Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge

Oleg Gusev1,2,3,*, Yoshitaka Suetsugu1,*, Richard Cornette1,*, Takeshi Kawashima4, Maria D. Logacheva5,6,7, Alexey S. Kondrashov5,8, Aleksey A. Penin5,7,9, Rie Hatanaka1, Shingo Kikuta1, Sachiko Shimura1, Hiroyuki Kanamori1, Yuichi Katayose1, Takashi Matsumoto1, Elena Shagimardanova2, Dmitry Alexeev10, Vadim Govoron10, Jennifer Wisecaver11, Alexander Mikheyev4, Ryō Koyanagi4, Manabu Fujie4, Tomoaki Nishiyama12, Shuji Shigenobu13,14, Tomoko F. Shibata13, Veronika Golygina15, Mitsuyasu Hasebe13,14, Takashi Okuda1, Nori Satoh4 & Takahiro Kikawada1

Anhydrobiosis represents an extreme example of tolerance adaptation to water loss, where an organism can survive in an ametabolic state until water returns. Here we report the first comparative analysis examining the genomic background of extreme desiccation tolerance, which is exclusively found in larvae of the only anhydrobiotic insect, Polypedilum vanderplanki. We compare the genomes of P. vanderplanki and a congeneric desiccation-sensitive midge P. nubifer. We determine that the genome of the anhydrobiotic species specifically contains clusters of multi-copy genes with products that act as molecular shields. In addition, the genome possesses several groups of genes with high similarity to known protective proteins. However, these genes are located in distinct paralogous clusters in the genome apart from the classical orthologues of the corresponding genes shared by both chironomids and other insects. The transcripts of these clustered paralogues contribute to a large majority of the mRNA pool in the desiccating larvae and most likely define successful anhydrobiosis. Comparison of expression patterns of orthologues between two chironomid species provides evidence for the existence of desiccation-specific gene expression systems in P. vanderplanki.
Several organisms have evolved the ability to withstand extreme abiotic stresses, which are lethal for most other forms of life. Anhydrobiosis is a unique ametabolic state that enables a living organism to maintain viability even after losing more than 97% of its body water. Anhydrobiosis is generally associated with extraordinary cross-tolerance to a large variety of extreme conditions, such as temperatures ranging from −270 to +102 °C, vacuum and hydrostatic pressures up to 1.2 GPa and extremely high doses of radiation (up to 7,000 Gy)\(^{1,2}\). Anhydrobiotes also exhibit surprising longevity, and certain species have survived tens or even thousands of years in the dry form before recovery on rehydration\(^3\).

Among metazoans, the ability to survive severe desiccation by entering an anhydrobiotic state is limited to several groups that include mostly microscopic organisms\(^1\). The largest and most complex anhydrobiotic animal is the larva of a non-biting midge, which is the sleeping chironomid \textit{Polypedilum vanderplanki}\(^7\) (Fig. 1a). Chironomid midges are known for their capacity to adapt to a wide variety of extreme environments and constitute a unique group among insects\(^6\). However, anhydrobiosis is a complex adaptive trait and \textit{de novo} acquisition of such an extreme desiccation tolerance is most likely a unique evolutionary event. \textit{P. vanderplanki} is the only anhydrobiotic species known to date among both chironomid midges and the entire insect lineage. In contrast to other anhydrobiotes (such as rotifers, tardigrades, nematodes or even plants) that are found in phyla showing widespread desiccation tolerance in a large array of species, \textit{P. vanderplanki} is an isolated case among anhydrobiotes\(^1\). This finding suggests that the sleeping chironomid is a promising model for comparative genomics and should allow a precise dissection of the genetic background underlying the development of tolerance to complete desiccation. Over the last decade, investigations of the sleeping chironomid resulted in the identification of several groups of biomolecules, including late embryogenesis abundant (LEA) proteins, trehalose, antioxidants and heat–shock proteins contributing to desiccation tolerance\(^7\)\(^{\text{-}}\)\(^{13}\). In addition, \textit{in vitro} and \textit{in vivo} experiments have shown that these components are necessary but not sufficient to acquire complete desiccation tolerance. Therefore, whole-genome screening for anhydrobiosis-related features became an obligatory step in further understanding the molecular background of successful desiccation tolerance\(^\text{1,4,15}\).

As mentioned above, the adaptation of \textit{P. vanderplanki} is an isolated case of anhydrobiosis among all insects (Fig. 1a). A closely related species from the same genus, \textit{P. nubifer}, is sensitive to desiccation (Fig. 1b). Thus, the combination of \textit{P. vanderplanki} and \textit{P. nubifer} represents a uniquely informative model of comparative genomics for deciphering the entire genetic background of anhydrobiosis.

Our principal aim was to identify genome features specific to \textit{P. vanderplanki} that are lacking in \textit{P. nubifer} and other insects with sequenced genomes (including the fruit fly \textit{Drosophila} and mosquitoes of the genera \textit{Anopheles} and \textit{Aedes}; Fig. 1c). This strategy allowed us to successfully identify several key genomic features accounting for extreme desiccation tolerance in \textit{P. vanderplanki}. Among these features, the most obvious traits characterizing anhydrobiosis are the presence of specific genomic regions containing clusters of multi-copy protective genes involved in desiccation tolerance, the active utilization of protective proteins that most likely originate from horizontal gene transfer (HGT) and new desiccation-driven expression patterns for single genes that already exist in the \textit{P. vanderplanki} genome.

## Results

### Assembly and characteristics of the chironomid genomes

The ~600-fold coverage sequencing yielded a 104 Mbp for \textit{P. vanderplanki} (scaffold \(N_50 = 229\) kbp) and 107 Mbp for \textit{P. nubifer} (scaffold \(N_50 = 26\) kbp) genome assemblies (Supplementary Note 1) approximating to the estimated genome sizes (96 and 95 Mb, respectively). A spread of metaphase giant chromosomes revealed a chromosome number of \(2n = 8\) for both species (Supplementary Fig. 1). The \textit{P. vanderplanki} genome is characterized by higher AT content and a low number of known transposable elements. The \textit{P. vanderplanki} and \textit{P. nubifer} genomes are predicted to contain 17,137 and 16,553 protein-coding loci, respectively (Supplementary Note 1). The \textit{P. vanderplanki} and \textit{P. nubifer} genome contigs contain 97.18 and 95.56% of the complete core eukaryotic protein-coding sequences, which

---

**Figure 1** | Desiccation tolerance and phylogeny of two chironomid species. (a) Adult male of the sleeping chironomid, \textit{P. vanderplanki} (left), and anhydrobiotic cycle of the larvae (right). During the dry season, larvae desiccate slowly to reach an ametabolic, quiescent state, termed anhydrobiosis. On rehydration, dried larvae rapidly recover normal activity. (b) Adult male of the congeneric chironomid, \textit{P. nubifer} (left). \textit{P. nubifer} larvae can survive mild desiccation for 24 h like other chironomids, but they cannot enter anhydrobiosis and are killed by severe dehydration (right). Scale bar, 2 mm. (c) Phylogenetic tree inferred from the amino-acid sequence of cytochrome oxidase I (COI) showing the relationship between \textit{P. vanderplanki}, \textit{P. nubifer} and other Diptera. The scale shows the evolutionary distance between species in million years (MYA).
Table 1 | The statistics of the assembled genomes of *P. vanderplanki* and *P. nubifer*.

| Species          | Number of chromosomes | Size of genome assembly (Mbp) | Scaffold N50 (Mbp) | GC content (%) | Repetitive content (%) | Known transposable element (%) | Predicted number of genes | Number of genes in orthologous clusters with other species | Mean number of genes with >2 orthologues | Mean exon length (bp) | Mean intron length (bp) |
|------------------|-----------------------|-------------------------------|--------------------|----------------|------------------------|-------------------------------|--------------------------|-----------------------------------------------|------------------------------------------|----------------------------|------------------------|
| *P. vanderplanki*| 4                     | 104                           | 0.23               | 28             | 5.01                   | 0.26                          | 17,137                   | 13,529                                        | 5                                        | 324                        | 533                    |
| *P. nubifer*     | 4                     | 107                           | 0.03               | 39             | 3.30                   | 1.26                          | 16,553                    | 13,529                                        | 4                                        | 328                        | 452                    |

Figure 2 | Putative mechanism for the evolution of ARId in the *P. vanderplanki* genome. ARIds are genomic regions containing clusters of duplicated genes that are transcriptionally active during anhydrobiosis. (a) A gene of foreign origin (for example, LEA protein) is incorporated into *P. vanderplanki* genome by HGT and undergoes extensive duplications and shuffling. (b) A pre-existing *P. vanderplanki* gene originally not involved in anhydrobiosis and originating from another region of *P. vanderplanki* genome was inserted to a new locus by intragenomic duplication (IGD) and undergoes extensive duplications and shuffling to acquire or improve a specific function for desiccation tolerance. All the genes in the ARIds from both a,b become highly upregulated during anhydrobiosis (red arrows).

proteins and membranes from desiccation stress\(^\text{17}\). They have been reported in both plants and invertebrates characterized by tolerance to water depletion\(^\text{18}\). Four genes encoding LEA proteins have been reported in *P. vanderplanki*\(^\text{19,19}\), but not in other insects (including insects with sequenced genomes). Analysis of the *P. vanderplanki* genome revealed 27 LEA protein genes (Supplementary Data 3), but none in the *P. nubifer* genome (Supplementary Data 3). These data suggest that the presence and activity of LEA proteins is correlated with anhydrobiotic capability. While plants have multigene families encoding LEA proteins (for example, 51 genes in *Arabidopsis*\(^\text{20}\) of which the respective genes are distributed throughout the genome, only a few LEA proteins have been characterized in any particular invertebrate species\(^\text{21}\). At the same time, increasing numbers of transcriptome studies suggest that multiple LEA-like proteins are a feature of at least some anhydrobiotic animals\(^\text{22}\).

*P. vanderplanki* genes encoding LEA proteins (*PvLea* genes) are compactly arranged in two ARId clusters in the genome and there are some other genes interspersed\(^\text{16}\). None of the interspersed genes have orthologues in other insects or in *P. nubifer*. All predicted LEA-like genes in *P. vanderplanki* were expressed under non-desiccating conditions and most were strongly upregulated by desiccation (Supplementary Data 4)\(^\text{16}\). In most cases, desiccated larvae contained the highest level of each *PvLea* mRNA and the mRNA expression returned to control levels in larvae rehydrated for 24 h (Supplementary Data 4).

To obtain insight into the possible origin of LEA protein genes in the *P. vanderplanki* genome, we conducted phylogenetic analyses using the BlastX protocol to identify homologies of *PvLea* genes with other insect genes. However, these analyses did not identify any homologues (Supplementary Data 3). We
that result in the formation of splice variants specifically upregulated in the cycle of anhydrobiosis (see Supplementary Note 2 and Supplementary Fig. 3).

TRXs are small redox proteins present in all organisms\(^ {27}\). These proteins are involved in redox signalling and act as antioxidants by facilitating the reduction of other proteins via cysteine thiol–disulphide exchange\(^ {28}\). The number of TRXs in animal genomes ranges from one to five, and most isoforms are critical for normal organism function\(^ {29}\). The two chironomid genomes both contain three TRXs that are well conserved in number and structure among insect genomes (insect TRX-1 to TRX-3 in Fig. 4a). However, the \( P. \ vanderplanki \) genome contains 21 additional genes encoding TRXs arranged in two ARId sets unlinked to the classical TRX set of genes (\( P. \ vanderplanki \)-specific TRXs in Fig. 4a; Supplementary Data 6 and 8). These newly identified TRXs share key features of cytosolic TRX, including small size and a single TRX domain. In addition, all of the genes are strongly upregulated by desiccation. In contrast, the classical TRX genes in \( P. \ vanderplanki \) (\( PvTrx1–3 \)) respond only moderately to water loss (Supplementary Data 7).

**Unexpected diversity of protein-repair methyltransferases in ARId.** PIMT is an enzyme that recognizes and catalyses the repair of damaged \(-1\)-isoaspartyl and \(-\)aspartatyl groups in proteins\(^ {30}\). PIMT partially restores aspartic residues in proteins that have been non-enzymatically damaged due to age and extends the life of its substrates. This highly conserved enzyme is present in nearly all eukaryotes, Archaea and gram-negative eubacteria mostly as a single isoform (or as a few isoforms in certain plants and some bacteria)\(^ {31}\). Insects have a single copy of the PIMT-coding gene like plants, nematodes and mammals. PIMT activity in animals was found to be tightly linked to stress resistance and lifespan\(^ {30,31}\). The structure and number of PIMT-coding genes in the two chironomid species varied dramatically. Both species have the orthologues of PIMT shared by dipteran insects (PIMT-1 in Fig. 4b; Supplementary Data 9). \( P. \ nubifer \) has only one PIMT gene (\( PvPimt1 \)). However, the \( P. \ vanderplanki \) genome contains 13 additional genes paralogous to \( Pimt1 \) (\( PvPimt2–14 \) in Fig. 4b). These genes presumably code for functional PIMT proteins. The genes are arranged in a single cluster\(^ {16}\). Remarkably, the \( PvPimt1 \) location in \( P. \ vanderplanki \) is not within the single ARId constituting other \( Pimt \)-like genes. The expression of PIMT1-coding gene in both chironomids (\( PvPimt1 \) and \( PnPimt1 \)) did not change in response to desiccation, but the clustered \( PvPimt2–14 \) genes showed upregulation on entering anhydrobiosis (Supplementary Data 9). The abundance of \( PvPimt2–14 \) mRNAs was maximal in anhydrobiotic larvae and resembled plant seed traits where the accumulation of PIMT provides additional protection for proteins during long periods of dry storage by exerting their repair function on rehydration\(^ {12}\).

**Antioxidants and ARId-specific TRXs.** Antioxidants play an important role in the adaptation to extreme dehydration in anhydrobiotes\(^ {25}\). The expression of several key antioxidant genes is linked to anhydrobiosis in \( P. \ vanderplanki \). Desiccated larvae accumulate corresponding mRNA and proteins so that during rehydration they can efficiently scavenge reactive oxygen species\(^ {13}\). We identified 52 genes in \( P. \ vanderplanki \) and 29 genes in \( P. \ nubifer \) encoding core components of the insect enzymatic antioxidant systems\(^ {26} \) (Supplementary Data 6). The number of such genes in \( P. \ nubifer \) is similar to that of other insects\(^ {10} \). However, in \( P. \ vanderplanki \), several groups of antioxidant genes have expanded (Supplementary Data 6). In addition to the two cytoplasmic and single mitochondrial superoxide dismutases (SOD) that are well conserved among other insects including \( P. \ nubifer \), the \( P. \ vanderplanki \) genome possesses two additional genes encoding a Zn-Cu-SOD (Supplementary Data 6). On the basis of sequence similarity and genomic location, these genes are not paralogues of classical insect SOD. These SOD genes are highly expressed in response to desiccation and are most likely involved in anhydrobiosis (Supplementary Data 7). Another remarkable finding is the appearance of additional exons in glutathione peroxidase genes
proteins corresponding to *PvPimt*2–14 contain the conserved PIMT functional domain. In addition, the length and structure of the amino- and carboxy-terminal regions of the predicted proteins show marked variation. These findings suggest different substrate preferences or other specific properties of the various *PvPIMTs*. This multi-member family in *P. vanderplanki* is the first observation of large-scale expansion of *Pimt* genes in general and has not been reported in a single insect species.

**Haemoglobins and anhydrobiosis.** Chironomids are the only group of insects with haemolymph haemoglobins (Hbs) that act as the main respiratory proteins. Thus, chironomid Hbs have a respiratory function analogous to vertebrate Hb. Hbs enhance the O₂ capacity of the haemolymph and enable O₂ transport in the respiratory function. Chironomids, as the main respiratory proteins. Thus, chironomid Hbs have a remarkable difference between the two *Polypedilum* species is that the increased number of *Hb* genes in *P. vanderplanki* results from the insertion of a new cluster of paralogous intron-less *Hb* genes located in an ARId consisting of six members (see ARId sub-genome browser16). This gene set (*PvHb*11, 12, 17, 23, 24, 25, 32 and 33 in Supplementary Data 10) is strongly upregulated on desiccation. In contrast, all *PnHb* genes and their orthologous counterparts in *P. vanderplanki* are downregulated in the desiccating larvae (Supplementary Data 10)33. Four of the 6 anhydrobiotic chironomid-specific *Hb* genes (*PvHb*17, 23, 32 and 33) also showed high mRNA levels under non-stressed conditions and are 4 of the 10 most highly expressed *Hb* genes in the wet larvae (Supplementary Data 10).

Our data on developmental stage-specific and anhydrobiosis process-specific patterns of *Hb* gene expression suggest that the multiple members of this gene family in the chironomids are most likely involved in specialization of the function rather than a general increase in *Hb* protein dosage, as proposed by some authors33. One possibility is that the *P. vanderplanki*-specific Hbs may have specific properties allowing them to provide effective delivery of oxygen under conditions of increased molecular crowding in the larvae due to desiccation. Alternatively, the Hbs may protect larvae against free radicals generated during severe dehydration.

Other examples of the process generating ARIds in *P. vanderplanki* genome (Fig. 2b) are gene clusters coding small heat–shock proteins and several unknown genes16.

**Aquaporins and dehydration process en route to anhydrobiosis.** Water channels or aquaporins (AQPs) primarily control permeation of water across the phospholipid bilayer of the cell membrane34. Thus, AQPs most likely play pivotal roles in the dehydration process en route to anhydrobiosis. We have identified five AQP-coding genes in each species of the chironomid (*Aqp1*–5;...
Supplementary Table 3), which is a similar number to other dipteran insects. Aqp1 encodes the water-specific AQP and showed differences in the mRNA-level response to desiccation between the two species. In the anhydrobiotic species, the corresponding gene was strongly responsive to desiccation and its expression increased by more than threefold. The mRNA for Aqp1 represented more than 80% of all AQP mRNAs in the larva subjected to slow desiccation. In contrast, expression of Aqp1 in P. nubifer under desiccation did not increase (Supplementary Table 3). Aqp1 was previously assumed to play a key role in trafficking water out of the body of the anhydrobiotic P. vanderplanki larvae. The current comparative analysis of P. vanderplanki and P. nubifer Aqp1 orthologues revealed evolution of specific mRNA regulation in response to desiccation. In P. vanderplanki, the total abundance of Aqp1 mRNA in the larvae was higher under normal conditions and further drastically increased under slow desiccation (Supplementary Table 3).

**Anhydrobiosis-related trehalose metabolism pathway.** Trehalose is a disaccharide of glucose that stabilizes intact cells in the dry state and replaces water. P. vanderplanki larvae synthesize large amounts of trehalose (up to 20% dry mass) during dehydration en route to anhydrobiosis. In dehydrated larvae, trehalose stabilizes the structure of biomolecules. Recently, we isolated genes coding for trehalose-6-phosphate synthase (TPS) and trehalose-6-phosphate phosphatase (TPP). These genes govern trehalose synthesis and trehalase (TREH) hydrolyses trehalose into its component glucose units. Trehalose is abundantly synthesized from glycogen in the fat body in response to water loss. The elevated sugar concentration is achieved by increased TPS and TPP activities and suppression of TREH activity. In addition, in P. vanderplanki, TRET1 facilitates the transport of trehalose across cellular membranes of the fat body cell. TRET1 retains a high capacity for transport activity even when trehalose is highly concentrated in the dehydrating larval body during the final stage of entry into anhydrobiosis.

In both chironomids, the genes encoding members of the trehalose metabolism pathway (TMP) including TRET1, TPP, TPS and TREH are represented by single-copy genes (Supplementary Table 4) that are similar to those of other insects both in sequence and gene structure. However, the TMP genes in the two chironomid species responded very differently to desiccation. The expressions of both TPS and TREH were drastically elevated in P. vanderplanki but remained unchanged in P. nubifer. In contrast, the genes encoding TRET1 and TPP in both species showed a similar pattern of expression and were slightly increased in response to desiccation (Supplementary Table 4). These data suggest that in P. vanderplanki, the accumulation of trehalose during the onset of anhydrobiosis is mediated not by a special set of genes (the number and structure of TMP genes in P. vanderplanki is similar to that of other insects), but rather by the evolution of gene expression control mechanisms responsive to desiccation.

Another important question is how a simultaneous increase in TREH mRNA and protein in the larvae could be associated with a general decrease in the activity of this trehalose-hydrolysing enzyme. Post anhydrobiosis, the rapid decrease in trehalose concentration in the larvae is mediated by TREH activity. Our previous data suggest that the enzyme is stored in the larvae in advance of its requirement during rehydration.

As mentioned above, TRET1 has a pivotal role for transport of trehalose synthesized in the fat body in the desiccating larvae. Other desiccation-inducible transporters might be involved in trehalose uptake in the peripheral tissues on dehydration (see Supplementary Note 3; Supplementary Table 5).

**Discussion**

Elucidating the origins of ARIs and the amplification of genes in these regions are critical for understanding the genomics of anhydrobiosis. Desiccation causes extensive DNA damage and anhydrobiotic larvae require several days to repair the damage following rehydration. This DNA damage likely increases the frequency of genome rearrangements. Furthermore, cycles of anhydrobiosis might promote HGT as suggested for rotifers. A preliminary analysis identified at least 12 expressed genes in the P. vanderplanki genome with strong evidence for HGT and these genes are mostly from prokaryotes (Supplementary Table 6). The hypothetical scenario of HGT would be an integration of foreign DNA to the genome of the chironomid from consumed bacteria because it is the primary food source of the larvae. In addition, potential disruptions of cell membranes and severe DNA fragmentation associated with every cycle of anhydrobiosis are likely to facilitate this process.

The AQP and TMP genes are not located in ARId regions, but show desiccation-responsive expression patterns that are similar to what is observed with the genes from ARId clusters. Examining the P. vanderplanki genome and comparing ARIs and AQP or TMP gene-coding regions are promising models for uncovering the structure of yet unknown desiccation-inducible cis-elements and/or trans-regulation modules such as transcription factors and noncoding RNAs.

In summary, anhydrobiosis-associated genes (including chaperone-like proteins, antioxidants, aging-related proteins and unique globins) in P. vanderplanki have undergone massive expansion within the gene clusters or ARIds. For example, phylogenetic analysis of PvLea genes shows that the majority have no significant similarity (BlastP bit score < 100) to LEA protein genes in other organisms and most likely resulted from extensive gene duplication after a founding HGT event (Fig. 2; Supplementary Data 5). Finally, an important event for the evolution of anhydrobiosis in P. vanderplanki is the acquisition of new regulatory pathways that are strongly responsive to desiccation. These regulatory pathways control the expression of the gene clusters located in ARId regions and also control isolated genes co-opted for anhydrobiosis (TMP or AQP genes). All these evolutionary changes are likely to be further mediated by P. vanderplanki ecology and DNA-damaging effects of desiccation. The nature of P. vanderplanki habitats (large isolated rocks), the poor flying ability and strong selection pressure due to the long dry season in semi-arid areas of Africa facilitated microevolutionary patterns in this species. Our recent in vitro data show the direct contribution of the members in the expanded gene clusters (such as LEA proteins and antioxidants) to neutralize the effects of desiccation. Another possibility is a non-adaptive gene drift-based origin of the observed changes in P. vanderplanki genome. Future comparative investigations on isolated P. vanderplanki populations will certainly help to verify these hypotheses for the de novo acquisition and the evolution of anhydrobiosis in this unique insect.

**Methods**

**Insects.** Highly inbred lines of these sibling species that differ in their ability to resist complete desiccation were used for genomic DNA extraction. P. vanderplanki and P. nubifer larvae were reared on a 1% agar diet containing 2% commercial milk under controlled photoperiod (13 h light: 11 h dark) and temperature (27–28 °C) conditions.

**Number of chromosomes of P. vanderplanki and P. nubifer.** We observed polytene chromosomes in the salivary glands of fully hydrated larvae of P. vanderplanki and P. nubifer. Fourth instar larvae were fixed in a 3:1 mixture of 90% ethanol and glacial acetic acid and stored at −80 °C until use. Salivary glands were dissected out and stained in 1.0% orcein in 45% acetic acid. They were then washed lightly in 45% acetic acid and squashed in 50% lactic acid.

| Chromosome | P. vanderplanki | P. nubifer |
|------------|----------------|-----------|
| Number     | 21             | 20        |
Estimation of the chironomids' genome sizes by flow cytometry. The genome sizes of the two species were determined by flow cytometry (Cornette et al., in prep.). Briefly, heads of adult chironomids were homogenized into a solution of 0.5% Triton X-100 in 1 mM phosphate buffered saline buffer, before staining the nuclei with 5 µg ml$^{-1}$ of propidium iodide and filtering on a 30-µm mesh filter (Partec, Münster, Germany). The DNA content of stained nuclei was measured by a Coulter Epics Elite flow cytometry system (Beckman Coulter, Indianapolis, IN). The 2C DNA content of the sample was compared with the standard 0.36 pg DNA of D. melanogaster diploid nuclei.

Genomic DNA sampling. Genomic DNA from over 500 final instar larvae (of ~1 mg wet body weight) for construction of mate-pair libraries and other experiments was isolated with conventional cetrimonium bromide (CTAB) method$^{40}$ and NucleoSpin tissue (Machery-Nagel, Düren, Germany), respectively.

Genome sequencing. Genome sequences were obtained using paired-end and mate-pair protocols on Illumina HiSeq 2000, GAIIx and SOLiD 4 instruments. Genomic DNA was fragmented, libraries prepared and sequencing conducted according to the manufacturer's protocols. Mate-paired libraries for the SOLiD 4 system (Life Technologies, Carlsbad, CA) with inserts of ~2.5 kb were constructed from 5 µg of genomic DNA, and deposited on two quarters of a flow cell for each sample. Fifty base reads were obtained from each of the F3 and R3 tags, with 22 Gbp for both P. vanderplanki and P. nubifer libraries. To construct the libraries for whole-genome sequencing, DNA was processed using a TruSeq DNA Sample Preparation kit v.2 (Illumina, San Diego, CA) according to the manufacturer's instructions. Library lengths, as assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA), were 541 for P. nubifer and 577 for P. vanderplanki. Libraries were quantified using Qubit 2.0 (Life Technologies) and real-time PCR, and diluted to final concentration of 9 pM. Diluted libraries were clustered using a cBot instrument (Illumina) with a TruSeq PE Cluster Kit v3 (Illumina) and sequenced using a HiSeq 2000 sequencer (Illumina) with TruSeq SBS Kits v3-HS (Illumina), read length 101 from each end. Poly A-mRNA libraries were constructed using TruSeq RNA Sample Preparation kit v.2 (Illumina, San Diego, CA) according to the manufacturer's instructions. Library lengths, as assessed using an Agilent 2100 Bioanalyzer, were 901 for P. vanderplanki, and 1001 for P. nubifer. Libraries were quantified using Qubit 2.0 (Life Technologies) and real-time PCR, and diluted to final concentration of 9 pM. Diluted libraries were clustered using a cBot instrument (Illumina) with a TruSeq PE Cluster Kit v3 (Illumina) and sequenced using a HiSeq 2000 sequencer (Illumina) with TruSeq SBS Kits v3-HS (Illumina), read length 101 from each end. Poly A-mRNA libraries were constructed using TruSeq RNA Sample Preparation kit v.2 (Illumina) and quantified and sequenced in the same way as genomic DNA libraries. For the P. vanderplanki genome, sequencing with a single-molecule real-time sequencer was also performed. Approximately 6- and 10-kb insert libraries were constructed and sequenced with C2 chemistry using PacBio RS (Pacific Biosciences, Menlo Park, CA) for 34 cells (version 2).

Using two types of libraries, the GAIIx platform generated a total of 36.8 Gbp of P. vanderplanki sequence data (Supplementary Table 7). Furthermore, the HiSeq 2000 platform produced 6.9 Gbp of sequence data, the SOLiD 4 system generated 20.9 Gbp data and the PacBio RS yielded 1,479,033 reads with 1.7–2.7 kb mean maximum subread length, total 2.9 Gbp of independent fragment reads. On the basis of the genome size estimation of 100 Mbp (see above), the total of 67.5 Gbp of sequence data obtained corresponds to ~562-fold coverage of the P. vanderplanki genome (Supplementary Table 7). In the case of P. nubifer, the HiSeq 2000 platform generated 7.0 Gbp sequence data, providing ~58-fold coverage of that chironomid genome (Supplementary Table 8).

Genome sequence assembly. The shotgun, paired-end and mate-pair reads were assembled de novo by the Platania Assembler$^{41}$ (http://platania.bio.titech.ac.jp/) and the remaining gaps were filled with PacBio RS reads using the PBjelly pipeline$^{42}$.

Fosmid-end sequences. The fosmid library of P. vanderplanki genome was prepared by Takara Bio (Shiga, Japan). Randomly selected fosmid clones were end sequenced by the Sanger method using an ABI 3130xl sequencer (Life Technologies) and 397 for P. vanderplanki and 597 for P. nubifer. Libraries were quantified using Qubit 2.0 (Life Technologies) and real-time PCR, and diluted to final concentration of 9 pM. Diluted libraries were clustered using a cBot instrument (Illumina) with a TruSeq PE Cluster Kit v3 (Illumina) and sequenced using a HiSeq 2000 sequencer (Illumina) with TruSeq SBS Kits v3-HS (Illumina), read length 101 from each end. Poly A-mRNA libraries were constructed using TruSeq RNA Sample Preparation kit v.2 (Illumina, San Diego, CA) according to the manufacturer's instructions. Library lengths, as assessed using an Agilent 2100 Bioanalyzer, were 901 for P. vanderplanki, and 1001 for P. nubifer. Libraries were quantified using Qubit 2.0 (Life Technologies) and real-time PCR, and diluted to final concentration of 9 pM. Diluted libraries were clustered using a cBot instrument (Illumina) with a TruSeq PE Cluster Kit v3 (Illumina) and sequenced using a HiSeq 2000 sequencer (Illumina) with TruSeq SBS Kits v3-HS (Illumina), read length 101 from each end. Poly A-mRNA libraries were constructed using TruSeq RNA Sample Preparation kit v.2 (Illumina) and quantified and sequenced in the same way as genomic DNA libraries. For the P. vanderplanki genome, sequencing with a single-molecule real-time sequencer was also performed. Approximately 6- and 10-kb insert libraries were constructed and sequenced with C2 chemistry using PacBio RS (Pacific Biosciences, Menlo Park, CA) for 34 cells (version 2).

Using two types of libraries, the GAIIx platform generated a total of 36.8 Gbp of P. vanderplanki sequence data (Supplementary Table 7). Furthermore, the HiSeq 2000 platform produced 6.9 Gbp of sequence data, the SOLiD 4 system generated 20.9 Gbp data and the PacBio RS yielded 1,479,033 reads with 1.7–2.7 kb mean maximum subread length, total 2.9 Gbp of independent fragment reads. On the basis of the genome size estimation of 100 Mbp (see above), the total of 67.5 Gbp of sequence data obtained corresponds to ~562-fold coverage of the P. vanderplanki genome (Supplementary Table 7). In the case of P. nubifer, the HiSeq 2000 platform generated 7.0 Gbp sequence data, providing ~58-fold coverage of that chironomid genome (Supplementary Table 8).

Identification of chironomid genes. Predicted genes were annotated using a set of publicly available tools. We performed BlastP (version 2.2.25)$^{61}$ searches of gene models against NCBI-nr with an expectation value of 1.0e-05. As a result, for P. vanderplanki and P. nubifer, 2,558 out of 14,579 (17.5%) and 3,201 out of 13,352 (24.0%) genes did not have a significant hit, respectively. Protein domain annotation of gene models was done by combination of HMMER3 and domain models' Pfam$^{62,63}$. To obtain more comprehensive information on protein function, all deduced proteins were subjected to InterProScan (version 4.8) analysis. The result of annotation for all P. vanderplanki and P. nubifer genes, together with the expression data (RPKM), was prepared as a single MS-Excel file (Supplementary Data 11). The frequency of Gene Ontology terms and InterPro IDs$^{64}$ for P. vanderplanki and P. nubifer was also summarized (Supplementary Data 1).

Estimation of expression of the predicted genes. The mRNA expression levels for the entire transcript set (Supplementary Table 9) were estimated using the RPKM values. For confident comparison of the transcriptional response to desiccation in the two chironomids, only two sets of data (wet larvae versus larvae desiccated for 24 h, termed D0 and D24) were used. An increase in expression of more than threefold and an RPKM value > 10 for either data set was used as the criteria for placement of a gene in the 'desiccation-responsive' group. The expression data were represented by tracks in the genome browser.

Genome browser. A genome browser for the assembled genome sequences has been established using the Generic Genome Browser (GBrowse) 2.17 (ref. 65)
Pipeline for identification of horizontally transferred genes. We used a custom phylogenomic pipeline to build gene trees for all predicted coding regions in P. vanderplanki and P. nubifer. Scripts are available from the authors on request. Predicted amino-acid sequences were first queried using BlastP against a local database consisting of NCBI’s Reference Sequence and predicted protein sequences from recently sequenced microbial eukaryotes (IGI genome portal and Ghent University’s online genome annotation server BoGAS). For each Blast result, a bit score was considered significant if the E-value was < 1e−60 and the fraction conserved was > 0.3. If a hit met these sequence similarity thresholds, the associated sequence was extracted from the database using a custom Perl script. To reduce the number of paralogues in the analysis, only the top four hits per species were extracted. Extracted sequences were reordered based on global similarity to the query sequence with MAFFT using the minimum linkage clustering method and rough distance measurement (number of shared 6-mers)21. After reordering, the files were reduced to include only the top 200 sequences, and files with fewer than 4 sequences were eliminated. Alignments were performed with MAFFT using the automated strategy selection. Poorly aligned positions and sequences were removed from the alignment using REAP72, and trimmed alignments were further refined by a second MAFFT alignment using the same parameters as above. Phylogenetic trees were inferred using FastTree, assuming a JTT + CAT amino-acid model of substitution and 1,000 bootstrap replicates23. For each tree, the phylogenetic sister group to Polyplectidion was determined using SICLE27 (http://eceweb.arizona.edu/sicle/). Finally, the candidate genes were analysed manually to filter out potential false-positive cases. The results of final screening are summarized in Supplementary Table 6.
46. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl 1): i351–i358 (2005).

47. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

48. Cornette, R. et al. Identification of anhydrobiosis-related genes from an expressed sequence tag database in the cryptobiotic midge *Polyplethalmus vanderplanki* (Diptera; Chironomidae). *J. Biol. Chem.* **285**, 35889–35899 (2010).

49. Meyer, E. et al. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFLX. *BMC Genomics* **10**, 519 (2009).

50. Pertea, G. SeqClean <http://compbio.dfci.harvard.edu/tgi/software/> (2005–2006).

51. NCBI. The UniVec Database <http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/> (2013).

52. Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. * BMC Bioinformatics* **7**, 62 (2006).

53. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).

54. Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1873 (2005).

55. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. *Proceedings of the German Conference on Bioinformatics (GCB)* **99**, 45–56 (1999).

56. Kent, W. J.. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

57. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36 (2013).

58. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

59. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).

60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Bioinformatics* **30**, 208–210 (2014).

61. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local sequence comparison. *J. Mol. Biol.* **215**, 403–410 (1990).

62. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).

63. Finn, R. D. et al. The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).

64. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

65. Stein, L. D. et al. The generic genome browser: a building block for a model organism system database. *Genome Res.* **12**, 1599–1610 (2002).

66. Demin, A. G., Polukonova, N. V. & Mugue, N. S. Molecular phylogeny and the evolution of the Insects (Chironomidae, Diptera). *Russ. J. Genet.* **26**, 120–125 (2003).

67. Dixit, J. et al. Phylogenetic inference of Indian malaria vectors from multilocus DNA sequences. *Infect. Genet. Evol.* **10**, 755–763 (2010).

68. Papoucheva, E., Proviz, V., Lambkin, C., Goddeeris, B. & Blinov, A. Phylogeny of the endemic Baikalian *Sergentia* (Chironomidae, Diptera). *Mol. Phylogenet. Evol.* **29**, 120–125 (2003).

69. Grimaldi, D. & Engel, M. S. *Evolution of the Insects* 755 (Cambridge Univ. Press, 2005).

70. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

71. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

72. Hartmann, S. & Vision, T. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol. Biol.* **8**, 95 (2008).

73. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).

74. DeBlasio, D. & Wiscaver, J. SICLE: a high-throughput tool for extracting evolutionary relationships from phylogenetic trees. Preprint at http://arXiv:1303.5785 [q-bio.GN] (2013).

**Acknowledgements**

We extend our gratitude to the Federal Ministry of Environment of Nigeria for permitting research on *P. vanderplanki*. We acknowledge Y. Saito and T. Shiratori for their help with rearing and maintenance of midges and larvae, and Y. Sato for preparing sequencing libraries and for other technical assistance with sequencing. A part of sequencing was supported by OIST internal fund to the Marine Genomics Unit, NIAS internal fund and NIBB internal fund. The supercomputing was supported by the IT Section of OIST and NIAS. A part of computation was performed on the Supercomputer System in National Institute of Genetics. This work was supported in part by Grants-in-Aids from MEXT/SPS KAKENHI (Grant Number 21688004, 21282001, 23128512, 23780055, 24120006, 25128714 and 25252060), Japan; subsidy of the Russian Government to support the Program of competitive growth of Kazan Federal University among world class academic centres and universities; Russian Foundation for Basic Research (No 12-08-33157 mol_a_ved and No 14-04-01657_A); grant 11.G34.31.0008 from the Ministry of Education and Science of the Russian Federation; RFBR and Tatarstan government grant No 12-04-9707.

**Authors contributions**

T.Ki., N.S., O.G., R.C., A.S.K., T.O., and Y.S. designed and coordinated the project. M.D.L., Y.S., A.A.P., R.K., T.N., D.A., Va.G., Y.K., H.K., Sa.S., T.M., Ve.G., T.Ka., M.F., Sh.S., T.F.S. and M.H. performed genome and EST sequencing and assembly, and chromosome analysis; O.G., Y.S., R.C., T.Ka., R.H., S.K., E.S., T.N., Sh.S., A.M., J.W., M.H. and T.Ki. conducted annotation and analysis. O.G., T.Ka., N.S., T.N., R.C., Y.S. and T.Ki. prepared the manuscript.

**Additional information**

**Accession codes**: Sequence data for *P. vanderplanki* and *P. nubifer* have been deposited in GenBank/EMBL/DDJB nucleotide core database under the accession codes PRJDB1558 and PRJDB2914, respectively.

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

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

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**How to cite this article**: Gusev, O. et al. Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. *Nat. Commun.* 5:4784 doi: 10.1038/ncomms5784 (2014).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/4.0/