bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Draft genome of the Heterotardigrade Milnesium tardigradum sheds light on ecdysozoan evolution

Felix Bemm[1,2*], Laura Burleigh[3], Frank Förster[1,4], Roland Schmucki[5], Martin Ebeling[5], Christian J. Janzen[6], Thomas Dandekar[1], Ralph O. Schill[7], Ulrich Certa[5], Jörg Schultz[1,4]

[1] Department of Bioinformatics, University Würzburg, 97074 Würzburg, Germany

[2] Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

[3] Roche Products Limited, Hertfordshire, United Kingdom

[4] Center for Computational and Theoretical Biology (CCTB), 97074 Würzburg, Germany

[5] Pharmaceutical Sciences, Roche Pharma Research and Early Development, Roche Innovation Center Basel, 124 Grenzacherstrasse, Basel CH 4070, Switzerland

[6] Department of Zoology I, Biocenter, University Würzburg, 97074 Würzburg, Germany

[7] Institute of Biomaterials and biomolecular Systems, 70174 Stuttgart, Germany

* **E-Mail: felix.bemm@tuebingen.mpg.de**

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Tardigrades are among the most stress tolerant animals and survived even unassisted exposure to space in low earth orbit. Still, the adaptations leading to these unusual physiological features remain unclear. Even the phylogenetic position of this phylum within the Ecdysozoa is under debate. Complete genome sequences might help to address these questions as genomic adaptations can be revealed and phylogenetic reconstructions can be based on new markers. Here, we present a first draft genome of a Heterotardigrada species, namely *Milnesium tardigradum*. We consistently place *M. tardigradum* and the two previously sequenced Eutardigrada species, *Hypsibius dujardini* and *Ramazzottius varieornatus,* as sister group of the nematodes with the arthropods as outgroup. Based on this placement, we identify a massive gene loss thus far attributed to the nematodes which predates their split from the tardigrades. We provide a comprehensive catalog of protein domain expansions linked to stress response and show that previously identified tardigrade-unique proteins are erratically distributed across the genome of *M. tardigradum*. We suggest alternative pathways to cope with high stress levels that are yet unexplored in tardigrades and further promote the phylum Tardigrada as a rich source of stress protection genes and mechanisms.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

There is no life without water. Antony van Leeuwenhoek must have been well aware of this fact when in 1702 he collected some dry dust from a roof gutter. He was in for a surprise when he viewed the sample with one of his self-built microscopes. Soon after mixing with some clean water, he found tiny animals, which he called 'animalcules' [1]. Thus, seemingly dead animals came fully alive again after rehydration. In 1959, D. Keilin coined the term 'cryptobiosis', which can be triggered by low oxygen (anoxybiosis), low temperature (cryobiosis), high salt concentrations (osmobiosis) or desiccation (anhydrobiosis) [2]. One group of animals able to undergo cryptobiosis are Tardigrades (from latin tardus = slow and gradi = walk) [3]. These animals are about 0.1 to 1.2 mm in size with a peculiar shape reminiscent of bears. Accordingly, they have also been called 'kleine Wasserbärchen' (little water bear) in German. They were first identified at the end of the 18th century [4]. Today, more than 1,000 species are known [5,6]. As their German name already suggests, tardigrades are an aquatic life form and can only live covered by a water film. Still, most species inhabit terrestrial habitats like mosses and lichens which regularly fall completely dry. At these times, adults, juveniles and embryos [7] can only survive until the next rain period by changing from the active state into the anhydrobiotic tun state. As metabolic conversion of nutrients requires water, tardigrades in the tun state suspend life and do not age [8]. In this form, they survive being frozen [9,10], heated [11] and exposed to enormous levels of UV [12] or ionizing [13] radiation.

The first studies addressing the unique physiological peculiarities of tardigrades established different hypotheses regarding their underlying genomic basis. A genome wide analysis of the gene coding complement of the tardigrade *Hypsibius dujardini* found that horizontal gene transfer may have shaped the functional capacity of the animal much more than previously suspected [14]. The analysis identified several thousand genes likely to be derived from non-metazoan sources mostly from bacteria [15,16]. A second independent genome study of *H. dujardini* reported strong conflicts between the two assembled and annotated genomes although the biomaterial for both studies was taken from the same original stock culture [17]. Analysis of the second genome reference for *H. dujardini* indeed suggested a very low level of horizontal gene transfer. Lately, these studies were complemented by a similar analysis in a second tardigrade species, *Ramazzotius varieornatus* [18]. The authors leveraged their high-quality genome sequence of *R. varieornatus* and could show that only a small proportion of the gene coding complement represents putative foreign genes. The study further showed that the species (selectively) lost several members of pathways that promote stress damage (e.g., peroxisomal oxidative pathway, stress responsive pathway) during hypoxia, genotoxic or oxidative stress but simultaneously display expansion of gene families

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

related to ameliorating damage (e.g., superoxide dismutases). A close examination of gene expression profiles during dehydration and rehydration by the authors revealed only minor differences between the two states. Additionally, the study identified a tardigrade-unique DNA-associated protein that, when transferred to human cell culture, suppresses DNA damage and promotes viability following irradiation. In summary, the study suggested that a) tardigrades can enter a dehydrated state without a massive transcriptional turnover and b) that the genome provides mechanisms that prevent, extenuate or protect against damage caused by extreme environmental conditions. Lately, a comparative study of the tardigrades *H. dujardini* and *R. varieornatus* revealed contrasting gene expression responses to anhydrobiosis [19]. While *H. dujardini* experienced a major transcriptional turnover, *R. varieornatu*s showed only limited regulation when switching to anhydrobiosis. The study further confirmed that the *H. dujardini* genome encodes only for a few horizontally transferred genes. Surprisingly, some of these seemed to be involved in the entry of anhydrobiosis. A whole-genome molecular phylogeny found more evidence for a Tardigrada+Nematoda relationship than the previously supported Tardigrada+Arthropoda relationship but also argued that a full genome sequence of representatives of Onychophora, Heterotardigrada (the sister group to the Eutardigrada including Hypsibiidae) and basally arising Nematoda would be required to fully address the tardigrade placement.

Here, we present an early draft genome sequence of the heterotardigrade, *M. tardigradum* [20]. We have chosen this species because it is among the most stress resistant tardigrades [17]. Specimen even survived the exposure to space in low earth orbit [21]. Furthermore several transcriptomic [22–24], proteomic [25,26] and metabolomic [27] studies have already been performed within this species. Furthermore, this species is only related to the previously sequenced tardigrades. Its genome sequence not only enabled us to derive a more general view on the mechanisms of stress resistance. It also allowed us to corroborate the phylogenetic position of the tardigrades in the tree of life and, based on these results, to identify gene loss as a major evolutionary trend in both tardigrades and nematodes.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

A draft genome of *M. tardigradum*

The assembled genome of *M. tardigradum* comprises 75.1 Mb, in good agreement to the results of a flow cytometry based determination of 73.3 ±1.8 Mb (see Suppl. Fig. 1). Overall, 6654 contigs were assembled with a contig N50 size of 50 kb. The assembly was validated with two approaches. First, a prediction of 248 core eukaryotic genes with CEGMA [28] revealed a 96 % completeness of the genome. Second, a prediction of near- universal single-copy orthologs with BUSCO [29] was used to benchmark the *M. tardigradum* genome against three different lineages. Benchmarking against a nematode specific BUSCO set revealed a completeness of 41 % while benchmarks against an arthropod-specific and a metazoan-specific set revealed a completeness of 81 % and 82 % (see Fig. 1; Suppl. Table 1). Only 1.23 % of the assembly was classified as repetitive or low-complexity. Based on a metazoan repeat library, 1271 DNA transposons (mostly hobo-Activator and Tc1-IS630-Pogo) and 2033 LTR elements (mostly BEL/Pao and Gypsy/DIRS1) were identified. The integrative gene annotation approach predicted 19,401 protein coding genes. 1684 genes were putatively derived through tandem duplication while 43 genes probably originated through segmental duplication when compared to *H. dujardini* and *R. varieornatus*. The subsequent functional annotation found homologs for 12,518 genes (65 %) while 7534 had an ortholog within the reference species from Ensembl Metazoan Release 34 [26], *R. varieornatus* or *H. dujardini*. Based on the curated gene set, 10,966 genes were functionally assigned [30] to either a protein family, protein domain or a functional site, excluding low-complexity, transmembrane and coiled-coil assignments. 7357 genes had at least one associated gene ontology term. Out of 19,401 protein coding genes 261 were potentially derived through horizontal gene transfer while 665 were potential contaminations likely introduced during DNA extraction.

Phylogeny and hypothesis testing

Since the position of the phylum Tardigrada is still under discussion a phylogenomic analysis was performed to reconstruct the phylogeny underlying 56 Ensembl Metazoa species and the three tardigrades. Single copy clusters of orthologs with at least one tardigrade member were identified and used for phylogenomic reconstruction. Using all target sequences (selected species from Ensembl Metazoa Release 34 and the three tardigrades) 2245 of these clusters were detected. A Maximum Likelihood based phylogenetic reconstruction based on this supermatrix placed *M. tardigradum*, *R. varieornatus* and *H. dujardini* as the sister group of the nematodes (see Fig. 2). Next, we explicitly compared different hypotheses regarding the placement of the tardigrades, namely (i) as the sister group of the arthropods in a panarthropoda cluster, (ii) as the sister group of the nematodes grouping the tardigrades

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

into the Cycloneuralia and (iii) as the outgroup to both arthropods and nematodes (see Fig. 4). Here, we used four different data sets. First, the super matrix was used to extract the per-site log-likelihood calculated by RaxML [31]. For each hypothesis, the approximately unbiased test as implemented in CONSEL [32] was performed. The same approach was carried out on (ii) the domain repertoire, (iii) the domain architectures and (iv) shared orthologous groups. As in the sequence based reconstructions, three tardigrades and 56 Ensembl Metazoa species were considered. In each case, the presence / absence of the feature was encoded in a binary matrix. For the super matrix and domain occurrences, the placement of the tardigrades as sister group to the nematodes had the highest rank (see Fig. 4). For shared orthologous groups, the placement as sister group to the arthropods had the highest rank while domain architectures ranked both placements equally. Taken together, the three different tardigrade genomes support, although not with full confidence, the tardigrades as members of a cycloneuralia cluster and reject their placement within the panarthropoda.

Protein domain-ome comparison

The protein domain repertoire (domain-ome) of an organism establishes its functional capacity. Protein domains are evolutionarily conserved units usually with independent structural and functional properties [33–35]. They are widely distributed over all existing organisms [36] with some of them being universal and others being clade-specific. We tested whether the three tardigrade genomes shared patterns of domain family expansions, contractions or even total loss compared to the other 56 Ensembl metazoan species since these events could provide hints at which biological processes and molecular functions are most likely associated with tardigrade-specific traits. Protein domains were classified into Class I and II expansions as well as Class I and II contractions. Class I expansions and contractions are those where the query species experienced the highest or lowest protein domain occurrence count whereas Class II expansions and contractions indicated protein domains where the query species belongs to the group with the 5 % highest or lowest occurrence. Overall, 7939 protein domains were tested for contractions and expansions. In total, 169 protein domains tested significant for an expansion while 12 tested significant for a contraction (see Fig. 3, Suppl. Table 2) in at least one tardigrade. *M. tardigradum* showed 22 Class I expansions, 19 Class II expansions, 2 Class I contractions and 4 Class II contractions. *R. varieornatus* showed 36 Class I expansions, 26 Class II expansions, 0 Class I contractions and 3 Class II contractions while *H. dujardini* showed 35 Class I expansions, 31 Class II expansions, 2 Class I contractions and 1 Class II contractions. Only 12 domain expansions (ATPase_P-typ_cation-transptr_C, Hemopexin-like_repeat, Peptidase, metallopeptidase, Superoxide dismutase, copper/zinc binding domain, STAS_dom, BCS1_N, CLZ_dom, Glycine N-acyltransferase, N-terminal, LysM_dom, Nucleotide_cyclase, Peptidase M2, peptidyl-dipeptidase A, Peripla_BP_I; see Suppl. Table 2) overlapped between all three

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tardigrade species. Expanded domain sets from all three species were separately subjected to a gene ontology enrichment analysis (see Suppl. Table 3). Individual enrichments overlapped in 7 terms (cellular cation homeostasis, cellular ion homeostasis, multi-organism process, negative regulation of muscle contraction, response to organic substance, response to stimulus, response to stress). Surprisingly, all species showed a substantial amount of undetectable protein domains (*M. tardigradum*: 2239; *R. varieornatus*: 2146; *H. dujardini*: 2174) indicating potential loss events. The phylogenetic placement of tardigrades enabled us to examine the loss events in more detail. We used a Maximum Likelihood approach to reconstruct the domain repertoire of extinct ancestors. The approach enabled quantification of the domain loss using the phylogenetic tree while eliminating negative effects from contaminations or horizontal gene transfer events. We found that a major part of the losses happened before the split of tardigrades and nematodes followed by further tardigrade clade specific losses and only minor losses in the nematodes (see Fig. 5). No gene ontology terms were enriched in protein domains lost before the split of tardigrades and nematodes (Cycloneuralia losses) nor in protein (domains) specifically lost in tardigrades, arthropods or nematodes. Of note, the removal of putative contaminations and horizontally transferred genes had a larger impact on the numbers (e.g., nematode and arthropod specific losses). The assessment of domain family births revealed 2 domain families, namely SMC_ScpA (PF02616) and DUF612 (PF04747) gained before the split of tardigrades and nematodes. SMC_ScpA (PF02616) likely is an undetected contamination while DUF612 (PF04747) was only found in nematodes so far. Neither tardigrades nor nematodes showed further clade-specific gains.

Revisiting stress tolerance mechanisms

The enormous stress tolerance combined with the capability to undergo anhydrobiosis is arguably one of the most prominent features of tardigrades. Still, the underlying mechanisms are not fully clear. Even comparing the genomes of *H. dujardini* and *R. varieornatus* did not result in a consistent picture [19]. Rather, it was suggested, that already these closely related tardigrades utilize different strategies. The genome of *M. tardigradum* enabled us to revisit current hypotheses and to contrast the eutardigrade genomes with the genome of a heterotardigrade. During the last years, mainly two strategies for the conservation of cellular structure in anhydrobiosis have been discussed. One of the proposed mechanisms is the accumulation of a specific sugar, trehalose. Indeed, the genomes of all three tardigrades encode at least one biosynthetic pathway from D-glucose to trehalose (see Table 2). The second hypothesis highlights the importance of late embryo abundance (LEA) proteins to stabilze cellular structures. Still, we identified the typical LEA protein domains in only two of the tardigrades (*M. tardigradum*: 1, *H. dujardini*: 1). Contradictory, 10 such proteins in *R. varieornatus* were reported earlier [18]. However, closer inspection of these proteins

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

revealed that they did not contain typical LEA protein domains. Rather, some contained DUF883 (PF05957), a domain that frequently appears in proteins also containing LEA_4 (PF02987) protein domains. In addition to these most prominent hypotheses, a role of heat shock proteins (HSPs) in anhydrobiosis has been suggested. As these assist in protein folding and can refold denatured proteins, they could provide a self-evident mechanism to repair damage arising in anhydrobiosis. Indeed, all typical HSP protein families are encoded in the three tardigrade genomes (Table 1). Still, none of the families is significantly expanded in comparison to the other metazoan species. In addition to these hypotheses, Hashimoto et al. suggested the importance of several tardigrada-unique genes in the genome of *R. varieornatus* which are constitutively expressed and associated with stress tolerance. These included the previously identified Cytoplasmic Abundant Heat Soluble (CAHS) and Secretory Abundant Heat Soluble (SAHS) proteins as well as a newly characterized DNA Damage suppressor (Dsup). Based on the annotated *R. varieornatus* templates of these proteins we searched the genomes of *H. dujardini* and *M. tardigradum* using a profile and a homology based strategy. We found that all species encode several CAHS (see Table 1) but only *H. dujardini* and *R. varieornatus* encode SAHS proteins. Neither *M. tardigradum* nor *H. dujardini* has any homolog of Dsup.

The direct genome comparison of *H. dujardini* and *R. varieornatus* suggested that extensive loss in in the peroxisome pathway and in stress signaling pathways happened in both species and that loss of these resistance pathways may be associated with anhydrobiosis. We also searched for components that might act as reactive oxygen species (ROS) production suppressors. Surprisingly, we found an alternative oxidase (AOX) in each of the three tardigrades. AOX is erratically found in metazoan species but frequently present in many plants and bacteria. It can directly suppress ROS production and indirectly change the energy status of the cell through the alternative respiratory pathway [37]. In summary, the erratic distribution of tardigrade-unique proteins underlines their variable relevance for stress tolerance. Thus, there might indeed be a role of alternative pathways (e.g., ROS suppression through AOX) that are yet unexplored in tardigrades. Still, both facts support the notion that the phylum Tardigrada is a rich source of new protection genes, pathways and mechanisms involved in stress tolerance.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

The remarkable stress tolerance of tardigrades has fascinated scientists and nonscientists alike. Accordingly, the publication of the first genome sequence of a tardigrade, *H. dujardini*, generated large interest. Still, the results presented in the original publication have been defeated largely [14–16]. Here, the genome of another tardigrade, *R. varieornatus* was of substantial importance. Both Eutardigrada species are closely related and belong to the family of Hybsibiidae. However, the phylum Tardigrada consists of more than 1000 species with a huge phenotypic and presumably also genetic variability. Thus, clade-wide insights from comparative genomics are limited to the species spectra analysed. Here, we present an early draft genome of a Heterotardigrada, *M. milnesium*. Although fragmented, it enabled us to test current hypotheses regarding the molecular mechanisms of tardigrade stress tolerance of tardigrades in a wider scope.

More than 40 years ago, water replacement and vitrification were suggested as core mechanisms for stabilization of cellular structures [37]. According to this hypothesis, water is replaced by other biomolecules, resulting in a glass-like state of the cell. Mainly two types of molecules enabling this transition, sugars [38,39] and late embryo abundant (LEA) proteins [40,41] have been described. An analysis of the trehalose biosynthesis and metabolism pathway showed that necessary components to generate and reconvert trehalose are indeed present in all three tardigrades. Although small amounts of trehalose were found in eutartigrades like *R. varieornatus* and *H. dujardini*, no evidence of the relevance for anhydrobiosis was found in *M. tardigradum* [10,42]. This could indicate independent adaptations in different branches of the tardigrades. LEA proteins were first detected in plants [43]. Later, these proteins were shown to be of importance for anhydrobiosis in nematodes [40] and arthropods [44]. With a phylogenetic position between these two phyla, tardigrades might also leverage these proteins to prevent protein aggregation. Still, genes encoding for typical LEA domain could be identified in only two of the tardigrades (*M. tardigradum* and *H. dujardini*).

As further candidates, heat shock proteins have been suggested. As they assist in protein folding and can refold denatured proteins, they could provide a self-evident mechanism to repair damage arising in anhydrobiosis. Still, the relevance of heat shock proteins for tardigrades is controversial. HSP70 expression is increased at rehydration [12] but not increased in desiccated animals [12,45]. Directly comparing different variants of HSPs revealed complex patterns [46,47]. We found several HSP-related protein domains significantly altered in each of the three tardigrade species. Thus, expansions previously published for the *R. varieornatus [17]* genome might be assigned to the whole clade.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The emergence of ROS is of considerable danger for a cell, as it can damage all cellular components. Already a challenge for a 'standard' cell, this problem increases dramatically when a cell desiccates. Accordingly, genes involved in the reduction of ROS are up regulated upon entry into anhydrobiosis [48]. A screen for protein domains related to ROS production and scavenging unexpectedly revealed that all three tardigrade genomes code for an alternative oxidase (AOX, PF01786). This protein can lower the internal production of ROS at the mitochondria [49,50]. While common in bacteria, plants and fungi, AOX was thus far only found in a few metazoan species, mostly living in salt water. This includes the bdelloid rotifer *Adineta vaga* which is also capable of anhydrobiosis [51] respectively cryptobiosis. In addition to the inactivation of ROS, avoiding its emergence would be a complementary strategy. The presence of AOX proteins might indicate an overlooked mechanism for anhydrobiotic metazoa and tardigrades as the so far first terrestrial animal utilizing this mechanism for ROS defense.

The genome of *R. varieornatus* was the first that revealed tardigrade-unique proteins involved in stress response. Hashimoto et al. (2016) discovered a protein that shields DNA from radiation (Dsup). Indeed, expression of this protein in human cells gave them a survival advantage. Yamaguchi et al. (2012) further discovered genes constitutively expressed and associated with stress tolerance like Cytoplasmic Abundant Heat Soluble (CAHS) and Secretory Abundant Heat Soluble (SAHS) proteins as well as Mitochondrial Abundant Heat Soluble (MAHS) proteins (Yamaguchi et al. 2012) . Surprisingly, genes coding for these proteins are absent in *H. dujardini* and *M. tardigradum*. In summary, the scattered distribution of MAHS-like, SAHS-like and tardigrade-unique proteins like Dsup, suggests species or at least class specific adaptations towards typical tardigrade traits.

The early draft genome presented here also provided new data to address the phylogenetic placement of the tardigrades, which is still discussed controversially. Mainly two hypotheses exist, placing tardigrades either as the outgroup of the arthropods building the Panarthropoda clade [53–58], or together with the nematodes as a member of the Cycloneuralia [59–64]. We improved the first phylogenetic analysis based on the genomes of *R. varieornatus* and *H. dujardini* first by including *M. tardigradum* to rely on a broader taxonomic range and second by including a wider range of species. The resulting tree reproduces the results of the previous analysis and robustly places the tardigrades as a sister group of the nematodes with the arthropods as outgroup. Still, phylogenetic reconstructions only report the most likely tree and will not provide a comparison to other tree topologies. To address this, we explicitly tested the current taxonomic hypotheses based on different datasets (sites in the

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

supermatrix, orthologous groups, domain presence and domain architectures). This approach allowed us to provide further statistical support for each hypotheses. The hypothesis test on the supermatrix sites as well as the presence/absence of domain families supported the placement of the tardigrades as sister group to the nematodes consistent with the reconstruction of the tree. Testing shared domain architectures ranked none of two hypotheses (Tardigrada+Arthropoda and Tardigrada+Nematoda) highest while testing shared orthologous groups reversely ranked the Tardigrada+Arthropoda hypotheses highest. Still, we conclude that our analyses most strongly support the Cycloneuralia hypothesis.

With a reliable phylogenetic placement of the tardigrades at hand, we were able to reconstruct evolutionary events within the tree of the ecdysozoa. To this end, we used a maximum likelihood reconstruction of the ancestral nodes within the tree. We found, that the reduction of the domain repertoire starting from a complex Ur-ecdysozoan was a major trend in the evolution of the last common ancestor of nematodes and tardigrades was. This was followed by further lineage and species-specific losses, especially in tardigrades (see Fig. 5). Thus, the evolution of nematodes and tardigrades recalls the general trend of reduction already observed at the base of the Bilateria [65]. Notably, the reconstruction of evolutionary events was heavily influenced by contaminations and genes potentially derived through horizontal gene transfer. A more robust and contiguous genome reference for *M. tardigradum* and further support from transcriptomic, proteomic or other molecular experiments is necessary to ultimately link the adaptive genomic footprints to the unusual physiological capabilities of *M. tardigradum and* the tardigrades in general.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. van Leeuwenhoek A. On certain animalcules found in the sediment in gutters of the roofs of houses. Letter 144. The selected works of Antony van Leeuwenhoek (translated by S Hoole, 1807). 1807;2: 298–311.

2. Keilin D. The problem of anabiosis or latent life: history and current concept. Proc R Soc Lond B Biol Sci. 1959;150: 149–191.

3. Spallanzani L. Opuscoli di fisica animale e vegetabile. Dalla Società tipogr. de' classici italiani; 1826.

4. Bonnet C, Goeze JAE. Herrn Karl Bonnets Abhandlungen aus der Insektologie. Aus dem französischen übersetzt und mit einigen Zusätzen herausgegeben von Joh. August Ephraim Goeze. 1773.

5. Guidetti RG, Bertolani RB. Tardigrade taxonomy: an updated check list of the taxa and a list of characters for their identification. Zootaxa. 2005;845: 1–46.

6. Degma P, Guidetti R. Notes to the current checklist of Tardigrada. Zootaxa. 2007;1579: 41–53.

7. 3. Accessed 8 June 2014.31. Schill RO Fritz GB. Desiccation tolerance in embryonic stages of the tardigrade. J Zool. 2008;276: 103–107.

8. Schill RO MF. Anhydrobiosis in tardigrades and its effects on longevity traits. J Zool. 2008;275: 216–220.

9. Hengherr S, Reuner A, Brümmer F, Schill RO. Ice crystallization and freeze tolerance in embryonic stages of the tardigrade Milnesium tardigradum. Comp Biochem Physiol A Mol Integr Physiol. 2010;156: 151–155.

10. 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.

11. 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.

12. 19732016. Accessed 17 July 2014.36. Altiero T Guidetti R, Cesari M CV. Ultraviolet radiation tolerance in hydrated and desiccated eutardigrades. J Zoolog Syst Evol Res. 2011;49: 104–110.

13. 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.

14. Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Nishimura EO, et al. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci U S A. 2015;112: 15976–15981.

15. Bemm F, Weiß CL, Schultz J, Förster F. Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? Proc Natl Acad Sci U S A. 2016;113: E3054–6.

16. Delmont TO, Eren AM. Identifying contamination with advanced visualization and

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analysis practices: metagenomic approaches for eukaryotic genome assemblies. PeerJ. 2016;4: e1839.

17. Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, et al. No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proc Natl Acad Sci U S A. 2016;113: 5053–5058.

18. Hashimoto T, Horikawa DD, Saito Y, Kuwahara H, Kozuka-Hata H, Shin-I T, et al. Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. Nat Commun. 2016;7: 12808.

19. Yoshida Y, Koutsovoulos G, Laetsch DR, Stevens L, Kumar S, Horikawa DD, et al. Comparative genomics of the tardigrades Hypsibius dujardini and Ramazzottius varieornatus. PLoS Biol. 2017;15: e2002266.

20. Doyère L. Mémoire sur les Tardigrades. Annales des sciences naturelles. 1840. pp. 269–361.

21. 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.

22. 22281184. Accessed 29 February 2012.55. Wang C Grohme M a, Schill RO MB. Towards Decrypting Cryptobiosis-Analyzing Anhydrobiosis in the Tardigrade Milnesium tardigradum Using Transcriptome Sequencing. PLoS One. 2014;9: e92663.

23. Förster F, Beisser D, Grohme MA, Liang C, Mali B, Siegl AM, et al. Transcriptome analysis in tardigrade species reveals specific molecular pathways for stress adaptations. Bioinform Biol Insights. 2012;6: 69–96.

24. Mali B, Grohme MA, Förster F, Dandekar T, Schnölzer M, Reuter D, et al. Transcriptome survey of the anhydrobiotic tardigrade Milnesium tardigradum in comparison with Hypsibius dujardini and Richtersius coronifer. BMC Genomics. 2010;11: 168.

25. Schokraie E, Hotz-Wagenblatt A, Warnken U, Mali B, Frohme M, Förster F, et al. Proteomic analysis of tardigrades: towards a better understanding of molecular mechanisms by anhydrobiotic organisms. PLoS One. 2010;5: e9502.

26. Schokraie E, Warnken U, Hotz-Wagenblatt A, Grohme MA, Hengherr S, Förster F, et al. Comparative proteome analysis of Milnesium tardigradum in early embryonic state versus adults in active and anhydrobiotic state. PLoS One. 2012;7: e45682.

27. Beisser D, Grohme MA, Kopka J, Frohme M, Schill RO, Hengherr S, et al. Integrated pathway modules using time-course metabolic profiles and EST data from Milnesium tardigradum. BMC Syst Biol. 2012;6: 72.

28. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23: 1061–1067.

29. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31: 3210–3212.

30. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–1240.

31. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

32. Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 2001;17: 1246–1247.

33. Janin J, Chothia C. Domains in proteins: definitions, location, and structural principles. Methods Enzymol. 1985;115: 420–430.

34. Holm L, Sander C. Parser for protein folding units. Proteins. 1994;19: 256–268.

35. Buljan M, Bateman A. The evolution of protein domain families. Biochem Soc Trans. 2009;37: 751–755.

36. Levitt M. Nature of the protein universe. Proc Natl Acad Sci U S A. 2009;106: 11079–11084.

37. Van Aken O, Giraud E, Clifton R, Whelan J. Alternative oxidase: a target and regulator of stress responses. Physiol Plant. 2009;137: 354–361.

38. Crowe JH, Clegg JS. Anhydrobiosis. Stroudsburg, PA: Dowden, Hutchinson & Ross. Inc 477pp. 1973;

39. Crowe LM. Lessons from nature: the role of sugars in anhydrobiosis. Comp Biochem Physiol A Mol Integr Physiol. 2002;131: 505–513.

40. 11867276. Accessed 17 July 2014.43. Crowe JH Hoekstra FA. Anhydrobiosis. Annu Rev Physiol. 1992;54: 579–599.

41. Goyal K, Walton LJ, Browne JA, Burnell AM, Tunnacliffe A. Molecular anhydrobiology: identifying molecules implicated in invertebrate anhydrobiosis. Integr Comp Biol. 2005;45: 702–709.

42. Tunnacliffe A, Wise MJ. The continuing conundrum of the LEA proteins. Naturwissenschaften. 2007;94: 791–812.

43. Hengherr S, Heyer AG, Köhler H-R, Schill RO. Trehalose and anhydrobiosis in tardigrades--evidence for divergence in responses to dehydration. FEBS J. 2008;275: 281–288.

44. Galau GA, Bijaisoradat N, Hughes DW. Accumulation kinetics of cotton late embryogenesis-abundant mRNAs and storage protein mRNAs: coordinate regulation during embryogenesis and the role of abscisic acid. Dev Biol. 1987;123: 198–212.

45. Kikawada T, Nakahara Y, Kanamori Y, Iwata K-I, Watanabe M, McGee B, et al. Dehydration-induced expression of LEA proteins in an anhydrobiotic chironomid. Biochem Biophys Res Commun. 2006;348: 56–61.

46. Rizzo AM, Negroni M, Altiero T, Montorfano G, Corsetto P, Berselli P, et al. Antioxidant defences in hydrated and desiccated states of the tardigrade Paramacrobiotus richtersi. Comp Biochem Physiol B Biochem Mol Biol. 2010;156: 115–121.

47. Schill RO, Steinbrück GHB, 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.

48. Reuner A, Hengherr S, Mali B, Förster F, Arndt D, Reinhardt R, et al. Stress response in tardigrades: differential gene expression of molecular chaperones. Cell Stress Chaperones. 2010;15: 423–430.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

49. Dry up and survive: the role of antioxidant defences in anhydrobiotic organisms. J Limnol. 2013;72: 62–72.

50. Ito Y, Saisho D, Nakazono M, Tsutsumi N, Hirai A. Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature. Gene. 1997;203: 121–129.

51. Vanlerberghe GC, McIntosh L. ALTERNATIVE OXIDASE: From Gene to Function. Annu Rev Plant Physiol Plant Mol Biol. 1997;48: 703–734.

52. Flot J-F, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EGJ, et al. Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature. 2013;500: 453–457.

53. Yamaguchi A, Tanaka S, Yamaguchi S, Kuwahara H, Takamura C, Imajoh-Ohmi S, et al. Two novel heat-soluble protein families abundantly expressed in an anhydrobiotic tardigrade. PLoS One. 2012;7: e44209.

54. The evolution of the Ecdysozoa. Philos Trans R Soc Lond B Biol Sci. 2008;363: 1529–1537.

55. Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, et al. Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biol Evol. 2010;2: 425–440.

56. Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, et al. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc Natl Acad Sci U S A. 2011;108: 15920–15924.

57. Giribet G, Edgecombe GD. Reevaluating the arthropod tree of life. Annu Rev Entomol. 2012;57: 167–186.

58. Mayer G, Martin C, Rüdiger J, Kauschke S, Stevenson PA, Poprawa I, et al. Selective neuronal staining in tardigrades and onychophorans provides insights into the evolution of segmental ganglia in panarthropods. BMC Evol Biol. 2013;13: 230.

59. Smith MR, Ortega-Hernández J. Hallucigenia's onychophoran-like claws and the case for Tactopoda. Nature. 2014;514: 363–366.

60. Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, von Haeseler A, Kube M, et al. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Mol Phylogenet Evol. 2007;45: 942–951.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

61. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 2008;452: 745–749.

62. Lartillot N, Philippe H. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci. 2008;363: 1463–1472.

63. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci. 2009;276: 4261–4270.

64. Rehm P, Borner J, Meusemann K, von Reumont BM, Simon S, Hadrys H, et al. Dating the arthropod tree based on large-scale transcriptome data. Mol Phylogenet Evol. 2011;61: 880–887.

65. Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T. A transcriptome approach to ecdysozoan phylogeny. Mol Phylogenet Evol. 2014;80: 79–87.

66. Miller DJ, Ball EE. The gene complement of the ancestral bilaterian - was Urbilateria a monster? J Biol. 2009;8: 89.

67. Gregory TR, Johnston JS. Genome size diversity in the family Drosophilidae. Heredity . 2008;101: 228–238.

68. Ardila-Garcia AM, Umphrey GJ, Gregory TR. An expansion of the genome size dataset for the insect order Hymenoptera, with a first test of parasitism and eusociality as possible constraints. Insect Mol Biol. 2010;19: 337–346.

69. Sanger Institute. SMALT [Internet]. Available: https://www.sanger.ac.uk/resources/software/smalt/

70. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12: 59–60.

71. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol. 2016;12: e1004957.

72. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27: 722–736.

73. Chevreux B. MIRA: an automated genome and EST assembler. 2007; Available: http://archiv.ub.uni-heidelberg.de/volltextserver/7871/1/thesis_zusammenfassung.pdf

74. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013--2015. Institute for Systems Biology http://repeatmasker org. 2015;

75. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32: 767–769.

76. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001;29: 37–40.

77. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8: 785–786.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

78. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305: 567–580.

79. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40: e49.

80. Fang H. dcGOR: an R package for analysing ontologies and protein domain annotations. PLoS Comput Biol. 2014;10: e1003929.

81. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res. 2017;45: D635–D642.

82. Ekseth OK, Kuiper M, Mironov V. orthAgogue: an agile tool for the rapid prediction of orthology relations. Bioinformatics. 2014;30: 734–736.

83. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274.

84. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30: 1575–1584.

85. Chernomor O, von Haeseler A, Minh BQ. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst Biol. 2016;65: 997–1008.

86. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;30: 1188–1195.

87. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14: 587–589.

88. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44: D279–85.

89. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195.

90. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.

91. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59: 307–321.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Animal culture

Tardigrade specimens of *M. tardigradum* Doyere 1840 [20] (Eutardigrada, Apochela), cultured in the laboratory for a decade, were used to study the genome. Originally, they were collected from dry moss in Tübingen, Germany. The carnivorous tardigrade species was reared in plastic culture dishes on a small layer of 3 % agar, covered with Volvic™ water (Danone Waters Deutschland, Wiesbaden, Germany). Rotifers of the species *Philodina citrina* were provided as food twice a week. The cultures were maintained in an environmental chamber at 20 °C using an artificial light source with a 12 h light, 12 h dark cycle. For the DNA/RNA extraction exuvia with eggs and embryos were collected and cleaned by five washing steps with Volvic™ water. Subsequently, they were placed separately in a 24-well plate until they hatched. Juveniles were transferred into a reaction tube, frozen in liquid nitrogen and stored at -80°C.

Genome size estimation

The genome size of *M. tardigradum* was estimated using flow cell cytometry*. Drosophila melanogaster* was used as standard [67]. A culture of *M. tardigradum* was washed (4 times, M9 buffer) and placed into modified Galbraith's buffer. Nuclei were released with a tissue grinder (Kontes Dounce tissue grinder, "A" pestle) and filtered to a 30 µm Nylon mesh. The same procedure was carried out with a single head from *D. melanogaster* female. The nuclear suspension was stained with propidium iodid (PI) for 2 h and measured immediately with a FACScalibur flow cytometer (Becton Dickinson, USA) and analyzed with CellQuest Pro version 6.0. PI-positive cells were gated and fluorescence intensity was analyzed in FL2-H channel and displayed on a linear scale (Supplementary Figure S1A). Non-stained cells served as a negative control (Supplementary Figure S1C). The whole procedure was benchmarked by comparing *D. melanogaster* (Supplementary Figure S1B) and *Apis mellifera* (data not shown) [68]. Results indicated an error of 5 %.

Genome and transcriptome sequencing

DNA was extracted from approximately 1000 freshly hatched (to avoid bacterial contamination) animals using the Qiagen DNeasy kit according to the manufacturer's instructions for animal tissues (spin column protocol). Animals were washed five times in RNAse/DNAse-free water, resuspended in Qiagen buffer ATL and disrupted with a FastPrep-24 homogenizer (MP Biomedicals for 2 x 30 s at 4 m/s). Following overnight incubation in buffer ATL and proteinase K at 56 ºC, samples were treated with RNAse A and purified on a spin column. RNA was extracted from a similar sized animal culture as for the DNA. Animals were disrupted as described above, resuspended in buffer RLT, and RNA was extracted using the Qiagen RNeasy kit. Ribosomal RNA was depleted using the RiboMinus Kit for

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RNA-seq (Invitrogen) and reverse transcribed using random hexamers (Promega Im-Prom II Reverse Transcription System). cDNA was amplified using the GenomePlex Complete Whole Genome Amplification Kit (Sigma). The 95 ºC fragmentation step was omitted from the whole genome amplification, as RNA had been fragmented during homogenization. Bead libraries were prepared from DNA (1.8 µg) and cDNA (2 µg) using the GS FLX Titanium general library preparation kit (454 Life Sciences), followed by amplification using emulsion PCR with the LV emPCR kit (Lib-L) (454 life Sciences). Sequencing was performed on a 454 FLX instrument (454 Life Sciences). A second sequencing data set was produced from an additional batch of animals. DNA was extracted as above and subjected to whole genome amplification with Qiagen REPLI-g prior to sequencing. TruSeq DNA library prep and Illumina sequencing was carried out by GATC.

Genome and transcriptome assembly

Genomic and transcriptomic reads were prepared by masking vector contamination and adapters using SMALT [69]. All read sets were compared against NCBI-nr using diamond [70]. The resulting alignments were prepared for MEGAN using daa-meganizer [71]. MEGAN was used to compute the lowest common ancestor for each read individually. Reads assigned only to the superkingdom Bacteria or Archaea were removed from the data set if there GC content was smaller than 25% or larger than 55%. Remaining genomic reads were assembled with Canu (release 1.3; errorRate=0.035, genome- Size=75000000, minReadLength=50, corMinCoverage=0, corMaxEvidenceErate=0.15, minOverlapLength=50, trimReadsCoverage=2) [72]. Transcriptomic reads were assembled using MIRA4 with accurate settings [73]. Genome completeness was validated with CEGMA [28] and BUSCO [29].

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome feature annotation

Known repetitive elements were annotated with RepeatMasker (v.4.0.4, species=metazoa) [74]. Coding genes were annotated with Braker1 (version 1.9; default parameters) [75] by combining de novo gene predictions and evidence alignments from ESTs. Evidence alignments were generated by aligning all ESTs against the genome using BLAT. Resulting alignments were converted into intron boundaries and passed to Braker1. The resulting proteins were functionally classified using homology and profile based methods. Protein families, domains and important sites were assigned using InterproScan5 (release 5.20; default parameters) [30] and the Interpro database (release 59.0) [76]. Signal peptides and transmembrane regions were predicted with SignalP (v.4.0) [77] and TMHMM (v2.0) [78]. Gene ontology terms and basic functional descriptions were assigned by lifting protein domain gene ontology annotations to their respective gene/protein. Segmental and tandem duplicates were detected using MCScanX [79].

Protein Domain Expansions and Contractions

Significantly expanded and contracted protein domains (Pfam) were identified by comparing their occurrence in the three tardigrades (*H. dujardini* release 2.3.1 and *R. varieornatus* release Rv101 from http://ensembl.tardigrades.org) to all reasonable complete species present in the Ensembl Metazoa database (Release 34; see Suppl. Fig. 2) using a chi square test. The occurrence of a specific protein domain in the three tardigrades was compared to the occurrence of the same protein domain in each of the reference species individually. The number of all proteins associated with at least one protein domain was used as background for each species. The resulting p-values for each protein domain were combined into a weighted consensus p-value since they addressed the same null hypothesis, that a protein domain is not expanded or contracted significantly. For that, all p-values were z-transformed and a weighted consensus test was applied. The final weighted consensus p-value was adjusted using the Bonferroni method and considered significant at a level of 5%. Expansion and contractions were used to test for enriched gene ontology terms with dcGOR [80]. Enrichments were statistically verified with the hypergeometric test (Parent-Child algorithm). P-values were adjusted using Bonferroni's method. All proteins domains found in the 56 species were used as background.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phylogenetic reconstruction and hypothesis testing

The phylogenetic reconstruction was carried out using 56 selected species present in release 34 of the Ensembl Metazoa database [81]. Putative contaminations and horizontally transferred genes in all species were identified by comparing their predicted proteins against the Ensembl Metazoa database (excluding the query species) as well as all bacterial RefSeq non-redundant proteins using diamond [70]. The resulting alignments were prepared for MEGAN using daa-meganizer [71]. MEGAN was used to compute the lowest common ancestor for each read individually. Proteins with bacteria as lowest common ancestor which were encoded on contigs only containing bacterial proteins were flagged as putative contamination. Proteins with bacteria as lowest common ancestor but flanked by at least two eukaryotic proteins were flagged as horizontally transferred. Proteins flagged as contamination nor as horizontally transferred were excluded from all analysis if not stated otherwise. Potential in-paralogs, orthologs and co-ortholog pairs were identified using orthAgogue [82]. The species phylogeny was reconstructed using IQ-TREE (version 1.5.5; -alrt 1000 -bb 1000 -bspec GENESITE -m TEST) [83] based on the ortholog groups generated with MCL [84]. Only single copy ortholog groups which contained at least one tardigrade and a minimum number of 12 species were considered. The maximum-likelihood tree was inferred using the edge-linked partition model in IQ-TREE [85]. Branch support was obtained with the ultrafast bootstrap method [86]. Substitution models were selected using ModelFinder [87]. Alternative phylogenetic hypothesis for the placement of the tardigrades were tested with RAxML [31] and CONSEL [32]. Testing was done using binary representations of the absence-presence matrices for protein domain families, domain architectures and orthologous groups. Domain architectures were defined on Pfam domain families [88]. Repetitive stretches of domains were collapsed and the order not considered. RAxML (-f A -m BINGAMMA -T 2) was used to calculated per-site log likelihoods for each of the alternative hypothesis. Test statistics for alternative hypothesis were calculated using the ‘approximately unbiased test’ implemented in CONSEL with 100 replicates.

Protein domain loss estimation

Protein domain losses and gains were detected using the Tardigrada+Nematoda (Cycloneuralia) tree topology and the corresponding absence-presence matrices. Ancestral nodes were reconstructed using RAxML (-f A -m BINGAMMA -T 2).

Identification of domains associated with stress resistance

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SAHS/CAHS containing proteins previously identified in *R. varieornatus* were detected using a profile based approach. Template from *R. varieornatus* were aligned, the alignment manually curated and used to build a hidden Markov model (HMM) [89]. A reverse search of the model against *R. varieornatus* proteins was conducted and the results used to define an optimal inclusion e-value (CAHS = 1.2×10^{-22}; SAHS = 5.1×10^{-40}). The final model was used to screen proteins from all species. Dsup homologs were identified by a simple protein blast (BLASTP) [90] against the complete set of 56 Ensembl Metazoa Release 34 species and the three tardigrades. HSPs were identified by using predefined Pfam protein domains (PF00011, PF00012, PF00118, PF00166, PF00183, PF00226). Trehalose biosynthesis and metabolism components were identified by using predefined Pfam protein domains (PF00128, PF00358, PF00534, PF00982, PF01204, PF02056, PF02358, PF02922, PF03632, PF03633, PF03636, PF09071, PF11941, PF11975, PF16657). Late Embryogenesis Abundant proteins (LEA proteins) were identified using predefined Pfam protein domains (PF00477, PF02987, PF03168, PF03242, PF03760, PF10714). AOX proteins were identified using the predefined Pfam protein domain (PF01786).

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We would like to thank Steffen Hengherr for the maintenance of the tardigrade culture and Thomas Hegna for the drawing of *D. melanogaster* used in Figures 3 and 5.

Financial Disclosure

FB was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare that no competing interests exist.

Abbreviations

ROS – Reactive Oxygen Species; HSP – Heat Shock Protein; LEA – Late Embryo abundant; AOX – Alternative Oxidase; HGT – Horizontal Gene Transfer; HMM – hidden Markov model; CAHS – Cytoplasmic Abundant Heat Soluble: SAHS – Secretory Abundant Heat Soluble; MAHS – Mitochondrial Abundant Heat Soluble;

Accession Numbers

Raw sequencing data, the genome assembly and its gene annotation are deposited in EBI ENA under accession number PRJEB22082.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

Figure 1. Genome completeness, contamination rate and putative percentage of horizontal gene transfer (HGT) across all Ensembl Metazoa Release 34 species and the three tardigrades. A) Genome completeness values of BUSCO and CEGMA. CEGMA completeness values show less spread within and between phyla. BUSCO completeness (using a set of metazoa BUSCOs) shows lower values for Nematoda, Non-Ecdysozoa and Tardigrada. B) Percentage of contamination-derived genes (e.g., bacterial proteins only on an individual genome sequence such as a contig) and HGT-derived genes (e.g., bacterial proteins on an individual genome sequence such as a contig surrounded by eukaryotic genes).

Figure 2. Maximum Likelihood based reconstruction of ecdysozoan phylogeny using a supermatrix approach. Only bootstrap support values with less < 100% support are shown. Additionally, all branches had a SH-aLRT support >= 80% [91]. Arthropods are placed as sister taxon to nematodes and tardigrades.

Figure 3. Hypothesis testing using the phylogenetic position of tardigrades. A) According to three different hypotheses, tardigrades are either the sister taxon to the arthropods (1), the nematodes (2) or the outgroup to both (3). B) Rank and probability of acceptance for each of these hypotheses for the four different datasets.

Figure 4. Intersection analysis of expanded, contracted and lost protein domain families in *M. tardigradum*, *R. varieornatus* and *H. dujardini*. Expansions were generally species-specific and only a small number was shared across all three species. Lost protein domains were largely shared across all three species. A) Expanded protein domains families, B) Contracted protein domain families and C) potentially lost protein domain families compared across the three tardigrade genomes. *Abbreviations: HD = H. dujardini, MT = M. tardigradum, RV = R. varieornatus*

Figure 5. Domain loss in three Ecdysozoan lineages The domain repertoire of ancestral species was reconstructed using Maximum Likelihood based on the accepted Cycloneuralia hypothesis (see Fig. 3, hypothesis 2). Colors: Red Circles = Loss estimates based on the complete protein sets; Orange Circles = Loss estimates based on protein sets without potential contaminations; Green Circles = Loss estimates based on the complete protein sets without potential contaminations and horizontally transferred genes

Supplemental Figure 1. Genome size estimation for *M. tardigradum*. The histograms of

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Supplemental Figure 2. Ensembl Metazoa genome selection based on visual inspection of their completeness values. Empirical density and the cumulative distribution of CEGMA completeness values. 75% was chosen as the final threshold. species indicated in red were excluded based on the threshold.

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Tables

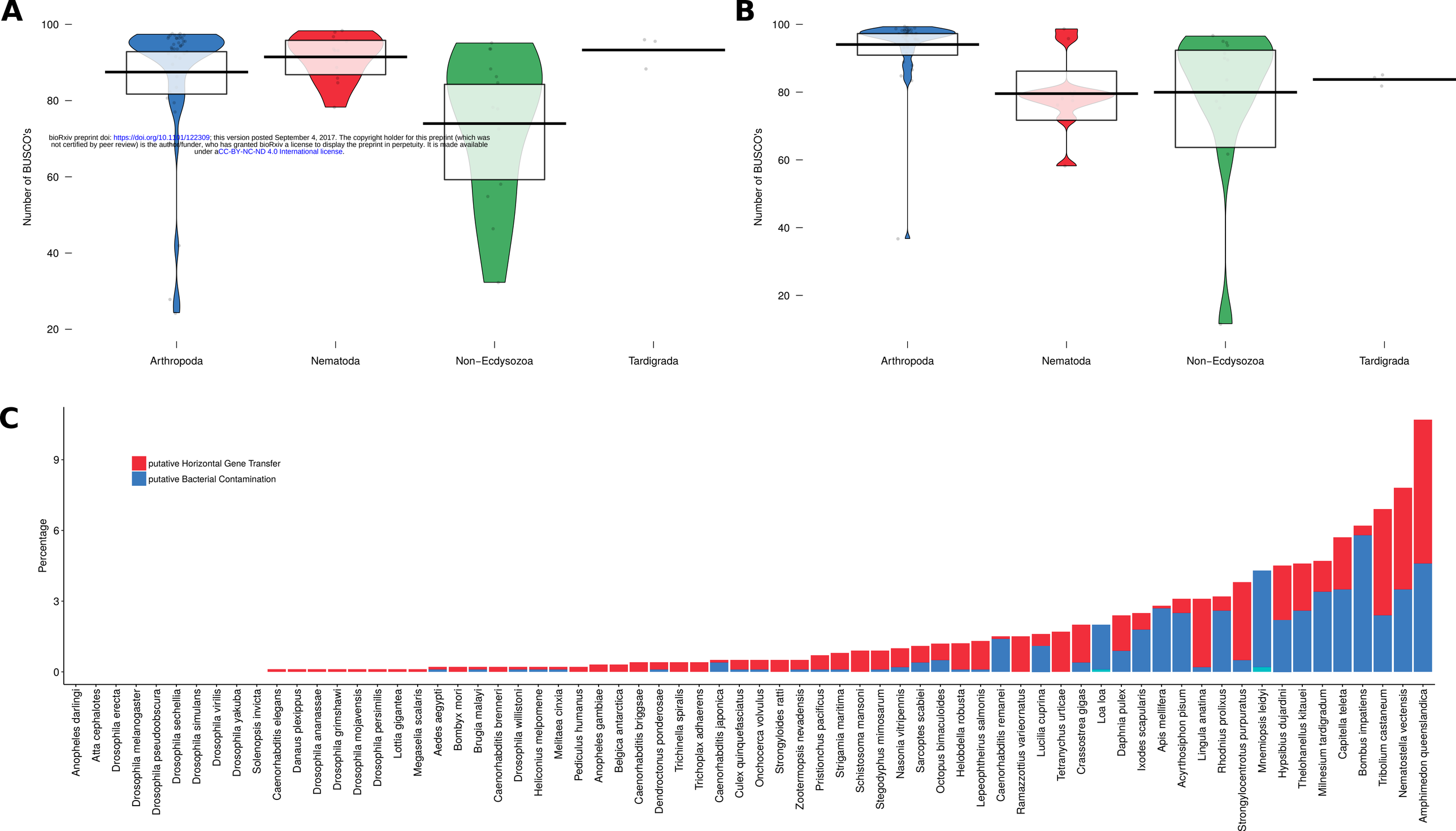
Table 1. Overview of components potentially involved in stress response identifiable in the three tardigrade genomes. First Block = Tardigrade-unique proteins present in the three tardigrade genomes. Second Block = Heat Shock Proteins present in the three tardigrade genomes. Third Block = Other stress-related proteins. *Abbreviations: MTAR = M. tardigradum; HDUJ = H. dujardini; RVAR = R. varieornatus.*

Name	MTAR	HDUJ	RVAR
CAHS	5	9	16
SAHS	0	5	13
MAHS	0	1	2
Dsup	0	0	1
HSP70	12	69	13
DNAJ	42	43	33
HSPB	9	11	7
HSPC	3	2	2
HSPD	17	13	10
HSPE	1	1	1
AOX	1	1	1
LEA	1	1	0

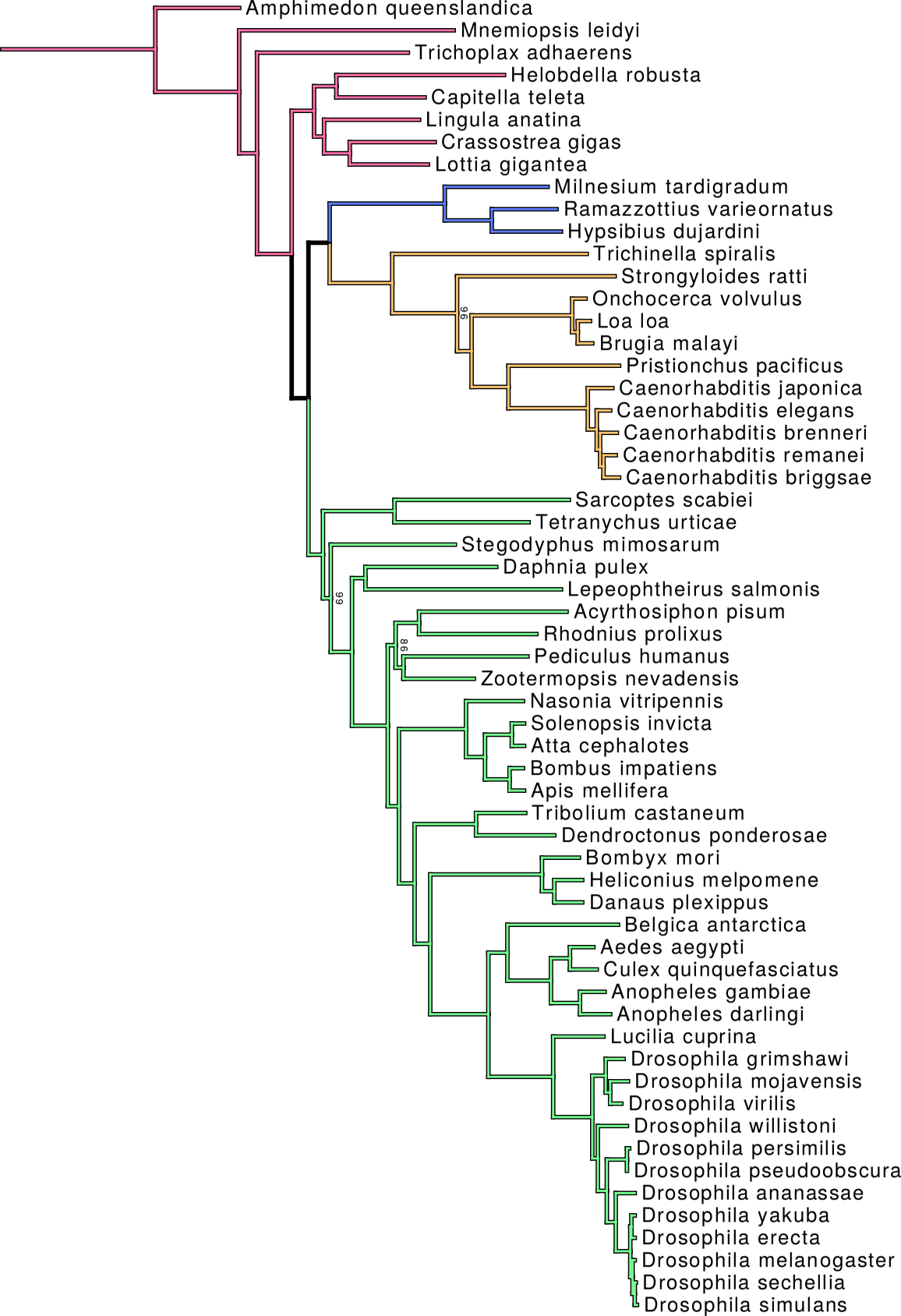
bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2. Overview of components potentially involved in the trehalose metabolism identifiable in the three tardigrade genomes. First Block = Trehalose Synthesis. Second Block = Trehalose Degradation. *Abbreviations: MTAR = M. tardigradum; HDUJ = H. dujardini; RVAR = R. varieornatus.*

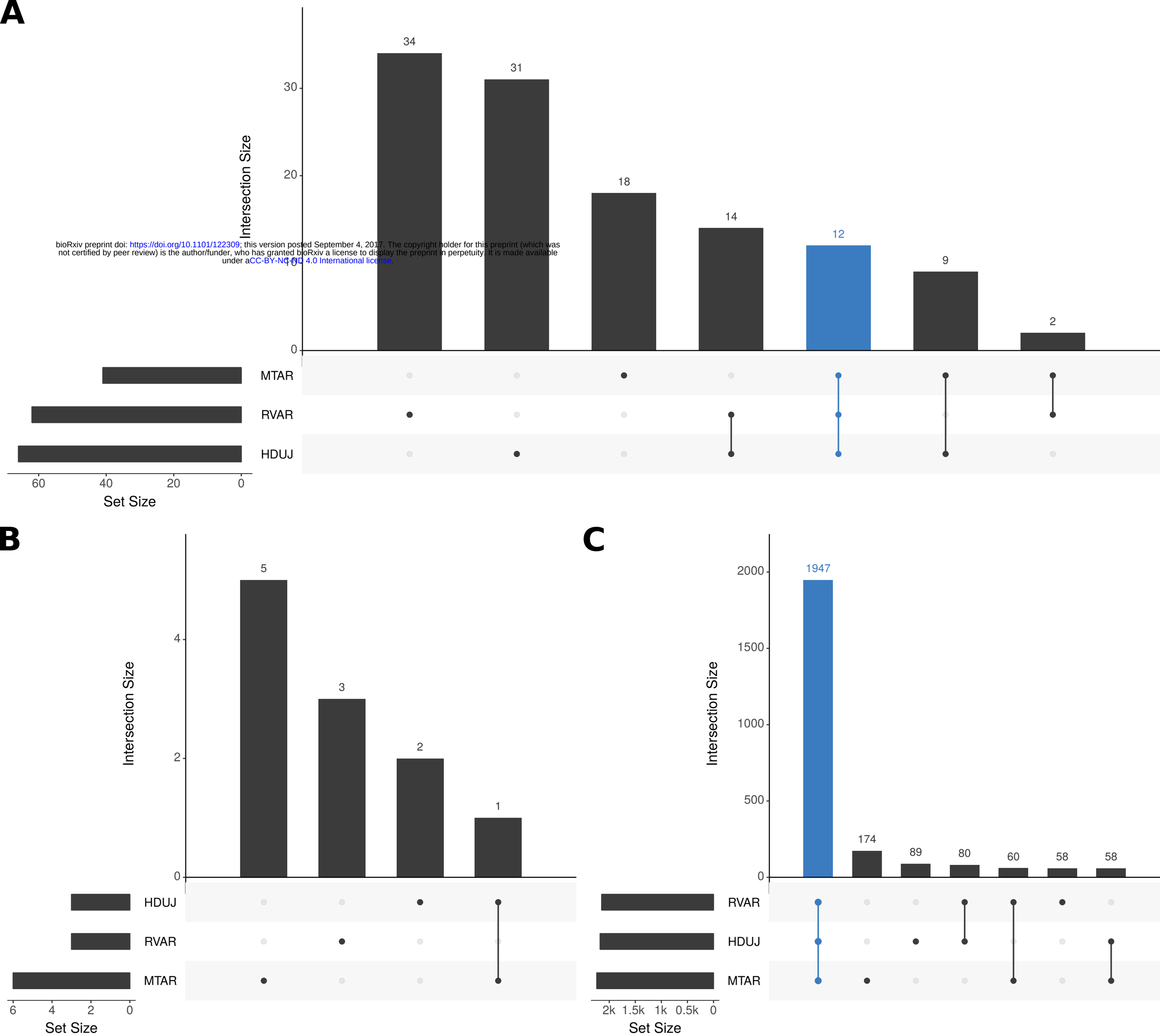
Name	Description	MTAR	HDUJ	RVAR
PF00128	Alpha amylase, catalytic domain	5	10	1
PF00534	Glycosyl transferases group 1	4	4	1
PF00982	Glycosyltransferase family 20	0	0	1
PF02358	Trehalose-phosphatase	0	0	1
PF03632	Glycosyl hydrolase family 65	7	3	1
PF01204	Trehalase	11	1	1



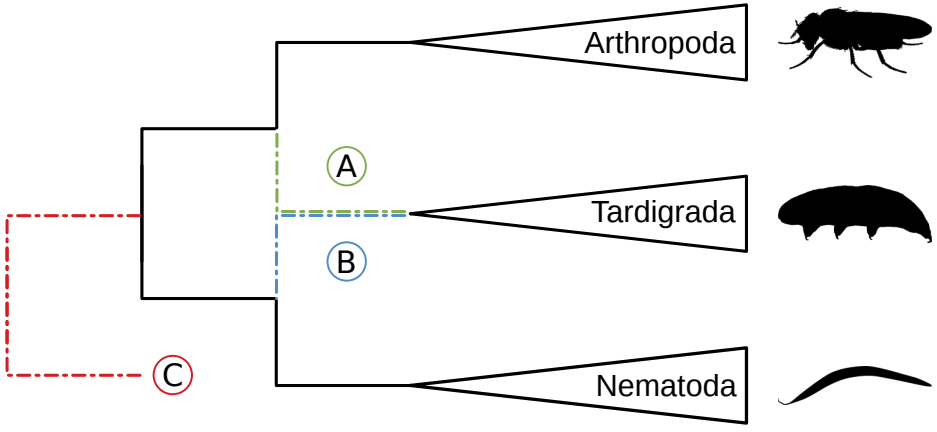
bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


Amphimedon queenslandica
Mnemiopsis leidyi
Trichoplax adhaerens
Helobdella robusta
Capitella teleta
Lingula anatina
Crassostrea gigas
Lottia gigantea
Milnesium tardigradum
Ramazzottius varieornatus
Hypsibius dujardini
Trichinella spiralis
Strongyloides ratti
Onchocerca volvulus
Loa loa
Brugia malayi
Pristionchus pacificus
Caenorhabditis japonica
Caenorhabditis elegans
Caenorhabditis brenneri
Caenorhabditis remanei
Caenorhabditis briggsae
Sarcoptes scabiei
Tetranychus urticae
Stegodyphus mimosarum
Daphnia pulex
Lepeophtheirus salmonis
Acyrthosiphon pisum
Rhodnius prolixus
Pediculus humanus
Zootermopsis nevadensis
Nasonia vitripennis
Solenopsis invicta
Atta cephalotes
Bombus impatiens
Apis mellifera
Tribolium castaneum
Dendroctonus ponderosae
Bombyx mori
Heliconius melpomene
Danaus plexippus
Belgica antarctica
Aedes aegypti
Culex quinquefasciatus
Anopheles gambiae
Anopheles darlingi
Lucilia cuprina
Drosophila grimshawi
Drosophila mojavensis
Drosophila virilis
Drosophila willistoni
Drosophila persimilis
Drosophila pseudoobscura
Drosophila ananassae
Drosophila yakuba
Drosophila erecta
Drosophila melanogaster
Drosophila sechellia
Drosophila simulans
99
99
99
98

bioRxiv preprint doi: https://doi.org/10.1101/122309; this version posted September 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



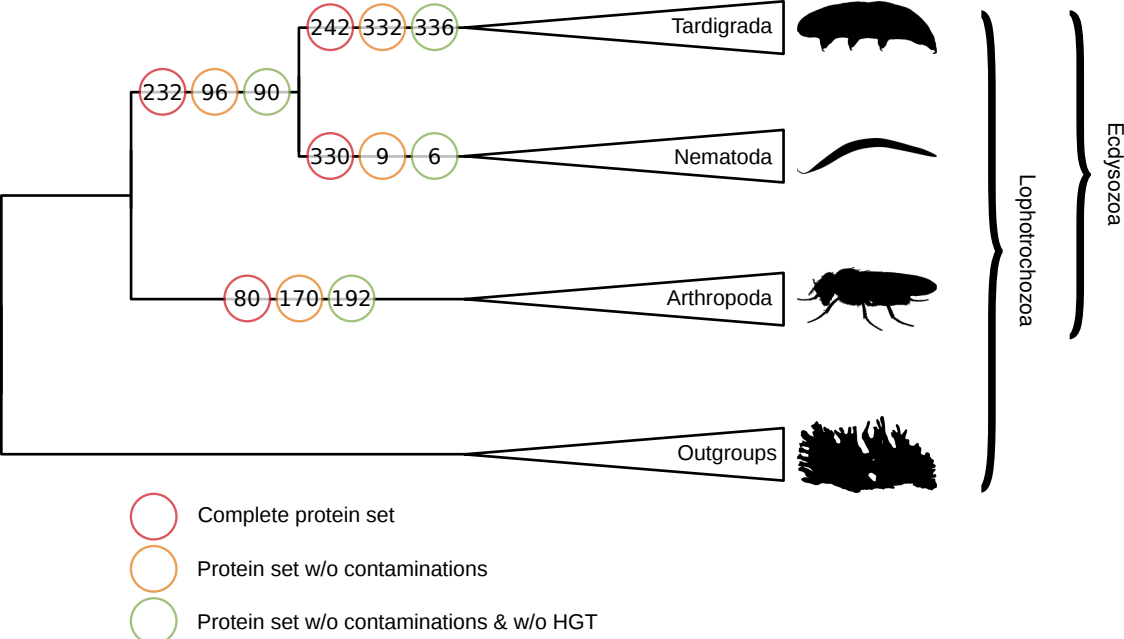
A



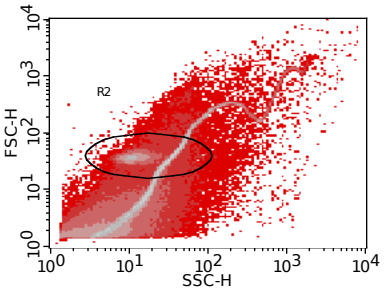
B

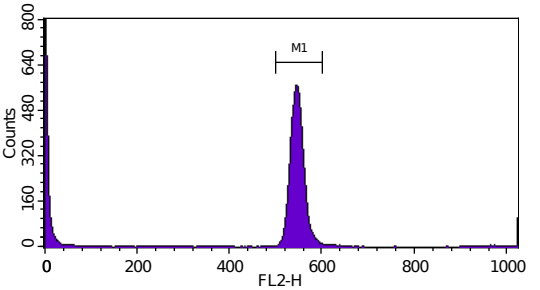
Consel Hypothesis Rank

Hypothesis	A	B	C
Sequence Supermatrix	2 ($1e^{-8}$)	**1 ($6e^{-9}$)**	3 (1.0)
Orthologous Groups	**1 ($3e^{-11}$)**	2 ($5e^{-11}$)	3 (1.0)
Domain Occurences	2 ($1e^{-53}$)	**1 ($3e^{-118}$)**	3 (1.0)
Domain Architectures	**1 ($5e^{-6}$)**	**1 ($5e^{-6}$)**	3 (1.0)


242
332
336
Tardigrada
232
96
90
330
9
6
Nematoda
80
170
192
Arthropoda
Outgroups
Lophotrochozoa
Ecdysozoa
Complete protein set
Protein set w/o contaminations
Protein set w/o contaminations & w/o HGT

A



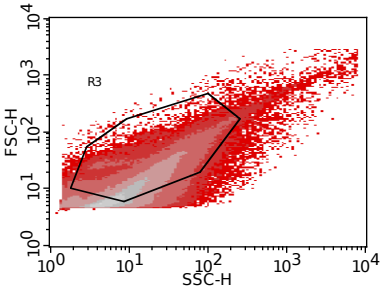


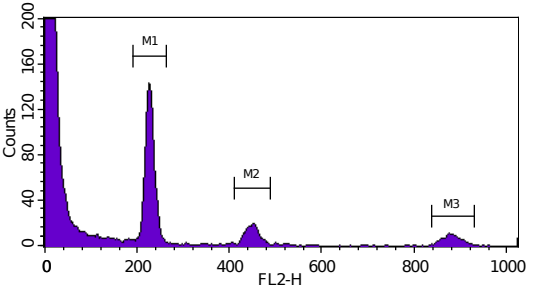
Histogram Statistics

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

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

B



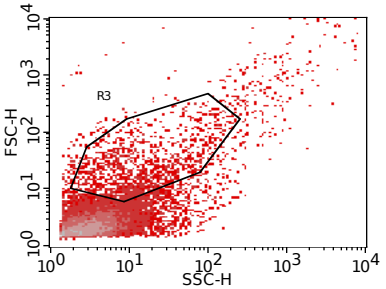


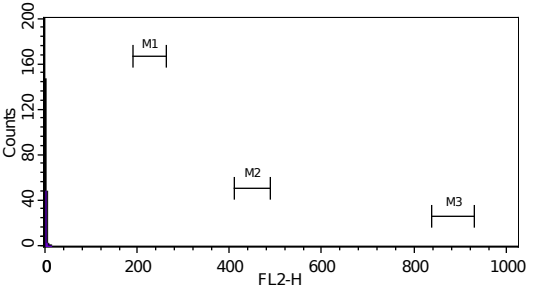
Histogram Statistics

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

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

C

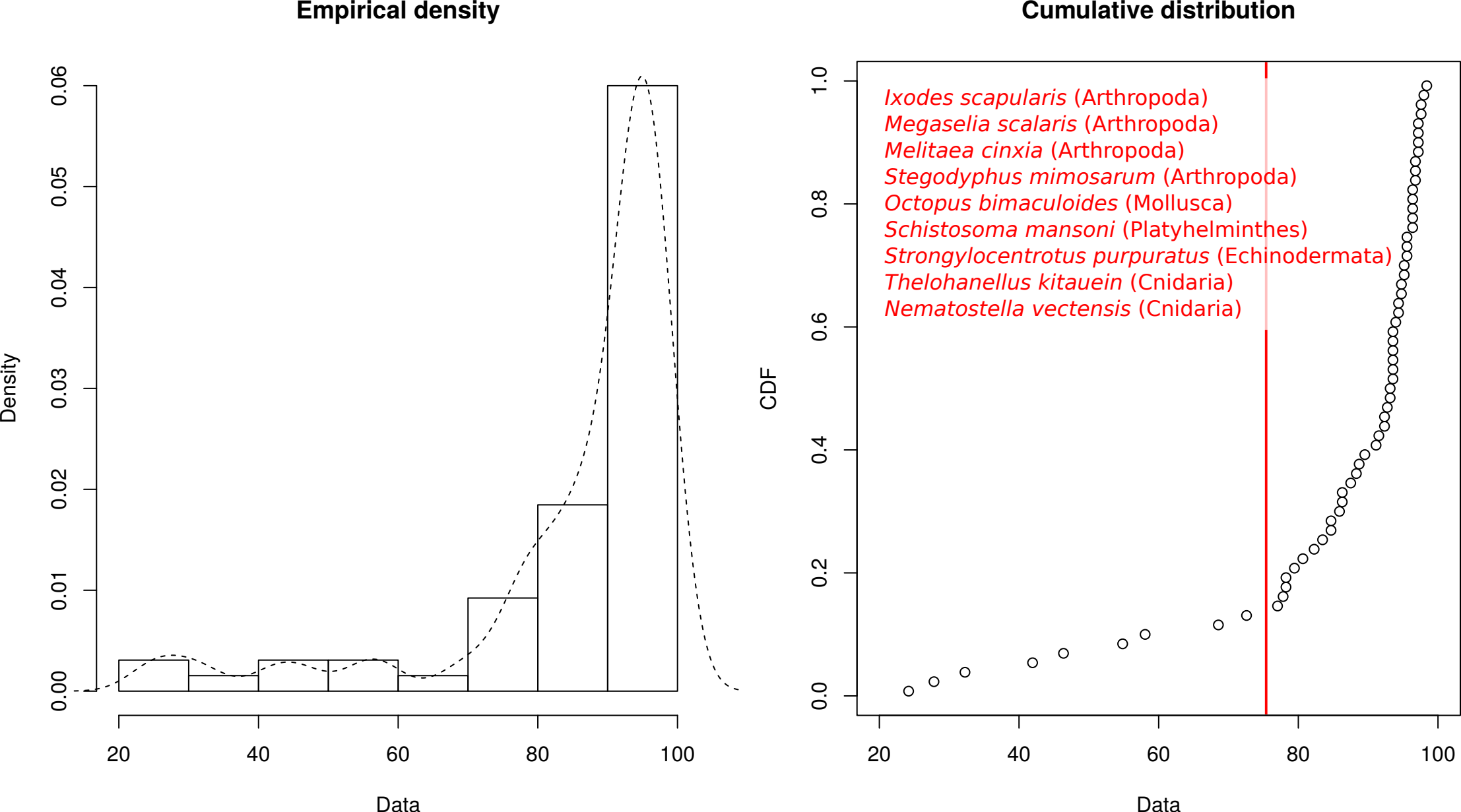




Histogram Statistics

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

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


Empirical density
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
20
40
60
80
100
Data
Cumulative distribution
CDF
0.0
0.2
0.4
0.6
0.8
1.0
Ixodes scapularis (Arthropoda)
Megaselia scalaris (Arthropoda)
Melitaea cinxia (Arthropoda)
Stegodyphus mimosarum (Arthropoda)
Octopus bimaculoides (Mollusca)
Schistosoma mansoni (Platyhelminthes)
Strongylocentrotus purpuratus (Echinodermata)
Thelohanellus kitauein (Cnidaria)
Nematostella vectensis (Cnidaria)
20
40
60
80
100
Data