A First Glimpse at the Genome of the Baikalian Amphipod *Eulimnogammarus verrucosus*☆

Lorena Rivarola-Duarte[a,b,1], Christian Otto[b,c,d,1], Frank Jühling[b], Stephan Schreiber[e,f], Daria Bedulina[g,h], Lena Jakob[i], Anton Gurkov[g,h], Denis Axenov-Gribanov[g,h], Abdullah H. Sahyoun[d,j], Magnus Lucassen[i], Jörg Hackermüller[e,f,b], Steve Hoffmann[c,b], Franz Sartoris[i], Hans-Otto Pörtner[i,*], Maxim Timofeyev[g,h,*], Till Luckenbach[a,*], Peter F. Stadler[b,c,f,d,k,l,m,n,*]

[a] *Department of Bioanalytical Ecotoxicology, UFZ – Helmholtz Centre for Environmental Research, Permoserstraße 15, D-04318 Leipzig, Germany*
[b] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[c] *LIFE, Leipzig Research Center for Civilization Diseases, University Leipzig, Philipp-Rosenthal-Strasse 27, D-04107 Leipzig, Germany*
[d] *Bioinformatics Group, Department of Computer Science, Härtelstraße 16-18, D-04107, Leipzig, Germany*
`studla@bioinf.uni-leipzig.de`
[e] *Young Investigators Group Bioinformatics and Transcriptomics, Department Proteomics, UFZ – Helmholtz Centre for Environmental Research, Permoserstraße 15, D-04318 Leipzig, Germany*
[f] *RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, D-04103 Leipzig, Germany*
[g] *Irkutsk State University, Karl Marx 1, 664003, Irkutsk, Russia*
[h] *Baikal Research Centre, Lenina 22-21, 664003, Irkutsk, Russia*
[i] *Alfred-Wegener-Institute Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, D-27570 Bremerhaven, Germany*
[j] *Doctoral School of Science and Technology, AZM Center for Biotechnology Research, Lebanese University, Mitein Street, Tripoli, Lebanon*
[k] *Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany*
[l] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[m] *Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg, Denmark*
[n] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Abstract

Eulimnogammarus verrucosus is an amphipod endemic to the unique ecosystem of Lake Baikal and serves in particular as an emerging model in ecotoxicological studies. We report here on a survey sequencing of its genome as a first step to establish sequence resources for this species. From a single lane of paired-end sequencing data we estimated the genome size as nearly 10 Gb and we obtained an overview of the repeat content. At least two thirds of the genome are non-unique DNA, and a third of the genomic DNA is composed of just five families of repetitive elements, including low-complexity sequences. Attempts to use off-the-shelf assembly tools failed on the available low-coverage data both before and after removal of highly repetitive components. Using a seed-based approach we nevertheless assembled short contigs covering 33 pre-microRNAs and the homeodomain-containing exon of nine Hox genes. The absence of clear evidence for paralogs implies that a genome duplication did not contribute to the large genome size. We furthermore report the assembly of the mitochondrial genome using a new, guided "crystallization" procedure. The initial results presented here set the stage for a more complete sequencing and analysis of this large genome.

Keywords: *Eulimnogammarus verrucosus*, Lake Baikal, survey sequencing, repetitive elements, genome evolution, mitochondrial genome, Hox genes, miRNA

☆Rapid Communication
[*] Corresponding authors
[1] Joint First Authors

INTRODUCTION

Lake Baikal, located in an intracontinental rift zone in the central region of southern Siberia, is the world's oldest (25-30 million years), by volume largest (23,000 km^3) and deepest (1642 m) lake, containing about 20% of the world's liquid freshwater (Rusinek, 2012a). The environmental conditions of open- and deep-water zones have remained close to their current state for the last 2-4 million years (Kozhova and Izmest'eva, 1998; Timofeyev, 2010). Lake Baikal features very specific abiotic characteristics that distinguish it from all other freshwater bodies in the Palearctic: (1) high oxygen content throughout the entire water column, (2) stable low water temperature with long seasonal ice coverage of the lake surface, and (3) super-oligotrophic conditions (Kozhov, 1963).

As a unique ecosystem with exceptionally high degrees of biodiversity and endemism, it was designated in 1996 a UNESCO World Heritage Site[2]. So far, 2595 animal species from Lake Baikal have been identified or described, of which 80% are endemics (Timoshkin, 2001; Rusinek, 2012b). This high degree of endemism reflects the long evolutionary history of Lake Baikal in isolation from other freshwater bodies (Timofeyev, 2010). Virtually nothing is known about the molecular basis behind the physiological adaptations of the endemic species to the specific abiotic conditions of Lake Baikal. Next-generation sequencing technologies now provide the possibility to obtain comparatively affordably comprehensive genome and transcriptome data that will be useful for addressing such questions.

In particular, the *Amphipoda*, an abundant macroinvertebrate taxon in Lake Baikal, comprises more than 350 species and sub-species, all endemic to this ecosystem (Rusinek, 2012b). DNA sequence information for amphipods in general is still scarce. Beyond mitochondrial genomes used for phylogenetic analyses (Cook *et al.*, 2005; Bauza-Ribot *et al.*, 2009; Ki *et al.*, 2010; Krebes and Bastrop, 2012; Shin *et al.*, 2012) and a few individual nuclear genes, the only systematic resources are a BAC library generated from genomic DNA of *Parhyale hawaiensis* (Parchem *et al.*, 2010) and two very recent transcriptome studies also in *P. hawaiensis* (Zeng *et al.*, 2011; Blythe *et al.*, 2012). The closest relative with a well-developed genomic resource is the water flea *Daphnia pulex* (Colbourne *et al.*, 2011), but belongs to a different Class (*Branchiopoda*).

Amphipods are among the clades with highly variable genome sizes. *C*-values range from $C = 0.68$ in *Caprella equilibra* (Libertini *et al.*, 2003) to $C = 64.62$ in *Ampelisca macrocephala* (Belzile *et al.*, 2007). Within the family *Gammaridae*, the genomes sizes reported in the Genome Size Database vary between $C = 1.58$ to $C = 3.80$ (Gregory, 2012), while Vergilino *et al.* (2012) reports $C = 14.06$ for *Gammarus lacustris*. The genome sizes of baikalean amphipods are at present not available in the literature.

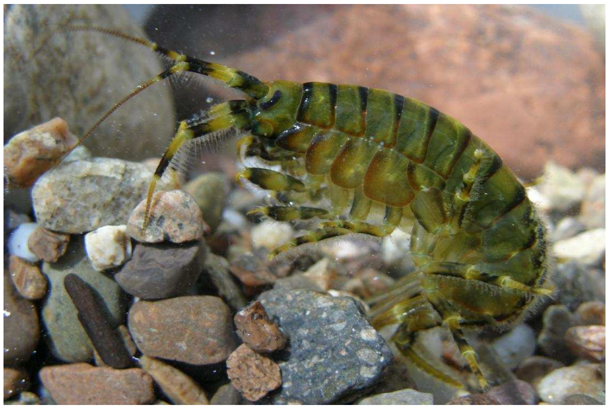

Figure 1: *Eulimnogammarus verrucosus* (Gerstfeldt, 1858) is an amphipod species endemic to Lake Baikal with a size up to 45 mm from telson to rostrum. Its typical color is green with black stripes across the body segments and antennae. The eyes are very slender and the segments of the meta- and urosome are armored by thorns. The karyotype is $n = 26$ (Salemaa and Kamaltynov, 1994a,b). This species is omnivorous and inhabits rocky substrata close to shore down to 10-15 meters water depth (Kravtsova *et al.*, 2004; Bazikalova, 1945). Photo by Vasiliy Pavlichenko.

Eulimnogammarus verrucosus (Gerstfeldt, 1858; Figure 1), an amphipod species endemic to Lake Baikal, is a typical inhabitant of the upper and sub-littoral zones of this lake, where it is found in relatively large numbers at water depths from 10 cm down to 15 m in substrate consisting of medium-size to large round pebbles (Bazikalova, 1945). This species has a temperature preference of 5-6°C with oxygen levels above 9mg O_2/l (Timofeyev and Shatilina, 2007). It is comparatively sensitive to higher temperatures. Juveniles are more tolerant to low oxygen levels and high temperature than adult stages (Shatilina *et al.*, 2011; Bedulina *et al.*, 2013). As *E. verrucosus* occurs close to the shore it may be particularly affected by water pollution and other human activities. Therefore it serves as a relevant model species for

[2] http://whc.unesco.org/en/list/754

ecotoxicological studies of the effects of anthropogenic contamination in the Lake Baikal amphipod fauna. Its genome size was unknown so far. We here present the first survey sequencing of the genome of a baikalian amphipod, *E. verrucosus*, from a single lane of an Illumina Hi-Seq 2000 as a way to obtain an overview and a baseline for deeper investigations in amphipod genomes.

MATERIALS & METHODS

Specimen

E. verrucosus specimen for sequencing was sampled by kick-sampling in 0.5 to 1 m deep water at the shore of Lake Baikal, close to the biological field station of Irkutsk State University in Bol'shie Koty (51°54′11.67″*N*, 105°04′07.61″*E*).

DNA isolation, Sequencing, and Data Preprocessing

A detailed description of the DNA isolation and sequencing protocols used in this work and the data preprocessing is given in the Supplementary Methods. Data that was preprocessed, clipped, and merged is hereafter denoted as "sequencing data" and used in all analyses, unless mentioned differently. Basic statistics on the data after processing with `CASAVA` (raw) as well as after both preprocessing steps is summarized in Table 1.

MicroRNA Annotation

MicroRNAs are rather easily detectable already in unassembled genomic data due to their small size and extreme level of sequence conservation. Using a simple `blastn` search, we identified reads that matched mature microRNA sequences from *P. hawaiensis* (21 queries, Blythe *et al.* 2012), *D. pulex* (45 queries from miRBase 19, Wheeler *et al.* 2009), and *Marsupenaeus japonicus* (48 queries, Huang *et al.* 2012). These reads were assembled using `SGA`, requiring a minimal overlap of 20 nt. Only contigs having trustworthy mature microRNA hits (≤ 2 errors) were kept. To check for precursors, the remaining contigs were analyzed with `cmsearch` of the Infernal package (Nawrocki *et al.*, 2009) using all miRNA-related covariance models from Rfam 11.0 (Gardner *et al.*, 2011). Being conservative, the best-scoring hit for each contig was only considered reliable if its `cmsearch` bit score exceeded 25, if the difference in bit score between best-scoring and second best-scoring hit was at least 10, and if the miRNA family of the mature sequence found on the contig matched the miRNA family of the covariance model that resulted in the best hit. We extracted the sequence of all reliable hits and compiled a set of 33 different precursor sequences of *E. verrucosus* containing 38 mature sequences, compiled in the electronic supplement[3]. We assigned -3p or -5p nomenclature to the mature microRNAs according to its position on the precursor. The strand of the precursor was selected according to the `infernal` hit.

With these mature sequences, we repeated the `blastn`-based search and the `SGA` assembly. Then, the entire set of preprocessed DNA-seq reads was mapped onto the assembled contigs using `bowtie2` (Langmead and Salzberg, 2012) in its local mode where partial overlaps can be detected as well. To control for spurious hits, any mapping was discarded if < 50% of the read length was mapped onto the contig or if the accuracy of the mapping, i.e., the fraction of matches in the alignment, was < 90%.

Hox gene annotation

In addition to microRNAs, we searched in the unassembled sequencing data for the ten Hox genes of the arthropod HOX cluster (Grenier *et al.*, 1997). These key developmental transcription factors contain the extremely well conserved homeodomain with a length of 60 amino acids. We therefore used `tblastn` to retrieve reads with high similarity to the Hox protein sequences of *Daphnia pulex* from Uniprot (UniProt Consortium, 2013) (accession numbers in 5'-3' order of the HOX cluster: *Abd-B* EFX86798.1, *Abd-A* EFX86800.1, *Ubx* EFX86802.1, *Antp* EFX86804.1, *Ftz* EFX86805.1, *Scr* EFX86800.1, *Dfd* EFX86808.1, *Zen*/*HOX3* EFX86809.1, *Pb* EFX86800.1, and *Lab* EFX86813.1). For each Hox gene, we aligned all matching reads with each other to get initial contigs. Subsequently, these genomic contigs of *E. verrucosus* were iteratively extended by searching for DNA-seq reads as extension candidates using `blastn`, aligning them to the contigs, and manual curating the sequence alignments to allow very few sequence errors (≤ 2) but to avoid invalid or ambiguous extensions. We stopped the iterations if no (unambiguous) extension was possible.

Since these genomic contigs of *E. verrucosus* were used in all subsequent Hox gene analyses, we outline in detail the precautions that were taken to prevent reads comprising biological variation to be used for extension. Here, biological variation might be present due to the minor contamination by congener species (inter-species) and due to the possibility of paralogs (intra-species). As a result of the low sequencing coverage, the distinction

[3] www.bioinf.uni-leipzig.de/publications/supplements/13-003

	Paired-end reads			Single-end reads		
	Reads (M)	Bases (Gb)	Avg. length (bp)	Reads (M)	Bases (Gb)	Avg. length (bp)
raw	352.7	35.6	101.0	-	-	-
clipped	352.7	33.3	94.3	-	-	-
clipped + merged	227.0	22.7	99.9	62.9	6.1	97.3

Table 1: Basic statistics of the data after sequencing and processing with the CASAVA pipeline (raw) as well as after preprocessing by clipping and merging steps. The clipped + merged data were cleaned after merging using an in-house script to resolve false positive and false negative clippings as well as illegitimate mergings (see text for details). These data were used in all subsequent analyses.

between sequencing errors and biological variation was generally difficult. Nevertheless, sequencing errors are expected to occur uniformly whereas biological variation is expected to be correlated with its selection pressure, i.e., in case of strong selection pressure, biological variation is repressed. Hence, even though the protein sequence of the homeodomain may be under strong selection pressure, 3rd codon positions and intronic sequences (if not overlapping non-coding genes) are under much weaker selection. We started the extension by searching for reads with high similarity to the Hox protein sequences of *D. pulex*. In this step, we might have captured reads from paralogous loci or congener species as well. However, we can expect that such “contaminating” reads either result in ambiguous extensions outside the highly conserved homeodomain and/or to exhibit “errors” primarily at 3rd codon positions. In this manner we can identify reads comprising biological variation and stop the iterative extension of the contig whenever multiple possibilities appear in the data. The absense of such reads thus provides us with upper bounds on the possible variability in a given region and argues against the existence of paralogous sequences or by congener contamination. The same arguments are true for miRNA contigs where the mature but not precursor sequence (except for the stem loop structure) is expected to be highly conserved.

Furthermore, we annotated the homeobox on the contigs. The contig sequences including the location of the homeobox can be found in the electronic supplement. To analyze whether one of these Hox genes appeared in multiple paralogous copies, the sequencing data was mapped onto them using segemehl (90% accuracy) and a position-wise coverage was calculated over the part of the homeobox present on the contigs.

For each contig, potential splice site donor and acceptor sites were predicted using the web service of MaxEntScan (Yeo and Burge, 2004). Acceptor and donor splice sites with *MaxEnt* scores above 7 and 4, respectively, are likely functional. To check for splice site conservation, we extracted the coding sequences of the Hox genes of *D. pulex* including the splice sites and manually aligned them to the corresponding contig of *E. verrucosus*. If the similarity of the alignment dropped significantly or if the homeobox on the contig ended prematurely, it is likely that the contig contains an intron-exon junction at this position. We therefore searched for a predicted splice site with good scores close to such locations. In addition, we searched for in-frame stop codons which indicate either the true end of the coding sequence or intronic sequence. If no splice site was identified with confidence, we mark the region as exonic until the first stop codon appeared despite the possibility of a true but less confident splice site before. By comparing the intron-exon structure on the contigs to the corresponding regions in *D. pulex*, we were able to determine whether there is splice site conservation, innovations or loss.

Repeat Analysis

Considering that repeated elements can be widely abundant in eukaryotic genomes (Richard *et al.*, 2008), we aimed to assess the classes and proportions of the most abundant repetitive elements present in the genome of *E. verrucosus*.

Using Jellyfishv1.1.6 (Marçais and Kingsford, 2011), 24-mers that appeared more than 500 000 times in our data were identified and assembled into 198 initial contigs using SGA (Simpson and Durbin, 2012), requiring a perfect overlap of at least 20 nt, followed by greedy extension. These initial contigs were further reduced to 97 extended contigs after manual curation to correct sequencing errors and collapse similar motifs using ClustalW algorithm for multiple alignment with default parameters (Thompson *et al.*, 1994) embedded in MEGA v.5.10 (Tamura *et al.*, 2011) and Sequencher software v4.8 (Gene Codes Corporation, Ann Arbor, U.S.A) with the parameters: dirty data assembly algorithm, optimize gap placement and use re-aligner, minimum match percentage between 85-95% and minimum overlap between 20-70 bp. In order to redefine the boundaries of these consensus sequences, we performed a mapping using segemehl (Hoffmann *et al.*, 2009) with 90% accuracy and trimming to segments with a coverage ≥ 1000X. It resulted in a set of 96 final repeat contigs and 41 core repeat sequences were identified within the repeat contigs. To identify low-complexity core repeats, we used the RepeatMasker web server (Smit *et al.*, 2013) with default parameters (except for cross_match search engine, slow speed/sensitivity, DNA sources: human, fruit fly, malaria mosquito and panicoids) and CENSOR-GIRI (Kohany *et al.*, 2006) with parameters: sequence source ALL, report simple repeats ON. We also calculated the relative entropy (or Kullback-Leibler divergence) of the dinucleotide content in the core repeat sequences compared to the background distribution in the entire sequencing data. The evolutionary history was inferred using MEGA v.5.10 with the Neighbor-Joining method (Saitou and Nei, 1987) including 1000 bootstrap samples. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura *et al.*, 2004). All positions containing gaps and missing data were eliminated. The core repeat sequences were classified into cluster according to the phylogenetic analysis. Annotation of these core repeat sequences was performed using the blast web service (Altschul *et al.*, 1990) at NCBI with parameters: nucleotide collection -nr/nt- database, optimized for ‘somewhat similar sequences’; blastn, and filter low complexity sequences OFF.

We further conducted a k-mer based repeat analysis on the survey sequencing data. Given k, let $S(k)$ be a k-mer set where $w \in S(k)$ iff w is a nucleotide sequence of length k and w is a substring of a read sequence or the reverse complementary of a read sequence in the sequencing data. The frequency $f(w)$ of each $w \in S(k)$ is defined as the total number of w in the read sequences or the reverse complement of read sequences. By definition, the frequency of k-mers not occurring in the sequencing data is 0. The total number n of k-mers in the read sequences (including the reverse complement) is given by $n = \sum_{w \in S(k)} f(w)$.

Assuming a random reference G of length m with $4^k \gg m$, there are $2 \cdot (m - k + 1)$ k-mers from both strands of G and each k-mer is expected to occur at most once. The probability of selecting a k-mer w from G is $p = 1/(2 \cdot (m - k + 1))$. The frequency X of a k-mer in the sequencing data after drawing read sequences with replacement from G can be modeled by the binomial distribution $X \sim B(n, p)$ where n and p is the total number of K-mers in the read sequences and the probability of selecting a k-mer from G, respectively.

We denoted a k-mer w with $f(w) = x$ repetitive iff $P(X \geq x) < 0.01$ since it would be highly unlikely that w is unique in G. Let x_{min} be the minimal value of x such that $P(X \geq x) < 0.01$ holds. The set $S_R(k)$ of repetitive k-mers is given by $S_R(k) = \{w \mid w \in S(k) \wedge f(w) \geq x_{min}\}$. Due to the classification, the repeat content C of $S(k)$ was calculated according to the following equation and it was used as an estimate of the repetitive content of G.

$$C = \frac{\sum_{w \in S_R(k)} f(w)}{\sum_{w \in S(k)} f(w)} = \frac{1}{n} \cdot \sum_{w \in S_R(k)} f(w)$$

Assembly of the Mitogenome

In order to identify and assemble the mitochondrial DNA, we followed a reference-based strategy since mitogenomic sequences of related amphipods are known. In total, we used the mitogenomic sequences of *Onisimus nanseni* (Ki *et al.*, 2010), *Gondogeneia antarctica* (Shin *et al.*, 2012), *Metacrangonyx longipes* (Bauza-Ribot *et al.*, 2009), *Parhyale hawaiensis*, and *Gammarus duebeni* (Krebes and Bastrop, 2012), where the latter within the suborder *Gammaridea* is expected to be the closest relative to *E. verrucosus* from the five species mentioned before.

To get initial seeds, or “crystals”, the sequencing data was mapped to the mitogenomes using segemehl with low minimum accuracy (80%). Due to the high divergence to other amphipods, the read alignments covered only a small portion of each mitogenome where the divergence was lower. The reference was subsequently mutated by means of consensus calling with the read alignments. In brief, if a position was covered by ≥ 20 reads and the majority of the aligned reads overlapping this position showed an alternative to the reference base, the reference base was replaced by the alternative. In similar manner, the reference sequence was altered in case of coherent insertions or deletions in the read alignments. The sequencing data was then mapped to the modified reference with segemehl, leading to an extension of the crystals.

This “crystallization strategy” of mapping and mutating was iteratively applied to each mitogenomic sequence until the extension of crystals stagnated. Subsequently, we collected reads which originated likely from the mitogenomic sequence of *E. verrucosus*. To identify these high-quality mitogenomic reads, the sequencing data was mapped to the last iteration of each mutated mitogenomic sequence using `segemehl`. All reads which mapped without errors to one of the mutated mitogenomes were considered mitogenomic and assembled to the set of 26 initial contigs using `SGA` (perfect overlaps ≥ 45 nt).

To obtain the entire mitogenome of *E. verrucosus*, we iteratively extended the initial contigs in both directions. In brief, we used a greedy extension method with majority voting. First, sequencing reads were identified which contained the first or last 45 nt of a contig and extended beyond the contig. The contig was extended by the sequence of maximal length that satisfied the following conditions: the extension must be supported by ≥ 80% of the reads with minimum of 100 reads. The search for reads with start or end overlaps and the subsequent extension was iteratively applied. Afterwards, the sequencing data was mapped to the extended contigs using `segemehl` and the perfectly mapping reads were again subjected to an assembly with `SGA` (perfect overlaps ≥ 45 nt). The resulting set of final contigs contained the putative mitogenomic sequence of *E. verrucosus* as the only contig with the expected mitogenomic length (> 14 kb) and a more than 20-fold higher mean position-wise coverage compared to all other contigs.

To reduce the computational effort, we used only a subset of reads pre-filtered by `blastn` during intermediate steps instead of mapping the entire sequencing data at each iteration step.

Annotation of the Mitogenome

The mitochondrial genome was annotated using the `MITOS` pipeline (Bernt *et al.*, 2012). All mitochondrial protein coding genes were detected. However, a subset of the tRNAs and the 16s ribosomal RNA was not found in the initial analysis. A manual comparison of the annotations to the mitogenome of *Gammarus duebeni* (Krebes and Bastrop 2012) showed that nearly all protein genes were annotated too long. We therefore ran `MiTFi` (Jühling *et al.*, 2012) independently to identify all mt-tRNAs. The missing 16S ribosomal RNA were thus identified easily within the raw data; it was rejected in the initial `MITOS` because of substantial overlaps with the adjacent protein sequence. The ribosomal RNAs matched only a subset of domains that were included in the general rRNA models of `MITOS`. The mitogenome pattern matches that of *G. duebeni*. The protein coding genes were compared to *Gammarus duebeni* and manually adapted by searching for adequate start and stop codons in the corresponding genome areas. Both ribosomal RNAs were annotated according to the *Gammarus duebeni* genome. The mitogenome was visualized with the help of `mtviz`[4].

Attempts to de novo *assembly*

We tested several approaches towards *de novo* genome assembly of the survey sequencing data of *E. verrucosus*. Prior to any assembly, read sequences containing ambiguous bases (e.g. N) and reads clearly derived from the mitogenome (≥ 90% accuracy) were excluded. The remaining reads constituted the first dataset used for assembly. For the second dataset, we additionally removed repetitive subsequences. More precisely, we identified 30-mers that appeared more than six times within the sequencing data using `Jellyfish` v.1.1.6 (Marçais and Kingsford, 2011) and excluded read sequences that comprised any of these 30-mers or their reverse complements. To assemble both data sets, we used `SOAPdenovo` v.1.05 (Li *et al.*, 2010b) and `Velvet` v.1.2.08 (Zerbino and Birney, 2008) with *k*-mer sizes 23 and 31. The assemblies were evaluated using common measures such as N50 and number of contigs longer than 100 bp and 1 kb.

RESULTS

Estimating the Genome Size

To estimate the size of the nuclear genome, it is common to use either experimental (c-values) or *k*-mer based approaches, e.g. in the giant panda genome project (Li *et al.*, 2010a). However, due to the low coverage of our survey sequencing data, the *k*-mer based approach was not successful in discriminating sequencing errors from the coverage in unique regions (data not shown). Instead, we assessed the coverage and applied the Lander-Waterman equation (Lander and Waterman, 1988) to estimate the genome size.

We used a small number of non-repetitive DNA and mRNA sequences published in previous studies: myosin heavy chain mRNA partial cds GenBank: AF474964.1 (Benson *et al.*, 2013), *gapdh* mRNA partial cds (Bedulina *et al.*, 2013) and heat shock protein 70 (*hsp70*) gene complete cds GenBank: JQ003919.1. After mapping the sequencing data with `segemehl` (90% accuracy) to those sequences, the position-wise coverage was calculated. The reads mapping at N loci were counted as $1/N$ to avoid counting multiply-mapped reads

[4] http://pacosy.informatik.uni-leipzig.de/mtviz/

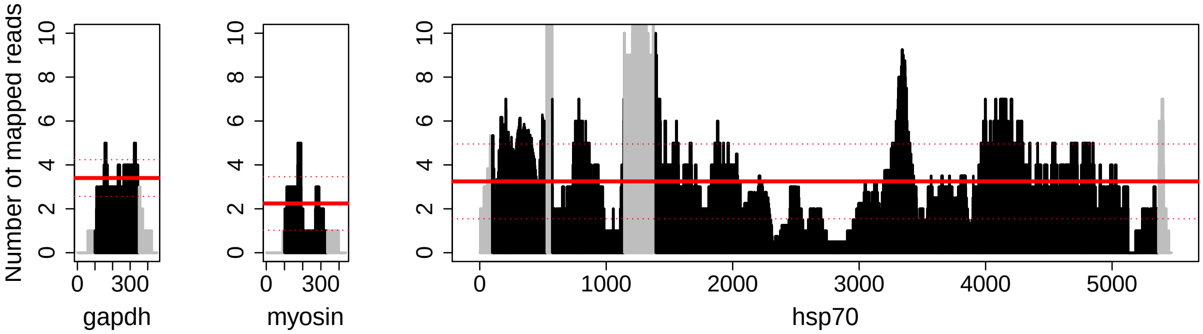

Figure 2: Coverage by sequence position after mapping the sequencing data to published DNA/mRNA sequences of *E. verrucosus*: *gapdh*, *myosin*, and *hsp70*. Mean and standard deviation for the coverage of each sequence is indciated as solid and dotted lines, respectively. In the coverage estimation, only the black colored segments were considered. The first and last 100 bases of each sequence were skipped due to the coverage decay towards the ends. Furthermore, two internal segments (positions 530-574 and 1143-1392) in *hsp70* with a coverage of around 44 and 16, respectively, were also excluded from the calculation (see text for further details).

multiple times. The position-wise coverage over the sequences is shown in Figure 2. To avoid a bias in the coverage calculation, the coverages at the first and last 100 bases were excluded (gray-shaded in Figure 2) since the coverage naturally decays towards the ends. Moreover, in *hsp70*, two internal segments (shaded in gray) were not considered in the analysis due to their high degree of multiple mappings. For example, the region 1143-1392 of the hsp70 gene is part of the promoter region containing the GAGA-factor binding site (GAGA) and two heat shock elements -HSE- (gaatgttcattttaaatag and gaatgatctgaaaag) (Bedulina *et al.*, 2013). HSE can be found in many stress-inducible genes and GAGA-factor binding site is also a common element in promoters (Gonsalves *et al.*, 2011). The extreme increase of coverage in the short region 530-574 cannot be explained by mapping full-length reads and hence was discarded as potential artifact. Taken together, we estimated the coverage as 2.96 ± 0.72 X.

To improve this preliminary estimate of the overall genome coverage, we searched for additional genomic regions that (i) can be identified already in unassembled data and (ii) are likely to be unique in the genome. MicroRNAs are particularly well-suited for this analysis because mature miRNA sequences are very short but highly conserved and thus can readily be detected in the sequencing data. In *E. verrucosus*, we identified 38 mature microRNAs in 33 distinct pre-miRNAs belonging to 30 different miRNA families. In three of them (mir-2, mir-184, mir-263), we found two divergent precursors, respectively. For the mir-184 family, we found distinct mature sequences that perfectly matches dpu-miR-263a and dpu-miR-263b from mirBase, respectively. The dataset comprised the majority of the microRNA families known to have been evolved before the split of crustaceans and insects (Hertel *et al.*, 2006; Tanzer *et al.*, 2010; Campbell *et al.*, 2011).

Position-wise coverage was estimated by mapping the sequencing data with `segemehl` (90% accuracy) to the contigs containing microRNA precursors. The coverage distribution of each mature miRNA of *E. verrucosus* is illustrated as box-whisker plots (see Figure 3). In addition, in the right-hand side panel, the distribution of the mean coverage values for the individual mature microRNAs is shown. The average of this distribution served an average of the overall genome coverage. We obtained a value of 2.92 ± 0.27 X, which was consistent with, but more accurate than, the estimate obtained from the few protein-coding loci. With a yield of 28.8 Gb preprocessed sequencing data (Table 1), we estimated the genome size of *E. verrucosus* at 9.96 ± 0.92 Gb. This value might be slightly overestimated due to presence of reads derived from contaminants in the sample. However, we estimate that these account for less than 5% of the data (see below), i.e., at most half of a standard deviation.

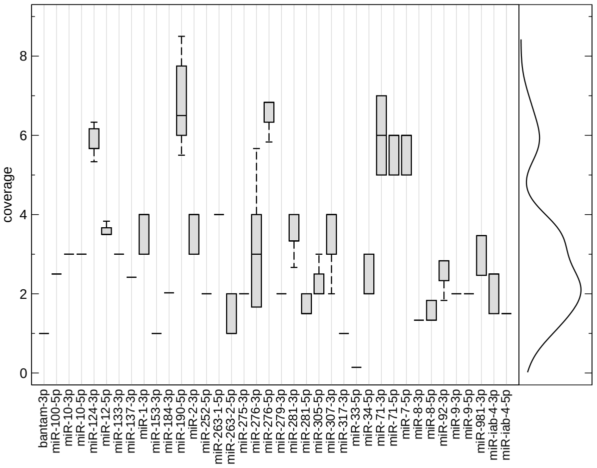

Figure 3: Box-whisker plots of the position-wise coverage for each newly annotated mature miRNA of *E. verrucosus*. The right panel shows a Gaussian kernel-smoothed density of the distribution of coverage means. The distribution of coverage means is used to estimate the genome size of *E. verrucosus* (see text for further details).

Hox genes

The Hox genes are a group of transcription factors with key roles in animal development that are arranged in a single gene cluster in most bilaterian animals. Arthropods typically harbour 10 genes (Grenier *et al.*, 1997). Hox genes have been used as indicators for large-scale duplications in the past (Ruddle *et al.*, 1999; Crow *et al.*, 2006) and hence they can also indicate whether genome duplication(s) has/have contributed to the large genome size of *E. verrucosus*. Using the sequencing data, we performed a semi-automatic extension procedure with hands-on work in order to assemble the genomic region of the homeodomain of ten Hox genes (see Methods section for details). The absense of ambiguous extensions or coherent sequence variation indicated that reads containing biological variation did not play a role during extension. The median size of the resulting contigs was 430 nt. With the exception of *Ubx*, peripheral regions of contig sequences showed only minor similarity to *D. pulex*, pointing to either a much less conserved exonic or more likely an intronic region (see Methods section for an explanation).

We were not able to identify any read in the sequencing data clearly originating from the Hox gene *Ftz*. This could be caused by insufficient coverage at the homeodomain motif of *Ftz*. A more plausible explanation is that *Ftz* might be absent in *E. verrucosus* since so far no *Ftz* gene has been identified in any member of the *Malacostraca* (Deutsch and Mouchel-Vielh, 2003).

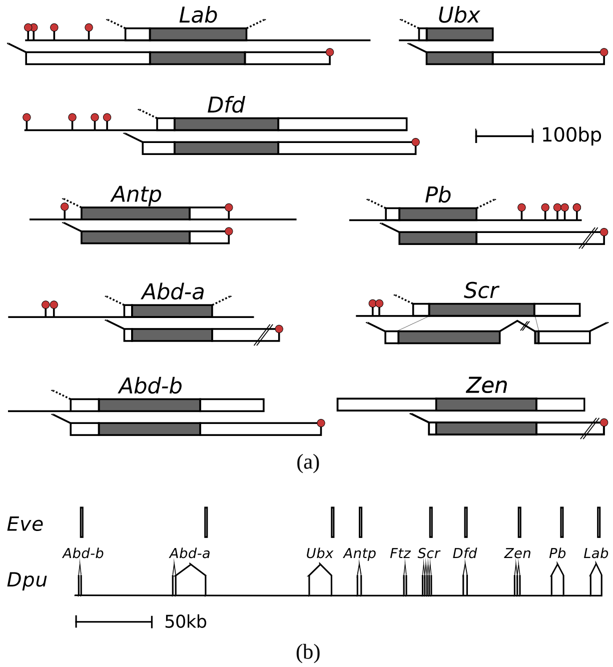

Figure 4: Comparison of Hox genes between *E. verrucosus* and *Daphnia pulex*. (a) Contigs containing parts of homeobox genes. In each case, the upper block represents the contigs obtained from *E. verrucosus* and the lower block, the corresponding exons from *D. pulex*. Shaded box indicate the homeobox sequence, red dots represent stop codons. Dashed lines indicate putative splice sites. (b) The genomic organization of the Hox genes in *D. pulex* (in 5′-3′ orientation) is shown in scale.

We investigated whether one of the Hox genes may be present in multiple paralogous copies in the genome of *E. verrucosus* by analyzing the sequencing coverage over the homeobox (see Methods for details) and illustrating it gene-wise as Box-Whisker plots (see Supplementary Figure S1). The average coverage of about 3 X was consistent with the microRNA-based estimate. Although the coverage for *Zen*/*HOX3* is twice as large, we did not observe any sequence variation (not even in 3rd codon positions) that would support a gene duplication, which (outside the vertebrates) has been observed e.g. for several Hox genes in *Chelicerata* (Schwager *et al.*, 2007) and the *Zen* gene in some flies (Stauber *et al.*, 1999). It is possible that a very recent gene duplication of *Zen* with virtually no sequence variation might have been missed due to the limited sequencing data. To detect such an event, it would be necessary to assemble the full *Zen* gene with distinct peripheral genomic regions. Nevertheless, it would still not be an indication of large-scale genome duplications as this would affect other Hox genes as well.

In order to demostrate the use of the genome sequencing data, we investigated the conservation of splice sites. Figure 4 summarizes the results of a comparison between the contigs and the corresponding regions of the Hox genes of *Daphnia pulex*. In the case of the acceptor splice sites, we found conservation in all Hox genes except for *Lab* and *Zen*. In the former case, a predicted splice site was identified 177 nt downstream of the annotated acceptor site in *Daphnia pulex* whereas no confident acceptor splice site was present in the case of *Zen*, indicating a splice site loss. Since the homeodomain is located in the last exon of *D. pulex* in all Hox genes except for *Scr*, we did not find conservations of donor sites. We found three splice site innovations, one in *Abd-A*, *Lab* and *Pb*. In addition to *Zen*, a potential splice site loss was identified in *Scr*. This may be an assembly error in the genome of *Daphnia pulex* or a lineage specific gain in it since there is also no splice site in *Drosophila melanogaster*. The coordinates of the potential exon-intron junctions in the contigs of *E. verrucosus* can be found in the electronic supplement.

Repeat Content

The most abundant repetitive sequences as determined from the most over-represented 24-mer sequences fell into just 5 classes denoted A to E, see Methods for details. A number of core repeat sequences and repeat contigs was associated with each repeat cluster (see Supplementary Table S3) and all core repeat sequences and repeat contigs can be found in the electronic supplement. Group C accounted for low-complexity sequences and comprises a variety of tandem repeats and other very low entropy sequences. Five of these matched microsatellite motifs also observed in other species. The results of the entropy calculation and tandem repeat search including the repeated motif and the `RepeatMasker` score is listed in Supplementary Table S4. A phylogenetic tree of the core repeat sequences (excluding cluster C) is shown in Figure 5 (see Supplementary Figure S2 for the complete tree). The four groups appeared to be unrelated, with the possible exception of group D and E. A `blastn`-based search against the NCBI and ENSEMBL genomes was conducted but, except for some weak similarities in ENSEMBL for A3 and cluster D (see Supplementary Table S5), did not result in any significant hits (%ID ≥ 70, E-value ≤ 0.001). None of these hyper-abundant sequences thus could be identified as member of a previously described repeat family with `CENSOR-GIRI`.

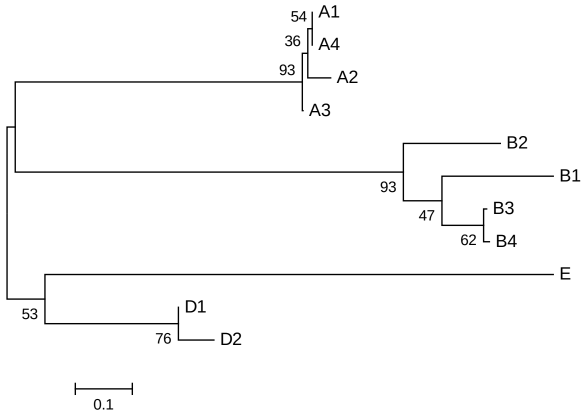

Figure 5: Phylogenetic tree of the core repeat sequences of the most abundant repeat clusters (A, B, D, and E). The tree was calculated and constructed using `MEGA` v.5.10 and the Neighbor-Joining algorithm. The evolutionary distances were computed using the Maximum Composite Likelihood approach (see Methods for details). Bootstrap values after 1000 samples are shown along the edges of the tree.

To estimate the approximate fraction of each repeat cluster in the genome of *E. verrucosus*, the sequencing data was mapped with `segemehl` to all repeat contigs and the number of reads mapping to each cluster was calculated (see Table 2). Depending on the accuracy (90 or 95%) during mapping, the major clusters A and B obtain mappings of 10.58-13.90% and 17.03-18.43% of the sequencing reads, respectively. Moreover, at least 4.40-5.82% of the reads originate from low-complexity sequences (cluster C) and 3.42-4.63% of the reads were mapped to the repeat contigs of cluster E. Overall, 36.83-45.40% of the reads were mapped to these most abundant 5 repeats clusters and hence represented the fraction of these repetitive elements in the genome of *E. verrucosus*.

In addition to the lower bound of the repeat content calculated using the most abundant repeat classes, we conducted another repeat content analysis based on 30-mers, described in the Methods section. The analysis classified the 30-mers into repetitive and non-repetitive 30-mers based on their frequency in the sequencing data. Here, any 30-mer occurring more than 6 times was unlikely to occur uniquely in the genomic sequence and

cluster	# reads mapping to clusters with			
	95% accuracy		90% accuracy	
A	30.66 M	10.58%	40.28 M	13.90%
B	49.37 M	17.03%	53.41 M	18.43%
C	12.74 M	4.40%	16.87 M	5.82%
D	4.07 M	1.40%	7.60 M	2.62%
E	9.90 M	3.42%	13.42 M	4.63%
total	106.74 M	36.83%	131.59 M	45.40%

Table 2: Frequency of reads in the survey sequencing data which can be mapped with `segemehl` and 90% or 95% accuracy to the repeat contigs of each cluster. Based on these statistics, 5 clusters of repetitive elements comprise between 36.83-45.40% of the genome of *E. verrucosus*.

hence denoted as repetitive. In consequence, the repeat content of the sequencing data and hence of the genomic sequence of *E. verrucosus* were estimated to 73.2%. The estimate might still not represent an upper bound on the repeat content since sequencing errors could not be corrected prior to the analysis due to the insufficient coverage and thus may lead to an underestimation of the actual fraction of repetitive sequence.

Mitogenome

The mitogenome was assembled using the "crystallization strategy" outlined in the Methods section. The complete sequence can be found in the electronic supplement and is available under the GenBank accession KF690638. The mitogenome is circular, as demonstrated by the presence of split reads connecting the 3′-end with the 5′-end of our assembly. It has a length of 15315 nt. It contains the expected 13 protein-coding genes, 22 tRNAs, and both ribosomal RNA subunits.

A phylogenetic analysis based on Cytochrome Oxidase Subunit 1 (COI) and 16s rRNA mitochondrial genes revealed the close relationship of *E. verrucosus* to *Gammarus duebeni* when compared to the other gammarids for which the complete mitogenome is available. This is in agreement with the fact that *E. verrucosus* and *G. duebeni* belong to the same Superfamily *Gammaroidea*, according to NCBI Taxonomy Browser (NCBI Resource Coordinators, 2013). Phylogenetic trees are shown in Supplementary Figure S3.

The gene order, Figure 6, is identical to that of *G. duebeni* which supports once more the close relationship of both species. The ribosomal RNAs are truncated and match previous descriptions for *G. duebeni* (Krebes and Bastrop, 2012). The mitogenome contains two intergenic spacers with a length of ∼250 bp. One is located between NAD2 and tRNA$^{\mathrm{Trp}}$, and another between tRNA$^{\mathrm{Ser2}}$ and NAD1. Similar large intergenic spacers have been described also in crustacean mitogenomes, for example, a relatively large spacer (177 bp) was found between srRNA and tRNA$^{\mathrm{Gly1}}$ in the *Upogebia major* mtDNA (Lin *et al.*, 2012).

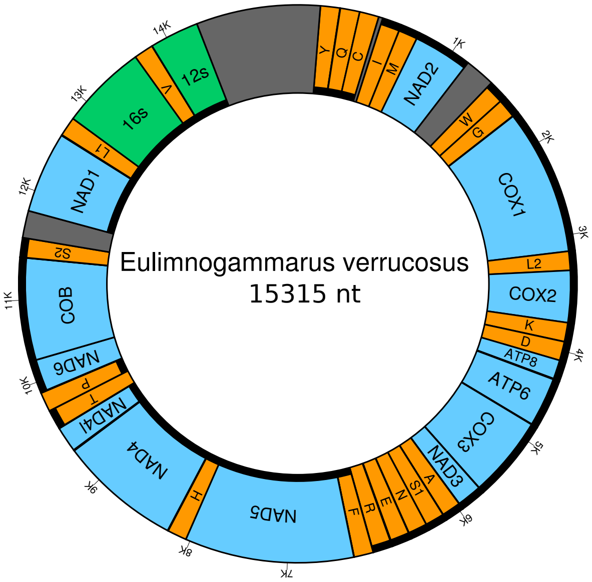

Figure 6: Map of the mitochondrial genome of *Eulimnogammarus verrucosus*. The 22 mt-tRNAs genes are highlighted in orange, both ribosomal RNAs in green, and the 13 protein-coding genes in blue. The control region and intergenic spacers are shown in gray.

The mitogenome appeared in the sequencing data with a coverage of approximately 1000 X. As observed also for other mitogenomes, the coverage in the control region was higher due to its lower complexity and high AT-content of 79.8% which is similar to other crustacean mitogenomes (Ki *et al.*, 2010). Tandem repeats within the control region have been observed previously in the amphipod *G. duebeni* (Krebes and Bastrop, 2012) and in the isopod *Ligia oceanica* (Kilpert and Podsiadlowski, 2006).

Surprisingly, apart from the mitogenome of *E. verrucosus*, we found a second, minor portion of different contigs with a coverage of only 10-50 X in our survey sequencing data, which also showed some similarity to the mitogenome. Blasting these distinct contigs against the NCBI Nucleotide collection (nt) (Altschul *et al.*, 1990) resulted in a perfect hit for a fragment of mitochondria ribosomal RNA of *Eulimnogammarus vittatus*, a related species also endemic to Lake Baikal and occurring in a similar habitat. As we found nearly a full sequence coverage for a second distinct mitogenome, we interpreted these reads as a contamination. It was introduced possibly as food (Gee, 2003) since the *E. verrucosus* specimen was not starved before DNA preparation. From the ratio of the coverage of the mitogenome (1000 X for *E. verrucosus* and 10-50 X for the other), we obtained an upper bound for the contamination by the congener species of 5%. Precautions were taken to avoid reads originating from contaminations to be included in the seed-based microRNA and Hox gene contigs (see Methods section on Hox genes). Care will be taken in future research to avoid such impurities, since they are much harder to identify by computational means than contamination with human or bacterial DNA (Longo *et al.*, 2011). We also assessed the contamination due to microorganisms in the sample, observing only a negligible fraction ($< 0.1\%$) of reads clearly originating from microorganisms. The overwhelming majority of sequences that match low-complexity regions thus also cannot stem from contaminating microorganism.

Attempts to de novo *assembly*

We tested two different popular genome assembly tools with different parameters and various strategies to filter highly repetitive reads. We were unable to reach N50 values exceeding 150 (compared to the read length of 101) and the number of contigs longer than 1000 bp remained very small. Details on representative assemblies can be found in the Supplementary Table S2. Even in the light of the low 3 X coverage of the data, the results were not satisfactory and did not reach the quality of the seed and crystallization based approaches that were used to obtain the contigs of Hox genes, miRNAs, and the full mitochondrial genome. However, no fully automatic pipeline of the cystalization approach was available yet, resulting in lot of manual work.

DISCUSSION

Our survey of the *E. verrucosus* genome presents the first large-scale investigation in the genomics of a species endemic to Lake Baikal. With a size of about 10 G, the *E. verrucosus* genome appears to be much more typical for crustaceans than the compact genome of *Daphnia pulex*, so far the only fully sequenced crustacean. Comprehensive genomic resources are of utmost importance for ecotoxicological and ecophysiological studies in an evolutionary context. In order to estimate, for instance, the effects of drastic environmental changes at Lake Baikal on the endemic species due to global warming, it is crucial to comprehensively investigate their stress response systems in comparison to that of their palaearctic relatives.

The large size of the genome together with the dominating contribution of just a few families of repetitive elements, however, poses a series of difficult methodological and technical problems. Not surprisingly, for instance, the attempts of *de novo* assembly of the survey sequencing data did not yield any useful results since the immense repeat content generated a long and very heavy tail in the *k*-mer distribution that made the usual strategies for data preprocessing inapplicable. Even after removing the repetitive portion of the sequencing data, any assembly attempt did not succeed due to fragmentation by removing the highly abundant repeats in conjunction with the low sequencing coverage. The next step, for which we are currently sequencing the genome to approximately 40 X, thus, consists in the development of efficient means for removing repetitive sequences prior to contig assembly. The extreme repeat content also renders mate pair sequencing inefficient as the majority of the mate pairs will have at least one end in the repetitive regions, making them less informative. In this case, expressed mRNAs of additional transcriptome sequencing data may provide long-range scaffolding information (outside repetitive regions) that can complement mate pair information. Although it appears infeasible to obtain a finished assembly for a genome of the size and repeat content of *E. verrucosus*, at least with the moderate resources available to most institutions, it is of utmost importance to gain insight into such very large genomes that are typical for many invertebrate groups. The restriction to model organisms with exceptionally small genomes, which is still the norm due to technical and methodological considerations, is likely to paint a quite distorted picture of genome evolution. A hint that very large genomes may exhibit organizational differences comes, for example, from the intriguing observation that the ∼ 13 Gb genome of the grasshopper *Chortippus parallelus* shows DNA methylation levels otherwise seen only for vertebrates (Lechner *et al.*, 2013).

The unassembled survey data presented here, nevertheless, conveyed interesting information even at low coverage. Beyond an estimate of the genome size and the analysis of the repeat content, it was possible to accomplish the annotation of 33 microRNA genes, the investigation of the highly conserved part of 9 Hox genes including conservation of intron-exon junctions by comparison with *Daphnia pulex*, and the assembly of the complete mitochondrial genome of *E. verrucosus*.

Acknowledgment

We thank Olga Zatsepina for consultations and Vasiliy Pavlichenko for the photograph shown as Figure 1. This work was performed within the *LabEglo* project HRJRG-221, funded by the bilateral funding programme "Helmholtz-Russia Joint Research Groups" (HRJRG) from the Helmholtz Association and the Russian Foundation for Basic Research (RFBR). It was further supported in part by the German BMBF (PTJ grant HNPCCSys 031 6065A), by the Helmholtz Impulse and Networking Fund through Helmholtz Interdisciplinary Graduate School for Environmental Research (HIGRADE), and by LIFE. Russian participants were also supported by the Russian Foundation for Basic Research (grants 11-04-91321-SIG_a, 11-04-00174-a, 12-04-00810-a) and President's of Russia grants (MD-2063.2012.4, MK-5466.2012.4). LIFE – Leipzig Research Center for Civilization Diseases, Universität Leipzig – is funded by means of the European Union, by the European Regional Development Fund (ERDF) and by means of the Free State of Saxony within the framework of the excellence initiative. AHS was supported by a grant for the Lebanese University from the AZM and Saade Association.

Supplemental Material consisting of additional information on methods, tables, and figures is available as a PDF document. Machine readable data files can be found at http://www.bioinf.uni-leipzig.de/publications/supplements/13-003/.

Conflict of interest

The authors declare no conflict of interest.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. J Mol Biol 215:403–10.

Bauza-Ribot MM, Jaume D, Juan C, Pons J, 2009. The complete mitochondrial genome of the subterranean crustacean *Metacrangonyx longipes* (Amphipoda): a unique gene order and extremely short control region. Mitochondrial DNA 20:88–99.

Bazikalova AY, 1945. The amphipods of Lake Baikal. Proc Baikal Limnological Station 11:1–440.

Bedulina DS, Evgen'ev MB, Timofeyev MA, Protopopova MV, Garbuz DG, Pavlichenko VV, Luckenbach T, Shatilina ZM, Axenov-Gribanov DV, Gurkov AN, Sokolova IM, Zatsepina OG, 2013. Expression patterns and organization of the hsp70 genes correlate with thermotolerance in two congener endemic amphipod species (*Eulimnogammarus cyaneus* and *E. verrucosus*) from Lake Baikal. Mol Ecol 22:1416–30.

Belzile C, Rees DJ, Dufresne F, 2007. Amphipod genome sizes : first estimates for Arctic species reveal genomic giants. Genome 158:151–158.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman D, Ostell J, Sayers EW, 2013. GenBank. Nucleic Acids Res 41:D36–D42.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF, 2012. MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol 69:313–319.

Blythe MJ, Malla S, Everall R, Shih YH, Lemay V, Moreton J, Wilson R, Aboobaker AA, 2012. High through-put sequencing of the *Parhyale hawaiensis* mRNAs and microRNAs to aid comparative developmental studies. PLoS One 7:e33784.

Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D, 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proc Natl Acad Sci U S A 108:15920–15924.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Cáceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Fröhlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kültz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Andrews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL, 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331:555–561.

Cook CE, Yue Q, Akam M, 2005. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. Proceedings of the Royal Society B Biological Sciences 272:1295–1304.

Crow KD, Stadler PF, Lynch VJ, Amemiya CT, Wagner GP, 2006. The fish specific Hox cluster duplication is coincident with the origin of teleosts. Mol Biol Evol 23:121–136.

Deutsch J, Mouchel-Vielh E, 2003. Hox genes and the crustacean body plan. Bioessays 25:878–87.

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A, 2011. Rfam: Wikipedia, clans and the ”decimal” release. Nucleic Acids Res 39:D141–D145.

Gee H, 2003. Zoology: You aren't what you eat. Nature 424:885–886.

Gerstfeldt G, 1858. Über einige zum Theil neue Arten Platoden, Anneliden, Myriapoden und Crustaceen Sibiriens: namentlich seines östlichen Theiles und des Amur-Gebietes. Buchdruckerei der Kaiserlichen Akademie der Wissenschaften.

Gonsalves SE, Moses AM, Razak Z, Robert F, Westwood JT, 2011. Whole-genome analysis reveals that active heat shock factor binding sites are mostly associated with non-heat shock genes in *Drosophila melanogaster*. PLoS ONE 6:e15934.

Gregory T, 2012. Animal genome size database. `http://www.genomesize.com`.

Grenier JK, Garber TL, Warren R, Whitington PM, Whitington S, 1997. Evolution of the entire arthropod Hox gene set predated the origin and radiation of the onychophoran/arthropod clade. Current Biology 7:547–553.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, The Students of Bioinformatics Computer

Labs 2004 and 2005, 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics 7:15.

Hoffmann S, Otto C, Kurtz S, Sharma C, Khaitovich P, Vogel J, Stadler PF, Hackermüller J, 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5:e1000502.

Huang T, Xu D, Zhang X, 2012. Characterization of host microRNAs that respond to DNA virus infection in a crustacean. BMC Genomics 13:159.

Jühling F, Pütz J, Bernt M, Donath A, Middendorf M, Florentz C, Stadler PF, 2012. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Res 40:2833–45.

Ki JS, Hop H, Kim SJ, Kim IC, Park HG, Lee JS, 2010. Complete mitochondrial genome sequence of the Arctic gammarid, *Onisimus nanseni* (Crustacea; Amphipoda): Novel gene structures and unusual control region features. Comp Biochem Physiol D Genomics Proteomics 5:105–115.

Kilpert F, Podsiadlowski L, 2006. The complete mitochondrial genome of the common sea slater, *Ligia oceanica* (Crustacea, Isopoda) bears a novel gene order and unusual control region features. BMC Genomics 7:241.

Kohany O, Gentles AJ, Hankus L, Jurka J, 2006. Annotation, submission and screening of repetitive elements in repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.

Kozhov M, 1963. Lake Baikal and its life, vol. XI of *Monographia Biologicae*. The Hague: W. Junk.

Kozhova OM, Izmest'eva LR, 1998. Lake Baikal – Evolution and Biodiversity. Leiden: Backhuys.

Kravtsova LS, Kamaltynov RM, Karabanov EB, Mekhanikova IV, Sitnikova TY, Rozhkova NA, Slugina ZV, Izhboldina LA, Weinberg IV, Akinshina TV, Yu SD, 2004. Macrozoobenthic communities of underwater landscapes in the shallow-water zone of southern Lake Baikal. Hydrobiologia 522:193–205.

Krebes L, Bastrop R, 2012. The mitogenome of *Gammarus duebeni* (Crustacea Amphipoda): A new gene order and non-neutral sequenceevolution of tandem repeats in the control region. Comp Biochem Physiol D Genomics Proteomics 7:201–211.

Lander E, Waterman M, 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2:231–9.

Langmead B, Salzberg SL, 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

Lechner M, Marz M, Stadler PF, Krauss V, 2013. Genome size, methylation rate, and CpG depletion in metazoans. Th Biosci 132:47–60.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder O, Leung F, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner C, Lam T, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford M, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam T, Yiu S, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong G, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J, 2010a. The sequence and de novo assembly of the giant panda genome. Nature 463:311–7.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J, 2010b. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–72.

Libertini A, Trisolini R, Eriksson-Wiklund AK, 2003. A preliminary survey on genome size in Amphipoda. XIth International Colloquium on Amphipoda, Tunis, Tunisia.

Lin F, Liu Y, Sha Z, Tsang L, Chu K, Chan T, Liu R, Cui Z, 2012. Evolution and phylogeny of the mud shrimps (Crustacea: Decapoda) revealed from complete mitochondrial genomes. BMC Genomics 13:631.

Longo MS, O'Neill MJ, O'Neill RJ, 2011. Abundant human DNA contamination identified in non-primate genome databases. PLoS ONE 6:e16410.

Marçais G, Kingsford C, 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27:764–70.

Nawrocki EP, Kolbe DL, Eddy SR, 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335–7.

NCBI Resource Coordinators, 2013. Database resources of the national center for biotechnology information. Nucleic Acids Res 41:D8–D20.

Parchem RJ, Poulin F, Stuart AB, Amemiya CT, Patel NH, 2010. Bac library for the amphipod crustacean, (*Parhyale hawaiensis*). Genomics 95:261–267.

Richard G, Kerrest A, Dujon B, 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686–727.

Ruddle FH, Amemiya CT, Carr JL, Kim CB, Ledje C, Shashikant CS, Wagner GP, 1999. Evolution of chordate hox gene clusters. Ann N Y Acad Sci 870:238–248.

Rusinek OT, editor, 2012a. Baicalogy, vol. 1. Novosibirsk: Nauka.

Rusinek OT, editor, 2012b. Baicalogy, vol. 2. Novosibirsk: Nauka.

Saitou N, Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–25.

Salemaa H, Kamaltynov R, 1994a. 7.4 Chromosomal relationships of the endemic Amphipoda (Crustacea) in the ancient lakes Ohrid and Baikal. Genetics and evolution of aquatic organisms (p. 405).

Salemaa H, Kamaltynov R, 1994b. The chromosome numbers of endemic Amphipoda and Isopoda - an evolutionary paradox in the ancient lakes Ohrid and Baikal. Ergebnisse der Limnologie 44:247–256.

Schwager EE, Schoppmeier M, Pechmann M, Damen WG, 2007. Duplicated Hox genes in the spider *Cupiennius salei*. Front Zool 4.

Shatilina ZM, Wolfgang Riss H, Protopopova MV, Trippe M, Meyer EI, Pavlichenko VV, Bedulina DS, Axenov-Gribanov DV, Timofeyev MA, 2011. The role of the heat shock proteins (HSP70 and sHSP) in the thermotolerance of freshwater amphipods from contrasting habitats. Journal of Thermal Biology 36:142–149.

Shin SC, Cho J, Lee JK, Ahn DH, Lee H, H P, 2012. Complete mitochondrial genome of the Antarctic amphipod Gondogeneia antarctica (Crustacea, amphipod). Mitochondrial DNA 23:25–27.

Simpson JT, Durbin R, 2012. Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22:549–56.

Smit AFA, Hubley R, Green P, 2013. Unpublished data. Current version open-4.0.1 (RMLib: 20120418 & Dfam: 1.1).

Stauber M, Jäckle H, Schmidt-Ott U, 1999. The anterior determinant *bicoid* of *Drosophila* is a derived *Hox* class 3 gene. Proc Natl Acad Sci USA 96:3786–3789.

Tamura K, Nei M, Kumar S, 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101:11030–5.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S, 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–9.

Tanzer A, Riester M, Hertel J, Bermudez-Santana CI, Gorodkin J, Hofacker IL, Stadler PF, 2010. Evolutionary genomics of microRNAs and their relatives. In: Caetano-Anolles G, editor, Evolutionary Genomics and Systems Biology, (pp. 295–327). Hoboken, NJ: Wiley-Blackwell.

Thompson JD, Higgins DG, Gibson TJ, 1994. CLUSTAL w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–80.

Timofeyev M, Shatilina Z, 2007. Different preference reactions of three Lake Baikal endemic amphipods to temperature and oxygen are correlated with symbiotic life. Crustaceana 80:129–138.

Timofeyev MA, 2010. Ecological and physiological aspects of adaptation to abiotic environmental factors in endemic Baikal and Palearctic amphipods. Tomsk: Tomsk State University.

Timoshkin OA, 2001. Lake Baikal: fauna diversity, problems of its “immiscibility” and origin, ecology and “exotic” communities. In: Timoshkin OA, editor, Index of animal species inhabiting Lake Baikal and its catchment area, (pp. 16–73). Novosibirsk: Nauka Publishers.

UniProt Consortium, 2013. Update on activities at the universal protein resource (uniprot) in 2013. Nucleic Acids Res 41:D43–7.

Vergilino R, Dionne K, Nozais C, Dufresne F, Belzile C, 2012. Genome size differences in *Hyalella cryptic* species. Genome 55:134–139.

Wheeler BM, Heimberg AM, Moy VN, Sperling EA, Holstein TW, Heber S, Peterson KJ, 2009. The deep evolution of metazoan microRNAs. Evol Dev 11:50–68.

Yeo G, Burge C, 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol 11:377–94.

Zeng V, Villanueva KE, Ewen-Campen BS, Alwes F, Browne WE, Extavour CG, 2011. *De novo* assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean *Parhyale hawaiensis*. BMC Genomics 12:581.

Zerbino D, Birney E, 2008. Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res 18:821–9.