

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

Manuscript Number:	GIGA-D-17-00124									
Full Title:	A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel <i>Limnoperna fortunei</i>									
Article Type:	Data Note									
Funding Information:	<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									
Abstract:	<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>									
Corresponding Author:	Marcela Uliano da Silva, Ph.D Universidade Federal do Rio de Janeiro Rio de Janeiro, RJ BRAZIL									
Corresponding Author Secondary Information:										
Corresponding Author's Institution:	Universidade Federal do Rio de Janeiro									
Corresponding Author's Secondary Institution:										
First Author:	Marcela Uliano da Silva, Ph.D									
First Author Secondary Information:										
Order of Authors:	<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
Order of Authors Secondary Information:	
Opposed Reviewers:	
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. 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>Resources</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 Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	Yes
<p>Availability of data and materials</p> <p>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.</p> <p>Have you have met the above requirement as detailed in our Minimum</p>	Yes

[Standards Reporting Checklist?](#)

1
2
3
4 **1 DATA NOTE**

5
6
7 **2 A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel**

8
9 **3 *Limnoperna fortunei***

10
11 **4 Authors:** Marcela Uliano-Silva^{1*}, Francesco Dondero², Thomas Dan Otto³, Igor Costa⁴, Nicholas
12 Costa Barroso Lima^{4,7}, Juliana Alves Americo¹, Camila Junqueira Mazzoni^{5,6}, Francisco
13 Prosdocimi⁴, Mauro de Freitas Rebelo^{1*}

14
15 7 Marcela Uliano-Silva: marcela.uliano@gmail.com

16 8 Francesco Dondero: francesco.dondero@uniupo.it

17 9 Thomas D. Otto: tdo@sanger.ac.uk

18
19 10 Igor Costa: igor.bioinfo@gmail.com

20 11 Nicholas Costa Barroso Lima: ncblima@gmail.com

21 12 Juliana Alves Americo: juliana.americo@gmail.com

22 13 Camila Mazzoni: mazzoni@izw-berlin.de

23 14 Francisco Prosdocimi: prosdocimi@bioqmed.ufrj.br

24 15 Mauro de Freitas Rebelo: mrebelo@biof.ufrj.br

25 16 Affiliations:

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 Science and Technological Innovation (DiSIT), Università del Piemonte

34
35
36 20 Orientale Amedeo Avogadro, Vercelli-Novara-Alessandria, Italy

37
38 21 3 Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, UK

39
40
41 22 4 Leopoldo de Meis Biomedical Biochemistry Institute (IBqM), Universidade Federal do Rio de
42 Janeiro, Rio de Janeiro, Brazil

43
44
45 24 5 Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, Berlin,
46 Germany

47
48
49 26 6 Berlin Center for Genomics in Biodiversity Research, Berlin, Germany

50
51
52 27 7 Bioinformatics Laboratory (LabInfo) of the National Laboratory for Scientific Computing,
53 Petrópolis, Rio de Janeiro, Brazil

54
55
56 28 *Correspondence: marcela.uliano@gmail.com; mrebelo@biof.ufrj.br

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

30 ABSTRACT

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

46 KEYWORDS: Amazon; binding domain; bivalves; genomics; TLR; transposon.

47 DATA DESCRIPTION

48 The golden mussel *Limnoperna fortunei* is an Asian bivalve that arrived in the southern
49 part of South America about 25 years ago [1]. Since then, it has moved ~5,000 km, invading
50 upstream continental waters and reaching northern parts of the continent [2] leaving behind a
51 track of great economic impact and environmental degradation [3]. The latest infestation was
52 reported in 2016 in the São Francisco River, one of the main rivers in the Northeast of Brazil,

1
2
3
4 53 with a 2,700 km riverbed that provides water to more than 14 million people. At Paulo Afonso,
5
6
7 54 one of the main hydroelectric power plants in the São Francisco River, maintenance due to
8
9 55 clogging of pipelines and corrosion caused by the golden mussel is estimated to cost US\$ 700,000
10
11 56 per year (*personal communication, Mizael Gusmã, Chief Maintenance Engineer for Centrais*
12
13 57 *Hidrelétricas do São Francisco – CHESF*).

14
15
16 58 A recent review has shown that, before arriving in South America, *L. fortunei* was
17
18
19 59 already an invader in China. Originally from the Pearl River Basin, the golden mussel has
20
21 60 traveled 1,500 km into the Yang Tse and the Yellow River basins, being limited further north
22
23
24 61 only by the extreme natural barriers of Northern China [4]. Today, *L. fortunei* is found in the
25
26 62 Paraguaizinho River, located only 150 km from the Teles-Pires River that belongs to the Alto
27
28
29 63 Tapajós River Basin and is the first to directly connect with the Amazon River Basin [5]. Due to
30
31 64 its fast dispersion rates, it is very likely that *L. fortunei* will reach the Amazon River Basin in the
32
33
34 65 near future.

35
36 66 The reason why some bivalves, such as *L. fortunei*, *Dreissena polymorpha*, and
37
38 67 *Corbicula fluminea*, are aggressive invaders is not fully understood. These bivalves present
39
40
41 68 characteristics such as (i) tolerance to a wide range of environmental variables, (ii) short life
42
43 69 span, (iii) early sexual maturation, and (iv) high reproductive rates that allow them to reach
44
45
46 70 densities as high as 150,000 ind.m⁻² over a year [6, 7] that may explain the aggressive behavior.
47
48
49 71 On the other hand, these traits are not exclusive to invasive bivalves and do not explain how they
50
51 72 outcompete native species and disperse so widely.

52
53 73 To the best of our knowledge, there are no reports of successful strategies to control the
54
55
56 74 expansion of mussel invasion in industrial facilities. Bivalves can sense chemicals in the water
57
58 75 and close their valves as a defensive response [8], making them tolerant to a wide range of
59
60
61
62
63
64
65

1
2
3
4 76 chemical substances, including strong oxidants like chlorine [9]. Microencapsulated chemicals
5
6 77 have shown better results in controlling mussel populations in closed environments [9, 10] but it
7
8 78 is unlikely they would work in the wild. Currently, there is no effective and efficient approach to
9
10
11 79 control the invasion by *L. fortunei*.

12
13
14 80 The genome sequence is one of the most relevant and informative descriptions of species
15
16 81 biology. The genetic substrate of invasive populations, upon which natural selection operates,
17
18
19 82 can be of primary importance to understand and control a biological invader [11].

20
21 83 We have partially funded the golden mussel genome sequencing through a pioneer
22
23 84 crowdfunding initiative in Brazil (www.catarse.me/genoma). In this campaign, we could raise
24
25
26 85 around US\$ 20,000.00 at the same time we promoted scientific education and awareness in Brazil.

27
28 86 Here we present the first complete genome dataset for the invasive bivalve *Limnoperna*
29
30 87 *fortunei*, assembled from short and long DNA reads and using a hybrid and hierarchical
31
32
33 88 assembly strategy. This high-quality reference genome represents a substantial resource for
34
35
36 89 further studies of genetics and evolution of mussels, as well as for the development of new tools
37
38 90 for plague control.

39
40
41 91

42 43 92 **Genome sequencing in short Illumina and long PacBio reads**

44
45 93 *Limnoperna fortunei* mussels were collected from the Jacui River, Porto Alegre, Rio
46
47
48 94 Grande do Sul, Brazil (29°59'29.3"S 51°16'24.0"W). Voucher specimens were housed at the
49
50
51 95 zoological collection (specimen number: 19643) of the Biology Institute at the Universidade
52
53 96 Federal do Rio de Janeiro, Brazil. For the genome assembly, a total of 3 individuals were
54
55 97 sampled for DNA extraction from gills. DNA was extracted using DNeasy Blood & Tissue Kit
56
57
58 98 (Qiagen, Hilden, Germany) to prepare libraries for Illumina Nextera paired-end reads, with
59
60
61
62
63
64
65

99 ~180bp and ~500bp of insert size, (ii) Illumina Nextera mate-pair reads with insert sizes from 3
 100 to 15 Kb, and (iii) Pacific Biosciences long reads (**Table 1**). Illumina libraries were sequenced
 101 respectively in a HiScanSQ or HiSeq 1500 machine, and Pacific Biosciences reads were
 102 produced with the P4C6 chemistry and sequenced in 10 SMRT Cells. All Illumina reads were
 103 submitted to quality analysis with FastQC (FastQC, RRID:SCR_014583) followed by trimming
 104 with Trimmomatic (Trimmomatic, RRID:SCR_011848) [12]. Pacific Biosciences adaptor-free
 105 subreads sequences were used as input data for the genome assembly.

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		

107
 108 *trimmomatic parameters for Illumina reads - ILLUMINACLIP:NexteraPE-PE.fa:2:30:10
 109 SLIDINGWINDOW:4:2 LEADING:10 TRAILING:10 CROP:101 HEADCROP:0 MINLEN:80

110
 111 For transcriptome sequencing, RNA was sampled from four tissues (gills, adductor
 112 muscle, digestive gland, and foot) of three different golden mussel specimens. RNA was
 113 extracted using NEXTflex Rapid Directional RNA-Seq Kit (Bioo Scientifics, TX, USA) and 12
 114 barcodes from NEXTflex Barcodes compatible with Illumina NexSeq Machine. Resulting reads

1
2
3
4 115 **(Supplementary Table S1)** were submitted to FastQC quality analysis (FastQC,
5
6 116 RRID:SCR_014583) and trimmed with Trimmomatic (Trimmomatic, RRID:SCR_011848) [12]
7
8
9 117 for all NEXTflex adaptors and barcodes. A total of 3 sets of *de novo* assembled transcriptomes
10
11 118 were generated using Trinity (Trinity, RRID:SCR_013048) **(Table 2)**; one set for each specimen
12
13
14 119 was a pool of the 4 tissue samples to avoid assembly bias due to intraspecific polymorphism
15
16 120 [13]. All generated sequences are deposited in the SRA Archive under the following accession
17
18
19 121 numbers: SRR5188384, SRR5195098, SRR5188200, SRR5195097, SRR5188315, and
20
21 122 SRR5181514. Also this Whole Genome Shotgun project has been deposited in the
22
23
24 123 DDBJ/ENA/GenBank under accession number NFUK00000000. The version described in this
25
26 124 paper is version NFUK01000000. Genome files are available in the Gigascience database.
27
28
29
30

31 126 **Table 2 - Trinity assembled transcripts used in the assembly and annotation of *L. fortunei***
32
33 127 **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

53 128
54
55 129
56
57 130
58
59
60
61
62
63
64
65

131 **Genome assembly using a hybrid and hierarchical strategy**

132 The Jellyfish software [14] was used to count and determine the distribution frequency of
133 lengths 25 and 31 k-mers (**Figure 1**) for the Illumina DNA paired-end and mate-pair reads
134 (**Table 1**). Genome size was estimated using the 25 k-mer distribution plot as total k-mer number
135 and then subtracting erroneous reads (starting k-mer counts from 12 times coverage), to further
136 divide by the homozygous coverage-peak depth (45 times coverage), as performed by Li *et al.*
137 (2010) [15]. A double-peak k-mer distribution was used as evidence of genome diploidy (**Figure**
138 **1**). The genome size of *L. fortunei* was estimated to be 1,6 Gb.

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

146 First, (i) trimmed paired-end and mate-pair DNA Illumina reads (**Table 1**) were
147 assembled into contigs using the software Sparse Assembler [19] with parameters *LD 0*
148 *NodeCovTh 1 EdgeCovTh 0 k 31 g 15 PathCovTh 100 GS 1800000000*. Next, (ii) the resulting
149 contigs were assembled into scaffolds using Pacific Biosciences long subreads data and the
150 PacBio-correction-free assembly algorithm DBG2OLC [20] with parameters *LD1 0 k 17*
151 *KmerCovTh 10 MinOverlap 20 AdaptiveTh 0.01*. Finally, (iii) resulting scaffolds were submitted
152 to 6 iterative runs of the program L_RNA_Scaffolder [21] that uses exon-distance information
153 from *de novo* assembled transcripts (**Table 2**) to fill gaps and connect scaffolds whenever

1
2
3
4 154 appropriate. At the end, (iv) the final genome scaffolds were corrected for Illumina and Pacific
5
6 155 Biosciences sequencing errors with the software PILON [22]: all DNA and RNA short Illumina
7
8
9 156 reads were re-aligned back to the genome with BWA aligner (BWA , RRID:SCR_010910) [23]
10
11
12 157 and resulting sam files were BAM-converted, sorted, and indexed with samtools package
13
14 158 (SAMTOOLS, RRID:SCR_002105) [24]. Pilon [22] identifies INDELS and mismatches by
15
16 159 coverage of reads and yields a final corrected genome draft. Pilon was run with parameters --
17
18
19 160 *diploid -duplicates*.

21 161 The final genome was assembled in 20,548 scaffolds, with an N50 of 312 Kb and a total
22
23
24 162 assembly length of 1.6 Gb (**Table 3**).

25
26 163

28
29 164 **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 %

30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46 165
47
48
49 166 An initial quality assessment revealed that 91% of all Illumina reads used to construct the
50
51 167 scaffolds mapped back to the final draft.

53
54 168 The golden mussel genome presents 81% of all Benchmarking Universal Single Copy
55
56 169 Orthologs (BUSCO version 3.3 analysis with Metazoa database) (BUSCO, RRID:SCR_015008)
57
58 170 (Table 4) and, compared to the mollusk genomes currently available [25, 26, 27, 28, 29] it
59
60
61
62
63
64
65

1
2
3
4 171 represents one of the best assemblies of molluscan genomes so far (**Table 5**). In fact, the
5
6 172 assembly of the *L. fortunei* genome presented here exhibits a slightest lower N50 than the
7
8
9 173 genomes of the mussel *Bathymodiolus platifrons* and the oyster *Crassostrea gigas* (**Table 5**).
10
11
12 174 Nevertheless, although *L. fortunei*'s genome is similar in size compared to *B. platifrons*, it is 3
13
14 175 times larger than the *C. gigas* genome (**Table 5**).

15
16 176 The main challenge of assembling bivalve genomes lies in the high heterozygosity and
17
18
19 177 amount of repetitive elements these organisms present: (i) the *Crassostrea gigas* genome was
20
21 178 estimated to have a heterozygosity rate 2.3% higher than other animal genomes [26], and (ii)
22
23
24 179 repetitive elements correspond to at least 30% of the genomes of all studied bivalves so far
25
26 180 (**Table 3**) [25, 26, 27, 28]. Also, retroelements might still be active in some species such as *L.*
27
28
29 181 *fortunei* (refer to the retroelements-related section of this paper) and *C. gigas* [26], allowing
30
31 182 genome rearrangements that may be obstacles for genome assembly. One exception seems to be
32
33
34 183 the deep-sea mussel *B. platifrons* which has lower heterozygosity rates compared to other
35
36 184 bivalves [28]. Sun *et al.*, (2017) [28] suggested it might be due to recurrent population
37
38 185 bottlenecks happened after events of population extinction and recolonization in the extreme
39
40
41 186 environment [28]. Nevertheless, most of the bivalve genome projects relying only on short
42
43 187 Illumina reads are likely to present fragmented initial drafts [25, 27]. PacBio long reads allowed
44
45
46 188 us to increase the N50 to 32 Kb and to reduce the number of scaffolds from millions to 61102,
47
48 189 using the DBG2OLC [20] assembler. Finally, interactive runs of L_RNA_scaffolder [21] using
49
50
51 190 the transcriptomes (**Table 2**) rendered the final result of N50 312 Kb in 20548 scaffolds. Thus,
52
53 191 our assembly strategy of Illumina contigs, low coverage of PacBio reads, transcriptome and
54
55 192 Illumina re-mapping for final correction (**Figure 2**) represents an option for cost-efficient
56
57
58 193 assembly of highly heterozygous genomes of nonmodel species such as 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%

14
15
16
17
18
19
20 211
21
22

	<i>Haliotis discus hannai</i>	<i>Crassostrea gigas</i>	<i>Pinctada fucata</i>	<i>Lottia gigantea</i>	<i>Aplysia californica</i>	<i>Mytilus galloprovincialis</i>	<i>Bathymodiolus platifrons</i>	<i>Modiolus philippinarum</i>	<i>Limnoperna fortunei</i>
Estimated genome size	1.65Gb	545 Mb	1.15 Gb	359.5 Mb	1.8Gb	1.6 Gb	1.64Gb	2.38 Gb	1.6 Gb
Number of scaffolds	80,032	11,969	800,982	4,475	8,766	1,746,447	65,664	74,575	20,548
Total size of scaffolds	1,865,475,499	558,601,156	1,413,178,538	359,512,207	715,791,924	1,599,211,957	1,659,280,971	2,629,649,654	1,673,125,894
Longest scaffold	2,207,537	1,964,558	698,791	9,386,848	1,784,514	67,529	2,790,175	715382	2,720,304
Shortest scaffold	854	100	100	1000	5001	100	292	205	558
Number of scaffolds > 1 K nt	79,923 (99.9%)	5,788 (48.4%)	142,882 (17.8%)	4,471 (99.9%)	8,766 (100.0%)	393,685 (22.5%)	38,704 (58.9%)	44,921 (60.2%)	20,547 (100%)
Number of scaffolds > 1 M nt	67 (0.1%)	60 (0.5%)	0 (0.0%)	98 (2.2%)	27 (0.3%)	0 (0.0%)	164 (0.2%)	0 (0%)	95 (0.5%)
Mean scaffold size	23,309	46,671	1,764	80,338	81,655	916	25,269	35,262	81,425
Median scaffold size	1,697	824	402	3,622	13,763	258	1,284	13,722	22,134
50% scaffold length	200,099	401,319	14,455	1,870,055	264,327	2,651	343,373	100,161	312,020
Sequencing coverage	322 X	155 X	40 X	8.87 X	11 X	32 X	319 X	209.5 X	60 X
Sequencing Technology	Illumina + PacBio	Illumina	454 + Illumina	Sanger	Sanger	Illumina	Illumina	Illumina	Illumina + PacBio

54 212
55 213
56
57 214
58
59
60
61
62
63
64
65

Table 5: Comparison of genome assembly statistics for molluscan genomes

1
2
3
4 **215 Around 10% of repetitive elements are transposons**

5
6 **216** Initial masking of *L. fortunei* genome was done using RepeatMasker program
7
8
9 **217** (RepeatMasker, RRID:SCR_012954) [30] with parameter *-species bivalves* and masked 3.4% of
10
11 **218** the total genome. This content was much lower than the masked portion of other molluscan
12
13
14 **219** genomes: 34% in *C. gigas* [26] and 36% in *M. galloprovincialis* [25], suggesting that the fast
15
16 **220** evolution of interspersed elements limits the use of repeat libraries from divergent taxa [31].
17
18
19 **221** Thus, we generated a *de novo* repeat library for *L. fortunei* using the program RepeatModeler
20
21 **222** (RepeatModeler, RRID:SCR_015027) [32] and its integrated tools (RECON [33], TRF [34], and
22
23
24 **223** RepeatScout [35]). This *de novo* repeat library was the input to RepeatMasker together with the
25
26 **224** first masked genome draft of *L. fortunei*, and resulted in a final masking of 33.4% of the genome.
27
28
29 **225** Even though more than 90% of the repeats were not classified by RepeatMasker
30
31 **226** (**Supplementary Table S2**), 8.85% of the repeats were classified as LINEs, Class I transposable
32
33 **227** elements. In addition, large numbers of reverse-transcriptases (824 counts, Pfam RVT_1
34
35
36 **228** PF00078), transposases (177 counts, Pfam HTH_Tnp_Tc3_2 PF01498), and integrases (501
37
38 **229** counts, Pfam Retroviral integrase core domain PF00665) and other related elements were
39
40
41 **230** detected; over 98% of these had detectable transcripts.

42
43 **231**
44
45 **232 More than 30,000 sequences identified by gene prediction and automated**
46
47 **233 annotation.**

48
49
50 **234** To annotate the golden mussel genome, we sequenced a number of transcriptomes (**Table S1**),
51
52
53 **235** *de novo* assembled (**Table 2**) and aligned these genomes to the genome scaffolds, and created
54
55 **236** gene models with the PASA pipeline [36]. These models were used to train and run the *ab initio*
56
57
58 **237** gene predictor AUGUSTUS (Augustus: Gene Prediction, RRID:SCR_008417) [37]
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

(Supplementary Figure S1). The complete gene models yielded by PASA [36] were BLASTed (e-value 1e-20) against the Uniprot database (UniProt, RRID:SCR_002380) 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 [37] 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 [37] parameters. Once the gene models were predicted, a final step was performed by using the PASA pipeline [36] once again in the *update* mode (parameters -c -A -g -t). This final step compared the 55,638 gene models predicted by AUGUSTUS [37] with the 40,780 initial transcript-based gene-structure models from PASA [36] 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 S2**) re-mapping, and 67% were annotated by homology searches against Uniprot or NCBI NR (**Table 6**).

Table 6: 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 minimum 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) [38] between *L. fortunei* proteins and the total protein set predicted for seven other mollusks: the mussels *Mytilus galloprovincialis*, *M. philippinarum* and *B. platifrons*, the pacific oyster *C. gigas*, the pearl oyster *Pinctada fucata*, 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 mussels analyzed. A total of 6,788 orthologs groups are shared among the five mussel species.

Of all the orthologous found for the total 8 species, 154 groups are composed of single-copy orthologs containing one representative protein sequence of each species. These sequences

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

were used to reconstruct a phylogeny: the 154 single-copy orthologs sequences were concatenated and aligned with CLUSTALW [39] with a resulting alignment of 117,787 sites in length (Figure 3B).

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 8 molluscan species. The complete protein sets of *L. fortunei*, *C. gigas*, *P. fucata*, *L. gigantea*, *M. galloprovincialis*, *H. discus*, *M. philippinarum*, *B. platifrons* (**Supplementary 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 seven mollusks, according to a model based on the Poisson cumulative distribution. Bonferroni correction ($p \leq 0.05$) was applied for false discovery and absolute frequencies of Pfam-assigned-domains were initially normalized by the total count number of Pfam-assigned-domains found in *L. fortunei* to compensate for discrepancies in genome size and annotation bias.

For *L. fortunei*, the annotation against Pfam.A classified 40127 domains in 24513 gene models of which 83 and 62 were respectively expanded or contracted in comparison with the other mollusks (**Supplementary Table S4 and S5; Figure 4A**). The 83 overrepresented domains were further analyzed for functional enrichment using domain-centric Gene Ontology (**Figure 4B**). The analysis shows a prominent expansion of binding domains in *L. fortunei*, such as Thrombospondin (TSP_1), Collagen, Immunoglobulins (Ig, I-set, Izumo-Ig 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 [40, 41, 42, 43, 44]. An evolutionary

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

299 pressure towards the development of a diversified innate immune system is also suggested by
300 the high amount of Leucine Rich Repeats (LRR) and Toll/interleukin-1 receptor homology
301 domains (TIR), both belonging to Toll Like Receptors (TLRs). Death, another over-represented
302 PFAM, is also part of TLR signalling, being present in several docking proteins such as Myd88,
303 Irak4 and Pelle [45]. Interestingly, Blast analysis of *L. fortunei* gene models against Uniprot
304 identified two types of TLRs: (i) 141 sequences with similarity to single cysteine clusters TLRs
305 (scc) typical of vertebrates, and (ii) 29 sequence hits with the multiple cysteine cluster TLRs
306 (mcc) typical of *Drosophila* [46]. Phylogenetic analysis of all sequences (**Supplementary**
307 **Figure S2**) shows evidence for TLRs clade separation in *L. fortunei*; the scc TLRs exhibit a
308 higher degree of amino acid changes, higher molecular evolution, and diversification than the
309 mcc TLRs.

Curiously, protein families involved in toxin metabolism, especially glutathione based
processes and sulfotransferases are contracted in the *L. fortunei* genome (**Table S5**).

Final considerations

Here we have described the first version of the golden mussel complete genome and its
automated gene prediction that were funded through a crowdfunding initiative in Brazil. 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 [47], or genes related to the reproductive system that have been shown to be
very effective for invertebrate plague control [48]. The golden mussel genome and the predicted
proteins are available for download in the Gigabase repository and the scientific community is
welcome to further curate the gene predictions.

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

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

325

326 **Availability of supporting data**

327 *Limnoperna fortunei*'s genome and transcriptome data are available in the Sequence
328 Read Archive (SRA) as BioProject PRJNA330677 and under the accession numbers
329 **SRR5188384, SRR5195098, SRR518800, SRR5195097, SRR5188315, SRR5181514**. Also
330 this Whole Genome Shotgun project has been deposited in the DDBJ/ENA/GenBank under
331 accession number NFUK00000000. The version described in this paper is version
332 NFUK01000000.

333

334 **Additional files**

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

336 **Supplementary Table S2:** RepeatMasker classification of repeats predicted in *L. fortunei*
337 genome.

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

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

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

342 **Supplementary Table S6:** Fantasy names given to *L. fortunei* genes and proteins from the
343 backers that have supported us through crowdfunding (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

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* genome.

List of Abbreviations

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

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.

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

367 We especially want to thank the 346 backers that supported the sequencing of the golden mussel
368 through crowdfunding, in a 2013 campaign that raised U\$ 20,000.00 (www.catarse.me/genoma).
369 We decided to give fantasy names to the genes and proteins that we found in the genome, to
370 thank the backers for their support. The name list is available in **Supplementary Table S6**.

371

372 Consent for publication

373 Does not apply.

374 Ethics approval

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

378

379

380 References

381 1. Pastorino G, Darrigran G, et al., *Limnoperna fortunei* (Dunker, 1857) (Mytilidae), nuevo
382 bivalvo invasor em águas Del Rio de la Plata. *Neotropica*. 1993;39:101–2.
383
384 2. Uliano-Silva M, Fernandes F da C, Holanda IBB, Rebelo MF. Invasive species as a threat
385 to biodiversity: The golden mussel *Limnoperna fortunei* approaching the Amazon River
386 basin. In: *Exploring Themes Aquat Toxicol Alodi, S*, editor. Research Signpost; 2013.
387
388 3. Boltovskoy D, Correa N. Ecosystem impacts of the invasive bivalve *Limnoperna fortunei*
389 (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

390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412

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.

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

413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435

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.

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

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

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.

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

459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481

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.

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

29. 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.

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

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

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

484

485 32. Smith AFA, Hubley R. RepeatModeler Open-1.0. [Internet]. 2014. Available from:
486 <http://www.repeatmasker.org>

487

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

490

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

493

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

496

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

500

501 37. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped
502 cDNA alignments to improve de novo gene finding. Bioinforma Oxf Engl. 2008
503 1;24(5):637–44.

504

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

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

39. 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.

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

41. 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.

42. 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.

43. 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.

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

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

528 45. Park HH, Lo YC, Lin SC, Wang L, Yang, JK, Wu H. The death domain superfamily in
530 intracellular signaling of apoptosis and inflammation. *Annu. Rev. Immunol.* 2007, 25,
531 561-586.

532 46. Leulier F, Lemaitre B..Toll-like receptors—taking an evolutionary approach. *Nature*
533 *Reviews Genetics*, 2008, 9.3: 165-178.

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

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

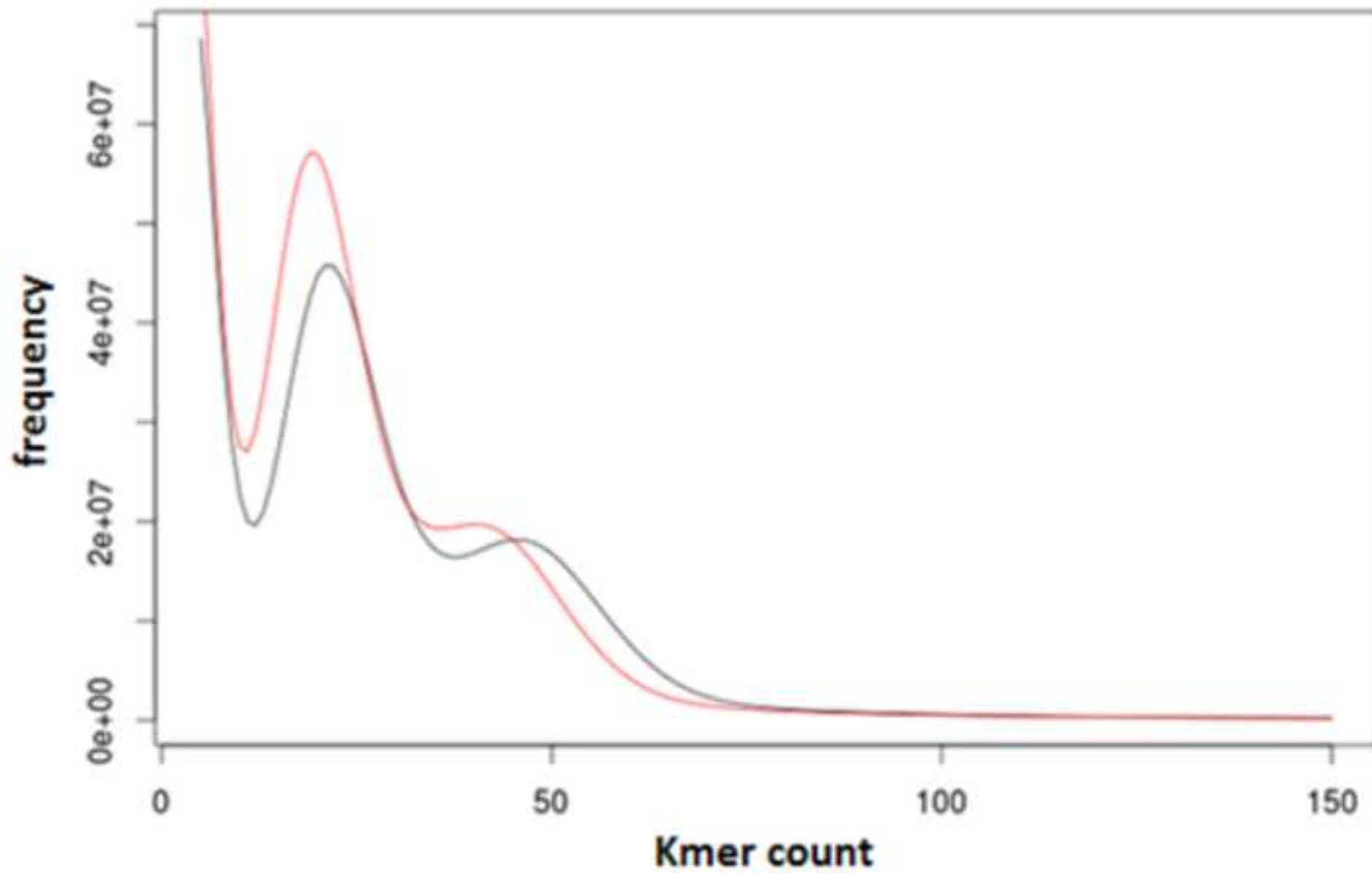
Figure 1: K-mer distribution of *Limnoperna fortunei* Illumina DNA reads (Table 1).

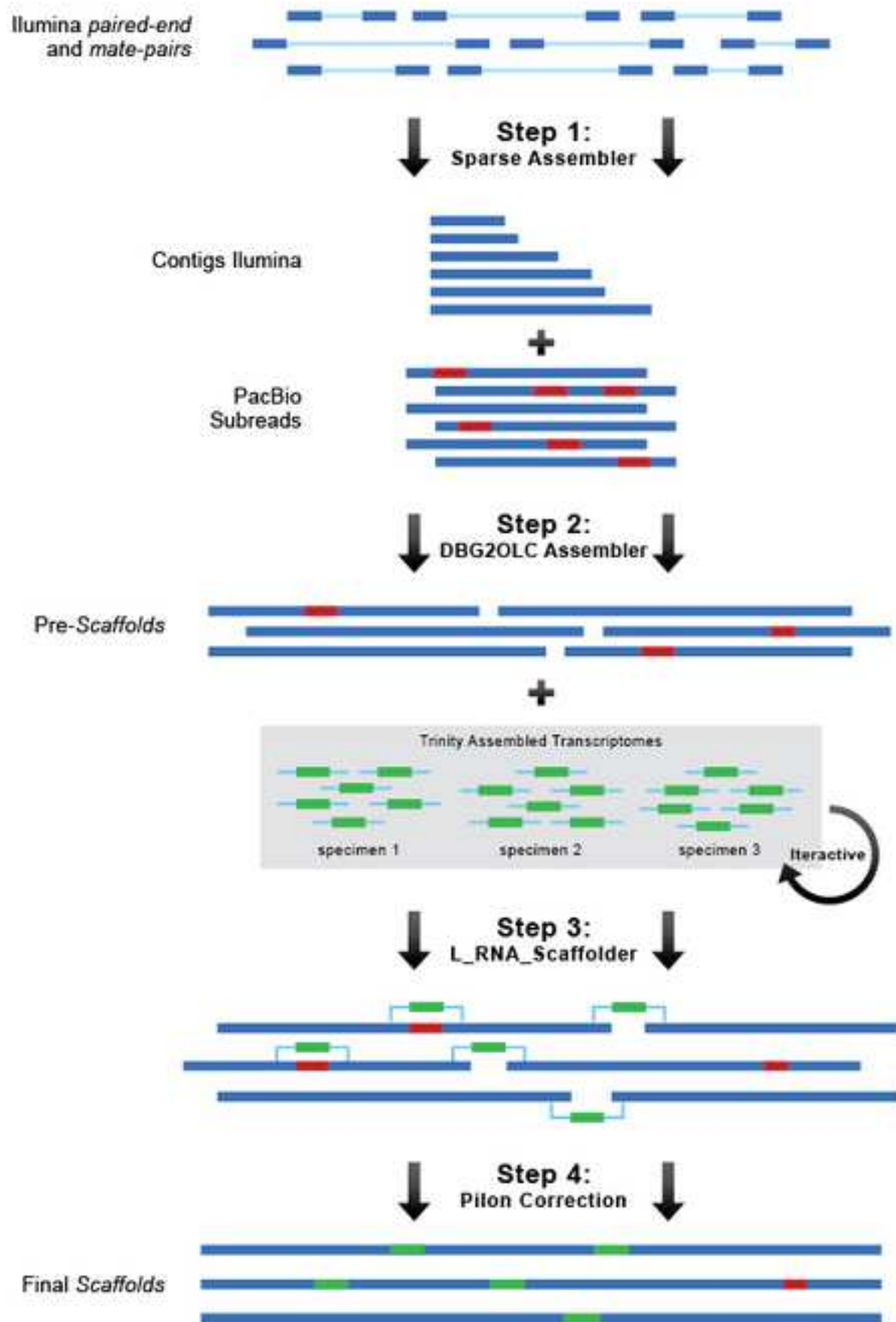
Figure 2: Hierarchical assembly strategy employed for the golden mussel genome assembly. Trimmed Illumina reads were assembled to the level of contigs with Sparse Assembler algorithm (**Step 1**). Then, Illumina contigs and PacBio reads were used to build scaffolds with DBG2OLC assembler, that anchors 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.

Figure 3A: Orthology assigned with OrthoMCL for the total set of proteins predicted from five mussel genome projects. Outside the Venn diagram its represented the species name and bellow it is the number of proteins / number of clustered proteins / number of clusters. **B:** Phylogeny of the concatenated data set using 154 single-copy orthologs extracted from eight molluscan genomes. Maximum likelihood tree nodes with 100 bootstrap resampling.

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

Figure 4: Gene family representation analysis in the *L. fortunei* genome. Panel A. PFAM hierarchical clustering, heatmap. Features were selected according to a model based on the Poisson cumulative distribution of each PFAM count in the golden mussel genome vs the normalized average values found in the other eight molluscan genomes (Bonferroni correction, $P \leq 0.05$). Transposable elements were included in the analysis. Colors depict the log2 ratio between PFAM counts found in each single genome and the corresponding mean value. The hierarchical clustering used the average dot product for data matrix and complete linkage for branching. Legend: Lf, *L. fortunei*; Mg, *M. galloprovincialis*; Pf, *P. fucata*; Lg, *L. gigantea*; Cg, *C. gigas*; Bp *B. platifrons*; Mp, *M. philippinarum*; Hd, *H. discus*. **Panel B. Gene ontology analysis of expanded gene families (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).





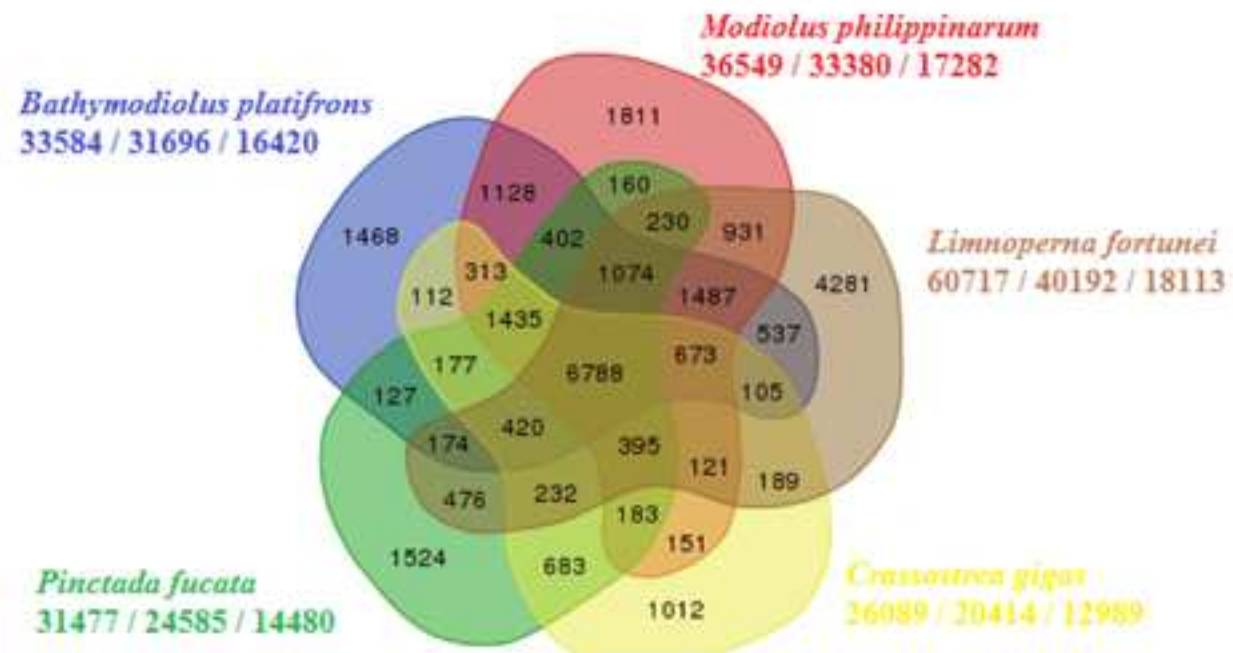
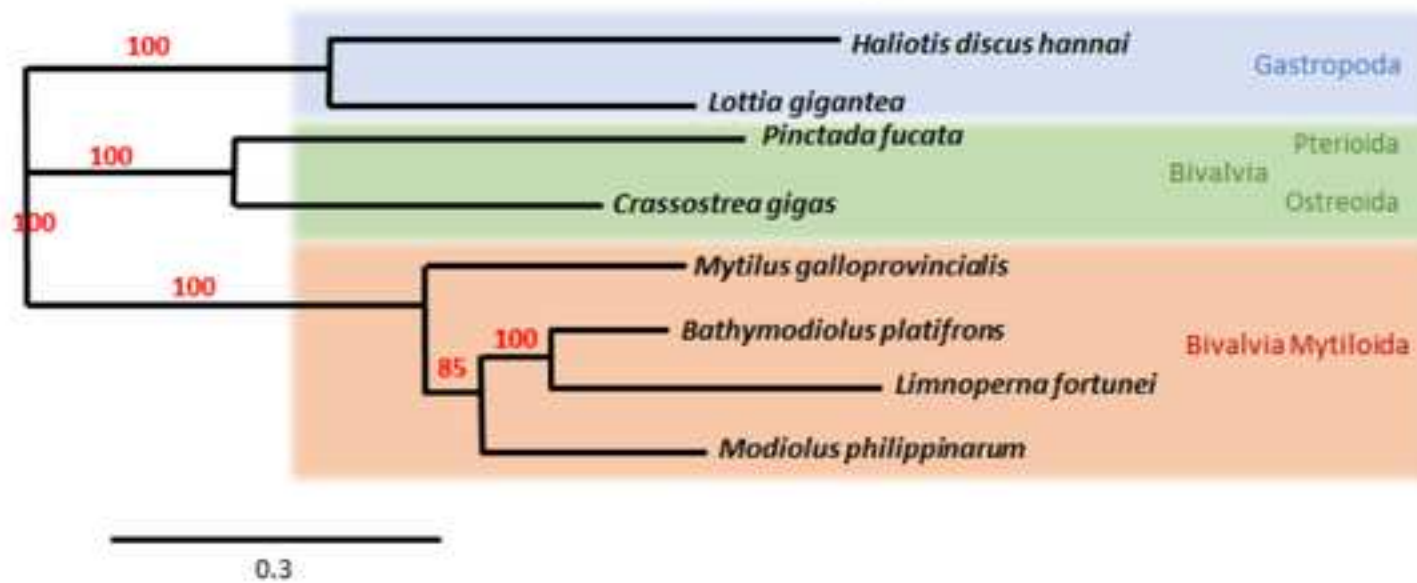
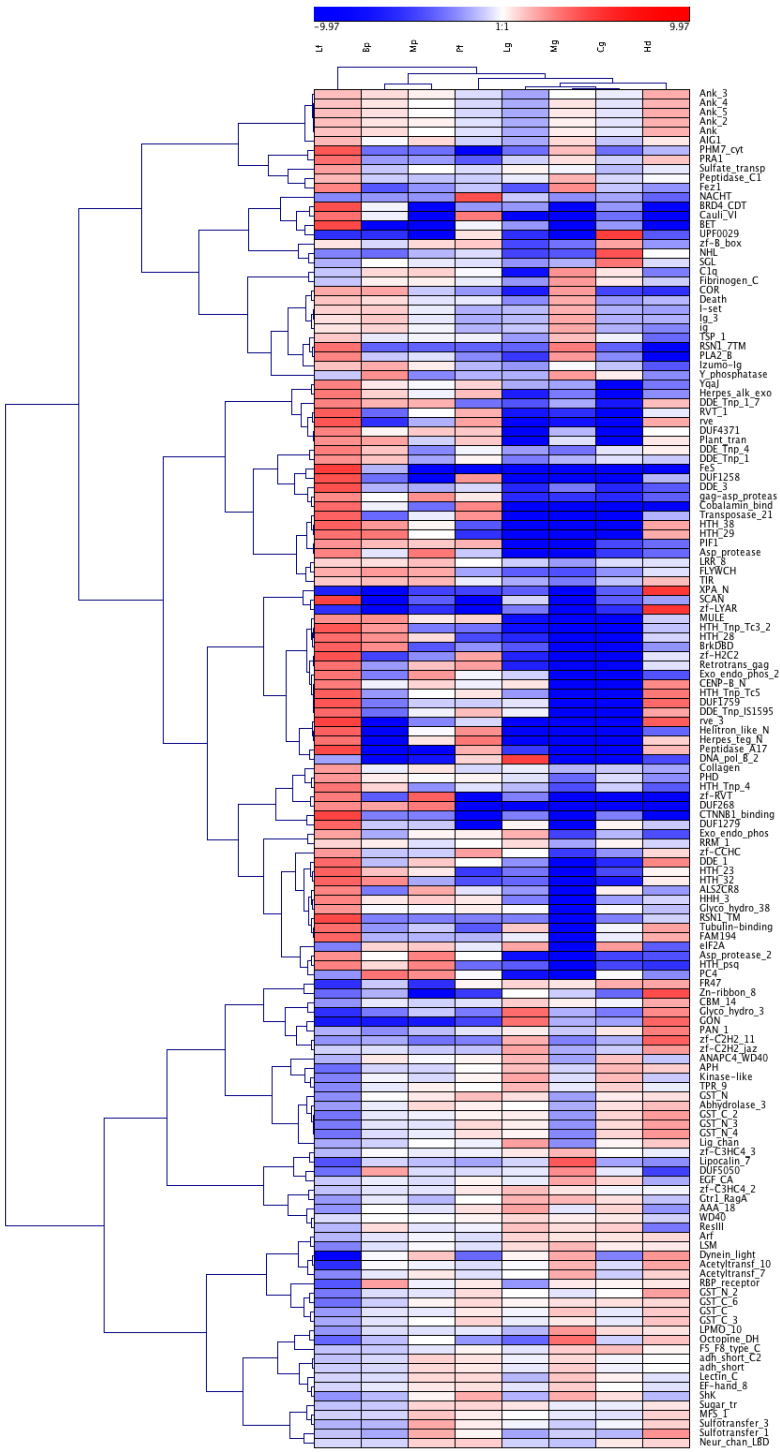
A**B**

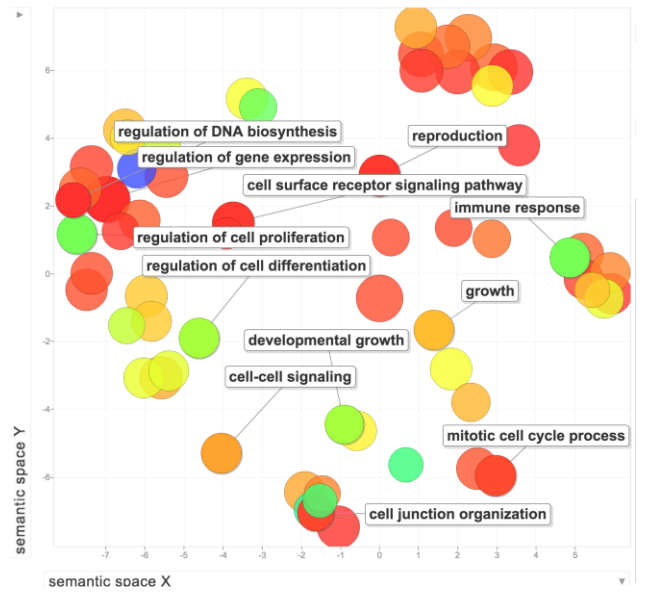
Figure 4


[Click here to download Figure figure4ed.pptx](#)

A




B





Click here to access/download
Supplementary Material
TableS1.docx

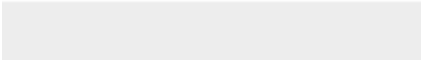





Click here to access/download
Supplementary Material
Figure-S1.png

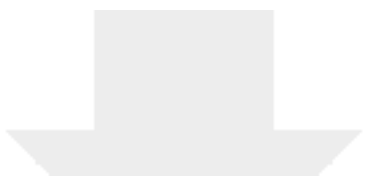


Click here to access/download
Supplementary Material
TableS2.docx




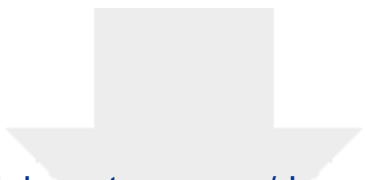


Click here to access/download
Supplementary Material
TableS3.docx




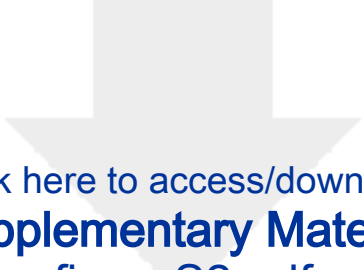
Click here to access/download
Supplementary Material
TableS4.xlsx



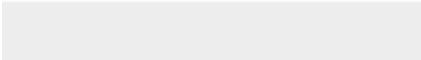



Click here to access/download
Supplementary Material
TableS5.xlsx





Click here to access/download
Supplementary Material
figureS2.pdf





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

