:[MG]Nucleoside-diphosphate-sugar epimerases (3/4)
KOG0543:[O]FKBP-type peptidyl-prolyl cis-trans isomerase (3/1)
KOG0549:[O]FKBP-type peptidyl-prolyl cis-trans isomerase (3/2)
KOG0730:[O]AAA+-type ATPase (3/10)
KOG0332:[A]ATP-dependent RNA helicase (2/6)
KOG0336:[A]ATP-dependent RNA helicase (2/6)
KOG0727:[O]26S proteasome regulatory complex, ATPase RPT3 (2/10)
KOG0658:[G]Glycogen synthase kinase-3 (1/8)
KOG0731:[O]AAA+-type ATPase containing the peptidase M41 domain (1/9)
KOG0734:[O]AAA+-type ATPase containing the peptidase M41 domain (1/8)
KOG0735:[O]AAA+-type ATPase (1/8)
KOG0736:[O]Peroxisome assembly factor 2 containing the AAA+-type ATPase domain (1/4)
KOG0737:[O]AAA+-type ATPase (1/6)
KOG0739:[O]AAA+-type ATPase (1/4)
KOG0751:[C]Mitochondrial aspartate/glutamate carrier protein Aralar/Citrin (contains EF-hand Ca2+-binding domains) (1/1)
KOG0923:[A]mRNA splicing factor ATP-dependent RNA helicase (1/2)
KOG0924:[A]mRNA splicing factor ATP-dependent RNA helicase (1/2)
KOG0925:[A]mRNA splicing factor ATP-dependent RNA helicase (1/2)
KOG0934:[U]Clathrin adaptor complex, small subunit (1/6)
KOG0935:[U]Clathrin adaptor complex, small subunit (1/6)
KOG0936:[U]Clathrin adaptor complex, small subunit (1/6)
KOG0983:[T]Mitogen-activated protein kinase (MAPK) kinase MKK7/JNKK2 (1/2)
KOG2733:[S]Uncharacterized membrane protein (1/2)
KOG3098:[S]Uncharacterized conserved protein (1/2)
KOG4431:[R]Uncharacterized protein, induced by hypoxia (1/2)
KOG4604:[S]Uncharacterized conserved protein (1/2)

Table III. Enzymes in metabolic pathways predicted to be present in *Hybsibius dujardini* and *Milnesium tardigradum* according to the EST data[i]

Hybsibius dujardini	*Milnesium tardigradum*
Glycolysis / Gluconeogenesis	**Glycolysis / Gluconeogenesis**
ec:1.2.4.1 pyruvate dehydrogenase (acetyl-transferring)	ec:1.2.1.12 glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)
ec:1.1.1.2 alcohol dehydrogenase (NADP+)	ec:5.3.1.1 triose-phosphate isomerase
ec:2.7.2.3 phosphoglycerate kinase	ec:6.2.1.1 acetate---CoA ligase
ec:5.3.1.1 triose-phosphate isomerase	ec:1.1.1.1 alcohol dehydrogenase
ec:4.1.2.13 fructose-bisphosphate aldolase	ec:5.4.2.2 phosphoglucomutase
ec:1.2.1.12 glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	
	ec:1.1.1.27 L-lactate dehydrogenase
ec:1.1.1.27 L-lactate dehydrogenase	ec:1.2.1.3 aldehyde dehydrogenase (NAD+)
	ec:4.2.1.11 phosphopyruvate hydratase
	ec:1.1.1.2 alcohol dehydrogenase (NADP+)
	ec:5.3.1.9 glucose-6-phosphate isomerase
Starch and sucrose metabolism	**Starch and sucrose metabolism**
ec:3.6.1.-	ec:3.6.1.-
ec:2.4.1.11 glycogen(starch) synthase	ec:3.2.1.20 alpha-glucosidase
ec:2.7.7.9 UTP---glucose-1-phosphate uridylyltransferase	ec:3.2.1.39 glucan endo-1,3-beta-D-glucosidase
ec:4.1.1.35 UDP-glucuronate decarboxylase	ec:3.2.1.1 alpha-amylase
ec:3.2.1.4 cellulase	ec:5.4.2.2 phosphoglucomutase
	ec:2.7.1.106 glucose-1,6-bisphosphate synthase
ec:3.2.1.1 alpha-amylase	ec:5.3.1.9 glucose-6-phosphate isomerase
Fatty acid biosynthesis	**Fatty acid biosynthesis**
ec:3.1.2.-	ec:3.1.2.-
	ec:1.3.1.-
Arachidonic acid metabolism	**Arachidonic acid metabolism**
ec:5.3.99.2 prostaglandin-D synthase	ec:1.14.14.1 unspecific monooxygenase
ec:1.14.14.1 unspecific monooxygenase	ec:3.1.1.4 phospholipase A2
ec:3.1.1.4 phospholipase A2	ec:5.3.99.2 prostaglandin-D synthase
ec:1.13.11.-	ec:3.3.2.6 leukotriene-A4 hydrolase
ec:3.3.2.10 soluble epoxide hydrolase	ec:1.11.1.9 glutathione peroxidase
	ec:2.3.2.2 gamma-glutamyltransferase
	ec:5.3.99.5 thromboxane-A synthase
Aminophosphonate metabolism	**Aminophosphonate metabolism**
ec:2.1.1.-	ec:2.1.1.-
Fatty acid elongation in mitochondria	**Fatty acid elongation in mitochondria**
	ec:1.1.1.35 3-hydroxyacyl-CoA dehydrogenase
ec:2.3.1.16 acetyl-CoA C-acyltransferase	

ec:3.1.2.22 palmitoyl[protein] hydrolase
ec:1.1.1.35 3-hydroxyacyl-CoA dehydrogenase

ec:4.2.1.17 enoyl-CoA hydratase

Fatty acid metabolism
ec:1.14.14.1 unspecific monooxygenase

ec:2.3.1.16 acetyl-CoA C-acyltransferase
ec:1.1.1.35 3-hydroxyacyl-CoA dehydrogenase

Fatty acid metabolism
ec:1.14.14.1 unspecific monooxygenase
ec:1.1.1.35 3-hydroxyacyl-CoA
dehydrogenase
ec:1.1.1.1 alcohol dehydrogenase
ec:5.3.3.8 dodecenoyl-CoA isomerase
ec:6.2.1.3 long-chain-fatty-acid---CoA
ligase
ec:1.3.99.7 glutaryl-CoA dehydrogenase
ec:1.3.99.2 butyryl-CoA dehydrogenase
ec:4.2.1.17 enoyl-CoA hydratase
ec:1.2.1.3 aldehyde dehydrogenase
(NAD+)

**Synthesis and degradation of ketone
bodies**
ec:2.3.3.10 hydroxymethylglutaryl-CoA
synthase

Linoleic acid metabolism
ec:1.1.1.-
ec:1.14.14.1 unspecific monooxygenase
ec:3.1.1.4 phospholipase A2

Linoleic acid metabolism
ec:1.14.14.1 unspecific monooxygenase
ec:3.1.1.4 phospholipase A2
ec:1.1.1.-

Glycerophospholipid metabolism
ec:2.3.1.51 1-acylglycerol-3-phosphate O-
acyltransferase
ec:2.7.1.82 ethanolamine kinase
ec:2.3.1.6 choline O-acetyltransferase

ec:2.3.1.-
ec:3.1.1.7 acetylcholinesterase

ec:3.1.1.4 phospholipase A2

Glycerophospholipid metabolism

ec:3.1.3.4 phosphatidate phosphatase
ec:3.1.1.4 phospholipase A2
ec:3.1.4.4 phospholipase D
ec:1.1.1.8 glycerol-3-phosphate
dehydrogenase (NAD+)
ec:2.3.1.-
ec:2.3.1.51 1-acylglycerol-3-phosphate O-
acyltransferase
ec:2.7.1.82 ethanolamine kinase

Sphingolipid metabolism
ec:1.14.-.-
ec:3.5.1.23 ceramidase
ec:2.4.1.-
ec:3.1.3.-

Sphingolipid metabolism
ec:3.1.3.-
ec:3.1.3.4 phosphatidate phosphatase
ec:3.5.1.23 ceramidase
ec:1.14.-.-
ec:3.2.1.23 beta-galactosidase
ec:3.1.4.12 sphingomyelin
phosphodiesterase
ec:2.4.1.-
ec:3.1.3.-

Biosynthesis of steroids

Biosynthesis of steroids

ec:5.3.3.2 isopentenyl-diphosphate Delta-isomerase

ec:1.14.21.6 lathosterol oxidase
ec:5.3.3.2 isopentenyl-diphosphate Delta-isomerase
ec:1.1.1.170 sterol-4alpha-carboxylate 3-dehydrogenase (decarboxylating)
ec:1.3.1.-
ec:2.7.4.2 phosphomevalonate kinase
ec:1.3.1.21 7-dehydrocholesterol reductase
ec:1.14.13.72 methylsterol monooxygenase
ec:2.5.1.10 geranyltranstransferase
ec:5.3.3.5 cholestenol Delta-isomerase

Terpenoid biosynthesis
ec:5.3.3.2 isopentenyl-diphosphate Delta-isomerase

Terpenoid biosynthesis
ec:5.3.3.2 isopentenyl-diphosphate Delta-isomerase
ec:2.5.1.10 geranyltranstransferase

Carotenoid biosynthesis
ec:1.14.-.-
ec:1.14.99.-
ec:2.3.1.-
ec:2.4.1.-
ec:1.-.-.-
ec:2.1.1.-

Carotenoid biosynthesis
ec:2.1.1.-
ec:1.14.-.-
ec:1.-.-.-
ec:2.3.1.-
ec:2.4.1.-

Ether lipid metabolism
ec:2.3.1.51 1-acylglycerol-3-phosphate O-acyltransferase
ec:2.3.1.-
ec:3.1.3.-

ec:3.1.1.4 phospholipase A2

Ether lipid metabolism

ec:3.1.3.-
ec:3.1.3.4 phosphatidate phosphatase
ec:3.1.1.4 phospholipase A2
ec:3.1.1.47 1-alkyl-2-acetylglycerophosphocholine esterase
ec:3.1.4.4 phospholipase D
ec:2.3.1.-
ec:2.3.1.51 1-acylglycerol-3-phosphate O-acyltransferase
ec:3.1.4.39 alkylglycerophosphoethanol-amine phosphodiesterase
ec:3.1.3.-

Bile acid biosynthesis
ec:1.1.1.-
ec:2.3.1.16 acetyl-CoA C-acyltransferase

Bile acid biosynthesis
ec:1.1.1.-
ec:1.1.1.1 alcohol dehydrogenase
ec:3.1.1.13 sterol esterase
ec:1.3.99.5 3-oxo-5alpha-steroid 4-dehydrogenase
ec:1.2.1.3 aldehyde dehydrogenase (NAD+)
ec:1.14.13.15 cholestanetriol 26-

monooxygenase

Retinol metabolism
ec:1.1.1.-
ec:1.14.14.1 unspecific monooxygenase
ec:3.1.1.-
ec:1.1.-.-
ec:2.3.1.20 diacylglycerol O-acyltransferase

Retinol metabolism
ec:1.14.14.1 unspecific monooxygenase
ec:1.1.-.-
ec:3.1.1.-
ec:1.1.1.-
ec:1.1.1.1 alcohol dehydrogenase

alpha-Linolenic acid metabolism
ec:3.1.2.-
ec:2.3.1.16 acetyl-CoA C-acyltransferase

ec:1.-.-.-
ec:3.1.1.4 phospholipase A2
ec:1.1.1.35 3-hydroxyacyl-CoA dehydrogenase

alpha-Linolenic acid metabolism
ec:3.1.1.4 phospholipase A2
ec:1.-.-.-
ec:1.1.1.35 3-hydroxyacyl-CoA dehydrogenase
ec:3.1.2.-
ec:4.2.1.17 enoyl-CoA hydratase

Table IVa. Specific stress pathway proteins predicted to be present in *Hypsibius dujardini* according to the analyzed EST data.*

DNA repair	e.g. MutS (gb|CD449386.1)
LEA proteins	e.g. gb|CK325833.1
DNA protection (RNA helicase)	e.g. gb|CD449793.1
Redox protection (peroxidase and superoxid dismutase)	e.g. gb|CK326879.1 and e.g. gb|CK326506.1
Heat shock proteins	e.g. gb|CF544577.1, gb|AAQ94878.1, gb|CD449707.1
Aquaporins	e.g gb|CD449847.1
Not detected: - pigments, - anti-freeze protein	

Key examples; more information in: *supplementary table Va on our tardigrade website

Table IVb. Specific stress pathway proteins predicted to be present in *Milnesium tardigradum* according to the translated EST data*

DNA repair	DNA repair protein RAD51, DnaJ Family
DNA protection (RNA helicase)	e.g. Helicase_C, DEAD, ATP-dependent RNA helicase, Spliceosome RNA helicase BAT1, Peroxiredoxin-5, Redoxin
Redox protection (superoxid dismutase and peroxidase)	e.g. Superoxide dismutase [Cu-Zn], Sod_Cu, AhpC-TSA, Peroxiredoxin-1,-2,-4,-6
Heat shock proteins	e.g. Heat shock protein 90, HSP 70, HSP 20, HSP 30
Cold shock protein	Cold shock-like protein cspC
Membrane permeability	e.g. MIP (Major Intrinsic Protein), Aquaporin-9, -10
Not detected: - pigments, - anti-freeze protein	

Key examples; more information in: *supplementary table Vb on our tardigrade website

Table IVc. Specific stress pathways derived from proteins predicted to be present in *M. tardigradum*

Bmh2 pathway	Bmh2 (key regulator), Sck1 (key regulator), Lcb1 (depending enzyme)[+], Lcb2 (depending enzyme)[+](Fig.5)[56]
Major DNA repair pathways	
NHEJ (non-homologous end joining):	Rad50 (regulated response)
HRR (homologous recombination):	RuvB, Rad51, Rad50 (regulated responses)
MMR (mismatch repair):	MutS, PCNA (regulated response)
NER (nucleotide excision repair):	Rad23 (regulated responses)
BER (base excision repair):	XRCC1 (regulated response)
Heat- shock response (HSR) pathways	10 kDa GroES/Hsp10 (regulated response) 20-30 kDa GrpE/Hsp27, Hsp20 (interaction partner of Mef2)[92]

	40 kDa DnaJ/Hsp40 (regulated response, stimulates Hsp70)
	70 kDa DnaK/Hsp70 (regulated response)
	90 kDa HtpG/Hsp90 (regulated response)
	Mef2 (key regulator), GAPDH (depending enzyme)[92]
	HSBP1 (regulator of TF HSF1)
LEA pathways	LEA
HOG signal pathway:	Hsp12; PBS2, HOG1 and
	PKA (regulators of Hsp12 in *S. cerevisiae,* found in tardigrades)[93; 94]
No Plant ABA responsive pathway:	vp1, vp5 (regulators in *Z. mays, A. thaliana*, not found in tardigrades)[95]
	B15C (peroxiredoxin), Rab 16 (rice, maize, barley, not found in tardigrades)[96]
	abi3 (*A. thaliana*, TF: regulates ABA-responsive genes, not found in tardigrades)[97; 98]
Protective pathways	SOD (enzyme), Prx-1,-2,-4,-6 (regulated enzyme by StRE)
	AP-1 and Nrf2 (redox regulated TFs for StRE, Nrf2 was not found, but its inhibitor INrf2)[99]
RNA helicase:	cgh-1, BAT1, BRR2[*], FAL1[*] (regulated responses)
DNA helicase:	HelicaseC, RAD5[*], TIP49[*] (regulated responses)

[*]identified in KOGs/COGs, see supplementary table II on our tardigrade website. [+]Lcb was found, but it is not clear which EST is Lcb1 and which Lcb2

Milnesium tardigradum
3050
214/234
(178)
Hypsibius dujardini
2105

Top 5 cluster in
Milnesium tardigradum

Top 5 cluster in
Hypsibius dujardini/Milnesium tardigradum

Top 5 cluster in
Hypsibius dujardini

1 2 3 4 5 6 7 8 9 10 11 12

94

gi|37204978
gi|50296889
gi|37212679
gi|40006518
Q9LF88_ARATH/354/397
gi|50272039
gi|39967540
gi|39967548
gi|37210843
gi|37213597
Q39871_SOYBN/348/391
Q39871_SOYBN/297/340
gi|39999831

Q9ZPW6_ARATH/77/120
Q9ZPW6_ARATH/322/365
Q1XI26_9DIPT/1/45
Q1XI26_9DIPT/453/496
Rc_MM1_01803
LEAD8_DAUCA/238/281
LEAD8_DAUCA/326/369
LEAD8_DAUCA/155/198
Q96246_ARATH/104/147
Q96246_ARATH/99/142
Q96246_ARATH/143/186
Q06431_BETVE/246/289
Q06431_BETVE/202/245
Q9LF88_ARATH/135/178
Q39871_SOYBN/158/201
Q2N1E0_PHAVU/119/162
Q06431_BETVE/78/121
Q96246_ARATH/220/263
Q2N1E0_PHAVU/163/206
Q9ZRW5_PHAVU/180/222
Q9ZTY1_SOYBN/170/213

Rc_MM1-92D19
Rc_MM1-93D19
Q7P280_FUSNV/15/58

gi|50297072

Q40334_MEDSA/2-106
O04132_SOYBN/2-106
Q9M4H2_VITVI/2-92
Q41092_PONTR/2-106
Q40711_ORYSJ/2-92
Q9M3V2_NICAL/2-87
O2623_VACCO/2-310
Q9M4N9_SECCE/1-108
Q41804_MAIZE/1-108
Q53WZ7_GOSHI/1-109
Q41684_PHAAU/1-109
Q5KT57_DAUCA/1-109
ERD10_ARATH/4-177
ERD14_ARATH/38-256
CO410_WHEAT/9-184
DHN1_PEA/12-194
Q8S626_WHEAT/75-391
DHN1_HORVU/11-139
DHR25_ORYSJ/13-228
Q9FRQ1_ARATH/21-186
BRHC_BRAPA/12-180
BRHC_BRAPA/5-117
DH11_GOSHI/13-145
DAS14_SOLLC/10-130
Q9FRQ1_ARATH/14-128
DHR16D_ORYSJ/14-168
DHR15_WHEAT/12-164
DHN1_MAIZE/14-168
DHN3_HORVU/7-161
DH16B_ORYSJ/12-151
DHR16C_RAPE/11-155

3.0

95

Tardigrades *H. sapiens*

81 (1) 11 58

172

0 *358 (1)*

5

C. elegans

Tardigrades (81) Yeast

46 35 136

Ubiquitin-like proteins, 20S proteasome,
chaperonin complex, ubiquitin protein
ligase, 26S proteasome regulatory
complex, thioredoxin-like protein,
AAA+-type ATPase, DNA repair
protein (RAD51/RHP55)

Shared with
man and
nematodes

Tardigrade adaptions

**DNA protection (MutS, DnaJ family,
helicases, recombinases), RNA protection
(helicases), Redox protection (strong;
glutaredoxin-related proteins), aquaporins,
LEA proteins, Heat shock proteins (hsp70),
protein turnover and integrity
(e.g. molecular chaperons)**

Metabolic adaptions (beyond central
pathways): Trehalose (*Richtersius coronifer*,
not *Milnesium*), Bile acid biosynthesis,
ether lipid metabolism, Terpenoids

Basic protection,
e.g. ABC transporters,
Zn proteases,
DnaJ-class chaperonins,
glutathion peroxidase,
ubiquitin, Bmh2 pathway

Shared with
yeast

Shared with
rotifers only

Hsp90 chaperonin

Shared
with man,
not in
nematodes

Bmh2 pathway
DNA repair (RHP57)
Proteasome maturation factor

Chapter 6.

Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms

Proteomic Analysis of Tardigrades: Towards a Better Understanding of Molecular Mechanisms by Anhydrobiotic Organisms

Elham Schokraie[1], Agnes Hotz-Wagenblatt[1], Uwe Warnken[1], Brahim Mali[2], Marcus Frohme[2], Frank Förster[3], Thomas Dandekar[3], Steffen Hengherr[4], Ralph O. Schill[4], Martina Schnölzer[1]*

1 Functional Proteome Analysis, German Cancer Research Center, Heidelberg, Germany, 2 Department of Molecular Biology and Functional Genomics, University of Applied Sciences Wildau, Wildau, Germany, 3 Department of Bioinformatics, University of Würzburg, Würzburg, Germany, 4 Department of Zoology, University of Stuttgart, Stuttgart, Germany

Abstract

Background: Tardigrades are small, multicellular invertebrates which are able to survive times of unfavourable environmental conditions using their well-known capability to undergo cryptobiosis at any stage of their life cycle. *Milnesium tardigradum* has become a powerful model system for the analysis of cryptobiosis. While some genetic information is already available for *Milnesium tardigradum* the proteome is still to be discovered.

Principal Findings: Here we present to the best of our knowledge the first comprehensive study of *Milnesium tardigradum* on the protein level. To establish a proteome reference map we developed optimized protocols for protein extraction from tardigrades in the active state and for separation of proteins by high resolution two-dimensional gel electrophoresis. Since only limited sequence information of *M. tardigradum* on the genome and gene expression level is available to date in public databases we initiated in parallel a tardigrade EST sequencing project to allow for protein identification by electrospray ionization tandem mass spectrometry. 271 out of 606 analyzed protein spots could be identified by searching ag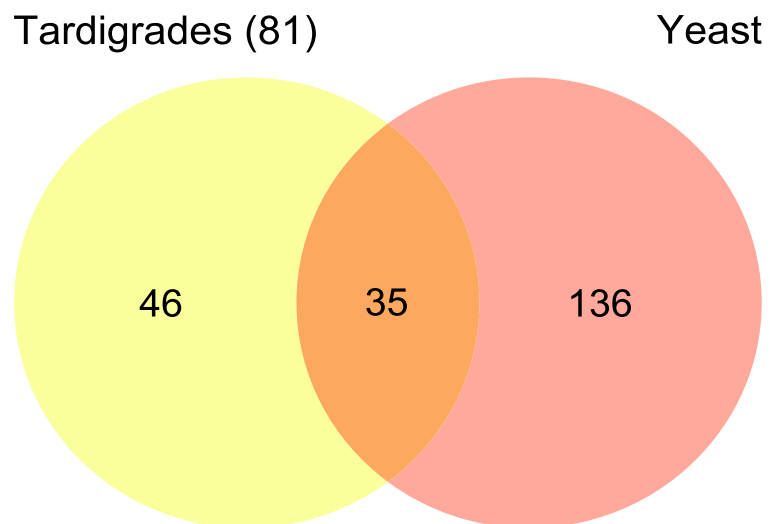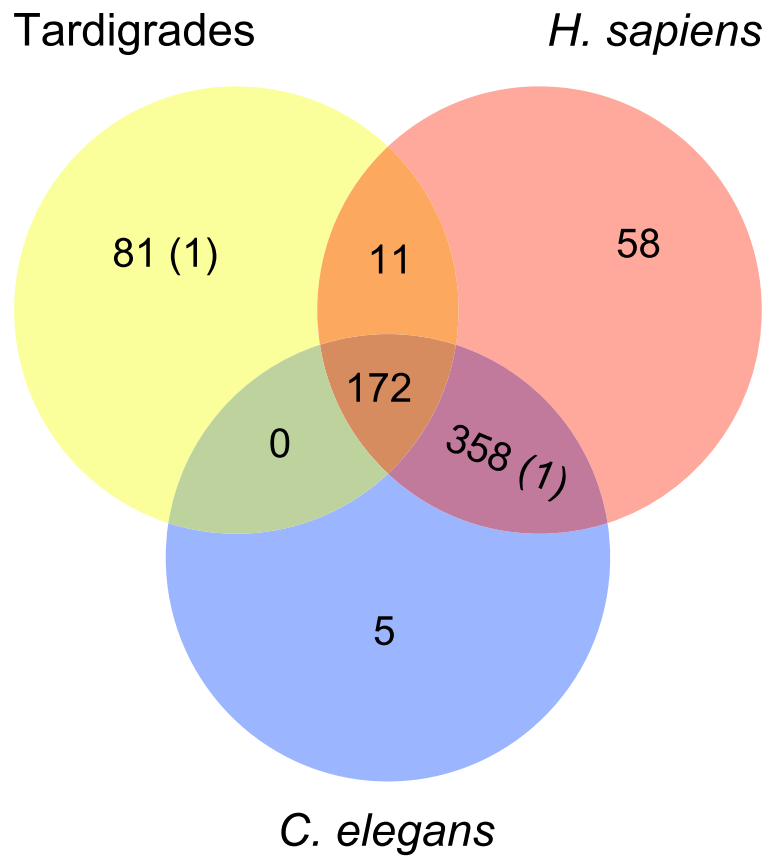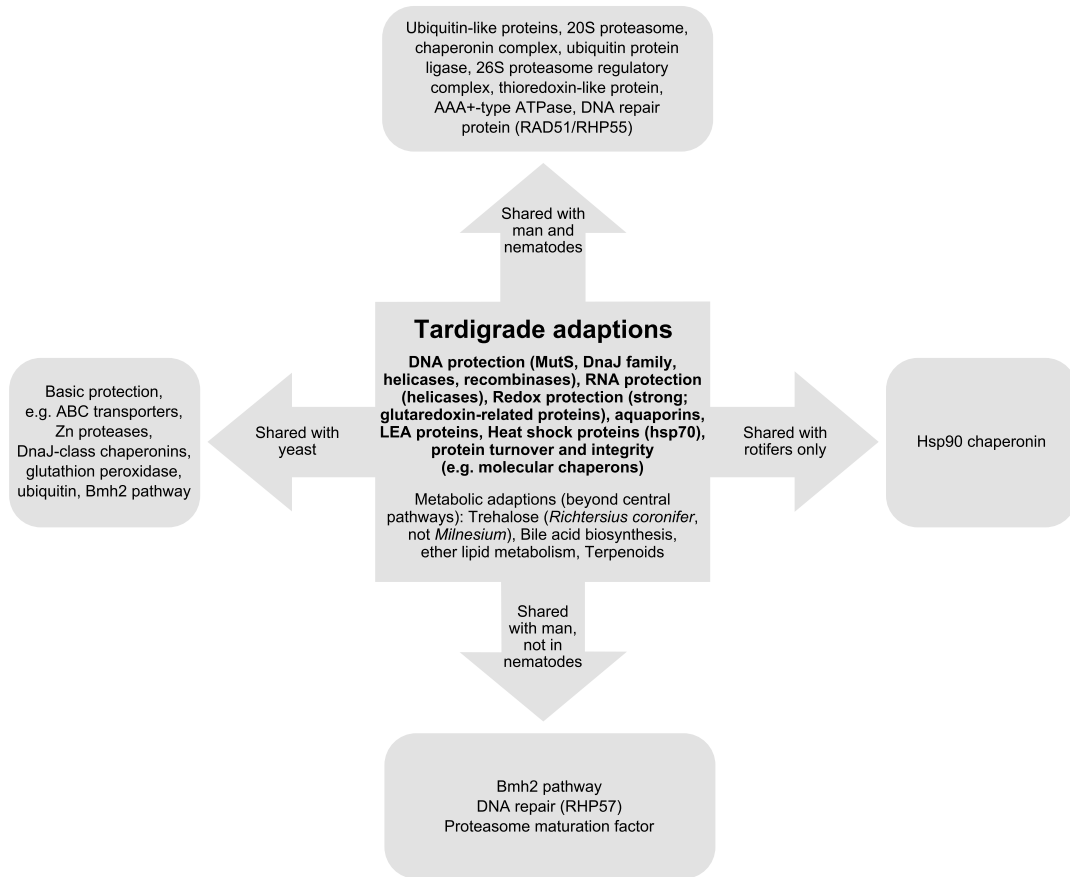ainst the publicly available NCBInr database as well as our newly established tardigrade protein database corresponding to 144 unique proteins. Another 150 spots could be identified in the tardigrade clustered EST database corresponding to 36 unique contigs and ESTs. Proteins with annotated function were further categorized in more detail by their molecular function, biological process and cellular component. For the proteins of unknown function more information could be obtained by performing a protein domain annotation analysis. Our results include proteins like protein member of different heat shock protein families and LEA group 3, which might play important roles in surviving extreme conditions.

Conclusions: The proteome reference map of *Milnesium tardigradum* provides the basis for further studies in order to identify and characterize the biochemical mechanisms of tolerance to extreme desiccation. The optimized proteomics workflow will enable application of sensitive quantification techniques to detect differences in protein expression, which are characteristic of the active and anhydrobiotic states of tardigrades.

Introduction

Many organisms are exposed to unfavourable, stressful environmental conditions, either permanently or for just certain periods of their lives. To survive these extreme conditions, they possess different mechanisms. One of amazing adaptation is anhydrobiosis (from the Greek for "life without water"), which has puzzled scientists for more than 300 years. For the first time the Dutch microscopist Anton van Leeuwenhoek (1702) gave a formal description of this phenomenon. He reported the revival of "animalcules" from rehydrated moss samples. In extreme states of dehydration, anhydrobiotic invertebrates undergo a metabolic dormancy, in which metabolism decreases to a non-measurable level and life comes to a reversible standstill until activity is resumed under more favourable conditions [1]. One of the best known anhydrobiotic organisms are tardigrades. Tardigrades remain in their active form when they are surrounded by at least a film of water. By loosing most of their free and bound water (>95%) anhydrobiosis occurs [2]. Tardigrades begin to contract their bodies and change their body structure into a so-called tun state (Figure 1). In the dry state these organisms are highly resistant to environmental challenge and they may remain dormant for a long period of time. Based on their amazing capability to undergo anhydrobiosis, tardigrades colonise a diversity of extreme habitats [3], and they are able to tolerate harsh environmental conditions in any developmental state [4]. Possessing the ability to enter anhydrobiosis at any stage of life cycle, tardigrades can extend their lifespan significantly [4,5]. Additionally, in the anhydrobiotic state, tardigrades are
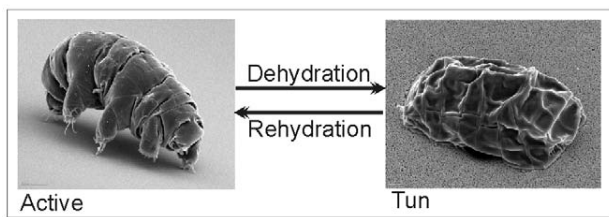
Figure 1. SEM images of *M. tardigradum* in the active and tun state. Tardigrades are in the active form when they are surrounded by at least a film of water. By loosing most of their free and bound water (>95%) anhydrobiosis occurs. Tardigrades begin to contract their bodies and change their body structure into a so-called tun.
doi:10.1371/journal.pone.0009502.g001

extraordinary tolerant to physical extremes including high and subzero temperatures [6,7,8], high pressure [6,9], and extreme levels of ionizing radiation [10,11]. Interestingly, tardigrades are even able to survive space vacuum (imposing extreme desiccation) and some specimens have even recovered after combined exposure to space vacuum and solar radiation [12].

Anhydrobiosis seems to be the result of dynamic processes and appears to be mediated by protective systems that prevent lethal damage and repair systems. However, the molecular mechanisms of these processes are still poorly understood. Up to now investigations of mechanisms of desiccation tolerance have focused mainly on sugar metabolisms, stress proteins and a family of hydrophilic proteins called LEA (late embryogenesis abundant). The presence of non-reducing trehalose and its expression during anhydrobiosis has been reported for different anhydrobiotic species [13,14], which indicates the important role of trehalose in anhydrobiosis. However, the existence of anhydrobiotic animals that exhibit excellent desiccation tolerance without having disaccharides in their system [15,16] shows that sugars alone do not sufficiently explain these phenomena.

Milnesium tardigradum Doyère (1840) is a very well known species of carnivorous tardigrade. Different aspects of the life history of this species have been already described [17]. While some genetic studies of *M. tardigradum* exist [18] almost nothing is known about the proteome. Partial sequences of three heat shock protein (hsp70 family) genes and the housekeeping gene beta-actin have been described [18] and the relation of hsp70 expression to desiccation tolerance could be shown by real time PCR [18] and by de novo protein synthesis [6]. Since no trehalose could be detected in *M. tardigradum* [19], investigating proteins and posttranslational modifications is of particular importance to clarify surviving mechanisms during desiccation.

To gain insight into the unique adaptation capabilities of tardigrades on the protein level we aimed to establish a comprehensive proteome reference map of active *M. tardigradum* employing optimized protocols for protein extraction, generation of high-resolution 2D gels and high-throughput protein identification by electrospray ionization tandem mass spectrometry (ESI-MS/MS). The proteome reference map of *M. tardigradum* provides the basis for further studies in order to understand important physiological processes such as anhydrobiosis and stress resistance. The optimized proteomics workflow will enable application of sensitive quantification techniques to detect differences in protein expression, which are characteristic of active and anhydrobiotic states. Thus, our proteomic approach together with in-depth bioinformatic analysis will certainly provide valuable information to solve the over 300 years existing puzzle of anhydrobiosis.

Results

Preparation of Protein Extracts from Active Tardigrades

To establish and optimize a reliable and robust protocol for the extraction of proteins from tardigrades in the active state we applied different workup protocols and evaluated them by one-dimensional (1D) gel electrophoresis. Figure 2 shows the separation of protein extracts from whole tardigrades without any precipitation step (lane 2), after trichloroacetic acid/acetone precipitation (lane 3), after chloroform/methanol precipitation (lane 4) and after using a commercially available clean-up kit (lane 5). When using trichloroacetic acid/acetone precipitation we lost many proteins especially in the low molecular weight range. Chloroform/methanol precipitation and application of clean-up kit delivered satisfying results but also using the whole protein lysate directly without any further purification resulted in high yields across the entire molecular weight range. This workup protocol was therefore used throughout our proteome study. To evaluate the quality of our protocol especially with respect to proteolysis we performed Western blot analysis to detect any protein degradation. Since no proteins have been identified so far, we have chosen two polyclonal antibodies directed against the highly conserved proteins actin and alpha-tubulin. As shown in Figure 3A and 3B both proteins could be detected at their expected molecular weight at approx. 40 and 50 kDa, respectively, which is in agreement with the protein bands of the control lysate of HeLa cells. Importantly, no protein degradation could be observed during our sample preparation.

Two Dimensional Gel Electrophoresis (2-DE)

The establishment of an optimized workup protocol was a prerequisite for high quality 2D gels from tardigrades in the active state. The proteomics workflow is depicted in Figure 4. One important step in the workflow is the collection and preparation of the samples. To avoid contamination with food-organisms,
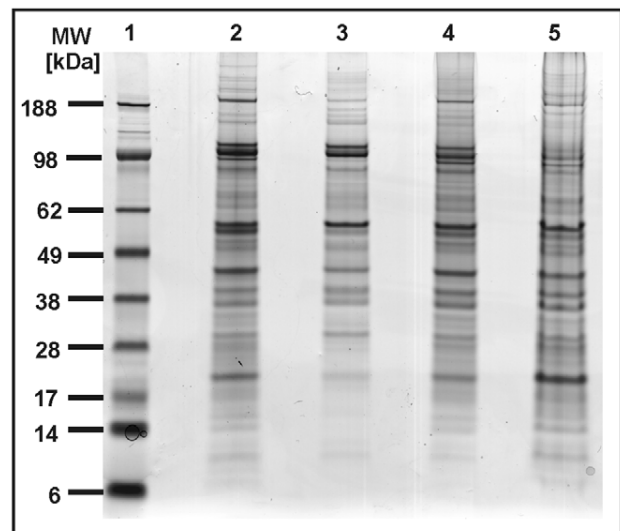
Figure 2. Comparison of different workup protocols for *M. tardigradum*. Total protein extract of tardigrades in the active state was separated on a one-dimensional polyacrylamide gel. Lane 1: Rainbow molecular weight marker. Lane 2: Protein extract of whole tardigrades without any precipitation step. Lane 3: Protein extract after TCA precipitation. Lane 4: Protein extract after chloroform/methanol precipitation. Lane 5: Protein extract using clean-up kit.
doi:10.1371/journal.pone.0009502.g002
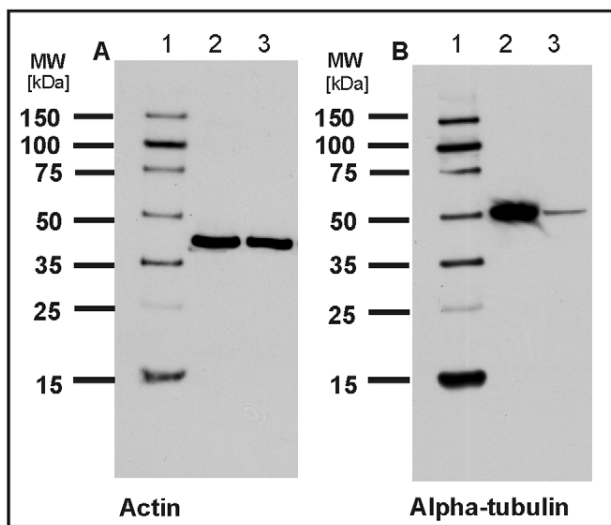
Figure 3. Analysis of protein degradation in total protein extracts of tardigrades by Western blot analysis. Actin (A) and alpha tubulin (B) were used as marker proteins for the detection of proteolysis. Lane 1A and 1B: DualVue Western blotting marker. Lane 2A and 2B: Total protein extract of HeLa cells as control. Lane 3A and 3B: Total protein extract of *M. tardigradum*. Notably, no protein degradation was observed during the workup procedure.
doi:10.1371/journal.pone.0009502.g003

Figure 4. The experimental workflow to developing the proteome map. Tardigrades were sonicated directly in lysis buffer. Total protein extracts were separated by two-dimensional gel electrophoresis. After silver staining protein spots were picked and in-gel digested with trypsin. MS/MS data obtained by LC-ESI-MS/MS analysis were searched against the NCBInr database, the clustered tardigrade EST database and the tardigrade protein database. Identified proteins with annotation were classified in different functional groups using the Blast2GO program. Identified proteins without annotation were analysed with the DomainSweep program to annotate protein domains.
doi:10.1371/journal.pone.0009502.g004

tardigrades were washed several times and starved over 3 days. Direct homogenization and sonication of deep-frozen tardigrades in ice cold lysis buffer without any previous precipitation step yielded protein extracts which were separated by high resolution 2D gel electrophoresis. For maximal resolution of protein spots and high loading capacity (330 μg proteins) we used pI 3–11 NL strips (24 cm) for the first dimension. Thus, high resolution separation coul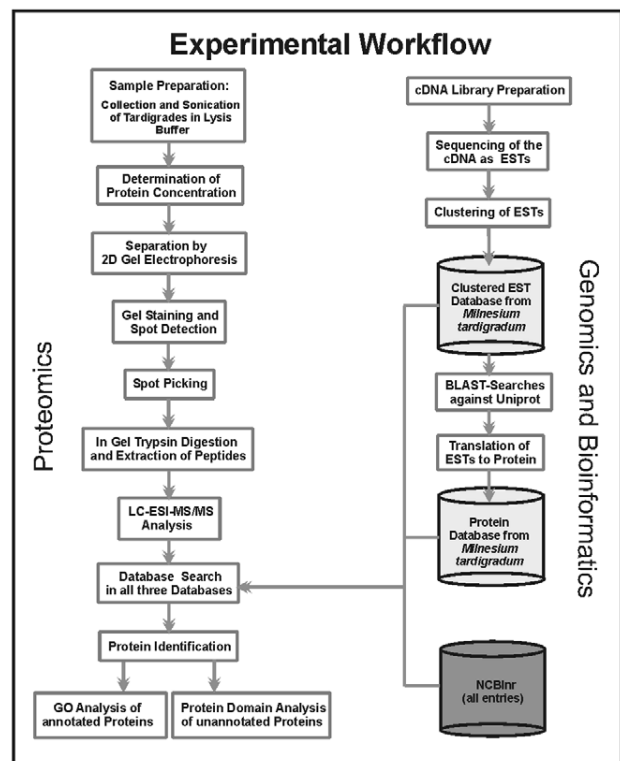d be achieved in the acidic as well as in the basic pH range as shown in the image of the silver stained preparative gel of whole protein extract (Figure 5).

Approximately 1000 protein spots were automatically detected on the 2D gel image using the Proteomweaver image software. A total of 606 protein spots were picked from the silver stained gel. These spots were digested with trypsin and after extraction of the tryptic peptides from the gel plugs peptide mixtures were analyzed by nanoLC-ESI-MS/MS.

Protein Identification

Identification of proteins depends on the representation of the sequence or a close homologue in the database. Since almost no genome or EST sequences of *M. tardigradum* are available to date in public databases we initiated the tardigrade EST sequencing project as outlined in figure 4 (Mali et al, submitted data). A cDNA library was prepared from tardigrades in different states (active, inactive, transition states). The cDNAs were sequenced as ESTs and clustered. Thereby, we obtained a nucleotide database containing 818 contigs and 2500 singlets. cDNA sequencing and generation of ESTs are still ongoing, thus the sequence coverage of *M. tardigradum* in the database is incomplete.

For protein identification we used the following databases: the database of *M. tardigradum* containing the clustered ESTs as outlined above, the tardigrade protein database, which was translated from the clustered EST database and thus represents a subdatabase containing only annotated proteins with known function and the publicly available NCBInr database. The selected

606 spots from the 2D gel correspond to some highly expressed proteins, but mostly to spots in the medium and low expression range. A total of 271 spots could be identified from the tardigrade protein and the NCBInr databases. Figure 6 demonstrates how identified proteins are distributed among these two databases. 56 unique proteins were successfully identified by searching the NCBInr database. It concerns proteins which are either highly conserved among different species e. g. actin or protein entries from *M. tardigradum* which are already available in the NCBInr database e.g. elongation factor 1-alpha. Further 73 unique proteins could be identified by searching the tardigrade protein database and another 15 unique proteins were present in both databases. Identical proteins that were identified from several spots were included only once in the statistics to avoid bias. Thus, the combination of the two databases was sufficient for the identification of 144 unique proteins. The corresponding protein spots are indicated by green circles in the 2D reference map shown in Figure 5. Table 1 shows an overview of identified proteins with annotation in different functional groups. In addition, detailed information about each of the identified 144 proteins including spot number, protein annotation, accession number (NCBInr and Tardigrade specific accession number), total protein score, number of matched peptides, peptide sequence and sequence coverage is
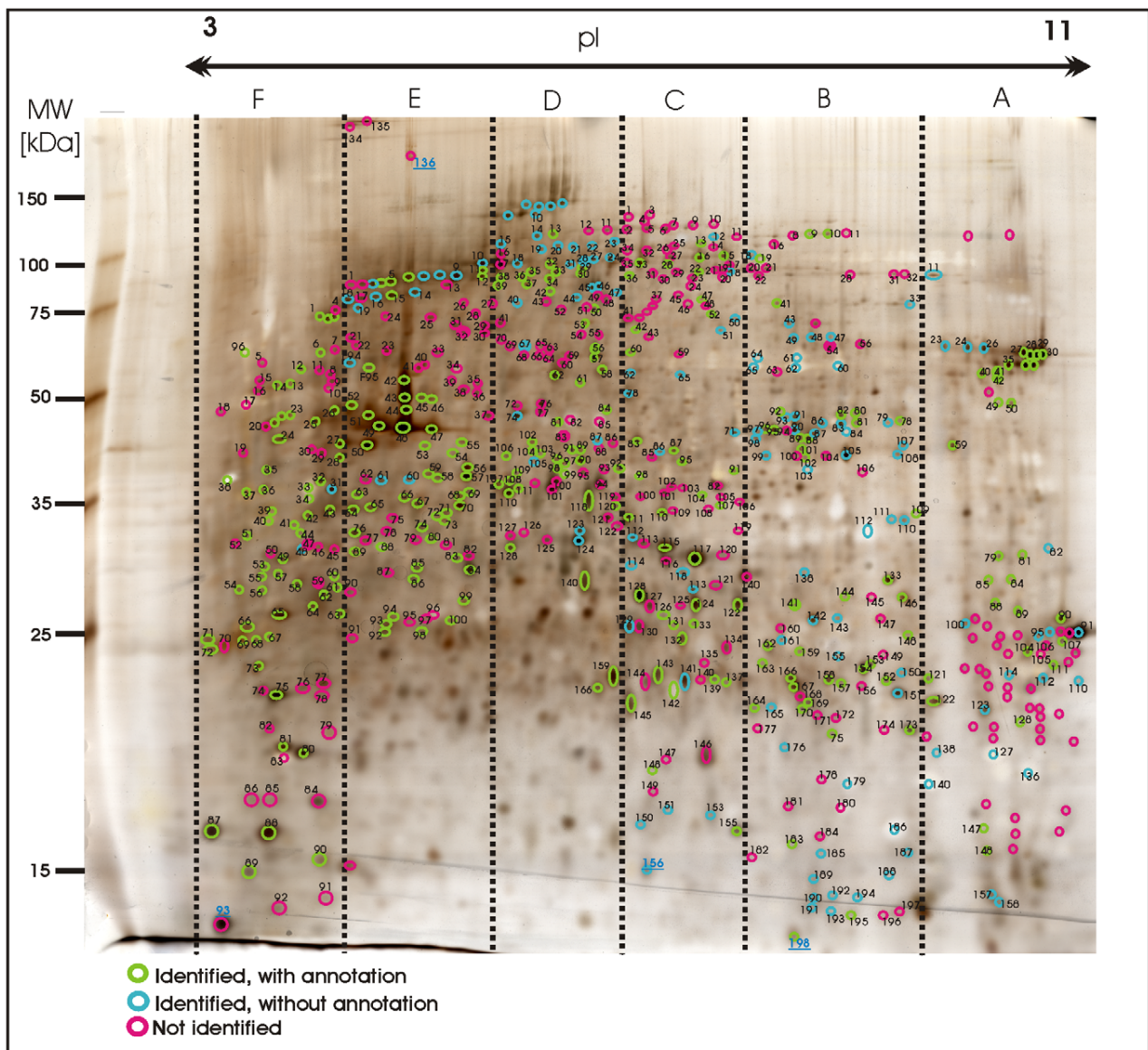
Figure 5. Image of a preparative 2D-gel with selected analysed protein spots. Total protein extract of 400 tardigrades in the active state corresponding to 330 µg was separated by high resolution two-dimensional gel electrophoresis. Proteins were visualised by silver staining. Three different categories are shown: Identified proteins with functional annotation are indicated in green, identified proteins without annotation are indicated in blue and not yet identified proteins are indicated in red.
doi:10.1371/journal.pone.0009502.g005

listed in Table 2. The individual ion score is included in brackets at the end of every peptide sequence. Following ion scores indicate a significant hit ($p < 0.05$): >53 for NCBInr searches, >14 for searches in the tardigrade protein database and >27 by searching the EST clustered database. Identical proteins identified in different spots are listed only once in Table 2. In these cases the spot with the highest protein score (in bold) is ranked at the top whereas the other spots are listed below. All further information such as accession numbers, peptide sequences and sequence coverage refer to the top-ranked spot.

The 15 proteins which were identified in both databases are indicated with asterisk (e.g. spot A30*) and both accession numbers are listed. In these cases the listed peptide sequences belong to the hit with the highest score. Protein spots below the bold one are marked with °, when only found in the NCBInr database or marked with ^, when only found in the tardigrade protein database.

Furthermore we were able to identify additional 150 protein spots by searching MS/MS data in the clustered EST database of *M. tardigradum*. These 150 proteins correspond to 36 unique contigs and ESTs. The protein information is listed in Table 3 and the protein spots are indicated by blue circles in the 2D reference map (Figure 5). Unfortunately, it was not possible to annotate them when performing a BLAST search. For these proteins of unknown function more information could be obtained by applying protein domain annotation methods. We ran all proteins through the DomainSweep pipeline which identifies the domain architecture within a protein sequence and therefore aids in finding correct functional assignments for uncharacterized protein sequences. It employs different database search methods to scan a number of
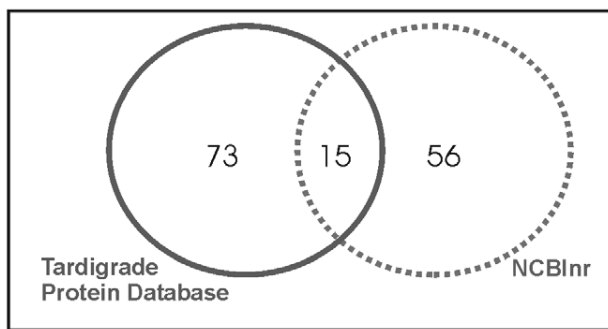
Figure 6. Comparison of database performance for protein identification. Protein spots were analysed by nanoLC-ESI-MS/MS and searched against the NCBInr database and the tardigrade protein database. The diagram illustrates the number of positive identifications in the respective database and the overlap between the two databases.
doi:10.1371/journal.pone.0009502.g006

protein/domain family databases. 2 out of the 36 unique proteins gave a significant hit, whereas 28 proteins were listed as putative and 6 proteins gave no hit at all.

In addition, we analyzed further 185 protein spots, which are indicated with red colour in Figure 5. Despite high quality MS/MS spectra, it was not possible to identify these protein spots in either of the databases used in our study.

In summary, we identified 421 (69.5%) out of 606 protein spots which were picked from the preparative 2D gel. 271 spots yielded 144 unique proteins with distinct functions whereas 150 spots were identified as proteins with yet unknown functions.

Functional Assignment of Proteins

The 144 unique proteins with annotation were further analysed using the Blast2GO program, which provides analysis of sequences and annotation of each protein with GO number to categorize the proteins in molecular function, biological process and cellular component. By analysing the proteins on the GO level 2 in the category molecular function we received a total of 9 subgroups as shown in Figure 7, upper middle chart. The majority of the identified proteins exhibit either binding (45%) or catalytic activity (33%). A more detailed analysis (GO level 3) revealed that 39% of the proteins with catalytic activity are involved in hydrolase activity (Figure 7, upper right chart) and 38% of binding proteins bind to other proteins (Figure 7, upper left chart).

Identified proteins are involved in diverse biological processes. A total of 16 subgroups of biological processes are represented (Figure 7, lower middle chart). 23% are involved in cellular processes and 18% in metabolic processes. Within the cellular processes a majority of 20% of tardigrade proteins are involved in cellular component organization and biogenesis. Within metabolic processes 28% of proteins are involved in cellular metabolic processes, 26% in primary metabolic processes and 21% in macromolecule metabolic processes (Figure 7, lower right chart). A detailed GO description of all identified and annotated tardigrade proteins is included in Table S1.

Identified Proteins and Protein Families

In our proteomic study several heat shock proteins have been identified, namely hsp-1 (spot F27), hsp-3 (spot F21), hsp60 (spot F57), hsp70 (spot B146, B173, C131, C133), hsp82 (spot F13), hsp86 (spot F24, F25), hsp90 alpha (spot E64), hsp90 beta (spot F24) and hsp108 (spot F12). Hsp70 is already described in *M. tardigradum* as a molecular chaperone which could play a role in desiccation

tolerance [18]. Hsp60 could be identified in spot F57 when searching the corresponding MS/MS data against the NCBInr database. No hit was obtained in the tardigrade EST or protein database which is surprising, because hsp60 is an abundant protein.

Several protein spots have been identified as cytoskeletal proteins, including actin as most abundant protein spot (E48) on the 2D gel and tubulin. Actin and tubulin are highly conserved proteins and were used to control proteolytic degradation during our workup procedure by Western blotting. Four different actin proteins are found by MS/MS analysis, which play important roles in muscle contraction, cell motility, cytoskeletal structure and cell division. Tubulin is a key component of the cytoskeletal microtubules. Both alpha- and beta-tubulin could be identified on the 2D gel in spot D107, D110 and F6. Further proteins involved in motor activity and muscle contraction were found, namely tropomyosin (e.g. spot F35), myosin (e.g. spot F81), annexin A6 (e.g. spot D90) and myophilin (e.g. spot A128), which is a smooth-muscle protein and was described in the tapeworm *Echinococcus granulosus* [20].

In addition, several proteins have been identified which are known to have important roles in embryonic or larval development. Mitochondrial malate dehydrogenase precursor (e.g. spot B109), vitellogenin 1 and 2 (e.g. spot D62 and B88), GDP-mannose dehydratase (spot C87), protein disulfide isomerase 2 (e.g. spot F3), hsp-3 (spot F21), hsp-1 (spot F27), tropomyosin (spot F35) and troponin C (spot F87) belong to this group of proteins. Vitellogenin, a major lipoprotein in many oviparous animals, is known as the precursor of major yolk protein vitellin [21]. Vitellogenin is a phospholipo-glycoprotein which functions as a nutritional source for the development of embryos [22]. During developing oocytes vitellogenin and vitellin are modified through cleavage and by different posttranslational modifications (PTMs) like glycosylation, lipidation and phosphorylation. Interestingly we could identify vitellogenin in several spots on the 2D gel showing vertical (pI) shifts most probably caused by PTMs.

Peroxiredoxins identified first in yeast [23] are conserved, abundant, thioredoxin peroxidase enzymes containing one or two conserved cysteine residues that protect lipids, enzymes, and DNA against reactive oxygen species. Different isoforms of peroxiredoxins could be identified on the 2D gel: peroxiredoxin-4 (spot C132), peroxiredoxin-5 (spot B183) and peroxiredoxin-6 (spot D159). An important aspect of desiccation tolerance is protection against free radicals [24,25]. Notably, the expression of 1-cysteine (1-Cys) peroxiredoxin family of antioxidants is reported in *Arabidopsis thaliana* and is shown to be related to dormancy [26]. Our results show the presence of important antioxidant systems, including superoxide dismutase (SOD) and peroxidases. Additionally different forms of glutathione S-transferases (spot A122, B153, B166, B169, D166, and D159) could be identified. Glutathione transferases (GSTs) constitute a superfamily of detoxifying enzymes involved in phase II metabolism. Detoxification occurs by either glutathione conjugation, peroxidase activity or passive binding [27]. Furthermore GSTs have cellular physiology roles such as regulators of cellular pathways of stress response and housekeeping roles in the binding and transport of specific ligands [28]. The consequence of this diversity in role is the expression of multiple forms of GST in an organism. It has been shown that the expression of the different isoenzymes is highly tissue-specific [29], and this heterogeneity of GSTs may be further complicated by posttranslational modifications such as glycosylation [30].

Some protein spots were identified as calreticulin (e.g. spot F14) which is a Ca^{2+}-binding protein and molecular chaperone. Calreticulin is also involved in the folding of synthesized proteins and glycoproteins [31].

Table 1. Overview of identified proteins classified in different functional groups.

Cytoskeleton elements and modulators	Enzymes	Proteases and protease inhibitors
Alpha-III tubulin	Glucan endo-1,3-beta-glucosidase	Cathepsin K
Beta-tubulin class-IV	Prostatic acid phosphatase	Cathepsin Z
Beta-tubulin class-I	Adenylate kinase isoenzyme 1	Cathepsin L1
Actin	Peptidyl-prolyl cis-trans isomerase	Neprilysin-2
Actin-5C	Glutamate dehydrogenase	Peptidase M17 precursor
Beta actin	Lysosomal acid phosphatase	Actinidain
Alpha actin	Mitochondrial malate dehydrogenase	Plasminogen
Actin, muscle-type (A2)	Arginine kinase	Aspartic protease inhibitor 8
muscle actin	Aconitase, mitochondrial	AFG3-like protein 2
Similar to alpha actinin CG4376-PB	Transaldolase	26S proteasome non-ATPase regulatory subunit 8
Myophilin	Aldolase A protein	Rab GDP dissociation inhibitor beta
Tropomyosin-1, isoforms 9A/A/B	Protein disulfide isomerase-3	Gamma-glutamyltranspeptidase
Tropomyosin	Matrix metalloproteinase-17	**Response to stress or heat**
Myosin regulatory light polypeptide 9	Mitochondrial long-chain enoyl-CoA hydratase/3-hydroxycyl-CoA	NADP-dependent isocitrate dehydrogenase
Myosin, essential light chain	Dehydrogenase alpha-subunit	Heat shock 70 kDa protein II
Heat shock proteins	Peroxidase	similar to heat shock cognate 70 protein isoform 2
Heat Shock Protein family member (hsp-3)	Methylmalonate-semialdehyde dehydrogenase	Short-chain dehydrogenase/reductase SDR YhdF
Heat Shock Protein family member (hsp-1)	Thioredoxin reductase 1	Aspartic protease inhibitor 8
Hsp 60	Succinyl-CoA ligase [GDP-forming] subunit beta, Mitochondrial E	UspA
Hsp 70	GTP-specific succinyl-CoA synthetase beta subunit	Rubber elongation factor protein (REF) (Allergen Hev b 1)
Heat shock cognate 70	Glycosyl transferase	Small rubber particle protein (SRPP) (22 kDa rubber particle protein)
Heat shock cognate 70 protein isoform 2	DEAD-box family (SNF2-like) helicase	Heat shock protein 90-beta
Heat shock 70 kDa protein II (HSP70 II)	Cysteine conjugate beta-lyase	Heat shock protein 83
Hsp 90-beta	26S proteasome non-ATPase regulatory subunit 13	Heat shock protein 60
Hsp90-alpha	GH19645	**Other Proteins**
Hsp90	**Glycolysis**	Translationally-controlled tumor protein homolog
Hsp 82	Glyceraldehyde-3-phosphate dehydrogenase	Elongation factor 1-alpha
Hsp 83	Triosephosphate isomerase	Elongation factor 1 gamma
Hsp108	Enolase	Elongation factor 2
Protein lethal(2)essential for life (member of Hsp20 family)	Phosphoglycerate kinase	Angiopoietin-related protein 1
Embryonic/larval development	**Transporters**	Spaghetti CG13570-PA
Vitellogenin-1	H(+)-transporting ATP synthase	Prohibitin
Vitellogenin-2	ATP synthase subunit d, mitochondrial	Proteasome subunit alpha type-4
Protein disulfide-isomerase 2	ATP synthase beta subunit	40S ribosomal protein S12
Heat Shock Protein family member (hsp-3)	Mitochondrial ATP synthase alpha subunit precursor	Periostin
Heat Shock Protein family member (hsp-1)	Annexin A6	Acetylcholine receptor subunit alpha-L1
Troponin C	**Antioxidant proteins**	Nucleosome remodelling factor – 38kD CG4634-PA
Putative LEA III protein isoform 2	Thiol-specific antioxidant protein	Coiled-coil domain-containing protein 25
GDP-Mannose Dehydratase	Superoxide dismutase [Cu-Zn]	Calreticulin
Tropomyosin	Peroxiredoxin-5, mitochondria	Lipoprotein-related protein
Dormancy related protein	Peroxiredoxin-4	14-3-3 protein beta/alpha-2 (Protein 14-3-3B2)
Putative LEA III protein isoform 2	Glutathione S-transferase	60S ribosomal protein L26-1
	Peroxiredoxin-6	Histone H4
		Histone H2B.2

Identified proteins with annotation are listed in 8 different groups with majority in protein enzymes. We also identified many heat shock proteins and proteins, which are involved in embryonic development, response to stress/heat and dormancy.
doi:10.1371/journal.pone.0009502.t001

Table 2. Identified proteins with annotation.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
A30*	elongation factor 1-alpha [Milnesium tardigradum]	gi|4530101	(EZ048811)	544	5	K.YAWVLDK.L (23)	51%
A27*, A28*, A29*, A32*,						R.LPLQDVYK.I (52)	
A33*, A35*, A40*, A41*,						K.IGGIGTVPVGR.V (56)	
A42*, A49*, A50*, A59*,						R.EHALLAYTLGVK.Q (65)	
A81*, A84^, A85^, A88*,						K.YYVTIIDAPGHR.D (67)	
A107^, B78^, B80*, B81*,						K.MDSSEPFSEDR.F + Oxidation (M) (72)	
C28°, C36°, D120°						R.NGYTPVLDCHTAHIACK.F (18)	
						K.MDSSEPFSEDRFNEIVK.E (12)	
						K.TLLEALDSISPPARPTDKPLR.L (69)	
						R.VETGVIKPGMVVTFAPTGLTTEVK.S (34)	
						K.NMITGTSQADCAVLVTPAPPGEFEAGISK.N (16)	
						K.SGDAAIVNLIPTKPLCVEAFSEYPPLGR.- (45)	
A79	DB:Swissprot Frame:3 orf:3 Homolog:Angiopoietin- related protein 1 Evalue:1e-29 Bitscore:130		EZ048825	35	1	R.VFYTSDVPDPNR.C (35)	5%
A84	DB:Swissprot Frame:1 orf:8 Homolog:Cathepsin K Evalue:1e-16 Bitscore:73.6		GH986829	58	1	K.LSEEFVR.D (13)	16%
A85						R.WSDVTRPGCK.G (46)	
A85	DB:Swissprot Frame:1 orf:7 Homolog:Actinidain Evalue:1e-11 Bitscore:70.5		EZ048769	33	1	R.NSWGPNWANK.G (33)	18%
	DB:Swissprot Frame:3 orf:3 Homolog:Glucan endo- 1,3-beta-glucosidase A1 Evalue:1e-23 Bitscore:110		EZ048807	28	1	K.EMFSVNDSPNKR.L + Oxidation (M) (28)	5%
A90	DB:Trembl Frame:-3 orf:1 Homolog:GF11309 Evalue:1e-06 Bitscore:57.8		EZ048774	192	3	R.GAVSCIDSFVNR.C (68)	20%
						R.FNPQQPASILQDR.K (74)	
						K.DSLSQTQFTELCTR.S (49)	
	spaghetti CG13570-PA [Drosophila melanogaster]	gi|17864228		55	1	K.ILGAGFDSDTFADLLR.T (55)	2%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
A104	DB:Swissprot Frame:2 orf:3 Homolog:Prostatic acid phosphatase Evalue:3e-24 Bitscore:112		GH986832	162	4	R.YSSYLGPK.F (53)	48%
						K.TWNNELGQLTSK.G (56)	
						K.FSIPEVLIVSSAVER.A (37)	
						R.AVQSTLVNAAGLFTPSGDTIWNSGSSEIGK.T (17)	
A89				200	4	R.YSSYLGPK.F (44)	37%
						R.SPIFTFFTDPYGK.T (63)	
						K.FSIPEVLIVSSAVER.A (48)	
						K.GMQQMYQLGQYISAR.Y + 2 Oxidation (M) (45)	
A111*	prohibitin [Aedes aegypti]	gi|57131967	(EZ048795)	121	2	K.FNASQLITQR.Q (54)	7%
						R.VLPSICNEVLK.G (67)	
A121	mitochondrial ATP synthase alpha subunit precursor [Strongylocentrotus purpuratus]	gi|47551121		97	1	R.VLSIGDGIAR.V (45)	3%
	ZK829.4 [Caenorhabditis elegans]	gi|17544676		57	1	R.VVDALGTPIDGK.G (54)	5%
						K.CAVVDVPFGGAK.G (53)	
						K.GFLGPGVDVPAPDMGTGER.E + Oxidation (M) (4)	
A122,	DB:Swissprot Frame:1 orf:3 Homolog:Glutathione S- transferase 1 Evalue:1e-39 Bitscore:164		EZ048812	439	8	K.LSQYIER.I (38)	45%
						K.VDGIIDFFK.D (65)	
B170, B175						K.QVAQSAAILR.F (65)	
						R.FNLSGKDEFEK.A (72)	
						K.FFSTDVHQYIK.T (42)	
						K.DMQSSMVTWYR.E (66)	
						R.FAFAYAGQQFEDNR.I (44)	
						K.EQMPFGQLPILEVDGK.Q + Oxidation (M) (47)	
B153	DB:Swissprot Frame:3 orf:2 Homolog: Glutathione S-		EZ048805	260	7	K.YILGNDVK.Y (19)	55%
B154	transferase Evalue:5e-44 Bitscore:177					R.YLLEYVGEK.Y (43)	
						K.SYDQFETQPK.W (31)	
						K.QYQNLADYHK.R (5)	
						R.LMYMSQDFEK.E + Oxidation (M) (32)	
						K.HYDMFSQFLGNK.K + Oxidation (M) (32)	
						K.LTQSTAIMHFLAR.K + Oxidation (M) (70)	
						K.QSLGLPFPNIPYYIDGNTK.L (2)	

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
B166	DB:Swissprot Frame:2 orf:1 Homolog: Glutathione S-transferase Evalue:3e-43 Bitscore:176		EZ048770	213	6	K.TEEEQQQCDMVEGALSDFR.Q + Oxidation (M) (25) K.QYLLGSDIK.Y (25)	31%
B158						R.YILEYVGEK.Y (43) K.LMYGSQDFEK.D + Oxidation (M) (23) K.LTQSNAILHHLAR.K (63) K.LMYGSQDFEKDK.S + Oxidation (M) (41) K.SEEEQQQCDMIEGALHDFR.M + Oxidation (M) (18)	
B169	DB:Swissprot Frame:1 orf:3 Homolog: Probable glutathione S-transferase 9 Evalue:8e-24 Bitscore:110		GH986911	96	2	R.LLFAAADQK.Y (48) K.VLAQTTSIVR.Y (48)	20%
D166	DB:Swissprot Frame:3 orf:3 Homolog: Glutathione S-transferase 1 Evalue:3e-30 Bitscore:131		GH986673	48	2	K.DMLVAMQR.W (14) K.LKGEEIMDYMK.D (11) K.DQTPYGQLPILEVDGMK.I (23)	18%
D159	DB:Swissprot Frame:1 orf:3 Homolog: Probable glutathione S-transferase 6 Evalue:2e-34 Bitscore:146		EZ048796	405	8	R.IIFDENDK.S (56) R.SFEQFFEK.Y (31) R.IIFDENDKSK.G (43) K.FTEATFPASLR.S (47) R.KFTEATFPASLR.S (63) R.LIFHGTGEDFEDVR.L (61) R.TEEEALADSVVDATNDILGDLIR.I (48) K.SRTEEALADSVVDATNDILGDLIR.I (58)	33%
A128	DB:Swissprot Frame:3 orf:2 Homolog:Myophilin Evalue:1e-33 Bitscore:143		EZ048783	273	6	R.NFSDEQLR.Q (35) R.LANEIQPGSIR.K (43) R.AAEVCEWVNK.I (38) K.IILGENVLSTSGK.M (84) R.QGETMISLQYGSNK.G (48) K.QNLNAVVICLESLGR.K (25)	40%
A147	DB:Swissprot Frame:3 orf:2 Homolog:Adenylate kinase isoenzyme 1 Evalue:6e-42 Bitscore:171		EZ048787	19	1	K.GFLIDGFPR.E (19)	4%
A148	DB:Swissprot Frame:1 orf:1 Homolog:Peptidyl-prolyl		EZ048822	140	3	K.TSKPVVIADCGQL.- (34)	25%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
	cis-trans isomerase Evalue:2e-75 Bitscore:282					K.TSKPVVIADCGQL.- (59)	
B19	DB:Swissprot Frame:2 orf:1 Homolog: Elongation factor 2 Evalue:1e-72 Bitscore:271		GH986944	22	1	K.HVVFGQVTEGLDIVK.K (49) R.VFSGTVQTGQK.V (22)	6%
B41	PREDICTED: similar to CG8036-PB, isoform B isoform 2 [Apis mellifera]	gi|66503776		125	1	K.LDSDLEGHPTPR.L (48) R.KLDSDLEGHPTPR.L (53)	3%
B82	glutamate dehydrogenase, short peptide [Drosophila melanogaster]	gi|458803		116	1	K.IIAEAANGPTPAADK.I (50) K.TFIVQGFGNVGLHTTR.Y (62)	9%
B88 B9, B10, B89, B95, B96, C22, C36, C47, C83, C87, C117, C122, C124	DB:Swissprot Frame:2 orf:1 Homolog: Vitellogenin-2 Evalue:1e-14 Bitscore:81.6		EZ048823	200	4	K.VSMINLR.L (41) R.AEDEYEWSR.A (40) K.TIVVLPSIYYK.N (51) K.IMVVLPGHSIEITAPQGR.T + Oxidation (M) (68)	14%
B92	NADP-dependent isocitrate dehydrogenase [Homo sapiens]	gi|3641398		69	0	K.DIFQEIYDK.Q (42) R.FKDIFQEIYDK.Q (16)	4%
B95	DB:Swissprot Frame:1 orf:1-3 Homolog:Uncharacterized protein C3orf33 homolog Evalue:3e-06 Bitscore:		GH986689	17	1	R.KDPEAMNETAK.W + Oxidation (M) (17)	6%
B102 B101	DB:Swissprot Frame:3 orf:1 Homolog:Lysosomal acid phosphatase Evalue:3e-10 Bitscore:65.1		EZ048780	90	2	K.FLEPVTVPR.A (52) K.FILYSAHDNTISALLAAFK.A (28) K.NNPNNVFDAPTTVIFPGCSEFCPLDQLR.K (10)	46%
109 A121	mitochondrial malate dehydrogenase precursor [Branchiostoma belcheri tsingtaunese]	gi|33439518		218	3	R.IQDAGTEVVNAK.A (64) R.DDLFNTNASIVR.D (66) K.AGAGSATLSMAYAGAR.F (87)	11%
B146 B141, B144	hypothetical protein TRIADDRAFT _63625 [Trichoplax adhaerens]	gi|195999922		342	4	R.VEIIANDQGNR.I (38) R.ITPSYVAFTADGER.L (91) R.IINEPTAAAIAYGLDK.K (79)	8%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
B173	heat shock cognate 70 [Aedes aegypti]	gi|94468966		235	2	K.NQLTSNPENTVFDVK.R (72) R.IINEPTAAAIAYGLDKK.E (62) K.IQVEYKGETK.N (38) K.MKETAEAYLGK.T + Oxidation (M) (62) R.IINEPTAAAIAYGLDK.K (28) K.STAGDTHLGGEDFDNR.L (50) R.IINEPTAAAIAYGLDKK.T (57)	8%
C131	Heat shock 70 kDa protein II (HSP70 II)	gi|123622		154	2	K.ETAEAYLGK.E (34) K.VEIIANDQGNR.T (60) R.TTPSYVAFTDTER.L (60)	8%
C133	PREDICTED: similar to heat shock cognate 70 protein isoform 2 [Acyrthosiphon pisum]	gi|193603576		153	2	K.VEIIANDQGNR.T (60) R.TTPSYVGFTDTER.L (62) R.IINEPTAAAIAYGLDK.K (16) K.STAGDTHLGGEDFDNR.M (16)	8%
B148	DB:Swissprot Frame:3 orf:2 Homolog:Malate dehydrogenase, cytoplasmic Evalue3e-66 Bitscore:251		GH986821	179	3	R.AIGQMAIQLK.N (52) K.DQGSALNQYAK.K (60) K.IILVVGNPANTNAYILSHYAPSLPK.E (67)	24%
B152	H(+)-transporting ATP synthase [Rattus norvegicus]	gi|57029		92	1	K.LELAQYR.E (31) R.EAYPGDVFYLHSR.L (61)	8%
B164*	ATPase subunit [Beta vulgaris subsp. Vulgaris]	gi|11263	(EZ048779)	64	0	K.LELAQYR.E (34) R.GIRPAINVGLSVSR.V (29)	4%
B167*	DB:Swissprot Frame:1 orf:1 Homolog:Arginine kinase Evalue:5e-90 Bitscore295	(gi|124264768)	EZ048827	254	6	R.FLQAAQAVR.F (41) K.LIDDHFLFK.E (39) K.LNFPNPDPEGK.Y (60) R.KYMTPEIIQK.L + Oxidation (M) (29) R.SLQGFPFNPLLNEQQYK.E (30)	33%
B133^, B157^, B159*^							
B162^, C91*, C98^, C104^							
C107^, C137^, C142^							
D98^, D159^							

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
B183	DB:Swissprot Frame:2 orf:5 Homolog:Peroxiredoxin-5, mitochondrial Evalue:3e-40 Bitscore:150		EZ048816	92	1	K.DLFYPIINDYHVGFDIEK.G (55); R.HLPSYVK.K (10)	24%
C28	pre-mRNA binding K protein, hnRNP K [Xenopus laevis, Peptide, 396 aa]	gi|299029		58	1	K.VHLLADPR.G (9); K.LNIEPDGTGVECSIADR.I (73); R.ILSISADIETIGEILK.K (58)	4%
C36							
C42	Heterogeneous nuclear ribonucleoprotein K [Mus musculus]	gi|13384620		67	0	R.ITAVLSPR.I (43); K.IILLLLSGAK.L (24)	7%
C47	PREDICTED: similar to aconitase, mitochondrial [Nasonia vitripennis]	gi|156537745		58	0	K.NTIVTSYNR.N (25)	2%
C52	peptidase M17 precursor [Clonorchis sinensis]	gi|118429525		55	1	K.IIYSHLDEPQK.Q (33); K.GITYDTGGADVK.A (55)	2%
C60	DB:Swissprot Frame:2 orf:1 Homolog:Gamma-glutamyltranspeptidase 1 Evalue:6e-49 Bitscore:194		GH986789	53	2	K.DMSSPEQDLYHQR.F + Oxidation (M) (31); K.LKEFLTSPQVAQSTR.R (22)	11%
C87	GDP-Mannose Dehydratase family member (gmd-2) [Caenorhabditis elegans]	gi|17507723		61	1	K.FYQASTSELYGK.V (61)	3%
C95	Short-chain dehydrogenase /reductase SDR YhdF [Bacillus licheniformis ATCC 14580]	gi|52079424		110	1	K.GAIVAFTR.S (51); K.TAIITGDSGIGR.A (59)	7%
	DB:Swissprot Frame:2 orf:1 Homolog:Uncharacterized oxidoreductase yhdF Evalue:3e-28 Bitscore:125		(GH986692)	31	1	K.TALITGASTGIGR.A (31)	6%
C98	DB:Swissprot Frame:3 orf:5 Homolog:Protein lethal(2)essential for life Evalue:2e-11 Bitscore:70.9		EZ048820	59	2	K.GYRPEEVTLK.T (15); K.DGVLSVECPLPQGNR.L (44)	30%
	DB:Swissprot Frame:-1 orf:1 Homolog:Probable transaldolase Evalue6e-34 Bitscore:144		GH986571	35	1	K.TVVMGASFR.N + Oxidation (M) (11); K.LLEELANSTAK.V (24)	18%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
C110	aldolase A protein [Homo sapiens]	gi|28595		71	1	K.GILAADESTGSIAK.R (71)	12%
C111	DB:Swissprot Frame:2 orf:1 Homolog;Glyceraldehyde-3-phosphate dehydrogenase Evalue6e-77 Bitscore:2		GH986712	344	7	K.ADVKEQDGQLSINGK.L(63); K.DVDVVAINDPFIDIK.Y(49); K.FGIVEGLMTTVHAFTATQK.V + Oxidation (M) (39); K.TMDIVSNASCTNCLAPLAK.V(77); R.AAIDKDVDVVAINDPFIDIK.Y(47); K.VIISAPSADAPMFVCGVNLDK.Y(33); K.VIISAPSADAPMFVCGVNLDKYDAK.T(35)	51%
C115	DB:Swissprot Frame:1 orf:6 Homolog:Plasminogen Evalue6e-36 Bitscore:84.3		EZ048798	63	2	K.GDFDEFIR.I(34); R.AYSGGISADMLCGAAPGK.D(29)	19%
D159				121	2	R.GCAQPNYPGVYGR.M(46); K.DSCQGDSGGPIVFLK.N(75)	21%
C126	DB:Swissprot Frame:2 orf:4 Homolog:Proteasome subunit alpha type-4 Evalue4e-81 Bitscore:300		GH986859	47	1	R.TTIFSPEGR.L(47)	4%
C128	expressed hypothetical protein [Trichoplax adhaerens]	gi|196010133		105	1	K.VGASEATLLNMLK.V(105)	4%
	F25H2.10 [Caenorhabditis elegans]	gi|17506815		97	1	K.TSFFQALQIPTK.I(97)	3%
	DEAD-box family (SNF2-like) helicase, putative [Theileria annulata]	gi|84996109		54	1	K.MLELISNIIK.K(54)	0%
	ResB family protein [Hydrogenobaculum sp. Y04AAS1]	gi|195953863		54	1	K.MLELISNIIK.K(54)	1%
C132	DB:Swissprot Frame:1 orf:1 Homolog:Peroxiredoxin-4 Evalue4e-86 Bitscore:318		EZ048818	393	8	R.GLFIDK.K(28); R.GLFIDKK.G(32); K.TQIGKPAPDFK.G(27); K.FENVNLSDYK.G(57); R.QITMNDLPVGR.S(51); K.GKFENVNLSDYK.G(59); R.CNVYGSGDVYPER.S(58)	31%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
C139	DB:Swissprot Frame:1 orf:1 Homolog:ATP synthase subunit d, mitochondrial Evalue:4e-27 Bitscore:121		EZ048797	81	2	K.DYGVYLEDAGHTLR.G (81) K.VLAFPESPAK.I (33)	13%
D166						R.VPVPGLVDQFR.K (48)	
C143	Glyceraldehydes-3-phosphate dehydrogenase [Achlya bisexualis]	gi|7274154		107	1	R.VPVPDVSVVDLITVR.L (107)	4%
C145^							
C143*	DB:Swissprot Frame:3 orf:2-4 Homolog:Triosephosphate Isomerase B Evalue:4e-69 Bitscore:260	(gi|1351273)	GH986530	281	5	K.AIADVISDWSK.V (67) R.EGNQETVVFR.Q (47)	33%
B163^						K.DVGAEWVILGHSER.R (82) K.VVIAYEPVWAIGTGK.T (47) K.EASGAFTGEISPAMLK.D (38)	
C145	DB:Swissprot Frame:2 orf:3 Homolog:Peroxiredoxin-4 Evalue:4e-65 Bitscore:247		EZ048781	113	3	K.IGSPAPDFK.A (41) R.GLFIIDQK.G (37) K.AVAVIDGQFQDIQLSTLK.G (35)	19%
	thiol-specific antioxidant protein [Homo sapiens]	gi|438069		54	1	R.QITVNDLPVGR.S	5%
C148	RecName: Full=Aspartic protease inhibitor 8	gi|124012		60	0	R.GALGGDVYLGK.S (12) R.GALGGDVYLGK.S (48)	10%
C155	DB:Swissprot Frame:1 orf:1 Homolog:Superoxide dismutase [Cu-Zn] Evalue:2e-48 Bitscore:192		GH986811	401	6	R.VTSAVAVMK.G (45) R.LACGIVGVVGGTK.- (69) R.VTSAVAVMKGDSPVK.A + Oxidation (M) (32) R.GLPAAESKIHGNSGGR.L (70) R.HVGDLGNLVADASGTAK.I (137) K.IDITDSLMSLMGEHSIVGR.A + 2 Oxidation (M) (48)	33%
D13	DB:Swissprot Frame:1 orf:1 Homolog:40S ribosomal protein S12 Evalue:2e-34 Bitscore:144		GH986534	22	1	K.LDADSLPR.K (22)	6%
D53	PREDICTED: similar to alpha actinin CG4376-PB [Tribolium castaneum]	gi|91080533		120	1	R.VGWEQLLTSINR.N (47) R.NINEVENQILTR.D (58)	3%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
D56	DB:Swissprot Frame:1 orf:3 Homolog:Periostin Evalue:1e-10 Bitscore:67.8		EZ048782	22	1	K.QTEGETVFIPDDAAFGK.M (22)	6%
D57	protein disulfide isomerase-3 [Haemaphysalis longicornis]	gi|148717319		68	0	K.HGVSGYPTLK.I (48)	3%
D61	Enolase (2-phosphoglycerate dehydratase) (2-phospho-D-glycerate hydro-lyase)	gi|1169533		94	1	R.GNPTVEEVTTDK.G(74)	6%
D109						K.VKIGMDVASSFYK.D + Oxidation (M) (20)	
D61	DB:Swissprot Frame:1 orf:1 Homolog:Matrix metalloproteinase-17 Evalue:2e-06 Bitscore:50.8		GH986535	17	1	R.FEVAEGFPK.S (17)	16%
D62	DB:Trembl Frame:1 orf:1 Homolog: Vitellogenin 1		EZ048784	477	9	K.FGNNIGQNIEK.Y (46)	34%
B19, C13, C15, C16, C28,	Evalue:2e-05 Bitscore:53.1					K.VLFDGNYVEIK.A (59)	
C33, D29, D30, (D32-D39),						K.KFGNNIGQNIEK.Y (75)	
D42, D50, D58, E11, E12,						K.EPILAIVSPVTGLK.V (74)	
E15, E45, E54, E65, E66,						R.AYLLQEGSCNAQIPQDK.K (42)	
E73, F26						R.AYLLQEGSCNAQIPQDK.V (36)	
						R.DELFAVLAANANPSASPLEIR.R (75)	
						K.VSEYTILYNGQPIPQPPTEGK.F (22)	
						-.DNSRDELFAVLAANANPSASPLEIR.R (48)	
D81	DB:Swissprot Frame:2 orf:1 Homolog:Actin-1 Evalue6e-87 Bitscore:319		GH986913	33	1	K.EISALAPNTIK.- (33)	5%
D84	UspA [Bacillus coagulans 36D1]	gi|124521548		56	1	R.IILVAIDGSK.M (56)	6%
	DB:Swissprot Frame:3 orf:1 Homolog:Aldehyde dehydrogenase, mitochondrial Evalue2e-47 Bitscore:188		EZ048791	160	4	K.ALYLSQGIR.A (38)	37%
						K.YGLAASVMTK.D + Oxidation (M) (22)	
						-.GYFIEPTVFADVK.D (48)	
						R.ELGEYGLDAYTEVK.T (53)	
D90	DB:Swissprot Frame:2 orf:1 Homolog:Annexin A6 Evalue:1e-37 Bitscore:143		EZ048803	223	6	K.DLFDDLKK.E (24)	48%
D89, D96						R.DHYNPTIR.A (21)	
						K.GIGTDEDTVIK.I (44)	

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
D91*	elongation factor 1 gamma [Bombyx mori]	gi|112983898	(EZ048793)	60	1	R.HLLFAITTR.R(47) R.AFQPFNPDNDAK.A(38) R.EVIDDIVSDTSGYFR.H(43) K.AIAGAGTSEEDLIEIMLTR.N + Oxidation (M) (6) K.VPAFFSADGK.V(58)	2%
D104°, D106°							
D92	mitochondrial long-chain enoyl-CoA hydratase/3-	gi|510108		57	0	K.ALTSFER.D(7)	3%
	hydroxycyl-CoA dehydrogenase alpha-subunit [Rattus n					K.DGPGFYTTR.C(34)	
	DB:Swissprot Frame:3 orf:1 Homolog: Peroxidase Evalue:7e-19 Bitscore:94.0		EZ048773	30	1	K.VIGMHYFSPVDK.M(16) R.TGFTTDQMAIIK.K + Oxidation (M) (30)	7%
D96	cysteine conjugate beta-lyase [Takifugu rubripes]	gi|5002565		84	1	K.ALVINTPNNPLGK.V(84)	3%
D96	DB:Swissprot Frame:3 orf:5 Homolog:26S proteasome non-ATPase regulatory subunit 13 Evalue:1e-29 Bit		GH986860	100	3	K.LLEEVEK.K(12) K.KLLEEVEK.K(26) R.SAGGMSELYK.N(32) R.LHGTYAEYFR.E(31)	15%
D103	DB:Swissprot Frame:1 orf:1 Homolog:Neprilysin-2 Evalue:9e-31 Bitscore:133		EZ048772	32	1	K.ITAQYSNFR.Y(32)	6%
D107	Tubulin alpha-3 chain (Alpha-III tubulin)	gi|3915094		252	2	R.LSVDYGK.K(27) R.QLFHPEQLITGK.E(31) R.LLGQIVSSITASLR.F(29) R.AVFVDLEPTVIDEIR.T(64) R.NLDIERPTYTNLNR.L(81) R.FDGALNVDLTEFQTNLVFYPR.I(21)	18%
D110	Tubulin beta-3 chain (Beta-tubulin class-IV)	gi|135464		152	1	R.FPGQLNADLR.K(70) K.LAVNMVPFPR.L(36)	7%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
D111	GH19645 [Drosophila grimshawi]	gi|19053606		120	2	R.AVIVDLEPGTMDSVR.S (46) K.KGIDAEVINLR.S (56) R.VFLLGEEVAQYDGAYK.V (64)	7%
D118	DB:Swissprot Frame:1 orf:4 Homolog:Histone H4 Evalue9e-39 Bitscore:160		GH986770	18	1	R.ISGLIYEETR.G (18)	12%
	DB:Swissprot Frame:-3 orf:1 Homolog:Histone H2B.2 Evalue:8e-32 Bitscore:135		EZ048778	15	1	K.LILPGELAK.H (15)	9%
D140	DB:Swissprot Frame:2 orf:3 Homolog:Acetylcholine		EZ048771	381	7	K.LGSWTFAK.D (51)	23%
D128	receptor subunit alpha-L1 Evalue:1e-16 Bitscore:87					R.LQYTDSAVK.K (34) R.LQYTDSAVKK.I (34) K.DELDVQTSQSK.F (68) R.AFLSLNWQDHR.L (80) K.FDDYFQSSVWK.F (61) K.LGSWTFAKDELDVQTSQSK.F (53)	
D159	DB:Swissprot Frame:2 orf:1 Homolog:Peroxiredoxin-6 Evalue:8e-61 Bitscore:233		GH986904	403	8	K.LAPEFEK.R (38) R.NFDELLR.V (27) R.VLDSLQLVSK.H (63) K.HSVVTPVDWK.- (69) K.LVLIYPATSGR.N (50) K.DLESYCGMGGGK.F + Oxidation (M) (48) K.MIALSCDDAQSHQGWIK.D + Oxidation (M) (40) K.FGMLDPDELNSNNMPVTAR.A + Oxidation (M) (68)	53%
E4	phosphoglycerate kinase[Verrucomicrobiae bacterium V4]	gi|161075769		54	1	K.AIGFLMEKELK.Y + Oxidation (M) (54)	2%
E5	Rubber elongation factor protein (REF) (Allergen)	gi|132270		104	1	R.SLASSLPGQTK.I (33)	18%
D99	Hev b 1)					K.FVDSTVVASVTIIDR.S (71)	
E5	Small rubber particle protein (SRPP) (22 kDa rubber particle protein) (22 kDa RPP) (Latex allergen	gi|14423933		87	0	K.AEQYAVITWR.A (43) R.IVLDVASSVFNTGVQEGAK.A (44)	14%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
	DB:Swissprot Frame:2 orf:1 Homolog:Liver						
	carboxylesterase Evalue:2e-33 Bitscore:142		EZ048809	21	1	K.AIVVAVNYR.V (21)	5%
E43*	actin [Heliothis virescens]	gi|14010639	(EZ048826)	667	7	K.EITALAPSTMK.I (41)	41%
						R.AVFPSIVGRPR.H (73)	
						K.IWHHTFYNELR.V (73)	
						K.QEYDESGPSIVHR.K (94)	
						K.SYELPDGQVITIGNER.F (79)	
						R.VAPEEHPVLLTEAPLNPK.A (90)	
						K.YPIEHGIITNWDDMEK..I (56)	
						K.DLYANTVLSGGTTMYPGIADR.M (45)	
						R.KDLYANTVLSGGTTMYPGIADR.M+Oxidation (M) (36)	
						R.TTGIVLDSGDGVSHTVPIYEGYALPHAILR.L (55)	
E43*	DB:Swissprot Frame:1 orf:2 Homolog: Actin-5C		EZ048826	471	9	R.DLTDYLMK.I (25)	37%
D99, D106, D108, E71,	Evalue:7e-155 Bitscore:547					R.GYSFVTTAER.E (38)	
E72, E84, E92, E94, E99,						K.EITALAPSTMK.I (41)	
E100, F44, F58, F61, F95						K.AEYDESGPSIVHR.K (112)	
						K.SYELPDGQVITIGNER.F (79)	
						K.DLYANTVLSGGTTMYPGIADR.M (45)	
						R.KDLYANTVLSGGTTMYPGIADR.M+Oxidation (M) (36)	
						K.LCYVALDFEQEMATAAASSLEK.S (39)	
						R.TTGIVLDSGDGVSHTVPIYEGYALPHAILR.L (55)	
E47*	Actin, muscle-type (A2)	gi|3121741	(EZ048826)	519	6	K.RGILTLK.Y (23)	30%
F64*, F95°						K.AGFAGDDAPR.A (62)	
						R.DLTDYLMK.I (24)	
						R.GYSFVTTAER.E (40)	
						K.EITALAPSTMK.I + Oxidation (M) (40)	
						R.AVFPSIVGRPR.H (59)	
						K.IWHHTFYNELR.V (60)	
						K.QEYDESGPSIVHR.K (58)	
						K.SYELPDGQVITIGNER.F (79)	
						R.VAPEEHPVLLTEAPLNPK.A (75)	

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
E48*	actin 5C [Lycosa singoriensis]	gi|161661023	(EZ048826)	644	6	K.AGFAGDDAPR.A (80)	35%
E44*						R.DLTDYLMK.I + Oxidation (M) (15)	
						R.GYSFVTTAER.E (44)	
						R.AVFPSIVGRPR.H (98)	
						K.IWHHTFYNELR.V (66)	
						K.QEYDESGPSIVHR.K (100)	
						K.SYELPDGQVITIGNER.F (70)	
						R.VAPEEHPVLLTEAPLNPK.A (83)	
						K.YPIEHGIITNWDDMEK.I + Oxidation (M) (25)	
						K.DIYANTVLSGGTTMYPGIADR.M (47)	
						R.KDLYANTVLSGGTTMYPGIADR.M+Oxidation (M) (18)	
E50*	beta-actin [Rachycentron canadum]	gi|161376754	(EZ048826)	501	4	K.AGFAGDDAPR.A (59)	33%
E6*, E42*, E49*, F27*,						K.EITALAPSTMK.I (49)	
F28*						R.AVFPSIVGRPR.H (73)	
						K.IWHHTFYNELR.V (46)	
						K.SYELPDGQVITIGNER.F (79)	
						M.EEEIAALVVDNGSGMCK.A (50)	
						R.VAPEEHPVLLTEAPLNPK.A (43)	
						K.DIYANTVLSGGTTMYPGIADR.M (24)	
						R.KDLYANTVLSGGTTMYPGIADR.M+Oxidation (M) (53)	
E52*	alpha-actin (aa 40-375) [Mus musculus]	gi|49864	(EZ048826)	106	0	R.GYSFVTTAER.E (34)	11%
E85*, D103*						K.EITALAPSTMK.I + Oxidation (M) (22)	
						K.SYELPDGQVITIGNER.F (50)	
E57	muscle actin	gi|797290		290	1	R.DLTDYLMK.I (24)	25%
D104*, D97*, E69*,						R.GYSFVTTAER.E (43)	
E68*,E67*, E59*						K.EITALAPSTMK.I (49)	
E58*, E55*, E53*, E93°						K.IWHHTFYNELR.V (39)	
						K.QEYDESGPSIVHR.K (58)	
						R.VAPEEHPVLLTEAPLNPK.A (39)	
						K.YPIEHGIITNWDDMEK.I (19)	
E63	DB:Swissprot Frame:1 orf:1 Homolog:AFG3-like protein 2 Evalue:9e-58 Bitscore:221		GH986706	35	1	K.CFELLSEK.K (9)	12%
						K.GLGYAQYLPR.E (27)	

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
E64	heat shock protein 90 alpha [Fundulus heteroclitus macrolepidotus]	gi|77999578		62	1	R.FYTSASGDEMVSLK.D+ Oxidation (M) (62)	6%
E70* D102*, E51*, E56*	actin [Paraphidippus aurantius]	gi|167683068	(EZ048826)	376	2	R.DLTDYLMK.I (15) R.GYSFVTTAER.E (41) K.EITALAPSTMK.I (46) K.IWHHTFYNELR.V (40) K.SYELPDGQVITIGNER.F (67) R.VAPEEHPVLLTEAPLNPK.A (59) K.YPIEHGIITNWDDMEK.I (8) R.TTGIVLDSGDGVSHTVPIYEGYALPHAILR.L (16)	42%
E73	PREDICTED: similar to Nucleosome remodelling factor – 38kD CG4634-PA [Apis mellifera]	gi|66507623		85	1	K.GDNDPIDVLEIGYK.V (85)	1%
E74	DB:Swissprot Frame:1 orf:4 Homolog: Protein		GH986548	118	3	K.SLAPEYAK.A (16) K.DNFEDALK.E (21) K.VDATVETDLATK.Y (80)	17%
E76, E88, F62	disulfide-isomerase Evalue:2e-44 Bitscore:178						
E80	DB:Swissprot Frame:1 orf:3-5 Homolog: Cathepsin Z Evalue:1e-63 Bitscore:243		GH986945	102	2	K.VGDFGPISGR.E (47) K.TFNQCGTCSEFGK.C (55)	11%
E83	DB:Swissprot Frame:3 orf:1 Homolog:26S proteasome non-ATPase regulatory subunit 8 Evalue:4e-48 Bits		EZ048799	154	4	K.DLIPDSSLR.T (40) R.IYYYDWK.D (5) K.CEALLNQIK.V (21) R.DVLEMGAQLAILK.R (45) R.ACPEVNLNSLCR.M (44)	23%
E86	DB:Swissprot Frame:3 orf:4 Homolog: Coiled-coil domain-containing protein 25 Evalue:2e-46 Bitscore:1		EZ048808	21	1	K.LSSAHVYLR.L (21)	6%
E89	ATP synthase beta subunit [Asteria miniata]	gi|46909233		346	3	K.AHGGYSVFAGVGER.T (33) R.FTQAGSEVSALLGR.I (97) R.VALTGLTVAEYFR.D (84) K.TVLIMELINNVAK.A (62)	24%
D140						K.VALVYGQMNEPPGAR.A + Oxidation (M) (38)	

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
E46				191	0	R.GIAELGIYPAVDPLDSTSR.I (26)	13%
						R.EGNDLYHEMIEGGVISLK.D + Oxidation (M) (7)	
						K.IGLFGGAGVGK.T (41)	
						R.IPVGPETLGR.I (34)	
						K.VVDLLAPYAK.G (40)	
						R.TIAMDGTEGLIR.G + Oxidation (M) (44)	
						R.FTQAGSEVSALLGR.I (33)	
E89	DB:Swissprot Frame:2 orf:1 Homolog:Rab GDP dissociation inhibitor beta Evalue:4e-51 Bitscore:201		GH986887	152	3	K.VALELLGPIR.Q (54)	45%
						R.GTGQVFDFTK.V (55)	
						R.CICLLDHPIPNTK.D (6)	
						K.DALSTQIIIPQNQVNR.N (33)	
						R.NNDIYISVVSYTHQVAAK.G (4)	
E92	DB:Swissprot Frame:1 orf:1		GH986892	75	2	K.TVTSLWR.E (22)	11%
						R.ASFAGDMNFYGK.A (54)	
E93	Homolog:Methylmalonate-semialdehyde dehydrogenase [acylating], mitochond						
E98	DB:Swissprot Frame:3 orf:1 Homolog:Thioredoxin reductase 1, cytoplasmic Evalue:2e-73 Bitscore:275		GH986518	50	1	R.TACTAEIGLDK.V (50)	5%
F3	DB:Swissprot Frame:2 orf:1 Homolog:Protein disulfide-isomerase 2 Evalue:3e-64 Bitscore:244		EZ048794	527	9	R.IDSFPTIK.L (43)	56%
F1, F2, F27, F43, F49						R.ITEFFGLTK.D (57)	
						K.NFDEVVMDK.S (56)	
						R.LISLADQIVK.Y (46)	
						K.GDNTVVEYGGER.T (38)	
						K.MDATANELEHTR.I + Oxidation (M) (72)	
						K.KGDNTVVEYGGER.T (62)	
						K.LSPIYDELGDHFK.D (79)	
						K.YKPEAGDLNPETLTK.F (64)	
						K.LKPHLNSQDVPEDWNAK.S (10)	
F6	Tubulin beta-1 chain (Beta-tubulin class-I)	gi|57429		54	0	R.YLTVAAIFR.G (15)	4%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
F12	hsp 108 [Gallus gallus]	gi|63509		84	0	R.FPGQLNADLR.K (39) R.ELISNASDALDK.I (50) K.GVVDSDDLPLNVSR.E (31)	4%
F13	hsp 82 [Drosophila pseudoobscura]	gi|9069		77	1	R.ELISNASDALDK.I (77)	3%
F14* B195^, B198^	DB:Swissprot Frame:3 orf:2 Homolog:Calreticulin Evalue:2e-69 Bitscore:262	gi|195107681	GH986835	365	7	K.FVWTAGK.F (23) K.FYGDADLNK.G (42) K.VFPSTMDQK.D (48) R.FYGLSAAFPK.F (41) K.DKPLVIQFSVK.H (59) K.HEQNIDCGGGYVK.V (88) K.EQFLDNKWEDR.W (65)	40%
F38	DB:Swissprot Frame:2 orf:1 Homolog:Calreticulin Evalue4e-19 Bitscore:94.7		GH986920	159	3	R.EIPNPAYK.G (19) K.AAEDFANDTWGK.T (85) K.SGTIFDDIIITDDIK.A (54)	47%
F21	Heat Shock Protein family member (hsp-3) [Caenorhabditis elegans]	gi|17568549		180	0	R.LSPEDIER.M (39)	6%
F22, F23						K.FDLTGLPPAPR.G (43) K.FEELNMDLFR.A (47)	
F24	heat shock protein 90-beta [Danio rerio]	gi|18858875		83	0	K.HFSVEGQLEFR.A (42) K.EKYIDQEELNK.T (16)	5%
F24	heat shock protein 90 [Danio rerio]	gi|555574		62	1	K.ADLMNNLGTIAK.S + Oxidation (M) (62)	11%
F24	DB:Swissprot Frame:2 orf:3 Homolog:Heat shock protein 83 Evalue:1e-84 Bitscore:312		EZ048788	214	3	R.YMSLTDPK.Q + Oxidation (M) (6) K.ADMINNLGTIAK.S (73) K.EDQMDYVEEK.K (47) R.ELISNSSDALDK.I (89)	21%
F25	Heat shock protein HSP 90-alpha	gi|17865490		275	1	R.ALLFVPR.R (20) R.APFDLFENR.K (39) R.RAPFDLFENR.K (41) R.ELISNSSDALDK.I (89) K.SLTNDWEDHLAVK.H (48)	8%
F27	heat shock protein 70 [Liriomyza huidobrensis]	gi|89892741		225	3	R.FEELCADLFR.S (68) K.NQIHDIVLVGGSTR.I (68) R.ARFEELCADLFR.S (58)	4%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
	Heat Shock Protein family member (hsp-1) [Caenorhabditis elegans]	gi|17541098		219	2	K.MDKNQIHDIVIVGGSTR.I + Oxidation (M) (31); R.LSKDDIER.M(52); R.FEELCADLFR.S(68); R.ARFEELCADLFR.S(58); K.SINPDEAVGAAVQAAILSGDK.S(41)	6%
F32	DB:Trembl Frame:1 orf:3 Homolog:Lipoprotein- related protein Evalue:7e-07 Bitscore:56.2		GH986605	19	1	K.VLASIDLTGK.T(19)	9%
F35*	tropomyosin	gi|42559676		273	2	K.IVELEEELR.V(55); K.LAMVEADLER.A(44); R.EDSYEEQIR.L(74); R.KLAMVEADLER.A(51); K.ALQREDSYEEQIR.L(18)	17%
F36*, F53*	DB:Swissprot Frame:1 orf:2 Homolog:Tropomyosin- 1, isoforms 9A/A/B Evalue:3e-30 Bitscore:131		GH986919	229	3	R.IQLLEDDLER.T(69); K.LSEASQAADESER.A(93); R.IQLLEDDLERTEER.L(67)	18%
	DB:Swissprot Frame:2 orf:1 Homolog:Tropomyosin Evalue:3e-11 Bitscore:66.6		GH986674	48	1	R.LEDDLVHEK.E(35); K.EVDRLEDDLVHEK.E(13)	38%
F36, F33, F37, F64, F69, F71, F72	DB:Trembl Frame:2 orf:1 Homolog:Putative uncharacterized protein Evalue:1e-18 Bitscore:97.1		EZ048810	109	2	R.DYPIYNDIPSTR.F(51); R.QAGFYADAEAQCQVIR.R(58)	28%
F39, F34, F42, E64, E90	DB:Trembl Frame:2 orf:3 Homolog: CG14304-PA Evalue9e-18 Bitscore:95.5		EZ048802	104	2	K.DYPTYNEIPNTR.F(55); K.QAGFYADIDAQCQAIR.R(49)	32%
F42	DB:Swissprot Frame:-2 orf:1 Homolog:Succinyl-CoA ligase [GDP-forming] subunit beta, mitochondrial E		GH986609	53	2	K.EEQVQEAFR.I(35); K.LPIIANSLEDAAAK.A(19)	17%
F51, E60, E63, E96, F31, F32,	DB:Trembl Frame:2 orf:1 Homolog:AGAP009479- PA Evalue:7e-15 Bitscore:83.2		EZ048819	58	1	K.QAGFYADTDAQCQVIR.R(58)	16%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
F52, F70, F73, F96							
F54	DB:Trembl Frame:1 orf:1 Homolog:Putative uncharacterized protein Evalue:2e-16 Bitscore:88.6		EZ048790	62	1	K.QAGFYADTEAQCQVIR.R (62)	18%
F40, F41, F66, F67, F68							
F55	DB:Swissprot Frame:2 orf:1 Homolog:CD109 antigen Evalue:2e-06 Bitscore:52.4		EZ048800	29	1	R.TVVVYDYNPQDR.K (29)	20%
F56	GTP-specific succinyl-CoA synthetase beta subunit [Homo sapiens]	gi|4406564		59	1	K.INFDDNAEFR.Q (59)	5%
F57	heat shock protein 60 [Salmo salar]	gi|16923167		63	1	K.VGGTSEVEVNEK.K (63)	7%
F65*	14-3-3 protein beta/alpha-2 (Protein 14-3-3B2) [Oncorhynchus mykiss]	gi|82089139	(GH986681)	90	0	R.NLLSVAYK.N (48) R.YDDMAGAMK.S (42)	6%
F73	DB:Swissprot Frame:2 orf:2 Homolog:60S ribosomal protein L26-1 Evalue:6e-45 Bitscore:179		GH986676	21	1	K.LSKDVSSSR.R (21)	6%
F75	DB:Swissprot Frame:2 orf:2 Homolog:Translationally-controlled tumor protein homolog Evalue:1e-57 Bi		EZ048806	234	3	R.LVEVPFLQDK.K (56) K.LVDNVLFEVTGK.Y (86) K.DAVTGDEMFSDSYK.Y (80)	33%
F81 / F80, F89	DB:Swissprot Frame:2 orf:4 Homolog:Myosin regulatory light polypeptide 9 Evalue:4e-50 Bitscore:198		EZ048792	331	6	K.RVQEQSPNEVDQFK.T (8) K.DTFASLGR.A (30) R.DLLGGVGDK.L (57) K.GQLDYVAFAK.L (42) K.LSADEMSQAFK.G + Oxidation (M) (67) K.EAFTMMDQNR.D + Oxidation (M) (38) K.VAGVDPEATITNAFK.L (97)	37%
F81	DB:Swissprot Frame:2 orf:1 Homolog:Cathepsin L1 Evalue:2e-65 Bitscore:249		GH986678	98	1	K.LPDLSEQNLVDCSK.K (98)	5%
F87	DB:Swissprot Frame:3 orf:1 Homolog:Troponin C Evalue:3e-60 Bitscore:231		GH986791	24	1	R.QIGTLLR.T (24)	4%
F88	putative LEA III protein isoform 2 [Corylus avellana]	gi|14148981		69	1	K.AGESQVQDTANAAK.N (69)	16%

Table 2. Cont.

Spot no.	Protein name	NCBInr Accession no. (°)	Tardigrade specific Accession no. (contig/EST) (^)	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage
	glycosyl transferase, family 2[Shewanella sediminis]	gi|157373461		59	1	R.HLLISLADK.Y (59)	1%
	HAW-EB3						
	DB:Swissprot Frame:2 orf:5 Homolog: Myosin,						
	essential light chain Evalue:8e-30 Bitscore:131		EZ048813	404	10	K.EVDEILR.L (20)	58%
						R.HLLISLGEK.L (59)	
						K.ESNGTIIAAELR.H (68)	
						K.DVGTLEDFMEAMR.V (73)	
						K.LTVEEFMPIYGQLSK.E (28)	
						R.VFDKESNGTIIAAELR.H (26)	
						K.EKDVGTLEDFMEAMR.V (23)	
						K.EVFGVYDMFFGDGTNK.V (45)	
						K.LLTVEEFMPIYGQLSK.E (26)	
						K.EVFGVYDMFFGDGTNKVDAMK.V (37)	
F95	SJCHGC06651 protein [Schistosoma japonicum]	gi|56759014		103	0	K.NTTCEFTGDILR.T (52)	9%
						R.TVSGVNGPLVILDDVK.F (51)	

Generated MS/MS data were searched against the NCBInr and tardigarde protein databases. Spot number, protein annotation, accession number, total protein score, number of matched peptides, peptide sequence and sequence coverage are listed. Identical proteins identified in different spots are listed only once and the spot with the highest protein score (in bold) is ranked at the top.

doi:10.1371/journal.pone.0009502.t002

Table 3. Identified proteins without annotation.

Spot no.	Accession no.	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage	DomainSweep analysis
A11	GH986700	52	1	-.VIAVSLPR.N (52)	3%	**No hits**
A82, A88, B33, B41, B43, C50,						
D99, D105, E72, F87						
A11	GH986755	32	1	-.LSISHNATLR.V (32)	4%	**Putative**
						IPR006210EGF
A94	GH986643	39	1	R.VDRSIPR.L (39)	3%	**Putative**
A91, A95, A110, A123, A140,						**IPR004077 Interleukin-1 receptor, type II**
B49, B64, B83, B90, B98, B105,						
B155, B165, B173, B176, B185,						
B186, B187, B188, B189, B190,						
B191, B192, B193, B194, B195,						
C51, C128, C141, C153, D45,						
D46, D56, D57, D74, D123						
A100	EZ048767	229	4	K.YDLIYK.G (15)	20%	**Putative**
				K.FLGFDTAGK.T (61)		**IPR017956 AT hook, DNA-binding,**
				K.IISFDVCNK.N (54)		**conserved site**
				K.TDSGVSCDVTD-KCDPIVK.A (39)		**IPR006689 ARF/ SAR superfamily**
				K.AVVDIEDPNN-SAGDSIDYGK.Y (60)		**IPR005464 Psychosine receptor**
A112	GH986667	317	5	R.EQFTQGCTVGR.N (61)	22%	**Putative**
A114				K.LEAAPNQCPEYK.K (89)		**IPR001749 GPCR, family 2, gastric**
				K.KLEAAPNQCPEYK.K (64)		**inhibitory polypeptide receptor**
				K.IMEVCNEPNTYENVNR.F + Oxidation (M) (44)		**IPR000372 Leucine-rich repeat, cysteine-**
				K.IQSLCTPADLQ-FFQSTHDR.I (60)		**rich flanking region, N-terminal**
						IPR004825 Insulin/ IGF/relaxin
A112	EZ048821	98	2	K.NADPLTILK.E (37)	14%	**Putative**
				K.IQSLCTPADLQ-FFQSTHDR.I (60)		**IPR008355 Interferon-gamma receptor**
						alpha subunit
A114	EZ048817	49	1	R.IGTETTSFDYLR.E (49)	3%	**Putative**
						IPR004354 Meiotic recombination protein
						rec114
A123	EZ048785	221	4	K.FLDFTR.G (28)	17%	**Putative**
				R.AADLDTLTK.L (57)		**IPR000762 PTN/MK heparin-binding**
				R.YLDMDQYDW-DTR.S + Oxidation (M) (54)		**protein**
				R.GTFDTAHIQG-LTALTTLR.L (60)		

Table 3. Cont.

Spot no.	Accession no.	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage	DomainSweep analysis
				R.IMSVDLTDINS-APGMFDAAK.T + 2 – Oxidation (M) (23)		
A136	EZ048814	55	1	R.IPAQFQSK.I (55)	**5%**	**Putative**
						IPR015874 4-disulphide core
B48	EZ048766	273	5	K.QVNAETFQK.A (36)	**24%**	**Putative**
A157, A158, B49, B65				K.YSETVHYEGGK.Q (39)		**IPR000507 Adrenergic receptor, beta 1**
				R.VDYVYSYHTK.M (4)		**IPR000463 Cytosolic fatty-acid binding**
				R.GDFWSTDKPHR.Y (32)		**IPR004825 Insulin/IGF/ relaxin**
				K.YDIALDTVEATLK.S (70)		
				R.LIPDELLGTYEFSGK.Q (93)		
B61	GH986621	231	6	R.VLNNGVLR.V (39)	**13%**	**Putative**
B60, B62, B64, B65, B79, B84,				R.VITVPEGIK.V (49)		**IPR001610 PAC motif (peptide matched in**
B93, B112, B143				R.SLLGEIPITK.G (38)		**frame 4)**
				R.RVITVPEGIK.V (46)		**IPR007758 Nucleoporin, Nsp1-like, C-**
				R.VITVPEGIKVESFK.S (26)		**terminal (peptide matched in frame 6)**
				K.GSLTAGSSSNTSGST-GSSSYSSGTTGSSGTSGGK.T (34)		
B62	EZ048776	230	6	R.VLNNGVLR.V (39)	**18%**	**Putative**
A138, B48, B60, B61, B64, B65,				R.VITVPEGIK.V (49)		**IPR007758 Nucleoporin, Nsp1-like, C-**
B84, B112, B138, B142, B143,				R.SLLGEIPITK.G (38)		**terminal**
B144, B161, B173				R.RVITVPEGIK.V (46)		
				R.VEAPIQVDQLTADQQR.S (93)		
				R.VLNNGVLRVEAPIQ-VDQLTADQQR.S (69)		
B79,	GH986933	38	1	K.NGDVSIPR.Q (38)	**6%**	**No hits**
D67, D109						
B91	GH986939	54	1	R.EALSAVTGGR.R (62)	**9%**	**No hits**
B43, B78-B80, B82, B83, B86,						
B87, B90, B92, B93, B97, B191,						
B193, C12, C51, C71, C112,						
C114, C123, C129, D2-D5, D8,						
D10, D21-D24, D27, D28, D31,						
D44, D47, D105, D118, D123,						
D124						
B102	EZ048815	403	6	K.QVNAETFNK.A (40)	**26%**	**Putative**
A23, A24, A26, A112, A127,				K.GGPAWPKDEK.F (17)		**IPR000507 Adrenergic receptor, beta 1**
B99, B103, B105, B107, B108,				K.ILFRPTLSAR.A (36)		**IPR006080 Mammalian defensin**

Table 3. Cont.

Spot no.	Accession no.	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage	DomainSweep analysis
B110, B111, B144				R.AQGLWEATTEGK.N(68)		IPR002181 Fibrinogen, alpha/beta/gamma
				R.LIPDELLGTFEFSGK.Q(92)		chain, C-terminal globular
				R.RLIPDELLGTFEFSGK.Q(36)		IPR000463 Cytosolic fatty-acid binding
				K.DYEFKEDGNMQMTAK.F + Oxidation (M) (20)		
				K.EVEYTSNYDMALDTVK.A(51)		
				R.MGLGVWESTSEQ ENMLEYLK.A(22)		
				R.GDKPGLAAFGDNIIEYSFTA-DSEGETGVLHGK.F(21)		
B103	EZ048768	40	1	R.VTTVSIPR.I(40)	**3%**	**No hits**
B185, C150, C151, C153						
B150	GH986581	108	3	R.VFVEEQLK.A(33)	**14%**	**Putative**
B151, B173				R.FNFLVFLGSTR.E(46)		**IPR000990 Innexin**
				R.GHTYEIMDPEK.V + Oxidation (M) (29)		
B152	EZ048775	42	1	R.KLEFILXFIF.-(42)	**5%**	**Putative**
						IPR003061 Colicin E1 (microcin) immunity
						Protein
						IPR000048 IQ calmodulin-binding region
B179	GH986603	53	1	R.AFEVPASECGK.S(53)	**5%**	**Putative**
						PR015880 Zinc finger, C2H2-like
						IPR008264 Beta-glucanase
B191	EZ048789	26	1	K.GSIGAPDVPK.N(26)	**4%**	**Putative**
						IPR001955 Pancreatic hormone
B186	GH986708	468	6	R.AFEVPASECGK.S(46)	**25%**	**Putative**
A140				R.AFEVPASECGKSPK.R(82)		**IPR015880 Zinc finger, C2H2-like**
				R.YRAFEVPASECGK.S(36)		**IPR000436 Sushi/SCR/CCP**
				K.IVSKDVCGSSPKPR.K(90)		**IPR008264 Beta-glucanase**
				R.SESGALWSEEQECTAK.F(62)		**IPR000008 C2 calcium-dependent**
				R.SESGALWSEEQ ECTAKFHPR.D(137)		**membrane targeting**
				R.VQVMDKDVGSSDDLVEQ-FECLTGPLVSSR.S+Oxidation (15)		
C18	EZ048777	46	1	R.NLADQAMSMGDGPLNFAK.A + 2 Oxidation (M)	**8%**	**Putative**
						IPR003569 Cytochrome c-type biogenesis
						Protein CcbS
						IPR002282 Platelet-activating factor

Table 3. Cont.

Spot no.	Accession no.	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage	DomainSweep analysis
						receptor
C78	GH986847	32	1	K.SEVFPRIR.S(32)	3%	Putative
B188, B173, C141						IPR003916 NADH-ubiquinone
						oxidoreductase, chain 5
C86	GH986916	196	4	K.NPYLELTDPK.-(38)	12%	Putative
				K.TPEESEAPQAIR.R(68)		IPR000863 Sulfotransferase
				K.TPEESEAPQAIRR.K(58)		IPR003504 Glial cell line-derived
				K.VEKTPEESEAPQAIR.R(32)		neurotrophic factor receptor alpha 2
C95	GH986921	35	1	-.VIAVSLPR.N(30)	2%	No hits
B18, B19, B47, B49, B138, C51,						
C62, C65, D107						
C95	GH986692	31	1	K.TALITGASTGIGR.A(31)	5%	Significant
						IPR002347 Glucose/ ribitol dehydrogenase
						IPR002198 Short-chain
						dehydrogenase/ reductase SDR
						Putative
						IPR003560 2,3-dihydro-2,3-
						dihydroxybenzoate dehydrogenase
						IPR002225 3-beta hydroxysteroid
						dehydrogenase/ isomerase
C110	GH986711	31	1	K.ERSPLANK.I(31)	4%	Putative
						IPR006210 EGF
C118	EZ048824	45	0	K.DSVAIGFPK.D(24)	7%	Putative
				K.ADEAGFTDAIK.A(21)		IPR003535 Intimin bacterial adhesion
						mediator protein
C141	EZ048801	395	6	R.NQVYQSMER.H(34)	22%	Putative
C117, C145				R.QNIDAIEIPR.L(78)		IPR002546 Myogenic basic muscle-
				K.DFLSAVVNSIQR.R(58)		specific protein
				R.LSQLAVDSVEIAK.D(74)		IPR000795 Protein synthesis factor, GTP-
				R.MTISEPFESAEALK.D + Oxidation (M) (72)		binding
				R.LEDVDDVLMSAFGMLK.A + 2 Oxidation (M) (26)		
				R.MTISEPFESAEALKDMIVR.L + 2 Oxidation (M) (15)		
				R.LQSSPTLSSL VDQDTFELIR.Q(37)		
C141	GH986597	27	1	-.TAVEAVVR.T(27)	4%	Putative
						IPR003065 Invasion protein B
C156	EZ048804	277	5	K.QFPFPISAK.H(43)	27%	Putative

Table 3. Cont.

Spot no.	Accession no.	Total protein score	No. of unique/ significant peptides	MS/MS peptide sequence (Indv. Ion score)	Sequence coverage	DomainSweep analysis
				R.NELGAQYNFK.I(44)		IPR001610 PAC motif
				R.VIQAATEILPGK.-(73)		IPR001713 Proteinase inhibitor
				K.LGHFQQYDVR.L(60)		IPR000010 Proteinase inhibitor I25,
				K.DRNELGAQYNFK.I(52)		cystatin
				K.HTGGSDFLI ADPEAQGVADAVR.S(4)		IPR001878 Zinc finger, CCHC-type
D87	GH986563	35	1	K.DNVPLFVGR.V(35)	4%	Putative
						IPR000215 Protease inhibitor I4, serpin
D110	EZ048786	46	1	R.FATPLILTGSK.D(3)	6%	Putative
				R.DVSPHPAACLTHSGR.V(43)		IPR002353 Type II antifreeze protein
						IPR002371 Flagellar hook-associated
						protein
						IPR000204 Orexin receptor
E9	GH986691	257	7	K.YANPQELR.Q(51)	31%	Putative
D2-D5, D8, D18, D10, D13,				K.SINVPQVEK.E(32)		IPR000980 SH2 motif
D14, D15, D19-D23, D27, D28,				K.QYWPYVDEKPR.M(46)		IPR000463 Cytosolic fatty-acid binding
D31, D40, D47, E3, E4, E6, E7,				K.KQYWPYVDEKPR.M(30)		
E8, E10, E11, E12, E14, E15,				R.DEDSFLYETPEA QNPIVQK.K(28)		
E16, E18, E19, E60, E61, E63,				K.RDEDSFLY ETPEAQNPIVQK.K(37)		
E64, F31, F94, F95				K.GLESETEDTAATTILIADMVHY- LK.Y(33)		
F6,	GH986624	35	1	R.ESLDFFR.V(35)	3%	No hits
F48						
F63	GH986878	38	1	K.AEETVPVLLTAEEK.L(38)	7%	Significant
						IPR007327 Tumor protein D52
						Putative
						IPR004077 Interleukin-1 receptor, type II

Generated MS/MS data were searched against the tardigrade clustered database. Spot number, protein annotation, accession number, total protein score, number of matched peptides, peptide sequence and sequence coverage are listed. Identical proteins identified in different spots are listed only once and the spot with the highest protein score (in bold) is ranked at the top. The significant or putative candidates found in Domain Sweep are also listed in the Table.
doi:10.1371/journal.pone.0009502.t003

Three different cathepsin proteins could be identified: cathepsin K (spot A84), cathepsin Z (spot E80) and cathepsin L1 (spot F81). Cathepsin L is a ubiquitous cysteine protease in eukaryotes and has been reported as an essential protein for development in *Xenopus laevis* [32], *Caenorhabditis elegans* [33] and *Artemia franciscana* [34].

Several protein spots are associated with ATP generation and consumption and may have important roles in the early development as described for *Artemia*, because many important metabolic processes require ATP [35,36]. ATP synthase (spot B152) regenerates ATP from ADP and Pi [37]. It consists of two parts: a hydrophobic membrane-bound part (CF0) and a soluble part (CF1) which consists of five different subunits, alpha, beta (spot E89), gamma, delta (spot C139) and epsilon. Arginine kinase (spot B167) is an ATP/guanidine phosphotransferase that provides ATP by catalyzing the conversion of ADP and phosphorylarginine to ATP and arginine [38]. The presence of arginine kinase has been shown in tissues with high energy demand [39].

Interestingly, we could identify the translationally controlled tumor protein (TCTP) (spot F75) on the 2D gel. TCTP is an important component of TOR (target of rapamycin) signalling pathway, which is the major regulator of cell growth in animals and fungi [40].
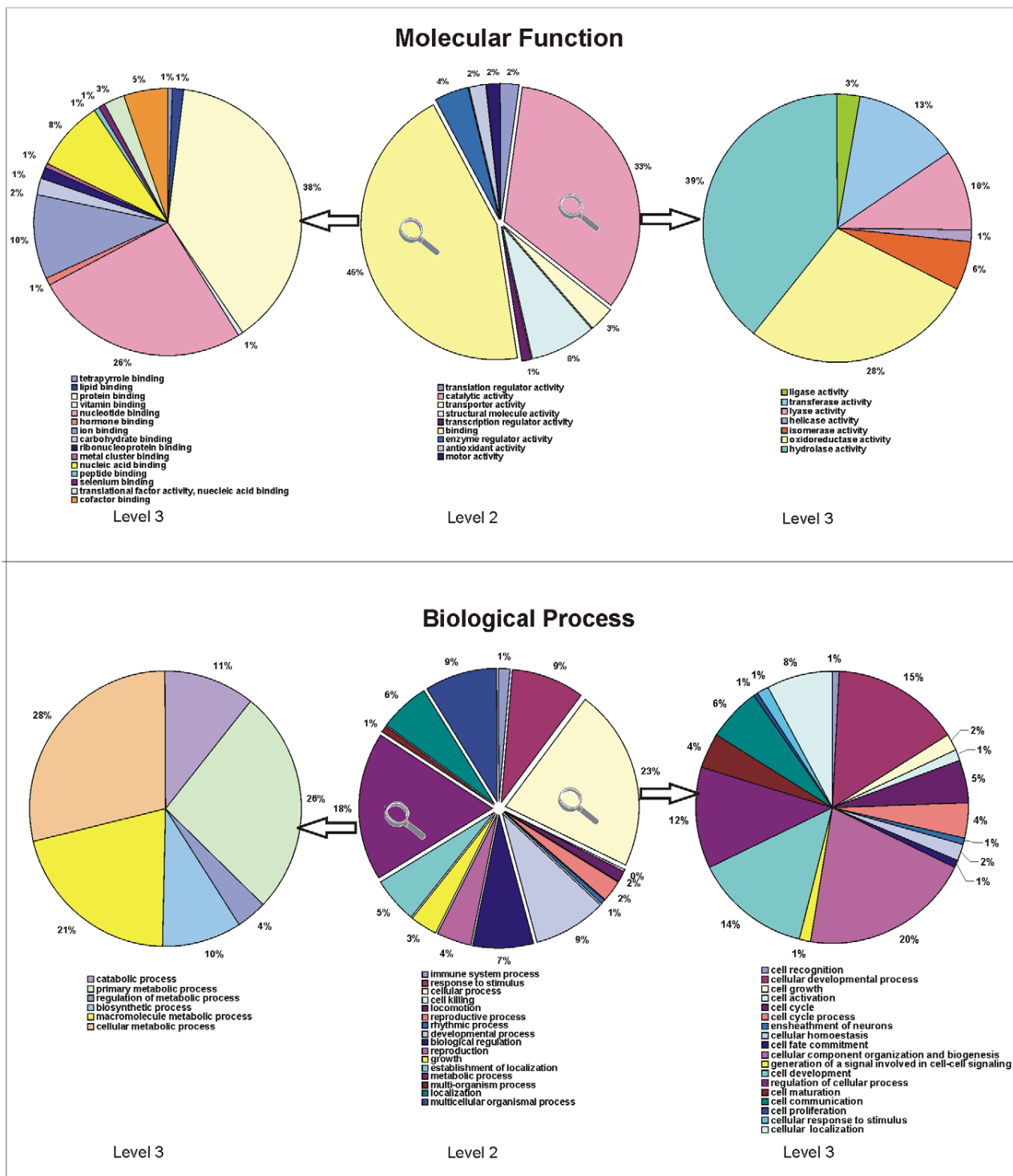
Figure 7. GO analysis of proteins identified in *M. tardigradum*. A total of 271 spots representing 144 unique proteins was analysed with the Blast2GO program. The GO categories "molecular function" and "biological process" are shown as pie charts. A total of 9 different molecular function groups and 16 groups for biological processes are present in our result. The major parts of these categories (level 2) are shown in more detail (level 3) on the left and right side.
doi:10.1371/journal.pone.0009502.g007

Evaluation of Heat Shock Proteins by Western Blot Analysis

To evaluate the highly conserved heat shock proteins 60 and 70, we performed Western blot analyses with antisera directed against these proteins. Hsp70 was found in several spots on the reference 2D proteome map, e.g. in spot B172, C31, C133 and F27. None of these spots fits well to the calculated molecular weight of approx. 70 kDa, most of them were considerably smaller. In contrast, the
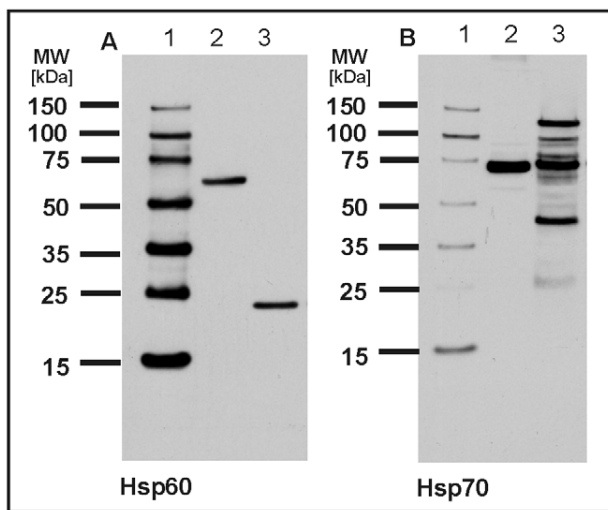
Figure 8. Detection of hsp60 and hsp70 by Western blotting. Total protein extract of *M. tardigradum* in the active state was separated on a one-dimensional polyacrylamide gel. Hsp60 (A) and hsp70 (B) could be immunodetected with high sensitivity. Lane 1A and 1B: DualVue Western blotting marker. Lane 2A and 2B: Total protein extract of HeLa cells. Lane 3A and 3B: Total protein extract of tardigrades. Notably, the protein bands in the HeLa control lysate show molecular weights of 60 and 70 kDa as expected. In contrast the detected protein band for hsp60 in *M. tardigradum* is considerably smaller. For hsp70 multiple bands are observed in *M. tardigradum* at higher as well as at lower molecular weights.
doi:10.1371/journal.pone.0009502.g008

immunoblot shows the strongest band at the expected position which is in agreement with the position of hsp70 in the control lysate of HeLa cells (Figure 8B). However, several additional bands can be observed at higher as well as at lower molecular weights. The lower bands might account for the identified spots on the 2D gel with lower molecular weight. The full-length protein might have escaped the spot picking procedure since only a limited number of detected spots were further processed.

Hsp60 was identified in spot F57 of the 2D map as described above. Since hsp60 was identified by only one peptide hit we confirmed this result by immunostaining using an antibody directed against a peptide in the C-terminal region of the entire protein. Only one band is visible on the Western blot at approx. 24 kDa whereas the protein band in the HeLa control lysate is located at its expected position (Figure 8A). The lower molecular weight is in accordance with the location of hsp60 (spot F57) on the 2D gel. Thus, in *M. tardigradum* hsp60 exists in a significantly shorter form. Whether the observed difference in the molecular weight indicates a different function and role of this protein in *M. tardigradum* needs to be investigated in future experiments. To test whether other tardigrade species show similar results we performed an immuno-blot with protein lysates from 5 other species namely *Paramacrobiotus richtersi*, *Paramacrobiotus* "*richtersi* group" 3, *Macrobiotus tonollii*, *Paramacrobiotus* "*richtersi* group" 2 and *Paramacrobiotus* "*richtersi* group" 1. Total protein lysate from HeLa cells was loaded as control (Figure 9A, lane 1). Actin served as loading control for all lysates (Figure 9B). Interestingly, some species also exhibit truncated forms of hsp60 on the Western blot whereas others show higher forms. The molecular weights are ranging from approx. 75 kDa for *P.* "*richtersi* group" 2 and *P.* "*richtersi* group" 1 lysates (Figure 9A, lane 4 and 6), 35 kDa for *P.* "*richtersi* group" 3 and *P. richtersi* lysates (Figure 9A, lane 5 and 8) down to 24 kDa for *M. tardigradum* and *M. tonollii* (Figure 9, lane 3 and 7).
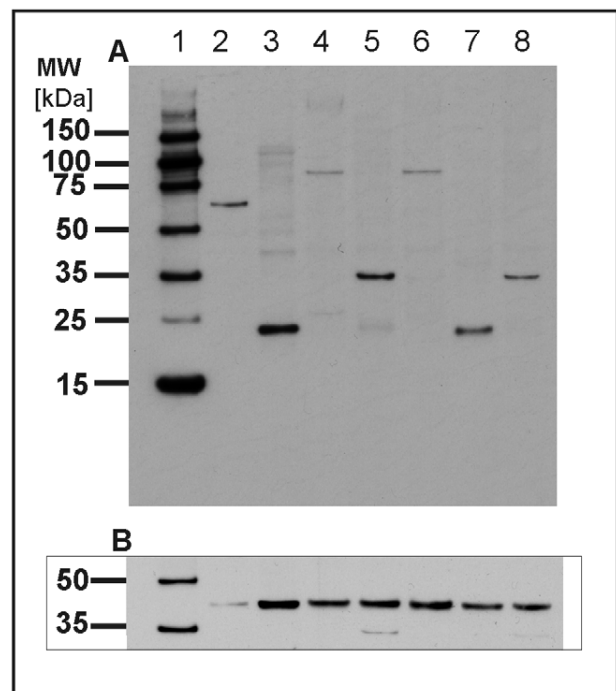
Figure 9. Detection of hsp60 in six different tardigrade species by Western blotting. Total protein extracts of tardigrades in the active state were separated on a one-dimensional polyacrylamide gel. Hsp60 (A) and actin (B) as loading control were immunodetected with high sensitivity. Lane 1: DualVue Western blotting marker. Lane 2: Total protein extract of HeLa cells. Lane 3: Total protein extract of *M. tardigradum*. Lane 4: Total protein extract of *Paramacrobiotus richtersi*. Lane 5: Total protein extract of *Paramacrobiotus* "*richtersi* group" 3. Lane 6: Total protein extract of *Macrobiotus tonollii*. Lane 7: Total protein extract of *Paramacrobiotus* "*richtersi* group" 2. Lane 8: Total protein extract of *Paramacrobiotus* "*richtersi* group" 1. Interestingly, the detected protein bands were ranging from 100 kDa to less than 24 kDa. Only hsp60 in the HeLa control lysate was detected at its expected position at 60 kDa.
doi:10.1371/journal.pone.0009502.g009

Discussion

Establishing a Comprehensive Proteome Map of *Milnesium tardigradum*

The analysis of the proteome *of M. tardigradum* represents to our knowledge the first detailed study of tardigrades on the protein level. Our experimental strategy aimed to identify as many as possible proteins from tardigrades. Thus, we have not employed any subcellular fractionation steps to obtain specific subproteomes. We have tested various protocols for protein extraction from whole tardigrades. We could show that direct homogenisation of tardigrades in lysis buffer without any previous precipitation steps is most efficient and enables the generation of high quality 2D gels. Since nothing was known about the proteolytic activity in *M. tardigradum* special precautions were taken to avoid any protein degradation or proteolysis throughout the whole workup procedure. Integrity of proteins was carefully inspected by Western blot analysis of the two housekeeping proteins actin and tubulin where the sequence homology was assumed to be high enough to detect the proteins with commercially available antibodies. The development of a robust workup protocol laid the basis for the generation of a protein map from whole tardigrades in the active state. 56 unique proteins could be identified by searching high

quality MS/MS spectra against the publicly available NCBInr database. However, for many proteins we could not find any homologues in the NCBInr database and only by using our own newly generated tardigrade protein database it was possible to identify another 73 unique proteins. 15 proteins were present in both databases. In addition 36 unique proteins were found in the clustered tardigrade EST database which could not be annotated by BLAST search. This concerns new specific proteins of *M. tardigradum*.

Performance of Database Searches

When we started our study of the tardigrade proteome very little was known about tardigrades at the genome and gene expression level. To this day, only 12 proteins are recorded in the NCBInr database, which originate from *M. tardigradum*. For all of them only partial sequences ranging from as few as 43 amino acids for beta actin up to 703 amino acids for elongation factor-2 are available. Therefore, in parallel to our proteomic study a *M. tardigradum* EST sequencing project has been initiated. Subsequently, two tardigrade specific databases have been established: a clustered tardigrade EST database and a tardigrade protein database which was extracted from the clustered EST database and thus represents a subdatabase containing all tardigrade-specific proteins with annotated function. However, since cDNA sequencing is still ongoing sequence information remains incomplete. We assume that the tardigrade database currently covers approximately one tenth of the tardigrade specific genes comparing the unique clusters found in tardigrades to all known proteins of *Caenorhabditis elegans* or *Drosophila melanogaster* in Ensembl. This fact is greatly influencing our database searches. For most of the protein spots that were analysed by ESI-MS/MS high quality fragmentation spectra were obtained from MS/MS experiments. However, when we searched these MS/MS data against the tardigrade databases and the publicly available NCBInr database, only about 70% of the spots yielded in protein identification whereas the remaining spots gave no significant protein hit. In addition it was impossible to manually extract peptide sequences that were sufficient in length to perform BLAST searches with satisfactory results.

When we examined the protein hits obtained by the three databases in more detail we found that in the NCBInr database approximately one half of the proteins were identified by only one significant peptide hit (Figure 10). For about 25% of the proteins more than one significant peptide hit was obtained. For the remaining 25% only the protein score which is the sum of two or more individual peptides scores was above the significance threshold while none of the peptide scores alone reached this value. In contrast, proteins found in the tardigrade protein database were predominantly identified by more than one significant peptide hit whereas a smaller number was represented by only one peptide. In no cases a protein was identified by the sum of non-significant peptide matches. For proteins without annotation the number of proteins identified by only one peptide was only slightly higher than the number of proteins identified by two or more peptides.

These results are not surprising. Since the NCBInr database contains very few sequences originating from *M. tardigradum* e.g. elongation factor 1-alpha the identification relies predominantly on high homologies between tardigrade sequences and sequences from other more or less related species of other taxa. The chances for detecting more than one identical peptide is significantly higher when searching MS/MS data against the tardigrade EST and tardigrade protein databases since these databases contain only tardigrade specific sequences.
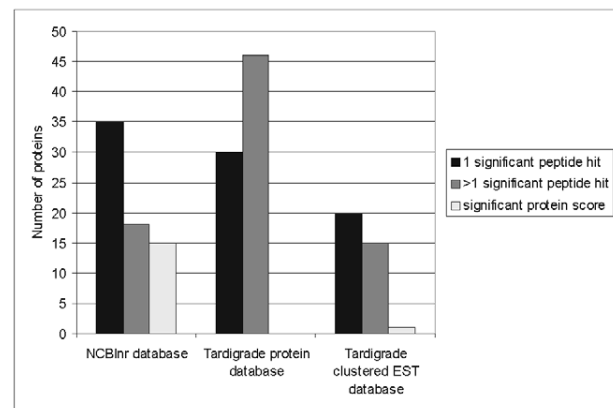
Figure 10. Statistical analysis of significant peptides found in the three different databases which were used to search the MS/MS data. The number of significant peptide hits is compared between the different databases. When searching against the NCBInr database most proteins were identified with only one significant peptide hit. In contrast when using the tardigrade protein database most proteins were represented by two or more significant peptides.
doi:10.1371/journal.pone.0009502.g010

Overall, one might evoke a potentially high false positive rate especially since proteins are included in the reference map which are either identified by only one significant peptide hit or where two or more non-significant peptide scores are summed up to a significant protein score. On the other hand, proteins like LEA and heat shock protein 60 are identified by only one peptide match. Nevertheless they could be confirmed by Western blot analysis to be present in the tardigrade protein extract. Given the incomplete sequence data available to date many proteins might escape confirmation by orthogonal methods e.g. due to the lack of specific antibodies.

Proteins Associated with Anhydrobiosis

Among the numerous proteins which were identified in our proteomic study some proteins have already been reported to play an important role in anhydrobiotic organisms. Most importantly, spot F88 was identified as a protein belonging to the LEA (late embryogenesis abundant) family (group 3). This result was already known from Western blot analyses (Schill et al., 2005, poster presentation, ISEPEP, Denmark). At least six different groups of LEA proteins have been described so far. Group 1, 2 and 3 are the three major groups. Whereas group 1 is only found in plants and group 2 predominantly in plants, group 3 is reported in organisms other than plants. Although the precise role of LEA proteins has not yet been fully elucidated, different research groups have reported on their association with tolerance to water stress by desiccation [41,42]. LEA protein of group 3 could be already identified in nematodes *C. elegans*, *Steinernema feltiae* and *Aphelenchus avenae*, and the prokaryotes *Deinococcus radiodurans*, *Bacillus subtilis* and *Haemophilus influenzae* [43,44,45].

Proteins Exhibiting an Unusual Location on the 2D Map

In general we identified some proteins which show a lower molecular weight than expected. As described above hsp60 is detected as a protein band at 24 kDa by Western blotting and the location of the corresponding spot on the 2D gel shows the same molecular weight. Comparison of different tardigrade species indicates the existence of short as well as long forms of hsp60.

Unique proteins, when analyzed on the 2D gel, often show multiple spots due to posttranslational modifications. Proteins of the vitellogenin family are widely distributed on the 2D gel and show pI as well as molecular weight shifts, which are due to modification through cleavage and to different PTMs like glycosylation and phosphorylation during development of oocytes. Ongoing experiments to detect PTMs using different fluorescence staining methods like ProQ-Emerald for the detection of glycoproteins and ProQ-Diamond for the detection of phospho-proteins indicate that these modifications indeed occur in tardigrades (data not shown).

Prediction of Functional Domains in Proteins with Yet Unknown Functions

36 proteins which could not be identified by BLAST searches were further examined looking for matching functional protein domains with DomainSweep. The function of the following two spots could be revealed with high confidence (Table 3): spot F63 seems to belong to the "tumor protein D52" interpro family (IPR007327). The hD52 gene was originally identified through its elevated expression level in human breast carcinoma, but cloning of D52 homologues from other species has indicated that D52 may play roles in calcium-mediated signal transduction and cell proliferation. Regarding the taxonomic neighbours of the tardigrades, one member in *C. elegans* and 10 members in *Drosophila melanogaster* are reported by Interpro for this family. Spot C95 seems to belong to the family "glucose/ribitol dehydrogenase" (IPR002347). 80 members both in *C. elegans* and in *Drosophila melanogaster* are reported for this family. 28 putative hits were found associated with other spots. These protein hits are putative candidates and therefore less reliable. A comprehensive protein database of *M. tardigradum* as the result of our ongoing cDNA sequencing will help us to evaluate these candidates.

Conclusion

In this study we present for the first time a comprehensive proteome map of *M. tardigradum*. A full description of proteins present in the active state provides a valuable basis for future studies. Most importantly, the protein reference map allows us to undertake quantitative proteomics analysis to detecting proteins with different expression levels in the active versus the anhydrobiotic state. In particular, our workflow is fully compatible with the application of 2D difference gel electrophoresis (2D DIGE), which is one technique allowing sensitive analysis of differences in the protein expression levels. This differential analysis on the protein level will help us to understand survival mechanisms in anhydrobiotic organisms and eventually to develop new methods for preservation of biological materials.

Materials and Methods

Tardigrade Culture and Sampling

Tardigardes of the species *M. tardigradum* Doyère 1840 were maintained in a laboratory culture. The culture was grown on agarose plates (3%) (peqGOLD Universal Agarose, peqLAB, Erlangen Germany) covered with Volvic™ water (Danone Waters, Wiesbaden, Germany) at 20°C. The juveniles were fed on green algae *Chlorogonium elongatum*, the adults with bdelloid rotifers *Philodina citrina*. The specimens for the experiments were all of middle-age, thus effects of age can be excluded. Tardigrades were starved over 3 days and washed several times with Volvic™ water to avoid contamination with food-organisms. Subsequently the animals were transferred to microliter tubes (200 individuals per tube) and surrounding water was reduced to approx. 1–2 µl.

An active state (I) according to Schill et al. [18] was investigated in this work. All samples were shock frozen in liquid nitrogen and stored at −80°C. 200 individuals are defined as one aliquot. Other tardigrade species (*Paramacrobiotus richtersi*, *Paramacrobiotus* "*richtersi* group" 3, *Macrobiotus tonollii*, *Paramacrobiotus* "*richtersi* group" 2 and *Paramacrobiotus* "*richtersi* group" 1) used for immunodetection of hsp60 were prepared in the same way.

Sample Preparation for Gel Electrophoresis

To optimize the sample preparation different precipitation methods have been tested. Chloroform/methanol and TCA/acetone precipitations were performed as described by Wessel, Fluegge [46] and Görg [47], respectively. We used also the commercially available precipitation kit (clean-up kit from GE Healthcare). Comparing the result of different precipitation protocols on a 1D gel we decided to homogenise the tardigrades directly in ice cold lysis buffer and avoid any precipitation steps. The animals (200 individuals) were homogenised directly in 60 µl lysis buffer (containing 8 M urea, 4% CHAPS, 30 mM Tris, pH 8,5) by ultrasonication (SONOPULS, HD3100, Bandelin Electronic) with 45% amplitude intensity and 1–0.5 sec intervals. The lysis buffer contained a Protease Inhibitor Mix (GE Healthcare) to inhibit serine, cysteine and calpain proteases. After homogenisation the samples were stored at −80°C. For gel electrophoresis insoluble particles were removed by centrifugation for 2 min at 14,000 g and the supernatant was quantified using BCA mini-assay.

One Dimensional Gel Electrophoresis and Western Blotting

To compare the efficiency of different sample preparation methods we separated approx. 10 µg total protein extract on a 1D gel. The gel was stained with protein staining solution (PageBlue from Fermentas). For Western blotting a total protein extract of tardigrades (15–20 ug) was separated on a NuPAGE™ 4–12% Bis-Tris mini gel (Invitrogen) using MES running buffer. 200 V were applied until the bromophenol blue front had reached the bottom of the gel (approx. 40 min). Separated proteins were electro transferred onto PVDF membrane for 1.5 h at maximum 50 mA (0.8/cm²) in a semi-dry transfer unit (Hoefer™ TE 77) using following transfer solution: 24 mM Tris, 192 mM glycine and 10% methanol. The PVDF membrane was incubated in a blocking buffer containing 5% non-fat milk, 0.1% Tween20 in PBS. As primary antibodies we used anti actin pan Ab-5 (dianova), anti hsp 60 Ab (D307) (Cell signaling), anti hsp70 Ab (BD Biosciences Pharmingen) and anti α-Tubulin Ab (Sigma).

For molecular weight determination of the target proteins on film we used ECL DualVue marker (GE-Healthcare). Immuno-reaction was detected using the ECL Western Blotting Detection kit from GE Healthcare. Images were acquired using an Image Scanner Model UTA-1100 (Amersham Biosciences).

Two Dimensional Gel Electrophoresis

For 2D gel preparation we added 60 µl 2x sample buffer (7 M urea, 2 M thiourea, 2% CHAPS, 2% DTT, 2% IPG-buffer 3–11 NL) to each aliquot and incubated by shaking for 30 min at 25°C. To avoid streaking on the gels we used 330 µl destreaking buffer (GE Healthcare) instead of rehydration buffer, to which we added 2% IPG-buffer (pI 3–11). Samples were incubated by shaking for 30 min at 25°C. We loaded 100 µg protein on analytical gels and 330 µg on preparative gel.

Strip loading. Loading of proteins was performed during strip rehydration with the recommended volume (450 µl for 24 cm strips) over night.

IEF conditions. First dimension isoelectric focusing (IEF) was performed, using 24 cm long IPG strips with non-linear gradients from pH 3–11 and an Ettan IPGphor instrument and proceeded for 46.4 kVh with the following running protocol: 3 h at 300 V, 6 h at 500 V, 8-h gradient up to 1000 V, 3-h gradient up to 8000 V and 3 h at 8000 V. Strips were either immediately used for the second dimension or stored at $-80°C$.

Second dimension. Strips were equilibrated in 6 M urea, 2% SDS, 30% glycerol, 0.375 M Tris-HCl pH 8.8, 0.002% bromophenol blue and 10 mg/ml DTT for 15 min, followed by a second equilibration step with the same buffer containing 25 mg/ml iodoacetamide instead of DTT, also for 15 min.

Strips were loaded on 12% SDS-gels with an overlay of agarose solution (0,5 mg/100 ml electrophoresis buffer). The second dimension was performed using an Ettan Dalttwelve electrophoresis system (GE Healthcare). Separation was carried out at 1.5 watt/1.5 mm thick gel until the bromophenol blue reached the bottom of the gel (approx. 18 h).

Silver staining of proteins and image analysis. Proteins on analytical gels were visualized by destructive silver staining according to Blum [48]. Additionally, we performed a silver stain compatible with mass spectrometric analysis described by Sinha [49] for preparative gels. Images were acquired using an Image Scanner Model UTA-1100 (Amersham Biosciences).

Protein Identification

In-gel digestion. Protein spots were excised semi-manually with a spot picker (GelPal, Genetix) following non-destructive silver staining and stored at $-80°C$ after removing water. Gel pieces were reduced, alkylated and in-gel digested with trypsin. Briefly, after incubation with 150 μl water at 42°C for 8 min, water was removed (washing step) and gel pieces were shrunk by dehydration with 150 μl 40 mM NH_4HCO_3/ethanol 50:50 (v/v) at 42°C for 5 min in a thermo mixer (600 rpm). The solution was removed and the proteins were reduced with 50 μl 10 mM dithiothreitol in 40 mM NH_4HCO_3 for 1 h at 56°C. The solution was removed and gel pieces were incubated with 150 μl 40 mM NH_4HCO_3 for 5 min at 42°C. After removing the solution gel pieces were alkylated with 100 μl 55 mM iodoacetamide in 40 mM NH_4HCO_3 for 30 min at 25°C in the dark, followed by three alternating washing steps each with 150 μl of 40 mM NH_4HCO_3 and ethanol for 5 min at 37°C. Gel pieces were then dehydrated with 100 μl neat acetonitrile for 1 min at room temperature, dried for 15 min and subsequently rehydrated with porcine trypsin (sequencing grade, Promega, Mannheim, Germany) with the minimal volume sufficient to cover the gel pieces after rehydration (100 ng trypsin in 40 mM NH_4HCO_3). Samples were incubated over night at 37°C.

Extraction. After digestion over night the supernatant was collected in PCR-tubes while gel pieces were subjected to four further extraction steps. Gel pieces were sonicated for 5 min in acetonitrile/0.1% TFA 50:50 (v/v). After centrifugation the supernatant was collected and gel pieces were sonicated for 5 min in acetonitrile. After collecting the supernatant gel pieces were sonicated for 5 min in 0.1% TFA followed by an extraction step again with acetonitrile. The combined solutions were dried in a speed-vac at 37°C for 2 h. Peptides were redissolved in 6 μl 0.1% TFA by sonication for 5 min and applied for ESI-MS/MS analysis.

ESI-MS/MS analysis and database search. NanoLC-ESI-MS/MS was performed on a Qtof Ultima mass spectrometer (Waters) coupled on-line to a nanoLC system (CapLC, Waters). For each measurement 5 μl of the digested sample was injected. Peptides were trapped on a Trapping guard C18- AQ,

10 mm×0.3 mm, particle size 5 μm (Dr. Maisch). The liquid chromatography separation was performed at a flow rate of 200 nl/min on a Reprosil C18-AQ column, 150 mm×75 μm, particle size 3 μm (Dr. Maisch GmbH). The following linear gradient was applied: 5% B for 5 min, from 5 to 15% B in 5 min, from 15 to 40% B in 25 min, from 40 to 60% B in 15 min and finally 60 to 95% B in 5 min. Solvent A contains 94.9% water, 5% acetonitrile, 0.1% formic acid, solvent B contains 95% acetonitrile, 4.9% water and 0.1% μl formic acid. The LC-ESI-MS/MS device was adjusted with a PicoTip Emitter (New Objective, Woburn, MA) fitted on a Z-spray nanoESI interface (Waters). Spectra were collected in the positive ion mode. The capillary voltage was set to 2400 V and the cone voltage was set to 80 V. Data acquisition was controlled by MassLynxTM 4.0 software (Waters). Low-energy collision-induced dissociation (CID) was performed using argon as a collision gas (pressure in the collision cell was set to $5×10^{-5}$ mbar), and the collision energy was in the range of 25–40 eV and optimized for all precursor ions dependent on their charge state and molecular weight. Mass Lynx raw data files were processed with Protein Lynx Global Server 2.2 software (Waters). Deisotoping was performed using the MaxEnt3 algorithm.

The obtained MS/MS spectra were searched against the publicly available NCBInr database using the MASCOT algorithm version 2.0 (Matrix Science, London, UK). The mass tolerance was set to 0.1 Da for fragment ions and 200 ppm for precursor ions. No fragment ions score cutoff was applied. The following search parameters were selected: variable modification due to methionine oxidation, fixed cysteine modification with the carbamidomethyl-side chain, one missed cleavage site in the case of incomplete trypsin hydrolysis. The following settings were applied: minimum protein score >53, minimum number of peptides ≥1. Furthermore, protein hits were taken as identified if a minimum of one peptide had an individual ion score exceeding the MASCOT identity threshold. Under the applied search parameters a sum MASCOT score of >53 refers to a match probability of p<0.05, where p is the probability that the observed match is a random event. Redundancy of proteins that appeared in the database under different names and accession numbers was eliminated. Additionally we searched against the *M. tardigradum* EST and protein database (see below) to identify sequences not present in the NCBInr databases. The following settings were applied: minimum protein score >14 for the EST and >27 for the clustered EST database (p<0.05). Other parameters were as described for the NCBInr searches.

Generation of the Tardigrade EST Database

cDNA libraries from mRNA from tardigrades in different states (active, inactive, transition states) were prepared and sequenced (Mali et al, submitted data). The obtained EST sequences were cleaned from vector sequences using Seqclean against UniVec-database from NCBI (version 12. September 2008, Kitts et al., unpublished). Repeats within the cleaned ESTs were masked using the online service RepeatMasker (version 3.2.6, RM-20080801, Smit et al., unpublished data) followed by a second Seqclean run to eliminate low quality and short sequences. The assembly was performed using cap3 [50] with clipping enabled and resulted in 3318 Unigenes (2500 singlets, 818 contigs). Identification of ribosomal sequences was done using a BlastN-search [51] against the Silva-DB (only eukaryotic sequences, Silva95, [52]) and an E-value cutoff of 1e-3 and resulted in 46 sequences which showed high similarity to ribosomal sequences. Unigenes coding for known proteins were identified using a BlastX search against Uniprot/Swissprot (version 14.1, September 2008), Uniprot/TrEMBL

(version 56.1, September 2008, The UniProt Consortium, 2008) and NRDB (version 12. September 2008,) with an E-value cutoff of 1e-3 and a hmmer-search against PFAM database (release 22, [53]) with an E-value cutoff of 0.1. Translation of Unigen sequences which gave a BlastX or PFAM hit (1539/1889 sequences) into the corresponding frame and a six-frame translation was performed using Virtual Ribosome (version 1.1 Feb-Mar, 2006, [54]). For six frame translation the read through mode of Virtual Ribosome was used. Afterwards stop codons were substituted by an undefined amino acid (X). All new sequences have been deposited in GenBank. The accession numbers are indicated in the Tables 2, 3 and S1 in the column "Tardigrade specific Accession no.".

Classification of Proteins

For functional analysis of identified proteins we used Blast2GO software, which consists of three main steps: blast to find homologous sequences, mapping to collect GO-terms associated to blast hits and annotation to assign functional terms to query sequences from the pool of GO terms collected in the mapping step [55]. Function assignment is based on GO database. Sequence data of identified proteins were uploaded as a multiple FASTA file to the Blast2GO software. We performed the blast step against public database NCBI through blastp. Other parameters were kept at default values: e-value threshold of 1e-3 and a recovery of 20 hits per sequence. Furthermore, minimal alignment length (hsp filter) was set to 33 to avoid hits with matching region smaller than 100 nucleotides. QBlast-NCBI was set as Blast mode. Furthermore, we have chosen an annotation configuration with an e-value-Hit-filter of 1.0E-6, Annotation CutOff of 55 and GO weight of 5. For visualizing the functional information (GO categories: Molecular Function and Biological process) we used the analysis tool of the Blast2GO software.

Protein Domain Analysis of Proteins without Annotation

Six frame translations of the Unigenes were run through the DomainSweep pipeline [56] and the significant and putative hits were collected. For each of the protein/domain databases used, different thresholds and rules were established [56]. Domain hits are listed as 'significant'

i. if two or more hits belong to the same INTERPRO [57] family. The task compares all true positive hits of the different protein family databases grouping together those hits, which are members of the same INTERPRO family/domain.

ii. if the motif shows the same order as described in PRINTS [58] or BLOCKS [59]. Both databases characterize a protein family with a group of highly conserved motifs/segments in a well-defined order. The task compares the order of the identified true positive hits with the order described in the corresponding PRINTS or BLOCKS entry. Only hits in correct order are accepted.

All other hits above the trusted thresholds are listed as 'putative'. By comparing the peptides which were identified by mass spectrometry with the six translations, the correct frame and the associated domain information was listed.

Supporting Information

Table S1 Blast2GO analysis of identified proteins. Spot number, protein annotation, accession number and GO information in all three categories molecular function, biological process and cellular component are listed.

Found at: doi:10.1371/journal.pone.0009502.s001 (0.16 MB XLS)

Acknowledgments

Author Contributions

Conceived and designed the experiments: ES. Performed the experiments: ES. Analyzed the data: ES AHW. Contributed reagents/materials/analysis tools: AHW UW BM MF FF TD SH ROS. Wrote the paper: ES MS. Supervised the project: MS.

References

1. Keilin D (1959) The problem of anabiosis or latent life: history and current concept. Proc R Soc Lond B Biol Sci 150: 149–191.
2. Baumann H (1922) Die Anabiose der Tardigraden. Zoologischen Jahrbücher. pp 501–556.
3. Nelson DR (2002) Current status of the tardigrada: evolution and ecology. Integrative and Comparative Biology 42: 652–659.
4. Schill RO, Fritz GB (2008) Desiccation tolerance in embryonic stages of the tardigrade *Minesium tardigradum*. Journal of Zoology (London) 276: 103–107.
5. Hengherr S, Brümmer F, Schill RO (2008) Anhydrobiosis in tardigrades and its effects on longevity traits. Journal of Zoology (London): in press.
6. Ramløv H, Westh P (2001) Cryptobiosis in the eutardigrade *Adorybiotus (Richtersius) coronifer*: Tolerance to alcohols, temperature and de novo protein synthesis. Zoologischer Anzeiger 240: 517–523.
7. Ramlov H, Westh P (1992) Survival of the cryptobiotic eutardigrade *Adorybiotus coronifer* during cooling to minus 196°C: effect of cooling rate, trehalose level, and short-term acclimation. Cryobiology 29: 125–130.
8. Hengherr S, Worland MR, Reuner A, Brummer F, Schill RO (2009) Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades. J Exp Biol 212: 802–807.
9. Seki K, Toyoshima M (1998) Preserving tardigrades under pressure. Nature 395: 853–854.
10. Horikawa DD, Sakashita T, Katagiri C, Watanabe M, Kikawada T, et al. (2006) Radiation tolerance in the tardigrade *Milnesium tardigradum*. International Journal of Radiation Biology 82: 843–848.
11. Jönsson KI, Schill RO (2007) Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade richtersius coronifer. CBP 146: 456–460.
12. Jönsson KI, Rabbow E, Schill RO, Harms-Ringdahl M, Rettberg P (2008) Tardigrades survive exposure to space in low Earth orbit. Current Biology 18: R729–R731.
13. Clegg JS, Jackson SA (1992) Aerobic heat shock activates trehalose synthesis in embryos of Artemia franciscana. FEBS Lett 303: 45–47.
14. Caprioli M, Krabbe Katholm A, Melone G, Ramlov H, Ricci C, et al. (2004) Trehalose in desiccated rotifers: a comparison between a bdelloid and a monogonont species. Comp Biochem Physiol A Mol Integr Physiol 139: 527–532.
15. Lapinski J, Tunnacliffe A (2003) Anhydrobiosis without trehalose in bdelloid rotifers. FEBS Lett 553: 387–390.
16. Tunnacliffe A, Lapinski J (2003) Resurrecting Van Leeuwenhoek's rotifers: a reappraisal of the role of disaccharides in anhydrobiosis. Philos Trans R Soc Lond B Biol Sci 358: 1755–1771.
17. Suzuki AC (2003) Life history of *Milnesium tardigradum* Doyere (tardigrada) under a rearing environment. Zoological Science (Tokyo) 20: 49–57.
18. Schill RO, Steinbruck GH, Kohler HR (2004) Stress gene (hsp70) sequences and quantitative expression in Milnesium tardigradum (Tardigrada) during active and cryptobiotic stages. J Exp Biol 207: 1607–1613.
19. Hengherr S, Heyer AG, Kohler HR, Schill RO (2008) Trehalose and anhydrobiosis in tardigrades–evidence for divergence in responses to dehydration. FEBS J 275: 281–288.
20. Martin RM, Chilton NB, Lightowlers MW, Gasser RB (1997) Echinococcus granulosus myophilin–relationship with protein homologues containing "calponin-motifs". Int J Parasitol 27: 1561–1567.
21. Kato Y, Tokishita S, Ohta T, Yamagata H (2004) A vitellogenin chain containing a superoxide dismutase-like domain is the major component of yolk proteins in cladoceran crustacean Daphnia magna. Gene 334: 157–165.

22. Mommsen TP, Walsh PJ (1988) Vitellogenesis and oosyte assembly. Fish physiology 11, part A: 347–406.

23. Kim K, Kim IH, Lee KY, Rhee SG, Stadtman ER (1988) The isolation and purification of a specific "protector" protein which inhibits enzyme inactivation by a thiol/Fe(III)/O2 mixed-function oxidation system. J Biol Chem 263: 4704–4711.

24. Leprince O, Hendry G, McKersie B (1993) The mechanisms of desiccation tolerance in developing seeds. Seed Sci Res 3: 231–246.

25. Leprince O, Atherton NM, Deltour R, Hendry G (1994) The Involvement of Respiration in Free Radical Processes during Loss of Desiccation Tolerance in Germinating Zea mays L. (An Electron Paramagnetic Resonance Study). Plant Physiol 104: 1333–1339.

26. Haslekas C, Stacy RA, Nygaard V, Culianez-Macia FA, Aalen RB (1998) The expression of a peroxiredoxin antioxidant gene, AtPer1, in Arabidopsis thaliana is seed-specific and related to dormancy. Plant Mol Biol 36: 833–845.

27. Salinas AE, Wong MG (1999) Glutathione S-transferases–a review. Curr Med Chem 6: 279–309.

28. Tew KD, Ronai Z (1999) GST function in drug and stress response. Drug Resist Updat 2: 143–147.

29. Wilce MC, Parker MW (1994) Structure and function of glutathione S-transferases. Biochim Biophys Acta 1205: 1–18.

30. Kuzmich S, Vanderveer LA, Tew KD (1991) Evidence for a glycoconjugate form of glutathione S-transferase pI. Int J Pept Protein Res 37: 565–571.

31. Michalak M, Groenendyk J, Szabo E, Gold LI, Opas M (2009) Calreticulin, a multi-process calcium-buffering chaperone of the endoplasmic reticulum. Biochem J 417: 651–666.

32. Miyata S, Kubo T (1997) Inhibition of gastrulation in Xenopus embryos by an antibody against a cathepsin L-like protease. Dev Growth Differ 39: 111–115.

33. Britton C, Murray L (2002) A cathepsin L protease essential for Caenorhabditis elegans embryogenesis is functionally conserved in parasitic nematodes. Mol Biochem Parasitol 122: 21–33.

34. Warner AH, Perz MJ, Osahan JK, Zielinski BS (1995) Potential role in development of the major cysteine protease in larvae of the brine shrimp Artemia franciscana. Cell Tissue Res 282: 21–31.

35. Clegg JS (1964) The Control of Emergence and Metabolism by External Osmotic Pressure and the Role of Free Glycerol in Developing Cysts of Artemia Salina. J Exp Biol 41: 879–892.

36. Conte FP, Droukas PC, D ER (1977) Development of sodium regulation and De Novo synthesis of NA + K-activated ATPase in larval brine shrimp, Artemia salina. J Exp Zool 202: 339–362.

37. Boyer PD (1997) The ATP synthase–a splendid molecular machine. Annu Rev Biochem 66: 717–749.

38. Morrison JF (1973) Arginine kinase. In The Enzymes 8: 457–486.

39. Dumas C, Camonis J (1993) Cloning and sequence analysis of the cDNA for arginine kinase of lobster muscle. J Biol Chem 268: 21599–21605.

40. Berkowitz O, Jost R, Pollmann S, Masle J (2008) Characterization of TCTP, the translationally controlled tumor protein, from Arabidopsis thaliana. Plant Cell 20: 3430–3447.

41. Goyal K, Walton LJ, Tunnacliffe A (2005) LEA proteins prevent protein aggregation due to water stress. Biochemical Journal 388: 151–157.

42. Browne JA, Dolan KM, Tyson T, Goyal K, Tunnacliffe A, et al. (2004) Dehydration-specific induction of hydrophilic protein genes in the anhydrobiotic nematode aphelenchus avenae. Eukaryotic Cell 3: 966–975.

43. Dure I L (2001) Occurrence of a repeating 11-mer amino acid sequence motif in diverse organisms. Protein Pept Lett 8: 115–122.

44. Solomon A, Salomon R, Paperna I, Glazer I (2000) Desiccation stress of entomopathogenic nematodes induces the accumulation of a novel heat-stable protein. Parasitology 121 (Pt 4): 409–416.

45. Browne J, Tunnacliffe A, Burnell A (2002) Anhydrobiosis: plant desiccation gene found in a nematode. Nature 416: 38.

46. Wessel D, Flugge UI (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. Anal Biochem 138: 141–143.

47. Gorg A, Postel W, Gunther S (1988) The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 9: 531–546.

48. Blum H, Beier H, Gross HJ (1987) Improved staining of plant proteins, RNA and DNA in polyacrylamide gels. Electrophoresis 8: 93–99.

49. Sinha P, Poland J, Schnolzer M, Rabilloud T (2001) A new silver staining apparatus and procedure for matrix-assisted laser desorption/ionization-time of flight analysis of proteins after two-dimensional electrophoresis. Proteomics 1: 835–840.

50. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868–877.

51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

52. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, et al. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35: 7188–7196.

53. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, et al. (2008) The Pfam protein families database. Nucleic Acids Res 36: D281–288.

54. Wernersson R (2006) Virtual Ribosome–a comprehensive DNA translation tool with support for integration of sequence feature annotation. Nucleic Acids Res 34: W385–388.

55. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

56. del Val C, Ernst P, Falkenhahn M, Fladerer C, Glatting KH, et al. (2007) ProtSweep, 2Dsweep and DomainSweep: protein analysis suite at DKFZ. Nucleic Acids Res 35: W444–450.

57. Mulder NJ, Apweiler R (2008) The InterPro database and tools for protein domain analysis. Curr Protoc Bioinformatics Chapter 2: Unit 2 7.

58. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, et al. (2003) PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31: 400–402.

59. Henikoff S, Henikoff JG, Pietrokovski S (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. Bioinformatics 15: 471–479.

Chapter 7.

Stress response in tardigrades: differential gene expression of molecular chaperones

ORIGINAL PAPER

Stress response in tardigrades: differential gene expression of molecular chaperones

Andy Reuner · Steffen Hengherr · Brahim Mali · Frank Förster · Detlev Arndt ·
Richard Reinhardt · Thomas Dandekar · Marcus Frohme · Franz Brümmer ·
Ralph O. Schill

Abstract Semi-terrestrial tardigrades exhibit a remarkable tolerance to desiccation by entering a state called anhydrobiosis. In this state, they show a strong resistance against several kinds of physical extremes. Because of the probable importance of stress proteins during the phases of dehydration and rehydration, the relative abundance of transcripts coding for two α-crystallin heat-shock proteins (Mt-sHsp17.2 and Mt-sHsp19.5), as well for the heat-shock proteins Mt-sHsp10, Mt-Hsp60, Mt-Hsp70 and Mt-Hsp90, were analysed in active and anhydrobiotic tardigrades of the species *Milnesium tardigradum*. They were also analysed in the transitional stage (I) of dehydration, the transitional stage (II) of rehydration and in heat-shocked specimens. A variable pattern of expression was detected, with most candidates being downregulated. Gene transcripts of one *Mt-hsp70* isoform in the transitional stage I and *Mt-hsp90* in the anhydrobiotic stage were significantly upregulated. A high gene expression (778.6-fold) was found for the small α-crystallin heat-shock protein gene *Mt-sHsp17.2* after heat shock. We discuss the limited role of the stress-gene expression in the transitional stages between the active and anhydrobiotic tardigrades and other mechanisms which allow tardigrades to survive desiccation.

Keywords Alpha-crystallin protein · Anhydrobiosis · Cryptobiosis · Heat-shock protein · Tardigrada · *Milnesium tardigradum*

A. Reuner · S. Hengherr · F. Brümmer · R. O. Schill (✉)
Zoology, Biological Institute, Universität Stuttgart,
Pfaffenwaldring 57,
70569 Stuttgart, Germany
e-mail: ralph.schill@bio.uni-stuttgart.de

B. Mali · M. Frohme
Molecular Biology and Functional Genomics,
University of Applied Sciences Wildau,
Bahnhofstraße 1, Gebäude 15,
15745 Wildau, Germany

F. Förster · T. Dandekar
Department of Bioinformatics, Biocenter, Am Hubland,
University of Würzburg,
97074 Würzburg, Germany

D. Arndt
Developmental Biology Unit,
European Molecular Biology Laboratory,
69012 Heidelberg, Germany

R. Reinhardt
MPI for Molecular Genetics,
Ihnestrasse 63-73,
14195 Berlin-Dahlem, Germany

Introduction

Along with nematodes and rotifers, semi-terrestrial tardigrades exhibit a remarkable tolerance against almost complete desiccation by entering a state known as anhydrobiosis (Keilin 1959) in all developmental stages (Schill and Fritz 2008). To survive drought, which occurs frequently in the habitat of moss-dwelling tardigrades, they enter a so-called tun state (Baumann 1922) and show strong resistance to physical extremes, including high and low temperatures (Ramløv and Westh 1992; Sømme 1996; Ramløv and Westh 2001; Hengherr et al. 2009), high pressure (Seki and Toyoshima 1998), vacuum and ionising irradiation (Horikawa et al. 2006; Jönsson et al. 2008). In the anhydrobiotic state, their metabolism is barely measurable (Pigoń and Węglarska 1955). The longer the animals spend in this state of suspended animation, the longer their lifespan (Hengherr et al. 2008a). The animals resume activity after successful rehydration.

Due to the remarkable ability of tardigrades to survive extreme desiccation, few studies on stress proteins (Schill et al. 2004; Jönsson and Schill 2007) have been carried out to investigate the molecular mechanisms of anhydrobiosis. One isoform of a heat-inducible heat-shock protein (Hsp70) has been described by Schill et al. (2004) as upregulated in the transition phase from the active to the anhydrobiotic state in *Milnesium tardigradum*.

Additionally, in the species *Richtersius coronifer*, a higher protein level of Hsp70 was detected during the transition from the anhydrobiotic to the active state (Jönsson and Schill 2007), whereas a decreased level of Hsp70 was found in anhydrobiotic animals. The investigation of transcripts and encoded stress proteins is based on their well-known function as molecular chaperones (Gething and Sambrook 1992; Georgopoulos and Welch 1993; Jakob et al. 1993). Results derived from several other organisms that tolerate dehydration or suspended animation like nematodes (Chen et al. 2006), crustaceans (Liang et al. 1997b; MacRae 2003), insects (Tammariello et al. 1999; Hayward et al. 2004; Bahrndorff et al. 2008; Lopez-Martinez et al. 2009) and plants (Alamillo et al. 1995; Ingram and Bartels 1996) suggest a versatile role for the stress response in dormant stages.

The present study examines whether an hsp stress response in anhydrobiotic tardigrades operates during dehydration and rehydration. Therefore the expression of several *hsp* transcripts belonging to different Hsp groups was analysed in the eutardigrade *M. tardigradum*. The sequences were taken from our expressed sequence tag (EST) library based on mRNA originating from specimens of *M. tardigradum* from our tardigrade culture. Transcripts of the chaperonin *Mt-shsp10* gene, two α-crystallin small heat-shock protein genes (*Mt-sHsp17.2* and *Mt-sHsp19.5*), one *Mt-hsp60* gene, three *Mt-hsp70* genes, as well as one Mt-*hsp90* gene, were examined to cover a broad range of heat-shock protein genes.

Materials and methods

Tardigrade culture

The study was carried out on the cosmopolitan eutardigrade species *M. tardigradum* Doyère 1849 (Apochela, Milnesidae). Tardigrades were and reared on petri dishes (ø 9.4 cm) filled with a small layer of agarose (3%; peqGOLD Universal Agarose, peqLAB, Erlangen, Germany) and covered with spring water (Volvic™ water, Danone Waters Deutschland, Wiesbaden, Germany) at 21°C and a light/dark cycle of 12 h. Rotifers (*Philodina citrina*) and nematodes (*Panagrellus* sp.) were provided as a food source, and juvenile tardigrades were also fed with the green alga *Chlorogonium elongatum*.

For all experiments, adult animals in good physical condition were taken directly from the culture and starved for 3 days to avoid extraction of additional RNA from incompletely digested food in the intestinal system.

Experimental design

To investigate differences in the expression of stress genes during anhydrobiosis, four different groups of tardigrades were set up. Expression of stress transcripts was analysed during the transition from the active to the anhydrobiotic animals (transition stage I), both during the anhydrobiotic stage and during the transition from the anhydrobiotic to the active state (transition stage II). Active animals were used as a control group. An additional group of animals was analysed, in which the animals were exposed to thermal stress for 1 h at 37°C in a heating block (Thermomixer 5436, Eppendorf, Hamburg, Germany). The transition stages were defined as described earlier by Schill et al. (2004) as transitional stage I, in which animals had started tun formation by contracting their legs, and transitional stage II, in which animals showed distinct movements and had stretched their legs. To achieve desiccation, they were put in open microlitre tubes (Sarstedt, Nümbrecht, Germany) and into small plastic chambers with 85% relative humidity (RH) for 2 days. Subsequently, the tubes were transferred for seven days into chambers with 33% RH to desiccate the animals. The humidity levels described above were achieved by sustaining a constant saturation vapour pressure over a saturated salt solution of KCl and $MgCl_2$, respectively. The boxes had transparent tops to monitor the processes without changing the humidities.

RNA preparation and quantitative real-time PCR

Each experimental group described above consisted of 50 animals, which were subdivided into groups of ten animals. Before RNA isolation, the animals of the "active" and "transition I" group were washed three times in spring water (Volvic water™, Danone). "Anhydrobiotic" and "transition II" animals were washed before desiccation. The tardigrades were homogenised in lysis buffer using a bead mill (FastPrep 24, MP Biomedicals, Heidelberg, Germany). Total RNA was prepared with the RNeasy® Micro Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA was digested during the preparation with the included DNAse I. The RNA was eluted in RNAse-free water, and quantity and quality were checked with a NanoDrop® ND-1000 spectrophotometer (peqLab, Erlangen). Subsequently, cDNA was prepared with the cDNA Synthesis Kit from Bioline (Luckenwalde, Germany).

To measure relative expression of the transcripts of the experimental groups compared to the control group,

quantitative real-time PCR was performed with a MyiQ Single Color Real-Time PCR Detection System (Bio-Rad Laboratories GmbH, München, Germany), and 0.5 μL of the first strand cDNA synthesis reaction mixture was used as template in a total reaction mix of 25 μL (ImmoMix™, Bioline) with 2.0 mM $MgCl_2$. Due to the fact that ribosomal proteins represent adequate housekeeping genes in quantitative real-time PCR (qrtPCR) in general (de Jonge et al. 2007), a partial sequence of the ribosomal protein S13 gene (rps13) was used as reference gene in this study. The efficiency of the PCR reactions was calculated from the slope of the standard curve, which was derived from a dilution series (1:2, 1:20, 1:200 and 1:2,000). Every PCR reaction was performed in triplicate. Threshold cycles (C_t values) were calculated by the MyiQ 2.0 software and analysed with the freely available Relative Expression Software Tool (REST©; Pfaffl et al. 2002), which allows for a determination of significant differences between the expression ratios and the estimation of the standard errors. The $\log_2$ expression ratios of the experimental groups were plotted to compare with the control group; error bars represent the $\log_2$ values of the standard error. Significantly different expression between experimental groups and control groups was accepted for $P<0.05$ (Pair Wise Fixed Reallocation Randomisation Test©, implemented in REST©).

The following programme was routinely used to conduct the qrtPCRs: initial denaturation step (95°C for 10 min) followed by 40 cycles of denaturation (10 s), annealing (20 s) and elongation (20 s). A melt curve analysis was added (95°C to 55°C in steps of 0.5°C every 30 s), and the product size was subsequently examined by gel electrophoresis.

Primers were designed by using the free internet tools "Primer3" (Rozen and Skaletsky 2000) and "NCBI/Primer BLAST" (based on Primer3) with target sequences from M. tardigradum EST libraries. Computational sequence analysis of the deduced EST sequences was performed using the Basic Local Alignment Search Tool (Altschul et al. 1990) at the web pages of the National Center for Biotechnical information (NCBI). Sequences with the highest homology were aligned with ClustalW implemented in the software MEGA4 (Tamura et al. 2007), and ESTs were named after GenBank entries with the highest homologies: Mt-shsp10, Mt-Hsp17.2, Mt-Hsp19.5, Mt-hsp60 and Mt-hsp90. Three different ESTs resulted in significant alignments with proteins of the Hsp70 family and were named Mt-hsp70-1, Mt-hsp70-2 and Mt-hsp70-3. Primers were designed for the coding sequences, without considering 3′ and 5′ untranslated regions.

Results

Analyses of two stress-gene sequences in our EST library resulted in the complete open reading frames for putative proteins with a high homology to small heat-shock/α-crystallin proteins. Alignment of the two sequences (Fig. 1), termed Mt-shsp17.2 and Mt-shsp19.5 after their calculated size of 17.2 and 19.5 kDa, respectively, display considerable differences between the two sequences and similarities to other α-crystallin/sHsp proteins (Fig. 1). Heat-shock treatment resulted in a 778.6-fold higher transcription level of Mt-shsp17.2 in animals compared to the control group (Fig. 2), indicating that it codes for an inducible α-crystallin/sHsp protein. In contrast, Mt-shsp19.5 was not regulated during heat shock. Both α-crystallin/sHsp sequences were not significantly regulated during the process of anhydrobiosis, with one exception; Mt-shsp17.2 was downregulated in the transitional stage II.

In general, most of the genes under investigation were downregulated or regulated at all in the two transitional stages and the anhydrobiotic stage.

Significant down regulation was found for Mt-hsp60 ($P=0.031$) and Mt-hsp70-1 ($P=0.025$) in the transitional stage I, whereas Mt-hsp70-2 ($P=0.019$) was upregulated. In the anhydrobiotic stage, the gene Mt-hsp90 ($P=0.002$) was significantly upregulated, and all other stress genes showed no significant regulation. However, in the transitional stage II, Mt-shsp10 ($P=0.002$), Mt-hsp60 ($P=0.009$), Mt-shsp19.5 ($P<0.001$), Mt-hsp70-1 ($P=0.002$) and Mt-hsp70-2 ($P<0.001$) were significantly downregulated. All genes, except the upregulated Mt-shsp17.2 ($P=0.008$), showed no significant heat-inducible stress response in heat-shocked animals.

The two small heat-shock/α-crystallin protein sequences, Mt-shsp17.2 (with a strong induction of expression under heat shock) and Mt-shsp 19.5 (longer isoform with no induction), were analysed bioinformatically to obtain further insights into their function and difference in induction (Fig. 3). The domain analysis (Fig. 3a, b) shows that both proteins contain an alpha-crystallin domain and have a dimer interface. The sHsps are generally active as large oligomers consisting of multiple subunits and are believed to be ATP-independent chaperones that prevent aggregation and are important in refolding in combination with other Hsps. The potential for multimerization is confirmed for these two sequences by corresponding motifs. However, the longer form leads to a different protein and is no longer in the COG0071/IbpA gene family. Furthermore, the N-terminus (first 60 amino acids) of Mt-shsp 19.5 is tardigrade specific and has no relatives in other organisms. Prosite motifs support the Hsp signature for both proteins and include only often occurring modification motifs. The longer sHsp protein has several potential phosphorylation modification sites predicted in the N-terminus.

To obtain more insight into the differential behaviour of both proteins, potential interaction partners were predicted using the interaction database STRING (von Mering et al.
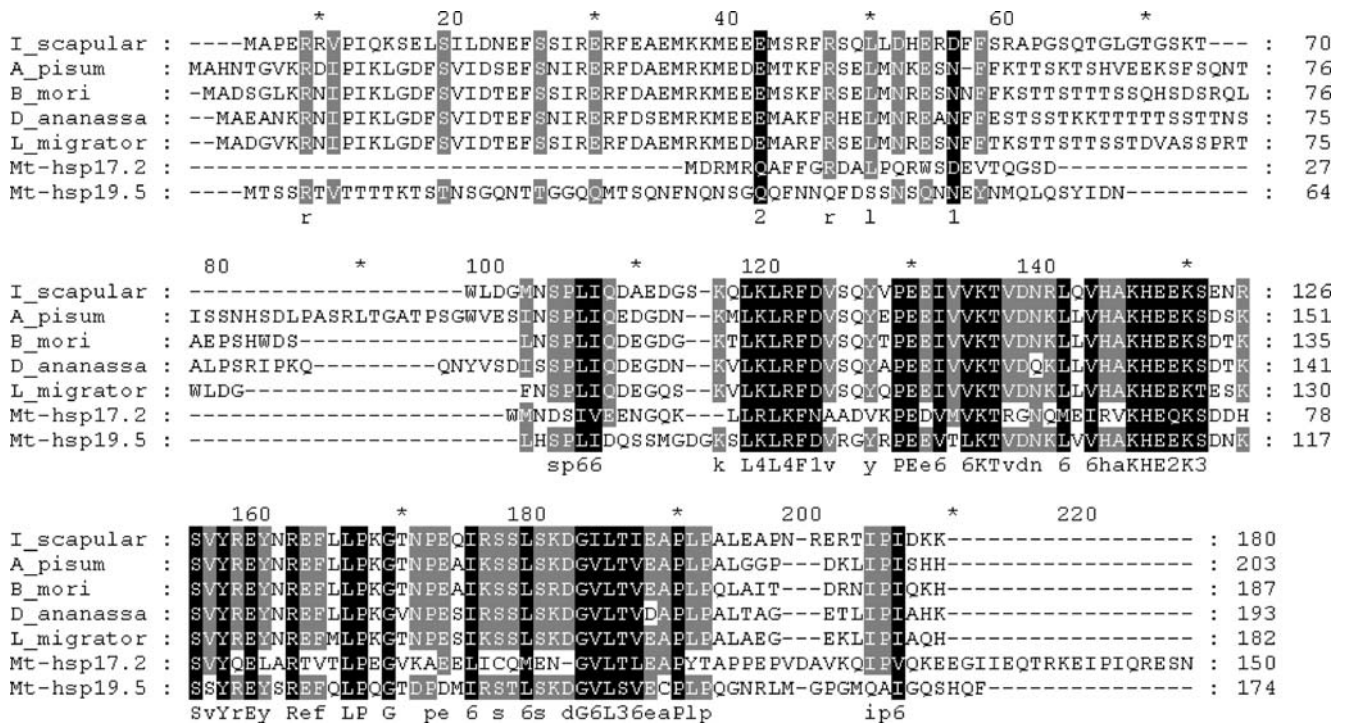
Fig. 1 Alignment of two α-crystallin/small heat-shock proteins from *M. tardigradum* with α-crystallin/small heat-shock proteins from *Ixodes scapularis* (EEC06453), *Acyrthosiphon pisum* (XP_001949446), *Bombyx mori* (NP_001036985), *Drosophila ananassae* (XP_001963454) and *Locusta migratoria* (ABC84493). *Black* high homology, *grey* weak homology, *blank* no homology

2005). Using the hsp homolog for *Mt-shsp 17.2* known from *Drosophila melanogaster* (protein CG14207-PB, isoform B), it appears that there is a tight interaction network in which the *Mt-shsp 17.2* homolog is involved. The protein Mef2 (Myocyte enhancing factor 2) is critical for the regulation of this network and one of the proteins regulated by it is glyceraldehyde 3-phosphate dehydrogenase.

In addition to their developmental function, a number of Mef2 target genes are involved in muscle energy production or storage and were identified in *Drosophila*. As it would be interesting to identify a similar adaptation in tardigrades, we searched by iterative sequence alignment techniques for tardigrade homologues of both proteins. Interestingly, we found the regulatory protein in the

Fig. 2 Relative expression of analysed stress-gene transcripts of *M. tardigradum* in different stages of anhydrobiosis and heat shock. The *asterisks* ($P \leq 0.05$) indicate different expression of transcripts compared with control specimens, which underwent no treatment (displayed by the base line)
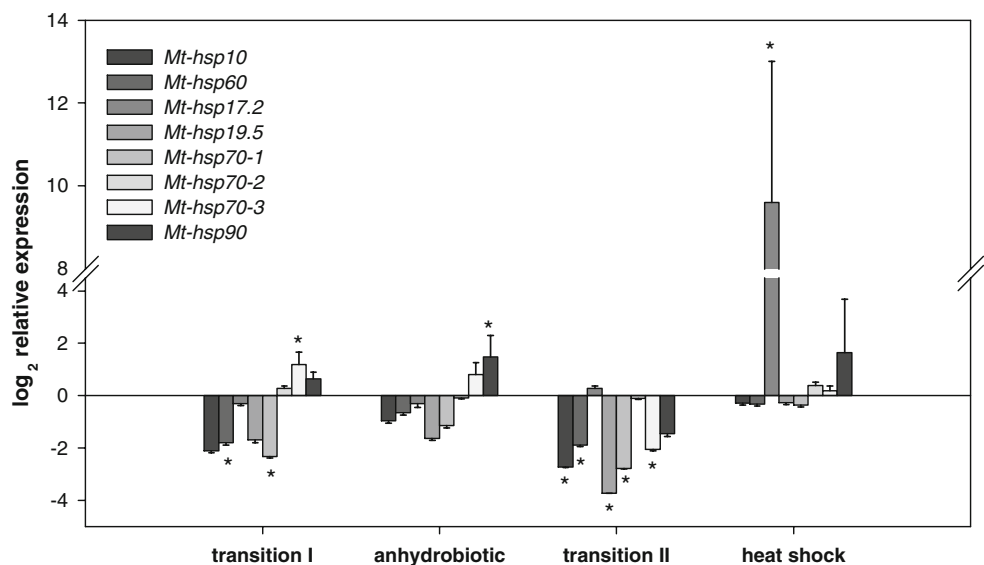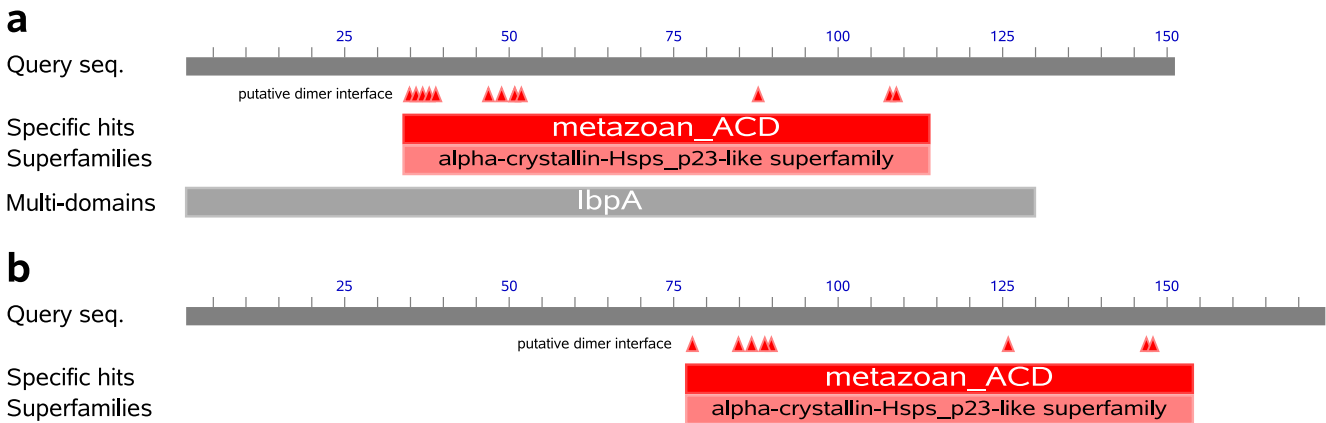
Fig. 3 **a** Domain analysis of *Mt-shsp 17.2* shows that it contains an alpha-crystallin domain (residues 34–113) from the Hsps-p23-like superfamily. There is a putative dimer interface predicted, and residues 1 to 127 belong to COG0071/IbpA, molecular chaperon COG. Compared to other known metazoan proteins, this is a small single domain protein (most others have multidomain context). The closest neighbour by sequence comparison is the heat-shock protein 20.6 (putative) from *I. scapularis* (*e*-value 4e−13) but there are also the well-characterised ones, e.g. from *B. mori* similar over most of the sequence (13–131) with an *e*-value of 2e−12. **b** Domain analysis of *Mt-shsp 19.5* Domain analysis shows that also this protein contains an alpha-crystallin domain (residues 76–154) from the Hsps-p23-like superfamily. At the N-terminal end of the domain, there is again a dimer interface predicted but somewhat weaker. There is highest similarity (1e−33 to heat-shock protein 20.6 isoform 2 from *Nasonia vitripennis*; residues 42–173) but again also to the *B. mori* version (residues 60–156) with 7e−33. Compared to other known metazoan proteins, this is again a small domain protein (most others have multidomain context). However, compared to the shorter version *Mt-shsp 17.2*, we have here a tardigrade-specific N-terminus (first 75 residues) not occurring in other organisms

eutardigrade species *Hypsibius dujardini*. Furthermore, a putative regulatory protein, which could be involved in the network in *M. tardigradum*, has a predicted dual specificity kinase function. Glyceraldehyde 3-phosphate dehydrogenase is found in *M. tardigradum*. In contrast, it turns out that the longer form *Mt-shsp19.5* is not predicted to be involved in this adaptive network. There are no interactions predicted by the STRING database, and furthermore, this is in accordance with our experimental observation that no induction in expression is observed.

Both small heat-shock protein genes were also investigated for regulatory motifs. They contain a number of insignificant motifs in the corresponding untranslated regions. Such patterns with a high probability of occurrence (and which have a high chance of false-positive predictions) include SeCys insertion sequences and GAIT (gamma interferon activated inhibitor of coeruloplasmin mRNA) elements. However, it cannot be ruled out that some type of similar regulation occurs in both of them. Furthermore, the long *shsp* mRNA contains an iron-responsive element structure at position 1188 (see Electronic supplementary material). Here, the chance of occurrence is sufficiently low to suggest functional significance. However, as nothing is known about iron-responsive element-binding proteins in tardigrades and the structure may also be targeted by other proteins, this merely suggests a stability prolonging regulatory element in this region, compatible with the stable, unchanging level of this heat-shock protein mRNA.

Discussion

In this study, the stress response of the eutardigrade *M. tardigradum* was analysed during anhydrobiosis by investigating the expression changes of stress-gene coding sequences for different classes of heat-shock proteins. Sequences were found with significant homologies to several proteins of stress response in EST libraries for *M. tardigradum*. Among them are complete coding sequences for a chaperonin Hsp10 and two α-crystallin/small heat-shock proteins of 17.2 kDa (150 amino acids) and 19.5 kDa (174 amino acids).

Small Hsps prevent protein aggregation and act as molecular chaperones during several kinds of stress (Haslbeck 2002). Studies on sHsp regulation in dormancies of different organisms revealed heterogenous patterns (Bonato et al. 1987; Yocum et al. 1991; Denlinger et al. 1992; Liang et al. 1997a; Yocum et al. 1998; Tammariello et al. 1999; Cherkasova et al. 2000; Goto and Kimura 2004; Rinehart et al. 2007; Gkouvitsas et al. 2008), indicating a diverse array of functions. An essential upregulation of *shsp* has been suggested for cold hardiness in the flesh fly *Sarcophaga crassipalpis* (Rinehart et al. 2007). One of the two *M. tardigradum shsp* sequences, *Mt-shsp17.2*, is strongly inducible by heat-shock treatment, but not regulated during anhydrobiosis. On the contrary, *Mt-shsp19.5* is not inducible by heat and is downregulated in animals in the transition from the anhydrobiotic to the

active state. This leads to the assumptions that both shsps feature different functions. Due to the low expression changes, their role in the anhydrobiosis of tardigrades is questionable, although it is not yet known if there is a sufficient basal level of sHsp proteins in *M. tardigradum*, so that upregulation is not necessary. However, the importance of small heat-shock proteins is clearly demonstrated in *Artemia franciscana*. A massive accumulation of the sHsp p26 occurs in diapausing embryos of this brine shrimp (Liang et al. 1997a; Liang et al. 1997b). The protein p26 is able to move into the nucleus (Clegg et al. 1995) and is thought to protect and/or chaperone, in cooperation with Hsp70, the nuclear matrix proteins (Willsie and Clegg 2002).

This study provides additional data towards the understanding of hsp70 expression during the anhydrobiosis of tardigrades. Schill et al. (2004) described three isoforms of inducible *hsp70* from *M. tardigradum*. The isoform 1 and the isoform 3 did not have a specific function for cryptobiosis. By contrast, transcription of isoform 2 was significantly induced in the transitional stage II between the anhydrobiotic and active stage in *M. tardigradum*. Assuming that a higher mRNA amount may lead to a higher protein content, a functional role of Hsp70 during anhydrobiosis can be suggested, either during anhydrobiosis or as part of a general stress-response mechanism. Since that assumption might not hold, an alternative role might be to prevent protein unfolding and aggregation resulting from the loss of cellular water that takes place during the entry to anhydrobiosis, or in the establishment of a system with refolding capacity to provide functional proteins during and after rehydration.

The lower expression of *Mt-hsp70-1* and *Mt-hsp70-2* during the transition to the active state supports the hypothesis that preceding the actual anhydrobiotic state there is preparation for the time of rehydration. In the eutardigrade species *R. coronifer*, a lower level of Hsp70 protein was found in desiccated animals when compared with active ones (Jönsson and Schill 2007). Assuming that *M. tardigradum* and *R. coronifer* share the same characteristics during desiccation, the upregulated *Mt-hsp70-3* transcript belongs to Hsp70 proteins, which contribute only a small part of the Hsp70 contingent in the cell. Because the antibody used by Jönsson and Schill (2007) was broadly reactive to a wide range of Hsp70 family members, a more prominent Hsp70 isoform might have a higher impact on the overall protein content. However, we note that the low expression of *hsp70* genes and low levels of proteins in tardigrades are similar to data derived from dehydration experiments with yeast containing different amounts of Hsp70 (Guzhova et al. 2008).

Research on Hsp90 revealed many different functions in cells. It acts as a controller of critical hubs in homoeostatic signal transduction, as a regulator of chromatin structure, gene expression, development and morphological evolution and is also involved in the secretory pathway (McClellan et al. 2007; Pearl et al. 2008). Focusing on the role of Hsp90 as a molecular chaperone (Richter and Buchner 2001), the expression changes of a partial putative *Mt-hsp90* sequence were analysed. *Mt-hsp90* was the only sequence investigated in our study that was more abundant in the anhydrobiotic state. However, an increase in its expression was not detected in transitional stage I. In the anhydrobiotic stage, no translation took place, due to the reduced metabolic activity (Pigoń and Węglarska 1955), but a significantly higher level of mRNA was observed, which subsequently decreased after rehydration. If or to what extent the *Mt-hsp90* mRNA was stored for translation into protein during and after rehydration is not known, nor do we know the level required to be effective.

During the whole process of anhydrobiosis, no increased expression was detected for transcripts with high homology to *hsp10* and *hsp60* sequences. Additionally, neither was induced by heat shock at 37°C. Hence, these stress genes, whose proteins are capable of binding and folding non-native proteins (Horwich et al. 2007), which may occur during desiccation, seem to play no relevant role in anhydrobiosis in *M. tardigradum*.

Our investigation of the stress-gene responses in *M. tardigradum* at the transcriptional level clearly shows that most mRNAs are less abundant during anhydrobiosis than in active animals, which may lead to a lower protein level. However, as already mentioned, the levels of stress protein needed for protection or repair in the tardigrade *M. tardigradum* are not known. Focusing on the expression of stress genes, our study suggests a minor role for stabilising and refolding stress proteins, leading to the assumption that denaturation of proteins due to drastic changes during desiccation is not a significant problem for *M. tardigradum*.

The question then arises as to what confers desiccation tolerance on *M. tardigradum* since trehalose (Hengherr et al. 2008b), and stress proteins do not seem to be directly involved. Recent studies showed the existence of other carbohydrates, for example sucrose, sorbitol, inositol and glycerol, have been found in *M. tardigradum* (unpublished results). Those molecules are able to form biological glasses, which may protect cellular structures according to the vitrification hypothesis (Crowe 2002; Crowe et al. 1998). Another important factor might be the presence of late-embryogenesis abundant proteins, which have been detected in *M. tardigradum* (Schill et al. 2004, 2005; McGee et al. 2005; Schokraie et al., submitted) and which are present in many organisms that survive desiccation (e.g. Wise and Tunnacliffe 2004; Goyal et al. 2005; Chakrabortee et al. 2007). A combination of proteins and carbohydrates may also play an important cellular protection role during desiccation in tardigrades.

References

Alamillo J, Almoguera C, Bartels D, Jordano J (1995) Constitutive expression of small heat shock proteins in vegetative tissues of the resurrection plant *Craterostigma plantagineum*. Plant Mol Biol 29:1093–1099

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bahrndorff S, Tunnacliffe A, Wise MJ, McGee B, Holmstrup M, Loeschcke V (2008) Bioinformatics and protein expression analyses implicate LEA proteins in the drought response of Collembola. J Insect Physiol 55:210–217

Baumann H (1922) Die Anabiose der Tardigraden. Zool Jahrb Abt Allg Zool Physiol Tiere 45:501–556

Bonato MCM, Silva AM, Gomes SL, Maia JCC, Juliani MH (1987) Differential expression of heat-shock proteins and spontaneous synthesis of HSP70 during the life cycle of *Blastocladiella emersonii*. Eur J Biochem 163:211–220

Chakrabortee S, Boschetti C, Walton LJ, Sarkar S, Rubinsztein DC, Tunnacliffe A (2007) Hydrophilic protein associated with desiccation tolerance exhibits broad protein stabilization function. PNAS 104:18073–18078

Chen S, Glazer I, Gollop N, Cash P, Argo E, Innes A, Stewart E, Davidson I, Wilson MJ (2006) Proteomic analysis of the entomopathogenic nematode Steinernema feltiae IS-6 IJs under evaporative and osmotic stresses. Mol Biochem Parasitol 145:195–204

Cherkasova V, Ayyadevara S, Egilmez N, Reis RS (2000) Diverse *Caenorhabditis elegans* genes that are upregulated in dauer larvae also show elevated transcript levels in long-lived, aged, or starved adults. J Mol Biol 300:433–448

Clegg JS, Jackson SA, Liang P, MacRae TH (1995) Nuclear-cytoplasmic translocations of protein p26 during aerobic-anoxic transitions in embryos of *artemia franciscana*. Exp Cell Res 219:1–7

Crowe LM (2002) Lessons from nature: the role of sugars in anhydrobiosis. Comp Biochem Physiol A Comp Physiol 131:505–513

Crowe JH, Carpenter JF, Crowe LM (1998) The role of vitrification in anhydrobiosis. Annu Rev Physiol 60:73–103

de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ, ter Elst A (2007) Evidence based selection of housekeeping genes. PLoS ONE 2:e898

Denlinger DL, Lee RE, Yocum GD, Kukal O (1992) Role of chilling in the acquisition of cold tolerance and the capacitation to express stress proteins in diapausing pharate larvae of the gypsy moth, *Lymantria dispar*. Arch Insect Biochem Physiol 21:271–280

Georgopoulos C, Welch WJ (1993) Role of the major heat shock proteins as molecular chaperones. Annu Rev Cell Biol 9:601–634

Gething M-J, Sambrook J (1992) Protein folding in the cell. Nature 355:33–45

Gkouvitsas T, Kontogiannatos D, Kourti A (2008) Differential expression of two small Hsps during diapause in the corn stalk borer *Sesamia nonagrioides* (Lef.). J Insect Physiol 54:1503–1510

Goto SG, Kimura MT (2004) Heat-shock-responsive genes are not involved in the adult diapause of *Drosophila triauraria*. Gene 326:117–122

Goyal K, Walton LJ, Tunnacliffe A (2005) LEA proteins prevent protein aggregation due to water stress. Biochem J 388:151–157

Guzhova I, Krallish I, Khroustalyova G, Margulis B, Rapoport A (2008) Dehydration of yeast: changes in the intracellular content of Hsp70 family proteins. Process Biochem 43:1138–1141

Haslbeck M (2002) sHsps and their role in the chaperone network. Cell Mol Life Sci 59:1649–1657

Hayward SAL, Rinehart JP, Denlinger DL (2004) Desiccation and rehydration elicit distinct heat shock protein transcript responses in flesh fly pupae. J Exp Biol 207:963–971

Hengherr S, Brümmer F, Schill RO (2008a) Anhydrobiosis in tardigrades and its effects on longevity traits. J Zool (Lond) 275:216–220 1–5

Hengherr S, Heyer AG, Köhler H-R, Schill RO (2008b) Trehalose and anhydrobiosis in tardigrades—evidence for divergence in responses to dehydration. FEBS J 275:281–288

Hengherr S, Worland MR, Reuner A, Brümmer F, Schill RO (2009) High temperature tolerance and vitreous states in anhydrobiotic tardigrades. Physiol Biochem Zool 82(6):749–755

Horikawa DD, Sakashita T, Katagiri C, Watanabe M, Kikawada T, Nakahara Y, Hamada N, Wada S, Funayama T, Higashi S, Kobayashi Y, Okuda T, Kuwabara M (2006) Radiation tolerance in the tardigrade *Milnesium tardigradum*. Int J Radiat Biol 82:843–848

Horwich AL, Fenton WA, Chapman E, Farr GW (2007) Two families of chaperonin: physiology and mechanism. Annu Rev Cell Dev Biol 23:115–145

Ingram J, Bartels D (1996) The molecular basis of dehydration tolerance in plants. Annu Rev Plant Physiol Plant Mol Biol 47:377–403

Jakob U, Gaestel M, Engel K, Buchner J (1993) Small heat shock proteins are molecular chaperones. J Biol Chem 268:1517–1520

Jönsson KI, Schill RO (2007) Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade *Richtersius coronifer*. Comp Biochem Physiol B Comp Biochem 146:456–460

Jönsson KI, Rabbow E, Schill RO, Harms-Ringdahl M, Rettberg P (2008) Tardigrades survive exposure to space in low Earth orbit. Curr Biol 18:R729–R731

Keilin D (1959) The Leeuwenhoek lecture. The problem of anabiosis or latent life: history and current concept. Proc R Soc Biol Sci Ser B 150:149–191

Liang P, Amons R, Clegg JS, MacRae TH (1997a) Molecular characterization of a small heat shock/alpha-crystallin protein in encysted artemia embryos. J Biol Chem 272:19051–19058

Liang P, Amons R, Macrae TH, Clegg JS (1997b) Purification, structure and in vitro molecular-chaperone activity of *artemia* P26, a small heat-shock α-crystallin protein. Eur J Biochem 243:225–232

Lopez-Martinez G, Benoit J, Rinehart J, Elnitsky M, Lee R, Denlinger D (2009) Dehydration, rehydration, and overhydration alter patterns of gene expression in the Antarctic midge, *Belgica antarctica*. J Comp Physiol B Biochem Syst Environ Physiol 179 (4):481–491

MacRae TH (2003) Molecular chaperones, stress resistance and development in *Artemia franciscana*. Semin Cell Dev Biol 14:251–258

McClellan AJ, Xia Y, Deutschbauer AM, Davis RW, Gerstein M, Frydman J (2007) Diverse cellular functions of the Hsp90 molecular chaperone uncovered using systems approaches. Cell 131:121–135

McGee B, Schill RO, Tunnacliffe A (2005) Hydrophilic proteins in invertebrate anhydrobiosis. Annual Meeting of the Society for Integrative and Comparative Biology (SICB), San Diego, USA

Pearl LH, Prodromou C, Workman P (2008) The Hsp90 molecular chaperone: an open and shut case for treatment. Biochem J 410:439–453

Pfaffl MW, Horgan GW, Dempfle L (2002) Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res 30:e36

Pigoń A, Węglarska B (1955) Rate of metabolism in tardigrades during active life and anabiosis. Nature 176:121–122

Ramløv H, Westh P (1992) Survival of the cryptobiotic eutardigrade *Adorybiotus coronifer* during cooling to −196°C: effect of cooling rate, trehalose level, and short-term acclimation. Cryobiology 29:125–130

Ramløv H, Westh P (2001) Cryptobiosis in the eutardigrade *Adorybiotus* (*Richtersius*) *coronifer*: tolerance to alcohols, temperature and de novo protein synthesis. Zool Anz 240:517–523

Richter K, Buchner J (2001) Hsp90: chaperoning signal transduction. J Cell Physiol 188:281–290

Rinehart JP, Li AQ, Yocum GD, Robich RM, Hayward SA, Denlinger DL (2007) Up-regulation of heat shock proteins is essential for cold survival during insect diapause. Proc Natl Acad Sci U S A 104:11130–11137

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365–386

Schill RO, Fritz GB (2008) Desiccation tolerance in embryonic stages of the tardigrade *Milnesium tardigradum*. J Zool (Lond) 276:103–107

Schill RO, Steinbrück GHB, Köhler H-R (2004) Stress gene (*hsp70*) sequences and quantitative expression in *Milnesium tardigradum* (Tardigrada) during active and cryptobiotic stages. J Exp Biol 207:1607–1613

Schill RO, McGee B, Tunnacliffe A (2005). Molecular adaptation to extreme dehydration in tardigrades: Hsp70 gene expression, and putative LEA protein induction during cryptobiosis. International Symposium on the Environmental Physiology of Ectotherms and Plants (ISEPEP), Roskilde, Denmark

Seki K, Toyoshima M (1998) Preserving tardigrades under pressure. Nature 395:853–854

Sømme L (1996) Anhydrobiosis and cold tolerance in tardigrades. Eur J Entomol 93:349–357

Tammariello SP, Rinehart JP, Denlinger DL (1999) Desiccation elicits heat shock protein transcription in the flesh fly, *Sarcophaga crassipalpis*, but does not enhance tolerance to high or low temperatures. J Insect Physiol 45:933–938

Tamura K, Dudley J, Nei M, Kumar S (2007) *MEGA4*: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res 33:D433–D437

Willsie JK, Clegg JS (2002) Small heat shock protein p26 associates with nuclear lamins and HSP70 in nuclei and nuclear matrix fractions from stressed cells. J Cell Biochem 84:601–614

Wise MJ, Tunnacliffe A (2004) POPP the question: what do LEA proteins do? Trends Plant Sci 9:13–17

Yocum GD, Joplin KH, Denlinger DL (1991) Expression of heat shock proteins in response to high and low temperature extremes in diapausing pharate larvae of the gypsy moth, *Lymantria dispar*. Arch Insect Biochem Physiol 18:239–249

Yocum GD, Joplin KH, Denlinger DL (1998) Upregulation of a 23 kDa small heat shock protein transcript during pupal diapause in the flesh fly, *Sarcophaga crassipalpis*. Insect Biochem Mol Biol 28:677–682

Chapter 8.

Tardigrade bioinformatics: Molecular adapations, DNA j family and dynamical modelling

Tardigrade bioinformatics: Molecular adapations, DNA j-family and dynamical modelling

F. Förster[1,*], D. Beisser[1,*], M. Frohme[2], R. O. Schill[3] and T. Dandekar[1,$]

[1]Dept of Bioinformatics, Biocenter University of Würzburg, 97074 Würzburg, Germany

[2]University of Applied Sciences, Bahnhofstraße 1, 15745 Wildau, Germany

[3]Dept. of Zoology, Institute for Biology, Universität Stuttgart, 70569 Stuttgart, Germany

[$]Corresponding author, phone: 0931-888-4551. fax -4552;

e-mail: dandekar@biozentrum.uni-wuerzburg.de

[*]Contributed equally

Abstract

Tardigrades are an independent animal phylum with remarkable adaptation against cold, heat, radiation and vacuum, which they outlast in a dormant stage (tun). Recent available information (e.g. from large EST sequencing projects) allows to investigate tardigrade-specific adaptations. We present several examples and techniques of bioinformatical analyses of the available tardigrade data. New integrative methods include ITS2 as a phylogenetic marker and identification of RNA stability motifs. Clustering of protein families reveals tardigrade-specific protein families and insights into mechanisms of oxidative stress tolerance, protein repair, protein turnover, DNA protection and stress pathways. An example we show in detail is the diversity of DNA-j like proteins that enhances the adaptation potential of *Milnesium tardigradum*. A sketch of adaptation dynamics in tardigrades is presented together with techniques for dynamical modelling, however, quantitative modelling requires still far more data and detail.

Introduction

Tardigrades show remarkable stress and desiccation tolerance (Neumann et al., 2009). Direct molecular information on different species and adaptations of this phylum is sparse. However, due to the large amount of large-scacle EST sequence information and other data which recently became available, there is now a potential for bioinformatics to improve insights into adaptations and state switching in tardigrades.

In the following, the general potential for bioinformatics to study molecular tardigrade adapatations is summarized. Large-scale EST data allow us furthermore to examine in detail tardigrade-specific adaptations. This is shown for a large family of DNA-j-like proteins prominent in *M. tardigradum*. Finally, we provide a first sketch on tools and necessary data for a dynamical view on the tardigrade life cycle.

Methods

Sequence sources

We obtained all ESTs of *M. tardigradum* from our ongoing sequencing project of *M. tardigradum.* For *H. dujardini* sequences we obtained all available EST sequences from GenBank. Proteins were predicted using a BLASTX search against UniProtKB/SwissProt-, UniProtKB/TrEMBL- and NR-database. The ORFs for nucleotide sequences which show a significant result (Evalue < 0.001) were extracted. Sequences without result in BLASTX were

searched against the next more extensive database. For sequences without homology the longest ORF was extracted.

Phylogenetic comparison of DnaJ proteins

All EST sequences (above) were searched for the DNA-j family profile hmm from PFAM (DNA-j). The domains were extracted and a multiple alignment was created using clustalw (version 2.0.12). The resulting sequences were bootstrapped and the maximum likelihood consensus tree was calculated using the phylip package (version 3.68).

Dynamic modeling

Kinetic data are sparse. Hence no modeling with differential equations was attempted. EST bank differences were calculated using Perl scripts, R statistics package and Bioconductor suite (Gentleman et al., 2004). Potential metabolite pathways are calculated with YANAsquare (Schwarz et al., 2007), involved protein families for regulatory adaptations used sequence analysis and domain databases as well as the tardigrade wokbench (Förster et al., 2009).

Results and Discussion

Options for bioinformatics to study tardigrade adaptations

Data resources for tardigrade bioinformatics are improving. These include *H. dujardini* focus site www.tardigrades.org (Daub et al., 2003; currently offline), Genbank (7450 nucleotide and 25 protein tardigrade sequence entries on the 30th October 2009 including species-specific, e.g., *H. dujardini* search option on the WEB; Wheeler et al., 2009) and large-scale EST data (Table 1). For detailed analysis including clusters of related proteins we introduced recently the tardigrade workbench http://waterbear.bioapps.biozentrum.uni-wuerzburg.de (Förster et al., 2009).

Phylogenetic markers include ribosomal RNA with its internal transcribed spacer 2 (ITS2; Koetschan et al., 2010). It combines sequence information (species distinction) and for larger evolutionary distances structure information (Schultz et al., 2006; Coleman et al., 2003). Considering compensatory base changes (CBCs) in the ITS2 indicate different species (Müller et al., 2007). We could assign four different species for *Paramacrobiotus* (Schill et al., 2009). *RNA sequence and structure motifs*: RFAM (Gardner et al., 2008) and UTRscan (Pesole and Liuni, 1999) provide surprisingly complete resources to study these. According to

our recent EST data, tardigrades avoid vertebrate motifs for RNA instability such as AU-rich elements (Shaw and Kamen, 1986) and prefer other motifs, e.g. 15-Lox Dice (Table 1).

Protein clusters include oxidative stress protection, dessication tolerance families (Alpert et al., 2006), repair proteins, protein turnover and DNA protection (Table 1) or are tardigrade-specific (TSPs; Förster et al., 2009). Conservation can be analyzed for different enzymes, for conserved orthologous groups or contrasting tardigrades with animals in general (Table 2). Thus the LEA protein family is a key adaptation for instance in *C. elegans* and a number of other animals (Browne, 2002) as well as in tardigrades (according to our recent EST census there are 13 identified LEA proteins in *H. dujardini* but only three LEA proteins in *R. coronifer,* Table 1). Domain analysis may suggest partial functions for TSPs (Table 2, top): In this TSP a chitin-binding domain type 2 is again found in *D. melanogaster* proteins though the sequence homology spans only the chitin-binding domain and no other parts of the protein. Further bioinformatical analysis including iterative sequence comparison but also gene context methods and structure information (Gaudermann et al., 2006) allow to identify pathways hidden in genomes or EST data (Table 1; Fig. 2). Recently also receptors were biochemically found in tardigrades (P2X nucleotide receptors; Bavan et al., 2009).

Analysis of the DNA-j family

Tardigrade protein clusters reveal details of specific molecular adaptations. Different members of the DNA-j family (Yamamoto, 1995) provide a good example. We use our latest transcriptome data on *M. tardigradum* as well as public sequences from other tardigrades. Both in *M. tardigradum* as well as in *H. dujardini* there are a number of such DNA protective proteins apparent from an unrooted phylogenetic tree (Fig. 1). However, we clearly see that protection by DNA-j family member proteins in *M. tardigradum* is higher then in *H. dujardini*. We found 59 proteins in *M. tardigradum* and in *H. dujardini* only eight. In addition they cluster in 20 (Mt) and 10 (Hd) COGs/KOGs, respectively. The latter classifies key protein groups and genes according to clusters of orthologous groups (Tatusov et al., 1997; Tatusov et al., 2003). This classification is also important in phylogenetic studies, it rapidly compares the distribution of observed protein families with other organisms (Table 2). Evidently, *M. tardigradum* has not only DNA-j-like proteins similar to those known from *R. coronifer* and *H. dujardini* but also a multitude of further and different ones. This diversity implies also a strongly enhanced adaptation potential and correlates well with the available data from physiology on *M. tardigradum* (Neumann et al., 2009).

Concerted changes in tardigrade adaptation

At present, there are no accurate time-resolved data on molecular adaptation of tardigrades available. However, the combination of all available data (ESTs, sequences, deduced proteins, pathways, further experimental data on proteins and metabolites) allows a sketch on dynamics of tardigrade adaptation (Fig. 2). Central carbohydrate metabolism can be involved in protection of the tun stage. Thus trehalose protects *Macrobiota* in tun stage. However, overall protection is not as good as in some tardigrades such as *M. tardigradum* with very low trehalose at all stages. Tardigrades survive only weeks in active state but survive in tun stage for up to 100 years by an almost complete reduction of their metabolic activity. Proper protein folding, transport and DNA protection is conducted by the DNA-j protein family (Förster et al., 2009 and this paper) during transition state II. Heat-shock protein protection (e.g. by Hsp 17.2) is involved in transition state I. Specifically, Hsp17.2 shows strong upregulation (778.6 fold) in transition state I and clear down regulation in transitions state II (Reuner et al., 2009). Gene context methods (Jensen et al., 2009) predict it is activated by the key regulator Mef2. Recent EST data show Mef2 is present in several tardigrades (Table 2). In contrast, heat shock protein 19.1 from *M. tardigradum* is predicted not to be involved in this regulatory protein interaction network. Other DNA and protein protective mechanisms (e.g. against oxidative stress) are activated many fold in transition state I and II. Further adaptation strategies include the increase of transporter mRNAs and formation of pores by aquaporins during cryptobiosis (Mali et al., 2009). The water content decreases from 80-90% to 2-3% by a comparatively slow, continuous process (Crowe, 1972). Trehalose levels increase to 0.153-0.472% of the dry weight in Microbiota approaching the tun state (Hengherr et al., 2008), whereas the level of trehalose in *Echiniscus* remains constant and is completely absent in *M. tardigradum.* Furthermore, several peptides, detected with mass spectrometry, suggest the occurrence of antifreeze proteins in *M. tardigradum,* the exact concentration is not known yet, but predicted to be enriched in tun stage. More detailed bioinformatical modelling of pathways requires direct metabolite measurements (Eisenreich et al., 2006) or at least a basic metabolic network topoplogy to fit EST expression information and predict enzyme activities and fluxes (Schwarz et al., 2007). Considering network information and in particular logical interactions such as that one between Mef2 and hsp 17.2 mentioned above, new methods allow to turn network descriptions into Boolean networks and subsequently into a dynamical model. Experimental data fitted well with the modelling in heptocytes (Philippi et al., 2009). In summary, we know already qualitatively a number of concomitant dynamic changes in tardigrade adaptation. Bioinformatics provides methods for the description of these processes

and their dynamics, however, we need still a lot more detailed and accurate dynamical data to fully apply them to tardigrades.

Conclusions

Fascinating views on the extreme adaptations of tardigrades turn to a molecular level such as DNA-j-family proteins and apply bioinformatics for tardigrade species and pathway analysis. Specific RNA motifs or splicing variants (e.g. hsp proteins) become visible. A general impression on state specific changes and tools for their detailed modelling became available, however, more quantitative data are required before detailed modelling of the transition from active tardigrade to tun stage is possible. Currently this is outlined by analysis of EST and first metabolite data.

Acknowledgements

German Summary

Bioinformatik der Tardigraden: Molekulare Anpassungen, DNA-j Protein Familie und dynamische Modellierung

Zusammenfassung

Tardigraden sind ein eigenständiger Tierstamm mit einer außergewöhnlichen Anpassungsfähigkeit gegenüber Kälte, Hitze, Strahlung und Vakuum in der inaktiven Tönnchenform. Neue Daten (z.B. von EST Sequenzierungsprojekten) machen es möglich, diese Tardigraden-spezifische Adaptionsstrategien zu untersuchen. Wir zeigen verschiedene Beispiele und bioinformatische Methoden zur Analyse dieser neuen Daten. Neue integrative Methoden beinhalten ITS2 als einen Marker in phylogenetischen Studien bei Tardigraden und die Identifizierung von RNA-Stabilitätsmotiven. Tardigraden-spezifische Proteinfamilien erlauben neue Erkenntnisse über oxidative Stresstoleranz, Reparaturproteine, Proteinumsatz, DNA-Schutz und Stress-pathways. Ein detailliertes Beispiel analysiert die Diversität der DNA-j ähnlichen Proteine, welche das Adaptionspotenzial von *Milnesium tardigradum* verstärken. Eine Skizze der Anpassungsdynamik in Tardigraden zeigen wir hier zusammen

mit dynamischen Modellierungstechniken, aber eine quantitative Modellierung braucht noch sehr viel mehr Daten und Details.

References

Alpert, P. (2006). Constraints of tolerance: why are desiccation-tolerant organisms so small or rare? J Exp Biol 209:1575-84.

Bavan, S., Straub, V. A., Blaxter, M. L. and Ennion, S. J. (2009). A P2X receptor from the tardigrade species Hypsibius dujardini with fast kinetics and sensitivity to zinc and copper. BMC Evol Biol 9:17.

Browne, J., Tunnacliffe, A., Burnell, A.. (2002). Anhydrobiosis: plant desiccation gene found in a nematode. Nature 416:38.

Coleman, A.W. (2003). ITS2 is a double-edged tool for eukaryote evolutionary comparisons. Trends in Genetics 19:370-5.

Crowe, J. H. (1972). Evaporative water loss by tardigrades under controlled relative ehumidities. Biol Bull 142:407-416.

Daub, J., Thomas, F., Aboobaker, A. and Blaxter, M.L. (2003). A survey of genes expressed in the tardigrade Hypsibius dujardini. See NCBI CK326773 and http://zeldia.cap.ed.ac.uk/tardigrades/tardibase.html.

Eisenreich, W., Slaghuis, J., Laupitz, R., Bussemer, J., Stritzker, J., Schwarz, C., Schwarz, R., Dandekar, T., Goebel, W. and Bacher, A. (2006). 13C isotopologue perturbation studies of Listeria monocytogenes carbon metabolism and its modulation by the virulence regulator PrfA. Proc Natl Acad Sci U S A 103:2040-2045.

Förster, F., Liang, C., Shkumatov, A., Beisser, D., Engelmann, J. C., Schnolzer, M., Frohme, M., Muller, T., Schill, R. O. and Dandekar, T. (2009). Tardigrade workbench: Comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades. BMC Genomics 10, 469.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R. and Bateman, A. (2009). Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136-D140.

Gaudermann, P., Vogl, I., Zientz, E., Silva, F. J., Moya, A., Gross, R. and Dandekar, T. (2006). Analysis of and function predictions for previously conserved hypothetical or putative proteins in Blochmannia floridanus. Bmc Microbiol 6:1.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H. and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80.

Guidetti, R., Schill, R.O., Bertolani, R., Dandekar, T. and Wolf, M. (2009). New molecular data for tardigrade phylogeny, with the erection of Paramacrobiotus gen. Journal of Zoological Systematics and Evolutionary Research, 47(4), 315-321.

Hengherr, S., Heyer, A. G., Kohler, H. R. and Schill, R. O. (2008). Trehalose and anhydrobiosis in tardigrades - evidence for divergence in responses to dehydration. Febs J 275:281-288.

Hengherr, S., Worland, M. R., Reuner, A., Brümmer, F. and Schill, R. O. (2009). High temperature tolerance in anhydrobiotic tardigrades is limited by glass transition. Physiological and Biochemical Zoology,in press.

Horikawa, D. D., Kunieda, T., Abe, W., Watanabe, M., Nakahara, Y., Yukuhiro, F., Sakashita, T., Hamada, N., Wada, S., Funayama, T., Katagiri, C., Kobayashi, Y., Higashi, S. and Okuda, T. (2008). Establishment of a rearing system of the extremotolerant tardigrade Ramazzottius varieornatus: a new model animal for astrobiology. Astrobiology 8:549-556.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. and von Mering, C. (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms.Nucleic Acids Res 37, D412-D416.

Jönsson, K. I. and Schill, R. O. (2007). Induction of Hsp70 by desiccation, ionising radiation and heat-shock in the eutardigrade Richtersius coronifer. Comp Biochem Physiol B Biochem Mol Biol 146:456-460.

Koetschan C., Förster F., Keller A., Schleicher T., Ruderisch B., Schwarz R., Müller T., Wolf M., Schultz J. (2010). The ITS2 Database III - sequences and structures for phylogeny. Nucleic Acids Research, in press.

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

Neumann, S., Reuner, A., Brümmer, F. and Schill, R. O. DNA damage in storage cells of anhydrobiotic tardigrades. (2009). Comp Biochem Physiol A Mol Integr Physiol 153:425-42.

Pesole, G. and Liuni, S. (1999). Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNAs. Trends Genet, 15: 378.

Philippi, N., Walter, D., Schlatter, R., Ferreira, K., Ederer, M., Sawodny, O., Timmer, J., Borner, C. and Dandekar, T. (2009). Modeling system states in liver cells: Survival, apoptosis and their modifications in response to viral infection. BMC Systems Biology 3:97.

Pigon, A. and Weglarska, B. (1955). Rate of metabolism in tardigrades during active life and anabiosis. Nature 176:121-122.

Reuner, A., Hengherr, S., Mali, B., Förster, F., Arndt, D., Reinhardt, R., Dandekar, T., Frohme, M., Brümmer, F. and Schill, R. O. (2009). The role of chaperones in anhydrobiotic tardigrades. Cell Stress and Chaperones, in press.

Schill, R. O., Förster, F., Dandekar, T. and Wolf, M. (2009). Distinguishing species in Paramacrobiotus (Tardigrada) via compensatory base change analysis of internal transcribed spacer 2 secondary structures, with the description of four new species. Organisms Diversity and Evolution, submitted.

Schultz, J., Müller, T., Achtziger, M., Seibel, P.N., Dandekar, T., Wolf, M. (2006). The internal transcribed spacer 2 database--a web server for (not only) low level phylogenetic analyses. Nucl Acids Res 34:W704-707.

Schwarz, R., Liang, C., Kaleta, C., Kuhnel, M., Hoffmann, E., Kuznetsov, S., Hecker, M., Griffiths, G., Schuster, S. and Dandekar, T. (2007). Integrated network reconstruction, visualization and analysis using YANAsquare. BMC Bioinformatics 8: 313.

Shaw, G. and Kamen, R. (1986). A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. Cell 46, 659-667

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D.A. (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997). A genomic perspective on protein families. Science 278:631-637.

Wheeler, D. L., , Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden T. L., Maglott, D. R., Ostell J., Pontius, J. U., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2009). Database resources of the National Center for Biotechnology Information. Nucl. Acids Res 33: D5-D15.

Yamamoto, T., Mori, Y., Ishibashi, T., Uchiyama, Y., Ueda, T., Ando, T., Hashimoto, J., Kimura, S. and Sakaguchi, K. (2005). Interaction between proliferating cell nuclear antigen (PCNA) and a DnaJ induced by DNA damage. J. Plant Res 118:91-97.

Tables and Figures

Table 1: Bioinformatics methods for phylogeny and species adaptations in tardigrades

Method	Example	Reference/Pointer
Phylogenetics		
Phylogenetic trees	ITS2	Koetschan et al., 2010
	CBC	Guidetti et al., 2009 and Reuner et al., 2009
Protein clusters	similar proteins	Tardigrade workbench (Förster et al, 2009),
		www.tardigrades.org (Daub et al., 2003)
	LEA proteins	Browne et al., 2002
		(13 LEA proteins in *H. dujardini*[1], 3 LEA proteins in *R. coronifer*[1])
RNA motifs	3' UTR	e.g. RFAM (Gardner et al., 2008)
		15-Lox dice preferred motif (tardigrade)
		AU-rich sequenes avoided motif (vertebrate)
Adaptations		
Pathway modelling	*M. tardigradum*	Tardigrade workbench (Förster et al, 2009)
	H. dujardini	Genbank
	Dessication tolerance	General overview (Alpert et al., 2006)
	Specific receptor	P2X receptor (Bavan et al. 2009)
Gene expression changes		large scale EST comparisons
		proteins or mRNAs activated only in tun stage, e.g. hsp17.2 or DNAj family (this paper)
Network model	*M. tardigradum*	Reuner et al., 2009 and this paper
DNA protection	*M. tardigradum*	Neumann et al., 2009
	H. dujardini	this paper; Förster et al., 2009
	R. coronifer	Jönnson and Schill, 2007

[1]Data from *M. tardigradum* are from an own ongoing EST sequencing effort and data from *H. dujardini* and *R. coronifer* are from Genbank,

Table 2: Typical examples of conserved proteins in tardigrades[1,2]

	MT[3]	RC[3]	HD[3]	HS[3]	DM[3]	CE[3]
Hypothetical protein containing chitin binding domain	✓	✓	✓	✗	✓	✗
Sucrase-isomaltase, intestinal	✓	✓	✗	✓	✓	✓
GlcNAx-1-P-transferase	✓	✗	✗	✓	✓	✓
Mef2	✓	✗	✓	✓	✓	✓
DNA j-family COG0484	✓	✗	✗	✓	✓	✓
DNA j-family KOG0714	✓	✓	✓	✓	✓	✓

[1]Data from *M. tardigradum* are from an own ongoing EST sequencing effort. [2]Data from *H. dujardini* and *R. coronifer* are from Genbank. [3]Abbreviations: MT, M.tardigradum, RC, *R. coronifer,* HD*, H.dujardini,* HS*, Homo sapiens,* DM*, Drosophila melanogaster,* CE*, C.elegans.* A red cross indicates species were no homolog could be found. A green hook indicates a homolog for the sequence was found within the corresponding dataset (details in Materials and Methods).
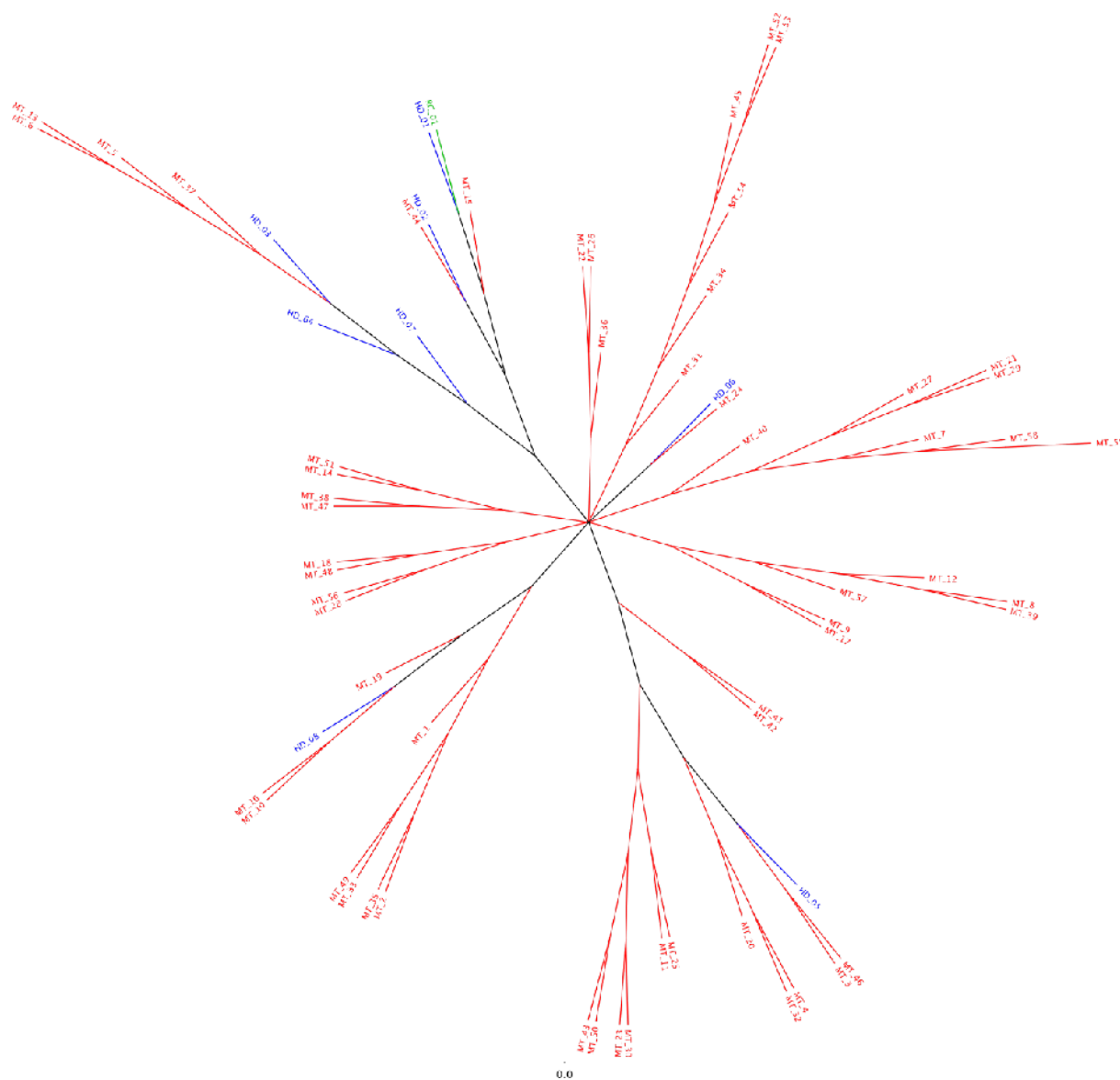
Figure 1: Maximum likelihood tree for the DNA j-family for tardigrades. The maximum likelihood tree is generated by Figtree (version 1.2.3). Besides several **DNA j-family** protein tardigrade sequences from *Hypsibius dujardini* (8 proteins, blue) and *Richertisius coronifer* (1 sequences, green) all Dna j-family members from *Milnesium tardigradum* (58 sequences, red) are shown. EST sequences for *M. tardigradum* were obtained by our ongoing sequencing project. EST sequence data for *H. dujardini* and *R. coronifer* were obtained from Genbank.
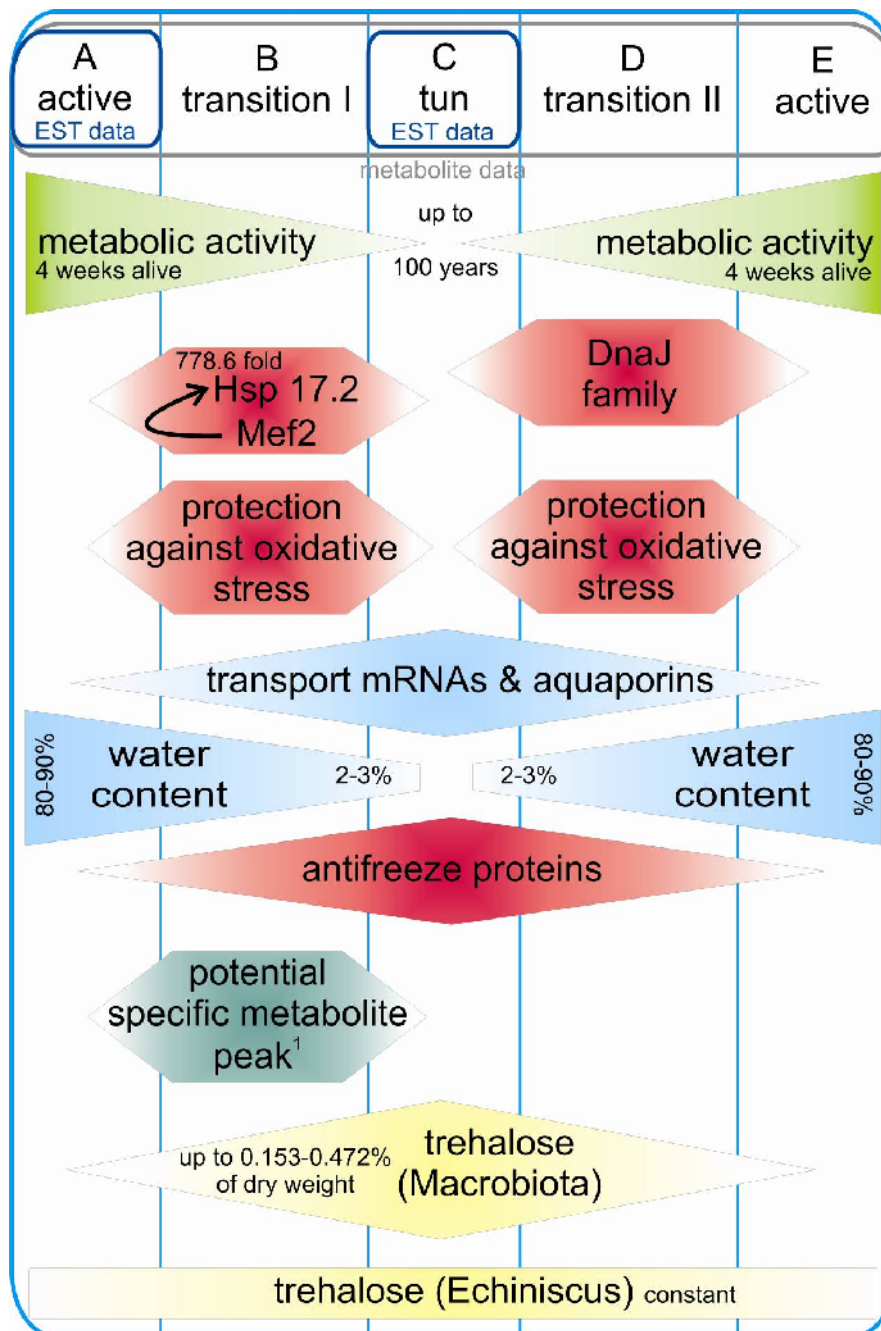
Figure 2: **Concerted changes in tardigrade adaptation**. Shown are the inactivation of central metabolic processes during the transition from an active to the cryptobiotic stage and the reactivation during transition state II (Pigon and Weglarska, 1955). A number of adaptations are involved in tardigrade transition from active stage to tun stage and provide first data for new efforts in dynamical modelling. The polygon shapes indicate the extant of activation either in active stage, tun stage or during transition. Details see results and methods. [1] no unambiguous data yet apart from trehalose, for other metabolites compare with (Horikawa et al., 2008; Neuman et al., 2009)

Chapter 9.

The ITS2 Database III—sequences and structures for phylogeny

The ITS2 Database III—sequences and structures for phylogeny

Christian Koetschan, Frank Förster, Alexander Keller, Tina Schleicher, Benjamin Ruderisch, Roland Schwarz, Tobias Müller*, Matthias Wolf* and Jörg Schultz*

Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland 97074 Wuerzburg, Germany

ABSTRACT

The internal transcribed spacer 2 (ITS2) is a widely used phylogenetic marker. In the past, it has mainly been used for species level classifications. Nowadays, a wider applicability becomes apparent. Here, the conserved structure of the RNA molecule plays a vital role. We have developed the ITS2 Database (http://its2.bioapps .biozentrum.uni-wuerzburg.de) which holds information about sequence, structure and taxonomic classification of all ITS2 in GenBank. In the new version, we use Hidden Markov models (HMMs) for the identification and delineation of the ITS2 resulting in a major redesign of the annotation pipeline. This allowed the identification of more than 160 000 correct full length and more than 50 000 partial structures. In the web interface, these can now be searched with a modified BLAST considering both sequence and structure, enabling rapid taxon sampling. Novel sequences can be annotated using the HMM based approach and modelled according to multiple template structures. Sequences can be searched for known and newly identified motifs. Together, the database and the web server build an exhaustive resource for ITS2 based phylogenetic analyses.

INTRODUCTION

The internal transcribed spacer 2 (ITS2) of the nuclear rDNA cistron is a widely used phylogenetic marker. In its early years it was specifically used for low-level phylogenetic analyses, i.e. of species within the same genus. At that time, only nucleotide information of the fast evolving sequence was used. With analyses of the two-dimensional structure of the molecule it became evident that the structure is highly conserved throughout the eukaryotes (1–3). The combination of a fast evolving sequence with a slow evolving structure within one molecule suggested its capability for higher level classifications (4). In the last years, the ITS2 has been revealed to be more than just an excellent phylogenetic marker. Its applications include usage as a marker for species identification in environmental samples (phylochips) (5,6), as a target molecule for barcoding (7,8) and for distinguishing species (9). In many of these cases, the structure plays a fundamental role.

Even though sequence databases typically include a large quantity of ITS2 sequences, no coherent information source existed so far including both sequence and structure information, with ITS2 specific annotations. As a consequence of this lack, every scientist had to predict the structure of each molecule in his/her dataset more or less manually. Even worse, in the majority of phylogenetic procedures as e.g. alignment or tree calculation the structure could not be used at all as the corresponding software was not capable of integrating the structure information. In order to tackle these problems and to be better able to exploit the power of this intriguing molecule, we have developed the ITS2 Database. Its goal is to provide a valid structure for every ITS2 sequence within GenBank and thereby to become an exhaustive data source for sequence/structure based phylogenetic analyses, as well as offering tools capable of exploiting the information surplus obtained by these secondary structures. In this article, we describe additions to the ITS2 Database in terms of (i) new developments in automated structure

prediction, (ii) new features for the access to the data via the Web interface and (iii) new tools for the analysis of ITS2 sequences.

DATA GENERATION

In the previous version of the database, we used a BLAST (10) based approach for the detection of ITS2 in un-annotated GenBank (11) sequences. We were able to predict the structure of more than 35 000 ITS2 where the start- and end-positions were either lacking or misidentified. As BLAST per se is a local alignment tool (hence the name) and the sequence length is very variable throughout the eukaryotes, heuristics had to be implemented to identify the start and end points of the ITS2. To improve this approach, we have recently developed a Hidden Markov Model (HMM) based method for the correct delineation of the ITS2 (12). Start and end position are inferred from the surrounding 5.8S and 28S regions, that are highly conserved. This method initiated a complete re-design of data generation for the ITS2 Database (Figure 1). In the initial step, we searched through the complete nucleotide database (nt) of GenBank for potential ITS2 sequences using hmmsearch (13). Simultaneously, all annotated ITS2 were extracted from GenBank. In cases where both methods were informative about the position of the ITS2, the HMM based information superseded that from GenBank. This led to 196 697 sequences with positional information of the ITS2 (Database accessed at the 22 June 2009). In the second step, all retained sequences were folded using UNAfold (14). Typical ITS2 features were shown by 63 645 structures, namely the conserved core of four helices with the third as the longest. This was a substantial increase compared to the previous approach where only

GenBank annotations were taken into account. This indicated the necessity of a correct delineation for the folding step. In the next step, these structures served as templates in the homology modelling process. In contrast to the previous approach, we iterated the homology modelling process until no further new correct structures were identified. This resulted in an additional 99 010 predicted full-length structures, further underlining the presence of a conserved structural core of the ITS2 throughout all eukaryotes. Remaining sequences which could either not be homology modelled or where start and end position could not be predicted run through a final step resulting in partial structures. A BLAST search against all identified sequence structure pairs was performed. All significant hits (E-value $< 10^{-10}$) were extended in both directions by five bases. Finally, we applied a less strict homology modelling which required at least two concatenated helices with a transfer larger than 75% each. This resulted in more than 50 000 partial structures. Using the modified pipeline, which would run in a single core 1221 days, we now provide structural information for over 210 000 ITS2, doubling the number of the previous version. As a detailed taxonomic breakdown (Table 1) the best coverage is found in fungi and plants with 80 and 93%, respectively. Only for $\sim$25% of the metazoan ITS2 sequences, a structure could be predicted. This could indicate a deviation from the 'common core'. It could also be caused by problems of UNAfold to identify the correct fold, leading to a paucity of templates for homology modelling. Additionally, the ITS2 Database now contains a record for each GenBank entry which was identified either via textual annotation or our HMM based annotation tool, rendering it as an exhaustive resource for ITS2 sequences and structures.
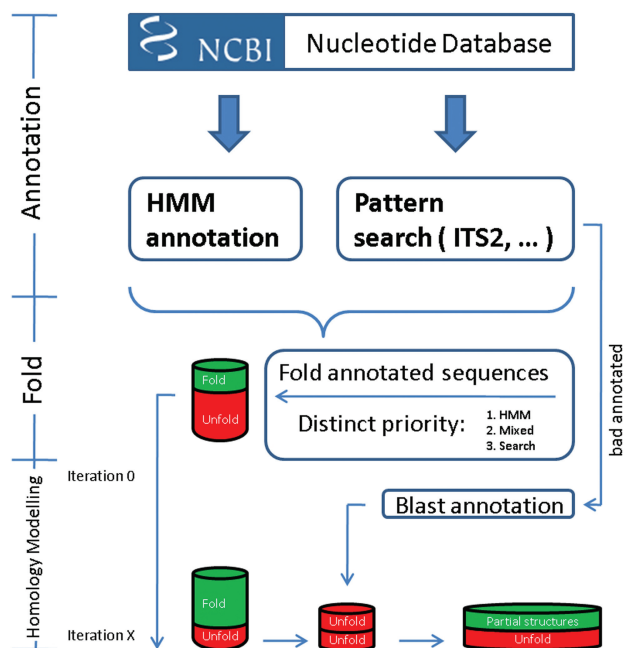
WEB INTERFACE

Search tab

In addition to a search for sequences and structures with GenBank identifiers or species information, we now also provide a BLAST based search. However, standard BLAST procedures are frequently not able to identify distantly related ITS2 sequences because of their high sequence divergence. To overcome this hindrance, we have implemented a sequence and structure based BLAST search that includes information about the highly conserved structure for the homology search. The sequence-structure BLAST uses an ITS2 specific 12×12 scoring matrix representing each nucleotide/structure combination as tuple. This matrix is also used in 4SALE (15) and, as corresponding rate matrix, in ProfDistS (16) for automatic sequence-structure alignment and phylogenetic reconstruction, respectively. Thus, species sampling that starts with any sequence of interest and covers broad taxonomic ranges has become as simple as a BLAST search.

Annotate tab

The web interface does not only present access to the information stored in the database. Further, it provides

Figure 1. Flow chart of the new pipeline for the ITS2 annotation.

Table 1. Taxonomic breakdown of predicted ITS2 structures

	Structure		Partials		All	
	Count	Percentage	Count	Percentage	Count	Percentage
Alveolata	1750	34.67	947	18.76	5048	53.43
Amoebozoa	19	13.01	9	6.16	146	19.18
Apusozoa	0	0.00	0	0.00	35	0.00
Choanoflagellida	0	0.00	0	0.00	1	0.00
Cryptophyta	25	38.46	17	26.15	65	64.62
Environmental samples	26	28.26	7	7.61	92	35.87
Euglenozoa	3	0.62	191	39.71	481	40.33
Fornicata	0	0.00	0	0.00	3	0.00
Fungi	79 251	59.14	28 124	20.99	134 005	80.13
Fungi/Metazoa incertae sedis	2	2.86	0	0.00	70	2.86
Haptophyceae	6	19.35	3	9.68	31	29.03
Heterolobosea	1	0.59	1	0.59	170	1.18
Metazoa	4754	20.14	1357	5.75	23 603	25.89
Nucleariidae	0	0.00	0	0.00	2	0.00
Parabasalidea	1	0.51	0	0.00	197	0.51
Rhizaria	12	2.66	2	0.44	451	3.10
Rhodophyta	27	3.52	28	3.65	768	7.16
Stramenopiles	4441	52.01	2537	29.71	8539	81.72
Viridiplantae	72 322	72.95	20 488	20.67	99 141	93.61
Sum	162 640	59.61	53 711	19.69	272 848	79.29

tools for researchers to process newly determined sequences and to integrate them with already published ones. As shown in the data generation pipeline, correct delineation of the ITS2 sequence can be crucial for structure prediction. We therefore have implemented a web-based interface for the HMM based annotation. It integrates five taxon-specific HMMs for searches and several individually selectable parameters, as e.g. cut-off *E*-value or size limitation. As a result, delimited ITS2 sequences are shown as well as the predicted hybrid of 5.8S and 28S rRNA as a confirmation of the HMM annotation's accuracy (12).

Model tab

After annotation of newly retained ITS2 sequences and selection of a taxon sampling from the ITS2 Database, secondary structures may be determined by two means: First, prediction may be accomplished by homology modelling with the complete set of sequences and structures of the database serving as templates (Predict tab). A second approach is to identify the best template structure within the taxon sampling and use it for homology modelling of the remainders (Model tab). To date, one had to manually run through all possible templates and select the one which resulted in the highest helix transfer percentages. To avoid this tedious and somewhat arbitrary procedure, we now provide the possibility to use multiple sequence-structure pairs to model multiple target sequences. The database will calculate all against all structures and select the template which resulted in the homology prediction with highest percentages of helix transfers for all target sequences.

Similarly, suboptimal structures of a sequence as e.g. retained from minimum free energy folding software, may be given as template input for a set of sequences. As a result, the database will model the structure for all requested sequences with the best fitting suboptimal secondary structure. This is needed, as sometimes the energetically best structure is not the biologically correct one. As the complete homology modelling approach is independent of the ITS2, it may be used to predict the secondary structure of any RNA given a homologous molecule with a known structure.

Motif tab

In addition to the overall structure, conserved motifs like an UGGU sequence preceding the apex of the third helix and a pyrimidine–pyrimidine mismatch in the second helix have been described for the ITS2 (2). In the aforementioned study, identification of these motifs was based on a small dataset and performed mainly by manual inspection. With the availability of the large set of ITS2 sequences in our database, we searched in an automatic way (17) for highly conserved motifs in the ITS2. From our pool of homology models, we randomly extracted a set of unique species. Analysing separately fungal and plant alignments, known and novel motifs were identified. Although the UGGU motif 5′ side to the apex of helix III differs in its composition for fungi, it is located in a corresponding position. For both kingdoms, the U–U mismatch is surrounded by two motifs: one to the left of helix II and one to the right between helix II and III with additional AAA (Figure 2). Having transformed these sequence motifs into HMMs, we now provide identification of these motifs in sequences of interest (Motif tab).

The ITS2 of *Dahlia brevis* as an example

As an example to illustrate the information that can be extracted from the database and the Web interface we analysed the ITS2 of *D. brevis* (18). Looking up the entry for the GenBank identifier 31281745 in the ITS2
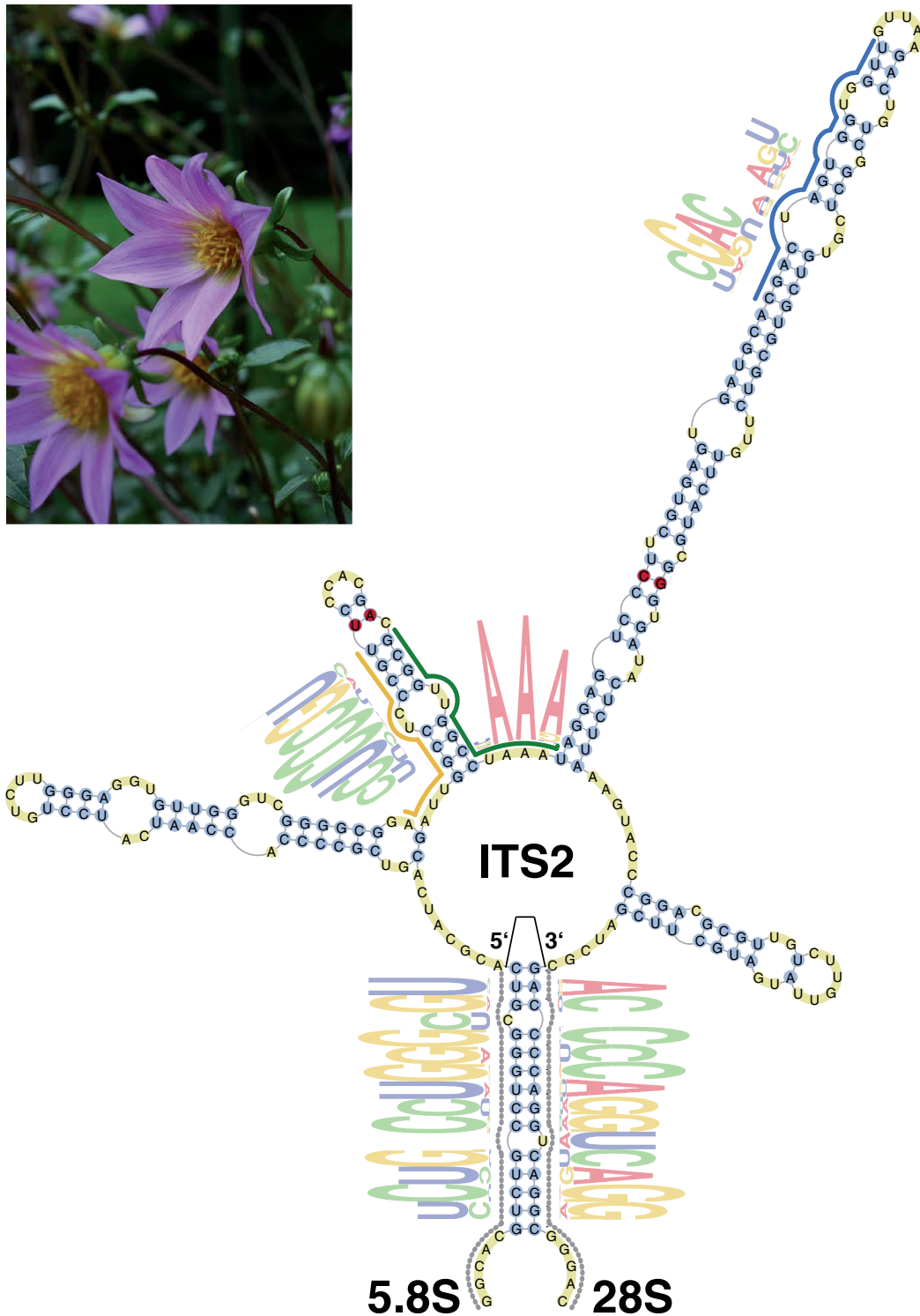
Figure 2. General ITS2 topology and visualization of plant HMM motifs for the secondary structure of *D. brevis* (gi: 31281745). Annotation from HMMs of 5.8S and 28S are displayed as dotted lines tracing the outline of their position, whereas the ITS2 motif HMMs are represented by coloured lines. In parts of these motifs, nucleotide frequencies are presented (21,22). Nucleotides are coloured yellow in unpaired regions, whereas paired nucleotides are blue. CBCs between secondary structures of *D. brevis* and *D. scapigeroides* (gi: 31281755) are shown in red.

Database revealed a stereotypical ITS2 structure (Figure 2). It adopts the common four helix structure with the third as the longest. Additionally, all sequence motifs characteristic for plants are present. In a comparison with another species, here *D. scapigeroides* (gi: 31281755), two Compensatory Base Changes (CBCs) could readily be identified. Indeed, two sequences belong with a probability of 93% to two different species, if at least one CBC is present (9). It should be mentioned, that the CBC criterion works only in one direction. The presence of more than one CBCs indicates with high probability two different species, if there is no CBC, there still could be two species. As *D. brevis* follows all the stereotypes of an ITS2 as the best scoring sequence resulting from all motif searches, it was selected as the 'May 2009' ITS2 in the newly added rubric 'ITS2 of the Month'.

CONCLUSIONS

With the new pipeline for structure prediction, the ITS2 Database now provides information about the structure of more than 210 000 ITS2 molecules, nearly 80% of all ITS2 sequences in GenBank, covering all major taxonomic units. Having the structure available is only the first step for a successful phylogenetic analysis. It would be a pity to use the structure only for the manual refinement of an alignment and neglect it in all other steps. We thus have developed additional stand-alone programs for the entire procedure, which includes automatic alignment calculation [4SALE (15)] as well as tree reconstruction [ProfDistS (16)] considering both, sequences AND secondary structures (these programs have to be downloaded separately). Together, they are seamlessly integrated into a pipeline from sequence through structure and finally to the phylogenetic tree (19). Finally, species boundaries in the dataset can be estimated using the CBCanalyzer [(20), meanwhile also implemented in 4SALE].

The application of secondary structures for the reconstruction of phylogenies improves not only the stability of resulting trees, but more importantly increases the accuracy of phylogenetic estimations (muanuscript under preparation). Thus, it would be desirable to include structural information not only for the ITS2, but also for other frequently used phylogenetic RNA markers.

Conflict of interest statement. None declared.

REFERENCES

1. Coleman,A.W. (2007) Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. *Nucleic Acids Res.*, **35**, 3322–3329.
2. Mai,J.C. and Coleman,A.W. (1997) The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *J. Mol. Evol.*, **44**, 258–271.
3. Schultz,J., Maisel,S., Gerlach,D., Muller,T. and Wolf,M. (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA*, **11**, 361–364.
4. Coleman,A.W. (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet*, **19**, 370–375.
5. Engelmann,J., Rahmann,S., Wolf,M., Schultz,J., Fritzilas,E., Kneitz,S., Dandekar,T. and Muller,T. (2008) Modeling cross-hybridization on phylogenetic rDNA microarrays increases the detection power of closely related species. *Mol. Ecol. Res.*, **9**, 83–93.
6. Landis,F.C. and Gargas,A. (2007) Using ITS2 secondary structure to create species-specific oligonucleotide probes for fungi. *Mycologia*, **99**, 681–692.
7. Moniz,M.B. and Kaczmarska,I. (2009) Barcoding of diatoms: nuclear encoded ITS revisited. *Protist*, Epub ahead of print.
8. Ben-David,T., Melamed,S., Gerson,U. and Morin,S. (2007) ITS2 sequences as barcodes for identifying and analyzing spider mites (Acari: Tetranychidae). *Exp. Appl. Acarol.*, **41**, 169–181.
9. Muller,T., Philippi,N., Dandekar,T., Schultz,J. and Wolf,M. (2007) Distinguishing species. *RNA*, **13**, 1469–1472.
10. Altschul,S.F., Madden,T.L., Schaffer,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**, 3389–3402.
11. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
12. Keller,A., Schleicher,T., Schultz,J., Muller,T., Dandekar,T. and Wolf,M. (2009) 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene*, **430**, 50–57.
13. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
14. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, **453**, 3–31.
15. Seibel,P.N., Muller,T., Dandekar,T. and Wolf,M. (2008) Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Res. Notes*, **1**, 91.
16. Wolf,M., Ruderisch,B., Dandekar,T., Schultz,J. and Muller,T. (2008) ProfDistS: (profile-) distance based phylogeny on sequence—structure alignments. *Bioinformatics*, **24**, 2401–2402.
17. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
18. Saar,D.E., Polans,N.O. and Sorensen,P.D. (2003) A phylogenetic analysis of the genus Dahlia (Asteraceae) based on internal and external transcribed spacer regions of nuclear ribosomal DNA. *Syst. Bot.*, **28**, 627–639.
19. Schultz,J. and Wolf,M. (2009) ITS2 sequence-structure analysis in phylogenetics: a how-to manual for molecular systematics. *Mol. Phylogenet. Evol.*, **52**, 520–523.
20. Wolf,M., Friedrich,J., Dandekar,T. and Muller,T. (2005) CBCAnalyzer: inferring phylogenies based on compensatory base changes in RNA secondary structures. *In Silico Biol.*, **5**, 291–294.
21. Byun,Y. and Han,K. (2009) PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics*, **25**, 1435–1437.
22. Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *Comput. Appl. Biosci.*, **13**, 583–586.

Chapter 10.

Including RNA Secondary Structures improves Accuracy and Robustness in Reconstruction of Phylogenetic Trees

–published in *Biology Direct*–

RESEARCH　　　　　　　　　　　　　　　　　　　　　　**Open Access**

Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees

Alexander Keller[†], Frank Förster[†], Tobias Müller, Thomas Dandekar, Jörg Schultz[*], Matthias Wolf[*]

Abstract

Background: In several studies, secondary structures of ribosomal genes have been used to improve the quality of phylogenetic reconstructions. An extensive evaluation of the benefits of secondary structure, however, is lacking.

Results: This is the first study to counter this deficiency. We inspected the accuracy and robustness of phylogenetics with individual secondary structures by simulation experiments for artificial tree topologies with up to 18 taxa and for divergency levels in the range of typical phylogenetic studies. We chose the internal transcribed spacer 2 of the ribosomal cistron as an exemplary marker region. Simulation integrated the coevolution process of sequences with secondary structures. Additionally, the phylogenetic power of marker size duplication was investigated and compared with sequence and sequence-structure reconstruction methods. The results clearly show that accuracy and robustness of Neighbor Joining trees are largely improved by structural information in contrast to sequence only data, whereas a doubled marker size only accounts for robustness.

Conclusions: Individual secondary structures of ribosomal RNA sequences provide a valuable gain of information content that is useful for phylogenetics. Thus, the usage of ITS2 sequence together with secondary structure for taxonomic inferences is recommended. Other reconstruction methods as maximum likelihood, bayesian inference or maximum parsimony may equally profit from secondary structure inclusion.

Reviewers: This article was reviewed by Shamil Sunyaev, Andrea Tanzer (nominated by Frank Eisenhaber) and Eugene V. Koonin.

Open peer review: Reviewed by Shamil Sunyaev, Andrea Tanzer (nominated by Frank Eisenhaber) and Eugene V. Koonin. For the full reviews, please go to the Reviewers' comments section.

Background

In the last decades, traditional morphological systematics has been augmented by novel molecular phylogenetics. One advantage of molecular data is the increased amount of parsimonious informative characters retained from genes that are usable for the inference of evolutionary relationships. This transition from few morphological features to abundant nucleotide or amino acid information has been a breakthrough for investigations of species relationships [1].

However, genetic data often inherits ambiguous information about phylogenetic relationships. Especially for very closely or distantly related taxa, certain parts of data sets may contradict each other or carry insufficient information. Phylogeneticists counter such problems e.g. by increase of the marker's size by inclusion of more nucleotides, thus increasing the amount of available data [2]. Moreover, different markers are combined, so that for example nuclear or mitochondrial genes are concatenated to increase the power of phylogenetic inferences [3,4]. These methods however face new problems. Increase of the number of nucleotides does not necessarily improve the accuracy of a tree reconstruction. Stochastically, only the robustness of the results is increased, if the complete elongated sequence evolved

* Correspondence: Joerg.Schultz@biozentrum.uni-wuerzburg.de;
matthias.wolf@biozentrum.uni-wuerzburg.de
† Contributed equally
Department of Bioinformatics, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

under the same evolutionary constraints [5]. The second method, marker concatenation, combines genes that result from different evolutionary processes and thus indeed include different evolutionary signals that may improve accuracy. However, they need to be investigated with marker-specific phylogenetic procedures as e.g. varying substitution models [6-8].

In this study we evaluate an alternative method applicable to ribosomal RNA (rRNA) genes that increases information content without addition of nucleotides. As non-coding RNA fragments of the genome, the rRNA gene is generally capable of folding into a secondary structure. In most cases, these structures are necessary for cell function and are thus evolutionarily conserved. Accordingly, structural information may be treated as a conserved marker. Secondary structures of ribosomal RNA therefore offer an additional source of information for tree reconstruction. In particular this is a major advantage in cases where secondary structures are very conserved, yet mutations of nucleotides occur frequently. This applies to the internal transcribed spacer 2 (ITS2) of the eukaryote ribosomal cistron [9,10]. Its secondary structure is evolutionarily maintained as it is of importance in ribogenesis. By contrast, the evolutionary rate of its sequence is relatively high and it is not present in the mature ribosome.

ITS2 sequences have been commonly used to infer phylogenies. Moreover, several studies already included secondary structures in their analyses either by morphometrical matrices or by sequence-structure alignments [11-16]. All these studies agree that the resulting reconstructions are improved by the secondary structures. However, no study has investigated and evaluated this benefit in detail. Evaluations of phylogenetic procedures are typically performed by two different means: the most commonly applied confidence measure in phylogenetics is non-parametric bootstrapping. Bootstrap support values are a measure of robustness of the tree and allow identification of trees or parts of trees that are not unambiguously supported by the data [17,18]. The second point of interest is accuracy measured by the distance between the real and the reconstructed tree. As the 'real' biological tree of life is not available, a switch to sequence simulations along 'real' artificial trees is necessary [19]. In this study we (1) simulate ITS2 sequences along evolutionary trees and (2) compare the results of tree reconstructions by sequence only data and combined sequence-structure data. Additionally, (3) the benefit of structural data is compared with that of sequence elongation. Furthermore, (4) a small biological example of plant phylogeny is presented in which reconstructions that either base on sequence-only or sequence-structure data are compared.

Results

The overall calculation time took 80,000 processor hours on our 40 nodes network cluster. Each node comprised four Xeon 2.33 GHz cores. In total 448 GB RAM were used by the cluster.

The shapes of bootstrap, Quartet distance and Robinson-Foulds distance distributions were similar for equidistant and variable distance trees. However, the branches of the trees for each underlying data set (sequence, sequence-structure and doubled sequence) received higher bootstrap support values and fewer false splits with constant branch lengths compared to variable distances, though differences were minimal (Figs. 1, 2, 3 and 4). Only Quartet distances are shown, since they are congruent with the results of the Robinson-Foulds distance (Additional file 1). Additionally, we included a relative per-branch representation of accuracy divided by the number of internal nodes in the Additional file 1. Bootstrap values and tree distances obtained by differing ancestor sequences were similar in their distributions and thus combined for each scenario during the analysis process. Naturally, with increasing branch lengths, all three investigated data sets (sequences, doubled sequences and sequence-structure) became less accurate and robust, i.e. Quartet distances increased and bootstrap support of nodes decreased. This effect was also observable with an increasing number of external nodes.

Differences between the three methods also increased with evolutionary distance and number of taxa. Thus, the three methods (especially sequence-structure and doubled sequence) yielded almost similar results with low divergence (e.g. branch length 0.05) and few taxa (e. g. 10 taxa), whereas the results were different with branch lengths above 0.25 and at least 14 taxa.

For the lowest branch length we simulated, i.e. 0.025, in comparison to medium divergences a decreased accuracy and bootstrap support was observable with all three methods. This is explainable by too few base changes as providing information for phylogenetic tree reconstruction.

Sequence data performed best in reconstruction of trees (as the maximum and minimum of the spline-curves for bootstraps and tree distances, respectively) at a divergence level between 0.05 and 0.1. Sequence-structure shifted the optimal performance to higher divergences. This effect was also observable for doubled sequence, however it was not as prominent as for sequence-structure.

In general, the robustness of recalculated trees was highest for doubled sequence information contents. However, inclusion of secondary structures largely increased the bootstrap support values of nodes in contrast to normal sequence data. There is thus a
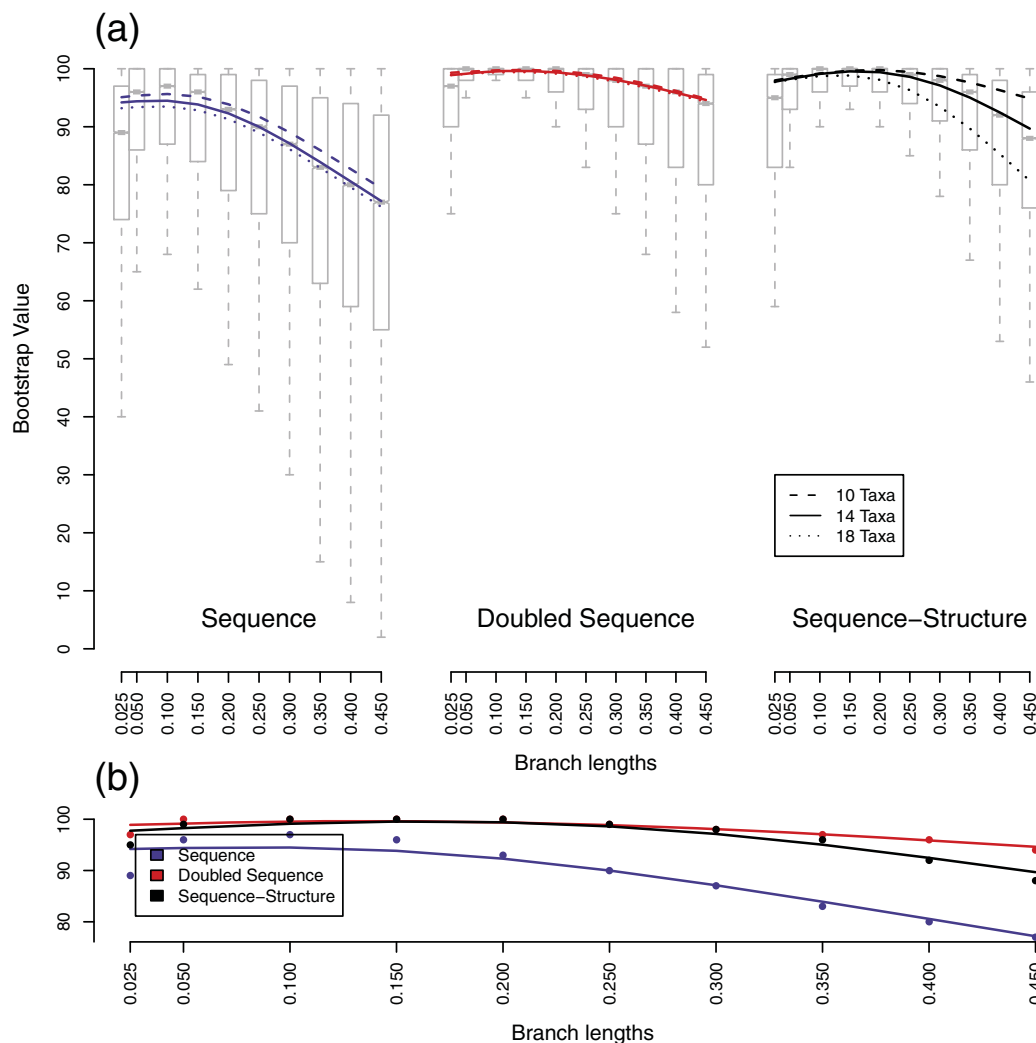
Figure 1 Bootstrap support values for equidistant trees. All five ancestral sequences were combined for a given scenario. (a) Boxplot and solid splines are for 14 taxa scenarios of the three methods. Dashed lines and dotted lines are splines of ten and 18 taxa, respectively. (b) Direct comparison of the 14 taxa splines and medians of all three methods. Sample sizes are 7,000, 11,000 and 15,000 for each of the ten, 14 and 18 taxa scenarios, respectively. Splines show a decrease of robustness with increased number of taxa used and increased branch lengths. Secondary structure and doubled sequences show an improvement in robustness in contrast to normal sequence information.

robustness benefit to using secondary structure that is not directly comparable to benefits achieved by marker elongation.

Additionally, the accuracy of the trees benefitted from secondary structures: the number of false splits was significantly reduced compared to sequence as well as doubled sequence data. Thus sequences-structures yielded the most accurate results in our comparisons.

The results of trees reconstructed with sequence data and sequence-structure data for the plant example were very different. Sequence only information resulted in a correct topology reconstruction of genera (Fig. 5). However, the family of the Malvaceae could not be resolved. This supports the notion that the optimum divergence

level of ITS2 sequences is at the species/genus level (see as well Additional file 2). By contrast, all genera and families could be resolved with secondary structures. This results in a flawless tree topology and highlights the improved accuracy. Furthermore, the robustness of the tree has been enhanced and the optimal divergence level has been widened.

Discussion
Number of Taxa and Divergence
Based on the simulations, we draw several conclusions regarding phylogenetic tree reconstructions with and without secondary structures. First of all, the robustness of a tree and its accuracy were significantly negatively

Figure 2 Quartet distances values for equidistant trees. All five ancestral sequences were combined for a given scenario. (a) Boxplot and solid splines are for 14 taxa scenarios of the three methods. Dashed lines and dotted lines are splines of ten and 18 taxa, respectively. (b) Direct comparison of the 14 taxa splines and medians of all three methods. The samples size of each scenario is 1,000. The accuracy of tree topologies decreases with more taxa and greater evolutionary distances between sequences. Trees calculated with secondary structures or doubled sequences show greater accuracy than those determined with normal sequences.

correlated with number of taxa. This is the case even for normalized per-branch accuracy data (Additional file 1). Graybeal [20] argues that an increased taxon sampling enhances accuracy of a resolved tree in the 'Felsenstein zone'. We argue that such an enhancement is the case for special occurrences of long branch attraction, but not, according to our study, for general tree topologies.

This is in accordance with Bremer et al. [2] as well as Rokas and Carroll [21], who also notice a slight decrease in accuracy with increased taxon sampling.

Secondly, according to Yang [22], a gene has an optimum level of sequence divergence for phylogenetic studies. The upper limits are reached when the observed difference is saturated, whereas the lower boundary is
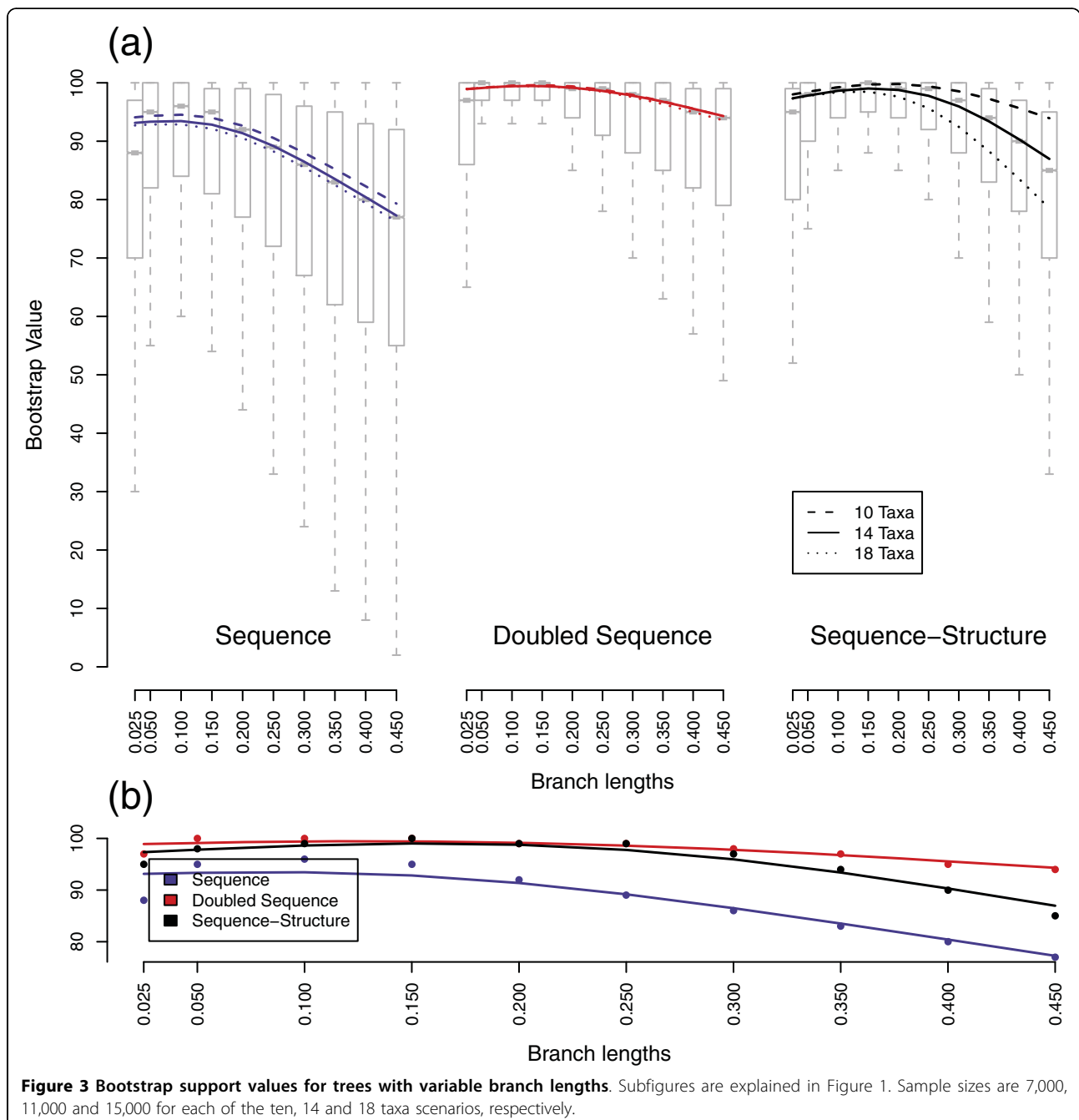
Figure 3 Bootstrap support values for trees with variable branch lengths. Subfigures are explained in Figure 1. Sample sizes are 7,000, 11,000 and 15,000 for each of the ten, 14 and 18 taxa scenarios, respectively.

lack of information content caused by too few substitutions. We observed a similar pattern so that we are able to estimate the divergence level of best performance for ITS2 sequences with and without secondary structures. However, these differ for sequence data and sequence-structure data in two ways: inclusion of secondary structures shifted the best performance to a higher level of divergence. Thus, organisms that are more distantly related can be included in phylogenies. Furthermore, the range of optimal performance is wider for sequence-

structure data. A shift to more distantly related sequences does not necessarily mean that relationships of closely related taxa are not any more resolvable. In a review Coleman [9] also identified this potential of ITS2 secondary structures by discussing several case studies. The small biological example of the Malvales and Sapindales in this study supports this notion. Our study mainly covers artificial data: a large scale comparison with biological data regarding the extension of the performance span is still desirable.
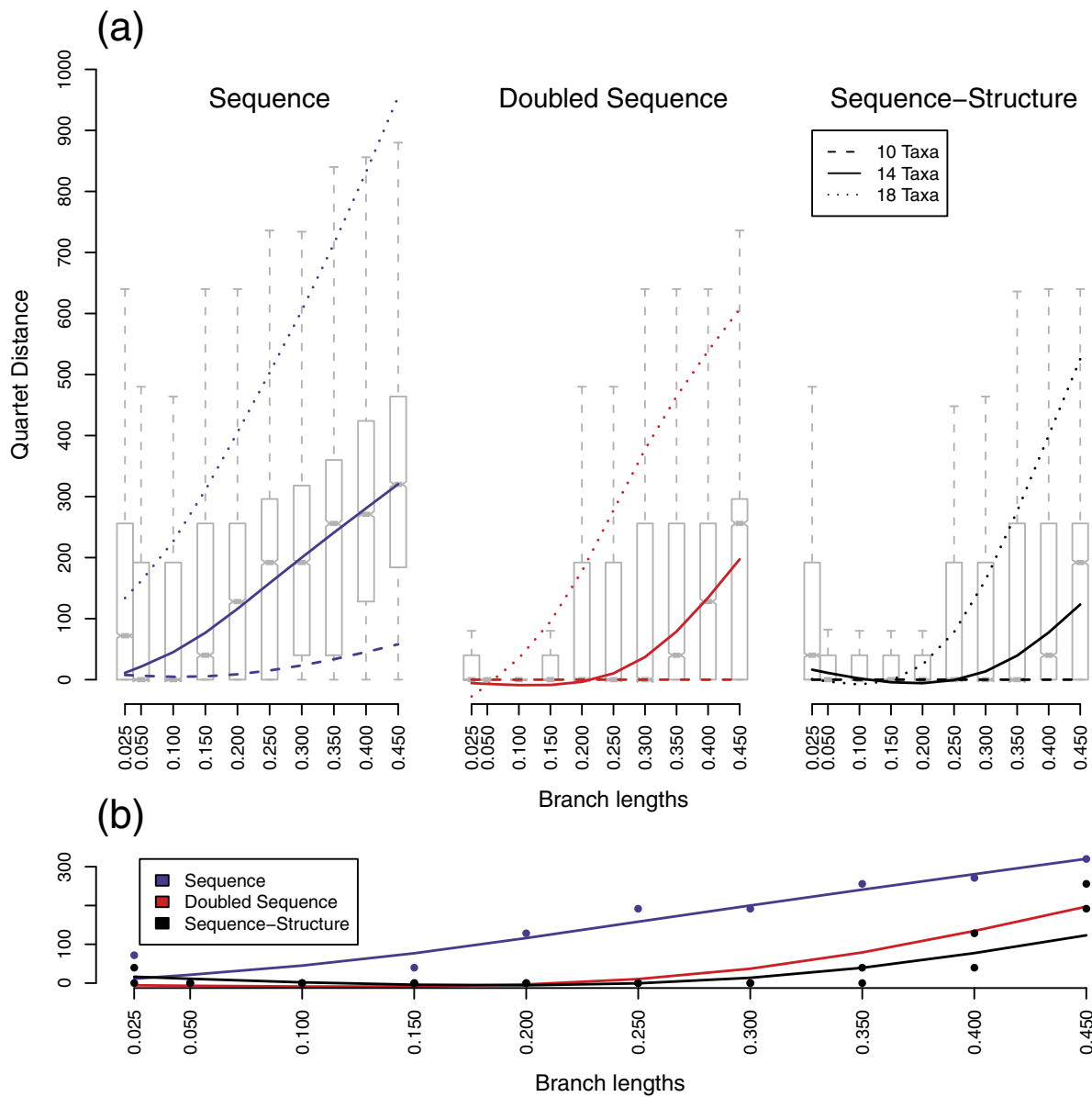
Figure 4 Quartet distances values for trees with variable branch lengths. Subfigures are explained in Figure 2. The samples size of each scenario is 1,000.

Robustness and Accuracy

A substantial benefit to tree robustness was observable when including secondary structure information. Trees reconstructed with secondary structures are generally better bootstrap-supported by the data than those resulting from sequence only data [18]. This is caused by a gain of information content due to increased number of states possible for each nucleotide (unpaired, paired). This information is extractable with a suitable combined score matrix as implemented in 4SALE [23] or similar by site partitioning as in PHASE [24].

The major benefit we identified for phylogenetics is the improvement of accuracy. Sequences-structures performed far better than sequences alone in matching the 'real' tree, especially for high divergences. The resulting immense profit for phylogeneticists is obvious. It is the most crucial property of a phylogenetic tree to be as accurate as possible.

Secondary structure vs. Marker elongation

Both, inclusion of secondary structures and increase of the number of nucleotides improved the reconstructed phylogenetic trees. However, inclusion of secondary

Figure 5 Tree topology of the plants example. Left side: topology and bootstrap values of sequence only data. Right side: corresponding tree with inclusion of secondary structure. Families of the species are given at the right end. GenBank identifiers are in parenthesis after the species names.
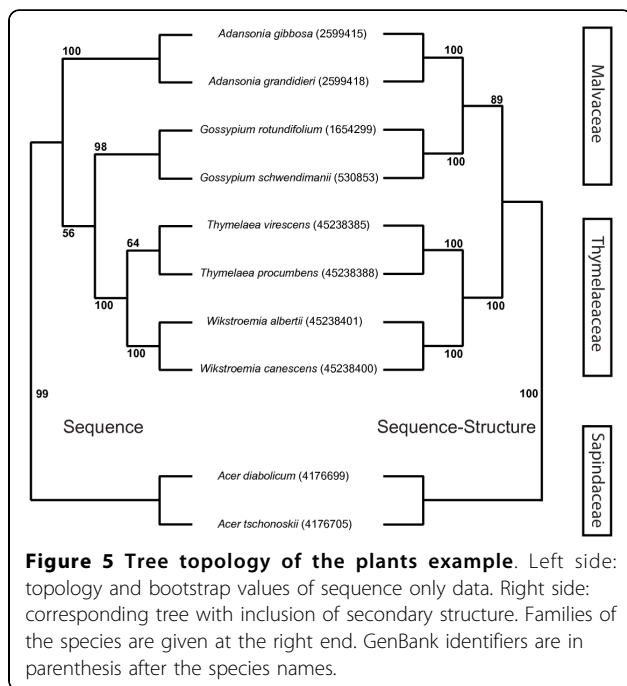
structure in the reconstruction process is not equivalent to marker elongation. The major effect of more nucleotides is to increase the bootstrap support values. This has already been demonstrated by other authors [2,5]. With a theoretical increase of marker's length to infinitely large, corresponding bootstraps within a tree will stochastically be maximized as they exactly represent the data. In contrast, the benefit of secondary structures is predominantly the improvement of a tree's accuracy. Thus, additional sequence elongation and secondary structures represent different types of information increase. As the secondary structure analysis already covers the whole marker region of the ITS2 sequence, sequence elongation is not possible for real biological data.

The results retained in this study for the ITS2 region may be transfered to other ribosomal genes. However, the combination of a conserved secondary structure with a variable sequence seems to be of major benefit in phylogenetic studies. Other ribosomal markers, as e.g. 5.8S or 28S rRNA genes may profit less from addition of secondary structures than the ITS2, as the markers themselves are relatively conserved.

Conclusions

Secondary structures of ribosomal RNA provide a valuable gain of information content that is useful for phylogenetics. Both, the robustness and accuracy of tree reconstructions are improved. Furthermore, this enlarges the optimal range of divergence levels for taxonomic inferences with ITS2 sequences. Thus, the usage of ITS2

sequence together with secondary structure for taxonomic inferences is recommended [25]. This pipeline is theoretically as well applicable to other reconstruction methods as maximum likelihood, bayesian inference or maximum parsimony. They may equally profit from secondary structure inclusion.
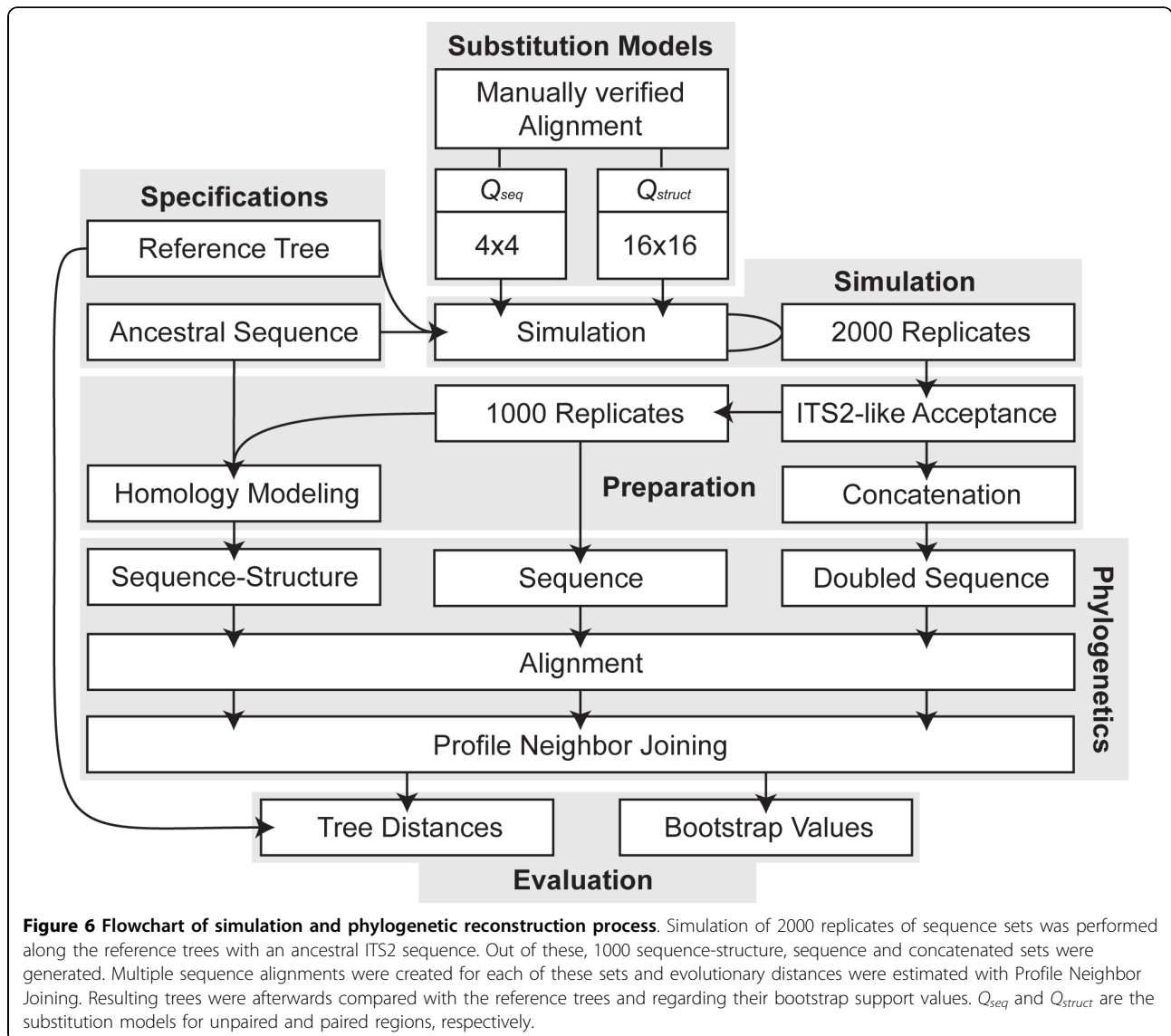
Methods

Simulation of ITS2 Sequences

Simulations of ITS2 sequences were performed with SISSI v0.98 [26]. Secondary structures were included in the simulation process of coevolution by application of two separate substitution models (Fig. 6, Additional file 3: Tab. 1 and Tab. 2): firstly we used a nucleotide 4×4 GTR substitution model Q_{seq} for the evolution of unpaired nucleotides and secondly a dinucleotide 16×16 GTR substitution model Q_{struct} for substitution of bases that form stem regions [11,27]. Q_{seq} and Q_{struct} were both estimated by a manually verified alignment based on 500 individual ITS2 sequences and structures with a variant of the method described by Müller and Vingron [28]. For lack of information about insertion and deletion events in the ITS2 region, such were not included into the simulations.

Simulations were started given (a) an ancestral sequence and (b) a reference tree that contained (c) specific branch lengths and (d) a certain number of taxa. In total, we used 10 different branch lengths, 5 ancestral sequences and 6 different trees (3 topologies for equal and variable branch length) resulting in 300 different combinatory conditions as evolutionary scenarios. (a) Ancestral sequences and structures were taken from the ITS2 database after HMM annotation [29-31]. They represented a cross section of the Eukaryota i.e. *Arabidopsis* (Plants) [GenBank:1245677], *Babesia* (Alveolata) [GenBank:119709754], *Gigaspora* (Fungi) [GenBank:3493494], *Gonium* (Green Algae) [GenBank:3192577] and *Haliotis* (Animals) [GenBank:15810877]. (b) The complete procedure was accomplished for two trees that shared a similar topology (Fig. 7). Tree shapes were chosen to resemble trees of a previously published simulation study [32]. The first was a tree that included constant branch lengths, whereas the second tree alternately varied +/- 50% of a given branch length. (c) The used branch lengths were 0.025, 0.05, 0.01, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4 and 0.45. For comparison, pairwise distances of a typical phylogenetic study with ITS2 sequences have been added as Additional file 2. (d) Reference trees were calculated for 10, 14 and 18 taxa. The ancestral sequence served as an origin of the simulated sequences, but was not included in the reconstruction process and resulting tree.

Each simulated sequence set contained sequences according to the number of taxa. Sequence sets were

Figure 6 Flowchart of simulation and phylogenetic reconstruction process. Simulation of 2000 replicates of sequence sets was performed along the reference trees with an ancestral ITS2 sequence. Out of these, 1000 sequence-structure, sequence and concatenated sets were generated. Multiple sequence alignments were created for each of these sets and evolutionary distances were estimated with Profile Neighbor Joining. Resulting trees were afterwards compared with the reference trees and regarding their bootstrap support values. Q_{seq} and Q_{struct} are the substitution models for unpaired and paired regions, respectively.

accepted as composed of ITS2-like sequences if the structure of each sequence had been determinable by homology modeling with a threshold of 75% helix transfer [33]. For homology modeling, the ancestral sequence served as a template. Thus, each structure had four helices with the third helix as the longest. This acceptance scheme has been introduced for two reasons: the data is very similar to biological samples [10] and the structure prediction method is equal to that used at the ITS2 database [30] as well as phylogenetic reconstructions [25]. In total, 2,000 valid sequence sets were obtained for each scenario, what corresponds to 600,000 sequence sets summarized over all scenarios.

The complete sequence set is downloadable at the Supplements section of the ITS2 Database http://its2.bioapps.biozentrum.uni-wuerzburg.de/.

Sequences and Structures of the Data Sets
Sequence data set: for each scenario, the order of the 2,000 simulated sequence sets retained from SISSI was shuffled. The first 1,000 were chosen and used as a sequence data set.

Sequence-structure data set: for each of the sequence sets used in the sequence data set, we determined the individual secondary structure of each sequence by homology modeling with at least 75% helix transfer [33]. The ancestral sequence was used as a template. Thus, for the sequence-structure data set we combined sequences with their respective secondary structures according to Seibel et al. [23]. Note, this approach using individual secondary structures is in contrast to alignments only guided by a consensus structure. Doubled nucleotide data set: The remaining 1,000 simulated
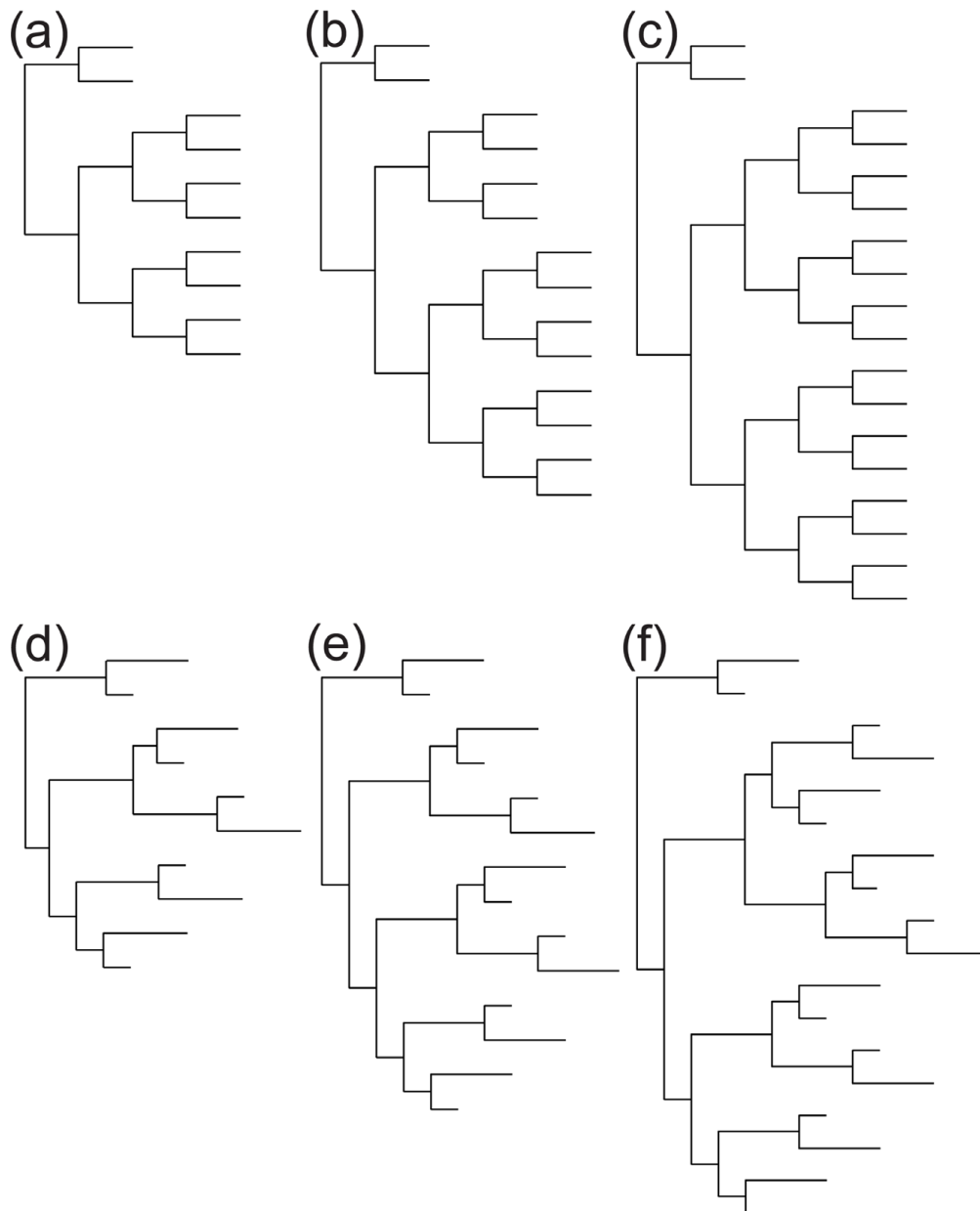
Figure 7 Reference tree topologies used for simulation process. Trees (a), (b) and (c) were trees with equidistance of branches. Trees (d), (e) and (f) were the corresponding variable trees with varying branch lengths. Trees (a) and (d) include ten taxa, (b) and (e) 14 taxa and (c) and (f) 18 taxa.

sequence sets were used to exemplify effects on phylogenetic analyses of a hypothetical ITS2 gene size duplication. Each sequence of these sets was concatenated with a corresponding sequence of the sequence data set (same taxon in the simulation trees). Thus we received a data set of doubled nucleotide content that includes as well 1,000 sequence sets.

Reconstruction of Simulated Phylogenetic Trees

For each simulated sequence set, ClustalW v2.0.10 [34] was used for calculation of multiple sequence alignments. In the cases of sequences and doubled sequences we used an ITS2 specific 4×4 scoring matrix [29,30]. For secondary structures, we translated sequence-structure information prior to alignment into pseudoproteins as described for 4SALE v1.5 [23,35]. Pseudoproteins were coded such that each of the four nucleotides may be present in three different states: unpaired, opening base-pair and closing base-pair. Thus, an ITS2 specific 12×12 scoring matrix was used for calculation of the alignment [23].

Reconstruction of phylogenetic trees for all trees has been performed with Profile Neighbor Joining (PNJ) of a console version of ProfDistS 0.9.8 [36,37]. With this we estimated improvements due to secondary structures, but keep the method of reconstruction constant. We decided in favor of PNJ and against other methods like maximum likelihood, Bayesian inference and parsimony for several reasons: the distance matrices are independent of insertion and deletion events, the algorithm is very fast and a pipeline for reconstructions with PNJ using secondary structures is already published [25]. However beneficial effects may be transferable to these methods. Profile building was allowed with default settings. General time reversible models (GTRs) were applied with the corresponding 4×4 and 12×12 substitution matrices for sequences and sequences-structures, respectively.

Robustness and Accuracy

Profile Neighbor Joining trees were bootstrapped with 100 pseudo-replicates to retain information about the stability of the resulting tree. Bootstrap support values of all tree branches obtained from the 1,000 sequence sets of a certain scenario were extracted and pooled. Furthermore, the resulting trees were compared to the respective reference tree. In this regard, two tree distance quantification methods were applied, Robinson-Foulds distances using the Phylip Package v3.68 [38] and Quartet distances using Qdist v1.0.6 [39]. Results of all sequence sets were combined for a given scenario to receive the distributions of bootstrap values, Quartet distances and Robinson-Foulds distances, respectively. The result of each 14-taxa-scenario was plotted as a boxplot with notches using R v2.9.0 [40]. An interpolating spline curve was added. For the remaining scenarios (10 and 18 taxa) only spline curves were added for the sake of clarity.

Short biological case study

Here we provide a short example of ITS2 secondary structure phylogeny, applied to biological data: we sampled sequences of three plant families using the ITS2-database browse feature (database accessed: June 2009): Thymelaeaceae (Malvales), Malvalceae (Malvales) and Sapindaceae (Sapindales). For each family we chose two sequences of the first two appearing genera. Tree reconstruction followed the methods described by Schultz and Wolf [25] and is equivalent to the reconstruction procedure used for the simulated sequence sets. Furthermore, the same procedure was applied without secondary structure information for comparison.

Reviewers' comments

Reviewer's report 1

Shamil Sunyaev, Division of Genetics, Dept. of Medicine, Brigham & Women's Hospital and Harvard Medical School

This manuscript demonstrates the utility of taking into account secondary structure in the phylogenetic analysis. Using comprehensive simulations and a real dataset of ITS2 sequences the authors demonstrated that for higher sequence divergence trees constructed with the help of secondary structure information improve accuracy and robustness. Another interesting result is that addition of taxa may reduce accuracy of tree reconstruction at least in terms of quartet distance between reconstructed and true trees.

Author's response

Thanks a lot for this positive report!

Reviewer's report 2

Andrea Tanzer, Institute for Theoretical Chemistry, University of Vienna (nominated by Frank Eisenhaber, Bioinformatics Institute (BII) Agency for Science, Technology and Research, Singapore)

General comments:

The manuscript "Ribosomal Secondary Structures improve Accuracy and Robustness in Reconstruction of Phylogenetic Trees" compares different methods to improve the quality of phylogenetic analysis. RNA secondary structure information has been included in a variety of previous phylogenetic analysis, but this is the first study exploring the effect on the resulting trees in detail.

The authors use internal transcribed spacer 2 of ribosomal RNAs, a well established set of markers, to simulate a broad spectrum of 300 different scenarios. In addition, they compare their results from the simulations to a set of biological examples from selected plant species.

Overall, the manuscript is carefully written and the authors chose analysis and method appropriately. The simulated sequence set could be used for future studies.

Minor comments:

*) The title might be a little bit miss-leading since 'Ribosomal Secondary Structures' do not improve the 'Accuracy and Robustness in Reconstruction of Phylogenetic Trees' in general and the method should be applicable to other RNA markers. Therefore, I suggest something like "Including Secondary Structures improve Accuracy and Robustness in Reconstruction of Phylogenetic Trees".

*) The setup for the simulations is quite complex. It might help the reader if you add a table or figure to the supplemental material that summarizes the individual conditions for each data set produced.

Alternatively, you could just add to the text that you use 10 different branch length, 5 ancestral sequences and 6 different trees (3 topologies for equal and variable branch length) resulting in 300 different conditions. If I understand this correctly, then you retrieved for each of these 300 conditions 2,000 sequence sets (a total of

600,000 sets), where each set contains 10, 14 and 18 taxa, resp., depending on the tree topology used. These numbers should be mentioned in the text.

*) The set of simulated sequences should be accessible, such that it can be downloaded and used by the community for further studies. Maybe put a link on the website of the ITS2 database.

*) Predicting secondary structures of single sequences occasionally results in (mfe) structures of unexpected shapes. One way to get around this problem is the calculation of consensus structures of a set of related sequences. The resulting consensus structures can then be used for contraint folding of those sequences that could not be folded correctly in the first place. Furthermore, the sequences might fold into a number of equally good structures, but folding programs present only the first result (under default settings). The 'true' structure could as well be among the best folds, but not necessarily the optimal one (suboptimal folding). After all, folding algorithms only make the most plausible predictions. In this study, prediction of RNA secondary structures includes homology modelling. It is of question weather this is the most efficient method. However, since the structures deposited at the ITS2 database were created that way, it seems legitimate to apply it here as well.

Author's response

Thank you for carefully reading the manuscript. We addressed the minor comments regarding text changes and included the necessary information within the text. The set of simulated sequences is now downloadable at the Supplement section of the ITS2 Database http://its2.bioapps.biozentrum.uni-wuerzburg.de/. We totally agree that there are other possibly more efficient methods concerning structure prediction. However, as already stated by Dr. Tanzer 'structures deposited at the ITS2 database were created that way [homology modelling], it seems legitimate to apply it here as well'. The big advantage of the ITS2 is, that the core folding pattern is already known. Therefore, we have an external criterium to check for the correctness of the predicted structures.

Reviewer's report 3

Eugene V. Koonin, National Center for Biotechnology Information, NIH, Bethesda

This is a useful method evaluation work that shows quite convincingly the inclusion of RNA secondary structure information into phylogenetic analysis improves the accuracy of neighbor-joining trees. My only regrets are about a certain lack of generality. It would be helpful to see a similar demonstration for for at least two different kinds of nucleic acid sequences not only ITS2. Also, at the end of the Conclusion section, the authors suggest that secondary structure could help also with other phylogenetic approaches (ML etc).

Showing this explicitly would be helpful, especially, given that NJ is hardly the method of choice in today's phylogenetics.

Author's response

Thank you for your encouraging report. For ITS2 the core structure is well known and there are about 200,000 individual secondary structures available. However, it is absolutely right that it would be helpful to perform an analysis also on other types of phylogenetic RNA markers. Unfortunately, today there is no comparable amount of data available concerning secondary structures of other RNAs. Similarly, there are no programs to run an analysis on other methods such as parsimony, maximum likelihood and/or bayesian methods simultaneously considering sequence and secondary structure information.

Additional file 1: Normalized Quartet distance and Robinson-Foulds plots. Similar to Figures 2 and 4, but showing per-branch Quartet distances as a normalized standard i.e. divided by number of splits. Robinson-Foulds Distances are given in absolute and normalized versions.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-6150-5-4-S1.PDF]

Additional file 2: Empirical pairwise distances. Pairwise distances of an ITS2 case study that integrates secondary structure.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-6150-5-4-S2.PDF]

Additional file 3: Substitution matrices. Nucleotide 4×4 GTR substitution model Q_{seq} for the evolution of unpaired nucleotides and a dinucleotide 16×16 GTR substitution model Q_{struct}.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1745-6150-5-4-S3.PDF]

Authors' contributions

AK, JS, MW and TD designed the study. FF and AK performed the simulation experiments and analyses. FF and TM estimated the substitution models used for simulations and reconstructions. AK, FF and MW drafted the manuscript. All authors contributed to writing the paper, read the final manuscript and approved it.

Competing interests

The authors declare that they have no competing interests.

References

1. Woese C, Kandler O, Wheelis M: **Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.** *Proc Natl Acad Sci USA* 1990, **87(12)**:4576-4579.

2. Bremer B, Jansen R, Oxelman B, Backlund M, Lantz H, Kim KJ: **More characters or more taxa for a robust phylogeny-case study from the Coffee family (Rubiaceae).** *Syst Biol* 1999, **48(3)**:413-435.
3. van Oppen M, McDonald B, Willis B, Miller D: **The evolutionary history of the coral genus** *Acropora* **(Scleractinia, Cnidaria) based on a mitochondrial and a nuclear marker: reticulation, incomplete lineage sorting, or morphological convergence?.** *Mol Biol Evol* 2001, **18(7)**:1315-1329.
4. Slowinski J, Lawson R: **Snake phylogeny: evidence from nuclear and mitochondrial genes.** *Mol Phylogenet Evol* 2002, **24(2)**:194-202.
5. Erixon P, Svennblad B, Britton T, Oxelman B: **Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics.** *Syst Biol* 2003, **52(5)**:665-73.
6. Whelan S, Liò P, Goldman N: **Molecular phylogenetics: state-of-the-art methods for looking into the past.** *Trends Genet* 2001, **17(5)**:262-72.
7. Posada D, Crandall KA: **The effect of recombination on the accuracy of phylogeny estimation.** *J Mol Evol* 2002, **54(3)**:396-402.
8. Egger B, Koblmüller S, Sturmbauer C, Sefc K: **Nuclear and mitochondrial data reveal different evolutionary processes in the Lake Tanganyika cichlid genus** *Tropheus.* *BMC Evol Biol* 2007, **7**:137.
9. Coleman AW: **ITS2 is a double-edged tool for eukaryote evolutionary comparisons.** *TIG* 2003, **19(7)**:370-375.
10. Coleman AW: **Pan-eukaryote ITS2 homologies revealed by RNA secondary structure.** *Nucleic Acids Res* 2007, **35(10)**:3322-3329.
11. Schöniger M, von Haeseler A: **A stochastic model for the evolution of autocorrelated DNA sequences.** *Mol Phylogenet Evol* 1994, **3(3)**:240-7.
12. Tillier ERM, Collins RA: **High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA.** *Genetics* 1998, **148(4)**:1993-2002.
13. Young I, Coleman AW: **The advantages of the ITS2 region of the nuclear rDNA cistron for analysis of phylogenetic relationships of insects: a** *Drosophila* **example.** *Mol Phylogenet Evol* 2004, **30**:236-242.
14. Biffin E, Harrington M, Crisp M, Craven L, Gadek P: **Structural partitioning, paired-sites models and evolution of the ITS transcript in** *Syzygium* **and** *Myrtaceae.* *Mol Phylogenet Evol* 2007, **43**:124-139.
15. Grajales A, Aguilar C, Sanchez J: **Phylogenetic reconstruction using secondary structures of Internal Transcribed Spacer 2 (ITS2, rDNA): finding the molecular and morphological gap in Caribbean gorgonian corals.** *BMC Evol Biol* 2007, **7**:90.
16. Keller A, Schleicher T, Förster F, Ruderisch B, Dandekar T, Müller T, Wolf M: **ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales).** *BMC Evol Biol* 2008, **8**:218.
17. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985, **39(4)**:1993-2002.
18. Hillis D, Bull J: **An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis.** *Syst Biol* 1993, **42(2)**:182-192.
19. Hillis DM, Huelsenbeck JP, Cunningham CW: **Application and accuracy of molecular phylogenies.** *Science* 1994, **264(5159)**:671-7.
20. Graybeal A: **Is it better to add taxa or characters to a difficult phylogenetic problem?.** *Syst Biol* 1998, **47**:9-17.
21. Rokas A, Carroll SB: **More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy.** *Mol Biol Evol* 2005, **22(5)**:1337-44.
22. Yang Z: **On the best evolutionary rate for phylogenetic analysis.** *Syst Biol* 1998, **47**:125-33.
23. Seibel PN, Müller T, Dandekar T, Schultz J, Wolf M: **4SALE - a tool for synchronous RNA sequence and secondary structure alignment and editing.** *BMC Bioinformatics* 2006, **7**:498.
24. Jow H, Hudelot C, Rattray M, Higgs P: **Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution.** *Mol Biol Evol* 2002, **19(9)**:1591-1601.
25. Schultz J, Wolf M: **ITS2 sequence-structure analysis in phylogenetics: a how-to manual for molecular systematics.** *Mol Phylogenet Evol* 2009, **52**:520-523.
26. Gesell T, von Haeseler A: **In silico sequence evolution with site-specific interactions along phylogenetic trees.** *Bioinformatics* 2006, **22(6)**:716-722.
27. Meyer S, von Haeseler A: **Identifying site-specific substitution rates.** *Mol Biol Evol* 2003, **20(2)**:182-189.
28. Müller T, Vingron M: **Modeling amino acid replacement.** *J Comput Biol* 2000, **37(6)**:761-776.
29. Schultz J, Müller T, Achtziger M, Seibel PN, Dandekar T, Wolf M: **The internal transcribed spacer 2 database-a web server for (not only) low level phylogenetic analyses.** *Nucleic Acids Res* 2006, **34(Supp 2)**:W704-707.
30. Selig C, Wolf M, Muller T, Dandekar T, Schultz J: **The ITS2 Database II: homology modelling RNA structure for molecular systematics.** *Nucleic Acids Res* 2008, , **36 Database:** D377-80.
31. Keller A, Schleicher T, Schultz J, Müller T, Dandekar T, Wolf M: **5.8S-28S rRNA interaction and HMM-based ITS2 annotation.** *Gene* 2009, **430(1-2)**:50-7.
32. Alfaro ME, Zoller S, Lutzoni F: **Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence.** *Mol Biol Evol* 2003, **20(2)**:255-266.
33. Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T: **Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures.** *RNA* 2005, **11(11)**:1616-1623.
34. Thompson J, Higgins D, Gibson T: **ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22)**:4673-4680.
35. Seibel PN, Müller T, Dandekar T, Wolf M: **Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE.** *BMC Res Notes* 2008, **1**:91.
36. Friedrich J, Dandekar T, Wolf M, Müller T: **ProfDist: a tool for the construction of large phylogenetic trees based on profile distances.** *Bioinformatics* 2005, **21(9)**:2108-2109.
37. Wolf M, Ruderisch B, Dandekar T, Schultz J, Müller T: **ProfDistS: (profile-) distance based phylogeny on sequence - structure alignments.** *Bioinformatics* 2008, **24**:2401-2402.
38. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5**:164-166.
39. Mailund T, Pedersen CNS: **QDist-quartet distance between evolutionary trees.** *Bioinformatics* 2004, **20(10)**:1636-7.
40. R Development Core Team: *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria 2009http://www.R-project.org.

Chapter 11.

ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales)

Research article

ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales)

Alexander Keller[†], Tina Schleicher[†], Frank Förster, Benjamin Ruderisch, Thomas Dandekar, Tobias Müller and Matthias Wolf*

Address: Department of Bioinformatics, University of Würzburg, Am Hubland, 97074 Würzburg, Germany

Email: Alexander Keller - a.keller@biozentrum.uni-wuerzburg.de; Tina Schleicher - tina.schleicher@biozentrum.uni-wuerzburg.de; Frank Förster - frank.foerster@biozentrum.uni-wuerzburg.de; Benjamin Ruderisch - b.ruderisch@gmx.de; Thomas Dandekar - dandekar@biozentrum.uni-wuerzburg.de; Tobias Müller - tobias.mueller@biozentrum.uni-wuerzburg.de; Matthias Wolf* - matthias.wolf@biozentrum.uni-wuerzburg.de

* Corresponding author   †Equal contributors

Abstract

Background: Within Chlorophyceae the ITS2 secondary structure shows an unbranched helix I, except for the '*Hydrodictyon*' and the '*Scenedesmus*' clade having a ramified first helix. The latter two are classified within the Sphaeropleales, characterised by directly opposed basal bodies in their flagellar apparatuses (DO-group). Previous studies could not resolve the taxonomic position of the '*Sphaeroplea*' clade within the Chlorophyceae without ambiguity and two pivotal questions remain open: (1) Is the DO-group monophyletic and (2) is a branched helix I an apomorphic feature of the DO-group? In the present study we analysed the secondary structure of three newly obtained ITS2 sequences classified within the '*Sphaeroplea*' clade and resolved sphaeroplealean relationships by applying different phylogenetic approaches based on a combined sequence-structure alignment.

Results: The newly obtained ITS2 sequences of *Ankyra judayi, Atractomorpha porcata* and *Sphaeroplea annulina* of the '*Sphaeroplea*' clade do not show any branching in the secondary structure of their helix I. All applied phylogenetic methods highly support the '*Sphaeroplea*' clade as a sister group to the 'core Sphaeropleales'. Thus, the DO-group is monophyletic. Furthermore, based on characteristics in the sequence-structure alignment one is able to distinguish distinct lineages within the green algae.

Conclusion: In green algae, a branched helix I in the secondary structure of the ITS2 evolves past the '*Sphaeroplea*' clade. A branched helix I is an apomorph characteristic within the monophyletic DO-group. Our results corroborate the fundamental relevance of including the secondary structure in sequence analysis and phylogenetics.

Background

Taxonomists face inconsistent or even contradictory clues when they examine the affiliation of organisms to higher taxonomic groupings. Several characters may yield alternative hypotheses explaining their evolutionary background. This also applies to the taxonomic position of the Sphaeropleaceae [1-23]. Different authors affiliate the green algal family by morphological characters to either ulvophytes or chlorophytes, until amendatory Deason et al. [10] suggested that the Neochloridaceae, the Hydrodic-

tyaceae and the Sphaeropleaceae should be grouped as Sphaeropleales within the chlorophytes, since all of them have motile biflagellate zoospores with a direct-opposite (DO) confirmation of basal bodies.

Subsequently, other taxonomic lineages (the '*Ankistrodesmus*' clade, the '*Bracteacoccus*' clade, the '*Pseudomuriella*' clade, *Pseudoschroederia*, the '*Scenedesmus*' clade, *Schroederia* and the '*Zofingiensis*' clade) were added to this biflagellate DO group, because they show molecular affiliation to either Neochloridaceae or Hydrodictyaceae [24].

Although nowadays most authors agree that the DO group is monophyletic, until now no study pinpointed the taxonomic linkage of the name-giving '*Sphaeroplea*' clade to the remaining 'core Sphaeropleales' persuasively with genetic evidence [6,23], i.e. the sister clade remains unclear [15,24]. Likewise, with respect to morphology, studies of 18S and 26S rRNA gene sequences neither resolve the basal branching patterns within the Chlorophyceae with high statistical power nor corroborate a monophyletic biflagellate DO group without ambiguity [6,23].

Müller et al. [25] obtained moderate statistical support for the close relationship of the '*Sphaeroplea*' clade and the 'core Sphaeropleales' with profile distances of 18S and 26S rDNA. In this study we followed and expanded their methodology with a very different phylogenetic marker. The internal transcribed spacer 2 (ITS2), the region of ribosomal RNA between the 5.8S rRNA gene and the large subunit (26S rDNA) has proven to be an appropriate marker for the study of small scale phylogenies of close relatives [26-29]. The sequence is in contrast to the bordering regions of ribosomal subunits evolutionary not conserved, thus genetic differentiation is detectable even in closely related groups of organisms. By contrast, the secondary structure seems to be well conserved and thus provides clues for higher taxonomic studies [27,30-33]. Secondary structure information is furthermore especially interesting within the Chlorophyceae, because van Hannen et al. [34] described an uncommon branching of ITS2 helix 1 within the genera *Desmodesmus*, *Hydrodictyon* [35] and *Scenedesmus*. It is not known when this feature evolved and whether it is, as we expect, an apomorphic feature for the DO-group. It is obvious that phylogenetic statements should be improvable by inclusion of structural information in common sequence analysis. For example, Grajales et al. [36] calculated morphometric matrices from ITS2 secondary structures for phylogenetic analyses, but treated information of sequence and structure as different markers. Here we combine sequence with structural information in just one analysis. Aside from the biological problem, we address the pivotal question of a

methodological pipeline for sequence-structure phylogenetics using rDNA data.

Methods

DNA extraction, amplification and sequencing

Extraction of genomic DNA from cultured cells of *Ankyra judayi*, *Atractomorpha porcata* and *Sphaeroplea annulina* was done using Dynabeads® (DNA DIRECT Universal, Dynal Biotech, Oslo, Norway) according to the manufacturer's protocol. PCR reactions were performed in a 50 μl reaction volume containing 25 μl FastStart PCR Master (Roche Applied Science), 5 μl gDNA and 300 nM of the primers ITS3 (5'-GCA TCG ATG AAG AAC GCA GC-3') and ITS4 (5'-TCC TCC GCT TAT TGA TAT GC-3') designed by White et al. [37].

Cycling conditions for amplification consisted of 94°C for 10 min, 30 cycles of 94°C for 30 s, 50°C for 30 s and 72°C for 45 s, followed by a final extension step of 10 min at 72°C. PCR products were analysed by 3% agarose gel electrophoresis and ethidium bromide staining.

PCR probes where purified with the PCR Purificaton Kit (Qiagen) and where quantified by spectrometry. Each sequencing probe was prepared in an 8 μl volume containing 20 ng DNA and 1.25 μM Primer. Sequencing was carried out using an annealing temperature of 50°C with the sequencer Applied Biosystems QST 3130 Genetic Analyzer by the Institute of Hygiene and Microbiology (Würzburg, Germany).

ITS2 secondary structure prediction

ITS2 secondary structures of the three newly obtained sequences were folded with the help of RNAstructure [38] and afterwards manually corrected. All available 788 chlorophycean ITS2 sequences were obtained from the NCBI nucleotide database. The ITS2 secondary structure of *Atractomorpha porcata* was used as template for homology modelling. Homology modelling was performed by using the custom modelling option as provided with the ITS2-Database [30-33] (identity matrix and 50% threshold for the helix transfer). Forty-nine species representing the chlorophycean diversity were retained and used as comparative taxa in inferring phylogenies (Table 1). For this taxon sampling, accurate secondary structures of sequences were now folded by RNAstructure and additionally corrected using Pseudoviewer 3 [39]. We standardized start and end of all helices according to the optimal folding of the newly obtained sequences.

Alignment and phylogenetic analyses

Using 4SALE [40,41] with its ITS2 specific scoring matrix, we automatically aligned sequences and structures simultaneously. Sequence-structure alignment is available at the ITS2 database supplements page. For the complete

Table 1: Chlorophyte species used for this investigation.

Clade	Species	Strain	GenBank
'Sphaeroplea '	*Ankyra judayi* (G.M. Smith) Fott 1957	SAG 17.84	EU352800
	Atractomorpha porcata Hoffman 1984 strain	SAG 71.90	EU352803
	Sphaeroplea annulina (Roth) C. Agardh 1824	SAG 377.1a	EU352801
	Sphaeroplea annulina (Roth) C. Agardh 1824	SAG 377.1e	EU352802
'Dunaliella'	*Haematococcus droebakensis* Wollenweber 1908	-	U66981
	Dunaliella parva Lerche 1937	-	DQ116746
	Dunaliella salina (Dunal) Teodoresco 1905	CCAP 19/18	EF473746
'Hydrodictyon '	*Hydrodictyon africanum* Yamanouchi 1913	UTEX 782	AY779861
	Hydrodictyon patenaeforme Pocock	CCAP 236/3	AY577736
	Hydrodictyon reticulatum (Linnaeus) B. de St.-Vincent 1824	CBS	AY779862
	Pediastrum braunii Wartmann 1862	SAG 43.85	AY577756
	Pediastrum duplex Meyen 1829	UTEX 1364	AY779868
	Pseudopediastrum boryanum (Raciborski) Sulek 1969	UTEX 470	AY779866
	Sorastrum spinulosum Nägeli 1849	UTEX 2452	AY779872
	Stauridium tetras (Ehrenberg) Ralfs 1844	EL 0207 CT	AY577762
'Oedogonium'	*Bulbochaete hiloensis* (Nordstedt) Tiffany 1937	-	AY962677
	Oedogonium cardiacum (Hassall) Wittrock 1870	-	AY962675
	Oedogonium nodulosum Wittrock 1872	-	DQ078301
	Oedogonium oblongum Wittrock 1872	-	AY962681
	Oedogonium undulatum (Brébisson) A. Braun 1854	-	DQ178025
'Reinhardtii'	*Chlamydomonas incerta* Pascher 1927	SAG 81.72	AJ749625
	Chlamydomonas komma Skuja 1934	-	U66951
	Chlamydomonas petasus Ettl	SAG 11.45	AJ749615
	Chlamydomonas reinhardtii Dangeard 1888	CC-620	AJ749638
	Chlamydomonas typica Deason & Bold 1960	SAG 61.72	AJ749622
	Eudorina elegans Ehrenberg 1831	ASW 107	AF486524
	Eudorina unicocca G.M. Smith 1930	UTEX 1215	AF486525
	Gonium octonarium Pocock 1955	Tex	AF054424
	Gonium pectorale O.F. Müller 1773	Chile K	AF054440
	Gonium quadratum E. G. Pringsheim ex H. Nozaki	Cal 3-3	AF182430
	Pandorina morum (O.F. Müller) Bory de Saint-Vincent 1824	Chile	AF376737
	Volvox dissipatrix (Shaw) Printz	-	U67020
	Volvox rousseletii G.S.West	-	U67025
	Volvulina steinii Playfair 1915	-	U67034
	Yamagishiella unicocca (Rayburn & Starr) Nozaki 1992	ASW 05129	AF098181
'Scenedesmus'	*Desmodesmus abundans* (Kirchner) Hegewald 2000	UTEX 1358	AJ400494
	Desmodesmus bicellularis (Chodat) An, Friedl & Heg. 1999	CCAP 276/14	AJ400498
	Desmodesmus communis (Hegewald) Hegewald 2000	UTEX 76	AM410660
	Desmodesmus elegans (Hortobágyi) Heg. & Van. 2007	Heg 1976–28	AM228908
	Desmodesmus opoliensis (P.G. Richter) Hegewald 2000	EH 10	AM410655
	Desmodesmus pleiomorphus (Hindák) Hegewald 2000	UTEX 1591	AM410659
	Desmodesmus quadricauda (Turpin) Hegewald		AJ400495
	Scenedesmus acuminatus (Lagerheim) Chodat 1902	UTEX 415	AJ249511
	Scenedesmus acutiformis (B. Schröder) F. Hindák 1990	SAG 276.12	AJ237953
	Scenedesmus basiliensis Chodat 1926	UTEX 79	AJ400489
	Scenedesmus dimorphus (Turpin) Kützing 1833	UTEX 417	AJ400488
	Scenedesmus longus Meyen 1829 ex Ralfs	NIOO-MV5	AJ400506
	Scenedesmus obliquus (Turpin) Kützing 1833	Tow 9/21P-1W	DQ417568
	Scenedesmus pectinatus Meyen 1828	An 111a	AJ237954
	Scenedesmus platydiscus (G.M. Smith) Chodat 1926	UTEX 2457	AJ400491
	Scenedesmus raciborskii Woloszynska 1914	An 1996–5	AJ237952
	Scenedesmus regularis Svirenko	Heg 1998–2	AY170857
	Scenedesmus wisconsinensis (G.M. Smith) Chodat 1996	An 41	AJ237950

Listed is the current clade classification of the species [69,70,24] and the GenBank accession numbers of the analyzed sequences. The four newly obtained sequences are of the '*Sphaeroplea*' clade.

alignment we tested for appropriate models of nucleotide substitution using the Akaike Information Criterion (AIC) as implemented in Modeltest [42]. The following PAUP-block was used for all maximum likelihood based phylogenetic analyses with PAUP* [43]: Lset Base = (0.2299 0.2415 0.2152) Nst = 6 Rmat = (1.4547 3.9906 2.0143 0.1995 3.9906) Rates = gamma Shape = 1.1102 Pinvar = 0.0931;. A maximum likelihood (ML) analysis was performed with a heuristic search (ten random taxon addition replicates) and nearest neighbour interchange (NNI) [44].

Maximum parsimony (MP) [45] was accomplished with gaps treated as missing data and all characters coded as "unordered" and equally weighted. Additionally, we clustered taxonomic units with neighbour-joining (NJ) [46] using maximum likelihood distances. Furthermore, with MrBayes [47] a Bayesian analysis (B) was carried out for tree reconstruction using a general time reversible substitution model (GTR) [48-50] with substitution rates estimated by MrBayes (nst = 6). Moreover, using ProfDist, a profile neighbour-joining (PNJ) tree [51,25] was calculated using the ITS2 specific substitution model available from the ITS2 Database. PNJ was also performed with predefined profiles (prePNJ) of all the clades given in Table 1.

For clade '*Scenedesmus*' two profiles were used for groups 'true *Scenedesmus*' (*Scenedesmus* except *S. longus*) and '*Desmodesmus*' (*Desmodesmus* and *S. longus*). We performed a sequence-structure profile neighbour-joining (strPNJ) analysis with a developmental beta version of ProfDist (available upon request). The tree reconstructing algorithm works on a 12 letter alphabet comprised of the 4 nucleotides in three structural states (unpaired, paired left, paired right). Based on a suitable substitution model [40], evolutionary distances between sequence structure pairs have been estimated by maximum likelihood. All other applied analyses were computed only on the sequence part of the sequence-structure alignment. For MP, NJ, PNJ, prePNJ and strPNJ analyses 1.000 bootstrap pseudoreplicates [52] were generated. One hundred bootstrap replicates were generated for the ML analysis. Additionally we used RAxML at the CIPRES portal to achieve 1.000 bootstraps with a substitution model estimated by RAxML [53]. All methods were additionally applied to a 50% structural consensus alignment cropped with 4SALE (data not shown). The individual steps of the analysis are displayed in a flow chart (Fig. 1).

Results
New ITS2 sequences
GenBank accession numbers for newly obtained nucleotide sequences are given in Table 1 (entries 1–4). The two ITS2 sequences of *Sphaeroplea annulina* (Roth, Agardh) strain SAG 377-1a and strain SAG 377-1e were identical

and thus only the first one was used for further analysis. According to folding with RNAstructure, ITS2 secondary structures of the three newly obtained sequences did not exhibit any branching in their helix I (Fig. 2) as it is described for the 'core Sphaeropleales', i.e. helix I was more similar to those of the CW-group and the '*Oedogonium*' clade. Helix I of *Sphaeroplea* annulina was explicitly longer (9 nucleotides) than those of the other newly obtained algae. Due to this insertion, for *Sphaeroplea*, a branching pattern was enforceable, but would have lower energy efficiency. However, the additional nucleotides are not homologous to the insertion capable of making an additional stem (Y-structure) found in the '*Scenedesmus*' and the '*Hydrodictyon*' clade (approximately 25 bases).

### ITS2 sequence and secondary structure information
ITS2 sequence lengths of all studied species ran from 202 to 262 nucleotides (nt), 235 nt on average. The GC contents of ITS2 sequences ranged from 36.84% to 59.92%, with a mean value of 52.42%. The number of base pairs (bp) varied between 64 and 89 bp and averaged 77 bp. The cropped alignment (50% structural consensus) showed that 23% of the nucleotides had at least a 50% consistency in their pairings. Compensatory base changes (CBCs) as well as hemi-CBCs (all against all) range from 0 to 16 with a mean of 6.6 CBCs (Fig. 2). Sequence pairs lacking CBCs were exclusively found within the same major clade.

Characteristics in a conserved part of alignment
In agreement with Coleman [28], the 5' side part near the tip of helix III was highly conserved including the UGGU motif [54,55,30], likewise the UGGGU motif in case of Chlorophyceae. We selected a part of the alignment at this position with adjacent columns (Fig. 2) to verify the suggested conservation. Having a closer look at this part of helix III, in our case, it showed typical sequence and structural characteristics for distinct groups. Studied species of the '*Oedogonium*' clade possess at position 3 in the selected part of the alignment an adenine and in addition at positions 3–5 paired bases. In contrast, the CW-group solely possessed three consecutively paired bases in this block, but not the adenine. A typical pattern for clades of the DO-group was a twofold motif of 3 bases: uracile, adenine and guanine at positions 7–9, which is repeated at positions 11–13. This could be a duplication, which results in a modified secondary structure. In addition, the '*core Sphaeropleales*' ('*Hydrodictyon*' clade and '*Scenedesmus*' clade) showed an adenine base change at position 6, compared to all other clades.

Phylogenetic tree information
The PAUP* calculation applying maximum Parsimony included a total of 479 characters, whereas 181 characters were constant, 214 variable characters were parsimony-
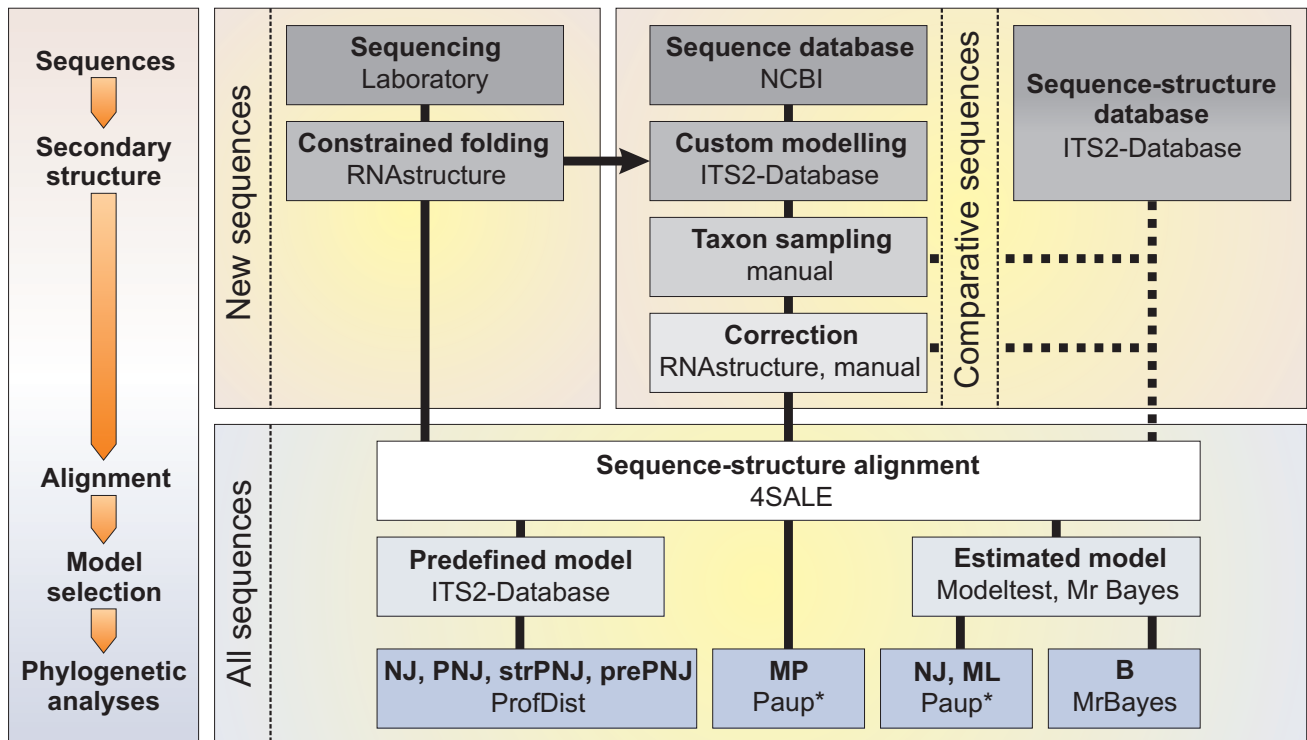
Figure 1
Flowchart of the methods applied in this study. Sequences were obtained from the laboratory and from NCBI and afterwards folded with RNAstructure [38] or custom modelling of the ITS2 Database [30-33]. An alternative way may pose to directly access sequences and structures deposed at the ITS2 Database. The sequence-structure alignment was derived by 4SALE [40]. Afterwards several phylogenetic approaches were used to calculate trees: NJ = neighbour-joining, PNJ = profile neighbour-joining, strPNJ = sequence-structure neighbour-joining, prePNJ = predefined profiles profile neighbour-joining, MP = maximum Parsimony, ML = maximum likelihood and B = Bayesian analysis.

informative compared to 84 parsimony-uninformative ones.

The resulting trees (Fig. 3 and 4, Table 2) of all performed analyses (NJ [PAUP* and ProfDist], PNJ, prePNJ, strPNJ, ML [PAUP* and RAxML], MP, B) yielded six major clades: the 'Dunaliella', the 'Hydrodictyon', the 'Oedogonium', the 'Reinhardtii', the 'Scenedesmus', and the 'Sphaeroplea' clade. All of them were separated and – except for the 'Scenedesmus' clade – highly supported by bootstrap values of 83–100%, respectively by Bayesian posterior probabilities of 0.86–1.0.

The 'Hydrodictyon' clade, the 'Scenedesmus' clade and the 'Sphaeroplea' clade form one cluster that was strongly supported by high bootstrap values of 67–96% (node "g"). The three clades composed the DO-group. The opposite cluster included the 'Dunaliella' and the 'Reinhardtii' clade, forming the CW-group. The 'Oedogonium' clade was chosen as the outgroup [56]. Both clusters (CW-group and 'Oedogonium' clade) were strongly supported by bootstrap values of 84–100% (nodes "i" and "h").

Except for the Bayesian analysis (least support for node "c"), all applied methods yielded node "e" as the weakest point within the basal (labelled) branches (Table 2), which presents the relationship between the 'Hydrodictyon' and the 'Scenedesmus' clade on the one hand and the 'Dunaliella', the 'Oedogonium', the 'Reinhardtii' and the 'Sphaeroplea' clade on the other hand. The phylogenetic tree resulting from neighbour-joining analysis by PAUP* (Fig. 3) did not support node "e" at all, but strongly supported the remaining labelled branches. The maximum likelihood analysis by PAUP* (Fig. 4) did not encourage node "e" either. Both maximum likelihood methods did not even support nodes "a" ('true *Scenedesmus*' compared to remaining clades) and "c" ('*Scenedesmus*' opposite to remaining clades). All other basal branches were supported by this method.

Varying neighbour-joining analyses by ProfDist (NJ, PNJ, prePNJ, strPNJ) supported all basal branches – except for the weakest node "e" (average support) – with very high bootstrap support values of 84–100%. The maximum Parsimony method gave average support (63 and 62%) for
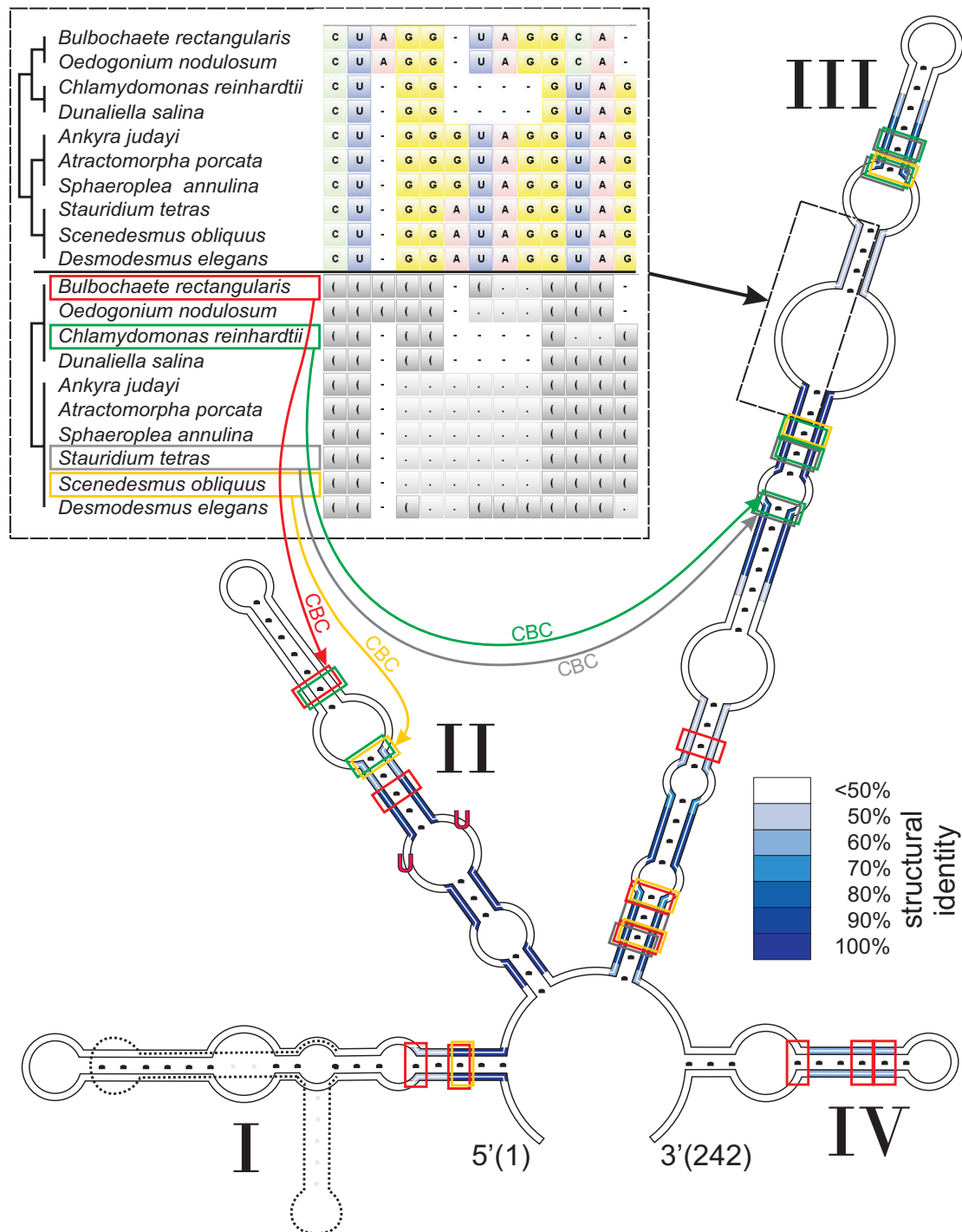
Figure 2
ITS2 structure of *Sphaeroplea annulina*, degrees of conservation and structure alignment. The structure of the internal transcribed spacer 2 of *Sphaeroplea annulina* shows the common four helices. Helix I is unbranched. Helix I of *Scenedesmus obliquus* with its branch is underlain in grey. The degree of conservation over the whole alignment is indicated in blue with different degrees of colour saturation. The structural consensus function of 4SALE [40] returns nucleotides on given percentages. In the upper left corner is the sequence-structure alignment of the conserved distal part of helix III showing a differentiation of the major clades with sequence and/or structure.

Table 2: Bootstrap support values for basal branches of all methods applied.

Software		ProfDist				PAUP*			MrBayes	RAxML
Model		ITS2				Modeltest		-	Estimated	
Analysis		NJ	PNJ	prePNJ	strPNJ	NJ	ML	MP	B	ML
Nodes	a	99	95	100¹	100	91	-	82	0.86	-
	b	96	96	100¹	96	99	93	86	1.00	98
	c	88	88	95	88	90	-	63	0.72	-
	d	100	99	100¹	100	100	92	100	1.00	96
	e	62	55	53	60	-	-	62	0.97	64
	f	100	100	100¹	100	100	99	100	1.00	100
	g	**87**	**91**	**88**	**96**	**86**	**67**	**80**	**0.98**	**93**
	h	99	99	100¹	99	100	100	100	1.00	100
	i	90	90	92	84	93	88	85	0.99	89
	j	97	98	100¹	98	93	91	91	0.99	98
	k	97	96	100¹	95	96	88	83	1.00	99
Figure		3							4	

The table supplements Fig. 3 and Fig. 4. Node "g" supports a monophyletic DO group and is printed in bold letters. Software used: ProfDist and PAUP*. Models of substitution: ITS2 = GTR with ITS2 substitution matrix, Modeltest: TVM+I+G with estimated parameters. Phylogenetic analysis: NJ = neighbour-joining, PNJ = profile neighbour-joining, prePNJ = profile neighbour-joining with predefined profiles, strPNJ = sequence-structure profile neighbour-joining, ML = maximum likelihood, B = Bayesian analysis (posterior probabilities), MP = maximum Parsimony. ¹Predefined profiles for profile neighbour-joining.

node "c" and "e" and high bootstrap values (80–100%) for the remaining basal clades. The Bayesian analysis offered posterior probabilities of 0.72 for node "c" and 0.86–1.0 for the remaining basal nodes. For further sister group relations see Fig. 3 and 4.

In comparison, the topology of the phylogenetic tree based on the 50% cropped alignment did not change, but the bootstrap support values were lower in all cases (data not shown).

Discussion

The internal transcribed spacer 2 (ITS2) is required in ribosome biogenesis [57-59] and its gradual removal from mature rRNA is driven by its specific secondary structure [60,59].

Using three newly obtained ITS2 sequences from *Ankyra judayi*, *Atractomorpha porcata* and *Sphaeroplea annulina* (Sphaeropleaceae) in this study we aimed to pursue two consecutive questions concerning the phylogenetic relationships within Chlorophyceae. (1) What is the phylogenetic position of the newly sequenced algae relative to the 'core Sphaeropleales' and could the biflagellate DO-group be regarded as monophyletic? (2) How does the secondary structure of the new ITS2 sequences look like and is an autapomorphic feature of the secondary structure associated with the monophyletic DO-group?

Considering the question (1) Buchheim et al. [6] and Wolf et al. [23] approached the problem with 18S + 26S

rDNA and 18S rDNA data, but the relationship between the 'core Sphaeropleales' and the Sphaeropleaceae remained unclear. However, in their studies, *Ankyra*, *Atractomorpha* and *Sphaeroplea* clustered in a monophyletic clade named Sphaeropleaceae. We confirm this '*Sphaeroplea*' clade with all three genera being strongly separated from other clades. As a result of a Bayesian analysis on a combined 18S and 26S rDNA dataset Shoup and Lewis [61] also found the Sphaeropleaceae as the most basal clade within the Sphaeropleales, but again the analysis lacked a strong backing. Beside these difficulties the 'core Sphaeropleales' were already shown to be monophyletic with high certainty [6,25,62,61,23].

The DO-group (Sphaeropleales including the '*Sphaeroplea*' clade) as emended by Deason et al. [10], for which the directly opposed basal body orientation and basal body connection features are verified [63-65], is now strongly supported by molecular phylogenetic analyses. There was already evidence of an extended DO-group [6,66,67], however, for some groups ultrastructural results are still lacking, and even though the collective basal body orientation and connection imply a monophyletic DO-group, until now no molecular phylogenetic analysis could show this with solid support [6,62,24,23]. We demonstrate for the first time with robust support values for the equivocal nodes that the 'core Sphaeropleales', the '*Sphaeroplea*' clade, and the Sphaeropleales are monophyletic.

Regarding question (2), for all structures of the '*Hydrodictyon*' and the '*Scenedesmus*' clade, helix I shows the typical
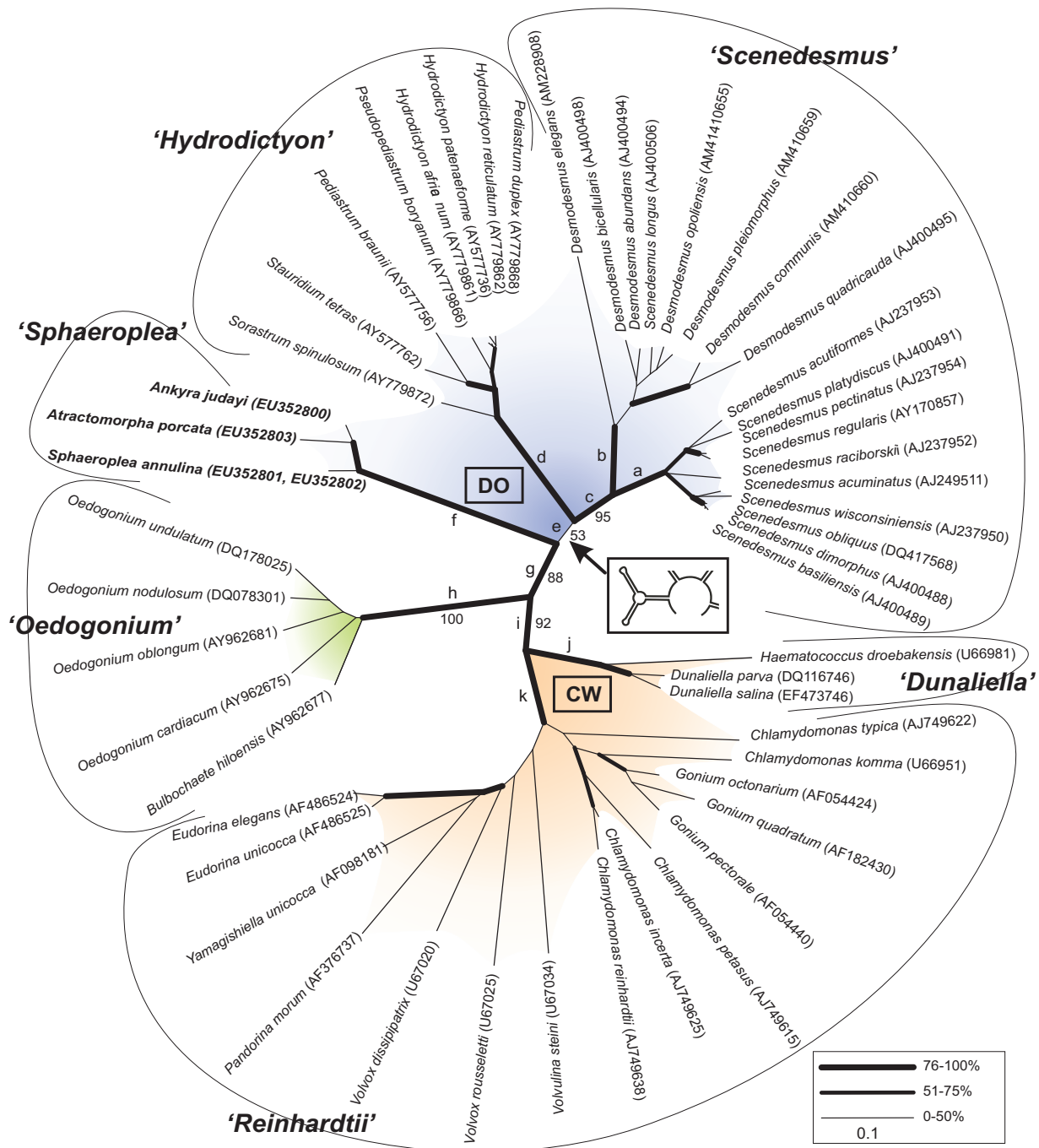
Figure 3
Neighbour-joining phylogeny of the Chlorophyceae based on comparison of ITS2 rRNA sequences and structures. The tree is unrooted, but the '*Oedogonium*' clade is most likely appropriate as outgroup [56]. Sequences of the '*Sphaeroplea*' clade were sequenced for this study and shown in bold letters. The phylogenetic tree is calculated by neighbour-joining with PAUP* [46,43] for an alignment with 52 taxa and 479 characters. The substitution model was set to TVM+I+G with parameters estimated by Modeltest [42]. Bootstrap values of basal branches are given for profile neighbour-joining with predefined profiles (ProfDist with ITS2 substitution model) [51,31]. Branch thickness is dependant of Bootstrap values calculated with four distance methods: neighbour-joining (PAUP*), neighbour-joining, complete profile neighbour-joining and sequence-structure profile neighbour-joining (all three ProfDist with ITS2 substitution model).
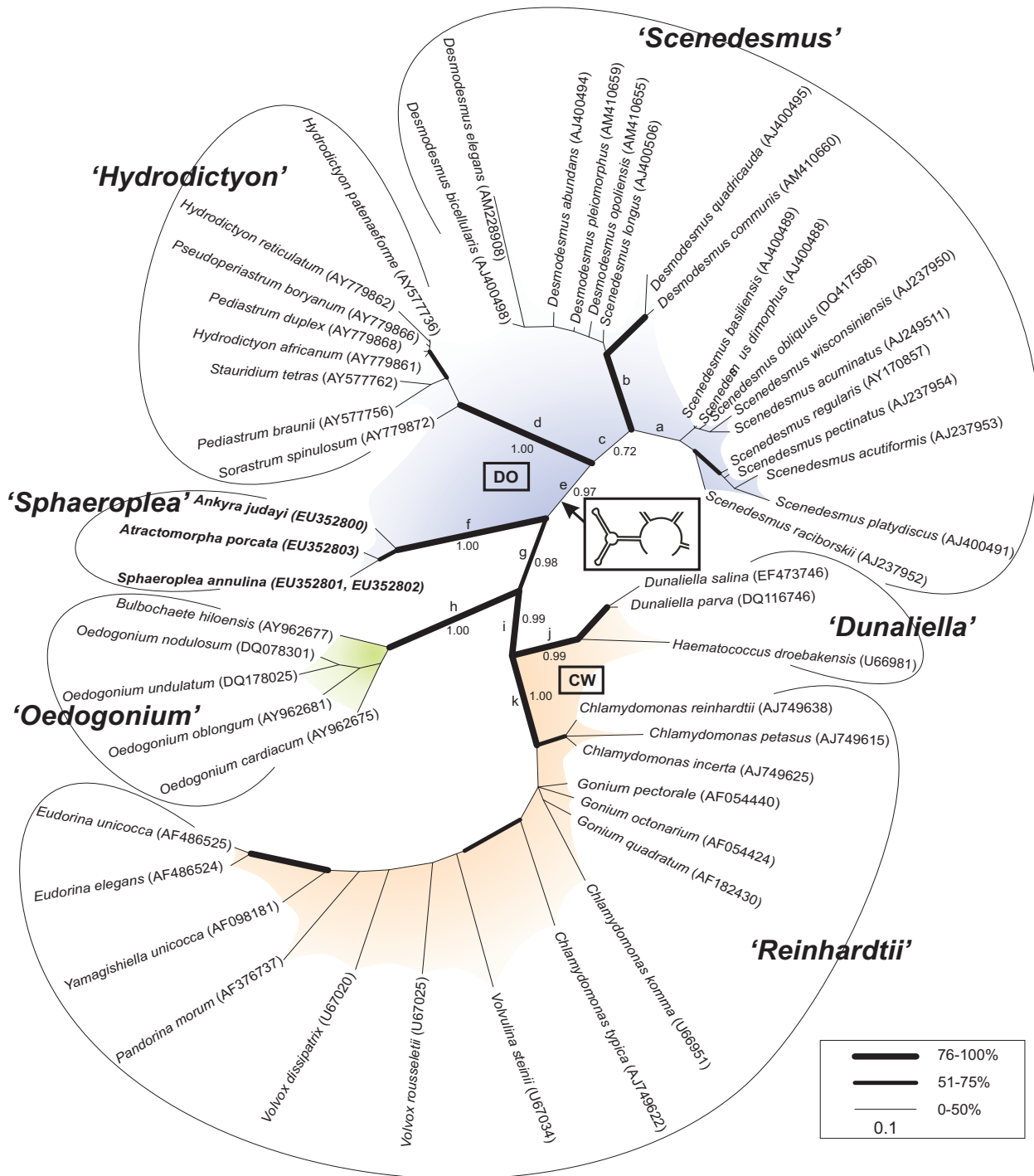
Figure 4
Phylogeny of chlorophyte ITS2 sequences and structures based on distances of a Bayesian analysis. The alignment contained 52 taxa and 479 characters. The suggested outgroup is the '*Oedogonium*' clade [56]. Sequenced species are shown in bold ('*Sphaeroplea*' clade). Substitution models and tree distances were calculated with MrBayes [47]. Posterior probabilities are shown for basal branches. Branch thickness is dependant of Bootstrap values calculated with maximum likelihood (PAUP* with TVM+I+G, RAxML) [42,53,43] and maximum Parsimony (PAUP*) (see legend). Resulting parameter of performing MP are L = 1231, CI = 0.4427, HI = 0.5573, RI = 0.7264, RC = 0.3216.

branching (Y-structure). Initially, An et al. [68] proposed a secondary structure model with an unbranched helix I for ITS2 sequences of 'Scenedesmus' clade members. Thereafter, van Hannen et al. [34] updated the model by folding the nucleotide sequences based upon minimum free energy and found a branched helix I as the most energetically stable option. The branching is result of an insertion of approximately 25 nucleotides capable of folding as an individual stem within the 5' end of the first helix. However, ITS2 sequence and secondary structure information of further 'core Sphaeropleales' members, e.g. the 'Ankistrodesmus' clade and the 'Bracteacoccus' clade, lacks hitherto. In contrast, the Y-structure is absent within the 'Sphaeroplea' clade and any other investigated group so far. Thus this feature is – contrary to our expectation – not an autapomorphic character for the biflagellate DO-group as a whole but for the 'core Sphaeropleales'.

Regarding future work, the resolution among the main clades of Chlorophyceae was statistically poorly supported in previous studies [68,15,6,23]. Pröschold and Leliaert [24] reviewed the systematics of green algae by applying a polyphasic approach, but did not yield a clear resolution regarding a sister taxon to the Sphaeropleales. Since they are not yet available, ITS2 sequences of chaetopeltidalean and chaetophoralean taxa could not be included in the present study and therefore the phylogenetic relationships between the main Chlorophyceae clades remain open. We recommend involving sequence and secondary structure information of chaetopeltidalean and chaetophoralean ITS2 sequences in future studies to find out if the monophyletic biflagellate DO-group could be further extended to a general monophyletic DO-group containing quadri- and biflagellate taxa. A genome-wide approach indicates that Sphaeropleales and Chlamydomonadales are sister taxa, however only a few organisms are included in this study [56]. An additional uprising question is when the Y has evolved within the 'core Sphaeropleales'. This could be resolved by inclusion of other members (e.g. *Bracteacoccus*) in further studies.

The two major reasons contributing to the robust results presented here are the change of the phylogenetic marker and the inclusion of secondary structure information. In contrast to previous phylogenetic work concerning Chlorophyceae, this study is based on the ITS2, which offers a resolution power for relationships from the level of subspecies up to the order level, because of their variable sequence but conserved secondary structure [26,30-33]. Hitherto commonly used markers in contrast are a lot more restricted. Using 4SALE [40] with implemented structure consideration, we could achieve for the first time a global simultaneously generated sequence-structure alignment (c.f. Fig. 1) yielding specific sequence and structural features distinguishing different algae lineages (c.f. Fig. 2).

Conclusion
In summary, the powerful combination of the ITS2 rRNA gene marker plus a multiple global alignment based synchronously on sequence and secondary structure yielded high bootstrap support values for almost all nodes of the computed phylogenetic trees. Thus, the relationship of Sphaeropleaceae is here resolved, being a part of the Sphaeropleales representing the monophyletic biflagellate DO-group. Furthermore, we could elucidate a branched helix I of ITS2 as an autapomorphic feature within the DO-group. This feature could be found only in the 'Hydrodictyon' and the 'Scenedesmus' clade. Our results corroborate the presented methodological pipeline, the fundamental relevance of secondary structure consideration, as well as the elevated power and suitability of ITS2 in phylogenetics. For a methodological improvement it is suitable to ameliorate the alignment algorithm in further considering horizontal dependencies of paired nucleotides, and moreover in future ITS2 studies it is suggested to include sequence and secondary structure information of hitherto not regarded taxa to resolve the chlorophycean phylogeny.

Authors' contributions
MW designed the study. FF determined the new sequences in the laboratory. BR implemented the strPNJ within ProfDist. TS and AK performed sequence analyses, structure prediction and phylogenetic analyses. TM developed the ITS2 sequence-structure substitution model and the ITS2 sequence-structure scoring matrix. TS, AK and MW drafted the manuscript. All authors contributed to writing the paper, read the final manuscript and approved it.

Acknowledgements

References
1. Agardh CA: *Systema algarum XXV* Lund, Sweden: Soc Physiogr; 1824.
2. Bold HC, Wynne MJ: *Introduction to the algae: structure and reproduction* Englewood Cliffs, NJ: Prentice Hall; 1985.
3. Booton GC, Floyd GL, Fuerst PA: **Polyphyly of tetrasporalean green algae inferred from nuclear small-subunit ribosomal DNA.** *J Phycol* 1998, **34**:306-311.
4. Booton GC, Floyd GL, Fuerst PA: **Origins and affinities of the filamentous green algal orders Chaetophorales and Oedogoniales based on 18S rRNA gene sequences.** *J Phycol* 1998, **34**:312-318.
5. Buchheim MA, Hoffman LR: **Ultrastructure of male gametes of *Sphaeroplea robusta* (Chlorophyceae).** *J Phycol* 1986, **22**:176-185.

6.   Buchheim MA, Michalopulos EA, Buchheim JA: **Phylogeny of the Chlorophyceae with special reference to the Sphaeropleales: a study of 18S and 26S rDNA data.** *J Phycol* 2001, **37:**819-835.
7.   Cáceres EJ, Robinson DG: **Ultrastructural studies on *Sphaeroplea annulina* (Chlorophyceae). Vegetative structure and mitosis.** *J Phycol* 1980, **16:**313-320.
8.   Cáceres EJ, Robinson DG: **Ultrastructural studies on *Sphaeroplea annulina* (Chlorophyceae). II. Spermatogenesis and male gamete structure.** *J Phycol* 1981, **17:**173-180.
9.   Chapman R, Buchheim MA, Delwiche CF, Friedl T, Huss VA, Karol KG, Lewis LA, Manhart J, McCourt RM, Olsen JL, Waters DA: **Molecular systematics of the green algae.** In *Molecular systematics of plants II: DNA sequencing* Edited by: Soltis DE, Soltis PS, Doyle J. Boston, MA: Kluwer Acad. Publ; 1998:508-540.
10.   Deason TR, Silva P, Watanabe S, Floyd GL: **Taxonomic status of the species of the green algal genus *Neochloris* .** *Plant Sys Evol* 1991, **177:**213-219.
11.   Engler A, Prantl K: *Die natürlichen Pflanzenfamilien I* Leipzig, Germany: Verlag von Wilhelm Engelmann; 1897.
12.   Hoffman LR: *Atractomorpha echinata* **gen. et ap.-nov., a new anisogamous member of the Sphaeropleaceae (Chlorophyceae).** *J Phycol* 1983, **19:**76-86.
13.   Hoffman LR: *Atractomorpha porcata* **ap. nov., a new member of the Sphaeropleaceae (Chlorophyceae) from California.** *J Phycol* 1984, **20:**225-236.
14.   Kützing FT: *Species Algarum* Leipzig, Germany: Brockhaus Verlag; 1849.
15.   Lewis LA, McCourt RM: **Green algae and the origin of land plants.** *Am J Bot* 2004, **91:**1535-1556.
16.   Mattox KR, Stewart KD: **Classification of the green algae: a concept based on comparative cytology.** In *Systematics of the green algae* Edited by: Irvine DEG, John DM. London, UK: Academic Press; 1984:29-42.
17.   Pascher A: **Systematische Übersicht über die mit Flagellaten in Zusammenhang stehenden Algenreihen und Versuch einer Einreihung dieser Algenstämme in die Stämme des Pflanzenreiches.** *Beihefte Bot Centralbl* 1931, **48:**317-332.
18.   Pascher A: **Über geißlbewegliche Eier, mehrköpfige Schwärmer und vollständigen Schwärmverlust bei *Sphaeroplea*.** *Beihefte Bot Centralbl* 1939, **59:**188-213.
19.   Rieth A: **Über die vegetative Vermehrung bei *Sphaeroplea wilmani* Fritsch et Rich.** *Flora* 1952, **139:**28-38.
20.   Rieth A: **Zur Kenntnis der Gattung *Sphaeroplea, Sphaeroplea caubrica* Fritsch.** *Flora* 1953, **140:**130-139.
21.   Stewart KD, Mattox KR: **Comparative cytology, evolution and classification of the green algae with some consideration of other organisms with chlorophylls a and b.** *Botanical Rev* 1975, **41:**104-135.
22.   West LW, Fritsch FE: *A treatise on the british freshwater algae, Reviewed edition* Cambridge, UK: Cambridge University Press; 1927.
23.   Wolf M, Buchheim MA, Hegewald E, Krienitz L, Hepperle D: **Phylogenetic position of the Sphaeropleaceae (Chlorophyta).** *Plant Sys Evol* 2002, **230:**161-171.
24.   Pröschold T, Leliaert F: **Systematics of the green algae: conflict of classic and modern approaches.** In *Unravelling the algae: the past, present, and future of algal systematics* Edited by: Brodie J, Lewis J. London, UK: CRC Press; 2007:123-135.
25.   Müller T, Rahmann S, Dandekar T, Wolf M: **Accurate and robust phylogeny estimation based on profile distances: a study of the Chlorophyceae (Chlorophyta).** *BMC Evol Biol* 2004, **4:**20.
26.   Coleman AW: **ITS2 is a double-edged tool for eukaryote evolutionary comparisons.** *TIG* 2003, **19:**370-375.
27.   Coleman AW, Vacquier VD: **Exploring the phylogenetic utility of ITS sequences for animals: a test case for Abalone (*Haliotis*).** *J Mol Evol* 2002, **54:**246-257.
28.   Coleman AW: **Pan-eukaryote ITS2 homologies revealed by RNA secondary structure.** *Nucl Acids Res* 2007, **35:**3322-3329.
29.   Young I, Coleman AW: **The advantages of the ITS2 region of the nuclear rDNA cistron for analysis of phylogenetic relationships of insects: a *Drosophila* example.** *Mol Phylogenet Evol* 2004, **30:**236-242.
30.   Schultz J, Maisel S, Gerlach D, Müller T, Wolf M: **A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota.** *RNA* 2005, **11:**361-364.
31.   Schultz J, Müller T, Achtziger M, Seibel P, Dandekar T, Wolf M: **The internal transcribed spacer 2 database – a web server for**

32.   **(not only) low level phylogenetic analyses.** *Nucl Acids Res* 2006, **34(Suppl 2):**W704-707.
32.   Selig C, Wolf M, Müller T, Dandekar T, Schultz J: **The ITS2 Database II: homology modelling RNA structure for molecular systematics.** *Nucl Acids Res* 2008, **36:**D377-D380.
33.   Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T: **Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures.** *RNA* 2005, **11:**1616-1623.
34.   van Hannen EJ, Fink P, Lürling M: **A revised secondary structure model for the internal transcribed spacer 2 of the green algae *Scenedesmus* and *Desmodesmus* and its implication for the phylogeny of these algae.** *Eur J Phycol* 2002, **37:**203-208.
35.   Buchheim MA, Buchheim JA, Carlson T, Braband A, Hepperle D, Krienitz L, Wolf M, Hegewald E: **Phylogeny of the Hydrodictyaceae (Chlorophyceae): inferences from rDNA data.** *J Phycol* 2005, **41:**1039-1054.
36.   Grajales A, Aguilar C, Sanchez J: **Phylogenetic reconstruction using secondary structures of internal transcribed spacer 2 (ITS2, rDNA): finding the molecular and morphological gap in Caribbean gorgonian corals.** *BMC Evol Biol* 2007, **7:**90.
37.   White TJ, Bruns T, Lee S, Taylor J: **Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics.** In *PCR protocols: a guide to methods and applications* Edited by: Innis MA, Gelfand DH, Sninsky JJ, White TJ. San Diego, CA: Academic Press; 1990:315-322.
38.   Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.** *Proc Natl Acad Sci USA* 2004, **101:**7287-7292.
39.   Byun Y, Han K: **PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures.** *Nucl Acids Res* 2006, **34:**416-422.
40.   Seibel P, Müller T, Dandekar T, Schultz J, Wolf M: **4 SALE – A tool for synchronous RNA sequence and secondary structure alignment and editing.** *BMC Bioinformatics* 2006, **7:**498.
41.   Thompson JD, Higgins DG, Gibson TJ: **ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucl Acids Res* 1994, **22:**4673-4680.
42.   Posada D, Crandall KA: **Modeltest: testing the model of DNA substitution.** *Bioinformatics* 1998, **14(9):**817-818.
43.   Swofford DL: *PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0b10* Sunderland, MA: Sinauer Associates; 2002.
44.   Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.
45.   Camin JH, Sokal RR: **A method for deducing branching sequences in phylogeny.** *Evolution* 1965:311-326.
46.   Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4(4):**406-425.
47.   Huelsenbeck JP, Ronquist F: **MrBayes: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17:**754-755.
48.   Lanave C, Preparata G, Sacone C, Serio G: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20:**86-93.
49.   Rodriguez F, Oliver JL, Marin A, Medina J: **The general stochastic model of nucleotide substitution.** *J Theor Biol* 1990, **142:**485-501.
50.   Tavaré S: **Some probabilistic and statistical problems in the analysis of DNA sequences.** In *Some mathematical questions in biology: DNA sequence analysis* Edited by: Lipman D, Miura RM. Providence, RI: American Mathematical Society; 1986:57-86.
51.   Friedrich J, Dandekar T, Wolf M, Müller T: **ProfDist: a tool for the construction of large phylogenetic trees based on profile distances.** *Bioinformatics* 2005, **21(9):**2108-2109.
52.   Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evolution* 1985:783-791.
53.   Stamatakis A, Hoover P, Rougemont J: **A rapid bootstrap algorithm for the RAxML web-servers.** *Systematic Biology* 2008 in press.
54.   Liu J, Schardl CL: **A conserved sequence in internal transcribed spacer 1 of plant nuclear rRNA genes.** *Plant Molecular Biology* 1994, **26:**775-778.
55.   Mai JC, Coleman AW: **The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants.** *J Mol Evol* 1997, **44:**258-271.

56. Turmel M, Brouard JS, Gagnon C, Otis C, Lemieux C: **Deep division in the Chlorophyceae (Chlorophyta) revealed by chloroplast phylogenomic analyses.** *J Phycol* 2008, **44:**739-756.

57. Côté CA, Peculis BA: **Role of the ITS2-proximal stem and evidence for indirect recognition of processing sites in pre-rRNA processing in yeast.** *Nucleic Acids Research* 2001, **29:**2106-2116.

58. Peculis BA, Greer CL: **The structure of the ITS2-proximal stem is required for pre-rRNA processing in yeast RNA.** *RNA* 1998, **4:**1610-1622.

59. Nues RW van, Rientjes JM, Morré SA, Mollee E, Planta RJ, Venema J, Raué Hendrik A: **Evolutionarily conserved structural elements are critical for processing of internal transcribed spacer 2 from *Saccharomyces cerevisiae* precursor ribosomal RNA.** *J Mol Biol* 1995, **250:**24-36.

60. Sande CA van der, Kwa MR, van Nues RW, van Heerikhuizen H, Raué Hendrik A, Planta RJ: **Functional analysis of internal transcribed spacer 2 of *Saccharomyces cerevisiae* ribosomal DNA.** *J Mol Biol* 1992, **223:**899-910.

61. Shoup S, Lewis LA: **Polyphyletic origin of parallel basal bodies in swimming cells of chlorophycean green algae (Chlorophyta).** *J Phycol* 2003, **39:**789-796.

62. Hegewald E, Hepperle D, Wolf M, Krienitz L: **Phylogenetic placement of *Chlorotetraedron incus, C. polymorphum* and *Polyedriopsis spinulosa* (Neochloridaceae, Chlorophyta).** *Phycologia* 2001, **40:**399-402.

63. Melkonian M: **Structural and evolutionary aspects of the flagellar apparatus in green algae and land plants.** *Taxon* 1982, **31:**255-265.

64. Melkonian M, Surek B: **Phylogeny of the Chlorophyta: congruence between ultrastructural and molecular evidence.** *Bull Soc Zool Fr* 1995, **120:**191-208.

65. Wilcox LW, Floyd GL: **Ultrastructure of the gamete of *Pediastrum duplex* (Chlorophyceae).** *J Phycol* 1988, **24:**140-146.

66. Lewis LA: **Diversity and phylogenetic placement of *Bracteacoccus tereg* (Chlorophyceae, Chlorophyta) based on 18S ribosomal RNA gene sequence data.** *J Phycol* 1997, **33:**279.

67. Krienitz L, Hegewald E, Hepperle D, Wolf M: **The systematics of coccoid green algae: 18S rRNA gene sequence data versus morphology.** *Biologia* 2003, **58:**437-446.

68. An SS, Friedl T, Hegewald E: **Phylogenetic relationships of *Scenedesmus* and *Scenedesmus*-like coccoid green algae as inferred from ITS-2 rDNA sequence comparisons.** *Plant biol* 1999, **1:**418-428.

69. **AlgaeBase** [http://www.algaebase.org]

70. Pröschold T, Marin B, Schlösser UG, Melkonian M: **Molecular phylogeny and taxonomic revision of *Chlamydomonas* (Chlorophyta). I. Emendation of *Chlamydomonas* Ehrenberg and *Chloromonas* Gobi, and description of *Oogamochlamys* gen. nov. and *Lobochlamys* gen. nov.** *Protist* 2001, **152:**265-300.

Chapter 12.

Distinguishing species in *Paramacrobiotus* (Tardigrada) via compensatory base change analysis of internal transcribed spacer 2 secondary structures, with the description of three new species

Running Head: DISTINGUISHING SPECIES IN *PARAMACROBIOTUS*

Using compensatory base change analysis of internal transcribed spacer 2 secondary

structures to identify three new species in *Paramacrobiotus* (Tardigrada)

Ralph O. Schill[1][§],

Frank Förster[2][§],

Thomas Dandekar[2],

Matthias Wolf[2]*

[1]Universität Stuttgart, Biological Institute, Zoology, Pfaffenwaldring 57, D-70569

Stuttgart, Germany,

[2]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, D-

97074 Würzburg, Germany,

[§] Authors contributed equally to this work

*Corresponding author: matthias.wolf@biozentrum.uni-wuerzburg.de

Abstract

Species of tardigrades within the genus *Paramacrobiotus* (Tardigrada) could be distinguished via an analysis of internal transcribed spacer 2 (ITS2) secondary structures. Sequences of four undescribed *Paramacrobiotus* "*richtersi* groups" (formely *Paramacrobiotus* "*richtersi* group" 1, *Paramacrobiotus* "*richtersi* group" 2, *Paramacrobiotus* "*richtersi* group" 3, and *Paramacrobiotus* "*richtersi* group" 4) and *Paramacrobiotus richtersi* from different continents were determined, annotated and their secondary structures predicted. A tree based on a combined sequence-structure alignment was reconstructed by Neighbor-Joining. The topology obtained is consistent with a tree based on a distance matrix of compensatory base changes (CBCs) between all ITS2 sequence-structure pairs in the global multiple alignment. The CBC analysis, together with 18S rDNA sequences, physiological, biochemical and biophysical data identified three new species, morphologically indistinguishable from *P. richtersi*. The species *Paramacrobiotus palaui* sp. nov., *Paramacrobiotus kenianus* sp. nov., and *Paramacrobiotus fairbanki* sp. nov. are new to science.

Using compensatory base change analysis of internal transcribed spacer 2 secondary structures to identify three new species in *Paramacrobiotus* (Tardigrada)

Introduction

Tardigrades have been discovered (Goeze 1773) and described (Spallanzani 1776) in the 18th century. The phyla Tardigrada refers to the animal's way of movement (Lat. *tardi* - slow, *grado* - walker) and these aquatic and semi-terrestrial metazoans are found in a variety of habitats within marine, freshwater and terrestrial ecosystems (Marcus 1929; Nelson 2002). Over the last decades, the number of described species of tardigrades has considerably increased meanwhile to more than 1,000 species, with new ones being discovered every year. They are subdivided into 3 classes (Eutardigrada, Heterotardigrada, Mesotardigrada), 4 orders, 21 families, 104 genera (Guidetti and Bertolani 2005). Tardigrades have a bilaterally symmetrical and segmented body, covered with a chitinous cuticle and with four pairs of lobopodous legs terminating in claws and/or in digits, and range in size from 0.1 mm to ca. 1.2 mm.

The general problem, which is frequently encountered in taxonomic studies in small invertebrates (including tardigrades), is that few morphological characters are available, depending on class, family and genus (Ramazzotti and Maucci 1983). In general, heterotardigrades, especially marine species, have a higher morphological diversity compared with the more conservative limno-terrestrial forms of eutardigrada (see Kristensen 1987; Jørgensen 2000; Guidetti and Bertolani 2001; Jørgensen and Kristensen 2004). As consequence, phylogenetic analysis of eutardigrades based only on morphological characters are difficult. Therefore phylogenetic relationships within and between the tardigrade orders, families and genera remain mostly unresolved, highlighting the increased need for molecular markers. Such molecular phylogenetic studies using protein-coding nuclear genes support the

classification currently adopted at order and family level (Garey et al. 1999; Jørgensen and Kristensen 2004; Regier et al. 2004; Guidetti et al. 2005; Nichols et al. 2006).

The family Macrobiotidae is a large and quite complex eutardigrade family, which includes 38% of species and 33% of genera of the order Parachela (Guidetti et al. 2005). Within this family, recently the genus *Paramacrobiotus* was erected using morphological characters and gene sequences (Guidetti et al. 2009). Within the genus the authors found several cryptic species from different places around the world, currently numbered from 1 to 3, which can not be clearly separated by morphological or common molecular markers like COI or 18S alone (Guidetti et al. 2009). This cryptic species have been already used for several physiological and biochemical studies and formally described as *Paramacrobiotus* "*richtersi* group*"*, consecutively numbered (Hengherr et al. 2008; Hengherr et al. 2009a; Hengherr et al. 2009b).

To get a deeper insight into the genus *Paramacrobiotus* we used for the first time in tardigrades, the sequence-structure information of the internal transcribed spacer 2 (ITS2) as marker to distinguish between tardigrade species. ITS2 sequences have been used already before to study sequence diversity in tardigrades over a large geographical area (Jørgensen et al. 2007; Møbjerg et al. 2007), but without correlating the ability of closely related taxa to interbreed with compensatory base changes (CBCs) that are observable in the secondary structure of the ITS2 region (Coleman 2000; Coleman and Vacquier 2002; Coleman 2003; Coleman 2007; Müller et al. 2007; Coleman 2009; Schultz and Wolf 2009). CBCs occur in the paired regions of a primary RNA transcript when both nucleotides of a paired site mutate while the pairing itself is maintained (e.g., G-C mutates to A-U; Gutell et al. 1994). Cross-fertilization experiments showed that taxa differing by even a single CBC usually cannot interbreed and should therefore be considered as different species (c.f. Schmitt et al. 2005). According to Coleman and Vacquier (2002, p. 255), "in all eukaryote groups where a broad

array of species has been compared for both ITS2 sequence secondary structure and tested for any vestige of interspecies sexual compatibility, an interesting correlation has been found. When sufficient evolutionary distance has accumulated to produce even one CBC in the relatively conserved pairing positions of the ITS2 transcript secondary structure, taxa differing by the CBC are observed experimentally to be totally incapable of intercrossing". This hypothesis was then subjected to large-scale testing by Müller, et al. (2007) for mainly fungi and plants. Meanwhile, the CBC criterion was also used to distinguish species within animals (Wolf et al. 2007). Müller, et al. (2007) showed that indeed there is a possibility to discriminate between two species by the existence of a single CBC between an ITS2 sequence-structure pair with a confidence of 93%, and therefore the method should be quite useful for cryptic species in tardigrades, too.

Material and Methods

Taxon Sampling

Seven eutardigrade species (Tab. 1) were used to investigate the internal transcribed spacer 2 secondary structures by compensatory base change analysis for identification of cryptic species (Müller et al. 2007). The species *Macrobiotus sapiens* Binda and Pilato, 1984 were collected in Rovinj, Croatia. *Paramacrobiotus richtersi* Murray, 1911 was collected in Tübingen, Germany. *Paramacrobiotus tonollii* Ramazzotti, 1956 was from Eugene, Oregon, USA. Furthermore we used three cryptic species formally described as *Paramacrobiotus* "*richtersi* group" 1 (from Naivasha, Kenya), *Paramacrobiotus* "*richtersi* group" 2 (from Nakuru, Kenya), and *Paramacrobiotus* "*richtersi* group" 3 (from Fairbanks, Alaska, USA) (Guidetti et al. 2009). Another tardigrade species was recently discovered in Koror, Palau and formally described as *Paramacrobiotus* "*richtersi* group" 4 (Schill 2007). All species were cultured with bdelloid rotifers, and raised on green algae, as described by Schill (2007).

ITS 2 gene amplification and sequencing

Genomic DNA was extracted from tardigrades according to Schill (2007) using the NucleoSpin Tissue method (Macherey-Nagel, Düren, Germany). The lysis was achieved by incubation of the specimens in a proteinase K/SDS solution at 56 °C for 12 h and subsequently at 70 °C for 10 min. During the lysis the whole animals were ground with a pestle to release the DNA. Subsequently the DNA was bound to a silica membrane in a NucleoSpin Tissue column. Contamination was removed by several steps of washing. Finally, the DNA was eluted twice from the membrane with pure water. For amplification of the ITS 2 rDNA gene we used the primers ITS3 5'-GCATCGATGAAGAACGCAGC-3' and ITS4 5'-TCCTCCGCTTATTGATATGC-3' published by White et al. (1990). Each PCR was carried out in a final volume of 25 µl: 80 ng DNA, 0.5 unit Taq DNA Polymerase (Genaxxon, Biberach, Germany) with 10x reaction buffer supplemented to a final concentration of 1.5 mM MgCl2, 0.25 mM dNTP, 1 µM oligonucleotide primers each. The ITS 2 region was amplified with an initial denaturation step of 3 min at 95 °C, followed by 30 cycles of 30 seconds at 95 °C, 30 seconds at 55 °C and 1.5 min at 72 °C and a final cycle of 10 min at 72 °C. All PCR products were directly sequenced at the company AGOWA (Berlin, Germany) with the PCR forward primer, their identity verified using the Basic Local Alignment Search Tool (1997), and submitted to the NCBI GenBank. The GenBank accession numbers and the lengths of the ITS2s are shown in Tab. 1.

Secondary Structure Prediction

The prediction of the secondary structure for all retained ITS2 sequences was performed according to Schultz and Wolf (2009): The ITS2 sequences were delimited and cropped with the HMM-based annotation tool present at the ITS2 Database (Keller et al. 2009; E-value < 0.001, metazoan HMMs). The secondary structure of the ITS2 of *P.*

"*richtersi* group" 2 from Nakuru, Kenya was predicted with RNAstructure 4.6 (Mathews et al. 2004) and ported to Vienna format with CBCanalyzer 1.0.3 (Wolf et al. 2005b). The structures of the remaining sequences were predicted by homology modelling at the ITS2 Database (Schultz et al. 2005; Wolf et al. 2005a; Schultz et al. 2006; Selig et al. 2008; Koetschan et al. 2010) with the *P.* "*richtersi* group" 2 structure as template. To obtain complete ITS2 secondary structures with four helices a threshold of 66% for the helix transfer (identity matrix) was used.

Alignment and Phylogenetic Analysis

Sequences and secondary structures were automatically and synchronously aligned with 4SALE (version 1.5) using an ITS2 specific scoring matrix (Seibel et al. 2006; Seibel et al. 2008). According to Keller et al. (2010) we determined maximum likelihood evolutionary distances between organisms simultaneously on sequences and secondary structures using an ITS2-sequence-structure-specific general time reversible substitution model (Seibel et al. 2006). A tree was calculated by Neighbour-Joining as implemented in ProfDistS 0.98 (Friedrich et al. 2005; Wolf et al. 2008). Bootstrap support was estimated based on 1.000 pseudo replicates (Felsenstein 1985). The alignment and the ProfDistS-tree are available for download at TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S10329) or the supplement section of the ITS2 Database (http://its2.bioapps.biozentrum.uni-wuerzburg.de/cgi-bin/index.pl?supplements).

Compensatory base change analysis

For a compensatory base change (CBC) analysis we utilized 4SALE (Seibel et al. 2006; Seibel et al. 2008) to detect CBCs between sequence-structure pairs of the alignment. BIONJ (Gascuel 1997) as implemented in the CBCanalyzer (Wolf et al. 2005b) was used to

calculate a so-called CBCtree based on the CBC matrix as derived from the global multiple sequence-structure alignment. The matrix is available for download at TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S10329) or the supplement section of the ITS2 Database (http://its2.bioapps.biozentrum.uni-wuerzburg.de/cgi-bin/index.pl?supplements).

Results

We obtained the complete ITS2 sequence from one strain each of *M. sapiens*, *P. tonollii*, *P. richtersi* and of the four unidentified *P.* "*richtersi*" groups to make the CBC analysis (Fig. 1a;b). Based on the folding patterns, *M. sapiens* and *P. tonollii* can be clearly separated as two discrete species, which we used as an outgroup in the Neighbor-joining tree (Fig. 2a). The NJ CBC tree shows a high level of resolution for the *P. richtersi* and the *P.* "*richtersi*" groups, which is given with bootstrap support values shown above the branches. The number of CBCs between the specimens ranges from zero to seven and are distributed over the whole structure of the ITS2 molecule (Fig. 1a; b, Fig. 2b; Tab. 2). *P.* "*richtersi* group" 1 and *P.* "*richtersi* group" 2 from Africa clustered together and show three CBCs compared to the cluster *P.* "*richtersi* group" 3 from the USA and *P. richtersi* from Germany. Compared with *P.* "*richtersi* group" 4 from Micronesia we detected five CBCs. Most CBCs are between the cluster *P.* "*richtersi* group" 3 and *P. richtersi* and the *P.* "*richtersi* group" 4. However, *P.* "*richtersi* group" 1 and *P.* "*richtersi* group" 2 show no CBC with respect to one another. Based on the existence of CBCs between an ITS2 sequence-structure pair we identified two new species besides *P. richtersi* and together with already available biochemical and biophysical, 18S rDNA, and COI sequence data three new species. It should be mentioned, that the CBC criterion appears to generally be a sufficient condition to establish species status, but not a necessary one: if there is a CBC then there are two species;

if there is no CBC like between *P. "richtersi* group" 3 and *P. richtersi* there still could be two species.

Discussion

Many tardigrade species with broad ecological requirements are considered to be cosmopolitan and very common, e.g. *P. richtersi* which is cited for many European localities, America, Africa, New Zealand, New Guinea, and diverse localities of Asia (Ramazzotti and Maucci 1983), whereas others with more narrow tolerances are rare or endemic (Nelson 2002). The "everything is everywhere" hypothesis is suggested especially for small organisms with high rates of dispersal and low rates of local extinction. The authors assume that species found in a given habitat are a function only of habitat properties and not of historical factors (Fenchel and Finlay 2004). However, determining the global distribution and phenotypic specialization of tardigrades is quite difficult, because of undersampling in many parts of the world. A large scale biogeographic study of limno-terrestrial tardigrades was done by McInnes and Pugh (2007) by achieving a 'best fit' of clustergrams with global plate tectonics. However, data are still too scarce to determine definitive biogeographical distributions of tardigrade species. A fine scale study was done by Guil and Giribet (2009) focusing on a small geographical region with the *Echiniscus blumi-canadensis* series. They discovered that the existence of a cryptic lineage, probably corresponding to a different species, can not be distinguished morphologically and further more, they could not confirm the "everything is everywhere" hypothesis.

Guidetti et al. (2009) found three cryptic species within the *Paramacrobiotus* "*richtersi* group," which were detected with 18S rRNA and COI gene sequences, but the calculations by minimum evolution, maximum parsimony (MP) and maximum likelihood (ML) analyses were not sufficient alone for erection of new species, and more supporting

biochemical and biophysical data were not available at that time. Nevertheless, within the study Guidetti et al. (2009) the new genus *Paramacrobiotus* was erected with altogether 22 known species. In this study again we added 18S rRNA gene sequences data from *P.* "*richtersi* group" 4 (data not shown). However, the better resolution benefits greatly from the usage of the sequence-structure information of the ITS2 to discriminate between two species by the existence of a single CBC between an ITS2 sequence-structure pair with a confidence of 93%. In our case study we found more than one CBC between the species which further increase the reliability.

In the CBC analysis *P.* "*richtersi* group" 1 and *P.* "*richtersi* group" 2 clustered together and we found significant differences in the ITS2 compared to the cluster *P.* "*richtersi* group" 3 and the originally described *P. richtersi*, as well as between the two clusters mentioned before and *P.* "*richtersi* group" 4. However, within the cluster of *P.* "*richtersi* group" 1 and *P.* "*richtersi* group" 2 and the cluster of *P.* "*richtersi* group" 3 and *P. richtersi* the CBC analysis was not strong enough to separate the species. As mentioned above, the CBC criterion generally is a sufficient condition to establish species status, but not a necessary one: if there is a CBC then there are two species; if there is no CBC there still could be two species.

The phylogenetic analysis from Guidetti et al. (2009), of the same animals as used in the current study, showed that the COI mtDNA and 18S rDNA marker separates the animals according their origin, even if the 18S rDNA cannot distinguish the two African specimens. However, the existence of different species was already assumed. Meanwhile physiological and biophysical data corroborate this hypothesis. A high-temperature treatment revealed clear differences between the upper-temperature tolerances of *P.* "*richtersi* group" 3 and *P. richtersi*. *P.* "*richtersi* group" 3 showed the lowest survival at higher temperatures, especially at temperatures above 60 °C which resulted in a sharp decrease in survival. The glass-

transition temperatures (T_g) as a measurement for the presence of a vitreous state between *P.* "*richtersi* group" 3 and *P. richtersi* were significantly different, too. This seems to be a species-specific difference between anhydrobiotic animals from Germany and USA and not just an adaptation to the different environment (Hengherr et al. 2009a). Freezing experiments with all used tardigrades (no data for *P.* "*richtersi* group" 4 available) also showed, that after cooling at different rates from room temperature (RT) down to -30 °C, no significant freezing behaviour was detectable (Hengherr et al. 2009b). The temperature of spontaneous freezing (supercooling point, SCP, temperature of crystallization, *Tc*) and quantity of water freezing, studied by differential scanning calorimeter calorimetry was comparable in all species, too, even if we expected a better adaptation from *P.* "*richtersi* group" 3 from Alaska to the cold.

In contrast, *P.* "*richtersi* group" 1 and *P.* "*richtersi* group" 2 showed in the before mentioned heat-tolerance study no significant differences in the glass-transition temperatures and survival. Only a biochemical study showed significant differences in the accumulation of compatible osmolytes (Hengherr et al. 2008). These results along with the phylogenetic analyses using 18S rRNA and COI gene sequences (Guidetti et al. 2009) are not sufficient to separate *P.* "*richtersi* group" 1 and *P.* "*richtersi* group" 2 in two distinct species. In fact the two groups seem to be closely related ecotypes within one new species.

Cross-breeding experiments, at least in principle, would give a theoretical base of the *Paramacrobiotus* species concept. For *P. richtersi* diploid bisexual biotypes (2n = 12) and triploid parthenogenetic biotypes (3n = 18) have been noted (Ramazzotti and Maucci 1983; Hohberg and Greven 2005; Guidetti et al. 2007). But all animals used in this study were parthenogenetic strains, with exception of the outgroups *M. sapiens* and *P. tonollii* (unpublished data). Therefore, the CBC analysis provides an objective method for defining species in asexual taxa, where breeding experiments could never be performed.

The morphological characters of *P. "richtersi* group" 1, *P. "richtersi* group" 2, *P. "richtersi* group" 3, and *P. "richtersi* group" 4 are in agreement with the description of the species *P. richtersi*. The species differ only slightly in body and egg sizes (supplement Tab. S1 & supplement Tab. S2). In this publication we identified three species of the genus *Paramacrobiotus* which are new to science: *Paramacrobiotus kenianus* sp. nov. (formerly *P. "richtersi* group" 1, and *P. "richtersi* group" 2 as ecotype), *Paramacrobiotus fairbanki* sp. nov. (formerly *P. "richtersi* group" 3), and *Paramacrobiotus palaui* sp. nov. (formerly *P. "richtersi* group" 4).

Taxonomic section

Paramacrobiotus kenianus sp. nov. (formerly *Paramacrobiotus "richtersi* group" 1)

Etymology. The species epithet is derived from Kenya, in reference to the sampling location (Nakuru, Kenya).

Type material. Holotype and paratypes, including eggs of the paratypes are deposited in the collection of R. O. Schill (Dept. of Zoology, Biological Institute, Universität Stuttgart, Germany).

Differential diagnosis. *Paramacrobiotus kenianus* sp. nov. can be differentiated from its congeners by compensatory base changes in the internal transcribed spacer 2 (ITS2) secondary structure. The holotype and paratypes are similar to its congeners in both qualitative and metric characters.

Description of the holotype. Body length 429.1 µm, body white or transparent, with eye spots, cuticle smooth without pores, buccal tube length 73.76 µm, outer buccal tube width 11.60 µm (*pt*=15,7), macroplacoid row length 27.3 µm (*pt*=37.1), first macroplacoid length 10.5 µm (*pt*=14.3), second macroplacoid length 10.0 µm (*pt*=14.8), external claw

length, leg 1, 9.6 µm (*pt*=13.0), external claw length, leg, 2 9.9 (*pt*=13.4), external claw length, leg 4, 13.5 µm (*pt*=18.2).

Remarks. Parthenogenetic species, eggs laid freely, egg projections are truncated cones or almost hemispherical, diameter of egg without processes 81.0 ± 4.1 µm, diameter of egg with processes 108.8 ± 6.4 µm, processes on the circumferences 17.7 ± 3.6, processes on the hemisphere 13.1. ± 1.4, processes height 13.5 ± 1.9 µm, basal diameter of processes 19.7 ± 2.7 µm, diameter of distal extremity 5.2 ± 1.3 µm. Within this species an ecotype (formerly *P.* "*richtersi* group" 2) from Nakuru, Kenya was identified, which differed in the trehalose accumulation during transition to an anhydrobiotic state (Hengherr et al. 2008).

Paramacrobiotus fairbanki sp. nov. (formerly *Paramacrobiotus* "*richtersi* group" 3)

Etymology. The species epithet is derived from Fairbanks, in reference to the sampling location (Fairbanks, Alaska, USA).

Type material. Holotype and paratypes, including eggs of the paratypes are deposited in the collection of R. O. Schill (Dept. of Zoology, Biological Institute, Universität Stuttgart, Germany).

Differential diagnosis. *Paramacrobiotus fairbanki* sp. nov. can be differentiated from its congeners by compensatory base changes in the internal transcribed spacer 2 (ITS2) secondary structure (except *Paramacrobiotus richtersi*), and all congeners showed differences in the 18S ribosomal DNA sequences (Guidetti et al. 2009). Calorimetry results (glas-transition temperatures) and survival after exposure to high temperatures are significantly different between *P. fairbanki* sp. nov. and *P. richtersi* (Hengherr et al. 2009a). The holotype and paratypes are similar to its congeners in both qualitative and metric characters.

Description of the holotype. Body length 473.5 µm, body white or transparent, with eye spots, cuticle smooth without pores, buccal tube length 56.4 µm, outer buccal tube width 12.8 µm (*pt*=22.6), macroplacoid row length 27.2 µm (*pt*=48.1), first macroplacoid length 9.8 µm (*pt*=17.4), second macroplacoid length 8.2 µm (*pt*=14.5), external claw length, leg 1, 10.1 µm (*pt*=18.0), external claw length, leg 2, 10.3 (*pt*=18.2), external claw length, leg 4, 10.5 µm (*pt*=18.7).

Remarks. Parthenogenetic species, eggs laid freely, egg projections are truncated cones or almost hemispherical, diameter of egg without processes 69.9 ± 3.4 µm, diameter of egg with processes 91.3 ± 3.8 µm, processes on the circumferences 15.1 ± 3.4, processes on the hemisphere 11.1 ± 1.4, processes height 11.5 ± 1.2 µm, basal diameter of processes 17.6 ± 2.3 µm, diameter of distal extremity 6.3 ± 0.7 µm.

Paramacrobiotus palaui sp. nov. (formerly *Paramacrobiotus* "*richtersi* group" 4)

Etymology. The species epithet is derived from Palau, in reference to the sampling location (Koror, Palau).

Type material. Holotype and paratypes, including eggs of the paratypes are deposited in the collection of R. O. Schill (Dept. of Zoology, Biological Institute, Universität Stuttgart, Germany).

Differential diagnosis. *Paramacrobiotus kenianus* sp. nov. can be differentiated from its congeners by compensatory base changes in the internal transcribed spacer 2 (ITS2) secondary structure. The holotype and paratypes are similar to its congeners in both qualitative and metric characters.

Description of the holotype. Body length 453.1 µm, body white or transparent, with eye spots, cuticle smooth without pores, buccal tube length 56.5 µm, outer buccal tube width 15.5 µm (*pt*=27.4), macroplacoid row length 29.9 µm (*pt*=52.9), first macroplacoid length

9.7 µm (*pt*=17.2), second macroplacoid length 7.5 µm (*pt*=13.3), external claw length, leg 1, 7.6 µm (*pt*=13.4), external claw length, leg 2, 10.2 (*pt*=18.1), external claw length, leg 4, 11.0 µm (*pt*=19.4).

Remarks. Parthenogenetic species, eggs laid freely, egg projections are truncated cones or almost hemispherical, diameter of egg without processes 55.3 ± 1.1 µm, diameter of egg with processes 73.6 ± 1.6 µm, processes on the circumferences 15.4 ± 1.4, processes on the hemisphere 11.4 ± 0.5, processes height 10.2 ± 1.3 µm, basal diameter of processes 13.4 ± 1.3 µm, diameter of distal extremity 5.0 ± 0.6 µm.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-3402.

Coleman, A. W. (2000). The significance of a coincidence between evolutionary landmarks found in mating affinity and a DNA sequence. *Protist*, 151(1), 1-9.

Coleman, A. W. (2003). ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics*, 19(7), 370-375.

Coleman, A. W. (2007). Pan-eukaryote ITS2 homologies revealed by RNA secondary structure. *Nucleic Acids Research*, 35(10), 3322-3329.

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

Coleman, A. W., & Vacquier, V. D. (2002). Exploring the phylogenetic utility of ITS sequences for animals: a test case for abalone (*Haliotis*). *Journal of Molecular Evolution*, 54(2), 246-257.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39, 783-791.

Fenchel, T., & Finlay, B. J. (2004). The Ubiquity of Small Species: Patterns of Local and Global Diversity. *Bioscience Biotechnology and Biochemistry*, 54(8), 777-784.

Friedrich, J., Dandekar, T., Wolf, M., & Müller, T. (2005). ProfDist: a tool for the construction of large phylogenetic trees based on profile distances. *Bioinformatics*, 21(9), 2108-2109.

Garey, J. R., Nelson, D. R., Mackey, L. Y., & Li, J. (1999). Tardigrade phylogeny: congruency of morphological and molecular evidence. *Zoologischer Anzeiger*, 238(3/4), 205-210.

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.

Goeze, J. A. E. (1773). *Herrn Karl Bonnets Abhandlungen aus der Insektologie aus d. Franz. übers. u. mit einigen Zusätzen hrsg. v. Joh. August Ephraim Goeze*. Halle: Gebauer.

Guidetti, R., & Bertolani, R. (2001). Phylogenetic relationships in the Macrobiotidae (Tardigrada: Eutardigrada: Parachela). *Zoologischer Anzeiger*, 240(3-4), 371-376.

Guidetti, R., & Bertolani, R. (2005). Tardigrade taxonomy: an updated check list of the taxa and a list of characters for their identification. *Zootaxa*, 845, 1-46.

Guidetti, R., Colavita, C., Altiero, T., Bertolani, R., & Rebecchi, L. (2007). Energy allocation in two species of Eutardigrada. *Journal of Limnology*, 66(Suppl. 1), 111-118.

Guidetti, R., Gandolfi, A., Rossi, V., & Bertolani, R. (2005). Phylogenetic analysis of Macrobiotidae (Eutardigrada, Parachela): a combined morphological and molecular approach. *Zoologica Scripta*, 34(3), 235-244.

Guidetti, R., Schill, R. O., Bertolani, R., Dandekar, T., & Wolf, M. (2009). New molecular data for tardigrade phylogeny, with the erection of *Paramacrobiotus* gen. nov. *Journal of Zoological Systematics and Evolutionary Research*, 47(4), 315-321.

Guil, N., & Giribet, G. (2009). Fine scale population structure in the Echiniscus blumi-canadensis series (Heterotardigrada, Tardigrada) in an Iberian mountain range-When morphology fails to explain genetic structure. *Mol Phylogenet Evol*, 51(3), 606-613.

Gutell, R. R., Larsen, N., & Woese, C. R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews*, 58(1), 10-26.

Hengherr, S., Heyer, A. G., Kohler, H. R., & Schill, R. O. (2008). Trehalose and anhydrobiosis in tardigrades-evidence for divergence in responses to dehydration. *FEBS Journal*, 275(2), 281-288.

Hengherr, S., Worland, M. R., Reuner, A., Brummer, F., & Schill, R. O. (2009a). High-temperature tolerance in anhydrobiotic tardigrades is limited by glass transition. *Physiological and Biochemical Zoology: PBZ*, 82(6), 749-755.

Hengherr, S., Worland, M. R., Reuner, A., Brümmer, F., & Schill, R. O. (2009b). Freeze tolerance, supercooling points and ice formation: comparative studies on the subzero temperature survival of limno-terrestrial tardigrades. *Journal of Experimental Biology*, 212(6), 802.

Hohberg, K., & Greven, H. (2005). Retention of embryonated eggs in parthenogenetic *Macrobiotus richtersi* Murray, 1911 (Eutardigrada). *Zoologischer Anzeiger*, 243(3), 211-213.

Jørgensen, A. (2000). Cladistic analysis of the Echiniscidae Thulin, 1928 (Tardigrada: Heterotardigrada: Echiniscoidea). *Steenstrupia*, 25, 11-23.

Jørgensen, A., & Kristensen, R. M. (2004). Molecular phylogeny of Tardigrada - Investigation of the monophyly of Heterotardigrada. *Mol Phylogenet Evol*, 32(2), 666-670.

Jørgensen, A., Møbjerg, N., & Kristensen, R. M. (2007). Molecular study of the tardigrade *Echiniscus testudo* (Echiniscidae) reveals low DNA sequence diversity over a large geographical area. *Journal of Limnology*, 66(Suppl. 1), 77-83.

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

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T., & Wolf, M. (2009). 5.8S-28S rRNA interaction and HMM-based ITS2 annotation. *Gene*, 430(1-2), 50-57.

Koetschan, C., Förster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R. et al. (2010). The ITS2 Database III--sequences and structures for phylogeny. *Nucleic Acids Res*, 38(Database issue), D275-279.

Kristensen, R. M. (1987). Generic revision of the Echiniscidae (Heterotardigrada), with a discussion of the origin of the family. *Biology of Tardigrades Selected Symposia and Monographs UZI*, 1, 261-335.

Marcus, E. (1929). Tardigrada. *HG Bronn's Klassen und Ordnungen des Tierreichs 5, IV, 3 Akademische Verlagsgesellschaft, Leipzig*.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., & Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19), 7287-7292.

McInnes, S. J., & Pugh, P. J. A. (2007). An attempt to revisit the global biogeography of limno-terrestrial Tardigrada. *Journal of Limnology*, 66(Suppl. 1), 90-96.

Møbjerg, N., Jørgensen, A., Eibye-Jacobsen, J., Agerlin Halberg, K., Persson, D., & Kristensen, R. M. (2007). New records on cyclomorphosis in the marine eutardigrade *Halobiotus crispae* (Eutardigrada: Hypsibiidae). *Journal of Limnology*, 66(Suppl. 1), 132-140.

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

Nelson, D. R. (2002). Current status of the tardigrada: evolution and ecology. *Integrative and Comparative Biology*, 42(3), 652-659.

Nichols, P. B., Nelson, D. R., & Garey, J. R. (2006). A family level analysis of tardigrade phylogeny. *Hydrobiologia*, 558(1), 53-60.

Pilato, G. (1981). Analisi di nuovi caratteri nello studio degli Eutardigradi. *Animalia*, 8, 51-57.

Ramazzotti, G., & Maucci, W. (1983). II phylum Tardigrada. *Memorie dell'Istituto Italiano di Idrobiologia*, 41, 1–1012.

Regier, J. C., Shultz, J. W., Kambic, R. E., & Nelson, D. R. (2004). Robust support for tardigrade clades and their ages from three protein-coding nuclear genes. *Invertebrate Biology*, 123(2), 93-100.

Schill, R. O. (2007). Comparison of different protocols for DNA preparation and PCR amplification of mitochondrial genes of tardigrades. *Journal of Limnology*, 66(Suppl. 1), 164-170.

Schmitt, S., Hentschel, U., Zea, S., Dandekar, T., & Wolf, M. (2005). ITS-2 and 18S rRNA gene phylogeny of Aplysinidae (Verongida, Demospongiae). *Journal of Molecular Evolution*, 60(3), 327-336.

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

Schultz, J., Müller, T., Achtziger, M., Seibel, P. N., Dandekar, T., & Wolf, M. (2006). The internal transcribed spacer 2 database--a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Research*, 34(Web Server issue), W704-707.

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

Seibel, P. N., Müller, T., Dandekar, T., Schultz, J., & Wolf, M. (2006). 4SALE--a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, 7, 498.

Seibel, P. N., Müller, T., Dandekar, T., & Wolf, M. (2008). Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Research Notes*, 1, 91.

Selig, C., Wolf, M., Müller, T., Dandekar, T., & Schultz, J. (2008). The ITS2 Database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Research*, 36(Database issue), D377-380.

Spallanzani, L. (1776). *Opuscoli di Fisica animale e vegetabile*. Modena.

White, T. J., Bruns, T., Lee, S., & Taylor, J. (1990). PCR protocols: a guide to methods and applications. *Chap Amplification and Direct Sequencing of Fungal Ribosomal RNA Genes for Phylogenetics*, 42(1-2).

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

Wolf, M., Friedrich, J., Dandekar, T., & Müller, T. (2005b). CBCAnalyzer: inferring phylogenies based on compensatory base changes in RNA secondary structures. *In Silico Biology*, 5(3), 291-294.

Wolf, M., Ruderisch, B., Dandekar, T., Schultz, J., & Müller, T. (2008). ProfDistS: (profile-) distance based phylogeny on sequence--structure alignments. *Bioinformatics*, 24(20), 2401-2402.

Wolf, M., Selig, C., Müller, T., Philippi, N., Dandekar, T., & Schultz, J. (2007). Placozoa: at least two. *Biologia*, 62(6), 641-645.

Table 1: GenBank accession no. and the lengths of the ITS2 sequences.

Species	Accession number	Length of ITS2 in basepairs
Paramacrobiotus kenianus sp. nov. (*Paramacrobiotus* "*richtersi* group" 1)	GQ403674	347
Paramacrobiotus kenianus sp. nov. (*Paramacrobiotus* "*richtersi* group" 2)	GQ403675	347
Paramacrobiotus palaui sp. nov. (*Paramacrobiotus* "*richtersi* group" 4)	GQ403676	351
Paramacrobiotus richtersi	GQ403677	344
Paramacrobiotus fairbanki sp. nov. (*Paramacrobiotus* "*richtersi* group" 3)	GQ403678	345
Paramacrobioutus tonollii	GQ403679	340
Macrobiotus sapiens	GQ403680	309

Table 2: Compensatory base changes.

Paramacrobiotus kenianus sp. nov. (*Paramacrobiotus* "*richtersi* group" 1)	0	0	5	3	3	6	5
Paramacrobiotus kenianus sp. nov. (*Paramacrobiotus* "*richtersi* group" 2)	0	0	5	3	3	6	5
Paramacrobiotus palaui sp. nov. (*Paramacrobiotus* "*richtersi* group" 4)	5	5	0	7	7	9	6
Paramacrobiotus richtersi	3	3	7	0	0	3	4
Paramacrobiotus fairbanki sp. nov. (*Paramacrobiotus* "*richtersi* group" 3)	3	3	7	0	0	4	5
Paramacrobiotus tonollii	6	6	9	3	4	0	3
Macrobiotus sapiens	5	5	6	4	5	3	0

Figure legends

Figure 1: Secondary structure of *Paramacrobiotus* ITS2 visualised by 4SALE. **A)** Template structure (*P.* "*richtersi* group" 2) used for homology modeling of the remaining ITS2 secondary structures. **B)** Consensus structure for all ITS2-sequences. The consensus structure shows all positions which are conserved in at least 51% over all sequence-structure pairs. Sequence conservation is indicated corresponding to the legend. Nucleotide-bonds which are 100% conserved are marked with dotted circles. CBCs distinguishing species in the *P.* "*richtersi* group" are surrounded by black boxes (c.f. Figure 2B).

Figure 2: Phylogenetic tree topologies and sampling locations. **A)** Neighbor-joining tree obtained by ProfDistS and supporting bootstrap values (1000 replicates) are shown in black. The CBC tree obtained by CBCanalyzer is underlain in dark grey. Corresponding sample locations are indicated by arrows. **B)** Numbers of CBCs distinguishing three species classified within *Paramacrobiotus*. The grey ovals correspond to the grey ovals in subfigure 2A and indicate the species groups which can be identified on the basis of CBCs.

Supplement Material

Supplement Table 1: Examination of specimens from the culture of R. O. Schill (Dept. of Zoology, Biological Institute, Universität Stuttgart, Germany). They are mounted in polyvinyl lactophenol, using phase contrast Microscopy for the measurements (μm). The pt index is the ratio between the lengh of a structure and the length of the buccal tube expressed as a percentage (Pilato 1981)

Supplement Table 2: Examination of the egg of *Paramacrobiotus kenianus* sp. nov. (formerly *Paramacrobiotus* "*richtersi* group" 1, and *Paramacrobiotus* "*richtersi* group" 2), *Paramacrobiotus fairbanki* sp. nov. (formerly *Paramacrobiotus* "*richtersi* group" 3), *Paramacrobiotus palaui* sp. nov. (formerly *Paramacrobiotus* "*richtersi* group" 4), and Paramacrobiotus richtersi (in μm) from the culture of R. O. Schill (Dept. of Zoology, Biological Institute, Universität Stuttgart, Germany).

A

I

II

III

IV

B

I

II

III

IV

Conservation
0.0   1.0
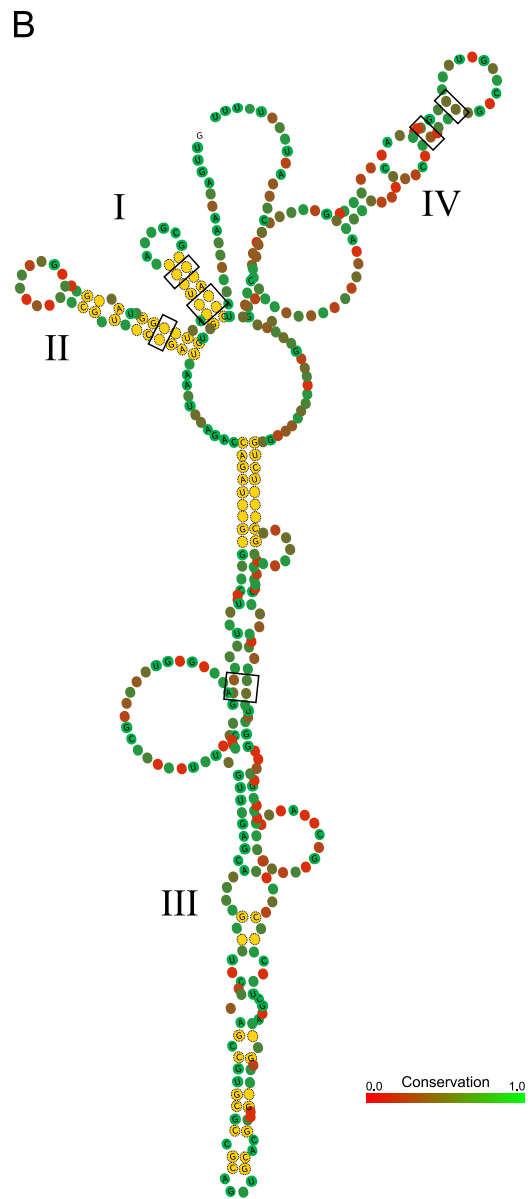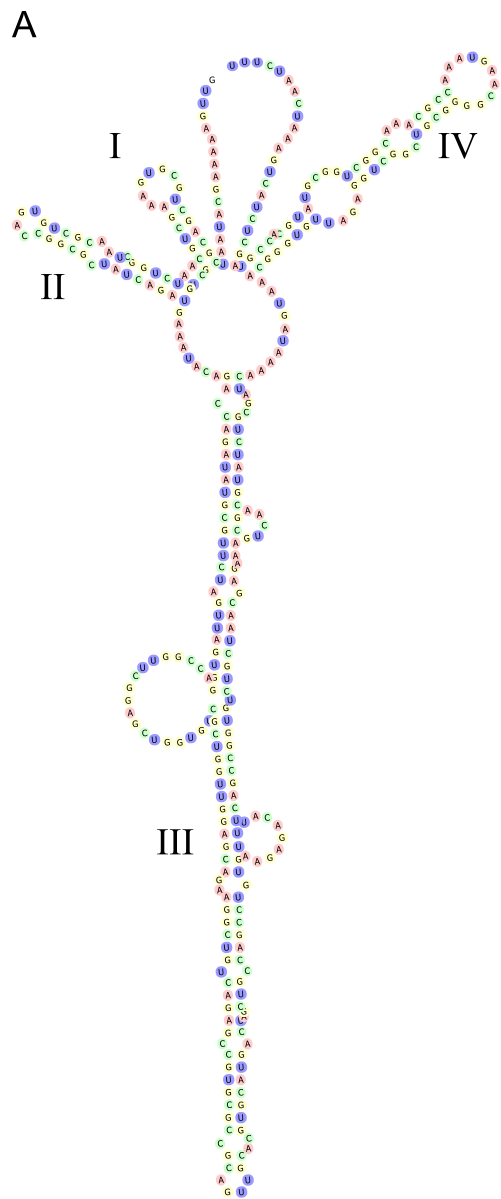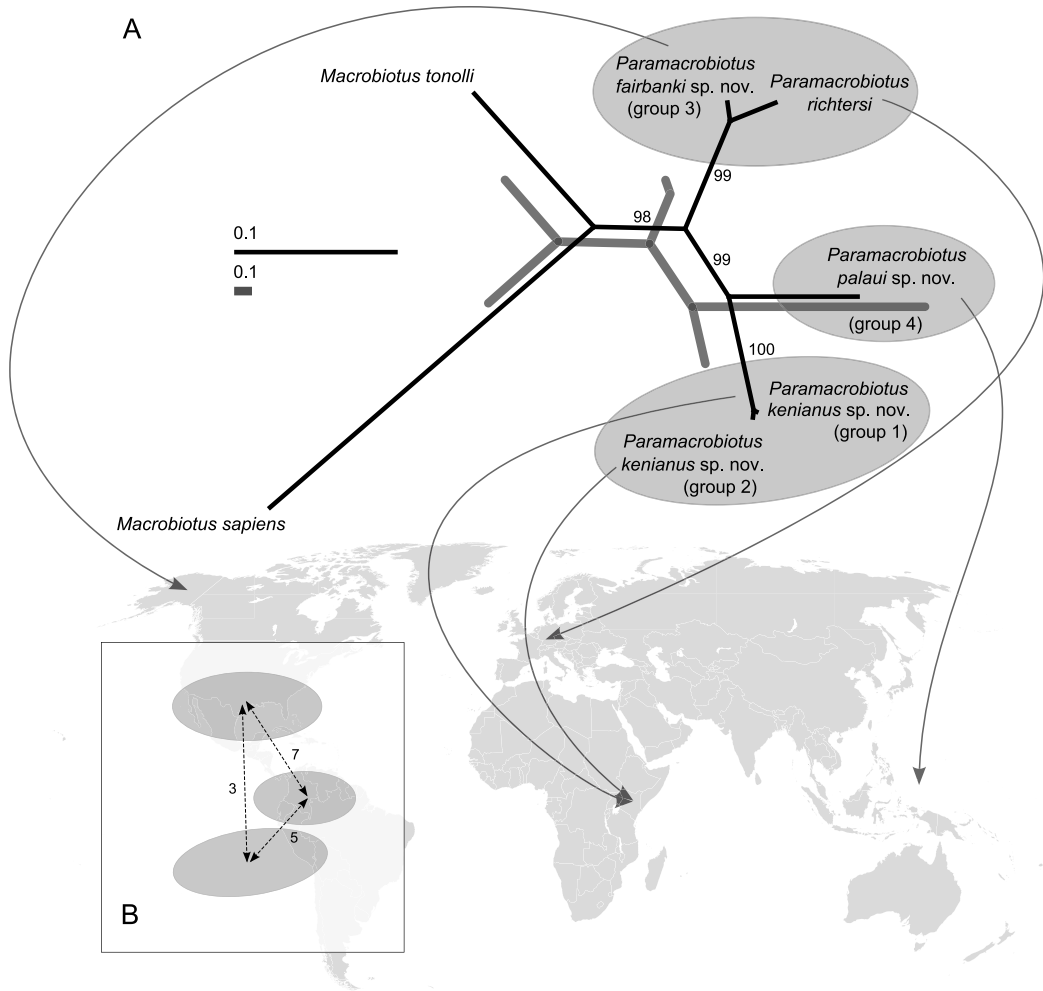
Part IV.

General Discussion and Conclusion

The phylum Tardigrada consists of animals with an interesting capability to resist extreme environmental conditions, like high and low pressure, high and low temperature or an amount of radiation, which is lethal to human beings. The question is, how can the members of the Tardigrada resist such conditions? Is it a kind of conservation or a repair mechanism? In this thesis, I tried to find answers to these questions. My part was the combination and the functional annotation of the data obtained by our laboratory work and the prediction of promising sequence candidates and pathways which might be involved in the adaption capabilities of the tardigrades.

Therefore, I established a flexible web-based database with various tools—the tardigrade workbench (published in Förster et al. 2009). The database contains the sequences obtained by our sequencing projects as well as other tardigrade sequences available at public databases, e.g. *H. dujardini* and *R. coronifer*. A part of the offered sequences is accessible at the common public databases, e.g. Genbank, but we offer a species specific BLAST search as well as a pattern search—a feature which so far no other database provides. Additionally, I developed an annotation pipeline for EST and protein data to allow their functional annotation. The sequence data were functionally classified using the cluster of orthologous group (COG) database (Tatusov et al. 1997; Tatusov et al. 2003) and were clustered with the CLANS algorithm.

To obtain stage dependent nucleotide data for *M. tardigradum* we generated two different directionally cloned complementary desoxyribonucleic acid (cDNA) libraries (published in Mali et al. 2010) and sequenced them using the Sanger method. The transcriptome sequencing project resulted in approximately 10,000 sequences. These were stage-specific assembled. This led to about 2000 unigenes for the active and the tun stage. A study in *C. elegans* under anoxia showed adaptions such as a cell cycle arrest, dephosphorylation of the histone H3 and morphological changes in the chromatin distribution (Padilla et al. 2002). Therefore, we performed an Gene Ontology (GO) enrichment analysis between both sets of ESTs to get information on the functional differences between the two stages. We identified 24 GO terms which were significantly underrepresented in the tun stage: 'nucleosome', 'nucleosome assembly', 'chromatin assembly or disassembly' and 'chromatin assembly'. The cellular component subset of the differential terms were associated with the structural components of the genome, e.g. 'nucleosome' or 'chromatin'. The identification of only underrepresented terms was consistent with the expected global metabolic arrest of animals undergoing cryptobiosis. Such a metabolic suppression could limit the genomic and cellular damage by minimisation of energy turnover. Further, GO terms involved in translation regulation were affected, e.g. 'regulation of translation' or 'translation factor activity, nucleic acid binding'. This implied a modulation of the translational activity as a response to desiccation.

Moreover, we analysed the most abundant ESTs in both *M. tardigradum* stages.

Possible explanations for the relative abundance of some transcripts in the tun stage might either be the storage for a translation on rehydration or a higher transcription of these ESTs during the desiccation. A high survival rate of *M. tardigradum* is accomplished only, when the dehydration occurs at high relative humidity (Hengherr et al. 2008; Horikawa and Higashi 2004). This might be necessary because the transcription of mRNAs coding for protective components has to take place first.

We identified intracellular fatty acid binding proteins (FABPs) at higher levels in tun stage. Described functions for FABPs are various: uptake, transport and delivery of fatty acids to β-oxidation (Haunerland and Spener 2004; Hittel and Storey 2001) or protective functions for fatty acids (Coe and Bernlohr 1998; Makowski and Hotamisligil 2004). FABPs may be involved in the protection of membranes and the storage of fatty acids during anhydrobiosis. We also found protease inhibitor levels increased in the tun stage of *M. tardigradum*. These mRNAs belonged to the Kazal-type serine proteinase inhibitor and Cystatin B. They might inhibit proteolytic reactions of proteases that could damage structures during the desiccation process. Also, a protection mechanism against microbial degradation is imaginable as this can occur at humidity levels at which tardigrades can not rehydrate. Therefore, tardigrades would be unable to activate any cellular defence mechanisms. The transcription of Cytochrome c oxidase subunit I (COI) was also increased in the inactive stage. A higher level of COI is described for the dehydrated antarctic nematode *Plectus murrayi* (Adhikari et al. 2009) and after heat shock of the yeast-like fungus *Cryptococcus neoformans* (Toffaletti et al. 2003). The upregulation of the COI may prevent damage to the electron transport chain and allows an increased energy production for the survival of *M. tardigradum*.

Our knowledge about the regulation of transcripts in tardigrades is limited. For that reason, we scanned the 5′ and 3′ UTRs of *M. tardigradum* and *H. dujardini* for regulatory RNA motifs (published in Förster et al. 2009, 2010*b*). These motifs are involved in the regulation of the turnover of mRNAs. In tardigrades mainly the 15-lipoxygenase differentiation control element (15-LOX-DICE) (Ostareck-Lederer et al. 1994), the K-box (Lai et al. 2005) and the brd-box (Lai 2002) were found. In contrast, instability motifs which are commonly found in vertebrates, e.g. the AU rich element (Chen and Shyu 1995), were not detected.

Due to the fact, that the transcriptome does not have to reflect the level of proteins in the organism, we also examined the proteome of *M. tardigradum* (published in Schokraie et al. 2010). Therefore, we developed a protocol for protein extraction and separation by high resolution two-dimensional gel electrophoresis. The identified peptides were scanned against NRdb from NCBI. One disadvantage of the NRdb is that only a few tardigrade proteins are known. Due to this reason, peptides can be identified only if they are very similar to the tardigrade proteins. On the basis of our own transcriptome sequencing project we were able to detect more significant hits within the proteome. The reason for this is on the one hand the reduced search

space and on the other hand that the sequences from the EST sequences are identical to the peptides.

In total we found 144 proteins with a known function and additional 36 proteins with significant hits against our EST dataset, but are not functionally annotated. Therefore, they were designated as new specific proteins of *M. tardigradum*. We identified some proteins which showed a lower molecular weight than expected, e.g. the HSP60 protein is detected as a protein at 24 kDa in Western Blot and the two dimensional gel electrophoresis. Moreover, proteins often showed multiple spots in two dimensional gels, due to post-translational modification like glycosylation or phosphorylation. Therefore, we are using fluorescence staining methods like ProQ-Emerald for glycosylation and ProQ-Diamond for phosphorylation detection in the ongoing analysis of the *M. tardigradum* proteome.

The obtained nucleotide and protein data from our transcriptome and proteome sequencing projects were stored in our tardigrade workbench. The CLANS clustering resulted in some clusters of stress related proteins like Ubiquitin, Cathepsin like proteins, small HSPs. Ubiquitin- and Cathepsin-like proteins may be involved in stress-induced protein degradation or regulation. Small HSPs play a role in the prevention of protein aggregation and act as molecular chaperons (Sun and MacRae 2005). Therefore we examined the stress response in *M. tardigradum* in detail (published in Reuner et al. 2008). We identified a complete cDNA of one HSP10 and two *α*-crystallin/small HSPs of 17.2 kDa (150 aa) and 19.5 kDa (174 aa) in our EST dataset which are differently regulated.

The smaller HSP17.2 is induced by heat shock, but not during anhydrobiosis. In contrast, the amount of the larger HSP19.5 is not affected by heat shock, but is downregulated in the transition from tun to active state in *M. tardigradum*. Due to the different expression pattern, we assume different functions for both transcripts. The role of the other HSPs during the anhydrobiosis is debatable, due to the little change in the expression levels, but it is not known whether a basal level of these proteins in tardigrades exists and thus upregulation may not be necessary or regulation involves not expression but allosteric regulation. Nevertheless, the important role of small HSPs was shown for *Artemia franciscana*. The protein p26 is massively accumulated in the diapause of embryos (Liang et al. 1997*a,b*). It is able to move into the nucleus (Clegg et al. 1995) and seems to protect or chaperone in cooperation with HSP70 (Willsie and Clegg 2002).

The analysis of the expression levels of the three isoforms of HSP70 described by Schill et al. (2004) showed for two of three isoforms no change in expression, but one isoform was significantly induced in the transitional stage between tun and active state. Due to the upregulation we suggest a functional role during anhydrobiosis of the HSP70 isoform. An alternative role might be prevention of proteins from aggregation or unfolding during the loss of cellular water or may be the establishment of a refolding system to provide protein functions during or after

rehydration.

The HSP90 of *M. tardigradum* was the only HSP in our study with a clear and significant increase in mRNA content in the anhydrobiotic stage. No increase was detected in transitional stage I. As there is no transcription in the stage of anhydrobiosis, due to the reduced metabolism (Pigon and Weglarska 1955), the reason for the storage of such an amount of mRNA of HSP90 in only the anhydrobiotic stage still remains to be elucidated.

The analysis of the HSPs clearly showed that the level of most mRNAs is decreased during anhydrobiosis. However the basal protein level of the stress proteins, which is necessary for protection or repair, in the tardigrade *M. tardigradum* is still unknown. Our study suggests a minor role of refolding or stabilisation of stress proteins. Thus it seems likely that denaturation of proteins is not a major problem for *M. tardigradum* during desiccation.

To find common mechanisms in tardigrades which provide their capabilities to resist extreme environmental conditions, we compared the available sequences of the three tardigrades *M. tardigradum*, *H. dujardini* and *R. coronifer* at sequence and annotation levels (published in Förster et al. 2009, 2010*a,b*; Mali et al. 2010). We found some interesting sequences responsible for detoxification , DNA repair, protein folding, Aquaporins (AQPs) and LEA proteins.

The oxidative stress proteins are an important component in many biological processes (França et al. 2007). They play a role as antioxidants such as Gluthathione S-transferase (GST), Thioredoxin and Superoxide dismutase (SOD). An over-expression of GST/Gluthathione peroxidase increases the resistance to oxidative and water stress in tobacco plants (Roxas et al. 1997). GST is also known to play a prominent role in the detoxification metabolism in nematodes (Lindblom and Dodd 2006). Therefore, we postulate that the tardigrade GST and SOD deal with oxidatively damaged cellular components during desiccation.

We identified a homologue of the recombination repair gene rhp57 (RAD57 homologue of *Schizosaccharomyces pombe*) in the EST datasets. Homologues within other species have been described, e.g. the human protein Xrcc3 (Liu et al. 1998). A deletion strain of *S. pombe* was more sensitive to methyl methanesulfonate, UV- and γ-rays (Tsutsui et al. 2000). Therefore, one might speculate that the tardigrade protein is involved in DNA repair and provides the capability to resist radiation.

The members of the DnaJ protein family stimulate the HSP70 chaperones and therefore are important for protein translation, folding and unfolding, translocation and degradation of proteins (Qiu et al. 2006). We identified 58 DnaJ-proteins in *M. tardigradum,* eight in *H. dujardini* and one in *R. coronifer* and reconstructed a maximum likelihood tree for these proteins (published in Förster et al. 2010*a*). The tree showed a large diversity of the proteins, which is supported by the number of DnaJ family members found in other organisms, e.g. 41 DnaJ-proteins in human. This diversity implies an enhanced adaption potential.

Many organisms activate AQPs to adapt to desiccation (Izumi et al. 2006; Philip et al. 2008). *Polypedilium vanderplanki* contains two AQPs. One is involved in anhydrobiosis and the other controls the water homoeostasis during normal environmental conditions (Kikawada et al. 2008). In *Eurosta solidaginis* one AQP is up-regulated and two other are downregulated following desiccation (Philip et al. 2008). We identified AQP-like proteins in all EST sets and suggest that these AQPs—together with other trans-membrane proteins—mediate the rapid transport of water across the plasma membrane during anhydrobiosis when the diffusion through the membrane is limited.

The group 3 LEA proteins (Tunnacliffe and Wise 2007) have been associated with anhydrobiosis in some nematodes (Adhikari et al. 2009; Browne et al. 2002; Browne et al. 2004; Gal et al. 2004) and other invertebrates (Bahrndorff et al. 2009; Hand et al. 2007; Kikawada et al. 2006; Pouchkina-Stantcheva et al. 2007). A molecular shield activity for many LEA proteins was postulated. Therefore, the LEAs proteins reduce inappropriate interactions between other proteins by an electrosteric mechanism (Goyal et al. 2005; Wise and Tunnacliffe 2004). So it might be as well that LEA proteins provide the resistance against the anhydrobiosis in tardigrades. We identified LEA protein homologues in *H. dujardini* and *R. coronifer* but not in *M. tardigradum*. However, in the proteomic approach we identified a protein spot which is associated with LEA protein which is supported by a Western Blot using LEA specific antibodies (published in Förster et al. 2010*b*; Schokraie et al. 2010). The reconstruction of a phylogenetic tree of the PFAM LEA-domain hits clearly showed that all obtained LEA proteins belong to the LEA group 3 proteins.

However, all these speculation about the functions in tardigrades have to be proven in the laboratory to understand the underlying mechanisms for the tardigrade adaption capability. Therefore, we are currently sequencing the transcriptome using next generation sequencing techniques, leading to quantitative data. Additionally, the annotation of the EST data will be improved by mapping to genome sequences. Unfortunately, neither a genome sequencing project for *M. tardigradum* was initiated nor the genome sequencing project for *H. dujardini* is finished.

In the second part of this thesis I examined the phylogenetic relationship of cryptic species within the tardigrade genus *Paramacrobiotus*. They were sampled from places in North America, Europe, Africa and Asia. Due to their uniform morphology, they can not be distinguished as different species based on morphology characters. However, a 18S and COI examination indicated the presence of different species within the sample set, but was not sufficient. To solve this phylogenetic question we used the ITS2 as phylogenetic marker, which resulted in the description of three new tardigrade species. In preparation of the usage of the ITS2 within the Tardigrada, we completely redesigned the ITS2 database and showed in an *in silico* simulation study the benefit we obtain by adding structure information for

phylogenetic reconstruction. Additional I was involved in a biological case study about the phylogenetic relationship within the Chlorophyceae.

For the first step we reorganised the comprehensive ITS2 database (published in Koetschan et al. 2010). New database content was assembled and the scheme of the complete database was redesigned. In contrast to the earlier version of the database (Selig et al. 2008) we do not use a local alignment algorithm for the detection of possible ITS2 sequences any more. Now, we included the HMM annotation in our pipeline to determine the starting and ending positions of the ITS2 (Keller et al. 2009). In older versions the homology modelling step was restricted to four iterations. In contrast, in the actual version of the ITS2 database we iterate until no further sequence structure pairs can be obtained. In our last step, we are using a BLAST search (Altschul et al. 1997) to find partial structures. The whole new pipeline obtained more then 210,000 sequence-structure pairs which doubles the number compared to the previous version (Selig et al. 2008). From a taxonomic view, we now achieve a coverage of at least 80 % for fungi and plants but only around 25 % for metazoans of all ITS2 sequences listed in Genbank. This might indicate that the structure of the metazoan ITS2 deviates from the common core. Nevertheless, the ITS2 database offers a unique and large comprehensive dataset on ITS2 structures of all kingdoms.

For the reason of a better phylogenetic tree reconstruction, it was necessary to show the benefit of adding structure information to the sequence information of the ITS2. The improvement of the reconstruction of a bilaterian tree by adding structure information to a sequence marker was shown for the SSU (Telford et al. 2005). They tested different phylogenetic models of evolution and built the final tree using a Bayesian framework. In contrast, we published a workflow based on a profile neighbour joining tree reconstruction for the ITS2 as phylogenetic marker (Schultz and Wolf 2009). Therefore, we performed a large simulation study to estimate the benefit, using sequence or sequence-structure information for the ITS2. In addition we compared our results against a marker with the doubled length.

We found that the robustness and the accuracy of a tree is decreased by an increased number of taxa. This is also described by Bremer et al. (1999) and Rokas and Carroll (2005). In our study we showed that the addition of secondary structure information leads to an enlarged range of optimal performance and to a shift of the best performance to a higher sequence divergence. Because of the impact of both parameters, the shift does not implicate that closely related taxa can not be resolved. This is a proof for the applicability of the ITS2 as wide range phylogenetic marker, as assumed by several case studies in Coleman (2003); Wiemers et al. (2009).

We found that the addition of secondary structure informations results in a large increase of tree robustness. This is expected as the additional data increase the information content due to a higher number of possible states for each marker-position. These extra information should be used if available and requires software

capable of dealing with this information, e.g. 4SALE (Seibel et al. 2006). For phylogeny the major benefit is the improvement of accuracy by using sequence-structure instead of sequence information only.

In comparison, both the addition of structure information and the extension of the marker length improved the reconstructed phylogenetic trees. In detail, longer markers primarily increased the bootstrap support values as has been shown previously (Bremer et al. 1999; Erixon et al. 2003), whereas the addition of secondary structure information predominantly improves the accuracy. Moreover, as the secondary structure of the ITS2 covers the whole sequence length, an elongation of the ITS2 is not possible for real life data.

We showed a clear benefit of adding secondary structure information to allow a more accurate phylogenetic tree reconstruction in a theoretical study. To transfer this knowledge to a real sequence dataset, we analysed the phylogenetic relationship within the Chlorophyceae (published in Keller et al. 2008). The clustering within the groups is known (Buchheim et al. 2001; Wolf et al. 2002), but the backbone was unclear. The sequence-structure information of the ITS2 as phylogenetic marker allowed to resolve the relationship of Sphaeropleaceae as a part of the Sphaeropleales representing the monophyletic biflagelate DO-group. In our study *A. judayi*, *A. porcata* and *S. annulina* clustered in a monophyletic clade named Sphaeropleaceae. We confirmed this clade with all three species clearly separated from other clades. In our study we verified the DO-group using molecular phylogenetic analysis as emended by Deason et al. (1991). We showed that the 'core Sphaeropleales', the Spheropleales and the Sphaeropleaceae are monophyletic. Furthermore, we investigated whether the secondary structure of the ITS2 is an autapomorphic feature for the monophyletic DO-group. Van Hannen et al. (2002) published a secondary structure model based upon minimum free energy calculation that showed a branched helix I. We found the branched structure is missing within the Sphaeroplea clade and all other investigated groups. Thus the feature does not seem to be an autapomorphic character for the biflagelate DO-group as whole, but for the 'core Sphaeropleales'. We did not include ITS2 sequences of chaetopeltidalean and chaetophoralean taxa. Therefore, the phylogenetic relationship between the main Chlorophyceae clades remained open.

Finally, we applied the ITS2 work-flow on a phylogenetic question concerning the tardigrade species (published in Schill et al. 2010). Guidetti et al. (2009) described three cryptic species within the *Macrobiotus 'richtersi* group', which were detected with 18S rRNA or COI gene sequences, and attributed them to the new genus *Paramacrobiotus*. However, the analysis using minimum evolution, maximum parsimony and maximum likelihood algorithms were not sufficient for the erection of new species. Since sequence-structure information improve the accuracy and robustness, we added sequence-structure information of the ITS2 and sequence data for the 18S rRNA from *P. 'richtersi* group 4' to our dataset.

In the CBC analysis we detected three clusters within the examined animals: one cluster consists of *P. 'richtersi* group 1' and *P. 'richtersi* group 2', one cluster consists of *P. 'richtersi* group 4' and the last cluster consists of *P. 'richtersi* group 3' and *Paramacrobiotus richtersi*. Between all these clusters at least three CBCs occur which allowed us to separate the species according to the CBC criteria published by Müller et al. (2007). Moreover, the original publication estimated the accuracy for one CBC. As we found at least three CBCs the reliability should further increase.

Additionally, the phylogenetic analysis of *P. 'richtersi* group 3' and *P. richtersi* using the 18S rDNA maximum likelihood tree (Guidetti et al. 2009) and physiological and biophysical data concerning the difference in high (Hengherr et al. 2009*b*) and low temperature (Hengherr et al. 2009*a*) tolerance showed evidence that they are different species. In contrast the *P. 'richtersi* group 1' and *P. 'richtersi* group 2' showed significant differences in the accumulation of osmolytes (Hengherr et al. 2008), but this and the phylogenetic analysis in Guidetti et al. (2009) is not sufficient to separate both as individual species. The two groups seem to be closely related ecotypes within one new species. Therefore, we described three new tardigrade species and named them according to their sample location.

Summary

The phylum Tardigrada consists of about 1000 described species to date. The animals live in habitats within marine, freshwater and terrestrial ecosystems all-over the world. Tardigrades are polyextremophiles. They are capable to resist extreme temperature, pressure or radiation. In the event of desiccation, tardigrades enter a so-called tun stage. The reason for their great tolerance capabilities against extreme environmental conditions is not discovered yet. Our FUNCRYPTA project aims at finding answers to the question what mechanisms underlie these adaption capabilities particularly with regard to the species *Milnesium tardigradum*.

The first part of this thesis describes the establishment of expressed sequence tags (ESTs) libraries for different stages of *M. tardigradum*. From proteomics data we bioinformatically identified 144 proteins with a known function and additionally 36 proteins which seemed to be specific for *M. tardigradum*. The generation of a comprehensive web-based database allows us to merge the proteome and transcriptome data. Therefore we created an annotation pipeline for the functional annotation of the protein and nucleotide sequences. Additionally, we clustered the obtained proteome dataset and identified some tardigrade-specific proteins (TSPs) which did not show homology to known proteins. Moreover, we examined the heat shock proteins of *M. tardigradum* and their different expression levels depending on the actual state of the animals. In further bioinformatical analyses of the whole data set, we discovered promising proteins and pathways which are described to be correlated with the stress tolerance, e.g. late embryogenesis abundant (LEA) proteins. Besides, we compared the tardigrades with nematodes, rotifers, yeast and man to identify shared and tardigrade specific stress pathways. An analysis of the 5′ and 3′ untranslated regions (UTRs) demonstrates a strong usage of stabilising motifs like the 15-lipoxygenase differentiation control element (15-LOX-DICE) but also reveals a lack of other common UTR motifs normally used, e.g. AU rich elements.

The second part of this thesis focuses on the relatedness between several cryptic species within the tardigrade genus *Paramacrobiotus*. Therefore for the first time, we used the sequence-structure information of the internal transcribed spacer 2 (ITS2) as a phylogenetic marker in tardigrades. This allowed the description of three new species which were indistinguishable using morphological characters or common molecular markers like the 18S ribosomal ribonucleic acid (rRNA) or the Cytochrome c oxidase subunit I (COI). In a large *in silico* simulation study we also

succeeded to show the benefit for the phylogenetic tree reconstruction by adding structure information to the ITS2 sequence. Next to the genus *Paramacrobiotus* we used the ITS2 to corroborate a monophyletic DO-group (Sphaeropleales) within the Chlorophyceae. Additionally we redesigned another comprehensive database—the ITS2 database resulting in a doubled number of sequence-structure pairs of the ITS2.

In conclusion, this thesis shows the first insights (6 first author publications and 4 coauthor publications) into the reasons for the enormous adaption capabilities of tardigrades and offers a solution to the debate on the phylogenetic relatedness within the tardigrade genus *Paramacrobiotus*.

Zusammenfassung

Der Tierstamm Tardigrada besteht aus derzeitig etwa 1000 beschriebenen Arten. Die Tiere leben in Habitaten in marinen, limnischen und terrestrischen Ökosystemen auf der ganzen Welt. Tardigraden sind polyextremophil. Sie können extremer Temperatur, Druck oder Strahlung widerstehen. Beim Austrocknen bilden sie ein so genanntes Tönnchenstadium. Der Grund für die hohe Toleranz gegenüber extremen Umweltbedingungen ist bis jetzt nicht aufgeklärt worden. Unser FUNCRYPTA Projekt versucht Antworten darauf zu finden, was die hinter dieser Anpassungsfähigkeit liegenden Mechanismen sind. Dabei steht die Art *Milnesium tardigradum* im Mittelpunkt.

Der erste Teil dieser Arbeit beschreibt die Etablierung einer *expressed sequence tags* (ESTs) Bibliothek für verschiedene Stadien von *M. tardigradum*. Aus unseren Proteomansatz konnten wir bislang 144 Proteine bioinformatisch identifizieren, denen eine Funktion zugeordnet werden konnte. Darüber hinaus wurden 36 Proteine gefunden, welche spezifisch für *M. tardigradum* zu sein scheinen. Die Erstellung einer umfassenden internetbasierenden Datenbank erlaubt uns die Verknüpfung der Proteom- und Transkriptomdaten. Dafür wurde eine Annotations-Pipeline erstellt um den Sequenzen Funktionen zuordnen zu können. Außerdem wurden die erhaltenen Proteindaten von uns geclustert. Dabei konnten wir einige Tardigraden-spezifische Proteine (*tardigrade-specific protein*, TSP) identifizieren die keinerlei Homologie zu bekannten Proteinen zeigen. Außerdem untersuchten wir die Hitze-Schock-Proteine von *M. tardigradum* und deren differenzielle Expression in Abhängigkeit vom Stadium der Tiere. In weiteren bioinformatischen Analysen konnten wir viel versprechende Proteine und Stoffwechselwege entdecken für die beschrieben ist, dass sie mit Stressreaktionen in Verbindung stehen, beispielsweise *late embryogenesis abundant* (LEA) Proteine. Des Weiteren verglichen wir Tardigraden mit Nematoden, Rotatorien, Hefe und dem Menschen, um gemeinsame und Tardigraden-spezifische Stoffwechselwege identifizieren zu können. Analysen der 5′ und 3′ untranslatierten Bereiche zeigen eine verstärkte Nutzung von stabilisierenden Motiven, wie dem *15-lipoxygenase differentiation control element* (LEA). Im Gegensatz dazu werden häufig benutzte Motive, wie beispielsweise AU-reiche Bereiche, gar nicht gefunden.

Der zweite Teil der Doktorarbeit beschäftigt sich mit den Verwandtschaftsverhältnissen einiger kryptischer Arten in der Tardigradengattung *Paramacrobiotus*. Hierfür haben wir, zum ersten Mal in Tardigraden, die Sequenz-Struktur-Informationen

der *internal transcribed spacer* 2 Region als phylogenetischen Marker verwendet. Dies erlaubte uns die Beschreibung von drei neuen Arten, welche mit klassischen morphologischen Merkmalen oder anderen molekularen Markern wie 18S ribosomaler RNA oder *Cytochrome c oxidase subunit I* (COI) nicht unterschieden werden konnten. In einer umfangreichen *in silico* Simulationsstudie zeigten wir den Vorteil der bei der Rekonstruktion phylogenetischer Bäume unter der Hinzunahme der Strukturinformationen zur Sequenz der ITS2 entsteht. ITS2 Sequenz-Struktur-Informationen wurden außerdem auch dazu benutzt, eine monophyletische DO-Gruppe (Sphaeropleales) in den Chlorophyceae zu bestätigen. Zusätzlich haben wir eine umfassende Datenbank, die ITS2-Datenbank, überarbeitet. Dadurch konnten die Sequenz-Struktur-Informationen verdoppelt werden, die in dieser Datenbank verfügbar sind.

Die vorliegende Doktorarbeit zeigt erste Einblicke (6 Erstautor- und 4 Koautor-Publikationen) in die Ursachen für die hervorragende Anpassungsfähigkeit der Tardigraden und beschreibt die erfolgreiche Aufklärung der Verwandtschaftsverhältnisse in der Tardigradengattung *Paramacrobiotus*.

Bibliography

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). 'Base-calling of automated sequencer traces using phred. I. Accuracy assessment'. *Genome Res* 8.3 (Mar. 1998), pp. 175–185.

Ewing, B. and Green, P. (1998). 'Base-calling of automated sequencer traces using phred. II. Error probabilities'. *Genome Res* 8.3 (Mar. 1998), pp. 186–194.

Adhikari, B. N., Wall, D. H., and Adams, B. J. (2009). 'Desiccation survival in an Antarctic nematode: molecular analysis using expressed sequenced tags'. *BMC Genom* 10, p. 69. DOI: 10.1186/1471-2164-10-69. URL: http://dx.doi.org/10.1186/1471-2164-10-69.

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., et al. (1997). 'Evidence for a clade of nematodes, arthropods and other moulting animals'. *Nature* 387.6632 (May 1997), pp. 489–493. DOI: 10.1038/387489a0. URL: http://dx.doi.org/10.1038/387489a0.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). 'Basic local alignment search tool'. *J Mol Biol* 215.3 (Oct. 1990), pp. 403–410. DOI: 10.1006/jmbi.1990.9999. URL: http://dx.doi.org/10.1006/jmbi.1990.9999.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., et al. (1997). 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs'. *Nucleic Acids Res* 25.17 (Sept. 1997), pp. 3389–3402.

Ammermann, D. (1967). 'The cytology of parthenogenesis in the tardigrade *Hypsibius dujardini*'. *Chromosoma* 23.2, pp. 203–213.

Bahrndorff, S., Tunnacliffe, A., Wise, M. J., McGee, B., Holmstrup, M., and Loeschcke, V. (2009). 'Bioinformatics and protein expression analyses implicate LEA proteins in the drought response of Collembola'. *J Insect Physiol* 55.3 (Mar. 2009), pp. 210–217. DOI: 10.1016/j.jinsphys.2008.11.010. URL: http://dx.doi.org/10.1016/j.jinsphys.2008.11.010.

Baker, W., Broek, A. v. d., Camon, E., Hingamp, P., Sterk, P., et al. (2000). 'The EMBL nucleotide sequence database'. *Nucleic Acids Res* 28.1 (Jan. 2000), pp. 19–23.

Baumann, H. (1922). 'Die Anabiose der Tardigraden'. *Zool Jahrb* 45, pp. 501–556.

Baumann, H. (1927). 'Bemerkungen zur Anabiose von Tardigraden'. *Zool Anz* 72, pp. 175–179.

Baumann, H. (1964). 'Über den Lebenslauf und die Lebensweise von *Milnesium tardigradum* Doyère (Tardigrada)'. *VeröEf. Überseemus. Bremen* 8, pp. 161–171.

Bavan, S., Straub, V. A., Blaxter, M. L., and Ennion, S. J. (2009). 'A P2X receptor from the tardigrade species *Hypsibius dujardini* with fast kinetics and sensitivity to zinc and copper'. *BMC Evol Biol* 9, p. 17. DOI: 10.1186/1471-2148-9-17. URL: http://dx.doi.org/10.1186/1471-2148-9-17.

Becquerel, P. (1950). 'La suspension de la vie au dessous de 1/20 K absolu par demagnetization adiabatique de l'alun de fer dans le vide les plus eléve'. *CR Hebd. Séances Acad. Sci* 231, pp. 261–263.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., and Wheeler, D. L. (2000). 'GenBank'. *Nucleic Acids Res* 28.1 (Jan. 2000), pp. 15–18.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2009). 'GenBank'. *Nucleic Acids Res* 37.Database issue (Jan. 2009), pp. D26–D31. DOI: 10.1093/nar/gkn723. URL: http://dx.doi.org/10.1093/nar/gkn723.

Bremer, B., Jansen, R. K., Oxelman, B., Backlund, M., Lantz, H., and Kim, K. J. (1999). 'More characters or more taxa for a robust phylogeny–case study from the coffee family (Rubiaceae)'. *Syst Biol* 48.3 (Sept. 1999), pp. 413–435.

Browne, J., Tunnacliffe, A., and Burnell, A. (2002). 'Anhydrobiosis: plant desiccation gene found in a nematode'. *Nature* 416.6876 (Mar. 2002), p. 38. DOI: 10.1038/416038a. URL: http://dx.doi.org/10.1038/416038a.

Browne, J. A., Dolan, K. M., Tyson, T., Goyal, K., Tunnacliffe, A., and Burnell, A. M. (2004). 'Dehydration-specific induction of hydrophilic protein genes in the anhydrobiotic nematode *Aphelenchus avenae*'. *Eukaryot Cell* 3.4 (Aug. 2004), pp. 966–975. DOI: 10.1128/EC.3.4.966-975.2004. URL: http://dx.doi.org/10.1128/EC.3.4.966-975.2004.

Buchheim, M., Michalopulos, E., and Buchheim, J. (2001). 'Phylogeny of the Chlorophyceae with special reference to the Sphaeropleales: a study of 18S and 26S rDNA data'. *Journal of Phycology* 37.5, pp. 819–835.

Byun, Y. and Han, K. (2009). 'PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots'. *Bioinformatics* 25.11 (June 2009), pp. 1435–1437. DOI: 10.1093/bioinformatics/btp252. URL: http://dx.doi.org/10.1093/bioinformatics/btp252.

Chen, C. Y. and Shyu, A. B. (1995). 'AU-rich elements: characterization and importance in mRNA degradation'. *Trends Biochem Sci* 20.11 (Nov. 1995), pp. 465–470.

Clegg, J. S., Jackson, S. A., Liang, P., and MacRae, T. H. (1995). 'Nuclear-cytoplasmic translocations of protein p26 during aerobic-anoxic transitions in embryos of *Artemia franciscana*'. *Exp Cell Res* 219.1 (July 1995), pp. 1–7. DOI: 10.1006/excr.1995.1197. URL: http://dx.doi.org/10.1006/excr.1995.1197.

Coe, N. R. and Bernlohr, D. A. (1998). 'Physiological properties and functions of intracellular fatty acid-binding proteins'. *Biochim Biophys Acta* 1391.3 (Apr. 1998), pp. 287–306.

Coleman, A. W. and Mai, J. C. (1997). 'Ribosomal DNA ITS-1 and ITS-2 sequence comparisons as a tool for predicting genetic relatedness'. *J Mol Evol* 45.2 (Aug. 1997), pp. 168–177.

Coleman, A. W. (2003). 'ITS2 is a double-edged tool for eukaryote evolutionary comparisons'. *Trends Genet* 19.7 (July 2003), pp. 370–375.

Coleman, A. W. (2007). 'Pan-eukaryote ITS2 homologies revealed by RNA secondary structure'. *Nucleic Acids Res* 35.10, pp. 3322–3329. DOI: 10.1093/nar/gkm233. URL: http://dx.doi.org/10.1093/nar/gkm233.

Deason, T., Silva, P., Watanabe, S., and Floyd, G. (1991). 'Taxonomic status of the species of the green algal genusNeochloris'. *Plant Systematics and Evolution* 177.3, pp. 213–219.

Degma, P. and Guidetti, R. (2007). 'Notes to the current checklist of Tardigrada'. *Zootaxa* 1579, pp. 41–53. ISSN: 1175-5326.

Degma, P., Bertolani, R., and Guidetti, R. (2010). *Actual checklist of Tardigrada species*. URL: http://www.tardigrada.modena.unimo.it/miscellanea/Actual%20checklist%20of%20Tardigrada.pdf.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge: Cambridge University Press. ISBN: 9780521629713.

Erixon, P., Svennblad, B., Britton, T., and Oxelman, B. (2003). 'Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics'. *Syst Biol* 52.5 (Oct. 2003), pp. 665–673.

Felsenstein, J. (2005). *PHYLIP (Phylogeny Inference Package) version 3.6*. Department of Genome Sciences, University of Washington, Seattle.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., et al. (2010). 'The Pfam protein families database'. *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D211–D222. DOI: 10.1093/nar/gkp985. URL: http://dx.doi.org/10.1093/nar/gkp985.

Förster, F., Liang, C., Shkumatov, A., Beisser, D., Engelmann, J. C., et al. (2009). 'Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades'. *BMC Genom* 10, p. 469. DOI: 10.1186/1471-2164-10-469. URL: http://dx.doi.org/10.1186/1471-2164-10-469.

Förster, F., Beisser, D., Frohme, M., Schill, R. O., and Dandekar, T. (2010*a*). 'Tardigrade bioinformatics: Molecular adapations, DNA j-family and dynamical modelling'. *J Zool Systemat Evol Res* conditionally accepted.

Förster, F., Beisser, D., Grohme, M., Liang, C., Mali, B., et al. (2010*b*). 'Transcriptome analyzed at different levels in *Hypsibius dujardini* and *Milnesium tardigradum*: Specific adaptations, motifs and clusters as well as general protective pathways'. *Genome Biol* in preparation.

França, M. B., Panek, A. D., and Eleutherio, E. C. A. (2007). 'Oxidative stress and its effects during dehydration'. *Comp Biochem Physiol A Mol Integr Physiol*

146.4 (Apr. 2007), pp. 621–631. DOI: 10.1016/j.cbpa.2006.02.030. URL: http://dx.doi.org/10.1016/j.cbpa.2006.02.030.

Frickey, T. and Lupas, A. (2004). 'CLANS: a Java application for visualizing protein families based on pairwise similarity'. *Bioinformatics* 20.18 (Dec. 2004), pp. 3702–3704. DOI: 10.1093/bioinformatics/bth444. URL: http://dx.doi.org/10.1093/bioinformatics/bth444.

Gabriel, W. N. and Goldstein, B. (2007). 'Segmental expression of Pax3/7 and engrailed homologs in tardigrade development'. *Dev Genes Evol* 217.6 (June 2007), pp. 421–433. DOI: 10.1007/s00427-007-0152-5. URL: http://dx.doi.org/10.1007/s00427-007-0152-5.

Gabriel, W. N., McNuff, R., Patel, S. K., Gregory, T. R., Jeck, W. R., et al. (2007). 'The tardigrade *Hypsibius dujardini*, a new model for studying the evolution of development'. *Dev Biol* 312.2 (Dec. 2007), pp. 545–559. DOI: 10.1016/j.ydbio.2007.09.055. URL: http://dx.doi.org/10.1016/j.ydbio.2007.09.055.

Gal, T. Z., Glazer, I., and Koltai, H. (2004). 'An LEA group 3 family member is involved in survival of *C. elegans* during exposure to stress'. *FEBS Lett* 577.1-2 (Nov. 2004), pp. 21–26. DOI: 10.1016/j.febslet.2004.09.049. URL: http://dx.doi.org/10.1016/j.febslet.2004.09.049.

Gesell, T. and Haeseler, A. v. (2006). 'In silico sequence evolution with site-specific interactions along phylogenetic trees'. *Bioinformatics* 22.6 (Mar. 2006), pp. 716–722. DOI: 10.1093/bioinformatics/bti812. URL: http://dx.doi.org/10.1093/bioinformatics/bti812.

Goeze, J. A. E. (1773). *Herrn Karl Bonnets Abhandlungen aus der Insektologie*. Halle.

Gorodkin, J., Heyer, L., Brunak, S., and Storomo, G. (1997). 'Displaying the in formation contents of structural RNA alignments: the structure logos'. *Comput. Appl. Biosci.* 13.6, pp. 583–586. DOI: 10.1093/bioinformatics/13.6.583. eprint: http://bioinformatics.oxfordjournals.org/cgi/reprint/13/6/583.pdf. URL: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/13/6/583.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., et al. (2008). 'High-throughput functional annotation and data mining with the Blast2GO suite'. *Nucleic Acids Res* 36.10 (June 2008), pp. 3420–3435. DOI: 10.1093/nar/gkn176. URL: http://dx.doi.org/10.1093/nar/gkn176.

Goyal, K., Walton, L. J., and Tunnacliffe, A. (2005). 'LEA proteins prevent protein aggregation due to water stress'. *Biochem J* 388.Pt 1 (May 2005), pp. 151–157. DOI: 10.1042/BJ20041931. URL: http://dx.doi.org/10.1042/BJ20041931.

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., et al. (2007). 'Eukaryotic genome size databases'. *Nucleic Acids Res* 35.Database issue (Jan. 2007), pp. D332–D338. DOI: 10.1093/nar/gkl828. URL: http://dx.doi.org/10.1093/nar/gkl828.

Grillo, G., Turi, A., Licciulli, F., Mignone, F., Liuni, S., et al. (2010). 'UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs'. *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D75–D80. DOI: 10.1093/nar/gkp902. URL: http://dx.doi.org/10.1093/nar/gkp902.

Guidetti, R and Bertolani, R (2005). 'Tardigrade taxonomy: an updated check list of the taxa and a list of characters for their identification'. *Zootaxa* 845 (Feb. 2005), pp. 1–46. ISSN: 1175-5334.

Guidetti, R., Schill, R. O., Bertolani, R., Dandekar, T., and Wolf, M. (2009). 'New molecular data for tardigrade phylogeny, with the erection of *Paramacrobiotus* gen. nov.' *J Zool Systemat Evol Res* 47.4, pp. 315–321.

Hand, S. C., Jones, D., Menze, M. A., and Witt, T. L. (2007). 'Life without water: expression of plant LEA genes by an anhydrobiotic arthropod'. *J Exp Zool A Ecol Genet Physiol* 307.1 (Jan. 2007), pp. 62–66. DOI: 10.1002/jez.a.343. URL: http://dx.doi.org/10.1002/jez.a.343.

Haunerland, N. H. and Spener, F. (2004). 'Fatty acid-binding proteins–insights from genetic manipulations'. *Prog Lipid Res* 43.4 (July 2004), pp. 328–349. DOI: 10.1016/j.plipres.2004.05.001. URL: http://dx.doi.org/10.1016/j.plipres.2004.05.001.

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'. *J Exp Biol* 212.Pt 6 (Mar. 2009), pp. 802–807. DOI: 10.1242/jeb.025973. URL: http://dx.doi.org/10.1242/jeb.025973.

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'. *Physiol Biochem Zool* 82.6, pp. 749–755. DOI: 10.1086/605954. URL: http://dx.doi.org/10.1086/605954.

Hengherr, S., Heyer, A. G., Köhler, H.-R., and Schill, R. O. (2008). 'Trehalose and anhydrobiosis in tardigrades–evidence for divergence in responses to dehydration'. *FEBS J* 275.2 (Jan. 2008), pp. 281–288. DOI: 10.1111/j.1742-4658.2007.06198.x. URL: http://dx.doi.org/10.1111/j.1742-4658.2007.06198.x.

Hershkovitz, M. A. and Lewis, L. A. (1996). 'Deep-level diagnostic value of the rDNA-ITS region'. *Mol Biol Evol* 13.9 (Nov. 1996), pp. 1276–1295.

Hittel, D. and Storey, K. B. (2001). 'Differential expression of adipose- and heart-type fatty acid binding proteins in hibernating ground squirrels'. *Biochim Biophys Acta* 1522.3 (Dec. 2001), pp. 238–243.

Horikawa, D. D. and Higashi, S. (2004). 'Desiccation tolerance of the tardigrade *Milnesium tardigradum* collected in Sapporo, Japan, and Bogor, Indonesia'. *Zoolog Sci* 21.8 (Aug. 2004), pp. 813–816.

Horikawa, D. D., Sakashita, T., Katagiri, C., Watanabe, M., Kikawada, T., et al. (2006). 'Radiation tolerance in the tardigrade *Milnesium tardigradum*'. *Int J Radiat Biol* 82.12 (Dec. 2006), pp. 843–848. DOI: 10.1080/09553000600972956. URL: http://dx.doi.org/10.1080/09553000600972956.

Hoshina, R., Hayashi, S., and Imamura, N. (2006). 'Intraspecific Genetic Divergence of Paramecium bursariaand Re-construc-tion of the Paramecian Phylogenetic Tree'. *Acta Protozool* 45, pp. 377–386.

Huang, X. and Madan, A. (1999). 'CAP3: A DNA sequence assembly program'. *Genome Res* 9.9 (Sept. 1999), pp. 868–877.

Izumi, Y., Sonoda, S., Yoshida, H., Danks, H. V., and Tsumuki, H. (2006). 'Role of membrane transport of water and glycerol in the freeze tolerance of the rice stem borer, *Chilo suppressalis* Walker (Lepidoptera: Pyralidae)'. *J Insect Physiol* 52.2 (Feb. 2006), pp. 215–220. DOI: 10.1016/j.jinsphys.2005.11.001. URL: http://dx.doi.org/10.1016/j.jinsphys.2005.11.001.

Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., et al. (2009). 'Infrastructure for the life sciences: design and implementation of the UniProt website'. *BMC Bioinformatics* 10, p. 136. DOI: 10.1186/1471-2105-10-136. URL: http://dx.doi.org/10.1186/1471-2105-10-136.

Jønsson, K. I., Harms-Ringdahl, M., and Torudd, J. (2005). 'Radiation tolerance in the eutardigrade *Richtersius coronifer*'. *Int J Radiat Biol* 81.9 (Sept. 2005), pp. 649–656. DOI: 10.1080/09553000500368453. URL: http://dx.doi.org/10.1080/09553000500368453.

Jønsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M., and Rettberg, P. (2008). 'Tardigrades survive exposure to space in low Earth orbit'. *Curr Biol* 18.17 (Sept. 2008), R729–R731. DOI: 10.1016/j.cub.2008.06.048. URL: http://dx.doi.org/10.1016/j.cub.2008.06.048.

Kaminuma, E., Mashima, J., Kodama, Y., Gojobori, T., Ogasawara, O., et al. (2010). 'DDBJ launches a new archive database with analytical tools for next-generation sequence data'. *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D33–D38. DOI: 10.1093/nar/gkp847. URL: http://dx.doi.org/10.1093/nar/gkp847.

Keilin, D. (1959). 'The Leeuwenhoek Lecture: The Problem of Anabiosis or Latent Life: History and Current Concept'. *Proceedings of the Royal Society of London. Series B - Biological Sciences* 150.939, pp. 149–191. DOI: 10.1098/rspb.1959.0013. eprint: http://rspb.royalsocietypublishing.org/content/150/939/149.full.pdf+html. URL: http://rspb.royalsocietypublishing.org/content/150/939/149.short.

Keller, A., Schleicher, T., Förster, F., Ruderisch, B., Dandekar, T., et al. (2008). 'ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales)'. *BMC Evol Biol* 8, p. 218. DOI: 10.1186/1471-2148-8-218. URL: http://dx.doi.org/10.1186/1471-2148-8-218.

Keller, A., Schleicher, T., Schultz, J., Müller, T., Dandekar, T., and Wolf, M. (2009). '5.8S-28S rRNA interaction and HMM-based ITS2 annotation'. *Gene* 430.1-2 (Feb. 2009), pp. 50–57. DOI: 10.1016/j.gene.2008.10.012. URL: http://dx.doi.org/10.1016/j.gene.2008.10.012.

Keller, A., Förster, F., Müller, T., Dandekar, T., Schultz, J., and Wolf, M. (2010). 'Including RNA Secondary Structures improves Accuracy and Robustness in Reconstruction of Phylogenetic Trees'. *Biol Direct* 5.1 (Jan. 2010), p. 4. DOI: 10.1186/1745-6150-5-4. URL: http://dx.doi.org/10.1186/1745-6150-5-4.

Kikawada, T., Nakahara, Y., Kanamori, Y., Iwata, K. i., Watanabe, M., et al. (2006). 'Dehydration-induced expression of LEA proteins in an anhydrobiotic chironomid'. *Biochem Biophys Res Commun* 348.1 (Sept. 2006), pp. 56–61. DOI: 10.1016/j.bbrc.2006.07.003. URL: http://dx.doi.org/10.1016/j.bbrc.2006.07.003.

Kikawada, T., Saito, A., Kanamori, Y., Fujita, M., Snigórska, K., et al. (2008). 'Dehydration-inducible changes in expression of two aquaporins in the sleeping chironomid, *Polypedilum vanderplanki*'. *Biochim Biophys Acta* 1778.2 (Feb. 2008), pp. 514–520. DOI: 10.1016/j.bbamem.2007.11.009. URL: http://dx.doi.org/10.1016/j.bbamem.2007.11.009.

Kinchin, I. (1994). *The biology of tardigrades*. Portland Pr.

Kitts, P., Madden, T., Sicotte, H., Black, L., and Ostell, J. (2009). 'UniVec'. *unpublished*. URL: http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html.

Koetschan, C., Förster, F., Keller, A., Schleicher, T., Ruderisch, B., et al. (2010). 'The ITS2 Database III–sequences and structures for phylogeny'. *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D275–D279. DOI: 10.1093/nar/gkp966. URL: http://dx.doi.org/10.1093/nar/gkp966.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994). 'Hidden Markov models in computational biology. Applications to protein modeling'. *J Mol Biol* 235.5 (Feb. 1994), pp. 1501–1531. DOI: 10.1006/jmbi.1994.1104. URL: http://dx.doi.org/10.1006/jmbi.1994.1104.

Lai, E. C. (2002). 'Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation'. *Nat Genet* 30.4 (Apr. 2002), pp. 363–364. DOI: 10.1038/ng865. URL: http://dx.doi.org/10.1038/ng865.

Lai, E. C., Tam, B., and Rubin, G. M. (2005). 'Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs'. *Genes Dev* 19.9 (May 2005), pp. 1067–1080. DOI: 10.1101/gad.1291905. URL: http://dx.doi.org/10.1101/gad.1291905.

Leinonen, R., Akhtar, R., Birney, E., Bonfield, J., Bower, L., et al. (2010). 'Improvements to services at the European Nucleotide Archive'. *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D39–D45. DOI: 10.1093/nar/gkp998. URL: http://dx.doi.org/10.1093/nar/gkp998.

Letunic, I., Doerks, T., and Bork, P. (2009). 'SMART 6: recent updates and new developments'. *Nucleic Acids Res* 37.Database issue (Jan. 2009), pp. D229–D232. DOI: 10.1093/nar/gkn808. URL: http://dx.doi.org/10.1093/nar/gkn808.

Liang, P., Amons, R., Clegg, J. S., and MacRae, T. H. (1997*a*). 'Molecular characterization of a small heat shock/α-crystallin protein in encysted *Artemia* embryos'. *J Biol Chem* 272.30 (July 1997), pp. 19051–19058.

Liang, P., Amons, R., Macrae, T. H., and Clegg, J. S. (1997*b*). 'Purification, structure and in vitro molecular-chaperone activity of *Artemia* p26, a small heat-shock/α-crystallin protein'. *Eur J Biochem* 243.1-2 (Jan. 1997), pp. 225–232.

Lindblom, T. H. and Dodd, A. K. (2006). 'Xenobiotic detoxification in the nematode *Caenorhabditis elegans*'. *J Exp Zool A Comp Exp Biol* 305.9 (Sept. 2006), pp. 720–730. DOI: 10.1002/jez.a.324. URL: http://dx.doi.org/10.1002/jez.a.324.

Liu, N., Lamerdin, J. E., Tebbs, R. S., Schild, D., Tucker, J. D., et al. (1998). 'XRCC2 and XRCC3, new human Rad51-family members, promote chromosome stability and protect against DNA cross-links and other damages'. *Mol Cell* 1.6 (May 1998), pp. 783–793.

Mai, J. C. and Coleman, A. W. (1997). 'The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants'. *J Mol Evol* 44.3 (Mar. 1997), pp. 258–271.

Makowski, L. and Hotamisligil, G. S. (2004). 'Fatty acid binding proteins–the evolutionary crossroads of inflammatory and metabolic responses'. *J Nutr* 134.9 (Sept. 2004), 2464S–2468S.

Mali, B., Grohme, M. A., Förster, F., Schnölzer, T. D. M., Reuter, D., et al. (2010). 'Transcriptome survey of the anhydrobiotic tardigrade *Milnesium tardigradum* in comparison with *Hypsibius dujardini* and *Richtersius coronifer*'. *BMC Genom* 11.1 (Mar. 2010), p. 168. DOI: 10.1186/1471-2164-11-168. URL: http://dx.doi.org/10.1186/1471-2164-11-168.

Mallatt, J. and Giribet, G. (2006). 'Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch'. *Mol Phylogenet Evol* 40.3 (Sept. 2006), pp. 772–794. DOI: 10.1016/j.ympev.2006.04.021. URL: http://dx.doi.org/10.1016/j.ympev.2006.04.021.

Marcus, E. (1928). 'Zur Ökologie und Physiologie der Tardigraden'. *Zool. Jahrb. Allg. Zool* 44, pp. 323–370.

Marcus, E. and Dahl, F. (1928). *Spinnentiere oder Arachnoidea IV. Bärtierchen (Tardigrada)*. Urban & Fischer Bei Elsevier.

Markham, N. R. and Zuker, M. (2005). 'DINAMelt web server for nucleic acid melting prediction'. *Nucleic Acids Res* 33.Web Server issue (July 2005), W577–W581. DOI: 10.1093/nar/gki591. URL: http://dx.doi.org/10.1093/nar/gki591.

Markham, N. R. and Zuker, M. (2008). 'UNAFold: software for nucleic acid folding and hybridization'. *Methods Mol Biol* 453, pp. 3–31. DOI: 10.1007/978-1-60327-429-6_1. URL: http://dx.doi.org/10.1007/978-1-60327-429-6_1.

Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999). 'Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure'. *J Mol Biol* 288.5 (May 1999), pp. 911–940. DOI: 10.1006/jmbi.1999.2700. URL: http://dx.doi.org/10.1006/jmbi.1999.2700.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). 'Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure'. *Proc Natl Acad Sci U S A* 101.19 (May 2004), pp. 7287–7292. DOI: 10.1073/pnas.0401799101. URL: http://dx.doi.org/10.1073/pnas.0401799101.

May, R. M., Maria, M., and Guimard, J. (1964). 'Action différentielle des rayons x et ultraviolets sur le tardigrade *Macrobiotus areolatus*, a l'état actif et desséché'. *Bull. Biol. Fr. Belg* 98, pp. 349–367.

Müller, T. and Vingron, M. (2000). 'Modeling amino acid replacement'. *J Comput Biol* 7.6, pp. 761–776. DOI: 10.1089/10665270050514918. URL: http://dx.doi.org/10.1089/10665270050514918.

Müller, T., Philippi, N., Dandekar, T., Schultz, J., and Wolf, M. (2007). 'Distinguishing species'. *RNA* 13.9 (Sept. 2007), pp. 1469–1472. DOI: 10.1261/rna.617107. URL: http://dx.doi.org/10.1261/rna.617107.

Myers, E. W. and Miller, W. (1988). 'Optimal alignments in linear space'. *Comput Appl Biosci* 4.1 (Mar. 1988), pp. 11–17.

Needleman, S. B. and Wunsch, C. D. (1970). 'A general method applicable to the search for similarities in the amino acid sequence of two proteins'. *J Mol Biol* 48.3 (Mar. 1970), pp. 443–453.

Nelson, D. R. (2002). 'Current Status of the Tardigrada: Evolution and Ecology'. *Integr Comp Biol* 42, pp. 652–659. eprint: http://icb.oxfordjournals.org/cgi/reprint/42/3/652.pdf.

Ono, F., Saigusa, M., Uozumi, T., Matsushima, Y., Ikeda, H., et al. (2008). 'Effect of high hydrostatic pressure on to life of the tiny animal tardigrade'. *Journal of Physics and Chemistry of Solids* 69.9. Study of Matter Under Extreme Conditions 2007, Study of Matter Under Extreme Conditions 2007, pp. 2297 –2300. ISSN: 0022-3697. DOI: DOI:10.1016/j.jpcs.2008.04.019. URL: http://www.sciencedirect.com/science/article/B6TXR-4S85DV3-3/2/c80856b5c0d0c150c431b873552ec552.

Ostareck-Lederer, A., Ostareck, D. H., Standart, N., and Thiele, B. J. (1994). 'Translation of 15-lipoxygenase mRNA is inhibited by a protein that binds to a repeated sequence in the 3' untranslated region'. *EMBO J* 13.6 (Mar. 1994), pp. 1476–1481.

Padilla, P. A., Nystul, T. G., Zager, R. A., Johnson, A. C. M., and Roth, M. B. (2002). 'Dephosphorylation of cell cycle-regulated proteins correlates with anoxia-induced suspended animation in *Caenorhabditis elegans*'. *Mol Biol Cell* 13.5 (May 2002), pp. 1473–1483. DOI: 10.1091/mbc.01-12-0594. URL: http://dx.doi.org/10.1091/mbc.01-12-0594.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., et al. (2003). 'TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets'. *Bioinformatics* 19.5 (Mar. 2003), pp. 651–652.

Philip, B. N., Yi, S.-X., Elnitsky, M. A., and Lee, R. E. (2008). 'Aquaporins play a role in desiccation and freeze tolerance in larvae of the goldenrod gall fly, *Eurosta solidaginis*'. *J Exp Biol* 211.Pt 7 (Apr. 2008), pp. 1114–1119. DOI: 10.1242/jeb.016758. URL: http://dx.doi.org/10.1242/jeb.016758.

Pigon, A. and Weglarska, B. (1955). 'Rate of metabolism in tardigrades during active life and anabiosis'. *Nature* 176.4472 (July 1955), pp. 121–122.

Pouchkina-Stantcheva, N. N., McGee, B. M., Boschetti, C., Tolleter, D., Chakrabortee, S., et al. (2007). 'Functional divergence of former alleles in an ancient asexual invertebrate'. *Science* 318.5848 (Oct. 2007), pp. 268–271. DOI: 10.1126/science.1144363. URL: http://dx.doi.org/10.1126/science.1144363.

Qiu, X.-B., Shao, Y.-M., Miao, S., and Wang, L. (2006). 'The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones'. *Cell Mol Life Sci* 63.22 (Nov. 2006), pp. 2560–2570. DOI: 10.1007/s00018-006-6192-6. URL: http://dx.doi.org/10.1007/s00018-006-6192-6.

Rahm, G. (1921). 'Biologische und physiologische Beiträge zur Kenntnis der Moosfauna'. *Zeitschr. allg. Physiol.* 20.1, pp. 1–34.

Rahm, G. (1937). 'A new order of tardigrades from the hot springs of Japan (Furu-Section, Unzen)'. *Annot Zool Jpn* 16, pp. 345–352.

Ramazzotti, G. and Maucci, W. (1983). 'Il phylum Tardigrada'. *Mem. Ist. Ital. Idrobiol* 41, pp. 1–1012.

Ramløv, H. and Westh, P. (1992). 'Survival of the cryptobiotic eutardigrade *Adorybiotus coronifer* during cooling to $-196\,°C$: effect of cooling rate, trehalose level, and short-term acclimation'. *Cryobiology(Print)* 29.1, pp. 125–130.

Ramløv, H. and Westh, P. (2001). 'Cryptobiosis in the Eutardigrade *Adorybiotus (Richtersius) coronifer*: Tolerance to Alcohols, Temperature and *de novo* Protein Synthesis'. *Zoologischer Anzeiger - A Journal of Comparative Zoology* 240.3-4, pp. 517–523. ISSN: 0044-5231. DOI: DOI:10.1078/0044-5231-00062. URL: http://www.sciencedirect.com/science/article/B7GJF-4DR7H96-1W/2/b78d38605ceeea0530be546cadf05c90.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). 'Automatic clustering of orthologs and in-paralogs from pairwise species comparisons'. *J Mol Biol* 314.5 (Dec. 2001), pp. 1041–1052. DOI: 10.1006/jmbi.2000.5197. URL: http://dx.doi.org/10.1006/jmbi.2000.5197.

Reuner, A., Bruemmer, F., Schill, R. O., Zantke, J., Kube, M., et al. (2008). 'Mechanisms and molecular adaptation to extreme dehydration in tardigrades: Hsp gene expression in *Milnesium tardigradum*'. *COMPARATIVE BIOCHEMISTRY AND PHYSIOLOGY A-MOLECULAR & INTEGRATIVE PHYSIOLOGY* 151.1, Suppl. S (Sept. 2008), S32. ISSN: 1095-6433. DOI: {10.1016/j.cbpa.2008.05.114}.

Reuner, A., Hengherr, S., Mali, B., Förster, F., Arndt, D., et al. (2009). 'Stress response in tardigrades: differential gene expression of molecular chaperones'. *Cell Stress & Chaperones* (Nov. 2009). DOI: 10.1007/s12192-009-0158-1. URL: http://dx.doi.org/10.1007/s12192-009-0158-1.

Rice, P., Longden, I., and Bleasby, A. (2000). 'EMBOSS: the European Molecular Biology Open Software Suite'. *Trends Genet* 16.6 (June 2000), pp. 276–277.

Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., Haeseler, A. v., et al. (2007). 'EST sequencing of Onychophora and phylogenomic analysis of Metazoa'. *Mol Phylogenet Evol* 45.3 (Dec. 2007), pp. 942–951. DOI: 10.1016/j.ympev.2007.09.002. URL: http://dx.doi.org/10.1016/j.ympev.2007.09.002.

Rokas, A. and Carroll, S. B. (2005). 'More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy'. *Mol Biol Evol* 22.5 (May 2005), pp. 1337–1344. DOI: 10.1093/molbev/msi121. URL: http://dx.doi.org/10.1093/molbev/msi121.

Roxas, V. P., Smith, R. K., Allen, E. R., and Allen, R. D. (1997). 'Overexpression of glutathione S-transferase/glutathione peroxidase enhances the growth of transgenic tobacco seedlings during stress'. *Nat Biotechnol* 15.10 (Oct. 1997), pp. 988–991. DOI: 10.1038/nbt1097-988. URL: http://dx.doi.org/10.1038/nbt1097-988.

Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., et al. (2009). 'Database resources of the National Center for Biotechnology Information'. *Nucleic Acids Res* 37.Database issue (Jan. 2009), pp. D5–15. DOI: 10.1093/nar/gkn741. URL: http://dx.doi.org/10.1093/nar/gkn741.

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 (Apr. 2004), pp. 1607–1613.

Schill, R. O., Förster, F., Dandekar, T., and Wolf, M. (2010). 'Distinguishing species in *Paramacrobiotus* (Tardigrada) via compensatory base change analysis of internal transcribed spacer 2 secondary structures, with the description of three new species'. *Organisms Diversity & Evolution* in press.

Schokraie, E., Hotz-Wagenblatt, A., Warnken, U., Mali, B., MarcusFrohme, et al. (2010). 'Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms.' *PLoS One* 5.3 (Mar. 2010), e9502. DOI: 10.1371/journal.pone.0009502. URL: http://dx.doi.org/10.1371/journal.pone.0009502.

Schuetz, G. (1987). 'A one-year study on the population dynamics of *Milnesium tardigradum* Doyere in the lichen *Xanthoria parietina* (L.) Th. Fr.' In: *Biology of tardigrades: proceedings of the 4th International Symposium on the Tardigrada, Modena, September 3-5, 1985*. Mucchi, pp. 217–228.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). 'SMART, a simple modular architecture research tool: identification of signaling domains'. *Proc Natl Acad Sci U S A* 95.11 (May 1998), pp. 5857–5864.

Schultz, J. and Wolf, M. (2009). 'ITS2 sequence-structure analysis in phylogenetics: a how-to manual for molecular systematics'. *Mol Phylogenet Evol* 52.2 (Aug. 2009), pp. 520–523.

Schultz, J., Maisel, S., Gerlach, D., Müller, T., and Wolf, M. (2005). 'A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota'. *RNA* 11.4 (Apr. 2005), pp. 361–364. DOI: 10.1261/rna.7204505. URL: http://dx.doi.org/10.1261/rna.7204505.

Schultz, J., Müller, T., Achtziger, M., Seibel, P. N., Dandekar, T., and Wolf, M. (2006). 'The internal transcribed spacer 2 database–a web server for (not only) low level phylogenetic analyses'. *Nucleic Acids Res* 34.Web Server issue (July 2006), W704–W707. DOI: 10.1093/nar/gkl129. URL: http://dx.doi.org/10.1093/nar/gkl129.

Seibel, P. N., Müller, T., Dandekar, T., Schultz, J., and Wolf, M. (2006). '4SALE–a tool for synchronous RNA sequence and secondary structure alignment and editing'. *BMC Bioinformatics* 7, p. 498. DOI: 10.1186/1471-2105-7-498. URL: http://dx.doi.org/10.1186/1471-2105-7-498.

Selig, C., Wolf, M., Müller, T., Dandekar, T., and Schultz, J. (2008). 'The ITS2 Database II: homology modelling RNA structure for molecular systematics'. *Nucleic Acids Res* 36.Database issue (Jan. 2008), pp. D377–D380. DOI: 10.1093/nar/gkm827. URL: http://dx.doi.org/10.1093/nar/gkm827.

Smit, A. F. A., Hubley, R., and Green, P. (1996–2004). 'RepeatMasker Open-3.0'. URL: http://www.repeatmasker.org.

Smith, T. F. and Waterman, M. S. (1981). 'Identification of common molecular subsequences'. *J Mol Biol* 147.1 (Mar. 1981), pp. 195–197.

Spallanzani, L. (1776). *Opuscoli di fisica animale e vegetabile*. Vol. 2. Societla Tipografica.

Sun, Y. and MacRae, T. H. (2005). 'Small heat shock proteins: molecular structure and chaperone function'. *Cell Mol Life Sci* 62.21 (Nov. 2005), pp. 2460–2476. DOI: 10.1007/s00018-005-5190-4. URL: http://dx.doi.org/10.1007/s00018-005-5190-4.

Suzuki, A. C. (2003). 'Life history of *Milnesium tardigradum* Doyére (tardigrada) under a rearing environment'. *Zoolog Sci* 20.1 (Jan. 2003), pp. 49–57. URL: http://www.bioone.org/doi/full/10.2108/zsj.20.49.

Suzuki, A. C. (2008). 'Appearance of males in a thelytokous strain of *Milnesium* cf. *tardigradum* (Tardigrada)'. *Zoolog Sci* 25.8 (Aug. 2008), pp. 849–853. DOI: 10.2108/zsj.25.849. URL: http://dx.doi.org/10.2108/zsj.25.849.

Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H., and Gojobori, T. (2000). 'DNA data bank of Japan (DDBJ) in collaboration with mass sequencing teams'. *Nucleic Acids Res* 28.1 (Jan. 2000), pp. 24–26.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). 'A genomic perspective on protein families'. *Science* 278.5338 (Oct. 1997), pp. 631–637.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., et al. (2003). 'The COG database: an updated version includes eukaryotes'. *BMC Bioinformatics* 4 (Sept. 2003), p. 41. DOI: 10.1186/1471-2105-4-41. URL: http://dx.doi.org/ 10.1186/1471-2105-4-41.

Telford, M. J., Wise, M. J., and Gowri-Shankar, V. (2005). 'Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: examples from the bilateria'. *Mol Biol Evol* 22.4 (Apr. 2005), pp. 1129–1136. DOI: 10.1093/molbev/msi099. URL: http://dx.doi.org/10.1093/molbev/msi099.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). 'CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice'. *Nucleic Acids Res* 22.22 (Nov. 1994), pp. 4673–4680.

TIGR (2010). *SeqClean: a script for automated trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences.* URL: http://compbio.dfci.harvard.edu/tgi/software/.

Toffaletti, D. L., Poeta, M. D., Rude, T. H., Dietrich, F., and Perfect, J. R. (2003). 'Regulation of cytochrome c oxidase subunit 1 (COX1) expression in *Cryptococcus neoformans* by temperature and host environment'. *Microbiology* 149.Pt 4 (Apr. 2003), pp. 1041–1049.

Tsutsui, Y., Morishita, T., Iwasaki, H., Toh, H., and Shinagawa, H. (2000). 'A recombination repair gene of *Schizosaccharomyces pombe*, rhp57, is a functional homolog of the *Saccharomyces cerevisiae* RAD57 gene and is phylogenetically related to the human XRCC3 gene'. *Genetics* 154.4 (Apr. 2000), pp. 1451–1461.

Tunnacliffe, A. and Wise, M. J. (2007). 'The continuing conundrum of the LEA proteins'. *Naturwissenschaften* 94.10 (Oct. 2007), pp. 791–812. DOI: 10.1007/s00114-007-0254-y. URL: http://dx.doi.org/10.1007/s00114-007-0254-y.

UniProt Consortium (2010). 'The Universal Protein Resource (UniProt) in 2010'. *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D142–D148. DOI: 10.1093/ nar/gkp846. URL: http://dx.doi.org/10.1093/nar/gkp846.

Van Hannen, E. J., Fink, P., and Lüring, M. (2002). 'A revised secondary structure model for the internal transcribed spacer 2 of the green algae *Scenedesmus* and *Desmodesmus* and its implication for the phylogeny of these algae'. *European Journal of Phycology* 37.02, pp. 203–208. DOI: 10.1017/S096702620200361X. eprint: http://journals.cambridge.org/article_S096702620200361X. URL: http:// journals.cambridge.org/action/displayAbstract?fromPage=online&aid= 105707&fulltextType=RA&fileId=S096702620200361X.

Wernersson, R. (2006). 'Virtual Ribosome–a comprehensive DNA translation tool with support for integration of sequence feature annotation'. *Nucleic Acids Res* 34.Web Server issue (July 2006), W385–W388. DOI: 10.1093/nar/gkl252. URL: http://dx.doi.org/10.1093/nar/gkl252.

Westh, P. and Ramløv, H. (1991). 'Trehalose accumulation in the tardigrade Adorybiotus coronifer during anhydrobiosis'. *J. exp. Zool* 258, pp. 303–311.

White, T. J., Bruns, T., Lee, S., and Taylor, J. (1990). 'Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics'. In: *PCR Protocols: A Guide to Methods and Applications*. Ed. by Innis, M. A., Gelfand, D. H., Shinsky, J. J., and White, T. J. San Diego: Academic Press, pp. 315–322.

Wiemers, M., Keller, A., and Wolf, M. (2009). 'ITS2 secondary structure improves phylogeny estimation in a radiation of blue butterflies of the subgenus *Agrodiaetus* (Lepidoptera: Lycaenidae: *Polyommatus*)'. *BMC Evol Biol* 9, p. 300. DOI: 10.1186/1471-2148-9-300. URL: http://dx.doi.org/10.1186/1471-2148-9-300.

Willsie, J. K. and Clegg, J. S. (2002). 'Small heat shock protein p26 associates with nuclear lamins and HSP70 in nuclei and nuclear matrix fractions from stressed cells'. *J Cell Biochem* 84.3, pp. 601–614.

Wise, M. J. and Tunnacliffe, A. (2004). 'POPP the question: what do LEA proteins do?' *Trends Plant Sci* 9.1 (Jan. 2004), pp. 13–17.

Wolf, M., Buchheim, M., Hegewald, E., Krienitz, L., and Hepperle, D. (2002). 'Phylogenetic position of the Sphaeropleaceae (Chlorophyta)'. *Plant Systematics and Evolution* 230.3, pp. 161–171.

Wolf, M., Achtziger, M., Schultz, J., Dandekar, T., and Müller, T. (2005). 'Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures'. *RNA* 11.11 (Nov. 2005), pp. 1616–1623. DOI: 10.1261/rna.2144205. URL: http://dx.doi.org/10.1261/rna.2144205.

Wolf, M., Ruderisch, B., Dandekar, T., Schultz, J., and Müller, T. (2008). 'ProfDistS: (profile-) distance based phylogeny on sequence–structure alignments'. *Bioinformatics* 24.20 (Oct. 2008), pp. 2401–2402. DOI: 10.1093/bioinformatics/btn453. URL: http://dx.doi.org/10.1093/bioinformatics/btn453.

Contributions

Chapter 3
'Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades'
I am an equal contributing first author of this publication.

I did the EST annotation. For that, I developed an annotation pipeline conducted the translation to the hypothetical protein data and performed the CLANS clustering based on the EST sequences and the search for RNA motifs. Chunguang Liang established the current version of the tardigrade workbench including programming new routines, data management and nucleotide motif analysis. Alexander Shkumatov did the initial setup of the server, of the virtual ribosome and the CLANS clustering. Daniela Beisser, Julia C. Engelmann, Martina Schnölzer and Marcus Frohme participated in tardigrade data analysis. Tobias Müller gave expert advice and input on statistics. Ralph O. Schill gave expert advice on tardigrade physiology and zoology. Thomas Dandekar led and guided the study including the analysis of data and program, supervision, and manuscript writing. All authors participated in the writing of the manuscript and approved the final version.

Chapter 4
'Transcriptome survey of the anhydrobiotic tardigrade *Milnesium tardigradum* in comparison with *Hypsibius dujardini* and *Richtersius coronifer*'
I participated as coauthor to this publication.

Brahim Mali established and optimised the tardigrade RNA extraction protocol and constructed and managed the cDNA clone libraries. Markus Grohme performed functional annotation and enrichment analysis, putative orthologue prediction and gave useful comments on sequence analysis. Marcus Frohme was responsible for supervision, budget, obtaining the funding for the project, and contributed advice at each step of the research. I performed quality control, processing and assembly of the ESTs and was involved in the data analysis. Thomas Dandekar contributed to the bioinformatic analysis. Weronika Welnicz performed the phylogenetic analysis, Ralph O. Schill provided the animals and coordinated the project and contributed comments on the candidate anhydrobiotic genes. Martina Schnölzer and Dirk Reuter supported the identification of the anhydrobiotic genes. Brahim Mali and

Markus Grohme wrote the main part of the manuscript. All authors read and approved the final manuscript.

Chapter 5
'Transcriptome analyzed at different levels in *Hypsibius dujardini* and *Milnesium tardigradum*: Specific adaptations, motifs and clusters as well as general protective pathways'
I am an equal contributing first author of this publication.

I and Markus Grohme prepared the EST databases for *M. tardigradum*, *H. dujardini* and *R. coronifer*. Moreover, I investigated the sequences for RNA stability motifs. Next, I clustered the *M. tardigradum* and *H. dujardini* ESTs using INPARANOID and CLANS to find orthologs. In addition, I searched the database for LEA proteins and reconstructed the tree. I was involved in the analysis of the stress pathways including PCRs. I participated in writing the manuscript. All authors approved the final version.

Chapter 6
'Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms'
I participated as coauthor to this publication.

I generated the tardigrade EST database including the deletion of vector sequences and sequences with low quality. I also performed repeat masking, assembly, remove of remaining contaminations, annotation and translation. I added minor parts to the manuscript. All authors approved the final version.

Chapter 7
'Stress response in tardigrades: differential gene expression of molecular chaperones'
I participated as coauthor to this publication.

I scanned the known tardigrade sequences for homologies to known HSPs. I designed degenerated and specific primers for the heat shock proteins. I added minor parts to the manuscript. All authors approved the final version.

Chapter 8
'Tardigrade bioinformatics: Molecular adaptions, DNA j family and dynamical modelling'
I am an equal contributing first author of this publication.

I annotated the ESTs and translated them to hypothetical proteins. I searched the ESTs for DnaJ-proteins and built the maximum likelihood tree. I, Thomas Dandekar and Daniela Beisser wrote the manuscript and all authors read and approved the final version.

Chapter 9
'The ITS2 Database III–sequences and structures for phylogeny'
I am an equal contributing first author of this publication.

I, Jörg Schultz and Christian Koetschan did a complete redesign of the database model. I and Christian Koetschan also redesigned the generation and update pipeline including programming and testing. I estimated new scoring matrices and gap costs for different alignment methods for ITS2 sequences, sequence-structure pairs together with Tobias Müller. I was involved in writing the manuscript. All authors approved the final version.

Chapter 10
'Including RNA Secondary Structures improves Accuracy and Robustness in Reconstruction of Phylogenetic Trees'
I am an equal contributing first author of this publication.

Alexander Keller, Jörg Schultz, Matthias Wolf and Thomas Dandekar designed the study. I and Alexander Keller performed the simulation experiments and analyses. I and Tobias Müller estimated the substitution models used for the simulations and reconstructions. Alexander Keller, I and Matthias Wolf drafted the manuscript. All authors contributed to writing the paper, read the final manuscript and approved it.

Chapter 11
'ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales)'
I participated as coauthor to this publication.

Matthias Wolf designed the study. I determined the new sequences in our laboratory. Benjamin Ruderisch implemented the strPNJ within ProfDist. Tina Schleicher and Alexander Keller performed sequence analyses, structure prediction and phylogenetic analyses. Tobias Müller developed the ITS2 sequence-structure substitution

model and the ITS2 sequence-structure scoring matrix. Tina Schleicher, Alexander Keller and Matthias Wolf drafted the manuscript. All authors contributed to writing the paper, read the final manuscript and approved it.

Chapter 12
'Distinguishing species in *Paramacrobiotus* (Tardigrada) via compensatory base change analysis of internal transcribed spacer 2 secondary structures, with the description of three new species'
I am an equal contributing first author of this publication.

I received the sequences from Ralph O. Schill. I cleaned the sequences and annotated them using the HMM-annotation tool. Afterwards, I performed the homology modelling and the multiple sequence-structure alignment, the estimation of CBCs and reconstructed the phylogenetic trees. All authors wrote the manuscript and finally approved it.

Curriculum Vitae

Fundamental information

Name:	Frank Förster
Gender:	Male
Date of Birth:	22nd January 1980
Place of Birth:	Hohenmölsen, Sachsen-Anhalt, Germany
Nationality:	German
Contact:	frank.foerster@biozentrum.uni-wuerzburg.de
Address:	Arndtstr. 18, Würzburg, Bayern, 97074, Deutschland

Education

2007/06–2010/04	PhD-Thesis at the Department of Bioinformatics at the University of Würzburg
2007/04/26	Diploma in Biochemistry
1999/10–2007/06	Studies in Biochemistry at the University of Hamburg
1998/11–1999/08	Bundeswehr
1998/07/10	Abitur
1991/09–1998/07	Agricola Gymnasium Hohenmölsen
1986/09–1991/08	Otto-Schlag-Oberschule Hohenmölsen

List of Publications

Publications associated with this thesis

Keller[*], A., **Förster, F.**[*], Müller, T., Dandekar, T., Schultz, J., Wolf, M. (2010). 'Including RNA Secondary Structures improves Accuracy and Robustness in Reconstruction of Phylogenetic Trees'. *Biol Direct* 5.1 (Jan. 2010), p. 4. DOI: 10.1186/1745-6150-5-4. URL: http://dx.doi.org/10.1186/1745-6150-5-4.

Keller, A., Schleicher, T., **Förster, F.**, Ruderisch, B., Dandekar, T., Müller, T., Wolf, M. (2008). 'ITS2 data corroborate a monophyletic chlorophycean DO-group (Sphaeropleales)'. *BMC Evol Biol* 8, p. 218. DOI: 10.1186/1471-2148-8-218. URL: http://dx.doi.org/10.1186/1471-2148-8-218.

Koetschan[*], C., **Förster, F.**[*], Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R., Müller, T., Wolf, M., Schultz, J. (2010). 'The ITS2 Database III—sequences and structures for phylogeny'. *Nucleic Acids Res* 38.Database issue (Jan. 2010), pp. D275–D279. DOI: 10.1093/nar/gkp966. URL: http://dx.doi.org/10.1093/nar/gkp966.

Mali, B., Grohme, M. A., **Förster, F.**, Schnölzer, T. D. M., Reuter, D., Welnicz, W., Schill, R. O., Frohme, M. (2010). 'Transcriptome survey of the anhydrobiotic tardigrade *Milnesium tardigradum* in comparison with *Hypsibius dujardini* and *Richtersius coronifer*'. *BMC Genom* 11.1 (Mar. 2010), p. 168. DOI: 10.1186/1471-2164-11-168. URL: http://dx.doi.org/10.1186/1471-2164-11-168.

Reuner, A., Hengherr, S., Mali, B., **Förster, F.**, Arndt, D., Reinhardt, R., Dandekar, T., Frohme, M., Brümmer, F., Schill, R. O. (2009). 'Stress response in tardigrades: differential gene expression of molecular chaperones'. *Cell Stress & Chaperones* (Nov. 2009). DOI: 10.1007/s12192-009-0158-1. URL: http://dx.doi.org/10.1007/s12192-009-0158-1.

Schill[*], R. O., **Förster, F.**[*], Dandekar, T., Wolf, M. (2010). 'Distinguishing species in *Paramacrobiotus* (Tardigrada) via compensatory base change analysis of internal transcribed spacer 2 secondary structures, with the description of three new species'. *Organisms Diversity & Evolution* in press.

Schokraie, E., Hotz-Wagenblatt, A., Warnken, U., Mali, B., MarcusFrohme, **Förster, F.**, Dandekar, T., Hengherr, S., Schill, R. O., Schnölzer, M. (2010). 'Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms'. *PlosOne* 5.3 (Mar. 2010), e9502. DOI: 10.1371/journal.pone.0009502. URL: http://dx.doi.org/10.1371/journal.pone.0009502.

Förster, F.[*], Liang[*], C., Shkumatov[*], A., Beisser, D., Engelmann, J. C., Schnölzer, M., Frohme, M., Müller, T., Schill, R. O., Dandekar, T. (2009). 'Tardigrade workbench: comparing stress-related proteins, sequence-similar and functional protein clusters as well as RNA elements in tardigrades'. *BMC Genom* 10, p. 469. DOI: 10.1186/1471-2164-10-469. URL: http://dx.doi.org/10.1186/1471-2164-10-469.

Förster, F.[*], Beisser[*], D., Frohme, M., Schill, R. O., Dandekar, T. (2010*a*). 'Tardigrade bioinformatics: Molecular adapations, DNA j-family and dynamical modelling'. *J Zool Systemat Evol Res* conditionally accepted.

Förster, F.[*], Beisser[*], D., Grohme[*], M., Liang, C., Mali, B., Reuner, A., Siegl, A. M., Engelmann, J., Shkumatov, A., Schokraie, E., Müller, T., Blaxter, M., Schnölzer, M., Schill, R. O., Frohme, M., Dandekar, T. (2010*b*). 'Transcriptome analyzed at different levels in *Hypsibius dujardini* and *Milnesium tardigradum*: Specific adaptations, motifs and clusters as well as general protective pathways'. *Genome Biol* in preparation.

[*] shared first authors

Other Publications

Bemm, F., Schwarz, R., **Förster, F.**, Schultz, J. (2009). 'A kinome of 2600 in the ciliate *Paramecium tetraurelia*'. *FEBS Lett* 583.22 (Nov. 2009), pp. 3589–3592. DOI: 10.1016/j.febslet.2009.10.029. URL: http://dx.doi.org/10.1016/j.febslet.2009.10.029.

Schwarz, R., Fletcher, W., **Förster, F.**, Merget, B., Wolf, M., Schultz, J., Markowetz, F. (2010). 'Evolutionary distances between divergent sequences—a rational kernel approach'. *Bioinformatics* submitted to PLoS Comput Biol.

Conference contributions

Achtziger, M., Dandekar, T., **Förster, F.**, Gerlach, D., Hammesfahr, B., Keller, A., Koetschan, C., Maisel, S., Müller, T., Philippi, N., Ruderisch, B., Schleicher, T., Schultz, J., Schwarz, R., Seibel, P. N., Selig, C. E., Wolf, M. (2009). 'ITS2—it's 2 in 1—Sequence-Structure Analyses'. Celebrating Darwin: From The Origin of Species to Deep Metazoan Phylogeny, Berlin, *(Poster)*. Mar. 2009.

Förster, F., Keller, A., Schill, R. O., Dandekar, T., Wolf, M. (2009*a*). 'Distinguishing species in *Paramacrobiotus* (Tardigrada, Macrobiotidae)'. 11th International Tardigrade-Symposium 2009, Tübingen, *(Poster)*. Aug. 2009.

Förster, F., Keller, A., Schill, R. O., Dandekar, T., Wolf, M. (2009*b*). 'Distinguishing species in *Paramacrobiotus* (Tardigrada, Macrobiotidae) via compensatory base change analysis of internal transcribed spacer 2 secondary structures'. Celebrating Darwin: From The Origin of Species to Deep Metazoan Phylogeny, Berlin, *(Poster)*. Mar. 2009.

Förster, F., Liang, C., Beisser, D., Schill, R., Dandekar, T. (2009*c*). 'Functional protein clusters and regulatory motifs in *Hypsibius dujardini* and *Milnesium tardigradum*'. 11th International Tardigrade-Symposium 2009, Tübingen, *(Poster)*. Aug. 2009.

Erklärung

Hiermit erkläre ich ehrenwörtlich, dass ich die vorliegende Dissertation selbständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Die Dissertation wurde bisher weder in gleicher noch ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt. Außer dem Diplom in Biochemie von der Universität Hamburg habe ich bisher keine weiteren akademischen Grade erworben oder versucht zu erwerben.

Würzburg, April 20, 2010

_____

Frank Förster