nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-018-0467-9

OPEN

In the format provided by the authors and unedited.

Clonal genome evolution and rapid invasive spread of the marbled crayfish

Julian Gutekunst[1], Ranja Andriantsoa[1], Cassandra Falckenhayn[1], Katharina Hanna[1], Wolfgang Stein[2], Jeanne Rasamy[3] and Frank Lyko[1]*

[1]Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center (DKFZ), Heidelberg, Germany. [2]School of Biological Sciences, Illinois State University, Normal, IL, USA. [3]Mention Zoologie et Biodiversité Animale, Université d'Antananarivo, Antananarivo, Madagascar.
*e-mail: f.lyko@dkfz.de

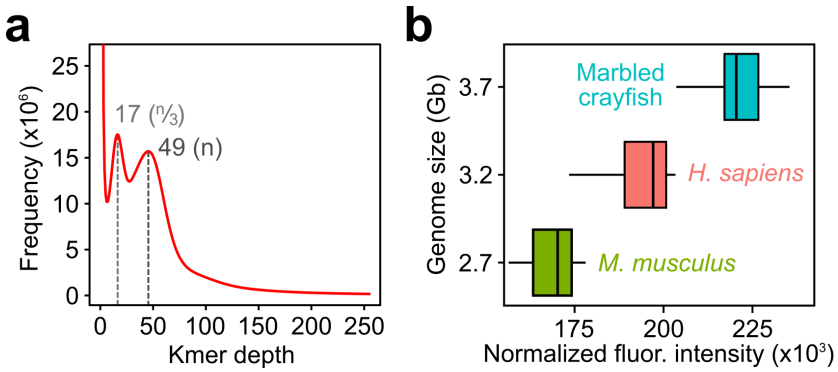

Figure S1. Genome size estimation of marbled crayfish. **a**, K-mer frequency (K=17) plot for *in silico* estimation of the genome size based on raw sequencing data. Peak depths are at 17x and 49x. **b**, Estimated genome sizes based on cell intensities measured by multiple flow cytometry experiments. Fluorescence intensities are shown after normalization to 1N genome equivalents.

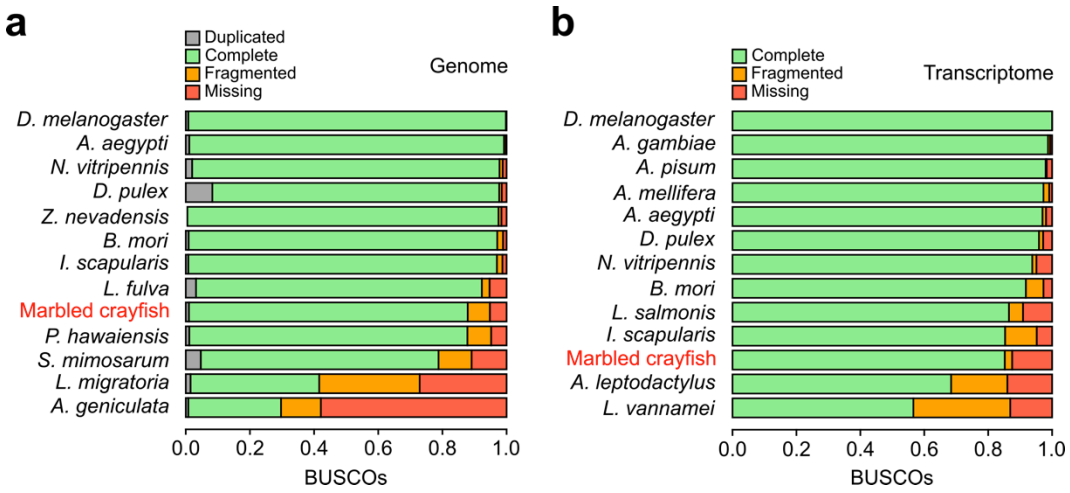

Figure S2. Benchmarking marbled crayfish genome and transcriptome assemblies of using 1066 universal single copy ortholog genes (BUSCOs) shared among arthropods. **a**, Out of 1066 BUSCOs 88% were found complete, 7% fragmented, and 5% had no sufficient evidence in the genome assembly. About 1% of the genes was found in multiple copies. **b**, In the transcriptome assembly 85% out of 1066 BUSCOs were found complete, 2% fragmented, and 13% had insufficient evidence.

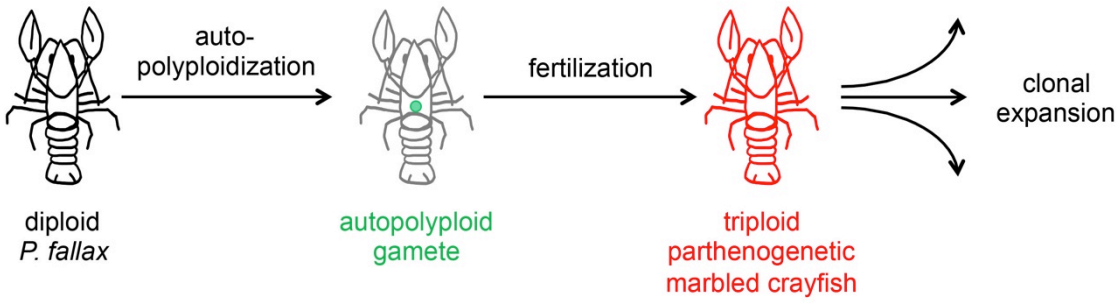

Figure S3. Model for the evolutionary origin of marbled crayfish. The autopolyploidization of a gamete in *P. fallax*, followed by fertilization could have resulted in a triploid genome and the formation of the clonal marbled crayfish population.

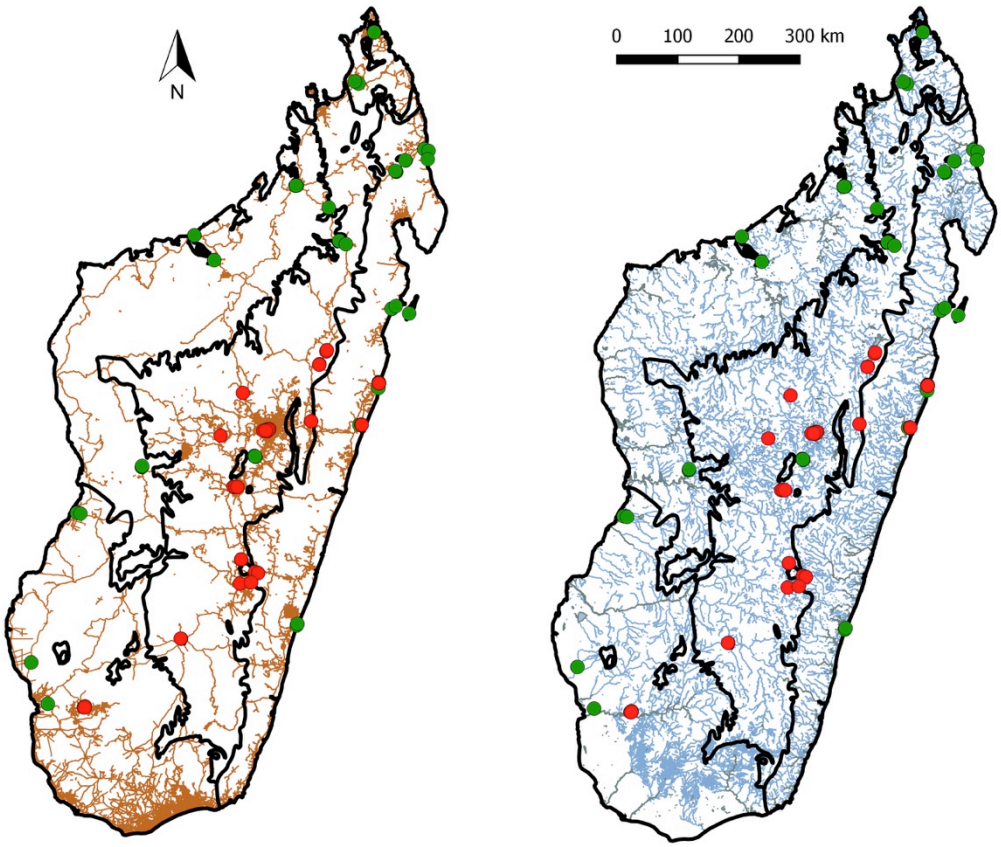

Figure S4. Presence and absence of marbled crayfish populations in the context of road (left) and freshwater (right) networks on Madagascar. Red dots indicate locations with a confirmed presence of marbled crayfish, green dots indicate locations where no marbled crayfish were found. Black lines indicate the boundaries of Madagascar bioclimatic zones[1]. Maps were drawn with QGIS Version 2.18.7. using OpenStreetMap data from Geofabrik.

Table S1 Raw sequencing statistics for the reference genome assembly (Petshop 1).

Type	Targeted insert size	Read pairs [Mio]	Yield [Mbp]	Sequencing protocol
Shotgun	500 bp	817	245,157	HiSeq PE150
Shotgun	550 bp	19	11,218	MiSeq PE300
LJD	3,000 bp	90	18,005	HiSeq PE100
LJD	8,000 bp	169	33,866	HiSeq PE100
LJD	20,000 bp	130	25,950	HiSeq PE100
LJD	40,000 bp	81	16,256	HiSeq PE100
Total		1,306	350,452	

Table S2 Analysis of pairwise collinear blocks using MCScanX.

alignment	score	E-value	hits	seq1	seq2	strand
1	150.0	0.018	3	scaffold113356	scaffold336218	-
2	150.0	0.034	3	scaffold14109	scaffold23998	-
3	150.0	0.075	3	scaffold14698	scaffold155431	-
4	150.0	0.085	3	scaffold14698	scaffold53260	+
5	150.0	0.085	3	scaffold14698	scaffold63477	+
6	150.0	0.066	3	scaffold15229	scaffold22404	+
7	150.0	0.066	3	scaffold15229	scaffold34289	+
8	150.0	0.075	3	scaffold15229	scaffold365170	-
9	150.0	0.075	3	scaffold15229	scaffold39143	-
10	150.0	0.066	3	scaffold15229	scaffold39842	+
11	150.0	0.066	3	scaffold15229	scaffold51465	+
12	150.0	0.066	3	scaffold15229	scaffold73233	+
13	150.0	0.075	3	scaffold15229	scaffold96846	-
14	150.0	0.066	3	scaffold155431	scaffold35251	+
15	150.0	0.066	3	scaffold155431	scaffold43482	+
16	150.0	0.075	3	scaffold155431	scaffold47145	-
17	150.0	0.075	3	scaffold155431	scaffold53260	-
18	150.0	0.075	3	scaffold155431	scaffold63477	-
19	150.0	0.066	3	scaffold155431	scaffold64592	+
20	150.0	0.075	3	scaffold155431	scaffold75525	-
21	150.0	0.066	3	scaffold22404	scaffold34289	+
22	150.0	0.075	3	scaffold22404	scaffold365170	-
23	150.0	0.075	3	scaffold22404	scaffold39143	-
24	150.0	0.066	3	scaffold22404	scaffold39842	+
25	150.0	0.066	3	scaffold22404	scaffold51465	+
26	150.0	0.066	3	scaffold22404	scaffold73233	+
27	150.0	0.075	3	scaffold22404	scaffold96846	-
28	150.0	0.075	3	scaffold35251	scaffold53260	-
29	150.0	0.075	3	scaffold35251	scaffold63477	-
30	150.0	0.075	3	scaffold43482	scaffold53260	-
31	150.0	0.075	3	scaffold43482	scaffold63477	-
32	150.0	0.085	3	scaffold47145	scaffold53260	+
33	150.0	0.085	3	scaffold47145	scaffold63477	+
34	150.0	0.085	3	scaffold53260	scaffold63477	+
35	150.0	0.075	3	scaffold53260	scaffold64592	-
36	150.0	0.085	3	scaffold53260	scaffold75525	+
37	150.0	0.075	3	scaffold63477	scaffold64592	-
38	150.0	0.085	3	scaffold63477	scaffold75525	+

In 38 alignments, we detected 66 collinear genes in 22 scaffolds.

Table S3 Comparative genome sequencing.

Species	Genome coverage	Source
marbled crayfish	71.8x	Petshop 1 laboratory strain
marbled crayfish	54.3x	Heidelberg laboratory strain
marbled crayfish	15.6x	wild catch from lake Moosweiher (Germany)
marbled crayfish	16.8x	MA1, market purchase from Madagascar (Antananarivo)
marbled crayfish	16.2x	Petshop 2 laboratory strain
marbled crayfish	30.0x	MA2, wild catch from Madagascar (Alaotra-Mangoro)
marbled crayfish	36.0x	MA3, wild catch from Madagascar (Analamanga)
marbled crayfish	29.1x	MA4, wild catch from Madagascar (Itasy)
marbled crayfish	27.6x	MA5, wild catch from Madagascar (Vakinankaratra)
marbled crayfish	38.9x	wild catch from Reilingen (Germany)
marbled crayfish	39.9x	Hannover aquarium lineage
P. fallax	19.9x	male from aquarium supply
P. fallax	41.6x	male from aquarium supply
P. fallax	17.8x	female from aquarium supply
P. fallax	38.7x	female from aquarium supply
P. alleni	16.6x	female from aquarium supply

Table S4 Discovery sites for marbled crayfish on Madagascar.

Region	District	Site	Latitude	Longitude	Habitat	Altitude (m)
Atsinanana	Toamasina	Ambodisaina	S18°09.176‘	E49°22.168‘	Rice field channel	12
		Tsarakofafa	S18°09.283‘	E49°22.137‘	Fish pond	9
	Brickaville	Andramagnala	S18°50.535‘	E49°05.022‘	Marsh	80
Alaotra Mangoro	Moramanga	Marovoay	S18°47.350‘	E48°14.764‘	Rice field channel	920
	Ambatondrazaka	Alaotra-Ambatosoratra	S17°37.816‘	E48°29.947‘	Lake	761
		Alaotra-Andreba Gara	S17°36.494‘	E48°30.499‘	Lake	757
		Ambohiboromanga	S17°50.374‘	E48°23.137‘	Stream	778
Analamanga	Antananarivo Renivohitra	Antsingita By Pass	S18°56.821‘	E47°31.280‘	Lake	1262
		Ankatso Hazo be Ravina	S18°54.696‘	E47°33.448‘	Rice field	1267
		Ambodiafotsy	S18°56.278‘	E47°27.680‘	Rice field	1267
		Anosy	S18°55.046‘	E47°31.388‘	Lake	1255
		Ikopa Anosizato Andrefana	S18°56.269‘	E47°29.912‘	Stream	1254
		Miandrarivo	S18°55.394‘	E47°32.278‘	Rice field channel	1270
	Ankazobe	Ankazobe	S18°18.862’	E47°07.272’	Rice field channel	1198
Itasy	Soavinandriana	Ampefy	S19°01.191’	E46°45.285’	Rice field channel	1230
Vakinankaratra	Antsirabe 1	Marovoalavo	S19°52.035‘	E46°57.882‘	Lake + Rice field	1518
		Sahalombo	S19°51.867‘	E46°59.744‘	Stream	1487
		Ranomaimbo	S19°52.142‘	E47°01.791‘	Lake	1491
Mahatsiatra Ambony	Ambohimahasoa	Iarindrano Ambodisahave	S21°04.051‘	E47°05.548‘	Stream	1104
			S21°03.824‘	E47°05.583‘	Rice field	1108
	Lalangina	Sahavondronina	S21°16.579‘	E47°20.111‘	Rice field channel	1192
		Andragnaroa	S21°17.567‘	E47°22.295‘	Stream + Rice field channel	1083
		Sahambavy	S21°26.808‘	E47°15.416‘	Lake	1086
	Fianarantsoa	Ambanimaso	S21°27.915‘	E47°04.324‘	Rice field channel + fish pond	1119
Ihorombe	Ihosy	Ihosy	S22°22.512‘	E46°06.016‘	Stream	711
		Vatrina	S22°22.440‘	E46°05.722‘	Rice field channel	718
Atsimo Andrefana	Bezaha	Thermal water	S23°30.674‘	E44°29.658‘	Rice field	116
		Taheza	S23°29.593‘	E44°30.801‘	Stream	119
		Andalasoingy-Ampasindava	S23°30.788‘	E44°30.621‘	Small permanent water point	108

Table S5 PCR genotyping of marbled crayfish samples from Madagascar**.**

Region	Catch size	Animals analyzed	Samples sequenced	No. of mismatches Cyt b	No. of mismatches Dnmt1
Atsinanana	501	19	8	1	0
Alaotra Mangoro	49	30	10	1	2
Analamanga	206	23	10	0	0
Itasy	20	2	2	0	0
Vakinankaratra	676	35	7	0	0
Mahatsiatra Ambony	473	28	5	0	0
Ihorombe	114	8	2	0	0
Atsimo Andrefana	43	11	5	0	2
Total	**2100**	**176**	**49**	2	4

No. of mismatches indicates the cumulative number of sequence mismatches for each sample group and PCR product.

Table S6 Animals included in phylogenetic analysis of marbled crayfish genomes.

Name	Description
Heidelberg	aquarium lineage from trade fair, since 1995, ref. 2
Hannover	aquarium lineage from trade fair, since 1998
Moosweiher	wild catch from Germany (2013), N48°01.844’ E07°48.368’, ref. 2
Reilingen	wild catch from Germany (2017), N49°17.649’ E08°32.672’
Madagacar 1	market purchase from Antananarivo, Madagascar (2013), ref. 2
Madagacar 2	wild catch from Madagascar (2016), S18°47.350' E48°14.764'
Madagacar 3	wild catch from Madagascar (2016), S18°56.821’ E47°33.280’
Madagacar 4	wild catch from Madagascar (2016), S19°01.191’ E46°45.285’
Madagacar 5	wild catch from Madagascar (2016), S19°52.142’ E47°01.791’
Petshop 1	aquarium lineage from Kölle-Zoo, since 2004, ref. 2
Petshop 2	aquarium lineage from Kölle-Zoo, since 2003, ref. 3

Table S7 Distribution of SNVs in marbled crayfish genomes.

location	position	ref	alt	ref aa	alt aa	animals									
						HD (219)	MW (245)	PET2 (142)	HAN (267)	REI (267)	MA1 (259)	MA2 (284)	MA3 (280)	MA4 (275)	MA5 (262)
scaffold12321	139,799	C	G	G	G	x			x	x	x	x	x	x	x
scaffold12321	139,841	C	G	P	P	x			x	x	x	x		x	x
scaffold13804	22,136	A	T	E	D		x	x	x		x		x	x	x
scaffold180344	24,983	A	T	Q	L	x	x		x			x	x	x	x
scaffold19008	137,128	G	A	A	A							x			
scaffold25978	34,826	A	G	K	R	x	x		x			x		x	x
scaffold305026	3,029	C	T	T	M					x					
scaffold33548	250,756	T	C	V	V			x	x	x			x	x	x
scaffold467	131,673	T	G	L	R					x					
scaffold96424	36,615	A	T	H	L		x		x	x		x	x	x	x

Gray highlighted rows mark non-synonymous SNVs. The number of total SNVs to the reference genome (Petshop 1) is provided in parentheses below each animal name. Animal names are abbreviated as follows: HD (Heidelberg), MW (Moosweiher), PET2 (Petshop 2), HAN (Hannover), REI (Reilingen), MA1 (Madagascar 1), MA2 (Madagascar 2), MA3 (Madagascar 3), MA4 (Madagascar 4), MA5 (Madagascar 5).

Supplementary Methods

Sample preparation for genome size estimation by flow cytometry

Hemocytes of marbled crayfish were extracted from the ventral abdominal artery or from the coxopodite of the 4th pereopod using a 0.5x25 mm needle (Terumo) and 1 ml syringe (Ersta) filled with 100 µl crayfish anticoagulant (100 mM glucose, 34 mM trisodium citrate, 26 mM citric acid, 15.8 mM EDTA, pH 4.6). After centrifugation for 5 min at 1,400 rpm the pellet was washed in 1x PBS (Gibco Life Technologies) and centrifuged again under the same conditions. The pellet was re-suspended in 1 x PBS with 10% DMSO (Sigma-Aldrich) and aliquoted for storage at -80°C.

Human and mouse whole-blood samples were mixed 1:1 with 1x PBS and then gently layered over a Ficoll-Hypaque solution (3 parts Ficoll : 10 parts blood mixture). The centrifugation was performed for 20 min at 400xg with the slowest acceleration rate and brake off. The upper aqueous plasma phase was removed and the lower phase was transferred to a new reaction tube. After adding 3 volumes of Hank's balanced salt solution (HBSS), the samples were centrifuged for 10 min at 400xg. The washing step with HBSS was repeated and the pellets were resuspended in 1x PBS. Aliquots were either used immediately for flow cytometry or stored at -80°C in 10% DMSO.

DNA isolation and quality control

Genomic DNA was isolated from frozen tissue using either DNeasy Blood & Tissue Kit (Qiagen) or Blood & Cell Culture Kit (Qiagen). The tissue was homogenized in lysis buffer of the corresponding protocol using the TissueRuptor (Qiagen). DNA extraction with the DNeasy Blood & Tissue Kit (Qiagen) was performed according to the manufacturer's instructions.

The quality of isolated genomic DNA was assessed via 8% TBE/TAE-Agarose Gel (1% (w/v) Agarose, 1x TBE) and/or via 2200 TapeStation (Agilent) following the manufacturer's

instructions for Genomic ScreenTape. The concentration was determined either via NanoDrop 2000 (Thermo Scientific) following the manufacturer's instructions or PicoGreen. A DNA standard serial dilution (1.56 ng, 3.125 ng, 6.25 ng, 12.5 ng, 25 ng, 50 ng and 100 ng) (Molecular Probes Life Technologies) in 1x TE was freshly prepared for each PicoGreen measurement. The DNA standard dilution and DNA samples (1 µl in 99 µl 1x TE) were measured in triplicates. To each 100 µl DNA solution, 100 µl freshly prepared PicoGreen (1:200 in 1x TE) was added and the fluorescence signals at 520 nm (extension 485 nm, emission 520 nm) were detected by a FLUOstar Optima (BMG Labtech). The DNA concentration of the sample was determined relative to the DNA standard serial dilution.

In silico genome size estimation

Genome size has been estimated *in silico* using raw shotgun reads. For this, occurrences of short substrings (kmers) with length 17 were counted using Jellyfish (1.1.11)[4]. Plotting kmer depths against frequencies revealed two major peaks. The first peak (at n/3, 17x) represents heterozygous kmers while the second peak (n, 49x) represents homozygous kmers. Genome size was estimated by calculating:

$$N = \frac{M \cdot L}{L - K + 1} \text{ and } G = \frac{T}{N}$$

Where N is the average coverage, M is the mean kmer coverage, L is the average read length, K is the kmer length, and T is the number of total bases.

Genome assembly

Before contig assembly, filtering, trimming, and preprocessing of SG and LJD reads was done. In detail, raw shotgun (SG) reads were trimmed for adapter sequences and low-quality bases using Trimmomatic (0.30)[4]. LJD reads were pre-processed with Eurofins software to remove adapter and linker sequences from raw reads. Trimmomatic was used for filtering read pairs

(Phred score: 5, window size: 8 bp, Phred score: 15, min length: 30 bp). Both SG and LJD reads, were further filtered using Prinseq (prinseq-lite-0.20.3)[5] for low complexity, gap containments, a GC content below 10% or above 90%, or when they were shorter than 90 bp (SG) and 30 bp (LJD), respectively. MiSeq data was preprocessed accordingly using Trimmomatic (0.32) and Prinseq. Contig assembly was done using SOAPdenovo2 (2.04-r240)[6] on shotgun reads with a kmer size of 75 and sparse graph construction. Contigs were deduplicated because of high heterozygosity and to reduce input complexity for scaffolding (BBTools v34.72 dedupe.sh). After deduplication, contigs associated with transcripts were connected by mapping transcriptome information to the assembly using L_RNA_SCAFFOLDER[7] which significantly increased transcript representation. LJD reads were used for connecting contigs over long distances using SOAPdenovo2 (2.04) and a kmer size of 35. Four iterations of gap filling using GapCloser (1.12)[6] with a subset of Petshop I shotgun reads dramatically reduced gap content. After assembly, genome quality was estimated using a benchmarking for universal single copy orthologs (BUSCO)[8].

For an automatic annotation using the MAKER pipeline (3.00)[9] sequences larger or equal 10 kb were extracted from the assembly. MAKER was run with standard parameter settings using the manually curated Uniprot/Swiss-Prot database (05/2016)[10] and annotated transcriptome information. Additionally, gene model information from BUSCO analyses were integrated. A hidden Markov model (HMM) for gene model prediction was trained in two iterations on transcriptome information. For functional annotation predicted proteins have been annotated using the Uniprot/Swiss-Prot database (06/2016). Furthermore, functional domains were predicted using InterproScan (5.20-59.0)[11] and integrated into annotation.

Transcriptome sequencing and assembly

Isolated total RNA from hepatopancreas, abdominal musculature, hematopoetic tissue and green glands was mixed to equal parts and sequenced by Eurofins MWG GmbH (Ebersberg,

Germany) using normalized cDNA pools. In parallel, total RNA from hepatopancreas was isolated and sequenced on an Illumina HiSeq platform. Both datasets were treated separately and assembled as follows. Reads were filtered for duplicates using FastUniq[12] and subsequently used to assemble the transcriptome with kmer sizes ranging from 19 to 63 and an insert size of 200. Gapless scaffold sequences were used for further filtering. Gap containing scaffolds were corrected for wrongly inserted gaps. Wrongly inserted gaps were identified by remapping contigs associated with a particular scaffold and evaluating gap structures. Subsequently, the 23 generated transcriptome assemblies from both data sets were merged together by clustering the transcript sequences with 97 % sequence identity using CD-HIT-EST[13]. The longest sequences of the clusters were kept as cluster representatives for further filtering. Overlapping transcripts from different assemblies were joined into one single transcript using CAP3[14]. Repetitive regions were identified with RepeatMasker and transcript sequences with repeats representing more than 10 % of the total sequence length were removed as source for possible miss-assembly. Finally, transcripts with a minimum length of 300 bp were used as the final assembly. Cis-self and trans-self chimeras were identified by splitting the transcript sequences in two equal parts and aligning them to each other using BLASTn. Identified chimera were corrected in the final transcriptome assembly. The quality of the assembled transcriptome was assessed by determining the completeness of 2,675 orthologous genes conserved among arthropods using BUSCO[8].

Transcriptome annotation

Using the automated annotation pipeline from TCW[15] the marbled crayfish transcript sequences were annotated with Universal Protein Rescource (UniProt) terms. UniProt terms were then linked to their Gene Ontology (GO) terms by applying QuickGO. For annotation with Clusters of Orthologous Groups (COG) the COG database was downloaded and marbled crayfish sequences were annotated using RSPBLAST[16]. Annotation with the Kyoto Encyclopedia of

Genes and Genomes (KEGG) was performed using the tool provided on the official website. For phylogenetic analysis of the assembled marbled crayfish transcriptome one species was selected to generate a database representing one of the following phylostrata: bilateria (*Xenopus laevis*, mRNA sequences Xenbase version), pancrustacea (*Drosophila melanogaster*), crustacea (*Daphnia pulex*), decapoda (*Litopenaeus vannamei*) and astacoidea (*Pontastacus leptodactylus/ Astacus leptodactylus*). The marbled crayfish transcripts were then aligned against the generated databases using BLASTx (e-value 10^{-10}). Sequences with significant BLAST hits in all databases were classified as bilaterian, sequences with hits only in *D. melanogaster*, *D. pulex*, *L. vannamei* and *P. leptodactylus* as pancrustacean, etc. The remaining sequences without significant sequence similarity to one of the species were classified as unique.

Coverage and collinearity analysis

Read depths were extracted from gene regions of Petshop 1 alignment files, as provided by the genome annotation file. Coverage depths were counted and plotted using R. Collinearity was estimated by extracting protein sequences from the marbled crayfish genome assembly. Protein sequences were blasted (blastp) against each other and MCScanX[17] was used to conduct collinearity analyses with default parameters, except that a minimum of 3 genes was required to call a collinear block. To detect any effect, the maximum E-value was increased to 1.

Analysis of genetic variants

Variant calling was performed using Freebayes (0.9.21-g7dd41db). Parameters were set for detecting variants with a probability of at least 0.7, reporting all haplotype alleles and including the reference as an allele, a minimum mapping quality of 30 and minimum base quality of 20, and a coverage of at least 6 reads. The availability of datasets from several independent animals for marbled crayfish and *P. fallax* enabled a multi-sample variant calling, while for *P.*

alleni only one sample was available. Furthermore, marbled crayfish samples were analyzed for a triploid ploidy level, while the other two species were analyzed in diploid mode. Phylogenetic trees were generated using the R package APE[18] by pairwise comparisons of variant sites (normalized by the number of total variants provided by the reference individual) with a coverage of at least 8 in all samples and clustering of their maximum distances. Violin plots for heterozygous and homozygous biallelic variants were created using the R package vioplot. The Circos plot was generated using the R package OmicCircos[19] using eight randomly chosen scaffolds from a subset of high nucleotide density genome sequences.

Supplementary References

1. Schatz GE. Endemism in the Malagasy tree flora. In: *Diversity and Endemism in Madagascar* (eds Lourenco WR, Goodman SM). ORSTOM (2000).

2. Vogt G*, et al.* The marbled crayfish as a paradigm for saltational speciation by autopolyploidy and parthenogenesis in animals. *Biol. Open* **4**, 1583-1594 (2015).

3. Seitz R, Vilpoux K, Hopp U, Harzsch S, Maier G. Ontogeny of the Marmorkrebs (marbled crayfish): a parthenogenetic crayfish with unknown origin and phylogenetic position. *J Exp Zool A Comp Exp Biol* **303**, 393-405 (2005).

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

5. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864 (2011).

6. Luo R*, et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).

7. Xue W*, et al.* L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* **14**, 604 (2013).

8. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.

Bioinformatics **31**, 3210-3212 (2015).

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

10. UniProt C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015).

11. Jones P*, et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

12. Xu H*, et al.* FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* **7**, e52249 (2012).

13. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

14. Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-877 (1999).

15. Soderlund C, Nelson W, Willer M, Gang DR. TCW: transcriptome computational workbench. *PLoS One* **8**, e69401 (2013).

16. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**, W327-331 (2004).

17. Wang Y*, et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene

synteny and collinearity. *Nucleic Acids Res* **40**, e49 (2012).

18. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).

19. Hu Y*, et al.* OmicCircos: A Simple-to-Use R Package for the Circular Visualization of Multidimensional Omics Data. *Cancer Inform* **13**, 13-20 (2014).