GigaScience

The genome of Antarctic-endemic Copepod, Tigriopus kingsejongensis

--Manuscript Draft--

Manuscript Number:	GIGA-D-16-00040R1
Full Title:	The genome of Antarctic-endemic Copepod, Tigriopus kingsejongensis
Article Type:	Data Note
Funding Information:	Korea Polar Research Institute (KR) (PE16070) — Dr. Hyun Park Korea Polar Research Institute (PE14260) — Dr Sanghee Kim
Abstract:	Background: The Antarctic intertidal zone is continuously subject to extreme fluctuations in biotic and abiotic stressors, and the West Antarctic Peninsula is the most rapidly warming region on earth. Organisms living in Antarctic intertidal pools are therefore of great interest for research on topics such as evolutionary adaptation to extreme environments and the effects of climate change. Findings: Here, we report the whole-genome sequence of the Antarctic endemic Harpacticoid copepod, Tigriopus kingsejongensis with a total of 37 Gb raw DNA sequence using Illumina Miseq platform and the libraries were prepared with 65-fold coverage with a total length of 295 Mb. The final assembly consists of 48,368 contigs with an N50 contig length of 17.5 kb and 27,823 scaffolds with N50 contig length of 159.2 kb and a total of 12,772 coding genes were inferred using the MAKER annotation pipeline approach. Comparative genome analysis revealed that T. kingsejongensis specific genes are enriched in transport and metabolism processes. Furthermore, rapidly evolving genes related to energy metabolism showed signatures of positive selection. Conclusions: The genome of T. kingsejongensis will provide an interesting example of an evolutionary strategy for Antarctic cold adaptation, and offers new genetic insights into Antarctic intertidal biota.
Corresponding Author:	Hyun Park KOREA, REPUBLIC OF
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Seunghyun Kang
First Author Secondary Information:	
Order of Authors:	Seunghyun Kang Do-Hwan Ahn Jun Hyuck Lee Sung Gu Lee Seung Chul Shin Jungeun Lee Gi-Sik Min Hyoungseok Lee Hyun-Woo Kim Sanghee Kim Hyun Park

Order of Authors Secondary Information:

Response to Reviewers:

Subject: Your submission to GigaScience - GIGA-D-16-00040

GIGA-D-16-00040
The First Copepod Genome Reveals Evolutionary Adaptation to Extreme Environments in the Antartic-endemic Tigriopus Seunghyun Kang; Do-Hwan Ahn; Jun Hyuck Lee; Sung Gu Lee; Seung Chul Shin; Jungeun Lee; Gi-Sik Min; Hyoungseok Lee; Hyun-Woo Kim; Sanghee Kim; Hyun Park GigaScience

Dear Hans,
We would like to thank you and all the reviewers for your kind help to revise our manuscript and consider our manuscript for publication in GigaScience. We have decided to change the article type from “Research” to “Data Note” following your precious suggestion. Accordingly, we have reduced biological analysis part (especially transcriptome response to temperature stress), focused on genome part in response to Data Note criteria and wrote down response to reviewer comments about remaining parts. As we have changed the article type, we also modified the title as follows: “The genome of Antarctic-endemic Copepod Tigriopus kingsejongensis”.
We appreciated all of the comments and suggestions and carefully considered all of them during the revision. All of inferred statements have been corrected, and also all of mistakes have been corrected in the revised manuscript. The corrected points were marked in yellow color in revised manuscript.
We did our best to address the comments from the reviewers. Hope the revised is acceptable for publication. We look forward to hearing your decision.
Thanks and best regards.
Hyun Park

Reviewer reports:
Reviewer #1: The paper describes the genome of a recently described species of the harpacticoid genus Tigriopus. The species is interesting as it lives in the cold Antarctic environment. If published, it may well be the first paper describing a copepod genome; however, it should be noted that other copepod genomes are already available online, including the congener Tigriopus californicus (https://i5k.nal.usda.gov/Tigriopus_californicus). Some reference to this and other copepod genomes (Eurytemora affinis and Salmon louse) might be appropriate.

Response) We added publically available copepod genome data in data description part line 30.

The abstract and background of this paper both open with what is clearly a false statement - there are not more species of copepods than insects or nematodes; there are over 1 million species of insects vs. ~12,000 species of copepods, so it's not clear what the authors are stating. Perhaps this is a language problem, but it results in a very significant error.

Response) We made clear the sentences in abstract and data description line 21.

Background line 24 - outdated refs for temperature adaptation. Full transcriptome response to heat stress in T. californicus was published 4 years ago (Schoville et al 2012 BMC Evolutionary Biology) and would seem to be an especially appropriate reference here as the authors could contrast response to cold with the response to heat. (also consider Barreto et al. 2011 Interpopulation patterns of divergence and selection across the transcriptome of the copepod Tigriopus californicus. Molecular Ecology. 20:560-572). It might be interesting to see if there is any overlap between genes identified as under positive selection in this species versus those identified as under selection between T. californicus populations (see Pereira et al. 2016 Molecular Ecology).

Response) We thank the reviewer’s valuable comments about transcriptome response to temperature stress. At this moment, we have changed article type from “research note” to “data note” and need to reduce biological analysis. Following your suggestions, we are going to compare not only T. japonicus but also T. californicus in our future study.

As for the genome data itself, the assembly reported is interesting as the genome size is substantially larger than that of T. californicus (size based on nuclear fluorescence is ~240Mb). The assembly is rather fragmented >11,000 scaffolds and I wonder if they might see significant improvement if they used a different assembler (maybe try ALLOPATHS instead of Celera - they have the necessary data)?

Response) We have used the three assemblers, Abyss, SOAP, and Celera, and all the assembly resulted in almost same genome size about 295 Mb in accordance with k-mer genome size estimation. The best statistical results were obtained from Celera and we have used Celera assembly results in the following analysis processes. We have constructed relatively long paired-end library datasets: 350bp, 400bp, 450bp, and 500bp. As far as we know, AllpathLG need at least 20 % overlapping “fragment library”. Unfortunately, AllpathLG cannot be applied to our datasets.

The completeness statistic based on coverage of CEGMA is not great (82%) but they used the larger set of 458 conserved proteins rather than the more conservative set of 248 proteins often used. The authors do not report what percentage of the genes are fully (vs. partially) covered in their assembly - this might make it easier for readers to better interpret the results. I think this may be an issue if their assembly is rather incomplete in total coverage, the reported gains and losses of gene families may be unreliable. The number of annotated gene models 12,772 is a bit low (10% lower than the smaller T. californicus genome at ~14,100).
Response) We have modified and added complete and partial annotated gene numbers and percentages in data description line 129.

Reviewer #2:

However, I quickly became a bit perturbed. The authors claim that genomes and genomic resources are lacking for copepods, and that this is the first copepod genome paper. This is an odd statement, given that there are more genomes freely available for copepods than for any other crustacean, and a plethora of genomic resources available, relative to other non-insect arthropods. The salmon louse genome project had an official press release five years ago, and is available for analysis. In addition, two other copepod genomes, those of Tigriopus californicus and Eurytemora affinis, are freely available from the Arthropod i5K website, and have been available for over two years. And this study does make a comparison with the genome of Tigriopus japonicus, which has been around for a while. I have seen some comparative studies that incorporate analyses of these other copepod genomes, without calling them the "first genome." There are also many copepod transcriptomes freely available.

Response) The meaning of our previous title was the first genome paper describing copepods. Following your suggestions, we modified the title and added publically available copepod genome data in data description part line 30.

Given the availability of the other copepod genomes, this study would benefit from comparisons with at least the congener Tigriopus californicus, in addition to the congener T. japonicus.

Response) Thank you so much for your informative comments. At this moment, we have changed article type from “research note” to “data note” and need to reduce biological analysis. Following your valuable suggestions, we are going to compare T. californicus genome and transcriptome with T. kingsejongensis and T. japonicus in our future study.

The authors applied PAML to test for signatures of selection. I recommend that they also use HyPhy, which is more powerful, and able to make greater inferences than PAML.

Response) We have used PAML in this study according to following articles and we will apply HyPhy in the future study.
Cao, Z., et al. (2013). "The genome of Mesobuthus martensii reveals a unique adaptation model of arthropods." Nature communications 4: 2602.

	Neafsey, D. E., et al. (2015). "Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes." Science 347(6217): 1258522. Qiu, Q., et al. (2012). "The yak genome and adaptation to life at high altitude." Nature Genetics 44(8): 946-949. Yim, H.-S., et al. (2014). "Minke whale genome and aquatic adaptation in cetaceans." Nature Genetics 46(1): 88-92.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript.	Yes

Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	

The genome of Antarctic-endemic Copepod, *Tigriopus kingsejongensis*

Seunghyun Kang[1¶], Do-Hwan Ahn[1¶], Jun Hyuck Lee[1,2], Sung Gu Lee[1,2], Seung Chul Shin[1], Jungeun Lee[1], Gi-Sik Min[3], Hyoungseok Lee[1], Hyun-Woo Kim[4*&], Sanghee Kim[5*&] & Hyun Park[1,2*&]

[1] Unit of Polar Genomics, Korea Polar Research Institute, Yeonsu-gu, Incheon, South Korea

[2] Polar Sciences, University of Science & Technology, Yuseong-gu, Daejeon, South Korea

[3] Department of Biological Sciences, Inha University, Incheon, South Korea

[4] Department of Marine Biology, Pukyong National University, Busan, South Korea

[5] Division of Polar Life Sciences, Korea Polar Research Institute, Yeonsu-gu, Incheon, South Korea

* Corresponding author

E-mail: kimhw@pknu.ac.kr; sangheekim@kopri.re.kr; hpark@kopri.re.kr

[¶]These authors contributed equally to this work.

[&]These authors also contributed equally to this work.

Abstract

Background: The Antarctic intertidal zone is continuously subject to extreme fluctuations in biotic and abiotic stressors, and the West Antarctic Peninsula is the most rapidly warming region on earth. Organisms living in Antarctic intertidal pools are therefore of great interest for research on topics such as evolutionary adaptation to extreme environments and the effects of climate change.

Findings: Here, we report the whole-genome sequence of the Antarctic endemic Harpacticoid copepod, *Tigriopus kingsejongensis* with a total of 37 Gb raw DNA sequence using Illumina Miseq platform and the libraries were prepared with 65-fold coverage with a total length of 295 Mb. The final assembly consists of 48,368 contigs with an N50 contig length of 17.5 kb and 27,823 scaffolds with N50 contig length of 159.2 kb and a total of 12,772 coding genes were inferred using the MAKER annotation pipeline approach. Comparative genome analysis revealed that *T. kingsejongensis* specific genes are enriched in transport and metabolism processes. Furthermore, rapidly evolving genes related to energy metabolism showed signatures of positive selection.

Conclusions: The genome of *T. kingsejongensis* will provide an interesting example of an evolutionary strategy for Antarctic cold adaptation, and offers new genetic insights into Antarctic intertidal biota.

Keywords: Copepoda, Genome, Antarctic, adaptation, *Tigriopus*

Data description

The subclass copepods are very diverse and approximately 12,000 copepod species have been described [1, 2]. They dominate the zooplankton community contributing about 70% of total zooplankton biomass [3] and play an important role in the marine meiobenthic food web linking between the phytoplankton and higher trophic levels [4]. Harpacticoid copepods of the genus *Tigriopus* Norman 1868 are dominant members of shallow supratidal rock pools worldwide. They are distributed among habitats that vary widely in salinity, temperature, desiccation risk, and UV radiation, and have been used as a model system to investigate topics such as osmoregulation [5], temperature adaptation [6, 7] and environmental toxicology [8]. As the genome resources of copepods has been publically available (*Tigriopus californicus* [http://i5k.nal.usda.gov/Tigriopus_californicus], *Tigriopus japonicus* [9], *Eurytemora affinis* [http://i5k.nal.usda.gov/Eurytemora_affinis] and salmon louse *Lepeophtheirus salmonis* [http://sealouse.imr.no/]), now it is possible to explore their fundamental biological processes and physiological responses to diverse environments.

Antarctica provides not only an extreme habitat for extant organisms, but also a model for research on evolutionary adaptations to cold environments [10, 11]. The Antarctic intertidal zone, particularly in the Western Antarctic Peninsula region, is one of the most extreme environments on earth. It also serves as a potential barometer for global climate changes, since it is the fastest-warming region on earth [12]. Antarctic intertidal species that have evolved stenothermal phenotypes through adaptation to a year-round climate of extreme cold may now face extinction by global warming. The response of these species to further warming in Western Antarctica is of serious concern; however, to date there have been few studies focusing on species from the Antarctic intertidal zone.

Tigriopus kingsejongensis was first found and recognized as a new endemic species in a rock pool in the Antarctic Peninsula, and is extremely cold-tolerant and can survive in frozen sea water [13]. We observed the morphological differences, such as increased numbers of caudal setae in nauplii, optimal growth temperature (ca. 8°C) and developmental characteristics have been compared to those of the congener *Tigriopus japonicus*, which is found in the coastal area of the Yellow Sea. *Tigriopus kingsejongensis* has evolved to overcome the unique environmental constrains of Antarctica, and therefore provides an ideal experimental model for all aspects of research on extreme habitats. This species may represent a case of rapid speciation, since the intertidal zone on King George Island and surrounding areas did not exist before 10,000 years ago [14]. *Tigriopus kingsejongensis* likely evolved as a distinct species within this relatively short time period. Thus, inter- and intraspecies comparative analyses of Antarctic *Tigriopus* species will help define the trajectory of adaptation to the Antarctic environment and also provide insights into the genetic basis of *Tigriopus* divergence and evolution.

Library construction and sequencing

Tigriopus kingsejongensis were collected from tidal pools in Potter Cove, near King Sejong Station, on the northern Antarctic Peninsula (62°14'S, 58°47'W) (Fig. S1 and S2 in additional file1) in January 2013 with a hand-nets. Water temperatures were 1.6 ± 0.8°C during this month. High-molecular-weight genomic DNA from pooled *T. kingsejongensis* was extracted using the DNeasy Blood & Tissue Kit (Qiagen). For Illumina Miseq sequencing, four library types were constructed with 350, 400, 450, and 500 bp for paired-end libraries, and 3 kb and 8 kb for mate-pair libraries, prepared using the standard Illumina sample preparation methods (Table 1). All sequencing processes were performed according to the manufacturer’s instructions (Illumina).

RNA was prepared from pooled *T. kingsejongensis* and *Tigriopus japonicus* specimens from two different temperature experiments (4°C and 15°C) using the RNeasy Mini Kit (Qiagen). For Illumina Miseq sequencing, subsequent experiments were carried out under the manufacturer’s instructions (Illumina). The *de novo* transcriptome assembly was performed with CLC Genomics Workbench, setting the minimum allowed contig length to 200 nucleotides. The assembly process generated 40,172 contigs with a max length of 23,942 bp and an N50 value of 1,093 bp. These generated contigs were used as reference sequences for mapping of trimmed reads, and fold changes in expression for each gene were calculated with a significance threshold of $P \leq 0.05$ using CLC Genomics Workbench (Table 2 and 3).

Genome assembly

First, k-mer analysis was conducted using jellyfish 2.2.5 [15] to estimate the genome size from DNA paired-end libraries. The estimateds genome size was 298 Mb with main peak at a depth of ~39x (Fig. 1). Then, assemblies were performed using a Celera Assembler with Illumina short reads [16]. Prior to assembly, Illumina reads were trimmed using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) with parameters -t 20, -l 70 and -Q 33, after which a paired sequence from trimmed Illumina reads was selected. Finally, trimmed Illumina reads with 65-fold coverage (insert sizes 350, 400, 450, and 500 bp) were obtained and converted to the FRG file format (required by the Celera assembler) using FastqToCA. Assembly was performed on a 96-processor workstation with Intel Xeon X7460 2.66 GHz processors and 1 terabyte RAM with the following parameters: overlapper = ovl, unitigger = bogart, utgGraphErrorRate = 0.03, utgGraphErrorLimit = 2.5, utgMergeErrorRate = 0.030, utgMergeErrorLimit = 3.25, ovlErrorRate = 0.1, cnsErrorRate = 0.1, cgwErrorRate = 0.1, merSize = 22, and doOverlapBasedTrimming = 1. The initial Celera assembly had a total size of 305 Mb, N50 contig size of 17,566 bp, and max contig size of 349.5 kb. Scaffolding was completed using the software SSPACE 2.0 scaffolder using mate-paired data [17]. Subsequently, we closed gaps using Gapfiller Ver.1.9 software with 65× trimmed Illumina reads with default settings [18]. *De novo* assembly of 203 million reads from paired-end libraries and mate-paired libraries yielded a draft assembly (65-fold coverage) with a total length of 295 Mb, and contig and scaffold N50 sizes of 17.6 kb and 159.2 kb, respectively

(Table 4 and Fig. 2).

Annotation

We used MAKER for genome annotation [19]. MAKER is a portable and easily configurable genome annotation pipeline. MAKER first identified repetitive elements using RepeatMasker [20]. This masked genome sequence was used for *ab initio* gene prediction with SNAP software [21], after which alignment of expressed sequence tags with BLASTn and protein information from tBLASTx were included. We used the *de novo* repeat library of *T. kingsejongensis* from RepeatModeler for RepeatMasker; proteins from five species with data from *D. melanogaster*, *D. pulex*, *T. japonicus*, and *Tigriopus californicus* were included in the analysis. RNA-seq-based gene prediction was performed by aligning all RNA-seq data against the assembled genome using TopHat [22], and Cufflinks [23] was used to predict cDNAs from the resultant data. Next, MAKER polished the alignments using the program Exonerate [24], which provided integrated information to synthesize SNAP annotation. MAKER then selected and revised the final gene model considering all information. A total of 12,772 genes were predicted using MAKER in *T. kingsejongensis*. Annotated genes contained an average of 4.6 exons, with an average mRNA length of 1,090 bp. Additionally, 12,562 of 12,772 genes were assigned preliminary functions based on automated annotation using Blast2GO (Ver. 2.6.0) [25] (Fig. S3 and S4 in additional file 1). The Infernal software package (Ver. 1.1) [26] and covariance models (CMs) from the Rfam database [27] were used to identify other non-coding RNAs in the *T. kingsejongensis* scaffold. We identified putative tRNA genes using tRNAscan-SE [28] (Table S1 in additional file 2). tRNAscan-SE uses a covariance model (CM) that scores candidates based on their sequence and predicted secondary structures.

Non-gap sequences occupied 284.8 Mb (96.5%), and simple sequence repeats (SSRs) were 1.2 Mb (0.4%) in total (Table S2 in additional file 2). Transposable elements (TEs) comprised 6.5 Mb, which is roughly 2.3% of the assembled genome (Table S2 and S3 in additional file 2). On the basis of homology and *ab initio* gene prediction, we found that the genome of *T. kingsejongensis* contains 12,772 protein-coding genes (Table 5). By assessing the quality of the annotated 12,772 gene models, we found that 11,686 protein-coding genes (91.5%) were supported by the RNA-seq data, of which, 7,325 (63%) showed similarity to proteins from other species. We also found that 179 of 248 CEGMA (Core Eukaryotic Genes Mapping Approach) core genes [29] were fully annotated (72.18 % completeness) and 197 of 248 genes were partially annotated (79.44 % completeness).

Gene families

The orthologous groups were identified from 11 species (*T. kingsejongensis, Aedes aegypti, D. melanogaster, Ixodes scapularis, M. martensii, Strigamia martima, Tetranychus urticae, D. pulex, Homo sapiens, Ciona intestinalis,* and *Caenorhabditis elegans*) (Table 6) using OrthoMCL [30] with standard parameters and options, and transcript variants other than the longest translation forms were removed. For *T. kingsejongensis*, the coding sequence from the MAKER annotation pipeline was used. The 1:1:1 single-copy orthologous genes were subjected to phylogenetic construction and divergence time estimation. Protein-coding genes were aligned using PRANK with the codon alignment option [31], and poorly aligned sequences with gaps were removed using Gblock under the codon model [32]. We constructed a maximum-likelihood phylogenetic tree using RAxML with 1,000 bootstrap values [33] and calibrated the divergence time between species with TimeTree [34]. Finally, the average gene gain/loss rate along the given phylogeny was identified using the program CAFÉ 3.1 [35]. We constructed orthologous gene clusters using four arthropod species (Antarctic copepod, *T. kingsejongensis*; scorpion, *Mesobuthus martensii*; fruit fly, *Drosophila melanogaster* and water flea, *Daphnia pulex*) to compare the genomic features and the adaptive divergence in the arthropods. In total, 2,063 gene families are shared by all four species, and 1,028 genes are specific to the Antarctic copepod. *Tigriopus kingsejongensis* shares 4,559 (73.5%) gene families with *D. pulex*, which belongs to the same crustacean lineage Vericrustacea, 3,531 (56.9%) with *D. melanogaster*, and 3,231 (52.1%) with *M. martensii* (Fig. 3A). Gene ontology (GO) analysis revealed that the 1,028 *T. kingsejongensis*-specific genes are enriched in transport (single-organism transport, GO: 0044765; transmembrane transport, GO: 0055085; ion transport, GO: 0006811; cation transport, GO: 0006812) and single-organism metabolic processes (GO: 0044710) (Table S4 and S5 in additional file 2). Subsequently, we performed gene gain-and-loss analysis on 11 representative species, and found that *T. kingsejongensis* gained 735 gene families and lost 4,401 gene families (Fig. 3B). Thus, this species exhibits a gene family turnover of 5,136, the largest value among the eight arthropods. We also analyzed expansion and contraction of the gene families (Table S6-S9 in additional file 2), and found 232 significantly expanded gene families in *T. kingsejongensis*; these gene families are significantly overrepresented in amino acid metabolism and carbohydrate metabolism in KEGG metabolic pathways.

Genome evolution

Adaptive functional divergence caused by natural selection is commonly estimated based on the ratio of nonsynonymous (*dN*) to synonymous (*dS*) mutations. To estimate *dN*, *dS*, and average *dN*/*dS* ratio (*w*), and lineage-specific PSGs in *T. kingsejongensis* and *T. japonicus*, protein-coding genes from *T. japonicus* were added to define orthologous gene families among the four species (*T. kingsejongensis, T. japonicus, D. pulex,* and *D. melanogaster*) using the program OrthoMCL with the same conditions previously described. We identified 2,937 orthologous groups shared by all four species, and single-copy gene families were used to construct a phylogenetic tree and estimate the time since divergence using the same methods described above. Each of the identified orthologous genes was aligned using the PRANK, and poorly aligned sequences with gaps were removed using Gblock. Alignments showing less than 40% identity and genes shorter than 150 bp were eliminated in subsequent procedures. The values of *dN*, *dS* and *w* were estimated from each gene using the Codeml program implemented in the PAML package with the free-ratio model [36] under F3X4 codon frequencies, and orthologs with $w \leq 5$ and $dS \leq 3$ were retained [37]. To examine the accelerated nonsynonymous divergence in either *T. kingsejongensis* or *T. japonicus* lineage, a binomial test [38] was used to determine GO categories with at least 20 orthologous genes. To define PSGs in *T. kingsejongensis* and *T. japonicus*, we applied basic and branch-site models, and Likelihood Ratio Tests (LRTs) were used to remove genes under relaxation of selective pressure. To investigate which functional categories and pathways were enriched in the PSGs, we performed DAVID Functional Annotation [39] with Fisher's exact test (cutoff: $P \leq 0.05$).

The average *w* value from 2,937 co-orthologous genes of *T. kingsejongensis* (0.0027) is higher than that of *T. japonicus* (0.0022). The GO categories that showing evidence of accelerated evolution in *T. kingsejongensis* are energy metabolism (generation of precursor metabolites and energy, GO: 0006091; cellular respiration, GO: 0045333) and carbohydrate metabolism (monosaccharide metabolic process, GO: 0005996; hexose metabolic process, GO: 0019318) (Figure 4A, Table S10 in Additional file 2). Branch-site model analysis showed that the genes belonging to the functional categories above have undergone a significant positive selection process by putative functional divergence in certain lineages. There are 74 and 79 positively selected genes (PSGs) in *T. kingsejongensis* (Table S11 in Additional file 2) and *T.*

japonicus (Table S12 in Additional file 2), respectively. The functional categories enriched in *T. kingsejongensis,* when compared to *T. japonicus,* support the idea that the functional divergence in *T. kingsejongensis* is strongly related to energy metabolism (oxidative phosphorylation, GO: 0006119; energy-coupled proton transport down electrochemical gradient, GO: 0015985; ATP synthesis-coupled proton transport, GO: 0015986; generation of precursor metabolites and energy, GO: 0006091) (Figure 4B, Table S13 and S14 in Additional file 2). In particular, three of the identified genes are involved in the oxidative phosphorylation (OxPhos) pathway, which provides the primary cellular energy source in the form of adenosine triphosphate (ATP). These three genes are nuclear-encoded mitochondrial genes: the catalytic F1 ATP synthase subunit alpha (*ATP5A*) (Fig. S5 in Additional file 1), a regulatory subunit acting as an electron transport chain such as ubiquinol-cytochrome *c* reductase core protein (*UQCRC1*) (Fig. S6 in Additional file 1), and an electron transfer flavoprotein alpha subunit (*ETFA*) (Fig. S7 in Additional file 1).

Availability of supporting data

The data for *T. kingsejongensis* genome and transcriptome has been deposited in the SRA as BioProject PRJNA307207 and PRJNA307513, respectively.

List of abbreviations

simple sequence repeats, SSRs; Transposable elements, TEs; CEGMA, Core Eukaryotic Genes Mapping Approach; Gene ontology, GO

Competing interests

The authors declare no competing interests.

Funding

This work was supported by an Antarctic organisms: Cold-adaptation mechanism and its application grant (PE16070) and the basic research program (PE14260) funded by the Korea Polar Research Institute (KOPRI).

Author contributions

H.P., Sanghee Kim and H.W.K. conceived and designed experiments and analyses; Seunghyun Kang, D.-H.A., S.G.L., S.C.S., J.L., G.S.M. and H.L. performed experiments and conducted bioinformatics. Seunghyun Kang, H.W.K., Sanghee Kim and H.P. wrote the paper.

Acknowledgements

We would like to thank Joseph A. Covi for comments and discussion.

References

1. Huys R, Boxshall GA: *Copepod evolution.* Ray Society; 1991.

2. Humes AG: **How many copepods?** *Hydrobiologia* 1994, **292:**1-7.

3. Wells P, Persoone G, Jaspers E, C. C: *Marine ecotoxicological tests with zooplankton. In: Persoone, G., Jaspers, E., Claus, C. (Eds.), Ecotoxicological Testing for the Marine Environment.* Inst. Mar. Sci. Res., Bredene; 1984.

4. Ruppert E, Fox R, Barnes R: **Invertebrate Zoology, A Functional Evolutionary Approach. Brooks/Cole-Thomson Learning.** *Belmont, CA* 2003.

5. Goolish E, Burton R: **Energetics of osmoregulation in an intertidal copepod: Effects of anoxia and lipid reserves on the pattern of free amino accumulation.** *Funct Ecol* 1989**:**81-89.

6. Lazzaretto I, Libertini A: **Karyological comparison among different Mediterranean populations of the genus *Tigriopus* (Copepoda Harpacticoida).** *Boll Zool* 2009, **53:**197-201.

7. Davenport J, Barnett P, McAllen R: **Environmental tolerances of three species of the harpacticoid copepod genus *Tigriopus*.** *J Mar Biol Assoc UK* 1997, **77:**3-16.

8. Raisuddin S, Kwok KW, Leung KM, Schlenk D, Lee J-S: **The copepod *Tigriopus*: A promising marine model organism for ecotoxicology and environmental genomics.** *Aquat Toxicol* 2007, **83:**161-173.

9. Lee J-S, Rhee J-S, Kim R-O, Hwang D-S, Han J, Choi B-S, Park GS, Kim I-C, Park HG, Lee Y-M: **The copepod Tigriopus japonicus genomic DNA information (574Mb) and molecular anatomy.** *Mar Environ Res* 2010, **69:**S21-S23.

10. Thorne MAS, Kagoshima H, Clark MS, Marshall CJ, Wharton DA: **Molecular analysis of the cold tolerant Antarctic Nematode, *Panagrolaimus davidi*.** *PLOS one* 2014, **9:**e104526.

11. Everatta MJ, Worlandb MR, Balea JS, Conveyb P, Hayward SAL: **Pre-adapted to the maritime Antarctic? – Rapid cold hardening of the midge, *Eretmoptera murphyi*.** *J Insect*

Physiol 2012, **58:**1104–1111.

12. Bromwich DH, Nicolas JP, Monaghan AJ, Lazzara MA, Keller LM, Weidner GA, Wilson AB: **Central West Antarctica among the most rapidly warming regions on Earth.** *Nature Geoscience* 2013, **6:**139-145.

13. Park E-O, Lee S, Cho M, Yoon SH, Lee Y, Lee W: **A new species of the genus *Tigriopus* (Copepoda: Harpacticoida: Harpacticidae) from Antarctica.** *Proc Biol Soc Wash* 2014, **127:**138-154.

14. Birkenmajer K: **Geology of Admiralty Bay, King George Island (South Shetland Islands). An outline.** *Pol Polar Res* 1980, **1:**29-54.

15. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764-770.

16. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA, et al: **A whole-genome assembly of *Drosophila*.** *Science* 2000, **287:**2196-2204.

17. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27:**578-579.

18. Nadalin F, Vezzi F, Policriti A: **GapFiller: a *de novo* assembly approach to fill the gap within paired reads.** *BMC Bioinformatics* 2012, **13:**S8.

19. Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics* 2011, **12:**491.

20. Smit AFA HR, Green, P.: **RepeatMasker Open-3.0. 1996-2004 (http://www.RepeatMakser.org).**

21. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.

22. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105-1111.

23. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated**

transcripts and isoform switching during cell differentiation. *Nat Biotech* 2010, **28:**511-515.

24. Slater GS, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6:**31.

25. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21:**3674-3676.

26. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25:**1335-1337.

27. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A: **Rfam: Wikipedia, clans and the "decimal" release.** *Nucleic Acids Res* 2011, **39:**D141-145.

28. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25:**955-964.

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

30. Li L, Stoeckert CJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.

31. Löytynoja A, Goldman N: **An algorithm for progressive multiple alignment of sequences with insertions.** *Proc Natl Acad Sci U S A* 2005, **102:**10557-10562.

32. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540-552.

33. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.

34. Hedges SB, Dudley J, Kumar S: **TimeTree: a public knowledge-base of divergence times among organisms.** *Bioinformatics* 2006, **22:**2971-2972.

35. Han MV, Thomas GW, Lugo-Martinez J, Hahn MW: **Estimating gene gain and loss rates in**

the presence of error in genome assembly and annotation using CAFE 3. *Mol Biol Evol* 2013, **30:**1987-1997.

36. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586-1591.

37. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW: **Comparative genomics reveals insights into avian genome evolution and adaptation.** *Science* 2014, **346:**1311-1320.

38. Consortium TCSaA: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437:**69-87.

39. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature protocols* 2008, **4:**44-57.

Figure legends

Figure 1 Estimation of the *T. kingsejongensis* genome size based on 33-mer analysis. The x-axis represents the depth (peak at 39X) and the y-axis represents the proportion. The genome size was estimated as 298 Mb (total k-mer number/volume peak).

Figure 2 Scaffold and contig size distributions of *T. kingsejongensis.* The percentage of the assembly included (y-axis) in contigs or scaffolds of a minimum size (x-axis, log scale) is shown for the contig (red) and scaffold (blue).

Figure 3 Comparative genome analyses of the *T. kingsejongensis* genome. **a** Venn diagram of orthologous gene clusters between the four arthropod lineages. **b** Gene family gain-and-loss analysis. The number of gained gene families (red), lost gene families (blue) and orphan gene families (black) are indicated for each species. Time lines specify divergence times between the lineages.

Table legends

Table 1 Statistics for each DNA library.

Table 2 Sequencing and assembly results of transcriptome analysis of *T. japonicus.*

Table 3 Sequencing statistics of RNA-seq analysis of *T. kingsejongensis*.

Table 4 Statistics of genome assembly.

Table 5 General statistics of genes in *T. kingsejongensis.*

Table 6 Summary of orthologous gene clusters in the 11 representative species.

Table 1 Statistics for each DNA library.

Library		Reads (n)	Ave. length	Sequences (bp) (n)	Reads (trimmed) (n)	Ave. length	Sequences (trimmed) (n)
Paired-end	Sum	99,710,266		29,271,916,613	65,644,374		14,668,956,871
	350S1	6,661,392	300	2,005,078,992	4,446,394	233	1,034,231,244
	350S2	4,933,058	265	1,311,700,122	4,618,711	211	975,471,763
	400S1	65,668,598	300	19,766,247,998	36,863,154	228	8,397,426,481
	450S1	3,418,988	300	1,029,115,388	2,812,455	230	646,302,159
	450S2	8,009,162	245	1,968,652,020	7,660,814	199	1,527,566,312
	500S1	11,019,068	289	3,191,122,093	9,242,846	226	2,087,958,911
Mate-Paired	Sum	103,373,998		7,753,049,850	73,515,391		5,169,006,268
	3KS1	8,374,238	75	628,067,850	6,745,546	73	493,099,413
	3KS2	9,250,994	75	693,824,550	5,281,513	65	344,618,723
	3KS3	51,349,594	75	3,851,219,550	39,147,167	72	2,816,638,666
	3KS4	3,063,232	75	229,742,400	1,740,986	65	112,554,745
	8KS1	9,847,636	75	738,572,700	7,887,612	73	572,246,251
	8KS2	16,322,038	75	1,224,152,850	9,653,293	65	630,842,698
	8KS3	5,166,266	75	387,469,950	3,059,274	65	199,005,774
Total		203,084,264		37,024,966,463	139,159,765		19,837,963,139
Coverage (folds)				120.7			64.7

Table 2 Sequencing and assembly results of transcriptome analysis of *T. japonicus*.

Sequencing	
Total reads (n)	37,956,160
Total bases (n)	7,714,415,316
Trimmed reads (n)	35,577,636
Trimmed bases (n)	5,989,188,343
Assembly	
Contigs (n)	40,172
Total contig length (bases)	28,850,726
N50 contig length (bases)	1,093
Max scaffold length (bases)	23,942
Annotation	
With blast results	20,392
Without blast hits	7,090
With mapping results	8,172
Annotated sequences	4,518

Click here to download Table Table 3.docx

Table 3 Sequencing statistics of RNA-seq analysis of *T. kingsejongensis*.

	4°C	**15°C**
Total reads (n)	15,786,118	16,417,072
Total bases (n)	3,567,662,668	3,763,295,032
Trimmed reads (n)	14,845,103	15,388,513
Trimmed bases (n)	2,761,189,158	2,833,805,442

Table 4 Statistics of genome assembly.

		Celera assembler (Version : 8.0)
Scaffold	Total scaffold length (bases)	295,233,602
	Gap size (bases)	10,474,460
	Scaffolds (n)	11,558
	N50 scaffold length (bases)	159,218
	Max scaffold length (bases)	3,401,446
Contig	Total contig length (bases)	305,712,242
	Contigs (n)	48,368
	N50 contig length (bases)	17,566
	Max contig length (bases)	349,507

Table 5 General statistics of genes in *T. kingsejongensis*.

Genes (n)	12,772
Gene Length Sum (bp)	82,293,116
Exons per genes (n)	4.6
mRNA Length Sum (bp)	43,306,342
Average mRNA length (bp)	1,090
Number of tRNA	1,393
Number of rRNA	215

Table 6 Summary of orthologous gene clusters in the 11 representative species.

Species	Source of data	No. of coding genes	No. of gene families	No. of genes in gene families	No. of orphan genes	No. of unique gene families	Average No. of genes in gene families
Aedes aegypti	Ensembl genome 25	15,797	7,958	12,792	7,839	854	1.61
Caenorhabditis elegans	Ensembl gene 78	20,447	6,536	13,737	13,911	1,528	2.10
Ciona intestinalis	Ensembl gene 78	16,671	7,017	9,058	9,654	503	1.29
Daphnia pulex	Ensembl genome 25	30,590	6,710	8,362	7,208	368	1.25
Drosophila melanogaster	Ensembl gene 78	13,918	9,673	21,917	20,917	2,408	2.27
Homo sapiens	Ensembl gene 78	20,300	8,696	17,186	11,604	1,065	1.98
Ixodes scapularis	Ensembl genome 25	20,486	8,097	11,277	12,389	873	1.39
Mesobuthus martensii	http://lifecenter.sgst.cn/main/en/scorpion.jsp	32,016	8,389	19,961	23,627	2,276	2.38
Strigamia maritima	Ensembl genome 25	14,992	7,727	11,012	7,265	583	1.43
Tetranychus urticae	Ensembl genome 25	18,224	6,602	11,788	11,622	939	1.79
T. kingsejongensis	this study	12,772	6,205	8,813	6,567	649	1.42

Click here to download Figure Fig1.tif

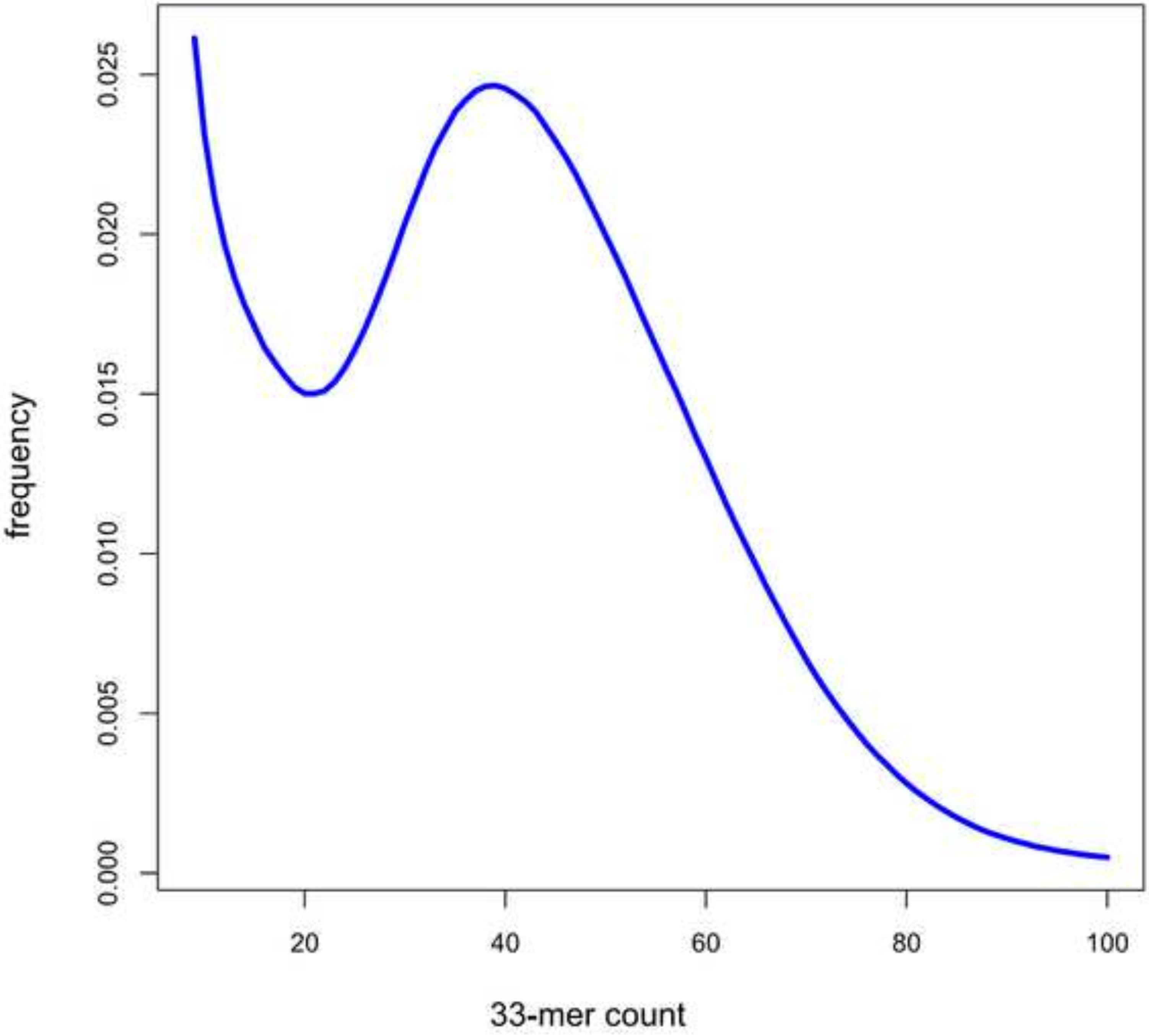


Figure 2

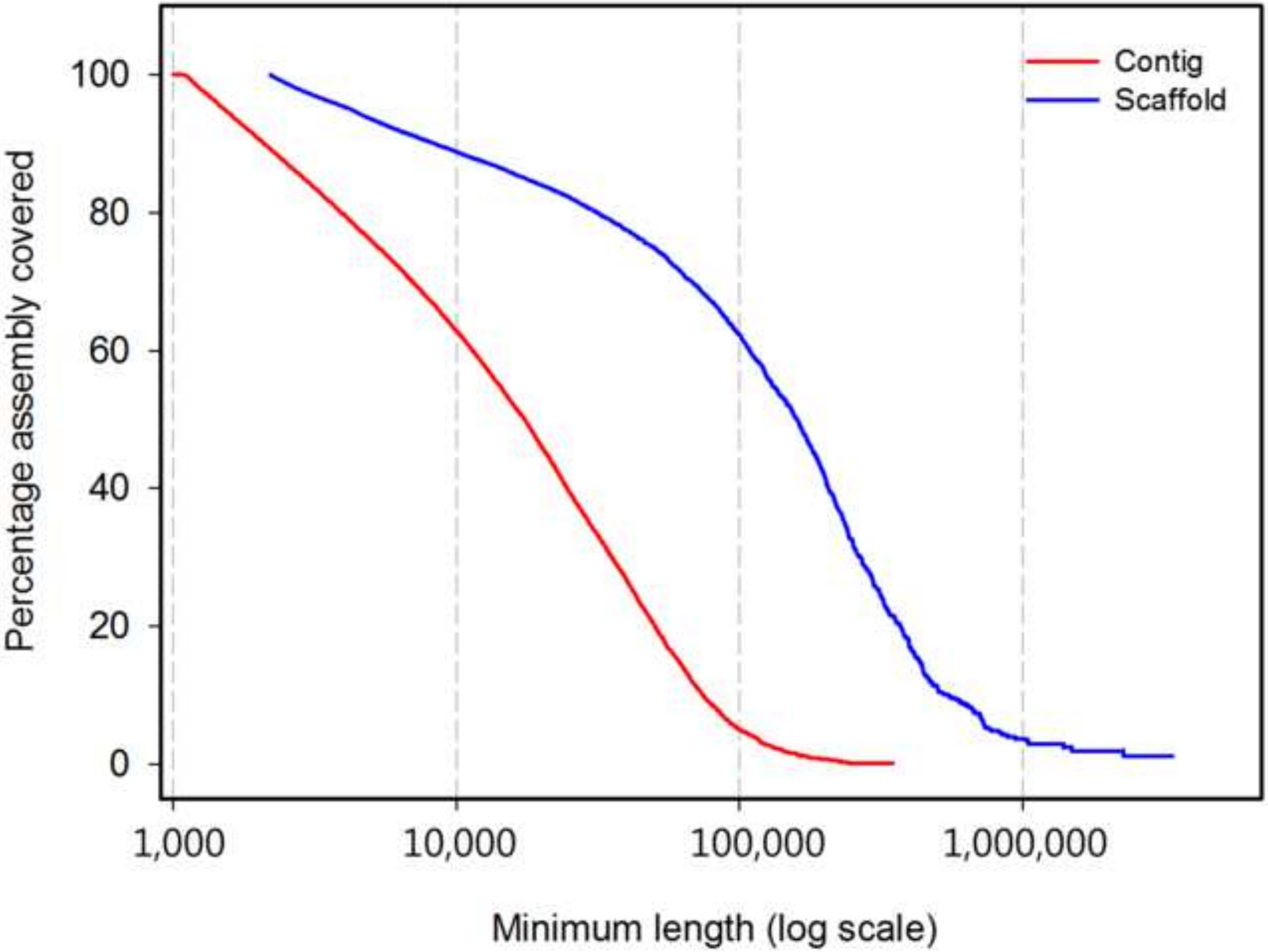

Contig
Scaffold
Percentage assembly covered
100
80
60
40
20
0
1,000
10,000
100,000
1,000,000
Minimum length (log scale)

Figure 3

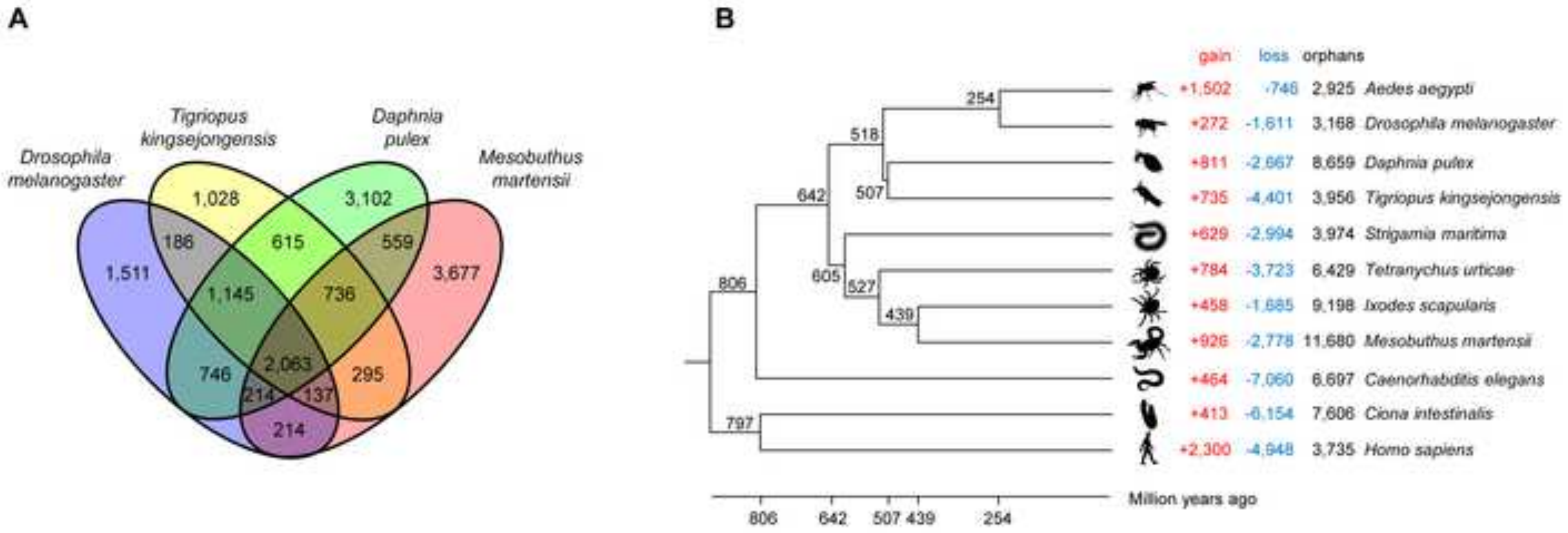
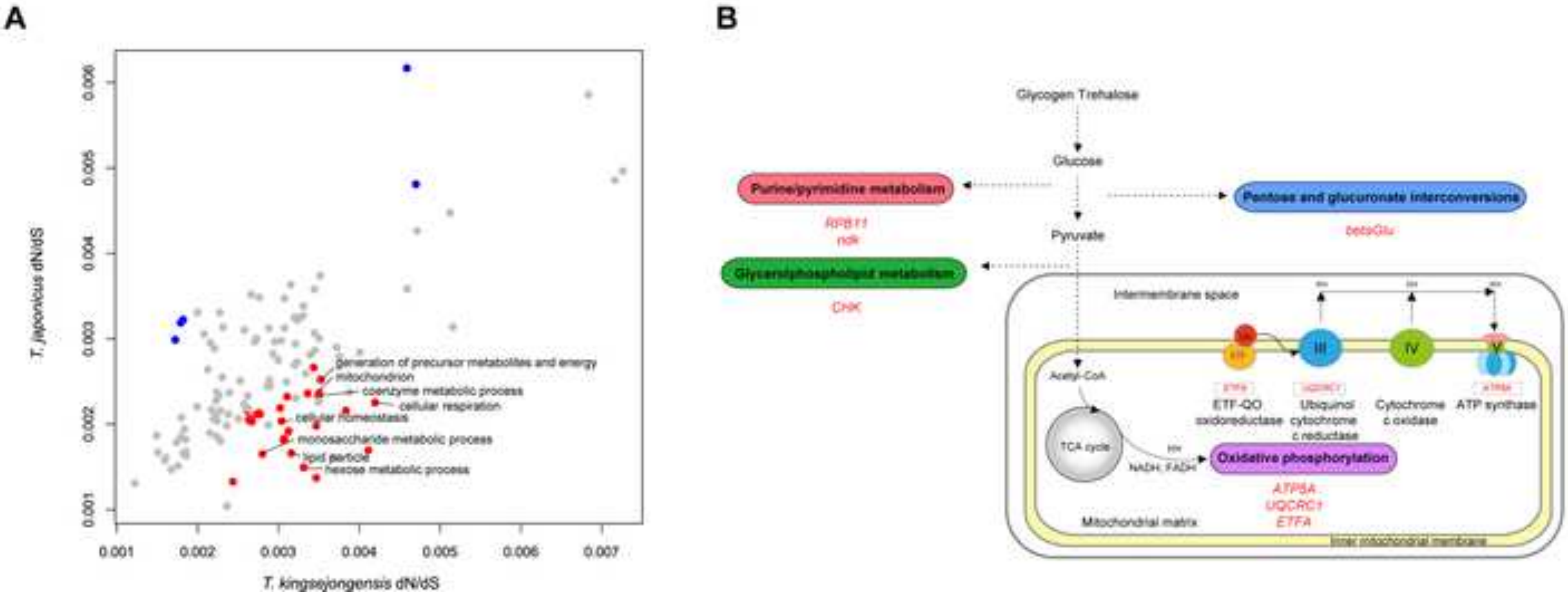

A
Drosophila melanogaster
Tigriopus kingsejongensis
Daphnia pulex
Mesobuthus martensii
1,511
1,028
3,102
3,677
186
615
559
1,145
736
746
2,063
295
214
137
214
B
gain
loss
orphans
+1,502 -746 2,925 Aedes aegypti
+272 -1,611 3,168 Drosophila melanogaster
+811 -2,667 8,659 Daphnia pulex
+735 -4,401 3,956 Tigriopus kingsejongensis
+629 -2,994 3,974 Strigamia maritima
+784 -3,723 6,429 Tetranychus urticae
+458 -1,685 9,198 Ixodes scapularis
+926 -2,778 11,680 Mesobuthus martensii
+464 -7,060 6,697 Caenorhabditis elegans
+413 -6,154 7,606 Ciona intestinalis
+2,300 -4,948 3,735 Homo sapiens
254
518
507
642
605
527
439
806
797
Million years ago
806
642
507 439
254

Figure 4


A
T. japonicus dN/dS
0.001
0.002
0.003
0.004
0.005
0.006
T. kingsejongensis dN/dS
0.001
0.002
0.003
0.004
0.005
0.006
0.007
generation of precursor metabolites and energy
mitochondrion
coenzyme metabolic process
cellular respiration
cellular homeostasis
monosaccharide metabolic process
lipid particle
hexose metabolic process
B
Glycogen Trehalose
Glucose
Pyruvate
Purine/pyrimidine metabolism
RPB11
ndk
Pentose and glucuronate interconversions
betaGlu
Glycerophospholipid metabolism
CHK
Intermembrane space
Acetyl-CoA
TCA cycle
NADH, FADH
Oxidative phosphorylation
ATP5A
UQCRC1
ETFA
III
IV
V
ETF-QO
oxidoreductase
Ubiquinol
cytochrome
c reductase
Cytochrome
c oxidase
ATP synthase
Mitochondrial matrix
Inner mitochondrial membrane

Additional file 1

Additional file 2