SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Genome

» Zoology

» Genomics

» Transcriptomics

Received: 26 April 2016

Accepted: 30 June 2016

Published: 16 August 2016

Data Descriptor: Genome sequencing of a single tardigrade *Hypsibius dujardini* individual

Kazuharu Arakawa[1,2], Yuki Yoshida[1,2] & Masaru Tomita[1,2]

Tardigrades are ubiquitous microscopic animals that play an important role in the study of metazoan phylogeny. Most terrestrial tardigrades can withstand extreme environments by entering an ametabolic desiccated state termed anhydrobiosis. Due to their small size and the non-axenic nature of laboratory cultures, molecular studies of tardigrades are prone to contamination. To minimize the possibility of microbial contaminations and to obtain high-quality genomic information, we have developed an ultra-low input library sequencing protocol to enable the genome sequencing of a single tardigrade *Hypsibius dujardini* individual. Here, we describe the details of our sequencing data and the ultra-low input library preparation methodologies.

Design Type(s)	protocol optimization objective
Measurement Type(s)	whole genome sequencing • transcription profiling assay
Technology Type(s)	DNA sequencing • RNA sequencing
Factor Type(s)	life cycle stage
Sample Characteristic(s)	Hypsibius dujardini

[1]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0052, Japan. [2]Systems Biology Program, Graduate School of Media and Governance, Keio University, Tsuruoka, Yamagata 997-0052, Japan. Correspondence and requests for materials should be addressed to K.A. (email: gaou@sfc.keio.ac.jp).

Background & Summary

Tardigrades are small (< 1 mm) aquatic ecdysozoans that occupy a distinct phylum, Tardigrada. This controversial phylogenetic positioning provides important insights into the evolution of arthropods and nematodes[1,2]. Limno-terrestrial tardigrades can withstand almost complete desiccation through a mechanism called anhydrobiosis[3,4], and tardigrades can survive extreme environments, including extreme temperature[5], pressure[6,7], radiation[8–10], and even exposure to the vacuum of space[11]. Although it is a weak anhydrobiote[12], *Hypsibius dujardini* is a model tardigrade due to its ease of culturing[13], transparent body color that is suited for developmental biology studies, and its feasibility for RNA interference-based genetic studies[14].

Recently, Boothby *et al.* reported the sequencing of *H. dujardini* and demonstrated that as many as 17.5% of the genes of this organism may have been acquired through horizontal gene transfer (HGT) and that these foreign origin genes may be the key to their extremotolerance[15]. However, the anhydrobiotic capabilities of *H. dujardini* are limited compared with those of other terrestrial tardigrades[12], and the non-axenic nature of tardigrade culture is prone to contamination, especially in light of the unusually large size of assembly (212.3 Mb) compared with previous estimates[13]. Hence, a calculation of the HGT rate requires comprehensive experimental evidence and careful data screening. Three research groups have provided evidence demonstrating that the HGT candidates presented by Boothby *et al.* were, in fact, mostly contaminant artifacts. Delmont and Eren employed metagenomics approaches and analyzed multiple DNA and RNA libraries, and Bemm *et al.* filtered Moleculo reads by k-mer frequencies, to identify contamination (including near-complete bacterial genomes) and effectively removed it from the assembly[16,17]. Koutsovoulos *et al.* presented a thorough comparison using an independently sequenced assembly of the same strain of *H. dujardini* and found no evidence of extensive HGT in this species[18].

We have also been studying the same strain, but with an ultra-low input sequencing approach. Tardigrades are small animals, mostly in the range of several hundred micrometers in length and comprising approximately one thousand cells, with only several hundred picograms of DNA per individual. Therefore, naive approaches require thousands of animals to extract sufficient DNA or RNA for the construction of libraries for high-throughput sequencing; the collection of that many animals from a non-axenic culture is inherently prone to contamination. Such contamination can be screened after assembly using bioinformatics and metagenomic approaches but would ideally be minimized in the library preparation steps. To this end, we employed an ultra-low input methodology to sequence a single individual of *H. dujardini* to minimize contamination and provided supporting evidence that there is no extensive HGT[19]. Here, we describe the details of our sequencing data and the ultra-low input library preparation methodologies. Although we also describe a series of single individual mRNA-Seq data in this work, due to the low reproducibility in some of the samples, we provide these data sets as supporting materials to possibly identify tardigrade-specific genes. Therefore, the rest of this descriptor is predominantly focused on the genome sequencing, reproducibility of which was confirmed by replications.

Methods

Tardigrade culture and sampling

The tardigrade *Hypsibius dujardini* Z151 was purchased from Sciento (Manchester, UK) in August 2012, and its culture was maintained in the laboratory using the culture method previously described for *Ramazzottius varieornatus*[20], with slight modifications. Briefly, tardigrades were fed *Chlorella vulgaris* (Chlorella Industry) on 2% Bacto Agar (Difco) plates prepared with Volvic water, and the plates were incubated at 18 °C under constant dark conditions. Culture plates were renewed every 7 ~ 8 days. Adults with body lengths >300 μm were selected. Cultures were manually inspected every day to obtain freshly laid eggs within 24 h. Eggs were transferred to new plates for egg samples. Similarly, newly laid eggs were observed every day to obtain newly hatched juvenile samples. Anhydrobiotic samples were prepared by placing adult samples in a chamber maintained at 85% relative humidity for 48 h. Successful anhydrobiosis was assumed when >90% of the samples prepared in the same chamber recovered after rehydration.

Genome sequencing

An adult individual was selected and placed in a sterile agar plate with 1% penicillin streptomycin (Invitrogen) prepared with autoclaved Volvic water for 48 h without food to clear the gut content and remove surface microbes. After 48 h, the animal was repeatedly washed with Milli-Q water on a sterile nylon mesh with 30 μm pores (Millipore), and any surface microbes were observed at × 500 ~ × 1000 magnification under a microscope (VHX-5000, Keyence). The cleaned animal was transferred to a low-binding PCR tube with minimal water carry-over (< 2 μl), to which a 200 μl genomic lysis buffer (Quick-gDNA MicroPrep kit, Zymo Research) with 0.5% beta-mercaptoethanol was added immediately. The animal was lysed using two freeze-thaw cycles of −80 °C and 37 °C incubation, and genomic DNA was extracted following manufacturer's protocol. Extracted DNA was then sheared to 550 bp target fragments with Covaris M220 using a 15 μl microTube, and the Illumina library was subsequently prepared using a Thruplex DNA-Seq kit (Rubicon Genomics) according to the manufacturer's instructions. Thruplex kit was previously reported to be efficient in low-input sequencing[21]. The purified library was quantified using a Qubit Fluorometer (Life Technologies), and the size distribution was

checked using TapeStation D1000 ScreenTape (Agilent Technologies). Library with sizes within the broad peak oberseved from TapeStation electropherogram between 400 and 1000 bp was selected by manually cutting them out of the agarose and purifying with a NucleoSpin Gel and PCR Clean-up kit (Clontech). The sample was then sequenced using one MiSeq v.3 600 cycles kit (Illumina) as 300 bp paired ends. Adapter sequences were removed using the MiSeq software (Illumina). In order to test the reproducibility of this single individual sequencing method, four more individuals were sequenced using the same procedures, except for the lysis step which was replaced by three freeze-thaw cycles of flash freezing with liquid nitrogen and 37 °C incubation, and the four libraries were multiplexed in a single MiSeq run. Genome extraction efficiency was assessed by extracting genomic DNA using the same method (with liquid nitrogen lysis) from 100 individuals with three replicates, and quantifying the amount using Qubit Fluorometer (Life Technologies).

mRNA sequencing

Active adults, adults in anhydrobiosis, eggs (1, 2, 3, 4, and 5 days after laying), and juveniles (1, 2, 3, 4, and 5 after hatching), were collected from the culture in three replicates and thoroughly washed with Milli-Q water on a sterile nylon mesh (Millipore). Each individual was placed in a low-binding PCR tube with minimal water carry-over ($< 2\,\mu l$). Pipette tips were used to crush animals by pressing them against the tube wall, and the individuals were then immediately lysed in TRIzol reagent (Life Technologies). RNA was extracted using a Direct-zol RNA kit (Zymo Research) following the manufacturer's instructions. RNA quality was checked using High Sensitivity RNA ScreenTape on TapeStation (Agilent Technologies), and the mRNA was amplified via the SMART-Seq approach[22,23] using the SMARTer Ultra Low Input RNA Kit for Sequencing v.3 (Clontech). Illumina libraries were prepared using a KAPA HyperPlus Kit (KAPA Biosystems). A purified library was quantified using a Qubit Fluorometer (Life Technologies), and the size distribution was checked using TapeStation D1000 ScreenTape (Agilent Technologies). Libraries with sizes above 200 bp were selected by manually cutting them out of the agarose and purifying with a NucleoSpin Gel and PCR Clean-up kit (Clontech). The samples were then sequenced using a NextSeq 500 High Output Mode 75 cycles kit (Illumina) as single ends, with 48 multiplexed samples per run. Only two replicates were sequenced for the juvenile day 5 sample. Adapter sequences were removed, and sequences were demultiplexed using the bcl2fastq v.2 software (Illumina).

Assembly and mapping of sequenced reads for technical validation

Genome sequence reads were assembled *de novo* using MaSuRCA 3.1.3 (ref. 24) with a mean insert length of 350 and a standard deviation of 150. Because the minimum range of size selection was 400 bp, insert size without Illumina adapters roughly corresponds to this length. Moreover, we have confirmed by mapping our reads to the assembly of Koutsovoulos *et al.* that the mean insert length was 363.5. Modifying this parameter to 550 bp produced identical assembly result, including the total number of scaffolds, average scaffold length, longest scaffold length, shortest scaffold length, scaffold N50, and scaffold N90. Only the total scaffold length was shortened by 551 bp. Using the assembly and genomic reads, the assembly of Boothby *et al.* (tg.genome.fsa, hereafter referred to as the UNC assembly), one pair of the Boothby reads (TG-500- SIPE_1_sequence.txt and TG-500- SIPE_2_sequence.txt), the assembly of Koutsovoulos *et al.* (nHd.2.3.abv500.fna, hereafter referred to as the Edinburgh assembly) and one pair of the Koutsovoulos reads (gHypdu_hiseq_pe_110427_3_1.fq and gHypdu_hiseq_pe_110427_3_2.fq) were mapped to assemblies in all combinations to assess the mapping rate. Mapping was performed using BWA v.0.7.11 MEM[25] with the default parameters, and the mapping percentage was calculated using the QualiMap build 11-11-13.

For the RNA-seq data, gene expression abundances were computed using the Kallisto software v.0.42.4 (ref. 26) with accompanying sleuth utility using the Augustus-based gene prediction data of the Edinburgh assembly (nHd.2.3.1.aug.transcripts.fasta), with the following parameters: -b 100 --bias --single -l 400 -s 50. Overall gene expression profiles were scaled with scale() function and visualized as a heatmap for genes with TPM>1 using hierarchical clustering based on the Spearman correlation with hclust() and plotting using the ggplot2 utility in R[27].

Data Records

36 raw Illumina reads were deposited into DDBJ under the BioProject ID PRJDB4575 for one set of whole genome shotgun reads and 35 sets of RNA-Seq reads comprised of three replicates each (two for the day 5 juveniles) for eggs collected on days 1–5 after laying and juveniles collected on days 1–5 after hatching, active adults, and tun (anhydrobiotic) adults (Data Citation 1; see the associated Metadata Record in Table 1 (available online only) for details). 4 additional whole genome shotgun reads were deposited into NCBI Sequence Read Archive (SRA) under the same BioProject ID. Assembled contigs from the genomic reads were deposited in Data Dryad (Data Citation 2), where the total scaffold length was 132,494,968 nt, the number of scaffolds was 54,960, the longest scaffold was 86,209 nt and the scaffold N50 was 4,851 nt.

Technical Validation

A very rough estimate of total genomic DNA from a single *H. dujardini* individual (1,000 cells harboring diploid 100 Mbp genome) would be around 200 pg. Our extraction method yielded 57.5 ± 3.7 pg per

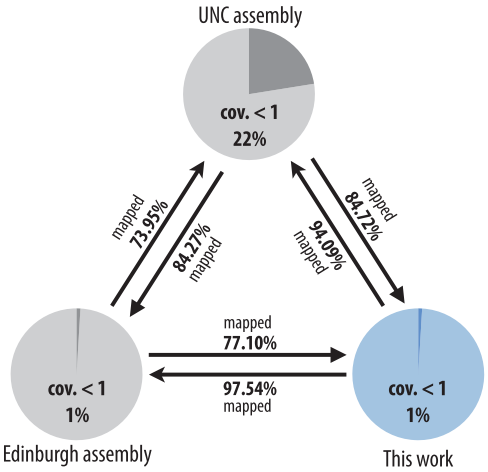

Figure 1. Comparison with existing assemblies and raw reads of *H. dujardini*. The ultra-low input sequencing data described in this work showed greater genomic coverage, as represented by a higher mapping rate of UNC and Edinburgh reads to our assembly, than between that of UNC and Edinburgh. UNC mapped to our assembly at 84.72%, whereas 84.27% mapped to the Edinburgh assembly. Similarly, 77.10% of the Edinburgh reads mapped to our assembly, but only 73.95% mapped to UNC assembly. While the UNC and Edinburgh reads presumably contained 10 ~ 25% contamination, as suggested by the percentage of unmapped reads, 94.09 or 97.54% of the study reads mapped to the UNC or Edinburgh assemblies, respectively. These data indicated minimal contamination in our reads.

individual, which would presumably represent genomic DNA from several hundred cells. This amount is marginally above the supported input range of Thruplex DNA-Seq kit (Rubicon Genomics), and therefore we have successfully constructed the sequencing libraries, and performed sequencing from a single individual.

To validate the comprehensiveness of the genomic coverage and to detect any contamination of our single animal protocol, reads and assembled sequences of the three independent projects (UNC, Edinburgh, and the current study) were compared by mapping all possible combinations (Fig. 1). Overall, 84.27% of UNC reads mapped to Edinburgh assembly, which represented reads from *H. dujardini*. The remaining sequences (approximately 10 ~ 15%) were likely a contamination of non-tardigrade sequences. A slightly higher percentage (84.72%) of reads mapped to the assembly described in this study, which implied a comparable or slightly greater genome coverage. In addition, 73.95% of the Edinburgh reads mapped to the UNC assembly, whereas a higher percentage (77.10%) mapped to our assembly. Thus, the reads described in this work obtained from a single animal using the ultra-low input protocol yielded genomic reads that were more comprehensive than the traditional methods, which start with thousands of tardigrades. The percentage of mapped reads was lower in the Edinburgh and UNC reads, with sequences likely to be derived from tardigrades in the range of 70 ~ 85%, with a contamination in the range of 10 ~ 25%. The Edinburgh reads seemed to contain a higher percentage of contamination than the UNC reads. These artifacts were successfully removed during the assembly steps, leaving only 149 contigs spanning 1.3 Mbp. This value represents approximately 1% of the total assembly length, with a coverage less than 1 when using the reads described in this work. By contrast, the UNC assembly contained 5,665 contigs, spanning 56.8 Mbp with a coverage less than 1, which amounted to 22% of the total assembly length. This result indicated that a large proportion of contaminations were contained in the final assembly. By contrast, our reads predominantly mapped to the existing assemblies (94.09% to the UNC assembly, and 97.54% to the Edinburgh assembly). The high rate of mapped reads indicated that contamination was minimal in our data. This result was also supported by the minimal percentage of contigs with coverage less than 1, which was 1% of the total assembly when either the UNC or Edinburgh reads were mapped. Note that our assembly is the unscreened raw output of MaSuRCA assembler and that this percentage would be much lower after screening and curation.

Reproducibility of this single individual sequencing approach was tested with four additional replicates but with fewer number of reads, as shown in Table 2. All replicates showed higher percentage of mapped reads compared to those of Edinburgh and UNC reads against all three scaffolds, and therefore the single individual sequencing approach is a reproducible and comprehensive method with minimum contaminations.

The mapping percentage and abundance estimates of the RNA-Seq data were visualized for active and tun (in anhydrobiosis) adult samples (Fig. 2) and for juvenile and egg samples (Fig. 3). Adult samples

Data ID	Number of reads	Mapping percentage to Edinburgh assembly	Mapping percentage to UNC assembly	Mapping percentage to assembly of this work
SRR3706607	10,642,574	93.52	89.87	93.69
SRR3706608	11,497,318	95.58	91.48	95.73
SRR3706609	11,855,730	92.91	89.64	93.14
SRR3706610	11,250,228	95.68	91.60	95.79

Table 2. Mapping percentages of four replicates of genome sequence reads.

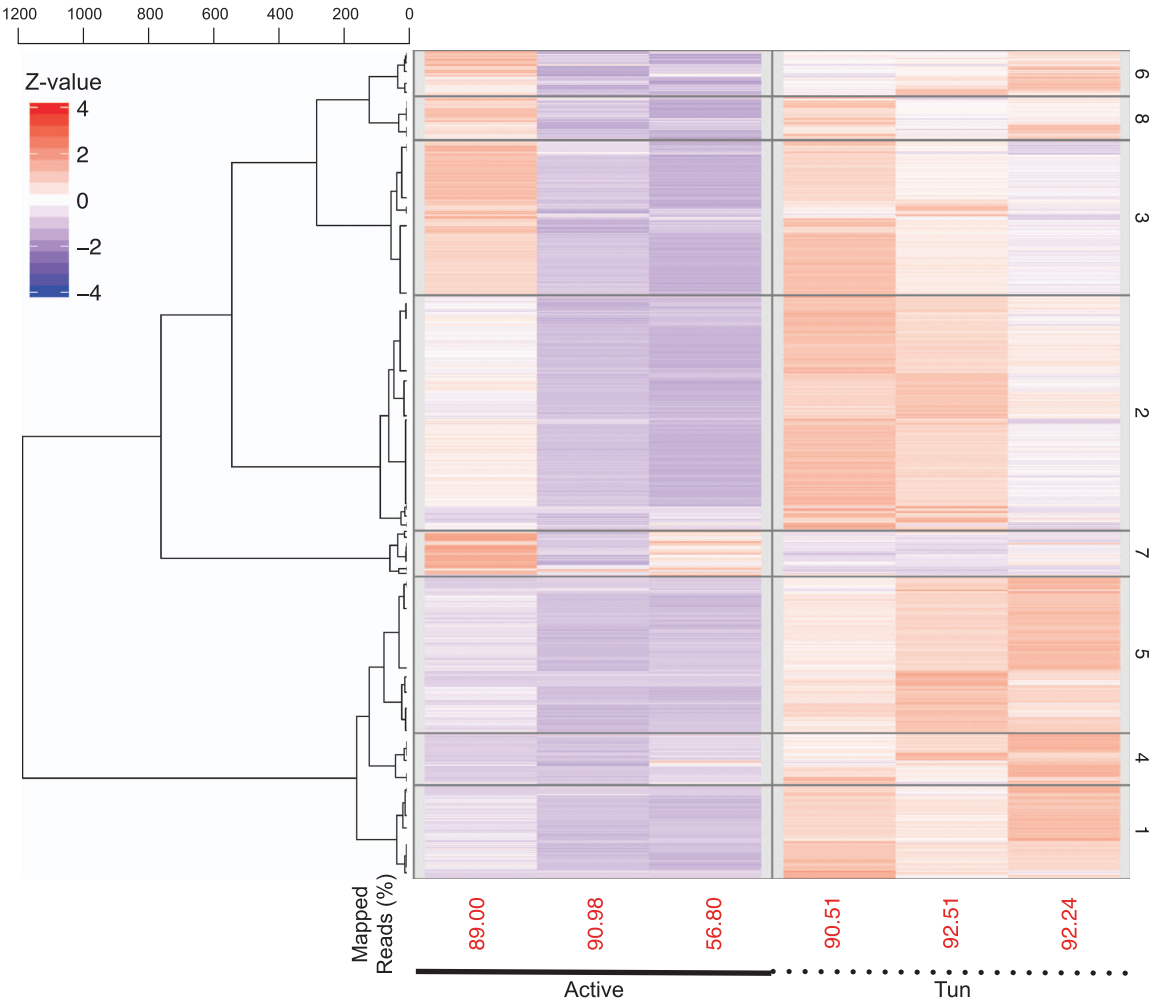

Figure 2. Transcriptome abundances of active and tun (in anhydrobiosis) adults in three biological replicates. Percentage of mapped reads is shown beneath the heatmap.

exhibited high mapping percentages and expression profiles, which mainly correlated within the biological replicates. However, both the reproducibility and mapping percentages were consistently low, with the exception of juvenile samples on the first and third days after hatching. This result is presumably due to the inefficiency of RNA extraction in these developmental stages.

Usage Notes

The genome sequencing reads described in this work were intended to confirm the published UNC and Edinburgh assemblies and to clarify the contigs derived from *H. dujardini,* as described previously[19]. Moreover, the assembly provided in this work was intended only for technical validation, as described above. The assembly was neither sufficiently scaffolded nor curated. Therefore, the Edinburgh assembly would be better suited for genomic studies. The RNA-Seq data described in this work were obtained to

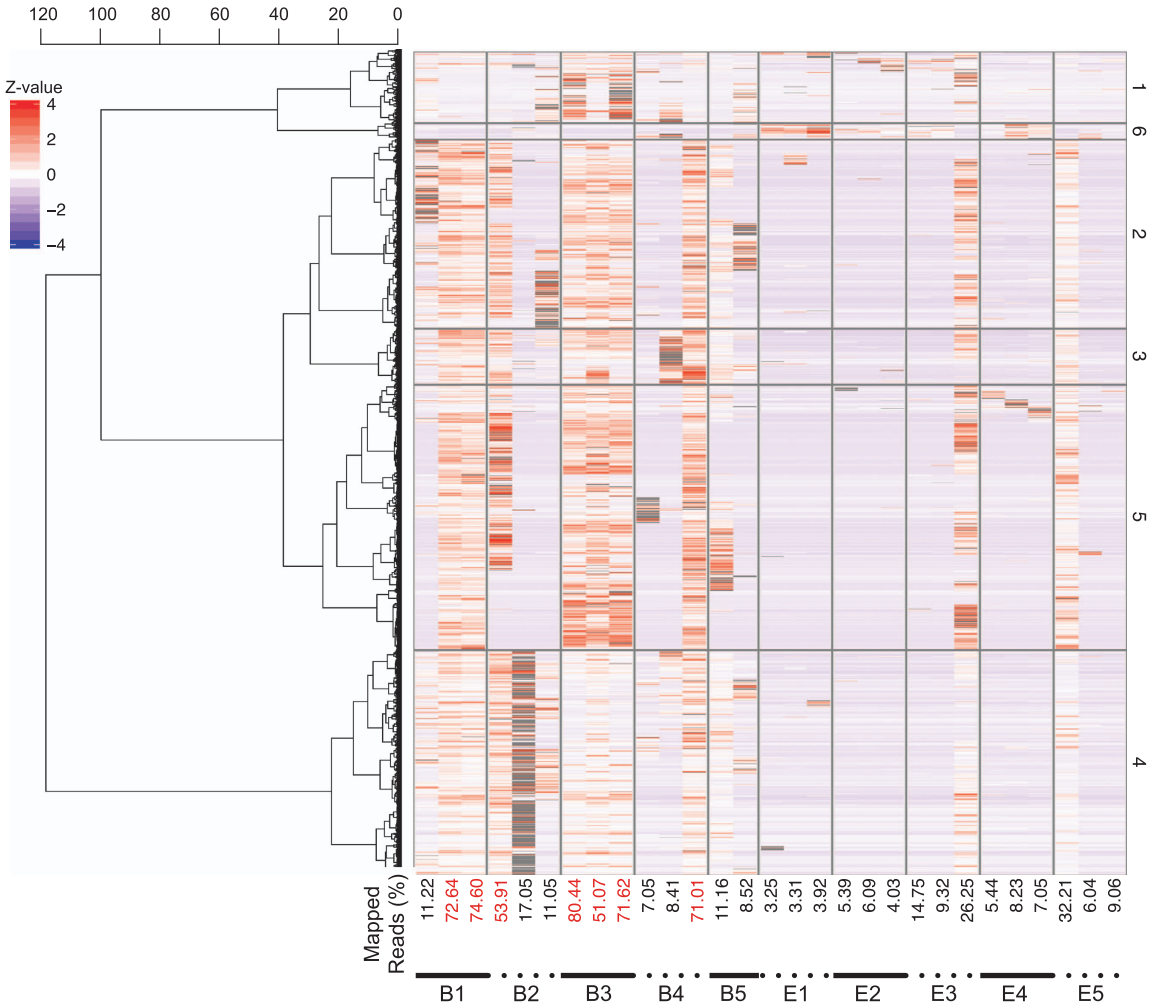

Figure 3. Transcriptome abundances of juveniles (B1 ~ B5) and eggs (E1E5) in the first 1 ~ 5 days after hatching or laying. Sequences were obtained in three biological replicates except for B5, which only has two replicates. The percentage of mapped reads is shown beneath the heatmap, and samples with a mapping percentage greater than 50% are colored in red.

further validate the confirmed data. As described in the technical validation section, the reproducibility was low for samples in the early developmental stages, presumably due to inefficient RNA extraction from eggs and juveniles. RNA-Seq data with low mapping percentage should not be used in comparative analyses.

References

1. Campbell, L. I. *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 USA* **108,** 15920–15924 (2011).
2. Mobjerg, N. *et al.* Survival in extreme environments—on the current knowledge of adaptations in tardigrades. *Acta Physiol (Oxf)* **202,** 409–420 (2011).
3. Crowe, J. H., Hoekstra, F. A. & Crowe, L. M. Anhydrobiosis. *Annu Rev Physiol* **54,** 579–599 (1992).
4. Keilin, D. The Leeuwenhoek Lecture. The problem of anabiosis or latent life: history and current concept. *Proceedings of the Royal Society of London, Series B* **150,** 149–191 (1959).
5. Becquerel, P. La suspension de la vie au dessous de 1/20 K absolu par demagnetization adiabatique de l'alun de fer dans le vide les plus eléve. *Comptes-rendus Hebdomadaires des Seances de l'Académie des Sciences de Paris* **231,** 261–263 (1950).
6. Horikawa, D. D. *et al.* Tolerance of anhydrobiotic eggs of the Tardigrade Ramazzottius varieornatus to extreme environments. *Astrobiology* **12,** 283–289 (2012).
7. Ono, F. *et al.* Effect of high hydrostatic pressure on to life of the tiny animal tardigrade. *Journal of Physics and Chemistry of Solids* **69,** 2297–2300 (2008).
8. Horikawa, D. D. *et al.* Analysis of DNA repair and protection in the Tardigrade Ramazzottius varieornatus and *Hypsibius dujardini* after exposure to UVC radiation. *PLoS ONE* **8,** e64793 (2013).
9. Horikawa, D. D. *et al.* Radiation tolerance in the tardigrade Milnesium tardigradum. *Int J Radiat Biol* **82,** 843–848 (2006).
10. May, R.-M., Maria, M. & Guimard, J. Action differentielle des rayons X et ultraviolets sur le tardigrade *Macrobiotus areolatus* a l'etat actif et desseche. *Bulletin biologique de la France et de la Belgique* **48,** 349–367 (1964).

11. Jonsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M. & Rettberg, P. Tardigrades survive exposure to space in low Earth orbit. *Curr Biol* **18,** R729–R731 (2008).
12. Wright, J. C. Desiccation tolerance and water-retentive mechanisms in tardigrades. *Journal of Experimental Biology* **142,** 267–292 (1989).
13. Gabriel, W. N. *et al.* The tardigrade *Hypsibius dujardini*, a new model for studying the evolution of development. *Dev Biol* **312,** 545–559 (2007).
14. Tenlen, J. R., McCaskill, S. & Goldstein, B. RNA interference can be used to disrupt gene function in tardigrades. *Dev Genes Evol* **223,** 171–181 (2013).
15. Boothby, T. C. *et al.* Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci USA* **112,** 15976–15981 (2015).
16. Bemm, F., Weiss, C. L., Schultz, J. & Forster, F. Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? *Proc Natl Acad Sci USA* **113,** E3054–E3056 (2016).
17. Delmont, T. O. & Eren, A. M. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **4,** e1839 (2016).
18. Koutsovoulos, G. *et al.* No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc Natl Acad Sci USA* **113,** 5053–5058 (2016).
19. Arakawa, K. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci USA* **113,** E3057 (2016).
20. Horikawa, D. D. *et al.* Establishment of a rearing system of the extremotolerant tardigrade Ramazzottius varieornatus: a new model animal for astrobiology. *Astrobiology* **8,** 549–556 (2008).
21. Rykalina, V. N. *et al.* Exome sequencing from nanogram amounts of starting DNA: comparing three approaches. *PLoS One* **9,** e101154 (2014).
22. Goetz, J. J. & Trimarchi, J. M. Transcriptome sequencing of single cells with Smart-Seq. *Nat Biotechnol* **30,** 763–765 (2012).
23. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9,** 171–181 (2014).
24. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29,** 2669–2677 (2013).
25. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
26. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34,** 525–527 (2016).
27. Project, R. *R: the R Project for Statistical Computing* https://www.r-project.org (2016).

Data Citations

1. Arakawa, K., Yoshida, Y. & Tomita, M. *DDBJ Sequence Read Archive* DRA004455 (2016).
2. Arakawa, K., Yoshida, Y. & Tomita, M. *Dryad Digital Repository* http://dx.doi.org/10.5061/dryad.t4k7h (2016).

Acknowledgements

The authors thank Nozomi Abe for technical support in tardigrade culturing and genomic sequencing, Yuki Takai for technical support in RNA sequencing, and Daiki Horikawa for the suggestion to use liquid nitrogen for lysis. *Chlorella vulgaris* used to feed the tardigrades was provided courtesy of Chlorella Industry Co. LTD. This work was supported by KAKENHI Grant-in-Aid for Young Scientists (No.22681029) from the Japan Society for the Promotion of Science (JSPS), by a Grant for Basic Science Research Projects from The Sumitomo Foundation (No.140340), and partly by research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan.

Author Contributions

K.A. designed and performed the experiments and drafted the manuscript. Y.Y. analyzed the RNA-seq data. M.T. managed the computer resources. All authors contributed to editing and revising the manuscript.

Additional Information

Table 1 is only available in the online version of this paper.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Arakawa, K. *et al.* Genome sequencing of a single tardigrade *Hypsibius dujardini* individual. *Sci. Data* 3:160063 doi: 10.1038/sdata.2016.63 (2016).