PeerJ Preprints NOT PEER-REVIEWED

A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel *Limnoperna fortunei*

Authors: Marcela Uliano-Silva, Francesco Dondero, Thomas D. Otto, Igor Costa, Nicholas Costa Barroso Lima, Juliana Alves Americo, Camila Mazzoni, Francisco Prosdocimi, Mauro de Freitas Rebelo

ABSTRACT

Background: For more than 25 years, the golden mussel *Limnoperna fortunei* has aggressively invaded South American freshwaters, having travelled more than 5,000 km upstream across five countries. Along the way, the golden mussel has outcompeted native species and economically harmed aquaculture, hydroelectric powers, and ship transit. We have sequenced the complete genome of the golden mussel to understand the molecular basis of its invasiveness and search for ways to control it. **Findings:** We assembled the 1.6 Gb genome into 20548 scaffolds with an N50 length of 312 Kb using a hybrid and hierarchical assembly strategy from short and long DNA reads and transcriptomes. A total of 60717 coding genes were inferred from a customized transcriptome-trained AUGUSTUS run. We also compared predicted protein sets with those of complete molluscan genomes, revealing an exacerbation of protein-binding domains in *L. fortunei*. **Conclusions:** We built one of the best bivalve genome assemblies available using a cost-effective approach using Illumina pair-end, mate pair, and PacBio long reads. We expect that the continuous and careful annotation of *L. fortunei*'s genome will contribute to the investigation of bivalve genetics, evolution, and invasiveness, as well as to the development of biotechnological tools for aquatic pest control.

DATA DESCRIPTION

The golden mussel *Limnoperna fortunei* is an Asian bivalve that arrived in the southern part of South America about 25 years ago [1]. Since then, it has moved ~5,000 km, invading upstream continental waters and reaching northern parts of the continent [2] leaving behind a track of great economic impact and environmental degradation [3]. The latest infestation was reported in 2016 in the São Francisco River, one of the main rivers in the Northeast of Brazil, with a 2,700 km riverbed that provides water to more than 14 million people. At Paulo Afonso, one of the main hydroelectric power plants in the São Francisco River, maintenance due to clogging of pipelines and corrosion caused by the golden mussel is estimated to cost U$ 700,000 per year (*personal communication, Mizael Gusmã, Chief Maintenance Engineer for Centrais Hidrelétricas do São Francisco – CHESF).*

A recent review has shown that, before arriving in South America, *L. fortunei* was already an invader in China. Originally from the Pearl River Basin, the golden mussel has traveled 1,500 km into the Yang Tse and the Yellow River basins, being limited further north only by the extreme natural barriers of Northern China [4]. Today, *L. fortunei* is found in the Paraguaizinho River, located only 150 km from the Teles-Pires River that belongs to the Alto Tapajós River Basin and is the first to directly connect with the Amazon River Basin [5]. Due to its fast dispersion rates, it is very likely that *L. fortunei* will reach the Amazon River Basin in the near future.

The reason why some bivalves, such as *L. fortunei*, *Dreissena polymorpha*, and *Corbicula fluminea*, are aggressive invaders is not fully understood. These bivalves present characteristics such as (i) tolerance to a wide range of environmental variables, (ii) short life span, (iii) early sexual maturation, and (iv) high reproductive rates that allow them to reach densities as high as 150,000 ind.m^{-2} over a year [6,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2995v1 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

7] that may explain the aggressive behavior. On the other hand, these traits are not exclusive to invasive bivalves and do not explain how they outcompete native species and disperse so widely.

To the best of our knowledge, there are no reports of successful strategies to control the expansion of mussel invasion in industrial facilities. Bivalves can sense chemicals in the water and close their valves as a defensive response [8], making them tolerant to a wide range of chemical substances, including strong oxidants like chlorine [9]. Microencapsulated chemicals have shown better results in controlling mussel populations in closed environments [9, 10] but it is unlikely they would work in the wild. Currently, there is no effective and efficient approach to control the invasion by *L. fortunei*.

The genome sequence is one of the most relevant and informative descriptions of species biology. The genetic substrate of invasive populations, upon which natural selection operates, can be of primary importance to understand and control a biological invader [11].

Here we present the first complete genome dataset for the invasive bivalve *Limnoperna fortunei,* assembled from short and long DNA reads and using a hybrid and hierarchical assembly strategy. This high-quality reference genome represents a substantial resource for further studies of genetics and evolution of mussels, as well as for the development of new tools for plague control.

Genome sequencing in short Illumina and long PacBio reads

Limnoperna fortunei mussels were collected from the Jacui River, Porto Alegre, Rio Grande do Sul, Brazil (29°59′29.3″S 51°16′24.0″W). Voucher specimens were housed at the zoological collection (specimen number: 19643) of the Biology Institute at the Universidade Federal do Rio de Janeiro, Brazil. For the genome assembly, a total of 3 individuals were sampled for DNA extraction from gills. DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) to prepare libraries for Illumina Nextera paired-end reads, with ~180bp and ~500bp of insert size, (ii) Illumina Nextera mate-pair reads with insert sizes from 3 to 15 Kb, and (iii) Pacific Biosciences long reads (**Table 1**). Illumina libraries were sequenced respectively in a HiScanSQ or HiSeq 1500 machine, and Pacific Biosciences reads were produced with the P4C6 chemistry and sequenced in 10 SMRT Cells. All Illumina reads were submitted to quality analysis with FastQC followed by trimming with Trimmomatic [12]. Pacific Biosciences adaptor-free subreads sequences were used as input data for the genome assembly.

PeerJ Preprints NOT PEER-REVIEWED

Table 1 - DNA reads produced for *L. fortunei* genome assembly

Library technology			**Raw data**		**Trimmed Data***	
	Reads insert size	**Pairs**	**Number of reads**	**Number of bases**	**Number of reads**	**Number of bases**
Illumina Nextera	Paired end – 180 bp	R1	209542721	21060365702	209036571	21001101404
		R2	209542721	21049308698	209036571	20991650008
	Paired end – 500 bp	R1	153948902	15472966961	153482290	15423123500
		R2	153948902	15462883157	153482290	15414813589
	Mate pair 3-12 Kb	R1	178392944	18017687344	58157933	5822572152
		R2	178392944	18017687344	58157933	5811310412
Pacific Biosciences	P4C - 10/SMTRC	Subreads	1663730	11171487485		

*trimmomatic parameters for Illumina reads - ILLUMINACLIP:NexteraPE-PE.fa:2:30:10 SLIDINGWINDOW:4:2 LEADING:10 TRAILING:10 CROP:101 HEADCROP:0 MINLEN:80

For transcriptome sequencing, RNA was sampled from four tissues (gills, adductor muscle, digestive gland, and foot) of three different golden mussel specimens. RNA was extracted using NEXTflex Rapid Directional RNA-Seq Kit (Bioo Scientifics, TX, USA) and 12 barcodes from NEXTflex Barcodes compatible with Illumina NexSeq Machine. Resulting reads (**Supplementary Table S1**) were submitted to FastQC quality analysis and trimmed with Trimmomatic [12] for all NEXTflex adaptors and barcodes. A total of 3 sets of *de novo* assembled transcriptomes were generated using Trinity **(Table 2**); one set for each specimen was a pool of the 4 tissue samples to avoid assembly bias due to intraspecific polymorphism [13].

Table 2 - Trinity assembled transcripts used in the assembly and annotation of *L. fortunei* genome

Sample	**Pooled tissues**	**Number of reads prior assembly**	**Number of Trinity Transcripts**	**Number of Trinity Genes**	**Average Contig Length**	**GC%**
Mussel 1	Gills, mantle, digestive gland, foot	406589144	433197	303172	854	34
Mussel 2	Gills, mantle, digestive gland, foot	376577660	435054	298117	824	34
Mussel 3	Gills, mantle, digestive gland, foot	334316116	499392	351649	844	34

Genome assembly using a hybrid and hierarchical strategy

The Jellyfish software [14] was used to count and determine the distribution frequency of lengths 25 and 31 k-mers (**Figure 1**) for the Illumina DNA paired-end and mate-pair reads (**Table 1**). Genome size was estimated using the 25 k-mer distribution plot as total k-mer number and then subtracting erroneous reads (starting k-mer counts from 12 times coverage), to further divide by the homozygous coverage-peak depth (45 times coverage), as performed by Li *et al.* (2010) [15]. A double-peak k-mer distribution was used as evidence of genome diploidy (**Figure 1**). The genome size of *L. fortunei* was estimated to be 1.6 Gb.

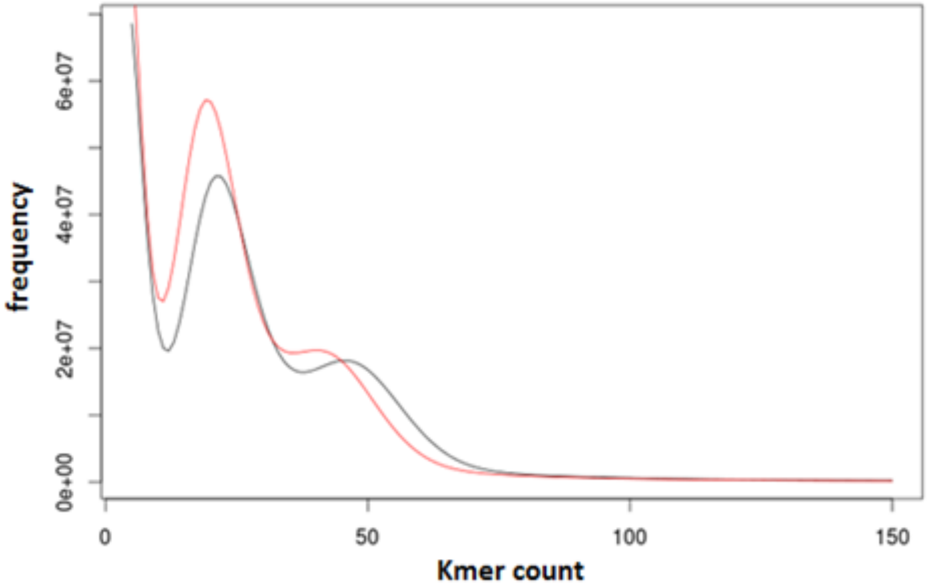

Figure 1: K-mer distribution of *Limnoperna fortunei* Illumina DNA reads (Table 1).

Initially, we attempted to assemble the golden mussel genome using only short Illumina reads of different insert sizes (paired-end and mate-pairs, Table 1) using traditional *de novo* assembly software such as ALLPATHS [16], SOAPdenovo [17], and Masurca [18]. All these attempts resulted in very fragmented genome drafts, with an N50 no higher than 5 Kb and a total of 4 million scaffolds. To reduce fragmentation, we further sequenced additional long reads (10 PacBio SMTR Cells, Table 1) and performed a hybrid and hierarchical *de novo* assembly described below and depicted in **Figure 2**.

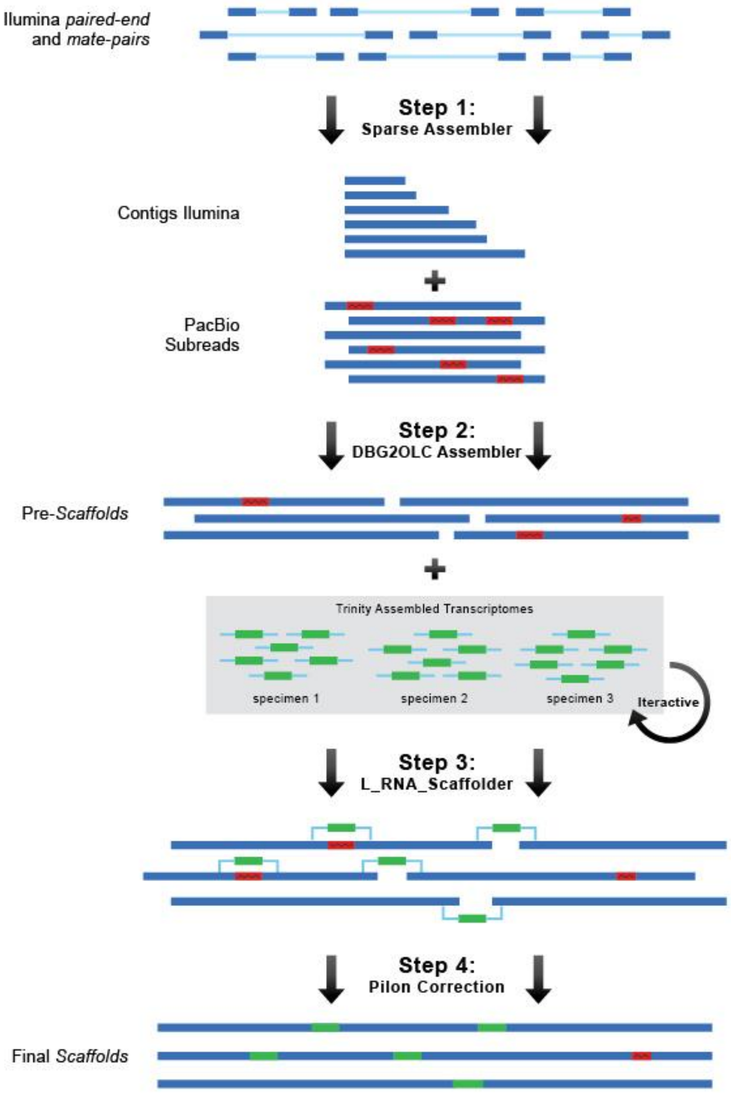

Figure 2: **Hierarchical assembly strategy employed for the golden mussel genome assembly.** Trimmed Illumina reads were assembled to the level of contigs with Sparse Assembler algorithm **(Step 1)**. Then, Illumina contigs and PacBio reads were used to build scaffolds with DBG2OLC assembler, that anchors Illumina contigs to erroneous PacBio subreads, correcting them and building longer scaffolds **(Step 2)**, followed by transcriptome joining scaffolds using L_RNA_scaffolder **(Step 3)**. Final scaffolds were corrected by re-aligning all Illumina DNA and RNA-seq reads back to them and calling consensus with Pilon software **(Step 4)**. In bold is bioinformatics software used in each step. Red blocks indicate PacBio errors, which are represented by insertions and/or deletions found in approximately 12% of PacBio subreads.

First, (i) trimmed paired-end and mate-pair DNA Illumina reads (**Table 1**) were assembled into contigs using the software Sparse Assembler [19] with parameters *LD 0 NodeCovTh 1 EdgeCovTh 0 k 31 g 15 PathCovTh 100 GS 1800000000*. Next, (ii) the resulting contigs were assembled into scaffolds using Pacific Biosciences long subreads data and the PacBio-correction-free assembly algorithm DBG2OLC [20] with parameters *LD1 0 k 17 KmerCovTh 10 MinOverlap 20 AdaptiveTh 0.01*. Finally, (iii) resulting scaffolds were submitted to 6 iterative runs of the program L_RNA_Scaffolder [21] that uses exon-distance information from *de novo* assembled transcripts (**Table 2**) to fill gaps and connect scaffolds whenever appropriate. At the end, (iv) the final genome scaffolds were corrected for Illumina and Pacific Biosciences sequencing errors with the software PILON [22]: all DNA and RNA short Illumina reads were re-aligned back to the genome with BWA aligner [23] and resulting sam files were BAM-converted, sorted, and indexed with samtools package [24]. Pilon [22] identifies INDELS and mismatches by coverage of reads and yields a final corrected genome draft. Pilon was run with parameters *--diploid –duplicates.*

The final genome was assembled in 20,548 scaffolds, with an N50 of 312 Kb and a total assembly length of 1.6 Gb (**Table 3**).

An initial quality assessment revealed that 91% of all Illumina reads used to construct the scaffolds mapped back to the final draft. The golden mussel genome presents 73% of all conserved core eukaryotic genes (CEGMA) and, compared to the published mollusk bivalve reviewed by Murgarella *et al.* (2016) [25], represents one of the best assemblies of molluscan genomes available so far **(Table 4)**. In fact, the assembly of the *L. fortunei* genome presented here exhibited a slightest lower N50 and higher scaffold number than the oyster *C. gigas* genome, even though the *L. fortunei* genome is 3 times the size of *C. gigas*. Based on these two parameters, and not taking into consideration genomes sequenced by the Sanger method (**Table 4**), our assembly surpasses the average bivalve genome assembly and may provide a robust dataset for the scientific community (**Table 4**).

Table 3: Assembly statistics for *Limnoperna fortunei*'s genome

Parameter	**Value**
Estimated genome size by k-mer analysis	1.6 Gb
Total size of assembled genome	1.673 Gb
Number of scaffolds	20548
Number of contigs	61093
Scaffold N50	312 Kb
Maximum scaffold length	2.72 Mb
Percentage of genome in scaffolds > 50 Kb	82,55%
Masked percentage of total genome	33 %

The main challenge related to assembling bivalve genomes lies in the high heterozygosity and amount of repetitive elements these organisms present: (i) the *Crassostrea gigas* genome was estimated to have a heterozygosity rate 2.3% higher than other animal genomes [26], and (ii) repetitive elements correspond to at least 30% of the genomes of all studied bivalves so far (**Table 3**) [25, 26, 27]. Also, retroelements might still be active in some species such as *L. fortunei* (refer to the retroelements-related section of this paper) and *C. gigas* [26], allowing genome rearrangements that may be obstacles for genome assembly. For this reason, bivalve genome projects relying only on short Illumina reads are likely to present fragmented initial drafts [25, 27]. PacBio long reads allowed us to move on from that stage, increasing N50 to 32 Kb and reducing scaffolds to 61102, using the DBG2OLC [20] assembler. Finally, interactive runs of L_RNA_scaffolder [21] using the transcriptomes (**Table 2**) rendered the final result of N50 312 Kb in 20548 scaffolds. Thus, our assembly strategy of Illumina contigs, low coverage of PacBio reads, transcriptome and Illumina re-mapping for final correction (**Figure 2**) represents an option for cost-efficient assembly of highly heterozygous genomes of nonmodel species such as bivalves.

PeerJ Preprints NOT PEER-REVIEWED

Table 4: **Comparison of genome assembly statistics for molluscan genomes**

	Mytillus galloprovincialis	*Crassostrea gigas*	*Pinctada fucata*	*Lottia gigantea*	*Aplysia california*	*Limnoperna fortunei*
Estimated genome size	1,600 Mb	545 Mb	1150 Mb	359,5 Mb	1,800 Mb	**1,600 Mb**
Number of scaffolds	1,746,447	11,969	800,982	4,475	8,766	**20,548**
Total size of scaffolds	1,599,211,957	558,601,156	1,413,178,538	359,512,207	715,791,924	**1,673,125,894**
Total scaffold length as percentage of known genome	100.0%	102.5%	122.9%	100.0%	39.8%	**104.6%**
Longest scaffold	67,529	1,964,558	698,791	9,386,848	1,784,514	**2,720,304**
Shortest scaffold	100	100	100	1000	5001	**558**
Number of scaffolds > 500 nt	676,492 (38.7%)	6,484 (54.2%2)	323,19 (40.4%)	4,475 (100%)	8,766 (100.0%)	**20,548 (100%)**
Number of scaffolds > 1 K nt	393,685 (22.5%)	5,788 (48.4%)	142,882 (17.8%)	4,471 (99.9%)	8,766 (100.0%)	**20,547 (100%)**
Number of scaffolds > 10 K nt	12,859 (0.7%)	3,172 (26.5%)	27,367 (3.4%)	1,318 (29.5%)	5,269 (23.7%)	**18,146 (88.3%)**
Number of scaffolds > 100 K nt	0 (0.0%)	1,353 (11.3%)	629 (0.1%)	291 (6.5%)	2,079 (23.7%)	**3,722 (18.1%)**
Number of scaffolds > 1 M nt	0 (0.0%)	60 (0.5%)	0 (0.0%)	98 (2.2%)	27 (0.3%)	**95 (0.5%)**
Mean scaffold size	916	46,671	1,764	80,338	81,655	**81,425**
Median scaffold size	258	824	402	3,622	13,763	**22,134**
N50 scaffold length	2,651	401,319	14,455	1,870,055	264,327	**312,02**
Percentage of assembly in scaffolded contigs	18.5%	95.7%	75.4%	99.0%	93.9%	**84.1%**
Percentage of assembly in unscaffolded contigs	81.5%	4.3%	24.6%	1.0%	6.1%	**15.9%**
Average number of contigs per scaffold	1.1	2.8	1.3	4.1	7.4	**2.9**
Sequencing coverage	32X	155X	40X	8,87X	11X	**60X**
Sequencing Technology	Illumina	Illumina	454 + Illumina	Sanger	Sanger	**Illumina + PacBio**

Around 10% of repetitive elements are transposons

Initial masking of *L. fortunei* genome was done using RepeatMasker [28] program with parameter *-species bivalves* and masked 3.4% of the total genome. This content was much lower than the masked portion of other molluscan genomes: 34% in *C. gigas* [26] and 36% in *M. galloprovincialis* [25], suggesting that the fast evolution of interspersed elements limits the use of repeat libraries from divergent taxa [29]. Thus, we generated a *de novo* repeat library for *L. fortunei* using the program RepeatModeler [30] and its integrated tools (RECON [31], TRF [32], and RepeatScout [33]). This *de novo* repeat library was the input to RepeatMasker together with the first masked genome draft of *L. fortunei*, and resulted in a final masking of 33.4% of the genome. Even though more than 90% of the repeats were not classified by RepeatMasker (**Supplementary Table S2**), 8.85% of the repeats were classified as LINEs, Class I transposable elements. In addition, large numbers of reverse-transcriptases (824 counts, Pfam PF00078), transposases (177 counts, Pfam PF01498), and integrases (501 counts, Pfam PF00665) were detected; over 98% of these had detectable transcripts.

More than 30,000 sequences identified by gene prediction and automated annotation

To annotate the golden mussel genome, we sequenced a number of transcriptomes (**Table S1**), *de novo* assembled (**Table 2**) and aligned these genomes to the genome scaffolds, and created gene models with the PASA pipeline [34]. These models were used to train and run the *ab initio* gene predictor AUGUSTUS [35] (**Supplementary Figure S1**). The complete gene models yielded by PASA [34] were BLASTed (e-value 1e-20) against the Uniprot database and those with 90% or more of their sequences showing in the BLAST hit alignment were considered for further analysis. Next, all the necessary filters to run an AUGUSTUS [35] personalized training were performed: (i) only gene models with more than 3 exons were maintained, (ii) sequences with 90% or more overlap were withdrawn and only the longest sequences were retained, and (iii) only gene models free of repeat regions, as indicated by BLASTN similarity searches with *de novo* library of repeats, were maintained. These curated data yielded a final set of 1,721 gene models on which AUGUSTUS [35] was trained in order to predict genes in the genome using the default AUGUSTUS [35] parameters. Once the gene models were predicted, a final step was performed by using the PASA pipeline [34] once again in the *update* mode (parameters -c -A -g -t). This final step compared the 55,638 gene models predicted by AUGUSTUS [35] with the 40,780 initial transcript-based gene-structure models from PASA [34] to generate the final set of 60,717 gene models for *L. fortunei.* Of those, 58% had transcriptional evidence based on RNA Illumina reads (**Table S1**) re-mapping, and 67% were annotated by homology searches against Uniprot or NCBI NR (**Table 5**).

Table 5: **Summary of gene annotation against various databases for *L. fortunei* whole genome-predicted genes**

Total number of genes	60,717
Total number of exons	220,058
Total number of proteins	60,717
Average protein size	304 aa
Number of protein BLAST hits* with Uniprot	26,198
Number of protein BLAST hits* with NR NCBI (no hits with Uniprot)	14,810
Number of protein HMMER hits* with Pfam.A	24,513
Number with proteins with KO assigned by KEGG	8,387
Number of proteins with BLAST hits* with EggNOG	36,868

*all considered hits had a maximum e-value of 1e-05

Protein clustering indicates evolutionary proximity among mollusks species

Orthology relationships were assigned using reciprocal best BLAST and OrthoMCL software (version 1.4) [36] between *L. fortunei* proteins and the total protein set predicted for three other mollusks: the pacific oyster *Crassostrea gigas*, the pearl oyster *Pinctada fucata*, and the gastropod *Lottia gigantea* (see **Supplementary Table S3** for detailed information on the comparative data) (**Figure 3A**). Around 70% of all protein sequences from all four species cluster in at least one orthologous group, demonstrating the evolutionary proximity among these mollusks species. A total of 6,220 orthologous groups are shared among all the species analyzed. Of all orthologous groups, 2,391 groups are composed of single-copy orthologs containing one representative protein sequence of each species. These sequences were used to reconstruct a phylogeny: the 2,391 single-copy orthologous sequences were concatenated and aligned with CLUSTALW [37] with a resulting alignment of 1,676,575 sites in length. The phylogenetic inferred tree (maximum likelihood, JTT model) shows a longer branch for *L. fortunei* (**Figure 3B**).

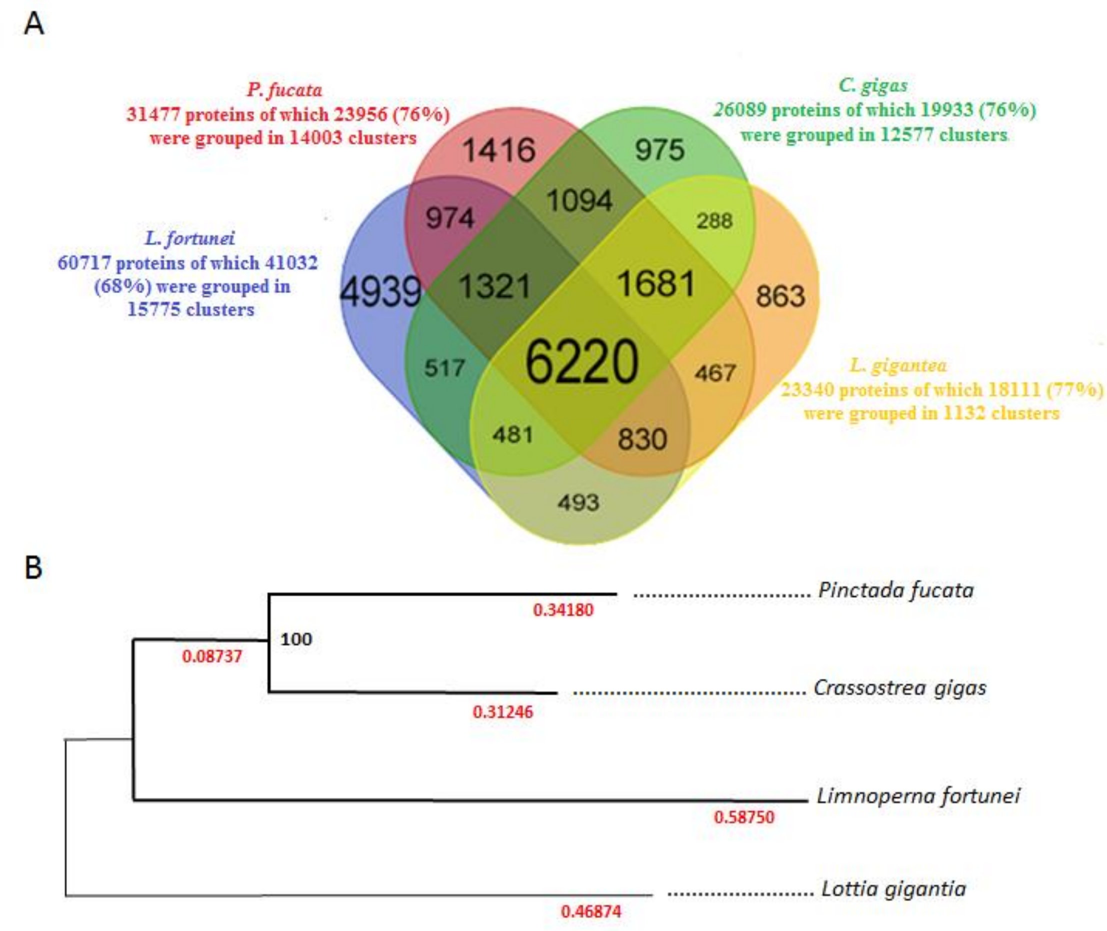

Figure 3A: Orthology assigned with OrthoMCL for the total set of proteins predicted from four Molluscan genome projects. **B**: Phylogeny of the concatenated data set using 2391 single-copy orthologs extracted from four molluscan genomes. Maximum likelihood tree nodes with 100 bootstrap resampling. The gastropod *L. gigantea* was placed as an outgroup inside the bivalve tree.

Protein domain analysis shows expansion of binding domain in *L. fortunei.*

We performed a quantitative comparison of protein domains predicted from whole genome projects of five molluscan species (**Table S3**). The complete protein sets of *L. fortunei*, *C. gigas*, *P. fucata*, *L. gigantean*, and *M. galloprovincialis* (**Table S3**) were submitted to domain annotation using HMMER against Pfam-A database (e-value 1e-05). Protein expansions in *L. fortunei* were rendered using the normalized Pfam count value (average) obtained from the other four mollusks, according to Poisson model. Bonferroni correction ($p<0.05$) was applied for false discovery rate. Absolute frequencies of Pfam-assigned-domains were initially normalized by total count number of Pfam-assigned-domains found in *L. fortunei* to compensate for discrepancies in genome size and bias on protein sets described for each species.

For *L. fortunei*, the annotation against Pfam.A classified 40127 domains in 24513 gene models of which 107 were overrepresented in comparison with the other mollusks (**Figure 4A**). The 107 overrepresented domains were analyzed for functional enrichment using domain-centric Gene Ontology (37) (**Figure 4B**). The analysis shows a prominent expansion of binding domains in *L. fortunei*, such as Thrombospondin (TSP_1), Collagen, Immunoglobulins (Ig, I-set, C1-set, Ig_3), and Ankyrins (Ank_2, Ank_3, and Ank_4). These repeats have a variety of binding properties and are involved in cell-cell, protein-protein and receptor-ligand interactions driving evolutionary improvement of complex tissues and immune defense system in metazoans [38, 39, 40, 41, 42].

Of note is the high amount of Leucine Rich Repeats (LRR_8, LRR_5, and LRR_4) and Toll/interleukin-1 receptor homology domains (TIR/TIR-2), both belonging to Toll Like Receptors (TLRs). Blast analysis of *L. fortunei* gene models against Uniprot identified two types of TLRs: (i) 141 sequences with similarity to single cysteine clusters TLRs (scc) typical of vertebrates, and (ii) 29 sequence hits with the multiple cysteine cluster TLRs (mcc) typical of *Drosophila*. Phylogenetic analysis of all sequences (**Supplementary Figure S2**) shows evidence for TLRs clade separation in *L. fortunei*; the scc TLRs exhibit a higher degree of amino acid changes, higher molecular evolution, and diversification than the mcc TLRs.

Final considerations

Here we have described the first version of the golden mussel complete genome and its automated gene prediction. This genome contains valuable information for further evolutionary studies of bivalves and metazoa in general. Additionally, our team will further search for the presence of proteins of biotechnology interest such as the adhesive proteins produced by the foot gland that we have described elsewhere [43], or genes related to the reproductive system that have been shown to be very effective for invertebrate plague control [44]. The golden mussel genome and the predicted proteins are available for download and the scientific community is welcome to further curate the gene predictions.

As the golden mussel advances towards the Amazon river basin, the information provided in this study may be used to help developing biotechnological strategies that may control the expansion of this organism in both industrial facilities and open environment.

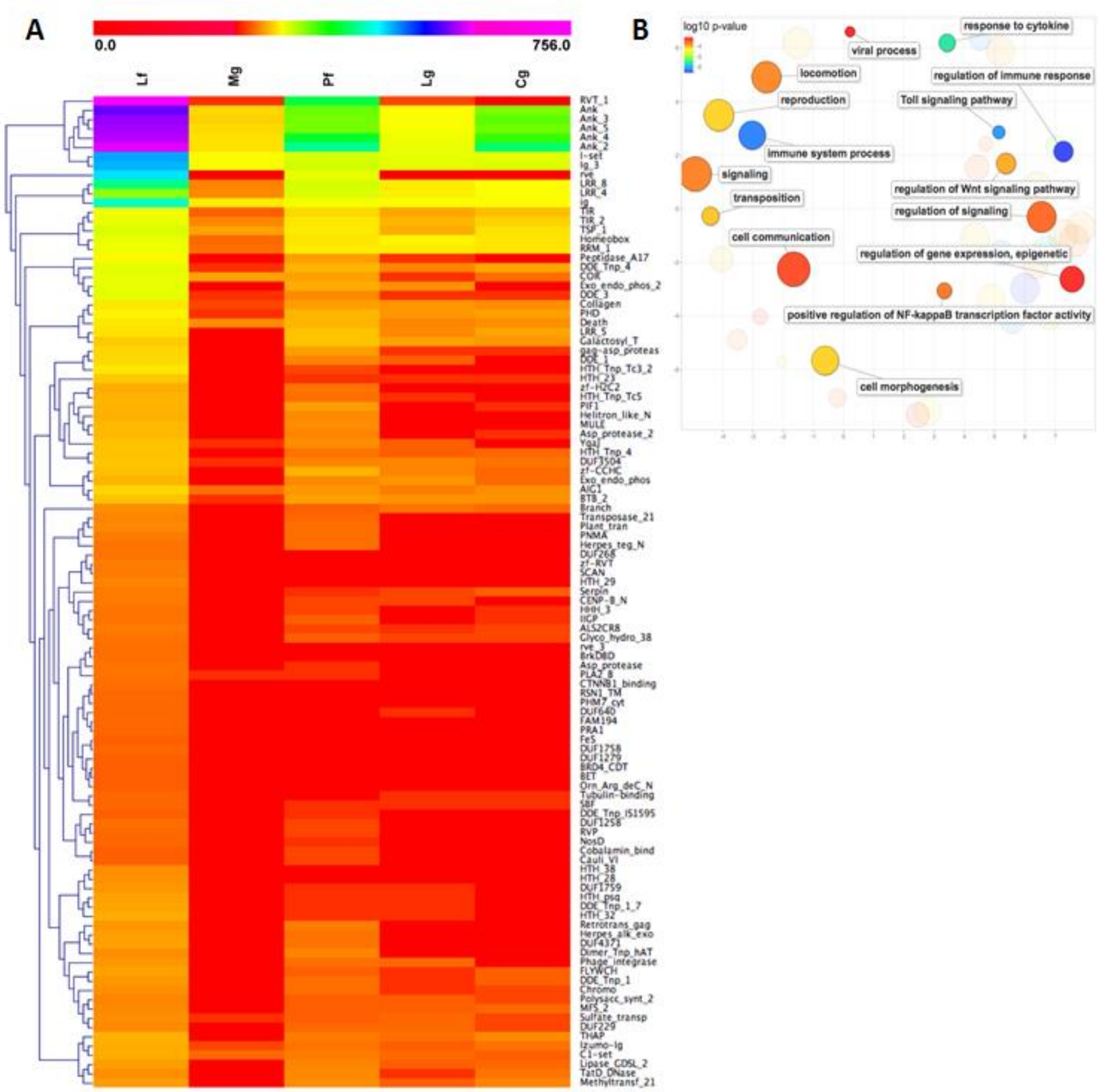

Figure 4: Gene expansions in the *L. fortunei* genome. A: PFAM hierarchical clustering, heatmap. Features were selected according to the Poisson cumulative distribution of each PFAM count in the golden mussel genome vs. the normalized average values found in the other four molluscan genomes (Bonferroni correction, $P < 0.05$). Legend: Lf, *L. fortunei*; Mg, *M. galloprovincialis*; Pf, *Pintada fucata*; Lg, *L. gigantea*; Cg, *C. gigas*. Colors depict normalized absolute counts. **B. Gene ontology analysis of enriched PFAMs, semantic scatter plot.** Shown are cluster representatives after redundancy reduction in a two-dimensional space applying multidimensional scaling to a matrix of semantic similarities of GO term. Color indicates the GO enrichment level (legend in upper left-hand corner); size indicates the relative frequency of each term in the UNIPROT database (larger bubbles represent less specific processes).

Additional files

Supplementary Table S1. RNA raw reads sequenced for 3 *L. fortunei* specimens, 4 tissues each.
Supplementary Table S2: RepeatMasker classification of repeats predicted in *L. fortunei* genome.
Supplementary Table S3: Details of the online availability of the data used for ortholog assignment and protein domain expansion analysis.
Supplementary Table S4: Fantasy names given to *L. fortunei* genes and proteins from the backers that have supported us through crowdfunding (www.catarse.me/genoma).
Supplementary Figure 1: Steps performed for the prediction and annotation of *L. fortunei* genome.
Supplementary Figure 2: Phylogenetic tree of Toll-like (TLRs) receptors found in *L. fortunei*

Competing interests

The authors declare that they have no competing interests.

Authors' contribution

Conceived and designed the experiments: MR, MU, TO, CM, FD. Performed the experiments: MU, JA. Analyzed the data: MU, TO, CM, FD, FP, NC, IC, MR. Contributed reagents/materials/analysis tools: MR, FP, CM. Wrote the paper: MU, FD, MR. All authors read and approved the final manuscript.

Funding

This work was supported by the Brazilian Government agencies CAPES (PVE 71/2013), FAPERJ APQ1 (2014), and FAPERJ/DFG (39/2014). Also, this work was funded through crowdfunding with the support of 346 people (www.catarse.me/genoma).

Acknowledgements

We thank Susan Mbedi and Kirsten Richter from BeGenDiv for RNA-Seq library preparation and sequencing. We thank Dr. Loris Bennett for IT support while performing bioinformatics analysis.
We especially want to thank the 346 backers that supported the sequencing of the golden mussel through crowdfunding, in a 2013 campaign that raised BRL 40k (www.catarse.me/genoma). We decided to give fantasy names to the genes and proteins that we found in the genome, to thank the backers for their support. The name list is available in **Supplementary Table S4**.

Ethics approval

Limnoperna fortunei specimens used for DNA extraction and sequencing were collected in the Jacuí River (29°59′29.3″S 51°16′24.0″W), southern Brazil. This bivalve is an exotic species in Brazil and is not characterized as an endangered or protected species.

References

1. Pastorino G, Darrigran G, et al.,. Limnoperna fortunei (Dunker, 1857) (Mytilidae), nuevo bivalvo invasor em águas Del Rio de la Plata. Neotropica. 1993;39:101–2.

2. Uliano-Silva M, Fernandes F da C, Holanda IBB, Rebelo MF. Invasive species as a threat to biodiversity: The golden mussel *Limnoperna fortunei* approaching the Amazon River basin. In: Exploring Themes Aquat Toxicol Alodi, S, editor. Research Signpost; 2013.

3. Boltovskoy D, Correa N. Ecosystem impacts of the invasive bivalve Limnoperna fortunei (golden mussel) in South America. Hydrobiologia. 2015;746(1):81–95.

4. Xu M. Distribution and Spread of Limnoperna fortunei in China. In: Limnoperna fortunei Boltovskoy D, editor. Cham: Springer International Publishing; 2015 p. 313–20.

5. Oliveira M, Hamilton S, Jacobi C. Forecasting the expansion of the invasive golden mussel Limnoperna fortunei in Brazilian and North American rivers based on its occurrence in the Paraguay River and Pantanal wetland of Brazil. Aquat Invasions. 2010;5(1):59–73.

6. Karatayev AY, Boltovskoy D, Padilla DK, Burlakova LE. The invasive bivalves Dreissena polymorpha and Limnoperna fortunei: parallels, contrasts, potential spread and invasion impacts. J Shellfish Res. 2007 1;26(1):205–13.

7. Orensanz JM (Lobo), Schwindt E, Pastorino G, Bortolus A, Casas G, Darrigran G, et al. No Longer The Pristine Confines of the World Ocean: A Survey of Exotic Marine Species in the Southwestern Atlantic. Biol Invasions. 2002 1;4(1–2):115–43.

8. Claudi R and Mackie GL. Practical manual for zebra mussel monitoring and control. Lewis Publishers, Boca. Raton, Florida, 1994. p 227

9. Calazans SHC, Americo JA, Fernandes F da C, Aldridge DC, Rebelo M de F. Assessment of toxicity of dissolved and microencapsulated biocides for control of the Golden Mussel Limnoperna fortunei. Mar Environ Res. 2013 91:104–8.

10. Aldridge DC, Elliott P, Moggridge G.D. Microencapsulated biobullets for the control of biofouling zebra mussels. Environ. Sci. Technol. 2006 40:975-979.

11. Cox GW. Alien species and evolution: the evolutionary ecology of exotic plants, animals, microbes, and interacting native species. Washington: Island Press; 2004. 377 p.

12. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics. 2014 1;170.

13. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 1;7(3):562–78.

14. Marçais G, Kingsford C. A Fast, Lock-free Approach for Efficient Parallel Counting of Occurrences of K-mers. Bioinformatics. 2011 Mar;27(6):764–770.

15. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010 Jan 21;463(7279):311–7.

16. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011 25;108(4):1513–8.

17. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience. 2012;1:18.

18. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013 1;29(21):2669–77.

19. Ye C, Ma Z, Cannon CH, Pop M, Yu DW. Exploiting sparseness in de novo genome assembly. BMC Bioinformatics. 2012;13(Suppl 6):S1.

20. Ye C, Hill CM, Wu S, Ruan J, Ma Z (Sam). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci Rep. 2016 30;6:31900.

21. Xue W, Li J-T, Zhu Y-P, Hou G-Y, Kong X-F, Kuang Y-Y, et al. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics. 2013;14:604.

22. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 2014 19;9(11):e112963.

23. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 15;25(14):1754–60.

24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 15;25(16):2078–9.

25. Murgarella M, Puiu D, Novoa B, Figueras A, Posada D, Canchaya C. A First Insight into the Genome of the Filter-Feeder Mussel Mytilus galloprovincialis. PLOS ONE. 2016 15;11(3):e0151561.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2995v1 | CC BY 4.0 Open Access | rec: 29 May 2017, publ: 29 May 2017

26. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 4;490(7418):49–54.

27. Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, et al. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA Res Int J Rapid Publ Rep Genes Genomes. 2012 19(2):117–30.

28. Smit AF., Hubley R, Green PJ. RepeatMasker Open-3.0. 1996 2010.

29. Fu H, Dooner HK. Intraspecific violation of genetic colinearity and its implications in maize. Proc Natl Acad Sci U S A. 2002 9;99(14):9573–8.

30. Smith AFA, Hubley R. RepeatModeler Open-1.0. [Internet]. 2014. Available from: http://www.repeatmasker.org

31. Bao Z, Eddy SR. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res. 2002 12(8):1269–76.

32. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999 1;27(2):573–80.

33. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005 1;21(Suppl 1):i351–8

34. Haas BJ, Delcher AL, Mount SM, Wortman JR, Jr RKS, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003 1;31(19):5654–66.

35. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinforma Oxf Engl. 2008 1;24(5):637–44.

36. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 2003 1;13(9):2178–89.

37. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994 11;22(22):4673–80.

38. Björklund ÅK, Ekman D, Elofsson A. Expansion of Protein Domain Repeats. PLoS Comput Biol. 2006;2(8):e114.

39. Rennemeier C, Hammerschmidt S, Niemann S, Inamura S, Zähringer U, Kehrel BE. Thrombospondin-1 promotes cellular adherence of gram-positive pathogens via recognition of peptidoglycan. FASEB J Off Publ Fed Am Soc Exp Biol. 2007 21(12):3118–32.

40. Schmucker D, Chen B. Dscam and DSCAM: complex genes in simple animals, complex animals yet simple genes. Genes Dev. 2009 15;23(2):147–56.

41. Pancer Z, Amemiya CT, Ehrhardt GRA, Ceitlin J, Larry Gartland G, Cooper MD. Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. Nature. 2004 Jul 8;430(6996):174–80.

42. Tucker RP. The thrombospondin type 1 repeat superfamily. Int J Biochem Cell Biol. 2004 36(6):969–74.

43. Uliano-Silva M, Americo JA, Brindeiro R, Dondero F, Prosdocimi F, Rebelo M de F. Gene discovery through transcriptome sequencing for the invasive mussel Limnoperna fortunei. PloS One. 2014;9(7):e102973

44. Hammond A, Galizi R, Kyrou K, Simoni A, Siniscalchi C, Katsanos D, et al. A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. Nat Biotechnol. 2015 7;34(1):78–83.