Supplement

Supplementary Note 1. Individuals and Sequencing

Three snails from the inbred line were used for DNA extraction. Pooling of tissue from whole snails, DNA extraction, library construction of paired libraries with insert sizes of 250, 500, 800, 2k, 5k and 10k as well as sequencing using Illumina HiSeq 2500 technology with read length of 125bp for small insert libraries 200, 500 and 800bp and Illumina HiSeq 2000 with read length 100bp for mate pair libraries 2 kb, 5 kb and 10 kb was performed by BGI, Hong-Kong. In total more than one billion reads (1,000,372,010) containing more than 116Gb (116,162,940,950bp) raw data was produced.

Supplementary Table 1. Sequenced raw data. The paired-end libraries with insert sizes of 250 bp, 500 bp and 800 bp were each sequenced with read lengths of 125 bp (HiSeq 2500). The mate pair libraries with insert sizes of 2 kb, 5 kb und 10 kb were sequenced with read lengths of 100 bp (HiSeq2000).

Insert size	Number of sequences	Number of nucleotides	%GC	Coverage
250 bp	289,883,600	36,235,450,000	39	23
500 bp	197,642,448	24,705,306,000	36	15
800 bp	157,503,550	19,687,943,750	36	12
2,000 bp	168,299,780	16,829,978,000	39	11
5,000 bp	135,963,092	13,596,309,200	39	8
10,000 bp	51,079,540	5,107,954,000	41	3

Supplementary Table 2. Summary statistics of different assembly steps.

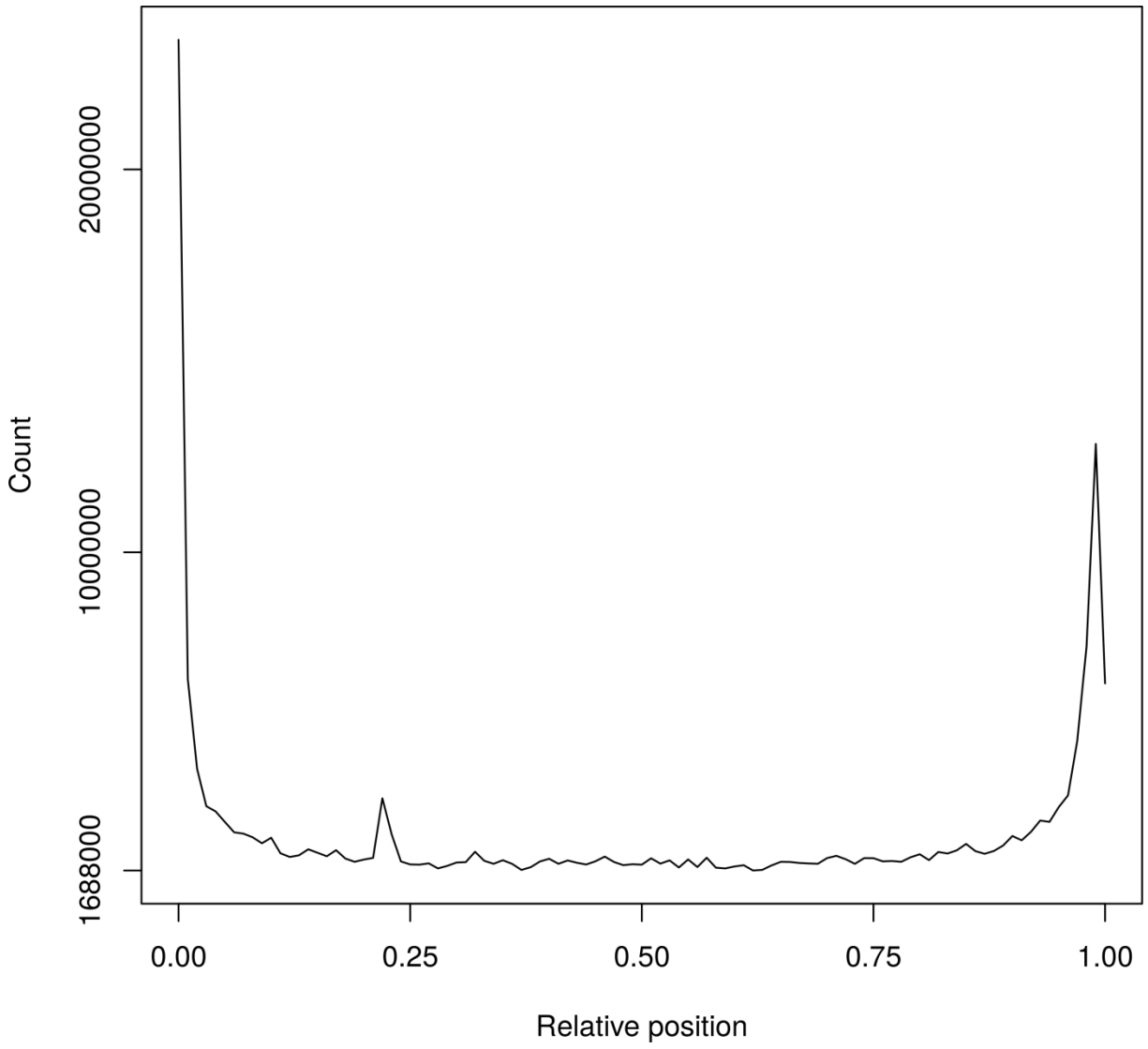
Assembly step	Number of Sequences	Total length [bp]	N50 [bp]	% Ns
Raw reads	1,000,372,010	116,162,940,950		0
Trimmed reads	994,535,287	115,378,553,461		0
Platanus				
Assembly	6,838,932	1,488,367,542	324	0
Scaffolding	193,639	966,366,534	259,302	15.27
Gap close	193,639	927,196,599	250,725	9.64
Length filter ≥ 500 bp	22,306	898,221,812	262,000	9.94
SSPACE	10,317	909,612,132	512,264	11.05
Removing mitochondrial scaffold	10,316	909,598,491	512,264	11.05
L_RNA_scaffolder				
MOTU4	10,268	909,604,080	518,249	11.05
MOTU5	10,036	909,629,612	555,879	11.06
MOTU2+3	9,965	909,636,872	575,006	11.06
Length filter ≥ 1 kb	4,825	906,300,918	576,630	11.09
GapFiller	4,825	909,751,983	578,730	6.42
Adding separate processed mitochondrial scaffold (cutting position)	4,826	909,765,727	578,730	6.42
Remove gap containing scaffolds < 1 kb	4,823	909,764,068	578,730	6.42

Supplementary Note 2. Re-mapping

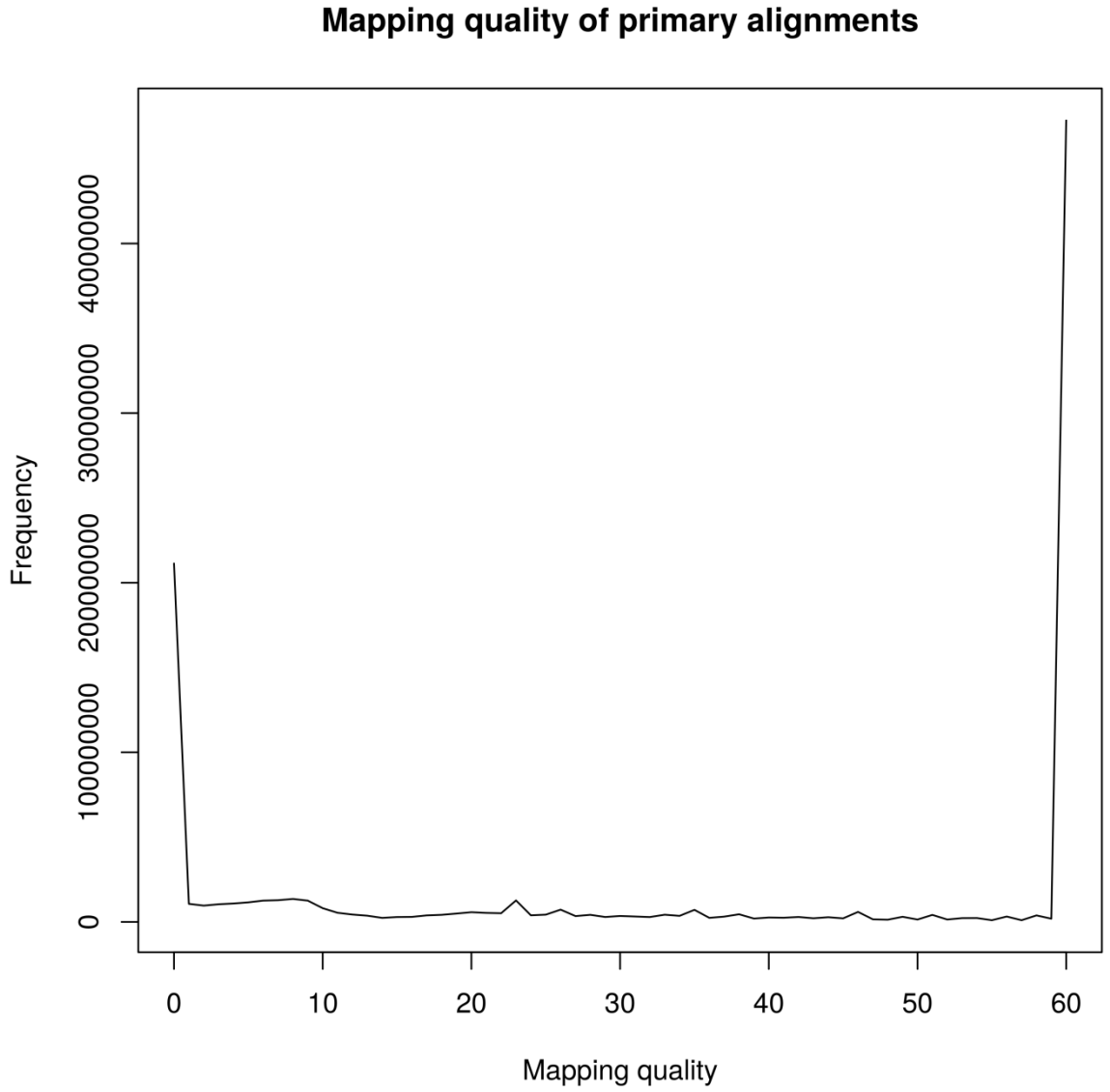
All trimmed genomic reads were mapped unpaired against the final genome assembly using BWA mem 0.7.12-r1039 (Li 2013) with the options -a -c 10000. All other parameters were kept as default. The coverage per position was calculated with samtools 1.1 mpileup and -A -C 50 -d 10000 except default parameters. Nearly all reads (97.64 %) mapped back but only 66.24 % of all nucleotides. The relatively low fraction of remapped nucleotides is due to the mapping quality cutoff of 50 during the mpileup step (94.9% without cutoff) and split mappings, which accumulate at the ends of contigs (Supplementary Figure 1). Split mappings are probably caused by incomplete assembled repeats or algorithmic problems to place a read correctly in a repeat. Per base coverage frequency distribution is shown in Figure 2A and the mapping quality frequency distribution in Supplementary Figure 2.

Re-mapping mate pairs

The trimmed mate pair reads of each library were mapped separately in paired mode against the final genome assembly using BWA mem with all default parameters. The mappings were sorted by position with samtools sort 1.1 (Li et al. 2009) and computation of statistics was realized in QualiMap bamqc 2.2 (Okonechnikov et al. 2015; Figure 2B). Accumulation of insert sizes on the lower end of the distribution and higher mean coverage than expected (Supplementary Table 3) can be explained by mate pairs that cannot span repetitive regions.



Supplementary Figure 1. Number of partially mapped reads along continuous parts of scaffolds.



Supplementary Figure 2. Mapping quality frequency distribution. Secondary and supplementary alignments are excluded.

Supplementary Table 3. Re-mapping statistics from mate pair libraries. Expected coverages were calculated from the number of nucleotides from each trimmed library and the estimated genome size of 1.6 Gb.

Library	Reads mapped [%]	Both in pair [%]	Mean / Expected Coverage
2k	98.42	97.88	17.27 / 10.51
5k	98.62	98.04	14.01 / 8.49
10k	98.28	97.67	5.24 / 3.19

Supplementary Note 3. Results from flow cytometric measurements.

Genome size estimations using different standards yielded comparable 2C-values. Mean 2C-value estimated using the standard *Glycine max* was 3.22 pg (± 0.02 s.d.), using *Lycopersicon esculentum* 3.19 pg (± 0.01 s.d.). These values correspond to 3149.16 Mb and 3119.82 Mb, respectively. The CVs for the G0/G1 peak of the analysed samples ranged from 1.74 to 3.92% (mean 2.56).

Supplementary Note 4. Genome size estimation from coverage

The number of total trimmed nucleotides which were used in the assembly divided by the maximum of the per-position coverage frequency distribution (Figure 2A) is an estimate for the genome size, assuming even sequencing coverage throughout the genome.

$$\frac{\text{Total trimmed nucleotides}}{\text{peak coverage}} = \frac{115.35\text{Gb}}{72} = 1.6025\text{Gb}$$

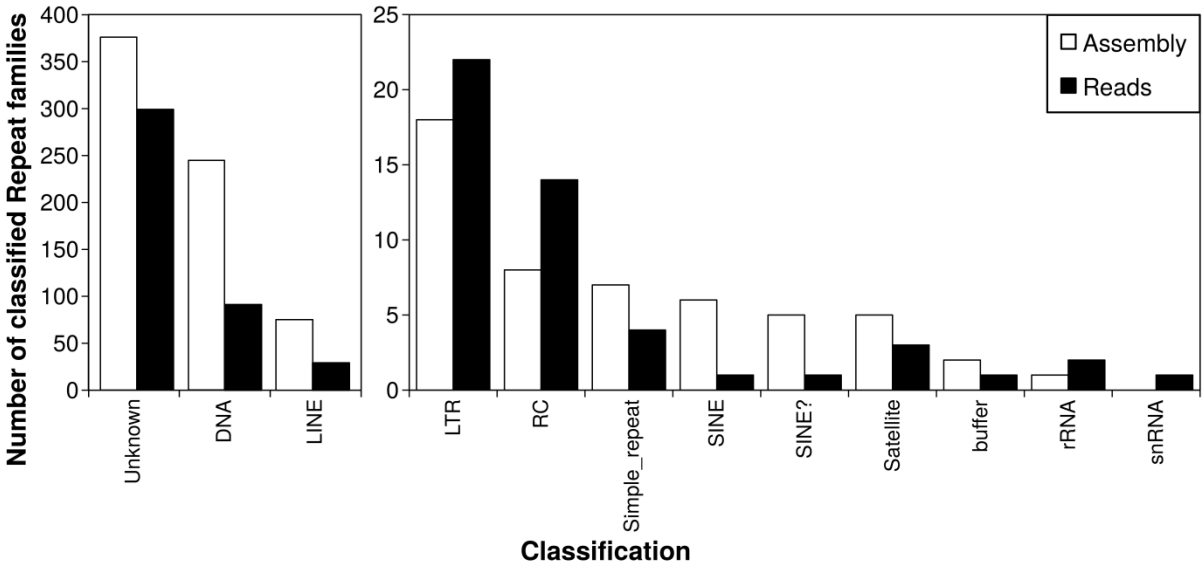
Supplementary Note 5. Repeat content

Re-mapping of all reads to the repeat library (same procedure as re-mapping on the genome assembly) revealed 244,177,765 reads (23.58 Gb) to be lying within repetitive sequences. A size estimation on the peak coverage results in 327.54Mb (20.5% of the 1.6 Gb estimated genome size) which is comparable to the length of annotated repeats in the assembly by the MAKER2 pipeline respectively RepeatMasker.

$$\frac{\text{Total nucleotides in repeats}}{\text{peak coverage}} = \frac{23.58\text{Gb}}{72} = 327.54\text{Mb}$$

$$367.14\text{Mb annotated repeats} + 58.45\text{Mb gaps} = 425.59\text{Mb repeats in assembly}$$

$$425.59\text{Mb repeats in assembly} + 692.74\text{Mb missing} = 1118.33\text{Mb (69.9\% possible repeats)}$$



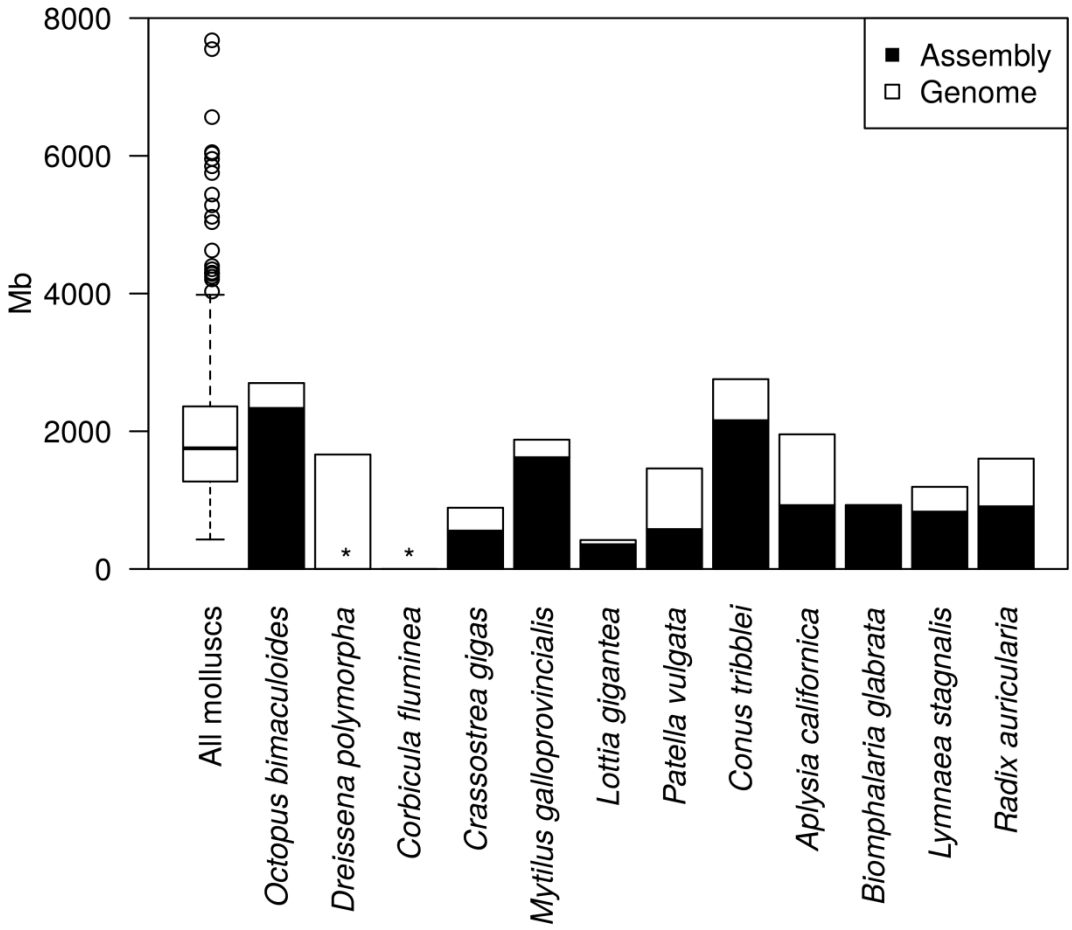
88

89 **Supplementary Figure 3.** Classified repeat families. RepeatModeler was executed on the

90 assembly and on the contigs originating from randomly drawing reads at certain coverage and their

91 assembly. Please note the different scales.

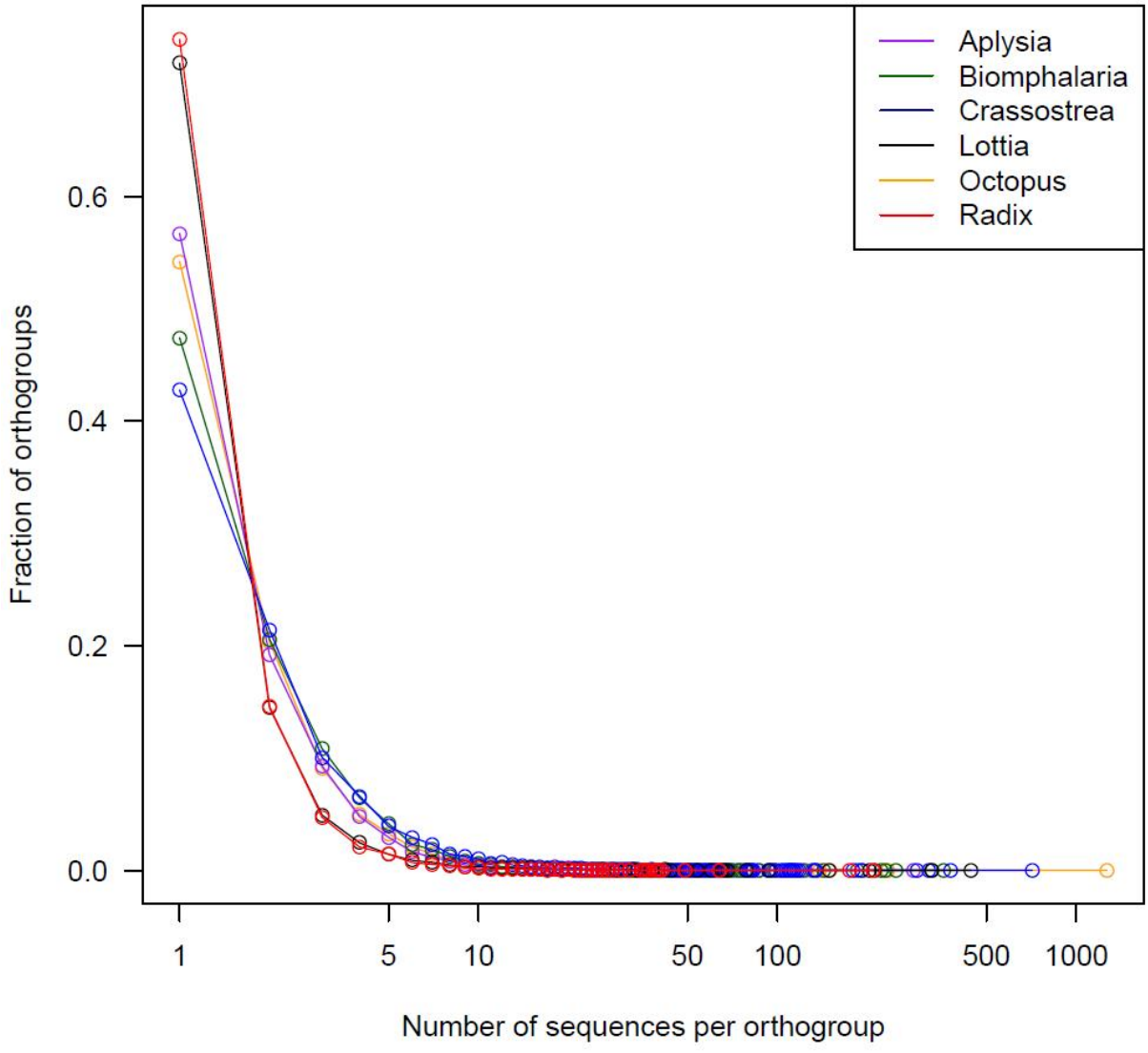
92



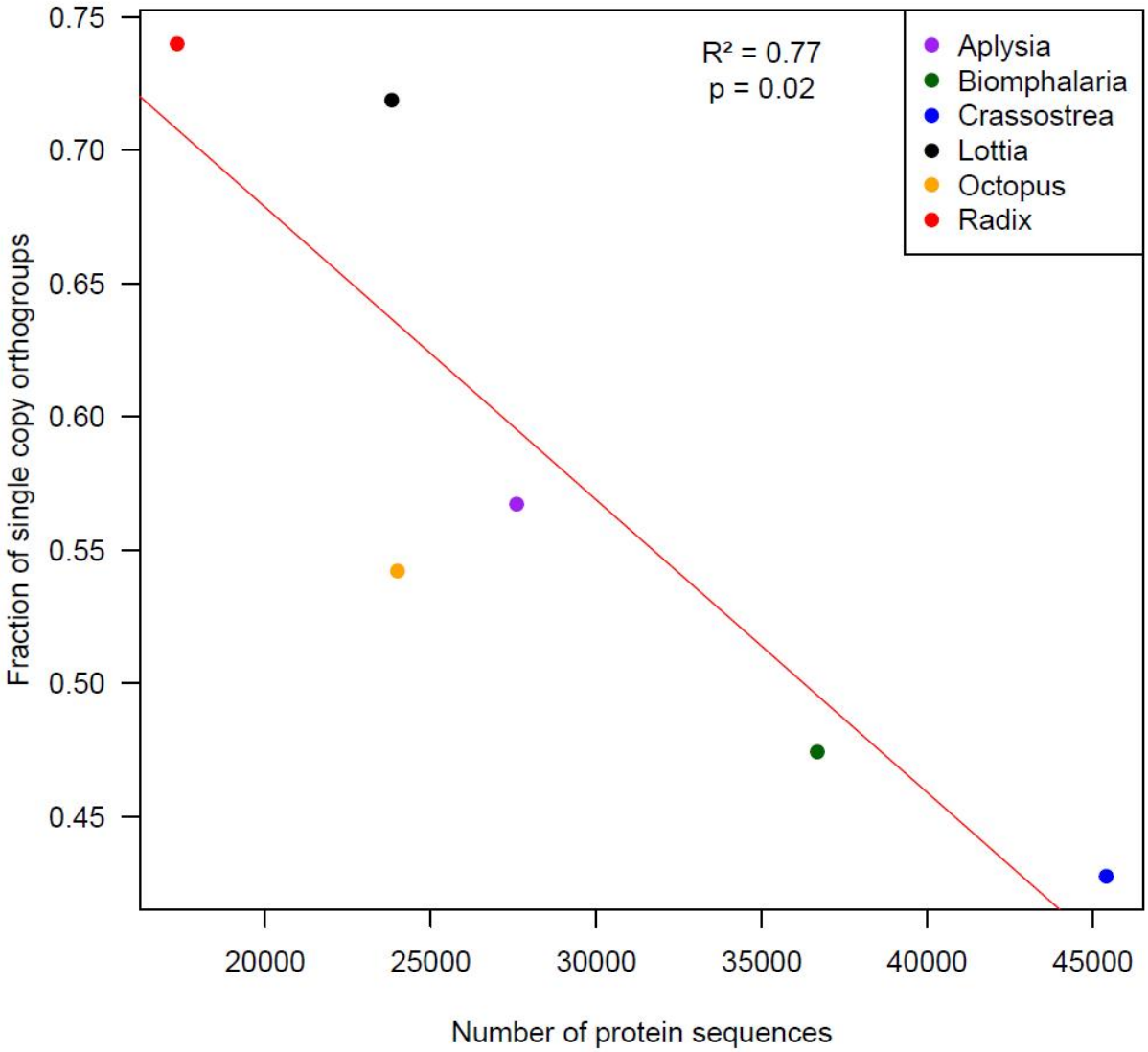
Supplementary Figure 4. Mollusc genome sizes. The first column shows the distribution of genome sizes of all molluscan records found in the Animal Genome Size Database (Gregory 2016; accessed on 28th of October 2016; N=263; circles are outliers). The total height of bars from mollusc species with available genome assemblies shows the estimated genome size and the black filled part the fraction represented in the corresponding assembly. *The total assembly length of *Dreissena* and *Corbicula* is below 1 Mb and therefore no black bar is visible in the graph. There is no estimated genome size for *Corbicula* and therefore no white bar is displayed. Citations from assemblies and genome sizes are given in Table 1.

Supplementary Table 4. Proteins similar to Swiss-Prot entries (accessed May 11th 2016; blastp e-value < 10^{-10}).

Species	Annotated proteins	Proteins similar to Swiss-Prot	Percentage
Octopus bimaculoides	23,994	19,217	80.1
Aplysia californica	27,591	19,920	72.2
Radix auricularia	**17,338**	**12,207**	**70.4**
Crassostrea gigas	45,406	30,917	68.1
Biomphalaria glabrata	36,675	24,422	66.6
Lottia gigantean	23,822	13,950	58.6



Supplementary Figure 5. Number of sequences from one species in orthogroups. The fraction of orthogroups (see Material and Methods) is the number of orthogroups containing this number of different protein sequences of one species relative to the number of orthogroups containing this species. Note the logarithmic x-axis.



Supplementary Figure 6. Correlation of the number of annotated protein sequences per species and the number of orthogroups containing only one sequence of this species. The fraction of orthogroups is relative to the number of orthogroups containing this species.

Supplementary Table 5. Protein sets used for ortholog clustering and GO-term enrichment. Protein sets from six different mollusc species and 16 different spiralian species from outside the Mollusca, which were used for orthologous clustering and Gene Ontology enrichment analysis. (# annotated OGs = number of ortholog groups, which were functionally annotated by GO-terms using a sequence of the according species)

Species	NCBI Genome ID	# sequences	# sequences in orthogroups	# unassigned sequences	# orthogroups containing species	# species specific orthogroups	# annotated sequences from InterProScan	# annotated OGs from species in reference list
Radix auricularia		17,338	15,880	1,458	9,128	5	10,436	6,105
Biomphalaria glabrata	357	36,675	33,044	3,631	10,594	43	21,813	1,977
Aplysia californica	443	27,591	25,463	2,128	10,208	44	17,495	863
Lottia gigantea	15113	23,822	20,444	3,378	9,696	56	12,387	886
Crassostrea gigas	10758	45,406	38,769	6,637	9,710	115	28,107	1,576
Octopus bimaculoides	41501	23,994	22,211	1,783	8,139	72	16,396	500
Caenorhabditis elegans	41	28,137	25,780	2,357	11,481	50	14,052	2,237
Caenorhabditis remanei	253	31,476	27,504	3,972	12,515	57	13,795	1,015
Caenorhabditis brenneri	254	30,670	25,398	5,272	11,387	94	13,154	646
Caenorhabditis briggsae	40	21,959	18,977	2,982	11,674	45	9,446	322
Wuchereria bancrofti	2616	19,323	15,942	3,381	8,641	14	7,768	683
Pristionchus pacificus	246	16,763	11,488	5,275	6,022	65	7,054	659
Loa loa	2686	16,281	12,623	3,658	8,439	14	6,646	238
Clonorchis sinensis	2651	13,634	12,479	1,155	6,750	2	6,590	808
Schistosoma mansoni	236	11,713	10,603	1,110	6,677	8	6,175	256
Opisthorchis viverrini	32471	16,356	13,425	2,931	6,475	6	6,120	129
Hymenolepis microstoma	24432	12,371	9,822	2,549	5,906	35	5,839	402
Schistosoma haematobium	10705	11,140	10,496	644	6,686	13	5,689	129
Echinococcus multilocularis	22333	10,656	9,368	1,288	6,559	1	5,619	162
Echinococcus granulosus	10706	11,319	8,611	2,708	6,094	5	5,274	110
Capitella teleta	15118	31,978	25,701	6,277	9,353	107	0	0
Helobdella robusta	15112	23,426	17,475	5,951	6,908	53	0	0

Supplementary Note 6. Ortholog clustering and Gene Ontology enrichment

Protein sets from six (including *Radix*) different mollusc species and 16 different spiralian species from outside the Mollusca (Supplementary Table 4) were used to predict ortholog cluster by OrthoFinder 0.7.1 (Emms & Kelly 2015) with default parameters. Functional annotation of all protein sets with Gene Ontology (GO) terms was performed with InterProScan 5 (Quevillon et al. 2005; Zdobnov & Apweiler 2001), using default parameters.

To construct the reference list for the GO enrichment analysis a step wise procedure was implemented to obtain GO annotations for as many orthogroups as possible. As *R. auricularia* is our focal species, all GO-annotations obtained for *R. auricularia* were assigned to orthogroups. In the following steps GO-annotations from one species after another were added in a phylogenetic order until all orthogroups had assigned GO-terms, where possible (see order of species from top to bottom and the corresponding number of orthogroups annotated with each corresponding species in Supplementary Table 5). The GO-term enrichment analysis was based on this reference list of annotated orthogroups compared to a test set containing orthogroups of certain sets of species. The analysis were performed using TopGO (Alexa & Rahnenfuhrer 2016).

Supplementary Table 6. Significantly enriched GO-terms for *Radix* specific and unassigned orthogroups (FDR < 5%).

Term	Annotated	Significant	Expected	q-value
Nucleoside transmembrane transport	15	6	0.29	2.10E-07
Glycolytic process	41	6	0.79	0.00012
Microtubule-based movement	138	10	2.66	0.00033
Carbohydrate metabolic process	399	18	7.7	0.00172
Tubulin complex assembly	4	2	0.08	0.00216
Post-chaperonin tubulin folding pathway	4	2	0.08	0.00216
Cilium or flagellum-dependent cell motility	15	3	0.29	0.00271
Neurotransmitter transport	59	5	1.14	0.00546
transport	1,617	42	31.19	0.00594
Sulfate transport	13	2	0.25	0.0251
cGMP biosynthetic process	14	2	0.27	0.02891
Peptidyl-glutamic acid carboxylation	2	1	0.04	0.0382
Inositol trisphosphate metabolic process	2	1	0.04	0.0382
Clathrin coat assembly	2	1	0.04	0.0382
glutamine biosynthetic process	2	1	0.04	0.0382
Clathrin-mediated endocytosis	2	1	0.04	0.0382
Chitin metabolic process	134	6	2.58	0.045

Supplementary Table 7. Significantly enriched GO-terms in orthogroups containing all five mollusc species but *Radix* (FDR < 5%).

Term	Annotated	Significant	Expected	q-value
Meiotic prophase I	1	1	0	0.0013
Synaptonemal complex assembly	2	1	0	0.0027
Actin ubiquitination	3	1	0	0.004
Carbohydrate transport	9	1	0.01	0.0119
G-protein coupled receptor signaling pathway	457	3	0.61	0.0206

Supplementary Note 7. Preprocessing and trimming

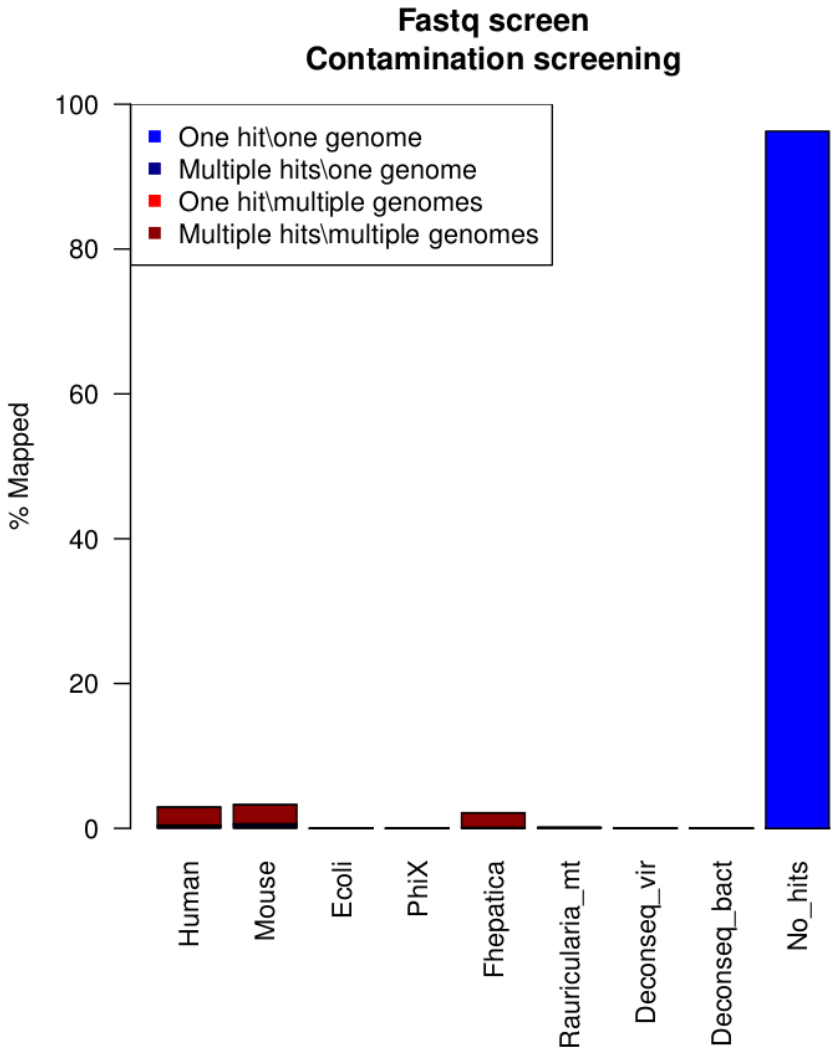
All read files were quality checked using FastQC 0.10.1 (Andrews 2010). Reads from small insert libraries were preprocessed with Trimmomatic 0.33 (Bolger et al. 2014) using the adapter trimming along with a custom adapter file (ILLUMINACLIP:<adapter.fasta>:2:30:10). FastQC reports from

mate pair libraries showed overrepresented k-mers at 3’ ends of reads that could be assembled into fragments of external adapters which is only possible if selection for fragments containing junction adapter was not completely free from other, smaller fragments. Mate pair reads with (partial) external adapter sequence do indeed overlap. Therefore, motifs from external adapter contamination only were used during adapter trimming additional to a minimum length threshold of 100 bp to discard false mate pair reads (ILLUMINACLIP:<motifs.fasta>:2:7:7 MINLEN:100).

99.4 % (994,535,287) of all reads and 99.3 % (115,378,553,461 bp) of all nucleotides survived trimming.

Supplementary Note 8. Contamination screening

All adapter trimmed reads were checked for possible contamination using FastqScreen 0.5.2 (Andrews et al. 2015) with libraries from human (*Homo sapiens* GRCh38), mouse (*Mus musculus* GRCm38), *E. coli* (U00096.3), Enterobacteria phage phiX174 (NC_001422.1), the known *Radix* parasite *Fasciola hepatica* (GCA_000947175.1), *R. auricularia* mt genome as positive control (NC_026538.1) and simulated bacterial and viral Databases from DeconSeq (Schmieder & Edwards 2011). Most sequences could not be mapped to the provided libraries (Supplementary Figure 5). Small fractions of reads could be mapped multiple times to multiple genomes which hint towards similar repeats in provided libraries and *Radix*.



Supplementary Figure 7. Results of contamination screening. All trimmed reads were mapped against different libraries containing possible contamination sources using FastqScreen.

Supplementary Note 9. Material and methods of flow cytometric analysis

Genome size (2C-values; Greilhuber et al. 2005) was estimated by flow cytometry using fresh foot tissue of six individuals of the same *Radix auricularia* inbred line as used for sequencing (Supplemtary Note S3) and the Partec CyFlow Space (Partec, Münster, Germany) equipped with a green solid-state laser (Partec, 532 nm, 30 mW). Sample preparation followed two-step Otto protocol (Otto 1990), with an internal standard *Glycine max* (L.) Merr. cv. Polanka (2C = 2.50 pg; Doležel et al. 1994) and *Lycopersicon esculentum* Mill. cv. Stupické polní tyčkové rané (2C = 1.96 pg; Doležel et al. 1992). The tissue of *Radix auricularia* and the internal reference standard were mixed and chopped with a razor blade in a Petri dish containing 1 ml of ice-cold Otto I buffer (0.1 M citric acid, 0.5%

Tween 20). The suspension was filtered through a 42-μm nylon mesh and incubated for approximately 15 min at room temperature. The staining solution consisted of 1 ml of Otto II buffer (0.4 M $Na_2HPO_4 \cdot 12\ H_2O$), β-mercaptoethanol (final concentration of 2 μl/ml), intercalating fluorochrome propidium iodide (PI) and RNase IIA (both at final concentrations of 50 μg/ml). Fluorescence intensities of 5000 particles (nuclei) were recorded. Sample/standard ratios were calculated from the means of the sample and standard fluorescence histograms, and only histograms with coefficients of variation (CVs) < 5% for the G0/G1 sample peak were considered. Four to five replicates were measured on three different days in order to minimize potential random instrumental error. If the between-day variation in fluorescence intensity was above 4% the most extreme value was discarded. The 2C-values estimated using different standards were inferred separately.

Supplementary Note 10. Transcriptome assemblies

In order to generate de novo transcriptomes (contigs) of available RNA-seq data, raw Roche 454 reads from *R. auricularia* and MOTU3 (Feldmeyer et al. 2015; accession numbers SRR1926149 and SRR1926204) were analysed in FastQC and trimmed with Trimmomatic 0.33 ("HEADCROP:35 TRAILING:20 MINLEN:200" and "ILLUMINACLIP:<overrepresented_k-mers.fa>:2:1:1:1 TRAILING:20 SLIDINGWINDOW:10:20 HEADCROP:35 CROP:540 MINLEN:50" respectively). The *R. auricularia* and MOTU3 transcriptomes were assembled using the Overlap Layout Consensus algorithms of MIRA 4.0.2 (Chevreux et al. 1999) (job = est,denovo,accurate; parameters = 454_SETTINGS -ALIGN:min_relative_score=75; technology = 454). Contiguous sequences were constructed from the above mentioned contigs and contigs from Feldmeyer et al. (2015) via meta-assembly using MIRA (job = genome,denovo,accurate; parameters = TEXT_SETTINGS --noclipping -AS:epoq=no -AS:mrpc=1 -OUT:sssip=yes; technology = text). Meta-contigs and unassembled contigs from the meta-assembly were merged to obtain the final transcriptomes.

Raw Illumina sequences for MOTU5 were obtained from Feldmeyer et al (2015) (SRR1926203), and *R. balthica* (MOTU2) from (Feldmeyer et al. 2011; Tills et al. 2015) and so far unpublished 50bp Illumina reads. All embryonic *R. balthica* reads are deposit in NCBI's BioProject PRJEB9533. Analyses in FastQC provided the overrepresented k-mers. Trimming using Trimmomatic 0.33

("ILLUMINACLIP:<adapter.fasta>:2:30:10 ILLUMINACLIP:<overrepresented.fasta>:2:1:1:1 SLIDINGWINDOW:5:20 HEADCROP:15 CROP:50 MINLEN:50"; "ILLUMINACLIP:<adapter.fasta>:2:30:10 ILLUMINACLIP:<overrepresented_k.mers.fa>:0:1:1:1 SLIDINGWINDOW:5:30 MINLEN:50" and "ILLUMINACLIP<adapter.fasta>:2:30:10 HEADCROP:10 MINLEN:35" respectively) survived 75.1 % for MOTU5 and 84.1 % for *R. balthica* respectively. The two datasets from MOTU5 and *R. balthica* were assembled unpaired with Trinity 2.0.6 (Grabherr et al. 2011) and Bridger 2014-12-01 (Chang et al. 2015) each. The Trinity and Bridger contigs from one species were meta-assembled in MIRA with the same parameters as the meta assembly from *R. auricularia* (MOTU4) and MOTU3.

Supplementary Note 11. Genome assembly

All trimmed reads were assembled using the Platanus 1.2.1 pipeline (Kajitani et al. 2014). The assembly was performed with default parameters except the initial k-mer size of 63 and a stepsize of 2. The automatically detected maximum k-mer size was 88, implying that the last stepsize was 1 (from 87 to 88). Scaffolding within the Platanus pipeline was computed with default parameters (e.g. minimum 3 links) using all six libraries in ascending order along with their average insert sizes according library preparation. Gapclose as last step of the Platanus pipeline was performed with standard parameters using all six libraries in ascending order regarding insert size. All sequences smaller 500 bp were excluded from following assembly steps.

The filtered and gapclosed Platanus scaffolds were scaffolded again using SSPACE Standard 3.0 (Boetzer et al. 2011). Since SSPACE/our hardware was not able to run the scaffolding successfully until the end, separately all six libraries were mapped unpaired using bowtie2 2.2.5 (Langmead & Salzberg 2012) against the filtered and gapclosed Platanus scaffolds. The Mappings were sorted by read name with samtools 1.1 (Li et al. 2009) and converted to the SSPACE readable TAB-format using the script sam_bam2tab.pl provided by SSPACE. All trimmed reads from all six libraries were used during the scaffolding via SSPACE using default parameters apart from contig extension switched on. The insert sizes and corresponding errors from paired reads were calculated from the re-mappings to the filtered and gapclosed Platanus scaffolds (250 bp: 236/0.17; 500 bp: 488/0.08;

800 bp: 771/0.06; 2 kb: 2147/0.12; 5 kb 4909/0.09; 10 kb 10010/0.10; "Library": "mean insert size"/"error" respectively). Unpaired reads for contig extension only were aligned by SSPACE using bowtie.

The SSPACE scaffolds were scaffolded again with L_RNA_scaffolder (Xue et al. 2013) using the transcriptomes of four *Radix* species (Supplementary Note 10). Therefore raw transcriptomic data from *R. auricularia* (MOTU4), *R. balthica* (MOTU2), MOTU3 and MOTU5 was preprocessed and assembled to obtain ESTs. Afterwards the ESTs were mapped with BLAT 35 (Kent 2002) and filtered for order and orientation on one scaffold using a custom perl script. Only ESTs that survived filtering were used as input in BLAT again to obtain the psl-file which is used as input for L_RNA_scaffolder. All four transcriptomic meta assemblies were mapped with BLAT 35 against the SSPACE scaffolds using standard parameters except *-extendThroughN* and *-out=blast8*. Only alignments with correct order and orientation of the parts of the split alignment on one scaffold, covering at least 20 bp and 80 % of the transcriptomic contig length were kept (Supplementary Figure 8).

Before transcriptomic scaffolding the mitochondrial scaffold was excluded to avoid misscaffolding caused by Mitochondrial DNA-like sequences in the nucleus (NUMTs). Afterwards the filtered transcriptomic contigs were used to sequencial scaffold the mitochondrial-free SSPACE scaffolds in three steps using L_RNA_scaffolder by first mapping the filtered transcriptomic contigs again with BLAT 35 (-extendThroughN -noHead) to create the correct input for L_RNA_scaffolder. All scaffoldings with L_RNA_scaffolder were performed with default parameters. First the mitochondrial-free SSPACE scaffolds were scaffolded with the filtered transcriptomic contigs from the same species *R. auricularia* (MOTU4). Second the output from the first scaffolding was used as input along with the filtered transcriptomic contigs from *Radix sp.* MOTU5. Third the output from the second scaffolding was used as input along with the filtered transcriptomic contigs from *R. balthica* (MOTU2) and *Radix sp.* MOTU3 together since hybridization is observed between the two MOTUs (Patel et al. 2015). All results from transcript filtering and scaffolding are summarized in Supplementary Table 8.

The excluded mitochondrial scaffold was cut at the same site as by Feldmeyer et al. (2015) and scaffolded without the shortest (250 bp) insert library in SSPACE (using the same parameters and

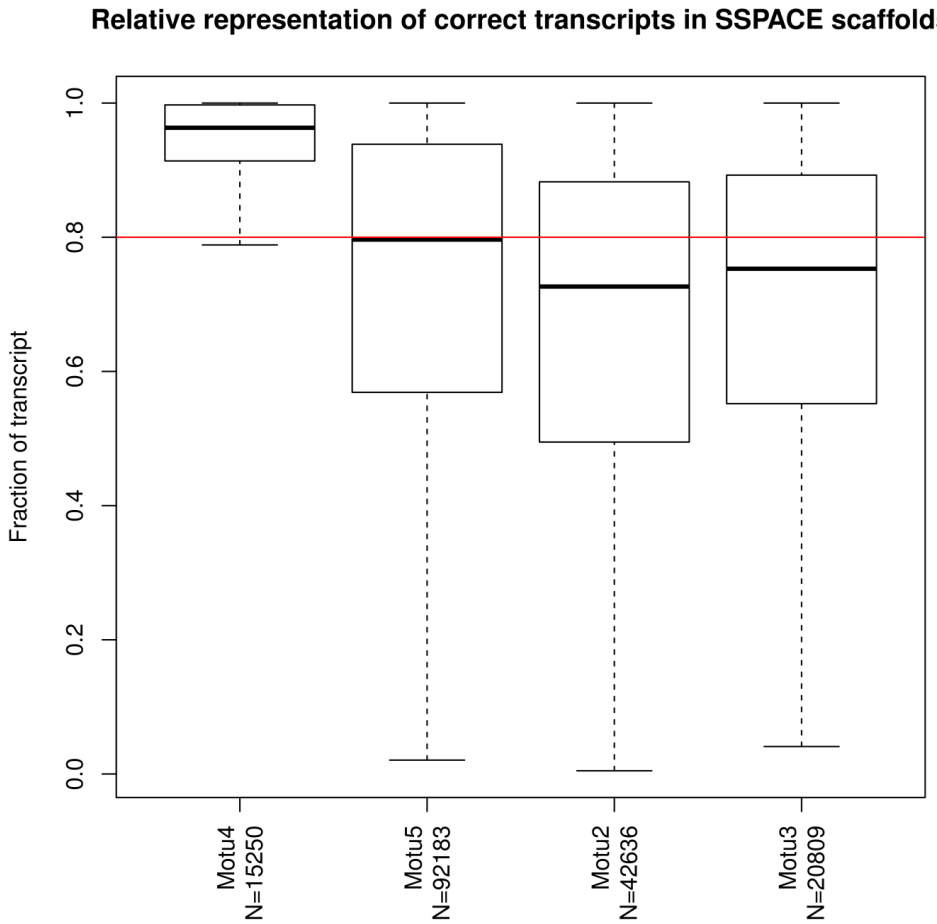
pipeline as before). Omitting the library with the smallest insert size ensures to keep the cutting position. Afterwards a gapfilling with GapFiller 1-10 (Boetzer et al. 2012) was performed using all genomic trimmed paired Illumina reads along with the insert sizes and errors used during genomic scaffolding with SSPACE except from default parameters. Mapping within GapFiller was executed via BWA for all libraries and closed all eight gaps and 750 missing nucleotide positions within one iteration.

After scaffolding with L_RNA_scaffolder all scaffolds smaller than 1 kb where excluded (5,140 scaffolds [51.58 %] / 3,335,954 bp total length [0.37 %]) and the mitochondrial scaffold was added.

All remaining scaffolds were used as input to GapFiller together with all trimmed paired reads from all libraries and the parameters used before during gapfilling of the mitochondrial scaffold.

Because of gap resizing ten scaffolds smaller than 1 kb remained in the assembly after gapfilling. Three of them still contained N's and were removed from the assembly.

A summary of the results of all assembly steps is in Supplementary Table 2 together with corresponding intermediate results.



Supplementary Figure 8. Transcriptome filtering. Fraction of mapped positions of each contig surviving order and orientation filtering. The red line marks the cut-off for contigs used in scaffolding with L_RNA_scaffolder at 80 %.

Supplementary Table 8. Genome scaffolding with transcriptomic data. Four *Radix* species (MOTUs) were used. The first line of each cell shows the number of sequences first and the total length in Mb second. In the second line the fraction of sequences and the fraction of total length relative to the unfiltered/raw meta transcriptome is shown respectively.

	MOTU4	MOTU5	MOTU2	MOTU3	Total
Raw meta transcriptome	22,798 / 16.3	145,687 / 75.2	91,728 / 101.4	34,418 / 27.9	294,631 / 220.8
Mapped with BLAT	22,108 / 16.0 97.0 / 98.3	116,566 / 66.6 80.0 / 88.5	74,411 / 95.6 81.1 / 94.3	30,686 / 26.4 92.9 / 94.3	243,771 / 204.5 82.7 / 92.6
Order / orientation	15,250 / 10.6 66.9 / 65.3	92,183 / 50.2 63.3 / 66.7	42,636 / 45.3 46.5 / 44.7	20,809 / 17.2 60.5 / 61.3	170,878 / 123.3 58.0 / 55.8
Relative representation ≥80%	14,687 / 10.4 64.4 / 63.6	45,672 / 31.2 31.3 / 41.5	16,584 / 22.2 18.1 / 21.9	8,864 / 7.6 26.8 / 27.1	85807 / 71.4 29.1 / 32.3
Scaffolded sites	49	232	71		352

Supplementary Note 12. Repeat library

The repeat library was created using dnaPipeTE 1.2 (Goubert et al. 2015) and RepeatModeler 1.0.4 (Simit & Hubley 2015).

dnaPipeTE

As suggested from the first author, mitochondrial and reverse reads were excluded from the trimmed reads as input for dnaPipeTE, thereby producing an input file containing only forward and unpaired reads from all libraries (499,530,440 reads / 57,943,393,315 bp). For 30 coverages from 0.0001 to 0.7 at an estimated genome size of 1.3 Gb two samples each were drawn. All parameters were set as default except the minimal contig length was lowered from 200 to 50 bp. After analysing the N50 distribution (Supplementary Figure 9) 0.025 was determined as the optimal sampling coverage and 50 repetitions were executed for this coverage with same parameters as above, resulting in 734,889 contigs with total length of 110,811,062 bp. The maximized N50 at 0.0075x coverage was not viewed to be the optimal sampling coverage because the samples are very small and a huge variation regarding N50 is expected, furthermore filter steps are induced afterwards to discard non-repeats.

Coverage of dnaPipeTE contigs

All trimmed reads were mapped unpaired against all contigs from the 50 repetitions using BWA mem with the options -t 80 -k 25 -a -y 26 -c 1000000000 apart from default settings. Using samtools 1.3, the

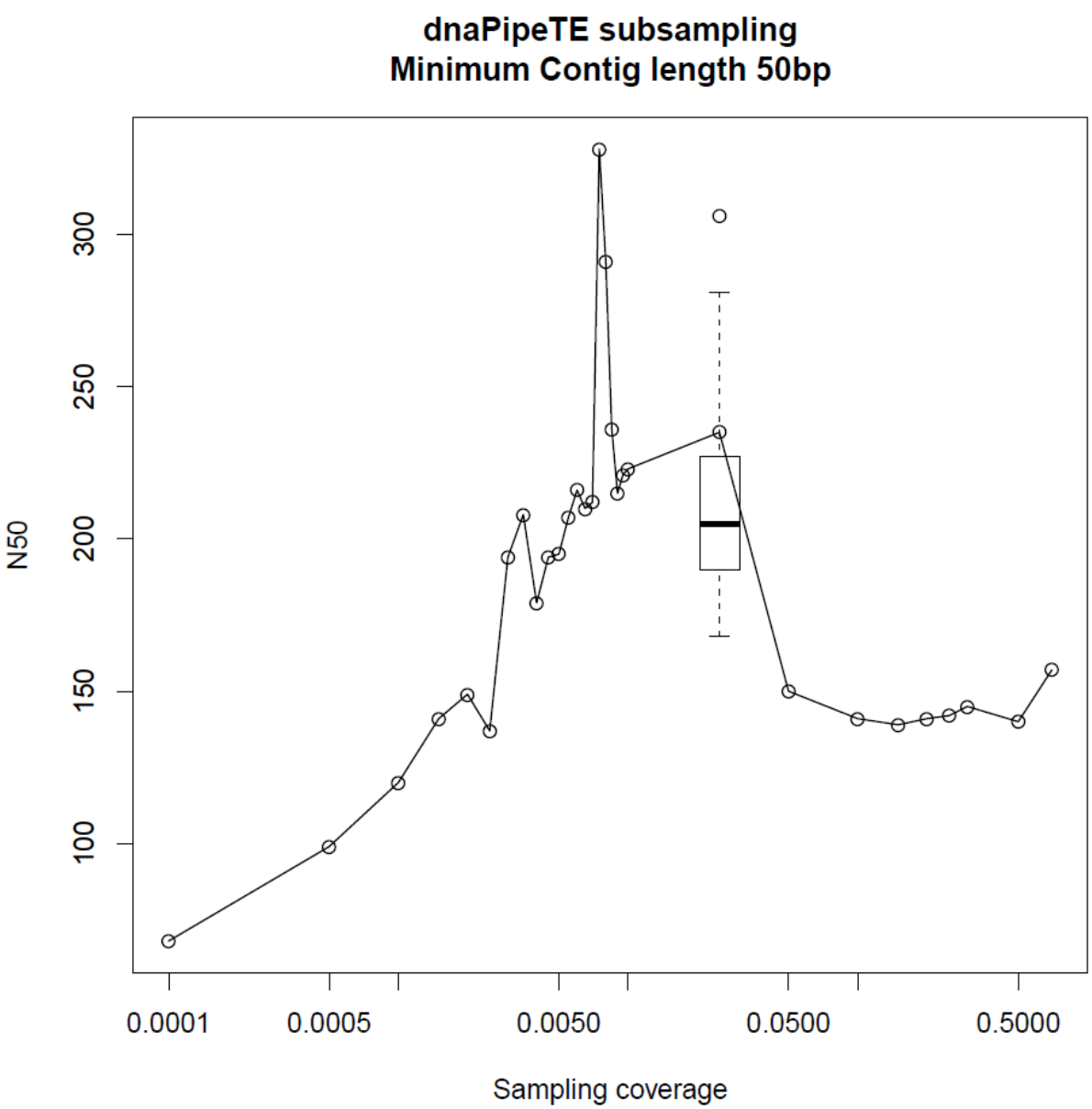
coverage per position was calculated running *mpileup* with options -A -C 50 -d 1000000 apart from standard parameters. Contigs with a median coverage smaller than the 90 % quantile (94x) of the per position coverage distribution from the re-mapping of all genomic reads used in the assembly to the final genome assembly were filtered out (669,284 contigs / 78,314,604 bp). The remaining 65,605 contigs (32,496,458bp total length) with sufficient coverage were used in the next steps.

RepeatModeler was then executed with default parameters on final genome assembly and on the high covered contigs from dnaPipeTE. The resulting fasta files containing the repeat families were concatenated into 1,216 sequences with a total length of 1,111,662bp.

Filtering out possible proteins from the repeat library

First a repeat-protein free protein database was built from Swiss-Prot database (Bateman et al. 2015) (accessed on May 11th 2016) and all repeat sequences from the Repbase database (Bao et al. 2015) (accessed on May 11th 2016). The repeat sequences were searched in the Swiss-Prot database via BlastX 2.3.0+ (Camacho et al. 2009) with an e-value cutoff of 10^{-11}. Protein sequences with hits from repeat sequences containing at least 20 bp in one hit were removed to obtain a repeat-protein free protein database.

The concatenated repeat families from RepeatModeler runs on the genome assembly and the contigs from read subsampling were blasted (BlastX; e-value cutoff 10^{-11}) against the repeat-protein free protein database. For six different families a hit was reported. These sequences were removed from final repeat library containing 1,210 sequences totalling a length of 1,101,332bp and an N50 of 1,492bp.



Supplementary Figure 9. Read subsampling. N50 distribution of 30 repeat sequence assemblies at different sampling coverages is shown by the connected circles. The boxplot shows the N50 distribution from the 50 repetitions at final sampling coverage at 0.025. Note the logarithmic x-axis.

Supplementary Table 9. Summarized results of the CEGMA analysis.

Statistics of the completeness of the genome based on 248 CEGs					
	#Prots	%Completeness	#Total	Average	%Ortho
Complete	152	61.29	165	1.09	8.55
Group 1	36	54.55	40	1.11	11.11
Group 2	36	64.29	40	1.11	11.11
Group 3	33	54.10	36	1.09	9.09
Group 4	47	72.31	49	1.04	4.26
Partial	233	93.95	328	1.41	32.62
Group 1	58	87.88	72	1.24	22.41
Group 2	54	96.43	82	1.52	37.04
Group 3	58	95.08	84	1.45	41.38
Group 4	63	96.92	90	1.43	30.16
These results are based on the set of genes selected by Genis Parra					

Key:
Prots = number of 248 ultra-conserved CEGs present in genome
%Completeness = percentage of 248 ultra-conserved CEGs present
Total = total number of CEGs present including putative orthologs
Average = average number of orthologs per CEG
%Ortho = percentage of detected CEGS that have more than 1 ortholog

Supplementary Note 13. Annotation

For annotation of the assembly we used the MAKER2 2.31.8 pipeline (Cantarel et al. 2008; Holt & Yandell 2011) combined with MPICH2 (http://www.mpich.org/) in three iterations combined with retraining of the species model in between.

Firstly an Augustus species model was computed on the Augustus webserver (Stanke et al. 2004; http://bioinf.uni-greifswald.de/webaugustus/training/create). The assembly, the species own ESTs and annotations in gff format from BUSCO 1.2 (Simão et al. 2015) were used as input. BUSCO was run on the assembly using the metazoan dataset together with the option --long apart from standard parameters. From a CEGMA 2.5 (Parra et al. 2007) run on the assembly a SNAP 2006-07-28 (Korf 2004) model was built using the script cegma2zff from the MAKER2 distribution and the SNAP scripts fathom (fathom genome.ann genome.dna -categorize 1000 && fathom -export 1000 -plus uni.ann uni.dna), forge (export.ann export.dna) and hmm-assembler.pl. A Genemark (GeneMark-ES suite 4.32; Lomsadze et al. 2005) model was built from a self-training (--ES) on the assembly.

The assembly, the Augustus species model, the ESTs, the ESTs from the other *Radix* species as alternative ESTs, the complete Swiss-Prot database (Accessed May 23rd 2016), the custom repeat library and the HMM models from SNAP and GeneMark were used as input for the first MAKER iteration. The options est2genome and protein2genome were switched off. Furthermore the minimum protein length being reported was set to 10 amino acids.

After the first iteration the gff file for the whole assembly was extracted using the MAKER gff3_merge, converted with maker2zff and a new HMM model was built for SNAP the same way as above. The Augustus species model was retrained locally by first converting with the SNAP script zff2gff3.pl (zff2gff3.pl genome.ann | perl -plne 's/\t(\S+)$/\t\.\t$1/') and second the autoAug.pl script from Augustus 3.2.2 (Stanke et al. 2006). The input for the autoAug.pl was the draft genome

assembly, the trained Augustus species model, the ESTs and the gff3 file created from the first MAKER iteration. Apart from standard parameters -v --useexisting were used.

For the second MAKER iteration the SNAP HMM model from CEGMA was exchanged to that created from the output of the first iteration as well as the updated Augustus species model were used as input. The minimum protein length was raised to 30 amino acids. Afterwards a retraining as above and a third MAKER iteration was realized.

The annotation pipeline resulted in 17,338 protein coding genes from lengths between 31 AA and 9,660 AA with a median of 332 AA, containing 7,968,643 AA in total. Gene lengths are between 141 and 127,541 bp with a mean of 11,570 bp and sum up to a total of 201 Mb which corresponds to 12.5% of the estimated genome size of 1.6 Gb. These genes contain 147,195 exons with an average of 8.5 exons per gene. Exonlengths reach from 3 bp to 10,740 bp with a mean of 171.9 bp and sum up to a total of 25 Mb which corresponds to 1.6% of estimated genome size of 1.6 Gb.

Literature cited

Alexa A, Rahnenfuhrer J. 2016. topGO: Enrichment Analysis for Gene Ontology. R package version 2.26.0.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Andrews S, Fiers M, Wingett S. 2015. FastQ Screen. http://www.bioinformatics.babraham.ac.uk/projects/fastq_screen.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob. DNA. 6:11. doi: 10.1186/s13100-015-0041-9.

Bateman A et al. 2015. UniProt: A hub for protein information. Nucleic Acids Res. 43:D204–D212. doi: 10.1093/nar/gku989.

Boetzer M et al. 2012. Toward almost closed genomes with GapFiller. Genome Biol. 13:R56. doi: 10.1186/gb-2012-13-6-r56.

Camacho C et al. 2009. BLAST plus: architecture and applications. BMC Bioinformatics. 10:1. doi: Artn 421\nDoi 10.1186/1471-2105-10-421.

Cantarel BL et al. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188–196. doi: 10.1101/gr.6743907.

Chang Z et al. 2015. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 16:1–10. doi: 10.1186/s13059-015-0596-2.

Chevreux B, Wetter T, Suhai S. 1999. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Comput. Sci. Biol. Proc. Ger. Conf. Bioinforma. '99, GCB, Hann. Ger. 45–56. doi: 10.1.1.23/7465.

Doležel J, Doleželová M, Novák FJ. 1994. Flow cytometric estimation of nuclear DNA amount in diploid bananas (Musa acuminata and M. balbisiana). Biol. Plant. 36:351–357. doi: 10.1007/BF02920930.

Doležel J, Sgorbati S, Lucretti S. 1992. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. Physiol. Plant. 85:625–631. doi: 10.1111/j.1399-3054.1992.tb04764.x.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157. doi: 10.1186/s13059-015-0721-2.

Feldmeyer B, Greshake B, Funke E, Ebersberger I, Pfenninger M. 2015. Positive selection in development and growth rate regulation genes involved in species divergence of the genus Radix. BMC Evol. Biol. 15:164. doi: 10.1186/s12862-015-0434-x.

Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M. 2011. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics. 12:317. doi: 10.1186/1471-2164-12-317.

Goubert C et al. 2015. De novo assembly and annotation of the Asian tiger mosquito (Aedesalbopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol. Evol. 7:1192–1205. doi: 10.1093/gbe/evv050.

Grabherr MG et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–52. doi: 10.1038/nbt.1883.

Gregory TR. 2016. Animal Genome Size Database.

Greilhuber J, Doležel J, Lysák MA, Bennett MD. 2005. The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. Ann. Bot. 95:255–260. doi: 10.1093/aob/mci019.

Holt C, Yandell M. 2011. MAKER2 : an annotation pipeline and genome- database management tool for second- generation genome projects. BMC Bioinformatics. 12:491. doi: 10.1186/1471-2105-12-491.

Kajitani R et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24:1384–1395. doi: 10.1101/gr.170720.113.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics. 5:59. doi: 10.1186/1471-2105-5-59.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv Prepr. doi: arXiv:1303.3997 [q-bio.GN].
Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079. doi: 10.1093/bioinformatics/btp352.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 33:6494–6506. doi: 10.1093/nar/gki937.
Okonechnikov K, Conesa A, García-alcalde F. 2015. Qualimap 2 : advanced multi-sample quality control for high- throughput sequencing data. Bioinformatics. 1–3.
Otto F. 1990. DAPI Staining of Fixed Cells for High-Resolution Flow Cytometly of Nuclear DNA. Methods Cell Biol. 33:105–110. doi: 10.1016/S0091-679X(08)60516-6.
Parra G, Bradnam K, Korf I. 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23:1061–1067. doi: 10.1093/bioinformatics/btm071.
Patel S, Schell T, Eifert C, Feldmeyer B, Pfenninger M. 2015. Characterizing a hybrid zone between a cryptic species pair of freshwater snails. Mol. Ecol. 24:643–655. doi: 10.1111/mec.13049.
Quevillon E et al. 2005. InterProScan: Protein domains identifier. Nucleic Acids Res. 33:116–120. doi: 10.1093/nar/gki442.
Schmieder R, Edwards R. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 6. doi: 10.1371/journal.pone.0017288.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.
Simit A, Hubley R. 2015. RepeatModeler Open-1.0.
Stanke M, Steinkamp R, Waack S, Morgenstern B. 2004. AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Res. 32:309–312. doi: 10.1093/nar/gkh379.
Stanke M, Tzvetkova A, Morgenstern B. 2006. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol. 7 Suppl 1:S11.1-8. doi: 10.1186/gb-2006-7-s1-s11.
Tills O, Truebano M, Rundle S. 2015. An embryonic transcriptome of the pulmonate snail Radix balthica. Mar. Genomics. 24:259–260. doi: 10.1016/j.margen.2015.07.014.
Zdobnov EM, Apweiler R. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 17:847–848. doi: 10.1093/bioinformatics/17.9.847.