Supplementary information: A First Glimpse at the Genome of the Baikalian Amphipod *Eulimnogammarus verrucosus*

Lorena Rivarola-Duarte[1], Christian Otto[1], Frank Jühling, Stephan Schreiber, Daria Bedulina, Lena Jakob, Anton Gurkov, Denis Axenov-Gribanov, Abdullah H. Sahyoun, Magnus Lucassen, Jörg Hackermüller, Steve Hoffmann, Franz Sartoris, Hans-Otto Pörtner*, Maxim Timofeyev*, Till Luckenbach*, Peter F. Stadler*

Supplementary Methods

DNA Isolation

Animals were placed in aquariums containing Baikal water at 6° C and sorted. They were washed once in ethanol in order to prevent contamination with epibioints and submerged in 800 μl of ethanol each. The tubes were kept in liquid nitrogen until the transportation to Germany.

One complete ethanol-preserved individual of *E. verrucosus*, was washed twice in molecular biology grade water to remove the ethanol. Total genomic DNA was extracted from the preserved material following a modified version of the DTAB-CTAB protocol [1]. The homogenization buffer containing 6% dodecyltrimethylammonium bromide (DTAB), 1.125 M NaCl, 75 mM TrisHCl and 37.5 mM EDTA at pH 8.0 was heated at 65° C and 500 μl were added to the specimen. The sample was homogenized with a MixerMill MM 400 for 5 min using two stainless steel beads, whose diameters were 3 mm and 5 mm. The homogenate was then centrifuged at maximum speed at 4° C for 5 min to eliminate the foam and a further 500 μl of homogenization buffer were added, followed by 30 min of incubation at 65° C. A digestion step using 50 μl of proteinase K (10 mg/ml) was performed overnight at 55° C. Subsequently, the homogenate was incubated at 70° C for 3 min to inactivate the enzyme. Ten microliters of RNase A (10 mg/ml) were added, followed by an incubation step at 37° C for 30 min. The DNA was extracted using one volume of chloroform/isoamyalcohol (24:1) vortexed for 20 s and centrifuged for 5 min at maximum speed in an Eppendorf table centrifuge. The aqueous phase was transferred to a new tube, mixed with 100 μl of 4 M LiCl and 400 μl of isopropanol by vortexing and kept overnight at −20° C. The DNA was precipitated by centrifugation at 4° C for 20 min at 12000 g and washed twice with 1 ml of 70% ethanol followed by centrifugation at 4°C for 10 min at 14000 g. The supernatant was decanted and the DNA pellet was dried for less than 5 min at 50° C and re-suspended in 50 μl of molecular biology grade water. Quantification and quality control of the DNA was conducted using a NanoDrop ND1000 (see Supplementary Tab. S1 for details about the chemicals and devices used).

Sequencing

Two technical replicates dual-indexed libraries were prepared using Nextera™ DNA sample preparation kit following the manufacturer's protocol. DNA Clean & Concentrator™-5 was used for clean-up of the tagmentated DNA and Agencourt® AMPure® XP kit for the last PCR clean-up step. The two technical replicates were pooled and then size-selected by electrophoresis using a 2% agarose gel and visualized by the addition of ethidium bromide. To obtain a length range of 150-600 bp, DNA was extracted and purified using MinElute Gel Extraction Kit. Agilent High

[1] Joint First Authors

*Corresponding authors

Sensitivity DNA Kit in an Agilent 2100 Bioanalyzer machine was used for quantification and quality control of the libraries in all steps. Clusters were generated on a cBot device using TruSeq PE Cluster Kit v3-cBot-HS. In the lane with the sample, the cluster density was 675000/mm^2. A multiplexed paired-end dual-index sequencing run with a number of cycles of 101-9-(7)-9-101 was conducted in an Illumina HiSeq 2000 instrument employing PhiX as quality and calibration control, TruSeq Dual Index Sequencing Primer Box Paired End and TruSeq SBS Kit v3-HS (200 Cycles + 50 cycles). Base calling (Bcl) conversion and demultiplexing was conducted using `CASAVA` (version 1.8.2), allowing one mismatch in the index sequence and keeping only the reads passing the quality filter. The overall yield for this lane was 35.6 Gigabases and 353 million reads, a mean quality score of 36.4 and 94% of the bases with a quality value ≥ 30 (see Supplementary Tab. S1 for details about the chemicals and devices used).

Data Preprocessing

After sequencing, 3'-adapter contaminations were removed using `cutadapt` v.1.1 [2] with option `-e` 0.15. Clipped reads shorter than 15 nt were discarded. We used `FLASH` v.1.0.3 [3] to merge paired-end reads that were generated from DNA fragments shorter than twice the read length. To prevent spurious mergings, an overlap of ≥ 30 nt with a mismatch ratio ≤ 0.15 was required. In addition, merging was prohibited in cases where the overlap comprises repetitive segments, i.e., an 24-mer occurring more than 100 times in the total clipped data. This prevents paired-end reads from being joined that overlap similar but genomically distant reoccurring motifs. Afterwards, we used a custom in-house script to clean up the merged data. It is based upon the characteristics of paired-end sequencing, where either 3' adapter sequences are not sequenced at all (i.e. if reads are generated from DNA fragments longer than the read length) or the length of 3' adapters contaminations is expected to be the same in both mates. In the latter case, after removing adapter sequences, both mate sequences are expected to be reverse complementary to each other. Hence the cleaning procedure can detect both cases of false positive and false negative clippings as well as illegitimate mergings and resolve them appropriately.

Supplementary Figures

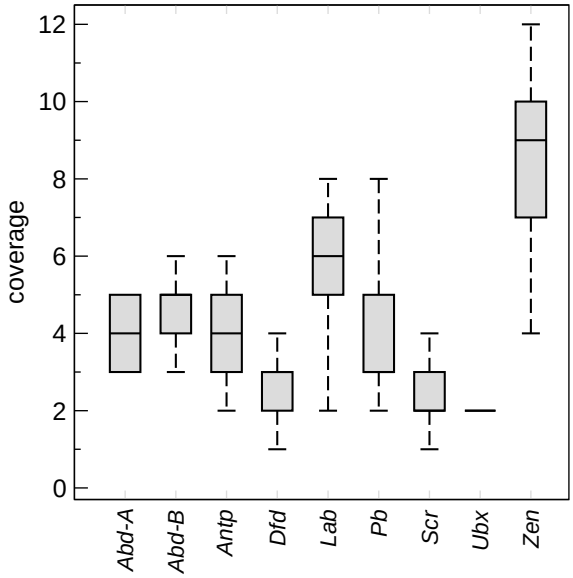


Figure S1: Box-whisker plots of the position-wise coverage over the part of the homeobox present on the Hox gene contigs of *E. verrucosus*.

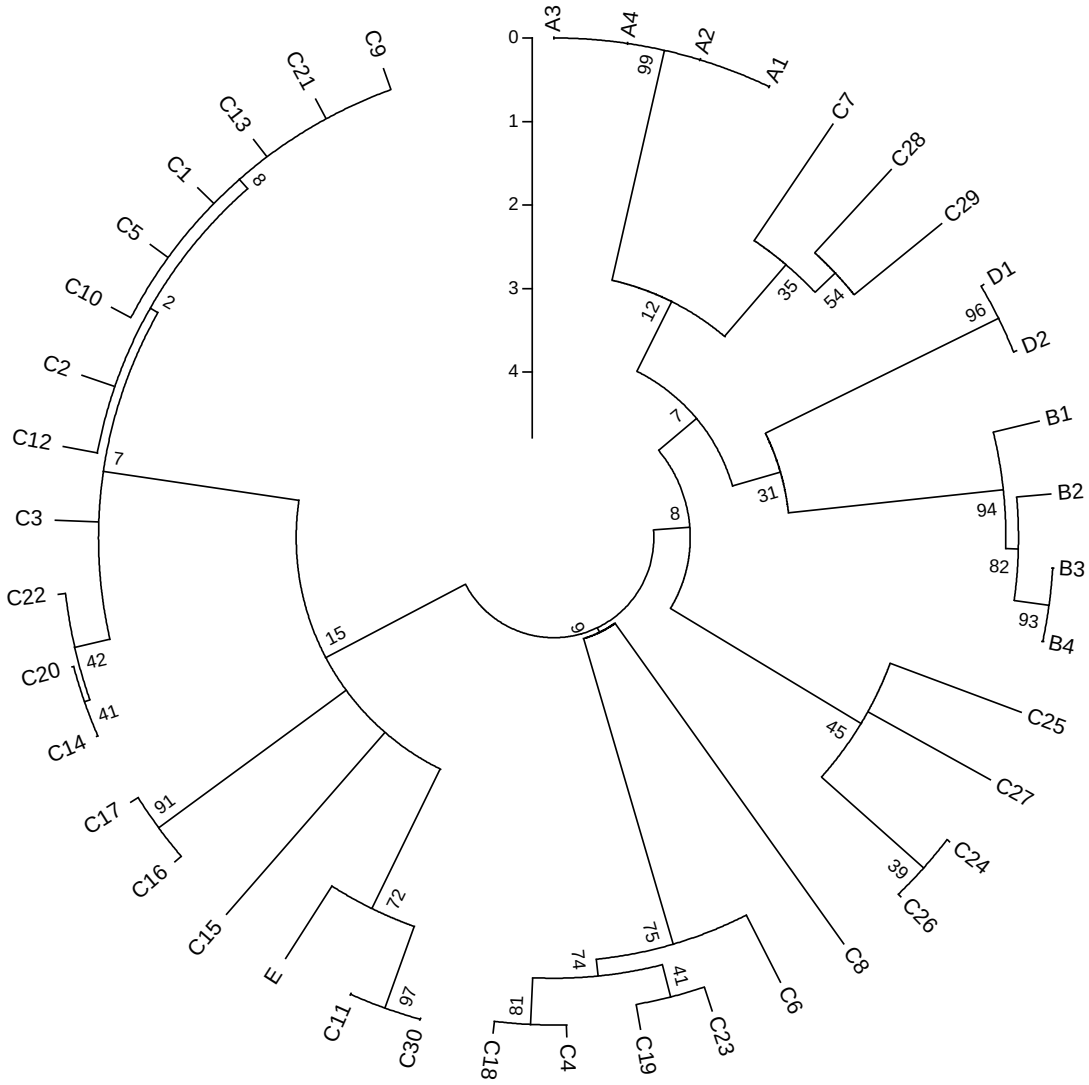
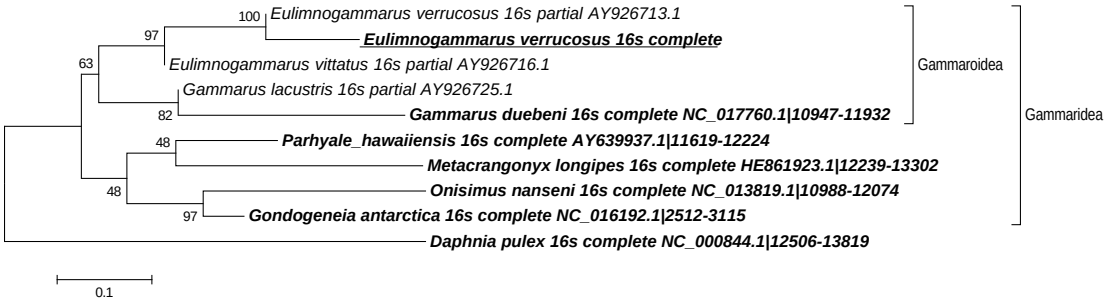
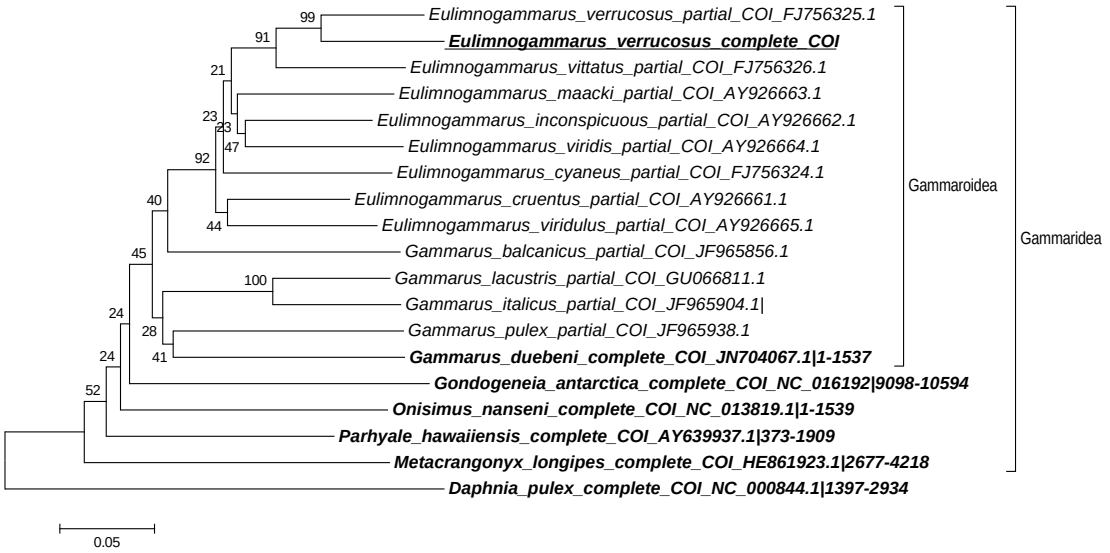


Figure S2: Phylogenetic tree of the core repeat sequences of the most abundant repeat clusters (A, B, C, D, and E). The tree was calculated and constructed using MEGA v.5.10 and the Neighbor-Joining algorithm. The evolutionary distances were computed using the Maximum Composite Likelihood approach (see Methods for details). Bootstrap values after 1000 samples are shown along the edges of the tree.



(a)



(b)

Figure S3: **Phylogenetic tree based on (a) 16s rRNA and (b) Cytochrome Oxidase Subunit 1 (COI) mitochondrial DNA sequences.** Taxons inside the cluster *Gammaroidea* belong to this Superfamily, and those inside the cluster *Gammaridea* to this Suborder under the *Amphipoda* Order and *Malacostraca* Class. *Daphnia pulex* works as out group, belonging to the *Branchiopoda* Class inside the *Crustacea* Subphylum according to NCBI Taxonomy Browser. Names in bold letters imply complete sequences for the genes. The underlined sequence is the one derived from our analysis. Numeric identifiers correspond to those in GenBank. Tree calculated using `MEGA` v.5.10 and the Neighbor-Joining algorithm. 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.

Supplementary Tables

Table S1: **List of chemicals and devices used.** This table comprises information on the companies and factory locations of the chemicals and devices used for DNA isolation and sequencing.

chemicals/device	company
Molecular biology grade water	AppliChem, Darmstadt, Germany
Dodecyltrimethylammonium bromide (DTAB)	AppliChem, Darmstadt, Germany
NaCl	VWR International GmbH, Darmstadt, Germany
TrisHCl	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
EDTA	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
MixerMill MM 400	Retsch GmbH, Haan, Germany
Stainless steel beads 3 mm	Retsch GmbH, Haan, Germany
Stainless steel beads 5 mm	QIAGEN GmbH, Hilden, Germany
Proteinase K	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
RNase A	AppliChem, Darmstadt, Germany
Chloroform	VWR International GmbH, Darmstadt, Germany
Isoamyalcohol	Sigma Aldrich, Steinheim, Germany
LiCl	Merck KGaA, Darmstadt, Germany
Isopropanol	VWR International GmbH, Darmstadt, Germany
Ethanol	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
NanoDrop ND1000	PeqLab Biotechnologie GmbH, Erlangen, Germany
Nextera™ DNA sample preparation kit	Illumina, Inc., San Diego, U.S.A
DNA Clean & Concentrator™-5	Zymo Research Corporation, Irvine, U.S.A
Agencourt® AMPure® XP kit	Beckman Coulter GmbH, Krefeld, Germany
Agarose	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
Ethidium bromide	Carl Roth GmbH + Co. KG, Karlsruhe, Germany
MinElute Gel Extraction Kit	QIAGEN GmbH, Hilden, Germany
Agilent High Sensitivity DNA Kit	Agilent Technologies, Santa Clara, U.S.A
Agilent 2100 Bioanalyzer machine	Agilent Technologies, Santa Clara, U.S.A
cBot device	Illumina, Inc., San Diego, U.S.A
TruSeq PE Cluster Kit v3-cBot-HS	Illumina, Inc., San Diego, U.S.A
Illumina HiSeq 2000	Illumina, Inc., San Diego, U.S.A
PhiX Control v3 kit	Illumina, Inc., San Diego, U.S.A
TruSeq Dual Index Sequencing Primer Box Paired End	Illumina, Inc., San Diego, U.S.A
TruSeq SBS Kit v3-HS (200 Cycles + 50 cycles)	Illumina, Inc., San Diego, U.S.A

Table S2: ***De novo* assembly statistics.** The assembly attempts were done using `SOAPdenovo` and `Velvet` with k-mer sizes of 23 and 31 nt on two different datasets. In dataset 1, reads with ambiguous bases (e.g. Ns) and reads clearly originated from mitogenome were filtered (see Methods for details). In dataset 2, addionally, reads containing repetitive subsequences were discarded. Overall, the assemblies were not satisfactory due to the very low N50 values and the small number of contigs larger than 1 kb.

			# contigs	
	k-mer	N50 (bp)	> 100 bp	> 1 kb
Dataset 1[1]				
`SOAPdenovo`	31	152	10 376 298	5445
	23	160	7 956 106	4671
Dataset 2				
`SOAPdenovo`	31	128	8 545 904	244
	23	149	7 257 495	424
`Velvet`[2]	31	145	1 052 758	171
	23	117	931 335	283

[1] `Velvet` crashed during the assembly of dataset 1 due to insufficient memory on the 800 GB-RAM machine.
[2] N50 and number of contigs > 100 bp of the `Velvet` assembly are biased since it reports only contigs larger than 200 bp.

Table S3: **Number of core repeat sequences and repeat contigs by cluster.** Cluster A and B are comprised of several repeat contigs with main differences on the margins but similar core repeat sequences. Due to the collection of low-complexity or tandem repeat-like sequences in cluster C, it contained a large number of core repeat sequences (see Methods for details).

cluster	core repeat sequences	repeat contigs
A	4	44
B	4	12
C	30	34
D	2	4
E	1	2
total	41	96

Table S4: **Results of the entropy calculation and tandem repeat search of the core repeat sequences.** The table summarized the relative entropies (or Kullback-Leibler divergence) of the core repeat sequences and the best hits given by `RepeatMasker` tandem repeat search. Dashes indicate core sequences w/o `RepeatMasker` hits. Core sequences of cluster C are deemed low-complexity due to high entropy values (> 0.7) and high-scoring `RepeatMasker` hits for most of them. Core sequences marked with an asterisk match microsatellite patterns that also have been observed in other species.

cluster	core repeat sequence	`RepeatMasker` pattern	`RepeatMasker` max score	rel. entropy
A	A1	-	-	0.31
	A2	-	-	0.285
	A3	-	-	0.263
	A4	-	-	0.249
B	B1	-	-	0.055
	B2	-	-	0.069
	B3	-	-	0.047
	B4	-	-	0.023
C	C1	(TACAGACA)n	160	1.229
	C2	(ATACGGAC)n	161	1.13
	C3	(G)n	116	2.349
	C4*	(ACGATG)n	116	1.597
	C5	(GACATGCA)n	116	1.103
	C6	(AACATT)n	116	1.191
	C7	(CCCACA)n	152	2.335
	C8	(GACTTA)n	116	1.192
	C9	(GGACACACGT)n	115	1.187
	C10	(ACACACGG)n	116	1.836
	C11	(ACAAAGTC)n	114	1.176
	C12	(CATGCGGA)n	116	0.789
	C13	(TACA)n	142	1.82
	C14	(CGAA)n	116	1.962
	C15	(A)n	80	3.331
	C16*	(GGAT)n	117	1.727
	C17	(TGAA)n	118	1.777
	C18	(ATG)n	116	1.965
	C19	(GAC)n	118	2.398
	C20	(GAGCGAAC)n	117	1.642
	C21*	(CA)n	606	2.255
	C22	(CGTGCAAA)n	178	0.987
	C23*	(CAA)n	352	1.845
	C24	(TTAAAATATG)n	146	1.341
	C25	(TCTTTAAACA)n	185	0.924
	C26*	(CTTAAAATAT)n	115	1.376
	C27	(CTAAC)n	114	1.149
	C28	-	-	0.797
	C29	-	-	1.522
	C30	-	-	1.17
D	D1	-	-	0.208
	D2	-	-	0.193
E	E	-	-	0.136

Table S5: **Results of blast-based comparison of core repeats against the ENSEMBL genomes.** Hits obtained with E-value ≤ 0.001 and %ID ≥ 70 using the web-based `blastn` comparison of the core repeat sequences against the ENSEMBL genomes. The table summarizes all relevant alignment information including bit score, E-value, and %ID. The search sensitivity was set to 'allow some local mismatches' and any repeat-based masking or filtering was disabled. Core repeat sequences w/o significant hits are omitted.

core repeat sequence	subject species	subject name	subject start	subject end	query start	query end	query length	alignment bit score	alignment E-value	alignment %ID
A3	*Mus musculus*	Chr:8	5169575	5169597	1	23	214	22.2	0.00028	86.96
A3	*Pongo abelii*	Chr:11	11427835	11427871	173	207	214	23.1	0.00026	78.95
A3	*Rattus norvegicus*	Chr:6	55576670	55576725	159	214	214	27.0	0.00089	71.67
A3	*Rattus norvegicus*	Chr:6	70199936	70199955	175	194	214	23.1	0.00089	95.00
D1	*Petromyzon marinus*	GL484935	1984	2112	23	147	168	53.4	3.00E-07	72.39
D1	*Petromyzon marinus*	GL484935	2132	2260	23	147	168	51.4	1.10E-06	71.85
D1	*Petromyzon marinus*	GL484935	1466	1594	23	147	168	49.5	4.40E-06	71.11
D1	*Petromyzon marinus*	GL484935	1836	1964	23	147	168	47.5	1.70E-05	70.37
D1	*Petromyzon marinus*	GL484935	1285	1409	23	144	168	46.6	3.40E-05	70.23
D2	*Petromyzon marinus*	GL484935	2058	2149	19	107	121	41.7	0.00099	73.96
D2	*Petromyzon marinus*	GL484935	1947	2038	19	107	121	41.7	0.00099	73.96

References

[1] S Gustincich, G Manfioleti, G Del Sal, C Schneider, and P A Carninci. Fast method for high-quality genomic DNA extraction from whole human blood. *Biotechniques*, 11:298–300, 1991.

[2] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

[3] T Magoč and S L Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27:2957–63, Nov 2011.