

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

<b>Manuscript Number:</b>	GIGA-D-17-00124R2									
<b>Full Title:</b>	A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel <i>Limnoperna fortunei</i>									
<b>Article Type:</b>	Data Note									
<b>Funding Information:</b>	<table border="1"> <tr> <td>CAPES (PVE (71/2013))</td> <td>Dr Mauro F Rebelo</td> </tr> <tr> <td>FAPERJ (APQ1 (2014))</td> <td>Dr Mauro F Rebelo</td> </tr> <tr> <td>FAPERJ/DFG (FAPERJ/DFG (39/2014))</td> <td>Dr Mauro F Rebelo</td> </tr> <tr> <td>Crowdfunding (www.catarse.me/genoma)</td> <td>Dr Mauro F Rebelo</td> </tr> </table>	CAPES (PVE (71/2013))	Dr Mauro F Rebelo	FAPERJ (APQ1 (2014))	Dr Mauro F Rebelo	FAPERJ/DFG (FAPERJ/DFG (39/2014))	Dr Mauro F Rebelo	Crowdfunding (www.catarse.me/genoma)	Dr Mauro F Rebelo	
CAPES (PVE (71/2013))	Dr Mauro F Rebelo									
FAPERJ (APQ1 (2014))	Dr Mauro F Rebelo									
FAPERJ/DFG (FAPERJ/DFG (39/2014))	Dr Mauro F Rebelo									
Crowdfunding (www.catarse.me/genoma)	Dr Mauro F Rebelo									
<b>Abstract:</b>	<p>Background: For more than 25 years, the golden mussel <i>Limnoperna fortunei</i> 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 <i>L. fortunei</i>. 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 <i>L. fortunei</i>'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.</p>									
<b>Corresponding Author:</b>	Marcela Uliano da Silva, Ph.D Universidade Federal do Rio de Janeiro Rio de Janeiro, RJ BRAZIL									
<b>Corresponding Author Secondary Information:</b>										
<b>Corresponding Author's Institution:</b>	Universidade Federal do Rio de Janeiro									
<b>Corresponding Author's Secondary Institution:</b>										
<b>First Author:</b>	Marcela Uliano da Silva, Ph.D									
<b>First Author Secondary Information:</b>										
<b>Order of Authors:</b>	<table border="1"> <tr> <td>Marcela Uliano da Silva, Ph.D</td> </tr> <tr> <td>Francesco Dondero, Ph.D</td> </tr> <tr> <td>Thomas D Otto, Ph.D</td> </tr> <tr> <td>Igor R Costa, Msc</td> </tr> <tr> <td>Nicholas CB Lima, Ph.D</td> </tr> <tr> <td>Juliana A Americo, Ph.D</td> </tr> <tr> <td>Camila J Mazzone, Ph.D</td> </tr> <tr> <td></td> </tr> </table>		Marcela Uliano da Silva, Ph.D	Francesco Dondero, Ph.D	Thomas D Otto, Ph.D	Igor R Costa, Msc	Nicholas CB Lima, Ph.D	Juliana A Americo, Ph.D	Camila J Mazzone, Ph.D	
Marcela Uliano da Silva, Ph.D										
Francesco Dondero, Ph.D										
Thomas D Otto, Ph.D										
Igor R Costa, Msc										
Nicholas CB Lima, Ph.D										
Juliana A Americo, Ph.D										
Camila J Mazzone, Ph.D										

	Francisco Prosdocimi, Ph.D
	Mauro F Rebelo, Ph.D
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>Reviewer reports:</p> <p>Reviewer #1: Thank you for providing an updated version of your manuscript and a detailed reply to all my previous concerns. The manuscript now appears to be nearly ready for publication. I still have a few comments, which could require some minor revisions, if the editor will deem them to be necessary.</p> <p>Overall, the main issue is the use of three different specimens for genome assembly, which is a non-standard procedure which is always preferable to avoid, especially in highly heterozygous genomes. However, given the circumstances, the strategy used is acceptable, since the "mixed" assembly certainly represents a major improvement compared to the assembly performed with Illumina reads only.</p> <p>With this respect however, the authors could expand a bit their thoughts around line 200, by briefly explaining this issue as a sort of "warning", which could be of help for future sequencing efforts on other bivalve species. In other words, the assembly statistics could have been even better with the preparation of mate-pair and PacBio libraries from the same specimen. Indeed, high heterozygosity might explain why the statistics of the <i>Limnoperna</i> genome are still lower than other genomes of similar size, but with lower heterozygosity rates and, probably, overall complexity (e.g. <i>P. yessoensis</i>), which have been assembled with Illumina reads only. Apart from SNPs and short indels, there is the possibility that particular regions of the genome present large scale rearrangements (i.e. CNVs, large indels and/or inversions), a phenomenon that has been already observed for <i>Ciona savignyii</i> and other species with large effective population size and broad dispersal of gametes by spawning (no need to report this in the manuscript though). Overall, I think this might partly explain the residual fragmentation of the genome, as the assembly will be complicated by reads originated from highly polymorphic regions across individuals.</p> <p>Response: We have added a comment about our use of DNA from different specimens and have advised that this is not the ideal way to go (lines: 203-212). However, it's good to note that some genome projects are unable to proceed otherwise. As examples, projects sequencing a rare sample from a threatened species, or from a species for which sample collection and access is difficult. So, it's good to note that, despite this extra difficulty, the use of hybrid approaches - especially with long reads - can now allow the genome assembly of various species even in these difficult scenarios.</p> <p>Lines 203-2012 are as follow: "... It's important to note that assembly statistics can perform better for genomes assembled with reads generated with DNA extracted from one unique individual. This, however, was not possible for <i>L. fortunei</i>'s genome, due to the high amount of high-quality-DNA necessary to produce Illumina mate-pair and PacBio long reads. In this study, the challenge of assembling the high polymorphic regions between haplotypes was enhanced by the difficulties of assembling reads originated from highly polymorphic regions across individuals. However, the golden mussel assembly presented here shows that the use of Illumina contigs, low coverage of PacBio long 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. "</p> <p>Reviewer 1: Thanks for including k-mer size in figure 1. However, for highly heterozygous genomes the use of shorter k-mers (17-20) is often appropriate for a better estimate of heterozygosity rates (the formula assumes the k-mers falling in the heterozygous peak differ from those of the homozygous peak just by 1 nucleotide, but this assumption might not be correct when long k-mers are used in highly heterozygous genomes). Please try to calculate the rate also with a shorter k-mer size (and plot it in figure 1, if necessary) and check whether the calculated heterozygosity rate changes significantly (it is possible that you are slightly underestimating it with the current k-mer size).</p>

	<p>Response: We have updated the heterozygosity rate estimated with kmer size of 17 in lines 153 and 190 of the new manuscript.</p> <p>Reviewer 1: Lines 277-282: please check the log file of your ProtTest analysis. The selection of the VT model (without a +G, or +F parameter) is somewhat odd. It is possible that your machine ran out of memory during the computation of the most complex models due to the large size of the input alignment, so that the LogL values could be computed just for the most simple models (such as VT). In any case the tree topology is exactly that one might expect, so the possible use of a different model will only have subtle effects.</p> <p>Response: ProtTest selected the VT+G+I+F model, and this was what we used in our phylogeny (line 286).</p> <p>Reviewer 1: Same % values are missing in table 5 for the newly added genomes. Also, check the comas to indicate thousands in all numbers.</p> <p>Response: the table was re-checked and corrected for %s and commas. Thank you.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p><b>Resources</b></p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes
<b>Availability of data and materials</b>	Yes

All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in [publicly available repositories](#) (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist?](#)

1  
2  
3  
4 **1 A hybrid-hierarchical genome assembly strategy to sequence the invasive**  
5  
6  
7 **2 golden mussel *Limnoperna fortunei***

9  
10 **3 Marcela Uliano-Silva<sup>1,2,3\*</sup>, Francesco Dondero<sup>4</sup>, Thomas Dan Otto<sup>5,6</sup>, Igor Costa<sup>7</sup>, Nicholas**  
11 **4 Costa Barroso Lima<sup>7,8</sup>, Juliana Alves Americo<sup>1</sup>, Camila Junqueira Mazzoni<sup>2,3</sup>, Francisco**  
12 **5 Prosdocimi<sup>7</sup>, Mauro de Freitas Rebelo<sup>1\*</sup>**  
13  
14 **6**

15 **7** Marcela Uliano-Silva: [marcela.uliano@gmail.com](mailto:marcela.uliano@gmail.com)

16 **8** Francesco Dondero: [francesco.dondero@uniupo.it](mailto:francesco.dondero@uniupo.it)

17 **9** Thomas D. Otto: [tdo@sanger.ac.uk](mailto:tdo@sanger.ac.uk)

18 **10** Igor Costa: [igor.bioinfo@gmail.com](mailto:igor.bioinfo@gmail.com)

19 **11** Nicholas Costa Barroso Lima: [ncblima@gmail.com](mailto:ncblima@gmail.com)

20 **12** Juliana Alves Americo: [juliana.americo@gmail.com](mailto:juliana.americo@gmail.com)

21 **13** Camila Mazzoni: [mazzoni@izw-berlin.de](mailto:mazzoni@izw-berlin.de)

22 **14** Francisco Prosdocimi: [prosdocimi@bioqmed.ufrj.br](mailto:prosdocimi@bioqmed.ufrj.br)

23 **15** Mauro de Freitas Rebelo: [mrebelo@biof.ufrj.br](mailto:mrebelo@biof.ufrj.br)

24 **16** Affiliations:

25  
26  
27  
28 **17** 1 Carlos Chagas Filho Biophysics Institute (IBCCF), Universidade Federal do Rio de Janeiro,  
29  
30  
31 **18** Rio de Janeiro, Brazil

32  
33 **19** 2 Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Berlin,  
34  
35  
36 **20** Germany

37  
38 **21** 3 Berlin Center for Genomics in Biodiversity Research, Berlin, Germany

39  
40 **22** 4 Department of Science and Technological Innovation (DiSIT), Università del Piemonte  
41  
42  
43 **23** Orientale Amedeo Avogadro, Vercelli-Novara-Alessandria, Italy

44  
45 **24** 5 Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK

46  
47  
48 **25** 6 Centre of Immunobiology, Institute of Infection, Immunity & Inflammation, College of  
49  
50  
51 **26** Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK.

52  
53 **27** 7 Leopoldo de Meis Biomedical Biochemistry Institute (IBqM), Universidade Federal do Rio de  
54  
55 **28** Janeiro, Rio de Janeiro, Brazil

56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 29 8 Bioinformatics Laboratory (LabInfo) of the National Laboratory for Scientific Computing,  
5  
6 30 Petrópolis, Rio de Janeiro, Brazil

7  
8  
9 31 \*Correspondence: [marcela.uliano@gmail.com](mailto:marcela.uliano@gmail.com); [mrebelo@biof.ufrj.br](mailto:mrebelo@biof.ufrj.br)

10  
11 32 **ABSTRACT**

12  
13  
14 33 **Background:** For more than 25 years, the golden mussel *Limnoperna fortunei* has aggressively  
15  
16 34 invaded South American freshwaters, having travelled more than 5,000 km upstream across five  
17  
18 35 countries. Along the way, the golden mussel has outcompeted native species and economically  
19  
20 36 harmed aquaculture, hydroelectric powers, and ship transit. We have sequenced the complete  
21  
22 37 genome of the golden mussel to understand the molecular basis of its invasiveness and search for  
23  
24 38 ways to control it.

25  
26  
27  
28 39 **Findings:** We assembled the 1.6 Gb genome into 20548 scaffolds with an N50 length of 312 Kb  
29  
30 40 using a hybrid and hierarchical assembly strategy from short and long DNA reads and  
31  
32 41 transcriptomes. A total of 60717 coding genes were inferred from a customized transcriptome-  
33  
34 42 trained AUGUSTUS run. We also compared predicted protein sets with those of complete  
35  
36 43 molluscan genomes, revealing an exacerbation of protein-binding domains in *L. fortunei*.

37  
38 44 **Conclusions:** We built one of the best bivalve genome assemblies available using a cost-  
39  
40 45 effective approach using Illumina pair-end, mate pair, and PacBio long reads. We expect that the  
41  
42 46 continuous and careful annotation of *L. fortunei*'s genome will contribute to the investigation of  
43  
44 47 bivalve genetics, evolution, and invasiveness, as well as to the development of biotechnological  
45  
46 48 tools for aquatic pest control.

47  
48  
49 49 **KEYWORDS:** Amazon; binding domain; bivalves; genomics; TLR; transposon.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 **51 DATA DESCRIPTION**

5  
6 **52** The golden mussel *Limnoperna fortunei* is an Asian bivalve that arrived in the southern  
7  
8  
9 **53** part of South America about 25 years ago [1]. Research suggests that *L. fortunei* was introduced  
10  
11  
12 **54** in South America through ballast water of ships coming from Hong Kong or Korea [2]. It was  
13  
14 **55** found for the first time in the estuary of the La Plata River in 1991 [1]. Since then, it has moved  
15  
16 **56** ~5,000 km, invading upstream continental waters and reaching northern parts of the continent [3]  
17  
18  
19 **57** leaving behind a track of great economic impact and environmental degradation [4]. The latest  
20  
21 **58** infestation was reported in 2016 in the São Francisco River, one of the main rivers in the  
22  
23  
24 **59** Northeast of Brazil, with a 2,700 km riverbed that provides water to more than 14 million  
25  
26 **60** people. At Paulo Afonso, one of the main hydroelectric power plants in the São Francisco River,  
27  
28  
29 **61** maintenance due to clogging of pipelines and corrosion caused by the golden mussel is estimated  
30  
31 **62** to cost US\$ 700,000 per year (*personal communication, Mizael Gusmã, Chief Maintenance*  
32  
33 **63** *Engineer for Centrais Hidrelétricas do São Francisco – CHESF*).

34  
35  
36 **64** A recent review has shown that, before arriving in South America, *L. fortunei* was  
37  
38  
39 **65** already an invader in China. Originally from the Pearl River Basin, the golden mussel has  
40  
41 **66** traveled 1,500 km into the Yang Tse and the Yellow River basins, being limited further north  
42  
43  
44 **67** only by the extreme natural barriers of Northern China [5]. Today, *L. fortunei* is found in the  
45  
46 **68** Paraguaizinho River, located only 150 km from the Teles-Pires River that belongs to the Alto  
47  
48  
49 **69** Tapajós River Basin and is the first to directly connect with the Amazon River Basin [6]. Due to  
50  
51 **70** its fast dispersion rates, it is very likely that *L. fortunei* will reach the Amazon River Basin in the  
52  
53 **71** near future.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

72           The reason why some freshwater bivalves, such as *L. fortunei*, *Dreissena polymorpha*,  
73 and *Corbicula fluminea*, are aggressive invaders is not fully understood. These bivalves present  
74 characteristics such as (i) tolerance to a wide range of environmental variables, (ii) short life  
75 span, (iii) early sexual maturation, and (iv) high reproductive rates that allow them to reach  
76 densities as high as 150,000 ind.m<sup>-2</sup> over a year [7, 8] that may explain the aggressive behavior.  
77 On the other hand, these traits are not exclusive to invasive freshwater bivalves and do not  
78 explain how they outcompete native species and disperse so widely.

79           To the best of our knowledge, there are no reports of successful strategies to control the  
80 expansion of mussel invasion in industrial facilities. Bivalves can sense chemicals in the water  
81 and close their valves as a defensive response [9], making them tolerant to a wide range of  
82 chemical substances, including strong oxidants like chlorine [10]. Microencapsulated chemicals  
83 have shown better results in controlling mussel populations in closed environments [10, 11] but  
84 it is unlikely they would work in the wild. Currently, there is no effective and efficient approach  
85 to control the invasion by *L. fortunei*.

86           The genome sequence is one of the most relevant and informative descriptions of species  
87 biology. The genetic substrate of invasive populations, upon which natural selection operates,  
88 can be of primary importance to understand and control a biological invader [12, 13].

89           We have partially funded the golden mussel genome sequencing through a pioneer  
90 crowdfunding initiative in Brazil ([www.catarse.me/genoma](http://www.catarse.me/genoma)). In this campaign, we could raise  
91 around USD\$ 20,000.00 at the same time we promoted scientific education and awareness in  
92 Brazil.



1  
2  
3  
4 93 Here we present the first complete genome dataset for the invasive bivalve *Limnoperna*  
5  
6 94 *fortunei*, assembled from short and long DNA reads and using a hybrid and hierarchical  
7  
8  
9 95 assembly strategy. This high-quality reference genome represents a substantial resource for  
10  
11 96 further studies of genetics and evolution of mussels, as well as for the development of new tools  
12  
13  
14 97 for plague control.

15  
16 98

### 19 99 **Genome sequencing in short Illumina and long PacBio reads**

21 100 *Limnoperna fortunei* mussels were collected from the Jacui River, Porto Alegre, Rio  
22  
23 101 Grande do Sul, Brazil (29°59'29.3"S 51°16'24.0"W). Voucher specimens were housed at the  
24  
25  
26 102 zoological collection (specimen number: 19643) of the Biology Institute at the Universidade  
27  
28  
29 103 Federal do Rio de Janeiro, Brazil. For the genome assembly, a total of 3 individuals were  
30  
31 104 sampled for DNA extraction from gills and to produce the three types of DNA libraries used in  
32  
33 105 this study. DNA was extracted using DNeasy Blood & Tissue Kit (Qiagen, Hilden, Germany) to  
34  
35  
36 106 prepare libraries for Illumina Nextera paired-end reads, with ~180bp and ~500bp of insert size,  
37  
38 107 (ii) Illumina Nextera mate-pair reads with insert sizes from 3 to 15 Kb, and (iii) Pacific  
39  
40  
41 108 Biosciences long reads (**Table 1**). Illumina libraries were sequenced respectively in a HiScanSQ  
42  
43 109 or HiSeq 1500 machine, and Pacific Biosciences reads were produced with the P4C6 chemistry  
44  
45  
46 110 and sequenced in 10 SMRT Cells. All Illumina reads were submitted to quality analysis with  
47  
48 111 FastQC (FastQC, RRID:SCR\_014583) followed by trimming with Trimmomatic (Trimmomatic,  
49  
50  
51 112 RRID:SCR\_011848) [14]. Pacific Biosciences adaptor-free subreads sequences were used as  
52  
53 113 input data for the genome assembly.

54  
55 114  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

115

116

117

118 **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
<b>Illumina</b> <b>Nextera</b>	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
<b>Pacific</b> <b>Biosciences</b>	P4C - 10/SMTRC	Subreads	1663730	11171487485		

119

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

122

123 For transcriptome sequencing, RNA was sampled from four tissues (gills, adductor  
 124 muscle, digestive gland, and foot) of three different golden mussel specimens. RNA was  
 125 extracted using NEXTflex Rapid Directional RNA-Seq Kit (Bioo Scientifics, TX, USA) and 12  
 126 barcodes from NEXTflex Barcodes compatible with Illumina NexSeq Machine. Resulting reads  
 127 (**Supplementary Table S1**) were submitted to FastQC quality analysis and trimmed with  
 128 Trimmomatic for all NEXTflex adaptors and barcodes. A total of 3 sets of *de novo* assembled  
 129 transcriptomes were generated using Trinity (Trinity, RRID:SCR\_013048) (**Table 2**); one set for

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 130 each specimen was a pool of the 4 tissue samples to avoid assembly bias due to intraspecific  
5  
6 131 polymorphism [15].  
7  
8

9 132

10  
11 133 **Table 2 - Trinity assembled transcripts used in the assembly and annotation of *L. fortunei***  
12  
13  
14 134 **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

135

### 136 Genome assembly using a hybrid and hierarchical strategy

137 The Jellyfish software (Jellyfish, RRID:SCR\_005491) [16] was used to count and  
138 determine the distribution frequency of lengths 25 and 31 k-mers (**Figure 1**) for the Illumina  
139 DNA paired-end and mate-pair reads (**Table 1**). Genome size was estimated to be 1,6 Gb by  
140 using the 25 k-mer distribution plot as total k-mer number and then subtracting erroneous reads  
141 (starting k-mer counts from 12 times coverage), to further divide by the homozygous coverage-  
142 peak depth (45 times coverage), as performed by Li *et al.* (2010) [17]. A double-peak k-mer  
143 distribution was used as evidence of genome diploidy (**Figure 1**) and high heterozygosity. The  
144 rate of heterozygosity was estimated to be 2.3% and it was calculated as described by Vij *et al.*

1  
2  
3  
4 145 (2016) [18], using as input data the 17-kmer distribution plot for reads from one unique  
5  
6 146 specimen.

7  
8  
9 147 Initially, we attempted to assemble the golden mussel genome using only short Illumina  
10  
11 148 reads of different insert sizes (paired-end and mate-pairs, **Table 1**) using traditional *de novo*  
12  
13 149 assembly software such as ALLPATHS (ALLPATHS-LG, RRID:SCR\_010742) [19],  
14  
15 150 SOAPdenovo (SOAPdenovo, RRID:SCR\_010752) [20], and MaSuRCA (MaSuRCA,  
16  
17 151 RRID:SCR\_010691) [21]. All these attempts resulted in very fragmented genome drafts, with an  
18  
19 152 N50 no higher than 5 Kb and a total of 4 million scaffolds. To reduce fragmentation, we further  
20  
21 153 sequenced additional long reads (10 PacBio SMTR Cells, **Table 1**) and performed a hybrid and  
22  
23 154 hierarchical *de novo* assembly described below and depicted in **Figure 2**.

24  
25  
26 155 First, (i) trimmed paired-end and mate-pair DNA Illumina reads (**Table 1**) were  
27  
28 156 assembled into contigs using the software Sparse Assembler [22] with parameters *LD 0*  
29  
30 157 *NodeCovTh 1 EdgeCovTh 0 k 31 g 15 PathCovTh 100 GS 1800000000*. Next, (ii) the resulting  
31  
32 158 contigs were assembled into scaffolds using Pacific Biosciences long subreads data and the  
33  
34 159 PacBio-correction-free assembly algorithm DBG2OLC [23] with parameters *LD1 0 k 17*  
35  
36 160 *KmerCovTh 10 MinOverlap 20 AdaptiveTh 0.01*. Finally, (iii) resulting scaffolds were submitted  
37  
38 161 to 6 iterative runs of the program L\_RNA\_Scaffolder [24] that uses exon-distance information  
39  
40 162 from *de novo* assembled transcripts (**Table 2**) to fill gaps and connect scaffolds whenever  
41  
42 163 appropriate. At the end, (iv) the final genome scaffolds were corrected for Illumina and Pacific  
43  
44 164 Biosciences sequencing errors with the software PILON [25]: all DNA and RNA short Illumina  
45  
46 165 reads were re-aligned back to the genome with BWA aligner (BWA, RRID:SCR\_010910) [26]  
47  
48 166 and resulting sam files were BAM-converted, sorted, and indexed with samtools package  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

167 (SAMTOOLS, RRID:SCR\_002105) [27]. Pilon [25] identifies INDELS and mismatches by  
168 coverage of reads and yields a final corrected genome draft. Pilon was run with parameters --  
169 *diploid –duplicates*.

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

**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 %
Mapping percentage of Illumina reads back to scaffolds	91 %

174  
175 The golden mussel genome presents 81% of all Benchmarking Universal Single Copy  
176 Orthologs (BUSCO version 3.3 analysis with Metazoa database; BUSCO, RRID:SCR\_015008)  
177 (**Table 4**) and, compared to the mollusk genomes currently available [28, 29, 30, 31, 32, 33, 34  
178 35] it represents one of the best assemblies of molluscan genomes so far also in terms of scaffold  
179 N50 and contiguity (**Table 5**).

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

180 One main challenges of assembling bivalve genomes lies in the high heterozygosity and  
181 amount of repetitive elements these organisms present: (i) the mussels *L. fortunei* and *Modiolus*  
182 *philippinarum* and the oyster *Crassostrea gigas* genomes were estimated to have heterozygosity  
183 rates of 2.3%, 2.02 % 1.95% respectively, which is substantially higher than other animal  
184 genomes [29], and (ii) repetitive elements correspond to at least 30% of the genomes of all  
185 studied bivalves so far (**Table 3**) [28, 29, 30, 31, 33, 34, 35 ]. Also, retroelements might be active  
186 in some species such as *L. fortunei* (refer to the retroelements-related section of this paper) and  
187 *C. gigas* [29], allowing genome rearrangements that may hinder for genome assembly. One  
188 exception seems to be the deep-sea mussel *B. platifrons* which has lower heterozygosity rates  
189 compared to other bivalves [31]. Sun *et al.*, (2017) [31] suggested it might be due to recurrent  
190 population bottlenecks happened after events of population extinction and recolonization in the  
191 extreme environment [31]. Nevertheless, most of the bivalve genome projects relying only on  
192 short Illumina reads are likely to present fragmented initial drafts [28, 30]. PacBio long reads  
193 allowed us to increase the N50 to 32 Kb and to reduce the number of scaffolds from millions to  
194 61102, using the DBG2OLC [23] assembler. Finally, interactive runs of L\_RNA\_scaffolder [24]  
195 using the transcriptomes (**Table 2**) rendered the final result of N50 312 Kb in 20548 scaffolds.  
196 It's important to note that assembly statistics can perform better for genomes assembled with  
197 reads generated with DNA extracted from one unique individual. This, however, was not  
198 possible for *L. fortunei*'s genome, due to the high amount of high-quality-DNA necessary to  
199 produce Illumina mate-pair and PacBio long reads. In this study, the challenge of assembling the  
200 high polymorphic regions between haplotypes was enhanced by the difficulties of assembling  
201 reads originated from highly polymorphic regions across individuals. However, the golden

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

202 mussel assembly presented here shows that the use of Illumina contigs, low coverage of PacBio  
203 long reads, transcriptome and Illumina re-mapping for final correction (**Figure 2**) represents an  
204 option for cost-efficient assembly of highly heterozygous genomes of nonmodel species such as  
205 bivalves.

**Table 4: Summary statistics of Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis for *L. fortunei* genome run for Metazoans**

Categories	Number of Genes	Percentage (%)
Total BUSCO groups searched	978	--
Complete BUSCOs	801	81.9%
Complete and single-copy BUSCOs	769	78.62%
Complete and duplicated BUSCOs	32	3.27%
Fragmented BUSCOs	72	7.36%
Missing BUSCOs	105	10.73%

210

14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

211 **Table 5: Comparison of genome assembly statistics for molluscan genomes.**

	<i>Haliotis discus hannai</i>	<i>Lottia gigantea</i>	<i>Aplysia californica</i>	<i>Ruditapes philippinarum</i>	<i>Patinopecten yessoensis</i>	<i>Crassostrea gigas</i>	<i>Pinctada fucata</i>	<i>Mytillus galloprovincialis</i>	<i>Bathymodiolus platifrons</i>	<i>Modiolus philippinarum</i>	<i>Limnoperna fortunei</i>
<b>Estimated genome size</b>	1.65Gb	359.5 Mb	1.8Gb	1.37 Gb	1.43 Gb	545 Mb	1.15 Gb	1.6 Gb	1.64Gb	2.38 Gb	1.6 Gb
<b>Number of scaffolds</b>	80,032	4,475	8,766	223,851	82,731	11,969	7,997	1,746,447	65,664	74,575	<b>20,548</b>
<b>Total size of scaffolds</b>	1,865,475,499	359,512,207	715,791,924	2,561,070,351	987,685,017	558,601,156	915,721,316	1,599,211,957	1,659,280,971	2,629,649,654	<b>1,673,125,894</b>
<b>Longest scaffold</b>	2,207,537	9,386,848	1,784,514	572,939	7,498,238	1,964,558	5,897,787	67,529	2,790,175	715382	<b>2,720,304</b>
<b>Shortest scaffold</b>	854	1,000	5,001	500	200	100	1,807	100	292	205	<b>558</b>
<b>Number of scaffolds &gt; 1 K nt</b>	79,923 (99.9%)	4,471 (99.9%)	8,766 (100%)	138,771 (61.9%)	16,004 (19.3%)	5,788 (48.4%)	7,997 (100%)	393,685 (22.5%)	38,704 (58.9%)	44,921 (60.2%)	<b>20,547 (100%)</b>
<b>Number of scaffolds &gt; 1 M nt</b>	67 (0.1%)	98 (2.2%)	27 (0.3%)	0 (0.0%)	248 (0.3%)	60 (0.5%)	27 (0.3%)	0 (0.0%)	164 (0.2%)	0 (0%)	<b>95 (0.5%)</b>
<b>Mean scaffold size</b>	23,309	80,338	81,655	11,441	11,939	46,671	114,508	916	25,269	35,262	<b>81,425</b>
<b>Median scaffold size</b>	1,697	3,622	13,763	1,327	362	824	14,683	258	1,284	13,722	<b>22,134</b>
<b>N50 scaffold length</b>	200,099	1,870,055	264,327	48,447	803,631	401,319	345,846	2,651	343,373	100,161	<b>312,020</b>
<b>Sequencing coverage</b>	322 X	8.87 X	11 X	39.7 X	297 X	155 X	234 X	32 X	319 X	209.5 X	<b>60 X</b>
<b>Sequencing Technology</b>	Illumina + PacBio	Sanger	Sanger	Illumina	Illumina	Illumina	Illumina + BACs	Illumina	Illumina	Illumina	<b>Illumina + PacBio</b>



1  
2  
3  
4 214  
5 215  
6  
7 216 **Around 10% of repetitive elements are transposons**  
8

9 217 Initial masking of *L. fortunei* genome was done using RepeatMasker program  
10  
11 218 (RepeatMasker, RRID:SCR\_012954) [36] with parameter *-species bivalves* and masked 3.4% of  
12  
13  
14 219 the total genome. This content was much lower than the masked portion of other molluscan  
15  
16 220 genomes: 34% in *C. gigas* [29] and 36% in *M. galloprovincialis* [28], suggesting that the fast  
17  
18  
19 221 evolution of interspersed elements limits the use of repeat libraries from divergent taxa [37].  
20  
21 222 Thus, we generated a *de novo* repeat library for *L. fortunei* using the program RepeatModeler  
22  
23 223 (RepeatModeler, RRID:SCR\_015027) [38] and its integrated tools (RECON [39], TRF [40], and  
24  
25  
26 224 RepeatScout [41]). This *de novo* repeat library was the input to RepeatMasker together with the  
27  
28  
29 225 first masked genome draft of *L. fortunei*, and resulted in a final masking of 33.4% of the genome.  
30  
31 226 Even though more than 90% of the repeats were not classified by RepeatMasker  
32  
33 227 (**Supplementary Table S2**), 8.85% of the repeats were classified as LINEs, Class I transposable  
34  
35  
36 228 elements. In addition, large numbers of reverse-transcriptases (824 counts, Pfam RVT\_1  
37  
38 229 PF00078), transposases (177 counts, Pfam HTH\_Tnp\_Tc3\_2 PF01498), and integrases (501  
39  
40  
41 230 counts, Pfam Retroviral integrase core domain PF00665) and other related elements were  
42  
43 231 detected; over 98% of these had detectable transcripts.  
44  
45  
46 232

47  
48 233 **More than 30,000 sequences identified by gene prediction and automated**  
49  
50 234 **annotation.**

51  
52  
53 235 To annotate the golden mussel genome, we sequenced a number of transcriptomes (**Table S1**),  
54  
55 236 *de novo* assembled (**Table 2**) and aligned these transcriptomes to the genome scaffolds, and  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 237 created gene models with the PASA pipeline [36]. These models were used to train and run the  
5  
6  
7 238 *ab initio* gene predictor AUGUSTUS (Augustus: Gene Prediction, RRID:SCR\_008417) [37]  
8  
9 239 (**Supplementary Figure S1**). The complete gene models yielded by PASA [42] were BLASTed  
10  
11  
12 240 (e-value 1e-20) against the Uniprot database (UniProt, RRID:SCR\_002380) and those with 90%  
13  
14 241 or more of their sequences showing in the BLAST hit alignment were considered for further  
15  
16 242 analysis. Next, all the necessary filters to run an AUGUSTUS [43] personalized training were  
17  
18  
19 243 performed: (i) only gene models with more than 3 exons were maintained, (ii) sequences with  
20  
21 244 90% or more overlap were withdrawn and only the longest sequences were retained, and (iii)  
22  
23  
24 245 only gene models free of repeat regions, as indicated by BLASTN similarity searches with *de*  
25  
26 246 *novo* library of repeats, were maintained. These curated data yielded a final set of 1,721 gene  
27  
28  
29 247 models on which AUGUSTUS [35] was trained in order to predict genes in the genome using the  
30  
31 248 default AUGUSTUS [43] parameters. Once the gene models were predicted, a final step was  
32  
33  
34 249 performed by using the PASA pipeline [42] once again in the *update* mode (parameters -c -A -g -  
35  
36 250 t). This final step compared the 55,638 gene models predicted by AUGUSTUS [43] with the  
37  
38  
39 251 40,780 initial transcript-based gene-structure models from PASA [42] to generate the final set of  
40  
41 252 60,717 gene models for *L. fortunei*. Of those, 58% had transcriptional evidence based on RNA  
42  
43 253 Illumina reads (**Table S2**) re-mapping, rate that was expected since our RNA-Seq libraries were  
44  
45  
46 254 constructed only for 4 tissues of adult golden mussel specimens without any environmental  
47  
48 255 stresses induction (**Table 2**). Therefore, these libraries lack transcripts for developmental stages,  
49  
50  
51 256 for some other cell types (i.e. hemocytes) and stress-inducible genes. Finally, 67% of the gene  
52  
53 257 models were annotated by homology searches against Uniprot or NCBI NR (**Table 6**).

54  
55 258  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

**Table 6: Summary of gene annotation against various databases for *L. fortunei* whole genome-predicted genes**

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

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

**Protein clustering indicates evolutionary proximity among mollusks species.**

Gene family relationships were assigned using reciprocal best BLAST and OrthoMCL software (version 1.4) [44] between *L. fortunei* proteins and the total protein set predicted for nine other mollusks: the mussels *M. galloprovincialis*, *M. philippinarum* and *B. platifrons*, the clam *Ruditapes philippinarum*, the scallop *Patinopecten yessoensis*, the pacific oyster *C. gigas*, the pearl oyster *Pinctada fucata* (genome version from Du *et al* [35]), and the gastropods *Lottia gigantea* and *Haliotis discus hannai* (see **Supplementary Table S3** for detailed information on the comparative data). **Figure 3A** presents orthologs relationships for five of the bivalves analyzed. A total of 6,337 orthologs groups are shared among the five bivalve species.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

274 Of all the orthologous found for the total 10 species, 44 groups are composed of single-  
275 copy orthologs containing one representative protein sequence of each species. These sequences  
276 were used to reconstruct a phylogeny: the single-copy orthologs sequences were concatenated  
277 and aligned with CLUSTALW [45] with a resulting alignment of 30755 sites in length (**Figure**  
278 **3B**). ProtTest 3.4.2 [46] was used to estimate the best fitting substitution model, which was  
279 VT+G+I+F [47]. With this alignment and model we reconstructed the phylogeny using PhyML  
280 [48] and 100 bootstrap repetition, the resulting tree is shown on **Figure 3B**.

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

281  
282  
283 We performed a quantitative comparison of protein domains predicted from whole  
284 genome projects of 10 molluscan species. The complete protein sets of *M. galloprovincialis*, *M.*  
285 *philippinarum* and *B. platifrons*, *Ruditapes philippinarum*, *Patinopecten yessoensis*, *C. gigas*,  
286 *Pinctada fucata*, *Lottia gigantea* and *Haliotis discus hannai* (**Supplementary Table S3**) were  
287 submitted to domain annotation using HMMER against Pfam-A database (e-value 1e-05).  
288 Protein expansions in *L. fortunei* were rendered using the normalized Pfam count value  
289 (average) obtained from the other nine mollusks, according to a model based on the Poisson  
290 cumulative distribution. Bonferroni correction ( $p \leq 0.05$ ) was applied for false discovery and  
291 absolute frequencies of Pfam-assigned-domains were initially normalized by the total count  
292 number of Pfam-assigned-domains found in *L. fortunei* to compensate for discrepancies in  
293 genome size and annotation bias.

294 For *L. fortunei*, the annotation against Pfam.A classified 40127 domains in 24513 gene  
295 models of which 83 and 67 were respectively expanded or contracted in comparison with the

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

296 other mollusks (**Supplementary Table S4 and S5; Figure 4A**). The 83 overrepresented domains  
297 were further analyzed for functional enrichment using domain-centric Gene Ontology (**Figure**  
298 **4B**). The analysis shows a prominent expansion of binding domains in *L. fortunei*, such as  
299 Thrombospondin (TSP\_1), Collagen, Immunoglobulins (Ig, I-set,Izumo-Ig Ig\_3), and Ankyrins  
300 (Ank\_2, Ank\_3, and Ank\_4). These repeats have a variety of binding properties and are involved  
301 in cell-cell, protein-protein and receptor-ligand interactions driving evolutionary improvement of  
302 complex tissues and immune defense system in metazoans [49, 50, 51, 52, 53]. An evolutionary  
303 pressure towards the development of a diversificated innate immune system is also suggested by  
304 the high amount of Leucine Rich Repeats (LRR) and Toll/interleukin-1 receptor homology  
305 domains (TIR). Death, another over-represented PFAM, is also part of TLR signaling, being  
306 present in several docking proteins such as Myd88, Irak4 and Pelle [54]. Interestingly, BLAST  
307 analysis of *L. fortunei* gene models against Uniprot identified two types of Toll Like Receptors  
308 (TLRs) whose prototypical architecture of N-terminal extracellular leucine-rich repeat (LRR)  
309 motifs including either a single or multiple cysteine cluster domain, a C-terminal TIR domain  
310 spaced by a single transmembrane-spanning domain [55] could be correctly identified using the  
311 Simple Modular Architecture Research Tool (SMART) [56]. Indeed, we confirmed 141  
312 sequences with similarity to single cysteine clusters TLRs (scc) typical of vertebrates, and 29  
313 sequence hits with the multiple cysteine cluster TLRs (mcc) typical of *Drosophila* [55].  
314 Phylogenetic analysis of all sequences (using PhyML [48], model JTT) (**Supplementary Figure**  
315 **S2**) shows evidence for TLRs clade separation in *L. fortunei*; the scc TLRs exhibit a higher  
316 degree of amino acid changes, higher molecular evolution, and diversification than the mcc  
317 TLRs. Overall, the expansion of these gene families might suggest an improved resistance to

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

318 infections. It is, however, equally curious that other immune-related gene families such as  
319 Fribinogen\_C and C1q seem to be contracted (**Supplementary Table S5**). This feature may  
320 depend on the evolutionary-driven, yet random, fate of the *L. fortunei* genome and consequence  
321 of different specific duplicate genes in other species. Also, other protein families involved in  
322 toxin metabolism, especially glutathione based processes and sulfotransferases are clearly  
323 contracted (**Table S5**).

**Final considerations**

325 Here we have described the first version of the golden mussel complete genome and its  
326 automated gene prediction that were funded through a crowdfunding initiative in Brazil. This  
327 genome contains valuable information for further evolutionary studies of bivalves and metazoa  
328 in general. Additionally, our team will further search for the presence of proteins of  
329 biotechnology interest such as the adhesive proteins produced by the foot gland that we have  
330 described elsewhere [57], or genes related to the reproductive system that have been shown to be  
331 very effective for invertebrate plague control [58]. The golden mussel genome and the predicted  
332 proteins are available for download in the Gigabase repository and the scientific community is  
333 welcome to further curate the gene predictions.

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

**Availability of supporting data**

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

339 *Limnoperna fortunei*'s genome and transcriptome data are available in the Sequence  
340 Read Archive (SRA) as BioProject PRJNA330677 and under the accession numbers  
341 **SRR5188384, SRR5195098, SRR518800, SRR5195097, SRR5188315, SRR5181514**. This  
342 Whole Genome Shotgun project has been deposited in the DDBJ/ENA/GenBank under accession  
343 number NFUK00000000. The version described in this paper is version NFUK01000000.  
344 Supporting data, also including annotations and BUSCO results, are available via the  
345 *GigaScience* repository GigaDB [59].

346

347 **Additional files**

348 **Supplementary Table S1.** RNA raw reads sequenced for 3 *L. fortunei* specimens, 4 tissues each.

349 **Supplementary Table S2:** RepeatMasker classification of repeats predicted in *L. fortunei*  
350 genome.

351 **Supplementary Table S3:** Details of the online availability of the data used for ortholog  
352 assignment and protein domain expansion analysis.

353 **Supplementary Table S4:** Expanded protein families in *L. fortunei* genome.

354 **Supplementary Table S5:** Contracted protein families in *L. fortunei* genome.

355 **Supplementary Table S6:** Fantasy names given to *L. fortunei* genes and proteins from the  
356 backers that have supported us through crowdfunding ([www.catarse.me/genoma](http://www.catarse.me/genoma)).

357 **Supplementary Figure 1:** Steps performed for the prediction and annotation of *L. fortunei*  
358 genome.

359 **Supplementary Figure 2:** Phylogenetic tree of Toll-like (TLRs) receptors found in *L. fortunei*  
360 genome.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

**361 List of Abbreviations**

**362** BUSCO: Benchmarking Universal Single-Copy Orthologs; SRA: Sequence Read Archive;  
**363** KEGG: Kyoto Encyclopedia of Genes and Genomes.

**364 Competing interests**

**365** The authors declare that they have no competing interests.

**366 Authors' contribution**

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

**371**

**372 Funding**

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

**376 Acknowledgements**

**377** We thank Susan Mbedi and Kirsten Richter from BeGenDiv for RNA-Seq library preparation  
**378** and sequencing. We thank Dr. Loris Bennett for IT support while performing bioinformatics  
**379** analysis.

**380** We especially want to thank the 346 backers that supported the sequencing of the golden mussel  
**381** through crowdfunding, in a 2013 campaign that raised US\$ 20,000.00 ([www.catarse.me/genoma](http://www.catarse.me/genoma)).



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

382 We decided to give fantasy names to the genes and proteins that we found in the genome, to  
383 thank the backers for their support. The name list is available in **Supplementary Table S6**.

384  
385 **Consent for publication**

386 Does not apply.

387 **Ethics approval**

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

391  
392  
393 **References**

- 394 1. Pastorino G, Darrigran G, et al., *Limnoperna fortunei* (Dunker, 1857) (Mytilidae), nuevo  
395 bivalvo invasor em águas Del Rio de la Plata. *Neotropica*. 1993;39:101–2.
- 396 2. Darrigran G. Potential impact of filter-feeding invaders on temperate inland freshwater  
397 environments. *Biol Invasions* 2002; 4:145–156.
- 398 3. Uliano-Silva M, Fernandes F da C, Holanda IBB, Rebelo MF. Invasive species as a threat  
399 to biodiversity: The golden mussel *Limnoperna fortunei* approaching the Amazon River  
400 basin. In: Exploring Themes Aquat Toxicol Alodi, S, editor. Research Signpost; 2013.
- 401  
402 4. Boltovskoy D, Correa N. Ecosystem impacts of the invasive bivalve *Limnoperna fortunei*  
403 (golden mussel) in South America. *Hydrobiologia*. 2015;746(1):81–95.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427

5. Xu M. Distribution and Spread of *Limnoperna fortunei* in China. In: *Limnoperna fortunei* Boltovskoy D, editor. Cham: Springer International Publishing; 2015 p. 313–20.
6. 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.
7. 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.
8. 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.
9. Claudi R and Mackie GL. Practical manual for zebra mussel monitoring and control. Lewis Publishers, Boca. Raton, Florida, 1994. p 227
10. 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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

430  
431 12. Cox GW. Alien species and evolution: the evolutionary ecology of exotic plants, animals,  
432 microbes, and interacting native species. Washington: Island Press; 2004. 377 p.

433  
434 13. Hall MR, Kocot KM, Baughman KW, Fernandez-Valverde SL, Gauthier MEA,  
435 Hatleberg WL, et al. The crown-of-thorns starfish genome as a guide for biocontrol of  
436 this coral reef pest. *Nature.* 2017 13;544(7649):231–4.

437  
438 14. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina  
439 Sequence Data. *Bioinformatics.* 2014 1;170.

440  
441 15. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and  
442 transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat*  
443 *Protoc.* 2012 1;7(3):562–78.

444  
445 16. Marçais G, Kingsford C. A Fast, Lock-free Approach for Efficient Parallel Counting of  
446 Occurrences of K-mers. *Bioinformatics.* 2011 Mar;27(6):764–770.

447  
448 17. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and de novo assembly of  
449 the giant panda genome. *Nature.* 2010 Jan 21;463(7279):311–7.

450

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

451 18. Vij S, Kuhl H, Kuznetsova IS, Komissarov A, Yurchenko AA, Heusden PV, et al.  
452 Chromosomal-Level Assembly of the Asian Seabass Genome Using Long Sequence  
453 Reads and Multi-layered Scaffolding. *PLOS Genet.* 2016 Apr 15;12(4):e1005954.

454  
455 19. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-  
456 quality draft assemblies of mammalian genomes from massively parallel sequence data.  
457 *Proc Natl Acad Sci U S A.* 2011 25;108(4):1513–8.

458  
459 20. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically  
460 improved memory-efficient short-read de novo assembler. *GigaScience.* 2012;1:18.

461  
462 21. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA  
463 genome assembler. *Bioinformatics.* 2013 1;29(21):2669–77.

464  
465 22. Ye C, Ma Z, Cannon CH, Pop M, Yu DW. Exploiting sparseness in de novo genome  
466 assembly. *BMC Bioinformatics.* 2012;13(Suppl 6):S1.

467  
468 23. Ye C, Hill CM, Wu S, Ruan J, Ma Z (Sam). DBG2OLC: Efficient Assembly of Large  
469 Genomes Using Long Erroneous Reads of the Third Generation Sequencing  
470 Technologies. *Sci Rep.* 2016 30;6:31900.

471  
472 24. Xue W, Li J-T, Zhu Y-P, Hou G-Y, Kong X-F, Kuang Y-Y, et al. L\_RNA\_scaffolder:  
473 scaffolding genomes with transcripts. *BMC Genomics.* 2013;14:604.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

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

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

482 28. Murgarella M, Puiu D, Novoa B, Figueras A, Posada D, Canchaya C. A First Insight into  
483 the Genome of the Filter-Feeder Mussel *Mytilus galloprovincialis*. PLOS ONE. 2016  
484 15;11(3):e0151561.

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

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

490 31. Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea  
491 chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol. 2017 Apr  
492 3;1(5):0121

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

32. Nam B-H, Kwak W, Kim Y-O, Kim D-G, Kong HJ, Kim W-J, et al. Genome sequence of pacific abalone (*Haliotis discus hannai*): the first draft genome in family Haliotidae. *GigaScience*. 2017 May;6(5):1–8.
33. Wang S, Zhang J, Jiao W, Li J, Xun X, Sun Y, et al. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat Ecol Evol*. 2017 Apr 3;1(5):0120.
34. Mun S, Kim Y-J, Markkandan K, Shin W, Oh S, Woo J, et al. The Whole-Genome and Transcriptome of the Manila Clam (*Ruditapes philippinarum*). *Genome Biol Evol*. 2017 Jun;9(6):1487–98
35. Du X, Fan G, Jiao Y, Zhang H, Guo X, Huang R, et al. The pearl oyster *Pinctada fucata martensii* genome and multi-omic analyses provide insights into biomineralization. *GigaScience*. 2017 Aug;6(8):1–12
36. Smit AF., Hubley R, Green PJ. RepeatMasker Open-3.0. 1996 2010.
37. 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.
38. Smith AFA, Hubley R. RepeatModeler Open-1.0. [Internet]. 2014. Available from: <http://www.repeatmasker.org>

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

526  
527 40. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*  
528 *Res.* 1999 1;27(2):573–80.

529  
530 41. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large  
531 genomes. *Bioinformatics.* 2005 1;21(Suppl 1):i351–8

532  
533 42. Haas BJ, Delcher AL, Mount SM, Wortman JR, Jr RKS, Hannick LI, et al. Improving the  
534 *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic*  
535 *Acids Res.* 2003 1;31(19):5654–66.

536  
537 43. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped  
538 cDNA alignments to improve de novo gene finding. *Bioinforma Oxf Engl.* 2008  
539 1;24(5):637–44.

540  
541 44. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for  
542 Eukaryotic Genomes. *Genome Res.* 2003 1;13(9):2178–89.

543  
544 45. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of  
545 progressive multiple sequence alignment through sequence weighting, position-specific  
546 gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994 11;22(22):4673–80.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

548 46. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models  
549 of protein evolution. *Bioinformatics*, 2011 27:1164-1165.

550

551 47. Müller T, Vingron M. Modeling amino acid replacement. *J. Comput Biol.* 2000. 7:761-  
552 776.

553

554 48. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large  
555 phylogenies by maximum likelihood. *Syst Biol.* 2003. 52: 696-704.

556

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

559

560 50. Rennemeier C, Hammerschmidt S, Niemann S, Inamura S, Zähringer U, Kehrel BE.  
561 Thrombospondin-1 promotes cellular adherence of gram-positive pathogens via  
562 recognition of peptidoglycan. *FASEB J Off Publ Fed Am Soc Exp Biol.* 2007  
563 21(12):3118–32.

564

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

567

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

571



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

572 53. Tucker RP. The thrombospondin type 1 repeat superfamily. *Int J Biochem Cell Biol.*  
573 2004 36(6):969–74.

574  
575 54. Park HH, Lo YC, Lin SC, Wang L, Yang, JK, Wu H. The death domain superfamily in  
576 intracellular signaling of apoptosis and inflammation. *Annu. Rev. Immunol.* 2007 25,  
577 561-586.

578  
579 55. Leulier F, Lemaitre B..Toll-like receptors—taking an evolutionary approach. *Nature*  
580 *Reviews Genetics*, 2008 9,3: 165-178.

581  
582 56. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture  
583 research tool: identification of signaling domains. *Proc Natl Acad Sci U S A.* 1998  
584 26;95(11):5857–64

585  
586 57. Uliano-Silva M, Americo JA, Brindeiro R, Dondero F, Prosdocimi F, Rebelo M de F.  
587 Gene discovery through transcriptome sequencing for the invasive mussel *Limnoperna*  
588 *fortunei*. *PloS One.* 2014;9(7):e10297.

589  
590 58. Hammond A, Galizi R, Kyrou K, Simoni A, Siniscalchi C, Katsanos D, et al. A CRISPR-  
591 Cas9 gene drive system targeting female reproduction in the malaria mosquito vector  
592 *Anopheles gambiae*. *Nat Biotechnol.* 2015 7;34(1):78–83.

1  
2  
3  
4 594 59. Uliano-Silva M, Dondero F, Otto TD, Costa I, Lima NC, Americo JA et al. Supporting data for  
5  
6 595 "A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel  
7  
8  
9 596 *Limnoperna fortunei*". *GigaScience Database* 2017. <http://dx.doi.org/10.5524/100386>

10  
11 597  
12  
13 598  
14  
15  
16 599 **Figure 1:** K-mer distribution of *Limnoperna fortunei* Illumina DNA reads (Table 1).  
17 600

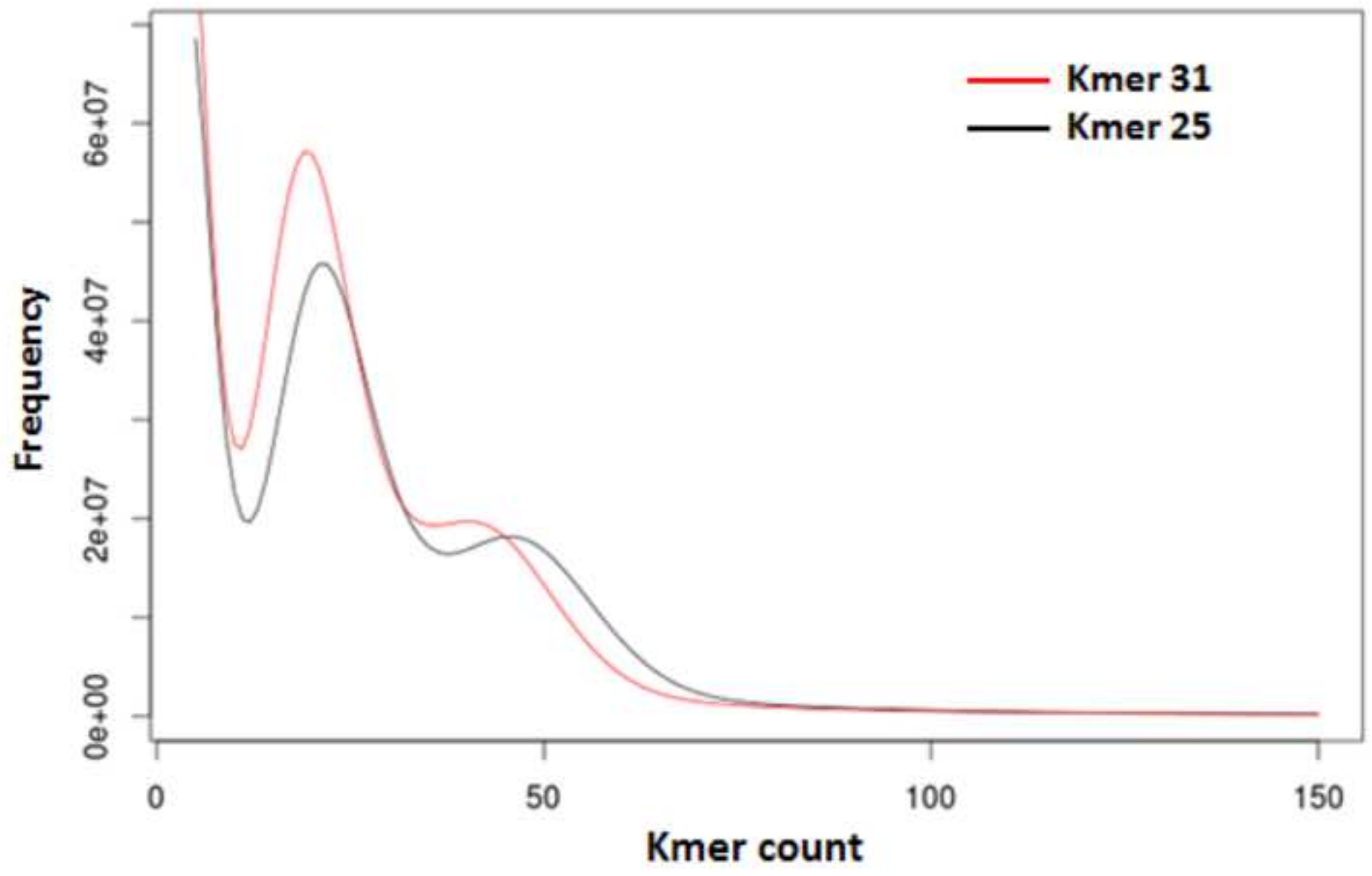
18 601  
19 602 **Figure 2: Hierarchical assembly strategy employed for the golden mussel genome**  
20 603 **assembly.** Trimmed Illumina reads were assembled to the level of contigs with Sparse  
21 604 Assembler algorithm (**Step 1**). Then, Illumina contigs and PacBio reads were used to build  
22 605 scaffolds with DBG2OLC assembler, that anchors Illumina contigs to erroneous PacBio  
23 606 subreads, correcting them and building longer scaffolds (**Step 2**), followed by transcriptome  
24 607 joining scaffolds using L\_RNA\_scaffolder (**Step 3**). Final scaffolds were corrected by re-  
25 608 aligning all Illumina DNA and RNA-seq reads back to them and calling consensus with Pilon  
26 609 software (**Step 4**). In bold is bioinformatics software used in each step. Red blocks indicate  
27 610 PacBio errors, which are represented by insertions and/or deletions found in approximately 12%  
28 611 of PacBio subreads.  
29  
30  
31  
32

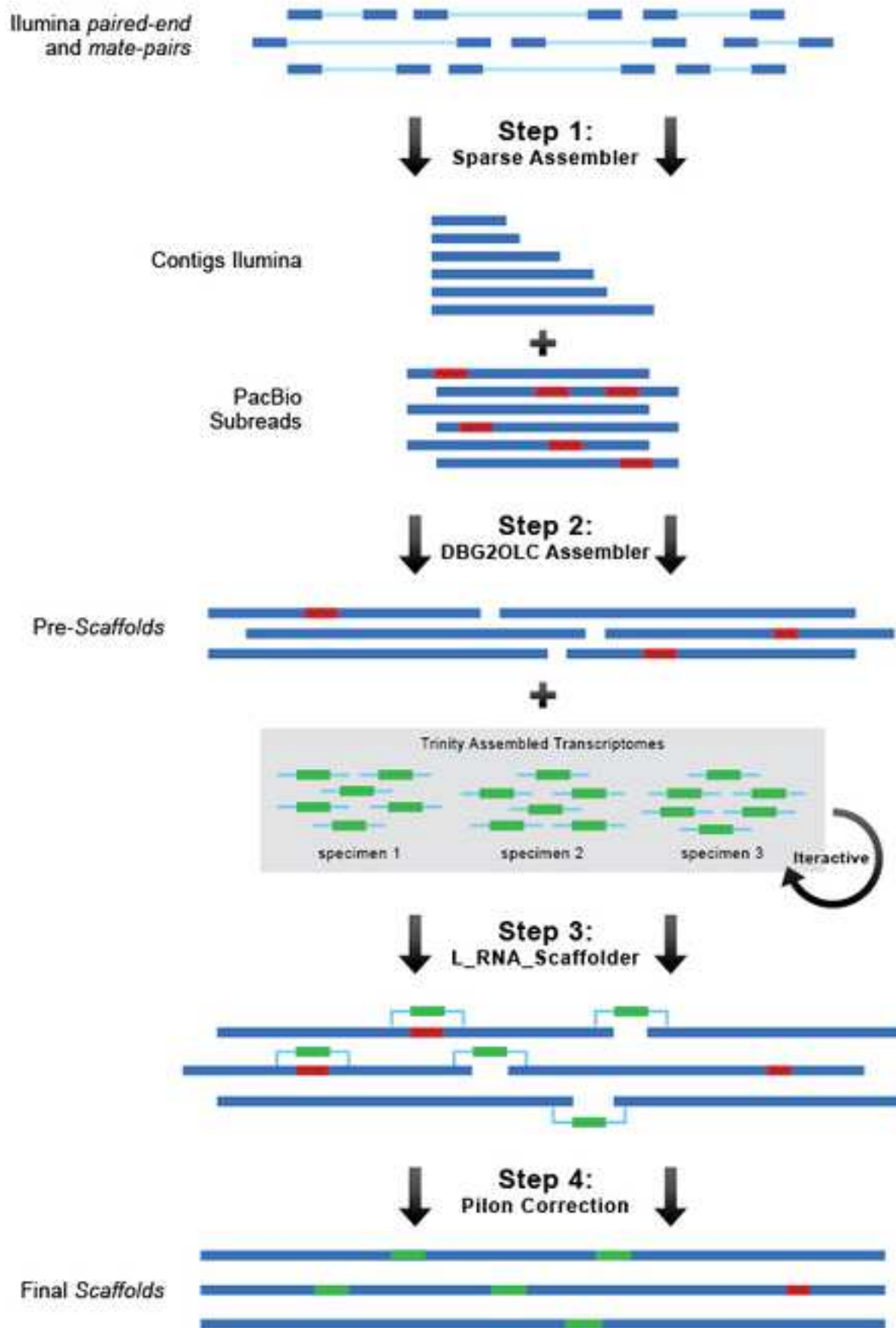
33 611  
34 612 **Figure 3A:** Gene family assigned with OrthoMCL for the total set of proteins predicted  
35 613 from five mussel genome projects. Outside the Venn diagram its represented the species name  
36 614 and below it is the number of proteins / number of clustered proteins / number of clusters. **B:**  
37 615 Phylogeny of the concatenated data set using 44 single-copy orthologs extracted from ten  
38 616 molluscan genomes. The VT model was estimated to be best fitting substitution model with  
39 617 ProtTest 3.4.2. We reconstructed the phylogeny using PhyML and 100 bootstrap repetition.  
40  
41  
42  
43  
44

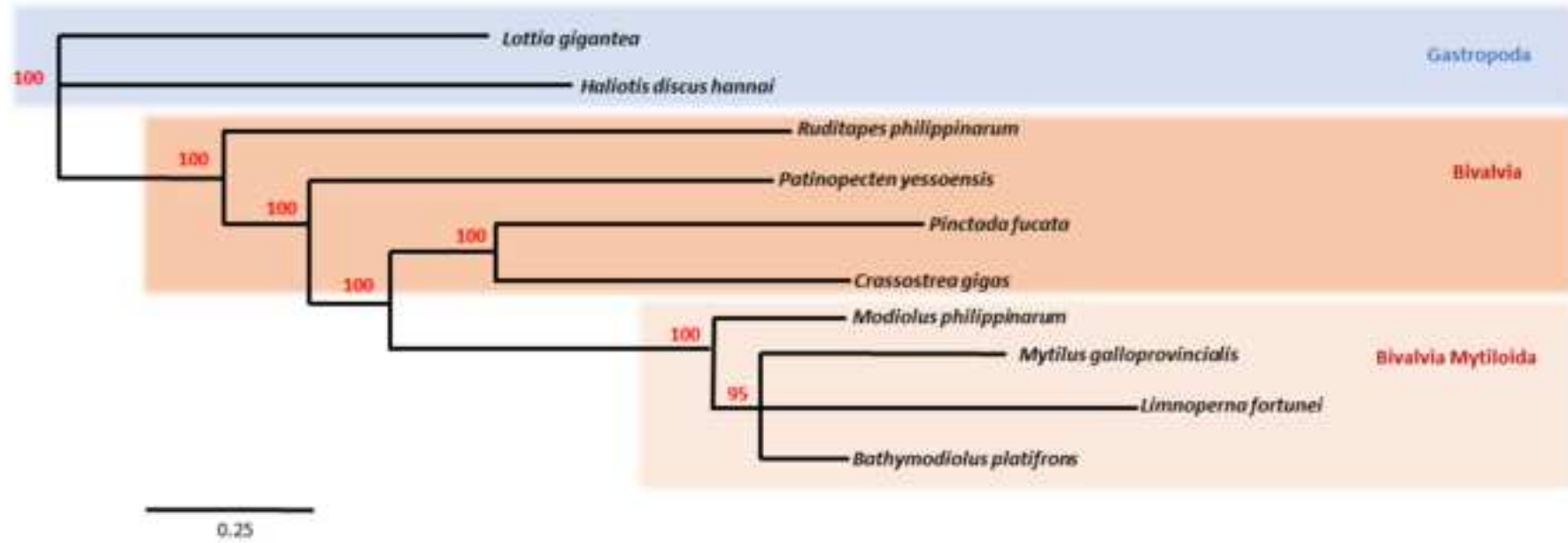
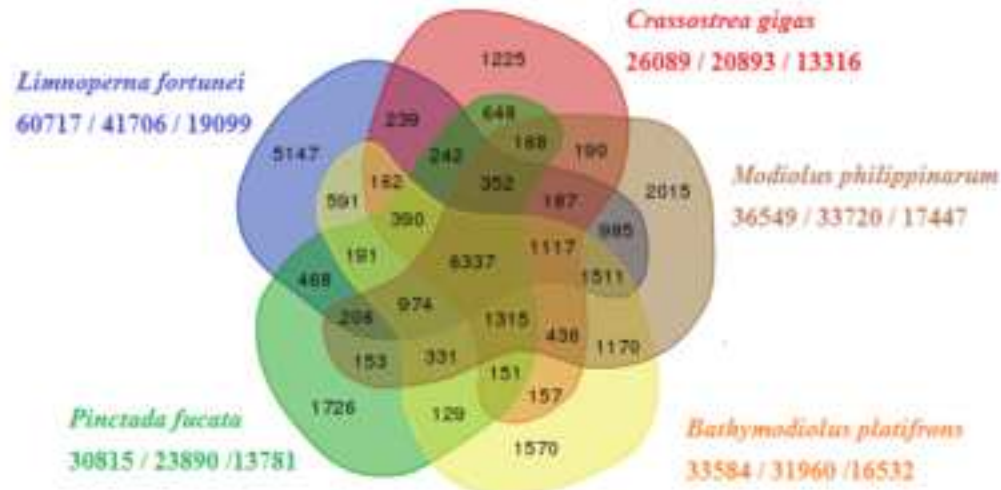
45 619  
46 620 **Figure 4: Gene family representation analysis in the *L. fortunei* genome. Panel A.**  
47 621 **PFAM hierarchical clustering, heatmap.** Features were selected according to a model based on  
48 622 the Poisson cumulative distribution of each PFAM count in the golden mussel genome vs the  
49 623 normalized average values found in the other nine molluscan genomes (Bonferroni correction,  $P$   
50 624  $\leq 0.05$ ). Transposable elements were included in the analysis. Colors depict the log2 ratio  
51 625 between PFAM counts found in each single genome and the corresponding mean value. The  
52 626 hierarchical clustering used the average dot product for data matrix and complete linkage for  
53 627 branching. Legend: Lf, *L. fortunei*; Bp, *Bathymodioulus platifrons*; Mg, *Mytilus*  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

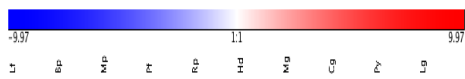
628 *galloprovincialis*; Mp, *Modioulus philippinarum*; Cg, *Crassostrea gigas*; Pf, *Pinctada fucata*;  
629 Py, *Patinopecten yessoensis*; Rp, *Ruditapes philippinarum*; Hd, *Haliotis discus hannai*; Lg,  
630 *Lottia gigantea* **Panel B. Gene ontology analysis of expanded gene families (PFAMs),**  
631 **semantic scatter plot.** Shown are cluster representatives after redundancy reduction in a two-  
632 dimensional space applying multidimensional scaling to a matrix of semantic similarities of GO  
633 term. Color indicates the GO enrichment level (legend in upper left-hand corner); size indicates  
634 the relative frequency of each term in the UNIPROT database (larger bubbles represent less  
635 specific processes).



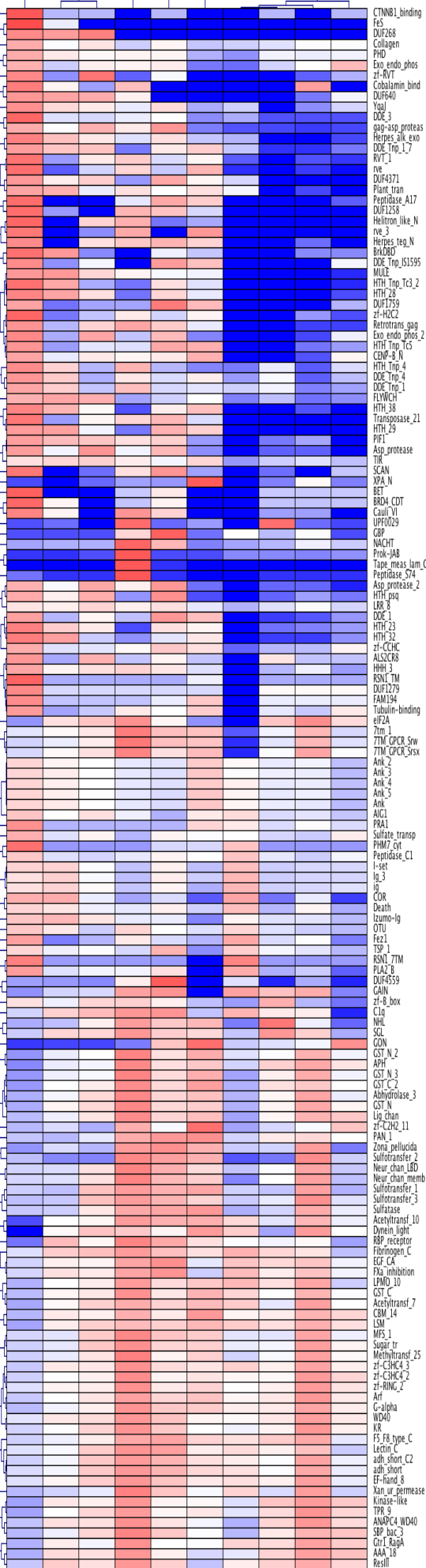




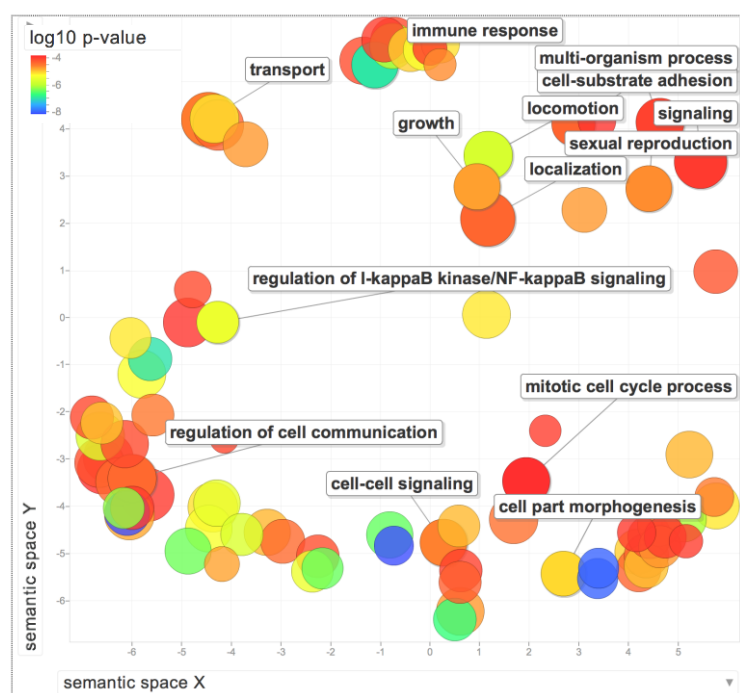
A




Li Bp MIP Pf Rp Hd Ma Cg Py Lg




B






Click here to access/download  
**Supplementary Material**  
TableS1.docx








Click here to access/download  
**Supplementary Material**  
TableS2.docx






Click here to access/download  
**Supplementary Material**  
Figure-S1.png

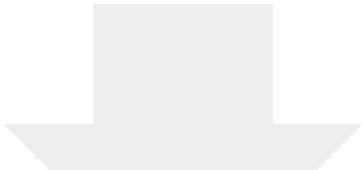


Click here to access/download  
**Supplementary Material**  
figureS2.pdf




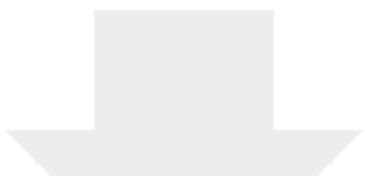


Click here to access/download  
**Supplementary Material**  
tableS6-.docx




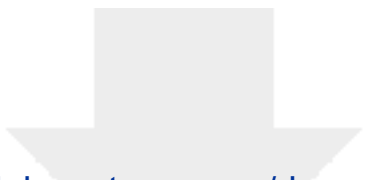
Click here to access/download  
**Supplementary Material**  
TableS3-revised.docx





Click here to access/download  
**Supplementary Material**  
Table-S4-final.xlsx





Click here to access/download  
**Supplementary Material**  
TableS5.xlsx

