GigaScience

The genome of golden apple snail Pomacea canaliculata provides insight into stress tolerance and invasive adaptation
--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00030
Full Title:	The genome of golden apple snail Pomacea canaliculata provides insight into stress tolerance and invasive adaptation
Article Type:	Research
Abstract:	Background: The golden apple snail (Pomacea canaliculata) is a worldwide fresh water snail listed in the top-100 worst invasive species, and a noted agricultural and quarantine pest causing huge economic loss, characterized with fast growth, strong stress tolerance, high reproduction rate, and adaptation to a broad range of environments. Results: Here, we used long-read sequencing to produce a 440-Mb high-quality chromosome level assembly for P. canaliculata genome. In total, 50 Mb (11.4%) repeat sequences and 21,533 gene models were identified in the genome. Major findings of this study include the recent explosion of DNA/hAT-Charlie TEs, the expansion of P450 gene family and the constitution of cellular homeostasis system, contributing to the ecological plasticity in the stress adaptation. In addition, the perivitellin gene expansion and high transcriptional level in ovary promotes the function of nutrients supplying and defense ability in the eggs. Furthermore, the gut metagenome also encodes rich genes for food digestion and xenobiotics degradation. Conclusions: These findings collectively provide novel insight into the molecular mechanisms of the ecological plasticity and high invasiveness. Our results not only strengthen the understanding of molluscs genomics and biological invasion, but also benefit preventing the invasion of apple snail and transmission of pathogenetic parasites.
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics 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. Have you included all the information requested in your manuscript?	Yes
Resources 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	Yes

Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials 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?	Yes

1 **The genome of golden apple snail *Pomacea canaliculata* provides insight into**

2 **stress tolerance and invasive adaptation**

3 Conghui Liu[1]*, Bo Liu[1]*, Yuwei Ren[1]*, Yan Zhang[1]*, Hengchao Wang[1], Shuqu Li[1],

4 Fan Jiang[1], Lijuan Yin[1], Guojie Zhang[2], Wanqiang Qian[1†] and Wei Fan[1†]

5 [1]Agricultural Genomic Institute, Chinese Academy of Agricultural Sciences,

6 Shenzhen, Guangdong, 518120, China.

7 [2]BGI-Shenzhen, Shenzhen, Guangdong, 518083, China

8

9 Conghui Liu: rapherlch@163.com; Bo Liu: lb_bobo@aliyun.com; Yuwei Ren:

10 xiaoshudaxia@126.com; Yan Zhang: milrazhang@163.com; Hengchao Wang:

11 wanghengchao000@qq.com; Shuqu Li: lishuqu1234@163.com; Fan Jiang:

12 greatjf@163.com; Lijuan Yin: yinlijuan1005@163.com; Guojie Zhang:

13 guojie.zhang@bio.ku.dk

14 *These authors contributed equally to this work.

15 [†]Correspondence should be addressed to Wanqiang Qian (qianwanqiang@caas.cn) or

16 Wei Fan (fanwei@caas.cn).

17

18 **Abstract**

19 **Background:** The golden apple snail (*Pomacea canaliculata*) is a worldwide fresh

20 water snail listed in the top-100 worst invasive species, and a noted agricultural and

21 quarantine pest causing huge economic loss, characterized with fast growth, strong

22 stress tolerance, high reproduction rate, and adaptation to a broad range of

23 environments.

24 **Results:** Here, we used long-read sequencing to produce a 440-Mb high-quality

25 chromosome level assembly for *P. canaliculata* genome. In total, 50 Mb (11.4%)

26 repeat sequences and 21,533 gene models were identified i~~n the genome~~. Major

27 findings of this study include the recent explosion of DNA/hAT-Charlie TEs, the

28 expansion of P450 gene family and the constitution of cellular homeostasis system,

29 contributing to the ecological plasticity in the stress adaptation. In addition, the

30 perivitellin gene expansion and high transcriptional level in ovary promote the

31 function of nutrients supplying and defense ability in the eggs. Furthermore, the gut

32 metagenome also encodes rich genes for food digestion and xenobiotics degradation.

33 **Conclusions:** These findings collectively provide novel insight into the molecular

34 mechanisms of the ecological plasticity and high invasiveness. ~~Our results not only~~

35 ~~strengthen the understanding of molluscs genomics and biological invasion, but also~~

36 ~~benefit preventing the invasion of apple snail and transmission of pathogenetic~~

37 ~~parasites.~~

38 **Keywords:** golden apple snail, *Pomacea canaliculata*, genome, adaptive evolution,

39 stress tolerance, P450, reproduction, perivitelline, metagenome

40 **Background**

41 The golden apple snail *Pomacea canaliculata* (family Ampullariidae; Order

42 Architaenioglossa) is a fresh water snail listed in the 100 of the world's worst invasive

43 species [1], and considered as a noted agricultural and quarantine pest worldwide [2].

2

44 Native to the tropical and subtropical South American, the *P. canaliculata* gradually

45 spread to the non-indigenous region, such as Southeast and East Asia [3], Africa [4],

46 North America [5], Oceania [6] and even Europe [7], and the successful

47 biological invasion ~~was due to~~ polyphagous feeding habits [8], voracious appetite [9],

48 broad environmental adaptability [10] and rapid growth and high rate of reproduction

49 [11]. Besides the ecological impact, the *P. canaliculata* ravaged a wide range of crops

50 including grain, fruit and vegetable [12], causing severe economic loss each year as a

51 result of yield loss, replanting cost and the funds of control

52 (https://www.cabi.org/isc/datasheet/68490). More seriously, *P. canaliculata* has

53 involved in the transmission of a human fatal disease, Eosinophilic meningitis, that

54 firstly appeared in East Asia where people take them as food frequently [13]. During

55 this pathophoresis, *P. canaliculata* acts as an important intermediate host of

56 pathogenic parasite *Angiostrongyulus cantonensis*, and the range of infectious regions

57 is still expanding, causing great challenge to human health [14, 15].

58 Molluscs is a highly diverse group and second only to arthropods in species

59 number [16], and the high biodiversity makes molluscs an excellent model to address

60 the issues such as biogeography, adaptability and evolution process [17], and the

61 worldwide invasive *P. canaliculata* provides valuable potential in these fields [18]. As

62 a primitive circumtropical species, *P. canaliculata* possesses strong ecology plasticity

63 to hold advantage on plenty of aspects, including low temperature resistance [19],

64 drought tolerance [20], which contributes to succeed in resource acquisition over the

65 competitive species. Additionally, *P. canaliculata* is tolerant with heavy metal

66 contamination. When living in contaminated water, its gill is enriched of high

67 concentration of heavy metal and histopathological changes in digestive tract is

68 detected, however, with extremely low mortality rate [21]. For protection of embryos,

69 the conspicuous coloration and neurotoxic lectin could confer the eggs a survival

70 advantage and defense against the potential predator [22]. Moreover, the

71 immune-neuroendocrine system can also be detected in *P. canalicula*, demonstrates

72 by the existence of a specific immune memory after the bacterial challenge [23, 24],

73 broadening the studies of invertebrate immunology.

74 During the past years, the genomic features of *P. canalicula* have been increasingly

75 studied. After the discovery of 14 pachytene bivalents in the karyotype [25],

76 molecular markers were identified to investigate the genetic diversity of *P.*

77 *canaliculata* population, including 369 amplified fragment length polymorphism

78 (AFLP) locis [26], 16,717 simple sequence repeats (SSR) [27, 28] and 15,412

79 single-nucleotide polymorphisms SNPs [29]. In addition, multiple transcriptome

80 analyses have been performed to investigate the adaptation, invasion and immune

81 mechanisms. For instance, Sun et al. reported 128,436 unigenes based on a de novo

82 assembly of Illumina reads [29], transcriptome changes in response to heat stress and

83 starving incubation was used to characterize invasive and adaptive abilities [30, 31], a

84 transcriptome analysis between invasive *P. canaliculata* and indigenous

85 *Cipangopaludina cahayensis* provides insights into biological invasion [28], and 402

86 immune-related differentially expressed genes (DEGs) by Lipopolysaccharide (LPyS)

87 challenge were used to explore the mechanisms against pathogens [32]. Furthermore,

4

88 proteomics tools such as Isobaric Tags For Relative, Absolute Quantitation (iTRAQ),

89 and Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS) were also

90 applied in the study of protein expression for the estivation and oviposition [33, 34],

91 together providing plentiful omics-data for the functional analysis of *P. canalicula*.

92 However, researches at whole genome level in *P. caniculata* still lags far behind

93 other molluscs species, due to the lack of a high-quality reference genome. By far,

94 multiple draft genomes of molluscs have been published, such as Califonia sea hare

95 [35], Pacific oyster [36], Pearl oyster [37], owl limpet [38], California two-spot

96 octopus [39], deep-sea mussel [40], *Biomphalaria* snails [41], greatly promoting the

97 research of molluscs genomics. In this study, we present a chromosome-level genome

98 assembly of *P. canalicula* with high-quality gene annotation, transcriptome data from

99 several tissues and under various conditions, as well as the metagenomic data from

100 the intestinal tracts, all of which were then applied to study the species-specific

101 invasive characters, such as cellular homeostasis system underlying strong stress, and

102 color and nutrient of the eggs. Our data will not only strengthen the understanding of

103 evolutionary mechanisms of molluscs and molecular basis of biological invasion, but

104 also foster developments to control the invasion of *P. canalicula* and interrupt the

105 transmission of pathogenetic nematode parasites.

106 **RESULTS**

107 **Complete genome assembly at chromosome level**

108 We generated 26.6 Gb (60.1 X) PacBio SMRT raw reads with average read length

109 10.1 Kb, and 291 Gb (652.4 X) Illumina HiSeq paired-end reads with read length

110 150-250 bp, using DNA extracted from one single adult *P. caniculate* (Table S1).

111 The 24.4 Gb (55.4 X) clean PacBio SMRT reads that passed quality filtering were

112 assembled by smartdenovo (https://github.com/ruanjue/smartdenovo), giving rise to

113 an assembly of 1234 raw contigs with total length 473.6 Mb and N50 length 1.0 Mb.

114 After filtering of alternatively heterozygous contigs, 745 resulting contigs with total

115 length 440.1 Mb and N50 length 1.1 Mb were taken as the final contigs. Previous

116 karyotype research shown that haploid P. *caniculate* genome consist of 14

117 chromosomes [25]. Based on Hi-C data, 439.5 Mb (99.9%) final contigs were

118 anchored and oriented into 14 large scaffolds, each corresponding to a natural

119 chromosome (Figure 1a and Figure 1b), with the longest 45.4 Mb and shortest 27.2

120 Mb. This assembly quality is much better than the other published mollucan genomes

121 so far (Table 1). Besides the length and continuity of assembled sequences, another

122 important aspect for evaluating genome assembly is the ratio of genome coverage.

123 With an estimated genome size of 446 Mb based on distribution of k-mer frequency

124 [42] (Figure S1), ~98.6 % of the genome has been assembled in *P. caniculata*. To

125 further confirm the accuracy and completeness of the assembly, we mapped the

126 Illumina shotgun reads to the assembled reference genome. Significantly, 97% and 95%

127 of the genome-derived and transcriptome-derived reads could be aligned to the

128 reference genome, respectively, ~~suggesting no obvious bias for sequencing and~~

129 ~~assembly~~. Additionally, the mitochondrial genome of *P. caniculata* was also

130 assembled as a single contig with 15,707 bp in length, which has 99.9 % sequence

131   identity to the published mitochondrial genome (GenBank: KJ739609.1) (Figure S2).

132   ~~The high-quality reference genome provides a good foundation for gene annotation.~~

133   The protein-coding genes were predicted on the reference genome by EVM,

134   integrating evidences from *de novo* prediction, transcriptome and homology data. In

135   total, 21,533 gene models were predicted as the reference gene set, with coding

136   regions spanning ~32.2 Mb (7.3 %) of the genome (Table 1 and Table S2). The

137   distribution of CDS length in *P. canaliculata* is similar to the closely related species

138   (Figure 1c). Overall, 97.5 % of the reference genes were supported by transcriptome

139   data, and 98.0 % of eukaryote core genes from OrthoDB (http://www.orthodb.org/)

140   were identified in the reference gene set by BUSCO, comparable to the other

141   published mollucan genomes (Table 1). For the functional annotation, a total of

142   19,815 (91.9 %) reference genes were annotated by at least one functional database.

143   Specifically, 15,662 (72.7 %), 13,769 (63.4 %), 17,081 (79.3 %), 18,847 (87.5 %) and

144   17,003 (79.9 %) reference genes were annotated with eggNOG, KEGG, NR, Interpro

145   and Uniprot database, respectively (Figure S3).

146   **Signs of Adaptive Evolution in P. canaliculata Genome**

147   To gain insight into evolutionary perspective of *P. canaliculata*, the phylogenetic tree

148   was built based on 471 high-confidence single-copy ortholog genes from seven

149   related species (*P. canaliculata, L. gigantea, A. california, B. glabrata, C. gigas, O.*

150   *bimaculoides and L. anatina*) by Phyml [43] and the divergence time was estimated

151   using mcmctree [44]. The result shows that *P. canaliculata* diverged from the ancestor

152 of *B. glabrata* and *A. California* 290 million years ago (Mya), and from *L. gigantea*

153 415 Mya (Figure 2a).

154 Then, the molluscan ortholog genes were investigated for adaptive evolution.

155 Utilizing pairwise protein sequence similarities, the gene family clustering was

156 conducted by orthfinder [45]. A total of 152,878 reference genes from the seven

157 species were clustered into 68,942 ortholog groups, amongst which 13,805 ortholog

158 groups with at least two genes each. In *P. canaliculata*, we identified 9,626 ortholog

159 groups, amongst which 117 and 5,462 ortholog groups undergone species-specific

160 expansion, thus may play important roles in adaption to the environment as an

161 invasive species. The functions of these orthologous groups are mainly related to

162 glycan biosynthesis, digestive, endocrine, signal transduction, immune, or

163 carbohydrate metabolism and so on (Figure S4).

164 The high-coverage genome assembly enables a comprehensive analysis of the

165 transposable elements (TEs), which plays multiple roles in driving genome evolution

166 in eukaryotes [46]. In total, we identified 49.6 Mb TE sequences in the assembled *P.*

167 *canaliculata* genome (Table 1), including 3.4 Mb long terminal repeats (LTR), 27.2

168 Mb long interspersed elements (LINE), 17.5 Mb DNA transposons and 1.5 Mb short

169 interspersed elements (SINE). Next, we analyzed the divergence rate of TEs for each

170 class of TEs among the available sequenced mollusk genomes, interestingly, only the

171 results of DNA transposons showed a unique peak at ~4% divergence rate for *P.*

172 *canaliculata* and *C. gigas* (Figure 2b), indicating a recent explosion of DNA

173 transposons in these two species. More than half of the DNA transposons belong to

8

174 the DNA/hAT-Charlie TE family, which is ~22.7% of total DNA/hAT-Charlie TEs in

175 the genome. TEs are powerful facilitators of evolution by generating "evolutionary

176 potential" to introduce small adaptive changes within a lineage, and the importance of

177 TEs to stress responses and adaptation has been reported in numerous researches [47,

178 48]. The recent explosion of DNA/hAT-Charlie TEs in *P. canaliculata* could also play

179 important roles to promote the potential plasticity in the stress adaptation.

180 **Investigation of Cellular homeostasis system underlying strong stress adaptation**

181 Homeostasis system plays a crucial role in the stress adaptability, providing the

182 molecular basis in re-establishing the dynamic equilibrium after the challenge of

183 various environmental stressors, including temperature, air exposure, anthropogenic

184 pollution and pathogens [49]. In the present study, we addressed three constituent

185 parts of the cellular homeostasis system, which contributes to the successful

186 ecological plasticity of *P. canaliculata* (Figure 3). Transcriptome data of the

187 hemocytes after stimulus (cold, heat, heavy and air exposure) was also sequenced and

188 analyzed to address the potential roles of the genes in Cellular homeostasis system.

189 Unfolded protein response (UPR) system makes the central part of protein

190 homeostasis [50]. Heat shock proteins (HSPs) acts as molecular chaperones to

191 maintain the correct folding, and heat shock transcription factor 1 (HSF1) are

192 responsible for the transcriptional induction of HSPs [51]. In *P. canaliculata* genome,

193 13 HSP70s，6 HSP90s, 7 HSP40s and 11 HSFs were identified (Table S3), and the

194 expression of HSP90s and HSFs were highly induced in response to the stress of heat,

195 cold, heavy metal and air exposure (Table S4). Inositol-requiring protein 1 (IRE1),

196 protein kinase RNA-like ER kinase (PERK), and activating transcription factor 6

197 (ATF6) are three mediators recruited by endoplasmic reticulum (ER) to regulated the

198 UPR [52]. We found putative coding genes of the three core mediators, their

199 respective downstream transcription factors, and the corresponding recognition

200 chaperons in *P. canaliculata* genome (Table S3).

201 Xenobiotic biotransformation system helps the mollusc adapt to toxicants, especially

202 the pesticide in aquatic environments [53]. Manual annotation on this genome

203 identified 157 cytochrome P450s (CYP450s), 15 flavin-containing monooxygenases

204 (FMOs), 53 glutathione S-transferases (GSTs) and 105 ATP binding cassette (ABC)

205 transporters, most of which showed an up-regulation in expression under stress (Table

206 S3, Table S4). These proteins are evidenced to function in contaminant detecting,

207 conjugative modification and expulsion for xenobiotic detoxification [54-56].

208 Massive production of reactive oxygen species (ROS) and reactive oxygen

209 intermediates (ROI) induced by stress lead to many pathological conditions, and

210 antioxidant system protect the organism from superoxide [57]. Four main antioxidant

211 enzyme classes, namely superoxide dismutase (SOD), catalase (CAT), peroxidase

212 (Prx), and glutathione peroxidase (GPX), were found in the *P. canaliculata* with an

213 elevating global expression in response to stress (Table S3, Table S4).

214 Apoptosis is a process of cell death when sensing stress and the regulation of

215 apoptosis maintains the dynamic homeostasis of internal environment. In *P.*

216 *canaliculata*, we propose the existence of both intrinsic and extrinsic apoptotic

217 signaling pathways, evidenced by the presence of homologous genes involve in both

218 pathways. It seems these two pathways could be activated by cytochrome C and

219 tumor necrosis factor receptor (TNFR), respectively (Table S3). The inhibitors of

220 apoptosis, such as XIAP, Bcl2 and Bak, are also detected with an increased expression

221 in response to the stress (Table S4), which are expected to delay the apoptosis process

222 and the cell death in stress response.

223 **The expansion of P450 gene family contribute to stress tolerance**

224 Cytochromes P450 (CYP) enzymes are a monooxygenase family with highly diverse

225 structures and functions, broadly identified in all kingdoms of life [58]. P450s

226 catalyze the reductive scission of molecular oxygen, and are responsible for the

227 synthesis and metabolism of various molecules, including drugs, hormones,

228 antibiotics, pesticides, carcinogens and toxins [59]. The synthesized hormones, such

229 as glucocorticoids, mineralocorticoids, progestins, and sex hormones, are critical to

230 stress response, growth and reproduction, and the endogenous and exogenous

231 chemical metabolism helps the host combat with the toxic compounds [60].

232 We found the *P. canaliculata* CYP gene family had greater level of expansion

233 compared to the other molluscs. We identified 157 genes in the genome of *P.*

234 *canaliculata*, and 128, 102, 135, 78, 52 and 94 genes from *A. California, B. glabrata,*

235 *C. gigas, L. gigantean, O. bimaculoides* and *P. fucata* respectively under the same

236 standard (Figure 4a). The expansive trend was also observed, compared with the

237 model species, such as *Homo sapiens* (57), *Mus musculus* (102), *Dario rerio* (94) and

11

238 *Drosophila melanogaster* (94) [61]. The gene expansion was mainly found in CYP2U

239 and CYP3A sub-families, and fewer genes expanded in CYP4F. In mammals, CYP2U

240 plays a role in the metabolism of fatty acid to generate bioactive eicosanoid

241 derivatives, potentially regulating the development of immune function [62]. In *P.*

242 *canaliculata*, 40 genes forged into the CYP2U clade, mainly expressing in

243 hepatopancreas (Figure 4b and Table S5_a, Table S5_b). CYP3A acts as a versatile

244 enzyme metabolizing a wide range of xenobiotics, and the productions promote the

245 growth of various cell types [63]. The 56 CYP3A genes have comprehensive

246 expression in hepatopancreas, gill and kidney (Figure 4b and Table S5_a, Table S5_b).

247 CYP4F possesses epoxygenase activity, metabolizing fatty acid to epoxides to

248 suppress hypertension, pain perception and inflammation [64]. 20 genes were

249 identified in CYP4F, and several CYP4F genes present highly induced expression

250 levels under the stress of cold, heat, heavy metal and air exposure, indicating their

251 critical roles in the stress tolerance (Figure 4b and Table S5_a, Table S5_b).

252 **The perivitellin gene expansion and high transcriptional level in ovary enhance**

253 **reproduction**

254 To adapt to the fast invasion life, besides the strong ability to stress tolerance, the *P.*

255 *canaliculata* possesses a high reproductive rate, and one important contributor is their

256 distinct eggs characterized with abundant nutrients, reddish or pinkish color, aerial

257 oviposition and neurotoxic [22, 34]. In most gastropod eggs, Pervitelline Fluid (PVF)

258 with large amounts of nutrients filled in space between the eggshell and the embryo,

259 is composed of carbohydrates, lipids and proteins termed perivitellins, which is not

260   only responsible for the major supply of material and energy during embryogenesis,

261   but also provide warning pigment and deadly toxicant against the predators [65].

262   Perivitellins of *P. canaliculata* (Pc) have been verified by proteomics approach and

263   was further divided into three categories called Pc Ovorubin (PcOvo), PcPV2, PcPV3,

264   which are all high-density lipoprotein (HDL) [66] (Figure 5a). We totally identified 18

265   perivitellin genes from the *P. canaliculata* genome, compared to 2 and 1 perivitellin

266   genes from *A. california* and *P. fucata* respectively, by aligning the seven reference

267   perivitellin gene sequences (NCBI accession AFQ23940.1, AFQ23939.1,

268   AFQ23938.1, AFQ23945.1, AFQ23937.1, P0C8G7.2, P0C8G6.2) to each genome

269   sequences with the same method (blastn e-value 10^{-20}). It is apparent that the copy

270   number of perivitellin genes was expanded in *P. canaliculata*, and our orthologous

271   and paralogous gene family data by orthoFinder confirmed this. Among the 20

272   perivitellin genes in *P. canaliculate*, there are 2 PcOvo, 13 PcPV2, and 3 unclassified

273   PVFs (Figure 5b and Table S6). The PcOvo carotenoprotein is responsible for the red

274   coloration of the eggs and antioxidant to protect against sun radiation and desiccation

275   [67, 68], while PcPV2 is reported to be neurotoxin implying lethal effect on rodents

276   [22]. The expansion of these genes may enhance the underlying functions of nutrition

277   and protection, offering the eggs an advantage of survival and improve the

278   reproduction rate.

279   The expression of 18 *P. canaliculata* perivitelline genes were detected in 7 tissues,

280   including embryo, testis, ovary, kidney, gill, hepatopancreas and hemocyte. The

281   highest expression of each gene concentrated in embryo and two sexual gland testis

282   and ovary, especially in the ovary (Figure 5b and Table S7), suggesting that their

283   decoding proteins might be of importance in germ cell production and embryo

284   development. Taken together, *P. canaliculata* distinguish its embryo development

285   from other seven species on the preponderance of perivitellin gene number and high

286   expression level, that further promotes corresponding function of nutrients supplying

287   and defense ability and eventually contribute to reproduction.

288   **Gut microbiome plays important roles in stress resistance and food digestion**

289   The gut microbiome is well known as the second genome of animals, which plays key

290   roles in food digestion, immune defense, etc that are essential to the animals. To

291   investigate whether the gut microbiome has influence on the invasive life style, we

292   collected gut digesta samples from 70 adults of *P. canaliculata*, and generated 31 Gb

293   high quality metagenomic data on Illumina HiseqX10 platform. To our knowledge,

294   this is the first high-depth sequencing of snail gut microbiome. A total of 1,142,095

295   non-redundant genes were obtained, with an average open reading frame (ORF)

296   length of 604 bp (Table S8). The taxonomic composition analysis showed that, at the

297   phylum level, Proteobacteria was the predominant, followed by Verrucomicrobia,

298   Bacteroidetes, Firmicutes, Spirochaetes, Actinobacteria, etc. (Table S9_a). At the

299   genus level, the most abundant genera include *Aeromonas*, *Enterobacter*,

300   *Desulfovibrio*, *Citrobacter*, *Comamonas*, *Klebsiella* and *Pseudomonas*. (Table S9_b),

301   most of which were also presented in the snails of *Achatina fulica* [69, 70].

302   It is interesting that some of the most abundant genera such as *Desulfovibrio*,

14

303　*Citrobacter* and *Pseudomonas* were reported to have strong abilities of removing

304　heavy metals, by mechanisms of bioprecipitation and bioabsorption [71-73]. For

305　example, the sulfur-reducing bacteria *Desulfovibrio* produced H_2S that precipitate

306　metals, and therefore reduced the toxic effects of dissolving metals [71]. Based on the

307　KEGG pathway database, the complete sulfate reduction metabolism pathway was

308　identified in the *P. canaliculata* gut microbiome. We suggested that the gut microbes

309　might help *P. canaliculata* to confront with the environmental stress of heavy metals

310　in hash conditions. In addition, a large number of genes in pathways of xenobiotics

311　biodegradation and metabolism were annotated, corresponding to 288 KEGG

312　orthologous groups (KOs) and 21 pathways (Table S10). As many of the pathways

313　such as benzoate degradation, toluene degradation, xylene degradation and steroid

314　degradation could not be identified in the host genome through KO analysis, we

315　suggested that the microbial detoxification abilities may contribute the *P. canaliculata*

316　to resist stresses caused by xenobiotics such as pesticides and environmental

317　pollutants.

318　In view of dietary digestion, the gut microbes were directly involved in breakdown of

319　the cellulose portion, and previous studies have isolated some cellulolytic bacteria and

320　evaluated the cellulolytic enzyme activities [74]. In our work, a broader range of

321　carbohydrate active enzymes (CAZymes) were found. Of the 208 annotated CAZyme

322　families, 99 were Glycoside Hydrolase (GH) families (Table S11). Enzymes that

323　could be classified as cellulases, endohemicelluloses, debranching enzymes,

324　oligosaccharide-degrading enzymes were all presented. These findings indicate that

15

325   the gut microbiome give assistance to digest a broad range of food sources, making *P.*

326   *canaliculata* grow fast to adapt to an invasive life style.

327   **Conclusion and discussion**

328   Given its environmental invasiveness, broad stress adaptability and rapid reproduction,

329   the golden apple snail *P. canaliculata* has received a vast of attention worldwide.

330   However, the underlying genetic mechanism has not been comprehensively

331   uncovered. The chromosome level genome of *P. canaliculata* presented in this study

332   sheds first lights into the genomic basis of the ecological plasticity to various stressors.

333   Major findings of this study include the recent explosion of DNA/hAT-Charlie TEs,

334   the expansion of P450 gene family and the constitution of Cellular homeostasis

335   system, contributing to the plasticity in the stress adaptation. Although the defined

336   function of the recently originated TEs could not be confirmed, the explosion of TEs

337   is deemed as powerful facilitators in adaptive evolution, indicating its important role

338   in *P. canaliculata*'s stress resistance. UPR system, Xenobiotic biotransformation

339   system and ROS system are major components of the Cellular homeostasis system,

340   and especially P450s expands with specific functions. In addition, exclusive

341   perivitellin genes are characterized from the *P. canaliculata* genome, contributing to

342   the high reproductive rate and the expansion of habitats. Furthermore, the gut

343   metagenome encodes rich genes for food digestion and xenobiotics degradation.

344   These findings collectively provide novel insight into the molecular mechanisms of

345   the ecological plasticity and high invasiveness.

16

346 The rich phenotypic and genetic diversity of molluscs make them an excellent species

347 group to address many valuable issues about evolution, ecology and function.

348 However, the genomic resource of Mollusca is still insufficient compared with other

349 close phylums, such as Arthropoda and Nematoda, and few molluscs could be

350 employed as model organism. *P. canaliculata* possesses potential to be a model

351 organism of molluscs because of several inherent characters. For example, *P.*

352 *canaliculata* is easy to acquire, for it has a broad global distribution originated from a

353 primarily circumtropical environment. Due to the high adaptability, rapid growth and

354 efficient reproduction, *P. canaliculata* also facilitate the cultivation in laboratory. We

355 report a fine reference genome of *P. canaliculata* in the present study, which is the

356 first chromosome level genome published in Mollusca. As the cellular complexity and

357 the conservation of pathways, *P. canaliculata* could be a representative of Mollusca,

358 so the genome described in this study can be used to advance our understanding of the

359 molecular mechanisms for various scientific issues in Mollusca.

360

361 **Methods**

362 **Samples collection and sequencing**

363 Adults of *P. canaliculata* were collected from a local paddy field in Shenzhen,

364 Guangdong province, China, and maintained in aerated freshwater at 15 ± 2 °C for a

365 week before processing. Genomic DNA was extracted from the foot muscles of a

366 single *P. canaliculata* for constructing PCR free Illumina 350-bp insert libraries and

367 PacBio 20-kb insert library, and sequenced on Illumina HiSeq 2500 and PacBio

368 SMRT platforms, respectively. The Hi-C library was prepared using the muscle tissue

369 of another single *P. canaliculate* by following methods: Nuclear DNA was

370 cross-linked in situ, extracted, and then digested with a restriction enzyme. The sticky

371 ends of the digested fragments were biotinylated, diluted, and then ligated to each

372 other randomly. Biotinylated DNA fragments were enriched and sheared again for

373 preparing the sequencing library, which was then sequenced on a HiSeq X Ten

374 platform (Illumina).

375 Seven tissues including embryos (2 days post fertilization), gill, hemocytes,

376 hepatopancreas, kidney, ovary and testis from six animals were collected as parallel

377 samples. Next, animals were cultivated in 37 °C and 10 °C for 24 hours heat and cold

378 tolerance, in Cr^{3+}(2mg L^{-1}), Cu^{2+}(0.2mg L^{-1}) and Pb^{2+}(1mg L^{-1}) for 24 hours heavy

379 metal tolerance, and in waterless tank for 7 days air exposure. Then the hemocytes

380 were harvested and stored, with three replicates for each group. In final, total

381 messenger RNAs (mRNA) were extracted from the stored tissues of *P. canaliculata*

382 materials for constructing cDNA libraries (insert 350-bp), and sequenced on an

383 Illumina HiSeq 2500 sequencer.

384 The intestinal digesta from 70 adult snails of *P. canaliculata* were collected, pooled

385 into 6 samples and stored at −20 °C until microbial DNA was extracted. A

386 combination of cell lysis treatments was applied, including five freeze-thaw cycles

387 (alternating between 65 °C and liquid nitrogen for 5 min), repeated beads-beating in

388 ASL buffer (cat. no. 19082; Qiagen Inc.), and incubated at 95 °C for 15 min. DNA

389 was isolated following the protocol reported protocol [75]. Paired-end libraries of

390 metagenomic DNA were prepared with an insert size of 350 base pairs (bp) following

391 the manufacture's protocol (cat. no. E7645L; New England Biolabs). Sequencing was

392 performed on Illumina HiSeq X10.

393

Genome assembly and annotation

395 The Illumina raw reads were filtered by trimming the adapter sequence and

396 low-quality part, resulting in a clean and high-quality reads data with average error

397 rate < 0.001. For the PacBio raw data, the short subreads (< 2 kb) and low-quality

398 (error rate > 0.2) subreads were filtered out, and only one representative subread was

399 retained for each PacBio read. The clean PacBio reads were assembled by the

400 software samrtdenovo (https://github.com/ruanjue/smartdenovo), then Illumina

401 reads were aligned to the contigs by BWA-MEM, and single base errors in the contigs

402 were corrected by Pilon (v1.16) with parameters "-fix bases, -nonpf, -minqual 20".

403 The P. *canaliculata* genome is highly heterozygous illustrated by the double peaks on

404 the distribution curve of K-mer frequency, and current assembly algorithm tends to

405 collapse homozygous regions and report heterozygous regions in alternative contigs.

406 To get a haploid reference contigs, we employed a whole-genome alignment (WGA)

407 strategy by MUMmer v3.23 to recognize and selectively remove alternative

408 heterozygous contigs, which were characterized by shorter length (less than 200 kb)

409 and most regions (larger than 50%) can be aligned to another larger contig with

410 confident identity (higher than 80%). Next, Hi-C sequencing data were aligned to the

411 haploid reference contigs by BWA-MEM, and then these contigs were clustered into

412 chromosomes with LACH-ESIS (http://shendurelab.github.io/LACHESIS/).

413 The gene models in *P. canaliculata* genome were predicted by EVidence Modeler

414 v1.1.1 [76], integrating evidences from ab initio predictions, homology-based

415 searches and RNA-seq alignments. Then, the protein-coding sequences were mapped

416 by RNA-seq data and functionally annotated using UniProt and InterProScan

417 (5.16-55.0) databases [77]. Finally, the gene models were retained if they had at least

418 one supporting evidence from UniProt database, InterProScan domain and RNA-seq

419 data. Gene functional annotation was performed by aligning the protein sequences to

420 NCBI NR, UniProt, COG and KEGG databases with BLASTP v2.3.0+ under E-value

421 cutoff of 10^{-5} and choosing the best hit. The pathway analysis and functional

422 classification were conducted based on KEGG database [78]. InterProScan was used

423 to assign preliminary GO terms, Pfam domains and IPR domains to the gene models.

424 A de novo repeat library for *P. canaliculata* was constructed by RepeatModeler

425 (v1.0.4; http://www.repeatmasker.org/RepeatModeler.html). TEs in the *P. canaliculata*

426 genome were also identified by RepeatMasker (v4.0.6; http://www.repeatmasker.org/)

427 using both Repbase library and the de novo library. Tandem repeats in the *P.*

428 *canaliculata* genome were predicted using Tandem Repeats Finder v4.07b [79]. The

429 divergence rates of TEs were calculated between the identified TE elements in the

430 genome and their consensus sequence at the TE family level.

431

432 **Evolutionary analysis**

433 Orthologous and paralogous groups were assigned from seven species (*P.*

434 *canaliculata*, *Lottia gigantea*, *Aplysia california*, *Biomphalaria glabrata*, *Crassostrea*

435 *gigas*, *Octopus bimaculoides* and *Lingula anatina*) by OrthoFinder [45] with default

436 parameters. Orthologous groups that contain only one gene for each species were

437 selected to construct the phylogenetic tree. The protein sequences of each gene family

438 was independently aligned by muscle v3.8.31 [80] and then concatenated into one

439 super-sequence. The phylogenetic tree was constructed by maximum likelihood (ML)

440 using PhyML v3.0 [43] with best-fit model (LG+I+G) that was estimated by ProtTest3

441 [81]. The Bayesian Relaxed Molecular Clock (BRMC) approach was adopted to

442 estimate the neutral evolutionary rate and species divergence time using the program

443 MCMCTree, implemented in PAML v4.9 package [44]. The calibration time (fossil

444 record time) interval (173-398 Mya) of *Octopus bimaculoides* was adopted from

445 previous results.

446

447 **Transcriptome data analysis**

448 Transcriptome reads were mapped to the reference genome of *P. canaliculata* using

449 TopHat (v. 2.1.0) with default settings. The expression level of each reference gene in

450 terms of FPKM was computed by cufflinks v2.2.1. A gene was considered to be

451 expressed if its FPKM >0. Differential gene expression analysis was conducted using

452 cuffdiff v2.2.1.

21

453

Metagenome data analysis

Raw reads were cleaned to exclude adapter sequences, low quality sequence, as well

as contaminated DNA. The adapter sequence in reads were identified and trimmed by

an ungapped dynamic programming algorithm; the low-quality part (head or tail) of

reads were trimmed off to ensure that the average error rate of the left reads is lower

than 0.001; the reads that mapped to the contaminated DNA by BWA-MEM [82] were

filtered out; finally, shorter reads (length < 75-bp) and unpaired reads were excluded

to form a clean reads data. The BWA database built for cleaning contamination

included genomes of 10 species: *P. canaliculata* genome, *Brassica rapa* genome,

Oryza sativa genome, 2 *Angiostrongylus cantonensis* genomes, *Caenorhabditis*

elegans genome, *schistosoma mansoni* genome, *clonorchis sinensis* genome, *fasciola*

hepatica genome, *Danio rerio* genome, and *human hg38* genome.

The clean reads were assembled by metaSPAdes (v3.11.1) [83] under pair-end mode

for each sample, then gene prediction was performed on contigs longer than 500 bp

by Prodigal (v2.6.3) [84] with parameter "-p meta", and gene models with cds length

less than 102 bp were filtered out. A non-redundant (NR) gene set (539,344 genes)

was constructed using the gene models predicted from each samples by cd-hit-est

(v4.6.6) [85] with parameter "-c 0.95 -n 10 -G 0 –a S 0.9", which adopts a greedy

incremental clustering algorithm and the criteria of identity > 95% and overlap > 90%

of the shorter genes. Then, the clean reads were mapped onto this NR gene set by

474 BWA-MEM with the criteria of alignment length ≥ 50bp and identity > 95%. The

475 unmapped reads from all samples were assembled together, and genes were predicted

476 again. The newly predicted genes were combined with the previous gene set by

477 cd-hit-est to get a new NR gene set (1,147,339 genes). After the taxonomic

478 assignments to the new NR gene set, 5244 genes classified as Eukaryota but not fungi

479 were removed, and the final NR gene set (1,142,095 genes) was obtained.

480 Taxonomic assignments for the final NR genes were made on the basis of DIAMOND

481 [86] protein alignment against the NCBI-NR database by CARMA3 [87]. Functional

482 annotation was performed by aligning all the protein sequences to the KEGG [88]

483 database (release 79) using DIAMOND and taking the best hit with the criteria of

484 E-value < 1e-5. CAZymes were annotated with dbCAN (release 5.0) [89] using

485 HMMER (v3.0) hmmscan [90] by taking the best hit with E-value < 1e-18 and

486 coverage > 0.35.

487 The clean reads from each sample were aligned against the gene catalog (1,142,095

488 genes) by BWA-MEM with the criteria of alignment length ≥ 50bp and identity >

489 95%. Sequence-based gene abundance profiling was performed as previously

490 described [91]. Taxonomic profiles of the samples were calculated by adding the gene

491 abundance together according to the taxonomic assignment result.

492

493

Abbreviations

P. Canaliculata, Pomacea canaliculata; L. gigantean, Lottia gigantean;

A. California, Aplysia California; B. glabrata, Biomphalaria glabrata; C. gigas,

Crassostrea gigas; O. bimaculoides, Octopus bimaculoides; L. anatine, Lingula

anatine; P. fucata, Pinctada fucata; Hem, hemocyte; Te, testis; Ov, ovary; Kn, kidney;

GI, gill; Hp, hepatopancreas, Em, embryo; SSR, simple sequence repeats; mya,

million years ago*; BLAST, basic local alignment search tool;* SNP, single nucleotide

polymorphism; PVF, Pervitelline Fluid; Ovo, ovorubin; AFLP, amplified fragment

length polymorphism; DEGs, differentially expressed genes; LPyS,

Lipopolysaccharide; iTRAQ, Isobaric Tags For Relative, Absolute Quantitation;

LC-MS/MS, Liquid Chromatography-tandem Mass Spectrometry; TEs, transposable

elements; LTR, long terminal repeats; LINE, long interspersed elements; SINE, short

interspersed elements; UPR, Unfolded protein response; HSPs, heat shock proteins;

HSF1, heat shock transcription factor 1; PERK, protein kinase RNA-like ER kinase;

ATF6,activating transcription factor 6; ER, endoplasmic reticulum; CYP450s,

cytochrome P450s; FMOs, flavin-containing monooxygenases; GSTs, glutathione

S-transferases; ABC, ATP binding cassette; ROS, reactive oxygen species; ROI,

reactive oxygen intermediates; SOD, superoxide dismutase; CAT, catalase; Prx,

peroxidase; GPX, glutathione peroxidase; TNFR, tumor necrosis factor receptor;

NR, non-redundant genes; ORF, open reading frame; Kos, orthologous groups;

CAZymes, carbohydrate active enzymes; GH, Glycoside Hydrolase.

515

Availability of data and materials

517   Tables S1 to S11 and Figures S1 to S4 are available in the supplementary information

518   file. The raw sequencing data has been deposited in DDBJ/EMBL/GenBank under

519   project accession PRJNA427478, SRR6425828 for genomic Illumina_PE125

520   sequencing data, SRR6425829 for genomic Illumina_PE150 sequencing data,

521   SRR6425827 for genomic Pacbio sequencing data, SRR6429132~SRR6429164 for

522   transcriptome sequencing data, and SRR6472920~SRR6472925 for gut microbiome

523   data. All the analysis data have also been released for public use and can be freely

524   accessed   at   AGIS

525   ftpsite: ftp://ftp.agis.org.cn/~fanwei/Pomacea_canaliculata_Genome/ .

Authors' contributions

527   WF and WQ conceived the study and designed the experiments. CL and YZ

528   performed the genome sequencing and assembly, BL performed annotation and

529   evolutionary analysis. CL performed the stress tolerance analysis, YR performed the

530   reproduction analysis, YZ performed the metagenome analysis. HW, SL, FJ, LY

531   provide suggestions and help checking. WF, CL, BL, YR, YZ wrote the manuscript,

532   and GZ help revise the manuscript. All authors read and approved the final

533   manuscript.

534

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

553   **Legends of Tables and Figures**

554   **Tables**

555   **Table 1. Summary of assembly and annotation of mollusk genomes**

Genome feature	*P. canaliculata*	*L. gigantea*	*A. california*	*B. glabrata*	*C. gigas*	*O. bimaculoides*
Assembled sequences (bp)	440,071,717	359,505,668	927,310,431	916,377,450	557,735,934	2,3381,887,882
Contig N50 size (bp)	1,072,857	94,165	9,817	18,978	37,218	5,982
Contig N90 size (bp)	303,904	10,180	1,626	5,132	11,109	1,606
Scaffold N50 size (bp)	31,531,291	1,870,055	917,541	48,059	401,685	475,182
Scaffold N90 size (bp)	23,662,357	74,480	207,390	817	68,181	79,088
GC content (%)	40.3	33.3	40.3	36.0	33.4	36
No. of gene models	21,533	23,824	19,909	14,224	28,402	15,814
Avg. CDS length (bp)	1,497	1,136	1,568	1,066	1,472	1,535
BUSCO (%)	98.9	98.4	98.7	72.8	99.4	98.7
Transposable elements (bp)	49,579,006	37,369,817	202,174,499	189,550,886	103,381,274	737,398,096
Tandem repeat (bp)	873,801	257,674	8,263,822	2,145,821	590,907	62,633,792

556

557   **Figures**

558   **Figure 1. The genome characteristics of *P. canaliculata*. (a)** Circos plot showing the

559   genomic features**. Track 1: 14 linkage groups of the genome; Track 2: distribution of

560   transposon elements in chromosomes; Track 3: protein-coding genes located on

561   chromosomes; Track 4: distribution of GC contents. **(b)** A genome-wide contacting

562   matrix from Hi-C data between each pair of the 14 chromosomes, using 100 kb

563   window size. The color value means the logarithm of valid reads to base 2 ($\log_2$(valid

564   reads)). **(c)** Distribution of CDS length in six closely related species.

565

566   **Figure 2. Evolutionary genomic analysis between *P. canaliculata* and other**

567   **molluscs. (a)** Phylogenetic placement of *P. canaliculata* within the molluscs dated

568   tree. The estimated divergence time were shown on each branching point, the species

569   marked with red color was *P. canaliculata*. **(b)** Distribution of divergence rate for the

570   class of DNA transposons in molluscs genomes. The divergence rate was calculated

571   by comparing all TE sequences identified in the genome to its corresponding

572   consensus sequence in each TE subfamily. The red arrow indicates the P. canaliculata

573   and C. gigas had a recent explosion of TEs at ~4% divergence rate.

574

575   **Figure 3. The cellular homeostasis system in *P. canaliculata*.** Unfolded protein

576   response (UPR) system included HSPs and HSF in the heat shock response and CNX,

577   NEF, GRP94, BIP, HSP40, ATF6, IRE1, PERK, COP2, XBP, ATF4, TRAM and

578   Derlin in the endoplasmic reticulum unfolded-protein response (UPR-ERAD).

579   Apoptotic pathways included XIAPs, Bcl2, caspases, TNFR, and FADD. The

580   antioxidant systems included PRX, SOD, CAT and GPX. The xenobiotic

581   biotransformation system included EPHX3, P450, FMO and ABC transpoter. Gene

582   boxes for gene families with the filled colors represent the degree of upregulation

583   (FPKM-stimulus/FPKM-control) by an ovreall result of stress including heat, cold,

584   heavy metal and air exposure. Pathways and genes were obtained based on KEGG

585   annotation.

586

587   **Figure 4. The expansion of P450 gene family in *P. canaliculata*.** (a) Phylogenetic

588   tree demonstrating orthologous and paralogous relationships of all P450 genes from 7

589   species including *P. canaliculata*, *A. california*, *B. glabrata*, *C. gigas*, *L. gigantea*,

590   *O.bimaculoides* and *P. fucata*. P450 genes from seven species were obtained based

591   Pfam annotation (Interpro) with the E-value 10^{-5}. Clades are labeled by P450

592   subfamily names. The tree was constructed using the Maximum likelihood method in

593   MEGA7, and branch length scale indicates average residue substitutions per site. (b)

594   Phylogenetic tree of P450 genes in *P. canaliculata*, which is a subset of the

28

595 phylogenetic tree for the 7 species, and their heat map of expression (FPKM) in seven

596 tissues (Hem, hemocyte; Te, testis; Ov, Ovary; Kn, kidney; Gl, gill; Hp,

597 hepatopancreas; Em, Embryo), and heat map of induced expression

598 (FPKM-stimulus/FPKM-control) under stress (Con: control; heat; cold; Hm: heavy

599 metal; Exp: air exposure).

600

601 **Figure 5. The *P. canaliculata* perivitellins composition and expression in different**

602 **tissues.** (a) Pervitelline Fluid (PVF) is under the eggshell and surrounds the embryo, it

603 contains carbohydrates, lipids, proteins, and the proteins is also known as perivitellins

604 and classified into three categories of PcOvo, PcPV2, PcPV3. (b) The shown

605 expression value is the logarithm of FPKM to base 2 ($\log_2$FPKM). The first 3 letters

606 in each gene ID refer to three classes of perivitellins, uPV means unclassified

607 perivitellins, PV2 means PcPV2, Ovo means PcOvo. Abbreviations were used for 7

608 tissues (Hem, hemocyte; Te, testis; Ov, Ovary; Kn, kidney; Gl, gill; Hp,

609 hepatopancreas; Em, Embryo).

610

611 **References**

612 1.   Lowe S, Browne M, Boudjelas S, de Poorter M. 100 of the World's Worst Invasive Alien

613   Species: A selection fromthe Global Invasive Species Database. Auckland, New Zeland:

614   World Conservation Union (IUCN); 2000.

615 2.   Ranamukhaarachchi SL, Wickramasinghe S. Golden apple snails in the world:

616 introduction, impact, and control measures. Global advances in ecology and

617 management of golden apple snails. 2006:133-52.

618 3. Naylor R. Invasions in Agriculture: Assessing the Cost of the Golden Apple Snail in Asia.

619 Royal Swedish Academy of Sciences. 1996;25:443-8.

620 4. Berthold T. Vergleichende Anatomie, Phylogenie und historische Biogeographie der

621 Ampullariidae: (Mollusca, Gastropoda). 1991.

622 5. Howells RG, Burlakova LE, Karatayev AY, Marfurt RK, Burks RL. Native and

623 introduced Ampullaridae in North America: History, status, and ecology. 2006:73-112.

624 6. Halwart M, Bartley DM. International mechanisms for the control and responsible use of

625 alien species in aquatic ecosystems, with special reference to the golden apple snail. Los

626 Baños, Philippines: Philippine Rice Research Institute (PhilRice); 2006.

627 7. López MA, Altaba CR, Andree KB, López V. First invasion of the Apple snail Pomacea

628 insularum in Europe. Tentacle. 2010;18:26-8.

629 8. Estebenet AL, Martín PR. Pomacea canaliculata (Gastropoda: Ampullariidae): life-history

630 traits and their plasticity. Biocell 2002;26:83-9.

631 9. Lach L. The spread of the introduced freshwater apple snail *Pomacea canaliculata*

632 (Lamarck) (Gastropoda Ampullariidae) on Oahu, Hawaii. Bishop Museum Occasional

633 Papers. 1999;58:66-71.

634 10. Yusa Y，Sugiura N, Wada T. Predatory Potential of Freshwater Animals on an Invasive

635 Agricultural Pest, the Apple Snail *Pomacea canaliculata* (Gastropoda: Ampullariidae), in

636 Southern Japan. Biol Invasions. 2006;8:137-47.

637 11. Lach L, Britton DK, Rundell RJ, Cowie RH. Food Preference and Reproductive Plasticity

638   in an Invasive Freshwater Snail. Biol Invasions. 2000;2:279-88.

639   12.   Mochida O. Spread of freshwater *Pomacea* snails (Pilidae, Mollusca) from Argentina to

640   Asia. Micronesica. 1991;3 51-62.

641   13.   Shan L, Zhang Y, Steinmann P, Zhou X. Emerging Angiostrongyliasis in Mainland China.

642   Emerg Infect Dis. 2008;14:161-4.

643   14.   Caldeira RL, Mendonca CL, Goveia CO, Lenzi HL, Graeff-TeixeiraC Lima WS, et al.

644   First record of molluscs naturally infected with *Angiostrongylus cantonensis* (Chen, 1935)

645   (Nematoda: Metastrongylidae) in Brazil. Memórias do Instituto Oswaldo Cruz.

646   2007;102:887-9.

647   15.   McMichael AJ, Beaglehole R. The changing global context of public health. Lancet

648   (London, England). 2000;356:495-9.

649   16.   Chapman A. Numbers of Living Species in Australia and the World. Australian Biological

650   Resources Study; 2009.

651   17.   Lindberg DR, Ponder WF, Haszprunar G. The Mollusca: relationships and patterns from

652   their first half-billion years. Oxford University Press, Oxford; 2004.

653   18.   Hayes KA, Cowie RH, Thiengo SC. A global phylogeny of apple snails: Gondwanan

654   origin, generic relationships, and the influence of outgroup choice (Caenogastropoda:

655   Ampullariidae). Biol J Linn Soc Lond. 2009;98:61-76.

656   19.   Matsukura K, Tsumuki H,Izumi Y, Wada T. Physiological response to low temperature in

657   the freshwater apple snail, *Pomacea canaliculata* (Gastropoda: Ampullariidae). J Exp

658   Biol. 2009;212:2558-63.

659   20.   Yusa Y, Wada T, Takahashi S. Effects of dormant duration, body size, self-burial and

660   water condition on the long-term survival of the apple snail, *Pomacea canaliculata*

661   (Gastropoda: Ampullariidae). Appl Entomol Zool. 2006;41:627-32.

662 21. Kruatrachue M, Sumritdee C, Pokethitiyook P, Singhakaew S. Histopathological effects

663   of contaminated sediments on golden apple snail (*Pomacea canaliculata*, Lamarck 1822).

664   Bull Environ Contam Toxicol. 2011;86:610-4.

665 22. Dreon MS, Frassa MV, Ceolín M, Ituarte S, Qiu JW, Sun J, et al. Novel animal defenses

666   against predation: a snail egg neurotoxin combining lectin and pore-forming chains that

667   resembles plant defense and bacteria attack toxins. PLoS One. 2013;8:e63782.

668   doi:10.1371/journal.pone.0063782.

669 23. Ottaviani E, Caselgrandi E, Fontanili P, Franceschi C. Evolution, immune responses and

670   stress: studies on molluscan cells. Acta Biol Hung. 1992;43:293-8.

671 24. Ottaviani E, Accorsi A, Rigillo G, Malagoli D, Blom JM, Tascedda F. Epigenetic

672   modification in neurons of the mollusc *Pomacea canaliculata* after immune challenge.

673   Brain Res. 2013;1537:18-26.

674 25. Mercado Laczkó AC, Lopretto EC. Estudio cromosómico y cariotípico de p*omacea*

675   *canaliculata* (Lamarck, 1801) (Gastropoda, Ampullariidae). Revista del Museo Argentino

676   de Ciencias Naturales "Bernardino Rivadavia" Hidrobiología. 1998;8:15-20.

677 26. Xu J, Han X, Li N, Yu J, Qian C, Bao Z. Analysis of genetic diversity of three geographic

678   populations of *Pomacea canaliculata* by AFLP. Acta Ecol Sin. 2009;29:4119- 26.

679 27. Chen L, Xu H, Li H, Wu J, Ding H, Liu Y. Isolation and characterization of sixteen

680   polymorphic microsatellite loci in the golden apple snail *Pomacea canaliculata*. Int J Mol

681   Sci. 2011;12:5993-8.

682   28.   Mu X, Hou G, Song H, Xu P, Luo D, Gu D, et al. Transcriptome analysis between

683   invasive *Pomacea canaliculata* and indigenous *Cipangopaludina cahayensis* reveals

684   genomic divergence and diagnostic microsatellite/SSR markers. BMC Genet. 2015;16:12.

685   29.   Sun J, Wang M, Wang H, Zhang H, Zhang X, Thiyagarajan V, et al. De novo assembly of

686   the transcriptome of an invasive snail and its multiple ecological applications. Mol Ecol

687   Resour. 2012;12:1133-44.

688   30.   Mu H, Sun J , Fang L, Luan T, Williams GA, Cheung SG, et al. Genetic Basis of

689   Differential Heat Resistance between Two Species of Congeneric Freshwater Snails:

690   Insights from Quantitative Proteomics and Base Substitution Rate Analysis. J Proteome

691   Res. 2015;14:4296-308.

692   31.   Yang L, Cheng TY, Zhao FY. Comparative profiling of hepatopancreas transcriptomes in

693   satiated and starving *Pomacea canaliculata*. BMC Genet. 2017;18:18.

694   32.   Xiong YM, Yan ZH, Zhang JE, Li HY. Analysis of albumen gland proteins suggests

695   survival strategies of developing embryos of *Pomacea canaliculata*. Molluscan Res.

696   2017:1-6.

697   33.   Sun J, Mu H , Zhang H, Chandramouli KH, Qian PY, Wong CK, et al. Understanding the

698   regulation of estivation in a freshwater snail through iTRAQ-based comparative

699   proteomics. J Proteome res. 2013;12:5271-80.

700   34.   Sun J, Zhang H, Wang H, Heras H, Dreon MS, Ituarte S, et al. First proteome of the egg

701   perivitelline fluid of a freshwater gastropod with aerial oviposition. J Proteome Res.

702   2012;11:4240-8.

703   35.   Aplysia Genome Project. Broad Institute. Vertebrate Biology Group. 2009.

704   https://www.broadinstitute.org/aplysia/aplysia-genome-project

705   36.   Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress

706   adaptation and complexity of shell formation. Nature. 2012;490:49-54.

707   37.   Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, et al. Draft genome

708   of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA

709   Res. 2012;19:117-30.

710   38.   Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, et al.

711   Insights into bilaterian evolution from three spiralian genomes. Nature. 2013;493:526-31.

712   39.   Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The

713   octopus genome and the evolution of cephalopod neural and morphological novelties.

714   Nature. 2015;524:220-4.

715   40.   Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea

716   chemosynthetic environments as revealed by mussel genomes. Nat Ecol Evol. 2017;1:121.

717   doi:10.1038/s41559-017-0121.

718   41.   Adema CM, Hillier LW, Jones CS, Loker ES, Knight M, Minx P, et al. Corrigendum:

719   Whole genome analysis of a schistosomiasis-transmitting freshwater snail. Nat Commun.

720   2017;8:16153.

721   42.   Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics

722   by analyzing k-mer frequency in de novo genome projects. Quantitative Biology

723   2013:arXiv:1308.2012 [q-bio.GN].

724   43.   Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms

725   and methods to estimate maximum-likelihood phylogenies: assessing the performance of

726   PhyML 3.0. Syst Biol 2010;59:307-21. doi:10.1093/sysbio/syq010.

727   44.   Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol.

728   2007;24:1586-91. doi:10.1093/molbev/msm088.

729   45.   Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome

730   comparisons dramatically improves orthogroup inference accuracy. Genome Biol.

731   2015;16:157. doi:10.1186/s13059-015-0721-2.

732   46.   Feschotte C, Wessler SR. Mariner-like transposases are widespread and diverse in

733   flowering plants. Proc Natl Acad Sci U S A 2002;99:280-5.

734   47.   Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. The struggle for life of the

735   genome's selfish architects. Biol Direct. 2011;6:19.

736   48.   Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc

737   Natl Acad Sci U S A. 2011;108 Suppl 2:10863-70.

738   49.   Chrousos GP. Stress and disorders of the stress system. Nat Rev Endocrinol.

739   2009;5:374-81.

740   50.   Vabulas RM, Raychaudhuri S, Hayer-Hartl M. Protein folding in the cytoplasm and the

741   heat shock response. Cold Spring Harbor perspectives in biology. 2010;2:a004390.

742   51.   Chen B，Retzlaff M，Roos T，Frydman J. Cellular Strategies of Protein Quality Control.

743   Cold Spring Harbor Perspectives in Biology. 2011;3:a004374.

744   52.   Korennykh A and Walter P. Structural basis of the unfolded protein response. Annu Rev

745   Cell Dev Biol. 2012;28:251-77.

746   53.   Chambers JE and Yarbrough JD. Xenobiotic biotransformation systems in fishes. Comp

747   Biochem Physiol C. 1976;55:77-84.

748 54. Mello DF, de Oliveira ES, Vieira RC, Simoes E, Trevisan R, Dafre AL, et al. Cellular and

749   Transcriptional Responses of *Crassostrea gigas* Hemocytes Exposed in Vitro to

750   Brevetoxin (PbTx-2) Mar Drugs. 2012;10: 583-97.

751 55. Boutet I, Tanguy A, Moraga D. Characterisation and expression of four mRNA sequences

752   encoding glutathione S-transferases pi, mu, omega and sigma classes in the Pacific oyster

753   *Crassostrea gigas* exposed to hydrocarbons and pesticides. Mar Biol 2004;146:53-64.

754 56. Deeley RG, Westlake C, Cole SP. Transmembrane transport of endo- and xenobiotics by

755   mammalian ATP-binding cassette multidrug resistance proteins. Physiol Rev.

756   2006;86:849-99.

757 57. Liu C, Zhang T, Wang L, Wang M, Wang W, Jia Z, et al. The modulation of extracellular

758   superoxide dismutase in the specifically enhanced cellular immune response against

759   secondary challenge of Vibrio splendidus in Pacific oyster (*Crassostrea gigas*). Dev

760   Comp Immunol. 2016;63:163-70.

761 58. Lamb DC, Lei L, Warrilow AG, Lepesheva GI, Mullins JG, Waterman MR, et al. The first

762   virally encoded cytochrome p450. J Virol. 2009;83:8266-9.

763 59. Urlacher VB, Girhard M. Cytochrome P450 monooxygenases: an update on perspectives

764   for synthetic application. Trends Biotechnol. 2012;30:26-36.

765 60. Sanderson T, van den Berg M. Topic 3.1: Interactions of xenobiotics with the steroid

766   hormone biosynthesis pathway. Pure Appl Chem. 2003;75:1957-71.

767 61. Goldstone JV, McArthur AG, Kubota A, Zanette J, Parente T, Jönsson ME, et al.

768   Identification and developmental expression of the full complement of Cytochrome P450

769   genes in Zebrafish. BMC Genomics. 2010;11:643.

770 62. Chuang SS, Helvig C, Taimi M, Ramshaw HA, Collop AH, Amad M, et al. CYP2U1, a

771   novel human thymus- and brain-specific cytochrome P450, catalyzes omega- and

772   (omega-1)-hydroxylation of fatty acids. J Biol Chem. 2004;279:6305-14.

773 63. Fleming I. The pharmacology of the cytochrome P450 epoxygenase/soluble epoxide

774   hydrolase axis in the vasculature and cardiovascular disease. Pharmacol Rev.

775   2014;66:1106-40.

776 64. Zhang G, Kodani S, Hammock BD. Stabilized epoxygenated fatty acids regulate

777   inflammation, pain, angiogenesis and cancer. Prog Lipid Res. 2014;53:108-23.

778 65. de Jong-Brink M, Boer HH, Joosse J. Mollusca. In: Adiyodi, K.G., Adiyodi,

779   R.G. (Eds.), Reproductive Biology of invertebrates. Oogenesis oviposition and

780   oosorption, vol. 1. John Wiley & Sons Ltd., New York, 1983; pp. 297-355.

781 66. Garin CF, Heras H, Pollero RJ. Lipoproteins of the egg perivitelline fluid of *Pomacea*

782   *canaliculata* snails (Mollusca: Gastropoda). J Exp Zool. 1996;276:307-14.

783 67. Dreon MS, Schinella G, Heras H, Pollero RJ. Antioxidant defense system in the apple

784   snail eggs, the role of ovorubin. Arch Biochem Biophys. 2004;422:1-8.

785 68. Dreon MS, Ituarte S, Heras H. The role of the proteinase inhibitor ovorubin in apple snail

786   eggs resembles plant embryo defense against predation. PLoS One. 2010;5:e15059.

787   doi:10.1371/journal.pone.0015059.

788 69. Cardoso AM, Cavalcante JJV, Vieira RP, Lima JL, Grieco MAB, Clementino MM, et al.

789   Gut Bacterial Communities in the Giant Land Snail *Achatina fulica* and Their

790   Modification by Sugarcane-Based Diet. Plos One. 2012;7   doi:ARTN

791   e3344010.1371/journal.pone.0033440.

792 70. Cardoso AM, Cavalcante JJV, Cantão ME, Thompson CE, Flatschart RB, Glogauer A, et

793   al. Metagenomic Analysis of the Microbiota from the Crop of an Invasive Snail Reveals a

794   Rich   Reservoir   of   Novel   Genes.   Plos   One.   2012;7   doi:ARTN

795   e4850510.1371/journal.pone.0048505.

796 71. Cabrera G, Pérez R, Gómez JM, Ábalos A, Cantero D. Toxic effects of dissolved heavy

797   metals on Desulfovibrio vulgaris and Desulfovibrio sp strains. J Hazard Mater

798   2006;135:40-6. doi:10.1016/j.jhazmat.2005.11.058.

799 72. Finlay JA, Allan VJ, Conner A, Callow ME, Basnakova G, Macaskie LE. Phosphate

800   release and heavy metal accumulation by biofilm-immobilized and chemically-coupled

801   cells of a Citrobacter sp. pre-grown in continuous culture. Biotechnol Bioeng.

802   1999;63:87-97.

803 73. Valls M, de Lorenzo V, Gonzalez-Duarte R, Atrian S. Engineering outer-membrane

804   proteins in *Pseudomonas putida* for enhanced heavy-metal bioadsorption. J Inorg

805   Biochem. 2000;79:219-23.

806 74. Pinheiro GL, Correa RF, Cunha RS, Cardoso AM, Chaia C, Clementino MM, et al.

807   Isolation of aerobic cultivable cellulolytic bacteria from different regions of the

808   gastrointestinal tract of giant land snail *Achatina fulica*. Front Microbiol. 2015;6
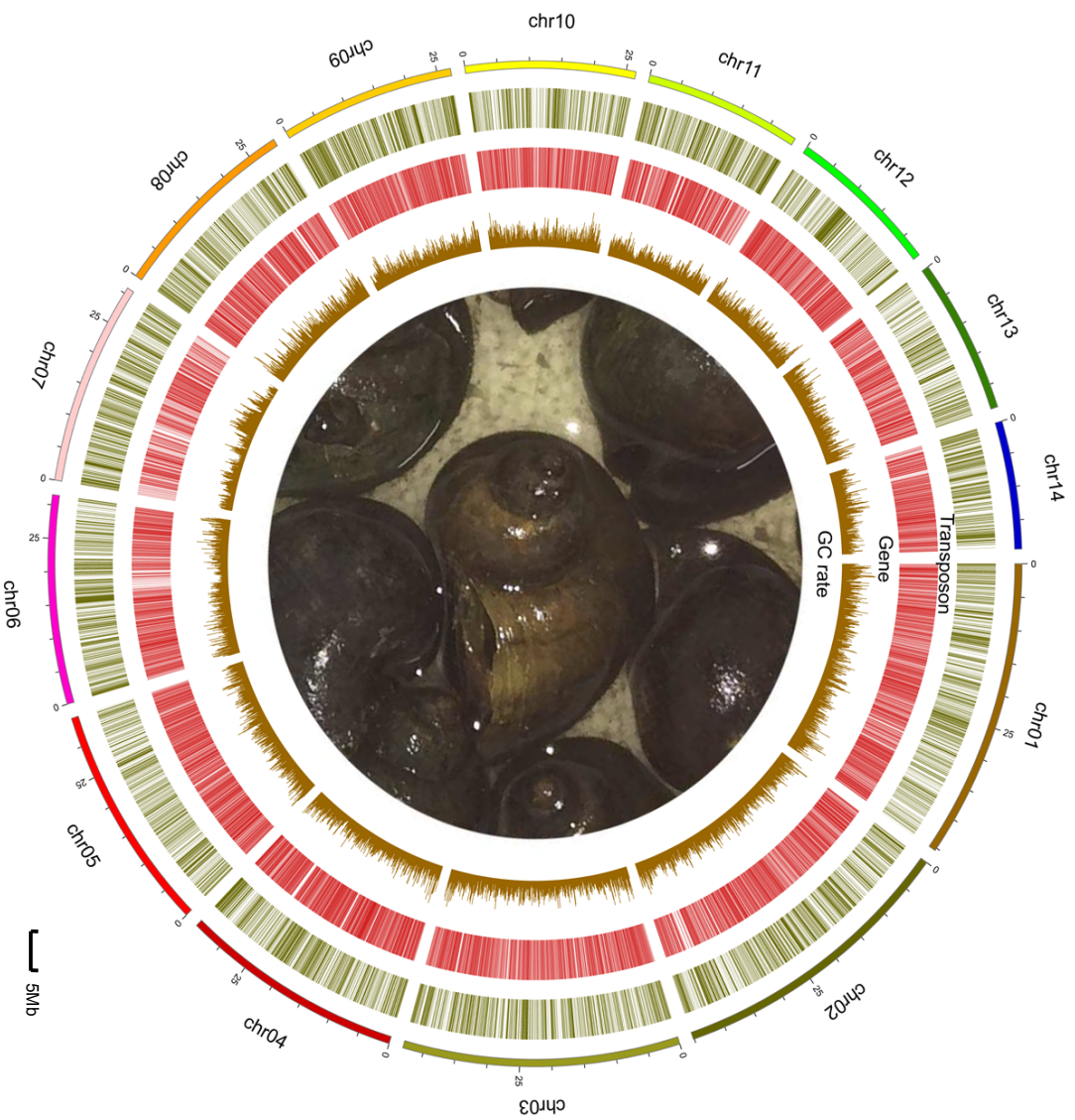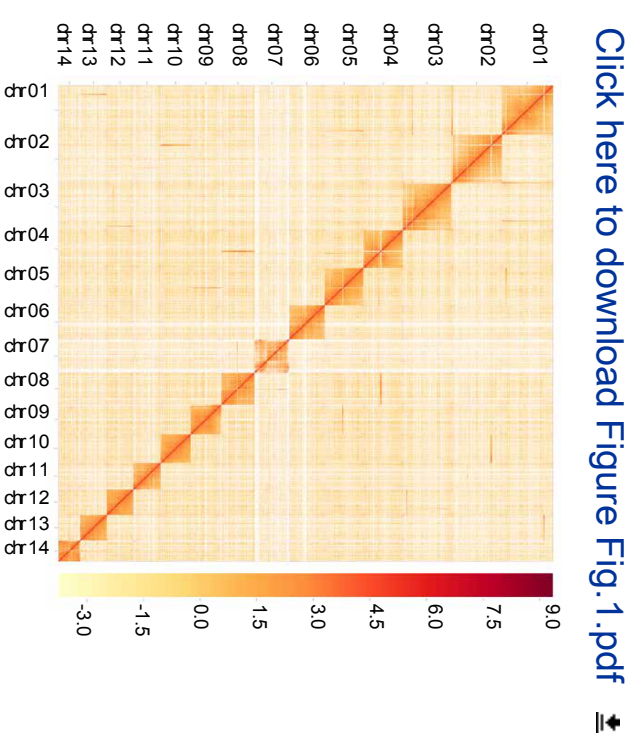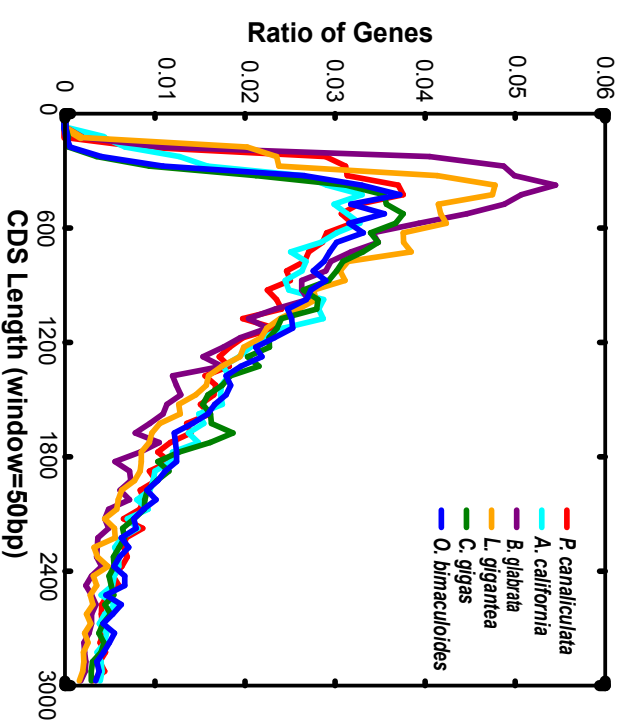
809   doi:Artn 86010.3389/Fmicb.2015.00860.

810 75. Zoetendal EG, Heilig HG, Klaassens ES, Booijink CC, Kleerebezem M, Smidt H, et al.

811   Isolation of DNA from bacterial samples of the human gastrointestinal tract. Nature

812   protocols 2006, 1(2): 870-873

813 76. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic
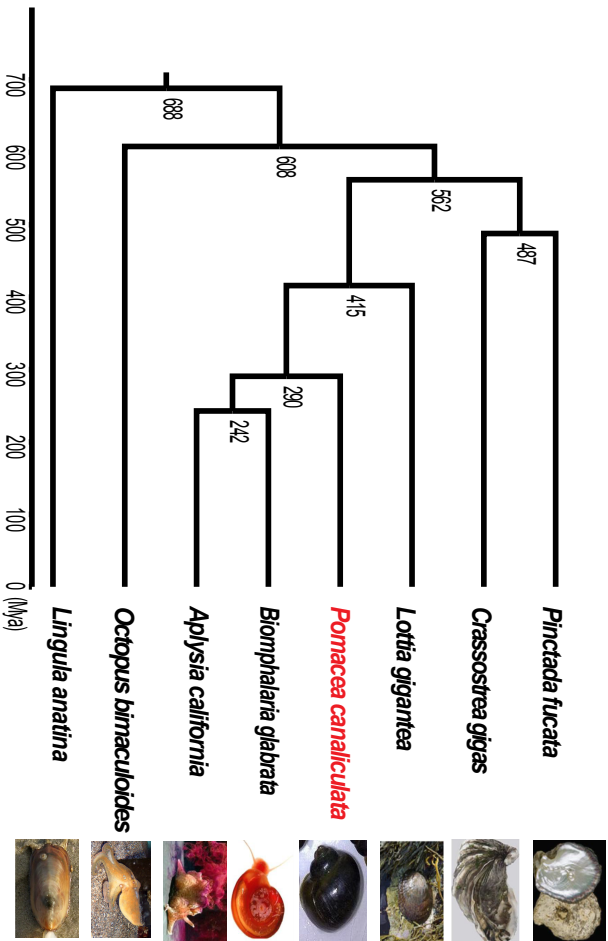
814   gene structure annotation using EVidenceModeler and the Program to Assemble Spliced

815   Alignments. Genome Biol. 2008;9:R7.

816   77.   Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan:

817   protein domains identifier. Nucleic Acids Res. 2005;33:W116-20.

818   78.   Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and

819   interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109-D14.

820   79.   Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids

821   Res. 1999;27:573-80.

822   80.   Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high

823   throughput. Nucleic Acids Res. 2004;32:1792-7.

824   81.   Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models

825   of protein evolution. Bioinformatics. 2011;27:1164-5. doi:10.1093/bioinformatics/btr088.

826   82.   Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler

827   transform. Bioinformatics. 2009;25:1754-60.

828   83.   Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile

829   metagenomic assembler. Genome Res. 2017;27:824-34.

830   84.   Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site

831   prediction in metagenomic sequences. Bioinformatics. 2012;28:2223-30.

832   85.   Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation

833   sequencing data. Bioinformatics. 2012;28:3150-2.

834   86.   Buchfink B, Chao X, Huson DH. Fast and sensitive protein alignment using DIAMOND.

835   Nat Methods. 2015;12:59-60.

836 87. Gerlach W and Stoye J. Taxonomic classification of metagenomic shotgun sequences

837   with CARMA3. Nucleic Acids Res. 2011;39 doi:Artn E9110.1093/Nar/Gkr225.

838 88. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for

839   deciphering the genome. Nucleic Acids Res. 2004;32:D277-80.

840 89. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated

841   carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012;40:W445-51.

842 90. Eddy SR. Accelerated Profile HMM Searches. Plos Comput Biol. 2011;7 doi:ARTN

843   e100219510.1371/journal.pcbi.1002195.

844 91. Qin JJ, Li YR, Cai ZM, Li SH, Zhu JF, Zhang F, et al. A metagenome-wide association

845   study of gut microbiota in type 2 diabetes. Nature. 2012;490:55-60.

846

Figure

a

b

c

Figure

a

b

Ratio of TE Pairs

Sequence Divergence Ratio

Crassostrea gigas
Octopus bimaculoide.
Lottia gigantea
Biomphalaria glabrata
Aplysia california
Pomacea canaliculata

a

CYP2F
CYP2B
CYP1A
CYP2C
CYP2U
CYP20A
CYP4A
CYP4V
CYP2J
CYP4F
CYP26A
CYP3A
CYP27A
CYP10
CYP49A

P. canaliculata
A. california
B. glabrata
C. gigas
L. gigantea
P. fucata
O. bimaculoides

b

Hem Te Ov Kn Gl Hp Em   Con Cold Heat Hm Exp

CYP1A
CYP2U
CYP3A
CYP4F

2.0

2^{-1} 2^1 2^3 2^5 2^7 0 5 10 15 20

P. canaliculata

Carbohydrates

Pervitellins

PcOvo PcPV2 PcPV3

Embryo
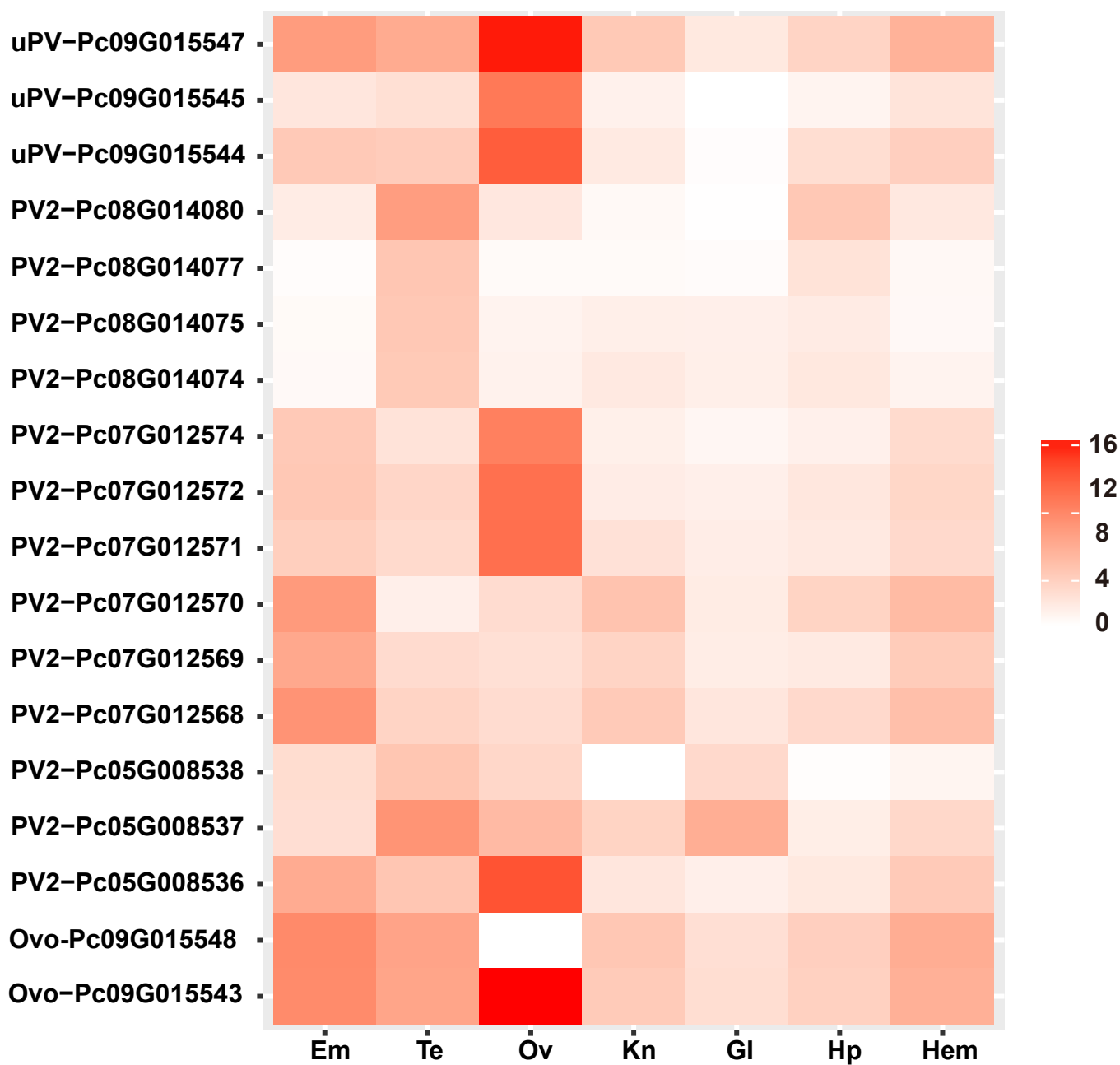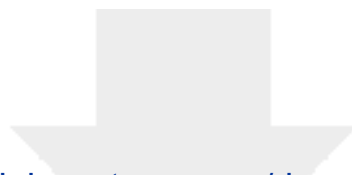
Lipids
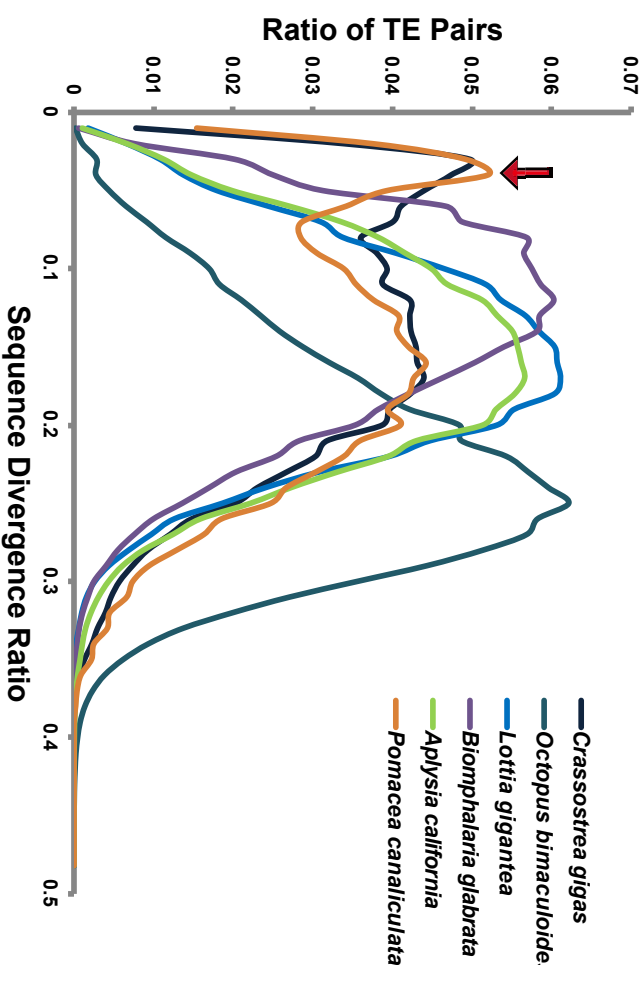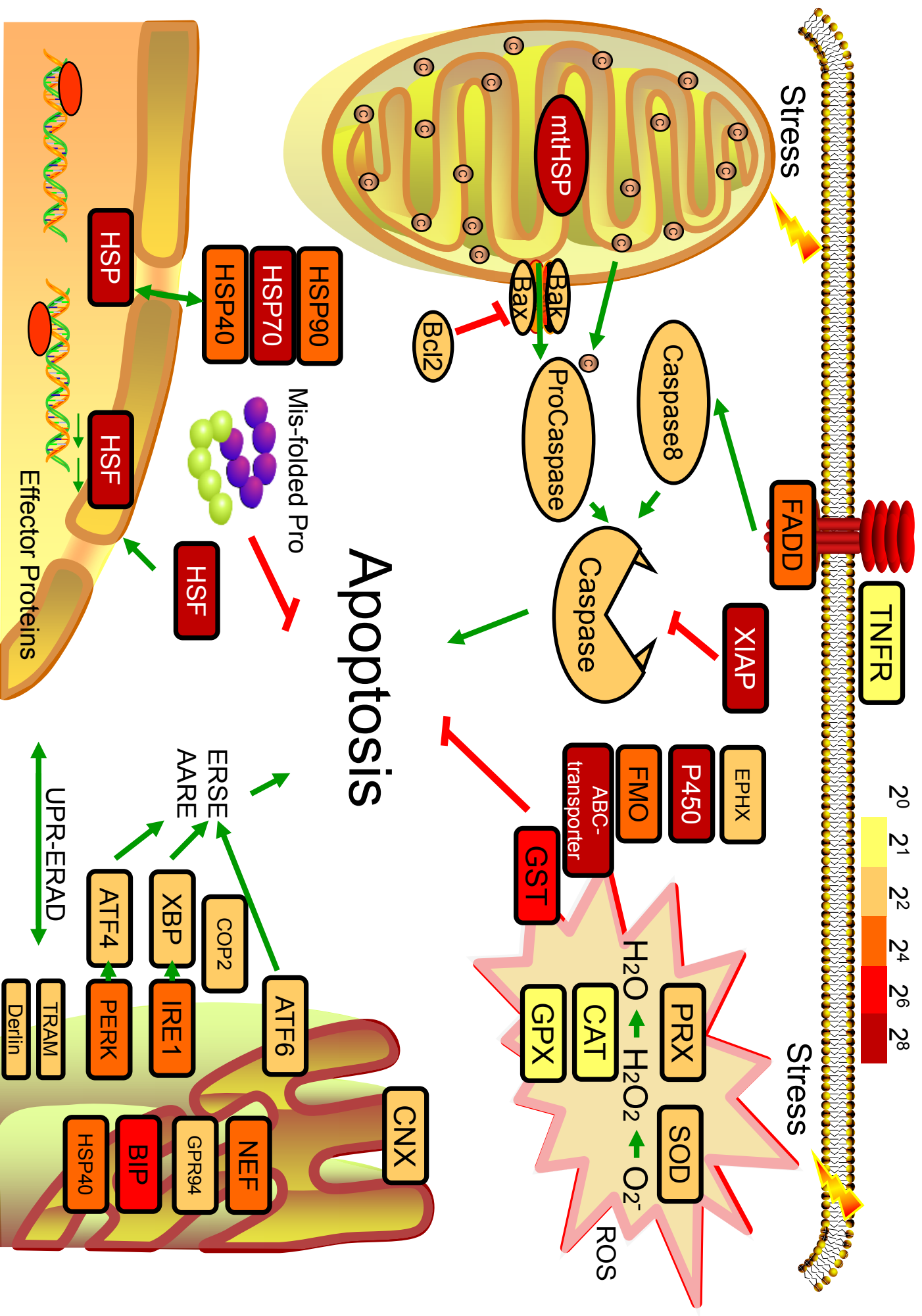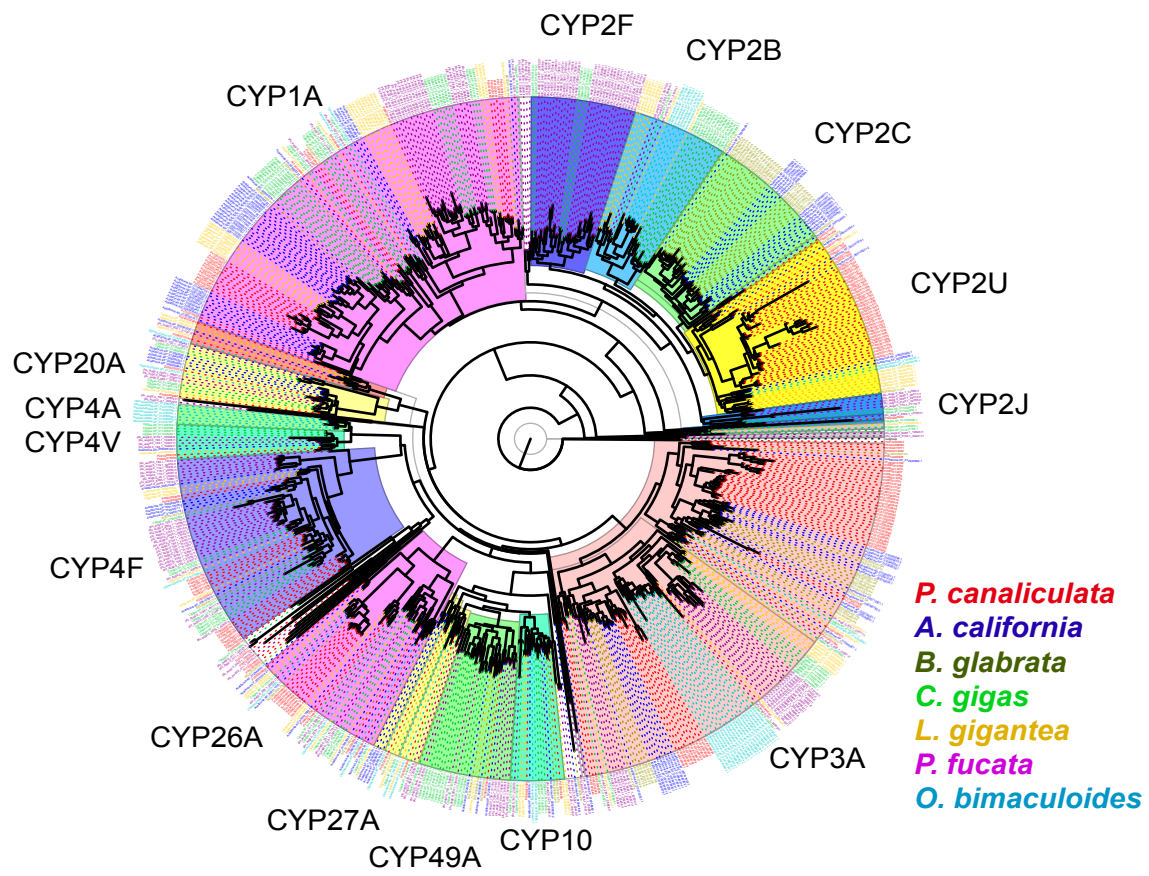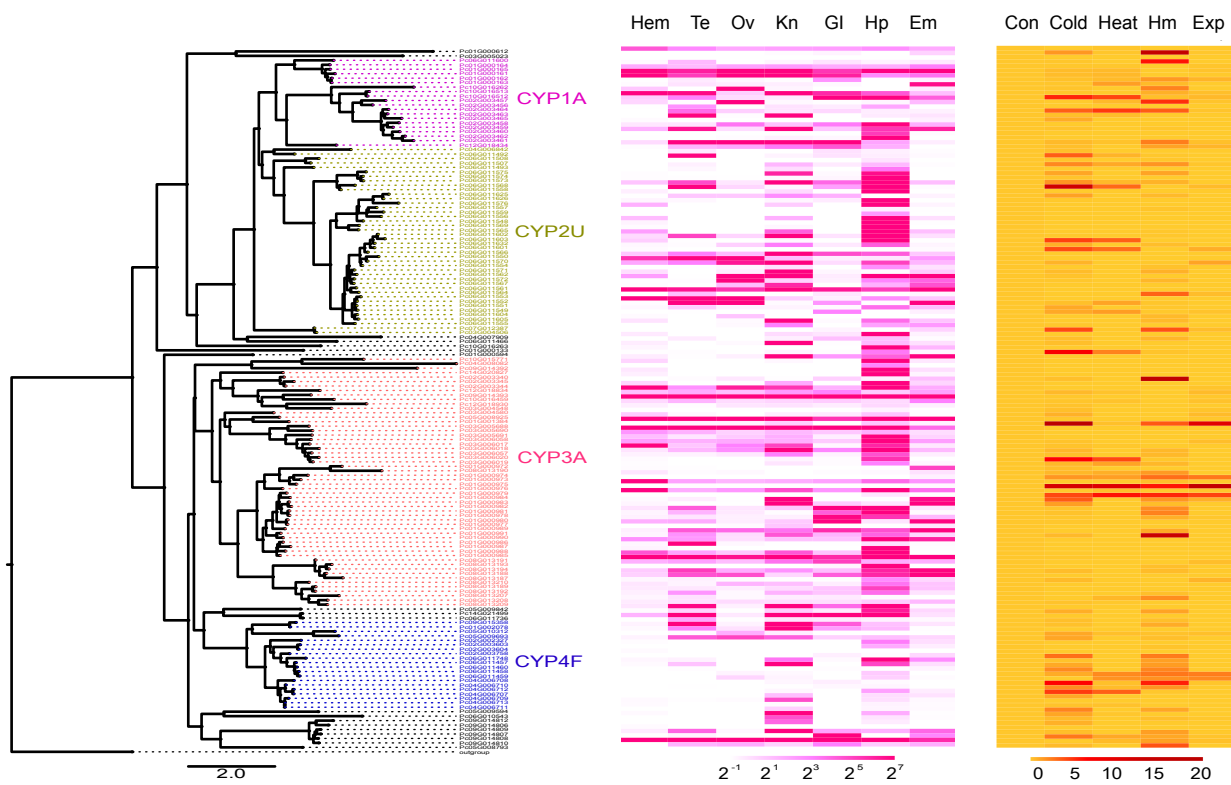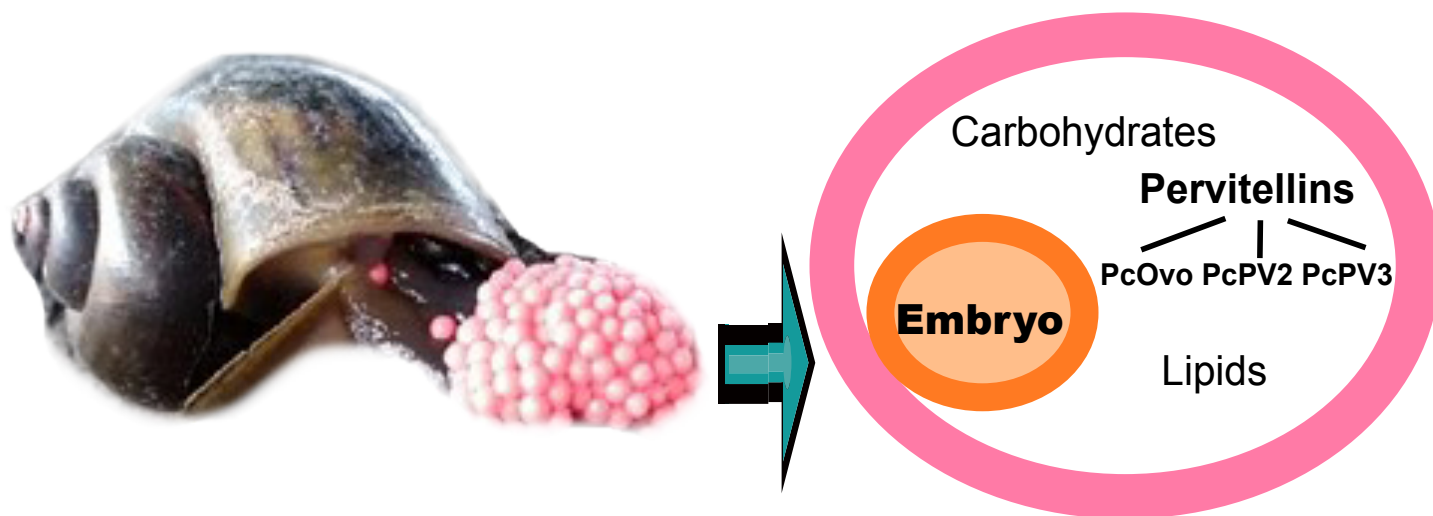
b

Click here to access/download
Supplementary Material
Supplemental Information.doc

Dear Laurie and Scott,

We are delighted to submit our genome paper of golden apple snail to GigaScience. We appreciate any of your advices.

It was 8 years ago that I worked on the panda genome project in BGI, and I have always been grateful for Laurie's kind revision of that manuscript published in *Nature*. It was 6 years ago that I wrote a review paper on the sequence assembly algorithm, and I was in debt to Scott for helping me revise that manuscript later published on Briefings in Functional Genomics. Now I am working at Agricultural Genomics Institute, Chinese Academy of Agricultural Sciences, being a PI researcher in agricultural genomics, and focusing mainly in pest animals and microbiome.

The golden apple snail is an important worldwide invasive animal, listed in the top-100 worst invasive species. It has become a major pest in the rice field, causing huge economic loss each year but lack of efficient preventing approaches. By PacBio sequencing and Hi-C technology, we have assembled the genome into 14 chromosomes, which is the best available genome sequence in Molluscs. Key findings include the recent explosion of DNA/hAT-Charlie TEs, the expansion of P450 gene family and the constitution of cellular homeostasis system, contributing to the ecological plasticity in the stress adaptation, as well as the perivitellin gene expansion and high transcriptional level in ovary that promotes the function of nutrients supplying and defense ability in the eggs. We also analyzed the gut metagenome and found rich genes for food digestion and xenobiotics degradation. The golden apple snail possesses potential to be a model organism of molluscs, and we believe that with a high-quality reference genome, it will become more important in molluscs researches.

Thank you for your consideration. We would really appreciate if you could accelerate the processing of our manuscript given a highly competitive situation.

Best wishes,

Wei Fan