

GigaScience

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

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00030R1	
Full Title:	The genome of golden apple snail <i>Pomacea canaliculata</i> provides insight into stress tolerance and invasive adaptation	
Article Type:	Research	
Funding Information:	National key research and development program of China (2016YFC1200600)	Dr Wei Fan
	Shenzhen science and technology program (JCYJ20150630165133395)	Dr Wei Fan
	Fund of Key Laboratory of Shenzhen (ZDSYS20141118170111640)	Dr Wei Fan
	The Agricultural Science and Technology Innovation Program (ASTIP) of Chinese Academy of Agricultural Sciences(CAAS) & Elite Youth Program of Chinese Academy of Agricultural Sciences	Dr Wei Fan
Abstract:	<p>Background: The golden apple snail (<i>Pomacea canaliculata</i>) is a fresh water snail listed among the top-100 worst invasive species, worldwide and a noted agricultural and quarantine pest that causes great economic losses. It is characterized by fast growth, strong stress tolerance, a high reproduction rate, and adaptation to a broad range of environments.</p> <p>Results: Here, we used long-read sequencing to produce a 440-Mb high-quality chromosome-level assembly for the <i>P. canaliculata</i> genome. In total, 50 Mb (11.4%) repeat sequences and 21,533 gene models were identified in the genome. The major findings of this study include the recent explosion of DNA/hAT-Charlie transposable elements (TEs), the expansion of the P450 gene family and the constitution of the cellular homeostasis system, which contributes to ecological plasticity in stress adaptation. In addition, the high transcriptional levels of perivitellin genes in the ovary and albumen gland promote the function of nutrient supply and defence ability in eggs. Furthermore, the gut metagenome also contains diverse genes for food digestion and xenobiotic degradation.</p> <p>Conclusions: These findings collectively provide novel insight into the molecular mechanisms of the ecological plasticity and high invasiveness.</p>	
Corresponding Author:	Wei Fan Chinese Academy of Agricultural Sciences CHINA	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Chinese Academy of Agricultural Sciences	
Corresponding Author's Secondary Institution:		
First Author:	Conghui Liu	
First Author Secondary Information:		
Order of Authors:	Conghui Liu	
	Yan Zhang	
	Yuwei Ren	
	Hengchao Wang	
	Shuqu Li	

	Fan Jiang
	Lijuan Yin
	Guojie Zhang
	Wanqiang Qian
	Bo Liu
	Wei Fan
Order of Authors Secondary Information:	
Response to Reviewers:	<p>GIGA-D-18-00030 The genome of golden apple snail <i>Pomacea canaliculata</i> provides insight into stress tolerance and invasive adaptation Conghui Liu; Bo Liu; Yuwei Ren; Yan Zhang; Hengchao Wang; Shuqu Li; Fan Jiang; Lijuan Yin; Guojie Zhang; Wanqiang Qian; Wei Fan GigaScience</p> <p>Dear Dr Fan,</p> <p>Your manuscript "The genome of golden apple snail <i>Pomacea canaliculata</i> provides insight into stress tolerance and invasive adaptation" (GIGA-D-18-00030) has been assessed by our reviewers. Although it is of interest, we are unable to consider it for publication in its current form. The reviewers have raised a number of points which we believe would improve the manuscript and may allow a revised version to be published in GigaScience.</p> <p>Reply: we have made revisions according to the reviewer's suggestions.</p> <p>Their reports, together with any other comments, are below. Please also take a moment to check our website at https://giga.editorialmanager.com/ for any additional comments that were saved as attachments.</p> <p>Please consider including more recent data on other molluscs in your analyses - see the report of reviewer 3 below. Reply: We have added a new mollusc species "golden mussel <i>Limnoperna fortunei</i>", and replaced the data for "pearl oyster <i>Pinctada fucata</i>" with the latest version.</p> <p>Please carefully revise the manuscript for language use and grammar, ideally with the help of a native speaker. Please note the attached file of one of the reviewers (available via Editorial Manager), which contains some suggestions for improvements Reply: We have revised the language and grammar, and asked a native speaker for polishing. We also adopted the suggestions from the attached file of one of the reviewers.</p> <p>Author roles: I note that you indicate four "equally contributing" first authors. Please note that we cannot indicate more than three "equally contributing" co-first authors, and please be aware that shared first authorship is reserved for exceptional cases where the contribution of two or three authors is indeed exactly equal. Reply: We have reduced the co-first authors to three.</p> <p>If you are able to fully address these points, we would encourage you to submit a revised manuscript to GigaScience. Once you have made the necessary corrections, please submit online at:</p> <p>https://giga.editorialmanager.com/</p> <p>Reply: After fully addressed all the points, we re-submitted the manuscript to GigaScience.</p> <p>If you have forgotten your username or password please use the "Send Login Details" link to get your login information. For security reasons, your password will be reset.</p> <p>Please include a point-by-point within the 'Response to Reviewers' box in the submission system. Please ensure you describe additional experiments that were</p>

carried out and include a detailed rebuttal of any criticisms or requested revisions that you disagreed with. Please also ensure that your revised manuscript conforms to the journal style, which can be found in the Instructions for Authors on the journal homepage.

The due date for submitting the revised version of your article is 21 Jun 2018.

I look forward to receiving your revised manuscript soon.

Best wishes,

Hans Zauner
GigaScience
www.gigasciencejournal.com

Reviewer reports:

Reviewer #1: In their manuscript Liu et al. reported the genome sequence of the golden apple snail *Pomacea canaliculata*. They constructed chromosomal-level genome assembly using HiSeq, PacBio, and Hi-C sequencing technologies. They also tested differential gene expression under various environmental stress, showing many genes are responded to maintain homeostasis. In addition, they sequenced gut metagenome of the snail for the first time, implying that microorganisms contribute to digestion and resistance to xenobiotics of the host animal.

I think the massive data provides fundamental information to understand the biology of the animal as well as molluscs, therefore the study is valuable to be published in the journal GigaScience after some corrections.

Overall, the methods are appropriate, but description and interpretation of the results look not sufficient in some points as shown below.

P. 5, lines 94-96

"such as California sea hare, Pacific oyster, Pearl oyster,..."
should be "such as the California sea hare, the Pacific oyster, the pearl oyster,..."
There are many mistakes like this. I won't mention all of them. Please consult professional English editor before submitting the revision.

Reply: We have corrected this mistake in the new submitted manuscript.

P. 7, lines 148-150

"genes from seven related species..."
In fact eight species including *Pinctada fucata* were analyzed in figures 2a and 4a. Takeuchi et al.(2016, *Zoological Letters*, 2:3) and Luo et al.(2015, *Nature Communications*, 6, 8301) should be referred for *P. fucata* and *Lingula anatina* genome data, respectively.

Reply: We have corrected the species number. Because a new species is added into analysis, now the total species number is nine. The reference paper of the new species "Uliano-Silva M, Dondero F, Dan Otto T, Costa I, Lima NCB, Americo JA, et al. A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel *Limnoperna fortunei*. *Gigascience*. 2017. doi: 10.1093/gigascience/gix128." were also added at line 104 in the new submitted manuscript.

In addition, please carefully correct scientific names in Abbreviations and figures.

"*Lottia gigantean*" should be "*Lottia gigantea*"

"*Aplysia californica*" should be "*Aplysia californica*."

"*Lingula anatine*" should be "*Lingula anatina*"

Reply: we have corrected all the mistakes on scientific names in Abbreviations and figures.

P. 9 178-179

From the results I could not understand how the idea that the "DNA/hAT-Charlie TEs... promote the potential plasticity in the stress adaptation" came. This hypothesis can be tested using the present RNA-seq data, by checking whether the TEs are up-regulated under the stresses.

Reply: Transposons can insert into any genomic regions, which may change the gene regulations, or modify the gene structure thus form new functions. If a genome has high transposon activity, then it has high ability to adapt to the changing environment, so the recent explosion of DNA TEs may benefit the fast evolution of *P. canaliculata* in the recent history. There were several previous studies (Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. The struggle for life of the genome's selfish architects. *Biol Direct.* 2011;6:19; Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proc Natl Acad Sci U S A.* 2011;108:10863-70) on this issue that provides evidences that TEs can introduce small adaptive changes for a species. Using the RNA-seq data to resolve this question is good idea. In our understanding, TEs can't be transcribed and translated as an independent element, except for some low and random transcriptions which are likely to be no functions. So we analyzed the expression of 709 genes including DNA elements that restricted to the 4% peak inside the gene region, compared with the other genes that outside the 4% peak. Differentially expressed genes (DEG) were defined here by P-value smaller than 0.05 for comparison of treatments (heat, cold, heavy metal and air exposure) and control data. The percent of DEGs in the 4% peak were higher than those of genes outside the peak (10.2% higher for heat, 8.6% higher for cold, 8.6% higher for heavy metal, and 7.3% higher for air exposure). Among the DEGs in the 4% peak, about half are up-regulated and the other half are down-regulated. Moreover, the DEGs in the 4% peak were mainly enriched in cellular metabolic process, response to stimulus, localization and signaling by GO annotation. These results indicated that genes in the 4% peak were likely to be more active in the response of stimulus, promoting the potential plasticity in the stress adaptation. The figure and related context was added in the new manuscript.

P.11 lines 232-236

The authors claimed that the *P. canaliculata* CYP gene family expanded compare to other molluscs. But the gene expansion of CYP looks common among molluscs. The number of the gene in *P. canaliculata* didn't significantly stand out from other molluscs (for example *P. canaliculata* has 157 genes and the Pacific oyster has 135). A molecular phylogeny in Fig 4a shows that lineage-specific gene expansion of CYP occurs not only in *P. canaliculata* but also other molluscs.

Reply: We appreciate the reviewer's comments. We claimed CYP450 family as expansion for two reasons. 1) Although the gene number of CYP450 in *P. canaliculata* was close to *C.gigas* (135) and *A. californica* (128), the gene ratio of "CYP450 genes/total genes in genome" was distinct, namely *P. canaliculata* (157/21553, ~0.0073), *C.gigas* (135/46748, ~0.0029) and *A. californica* (128/27591, ~0.0046). 2) The expansion was more obvious in special subfamily, such as CYP3A (*P.c* 56, *A.c* 24, *Bg* 21, *C.g* 12, *L.g* 10, *P.f* 23 and *O.b* 20) and CYP 2U (*P.c* 42 *A.c* 5, *Bg* 2, *C.g* 0, *L.g* 8, *P.f* 0 and *O.b* 2) (figure 4). However, we weaken the mood when refer to expansion and withdraw the adjective "great".

P. 17 lines 346-354

"The rich phenotypic... in laboratory."

These sentences should be move to Introduction.

Reply: We have moved "The rich phenotypic... in the laboratory." into introduction part between line 71 and 80.

P. 18 lines 380-381

"total messenger RNAs"

Total RNA or messenger RNA?

Reply: Here we mean "messenger RNA". The sentence was revised to be "In final, total RNAs were extracted from the stored tissues of *P. canaliculata* materials, and then mRNAs were pulled out by beads with poly-T for constructing cDNA libraries."

P. 21 line 445

Please cite the literature of "previous results."

Reply: We have added the literature "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. *Nature Ecology & Evolution.* 2017.1(5), 121; Benton MJ, Donoghue, PCJ,

Asher RJ. in *The Timetree of Life: Calibrating and Constraining Molecular Clocks* (eds Hedges, S. B. & Kumar, S.) 35–86 Oxford Univ. Press, 2009; Zapata F, Wilson NG, Howison M, Andrade SC, Jörger KM, Schrödl M, Goetz FE, Giribet G, Dunn CW. Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. *Proc Biol Sci.* 2014 281:20141739. doi: 10.1098/rspb.2014.1739” in revised version.

Figure 2

The title "Evolutionary genomic analysis between *P. canaliculata* and other molluscs" is not appropriate because *Lingula* is a brachiopod.

Reply: The title of Figure 2 is changed to "Evolutionary genomic analysis of *P. canaliculata*", because our focus is the species of *P. canaliculata*, other species were used for comparison.

Figure 4

Method for molecular phylogeny construction of CYP genes should be described.

Reply: The method was described in Figure 4 legend "The tree was constructed using the maximum likelihood method in MEGA7, and the branch length scale indicates the average number of residue substitutions per site".

Figure S1

Which K-mer size used?

Reply: here we used 17-mer, the K-mer size is 17.

Table S4, S5, and S7

It is not reader-friendly to show the huge data in a table. I couldn't recognize what is the message of the data. Why not visualize the data in a heat map like Fig4b.

Reply: A supplemental figure S6 corresponding to Table S4 was added.

Data in Table S5 was corresponding to the color in the heatmap of figure 4b, Data in Table S7 was corresponding to the color in the heatmap of figure 5b, so there is no need to add other heat map figures.

Table S9

What "Mean" and "SD" indicate? E-value of blast results? Please describe.

Reply: "Mean" and "SD" indicate the mean and standard deviation of relative abundance of a phylum or a genus from the 6 gut microbiota samples. We also added a note under the Table.

Reviewer 2: - [the reviewer has no specific comments to the authors at this point, but recommends careful improvements of language and grammar]

Reply: We have improved the language and grammar, and polished the text by native speakers.

Reviewer #3: This manuscript presents a high-quality genome assembly for the snail *P. canaliculata*. Such genome and further analysis presented will contribute deeply for future studies of the molecular evolution and adaptation of molluscs, as well as to the study of the molecular mechanisms leading to - or involved with - invasive species success. I also point out the relevance of a first qualitative description of a high-depth gut microbiome for a snail. For such reasons, I recommend the publication of this manuscript. Nevertheless, I would like to recommend some essential revision prior publication.

First, the English has to be revised. I'll give a few examples below, and authors will find major marks in purple concerning specifically the need of English revision in the revised pdf attached. However, the entire manuscript would benefit from a native English speaker revision.

Reply: we have revised the descriptions highlighted in the attached pdf, and also asked a native speaker to help polish the language.

Examples of sentences needing English revision:

Lines 50-51: "causing severe economic loss each year as a result of yield loss,

replanting cost and the funds of control." - rephrasing necessary.

Reply: This sentence is modified to "causing severe economic losses each year as a result of yield loss, replanting cost and expenditures on control."

Line 52: "More seriously, *P. canaliculata* has involved in the transmission of a human fatal disease."

Reply: This sentence is modified to "More seriously, *P. canaliculata* has been involved in the transmission of a fatal human disease,"

Line 57: "causing great challenge to human health"

- rephrasing advised.

Reply: This sentence is modified to "creating a great challenge in terms of human health."

Line 58: "Molluscs is ..." - English correction necessary.

Reply: This sentence is modified to "Molluscs are a highly diverse group, second only to arthropods in species number, and their high biodiversity makes them an excellent model to address issues such as biogeography, adaptability and evolutionary processes."

Lines 92-94: "However, researches at whole genome level in *P. canaliculata* still lags far behind other mollusks species, due to the lack of a high-quality reference genome. By far, multiple draft..." - rewriting necessary.

Reply: This sentence is modified to "However, research at the whole-genome level in *P. canaliculata* still lags far behind that in other mollusc species due to the lack of a high-quality reference genome. Multiple draft genomes of molluscs have been published, including the genomes of the California sea hare, Pacific oyster, pearl oyster, owl limpet, California two-spot octopus, golden mussel, and *Biomphalaria* snails, greatly promoting research on mollusc genomics."

Line 263: "was" should be "were".

Reply: Corrected in the new manuscript.

Data and analysis related comments:

Lines 36-37: The description of the genome and the several molecular expression data are great contributions for the further understanding of molluscan and invasive biology. Nevertheless, we should avoid direct jumps to conclusions such as in lines 36 and 37, as the results in the manuscript don't present tools or direct ways to prevent invasions or pathogen transmission. I advise the withdraw of such sentence.

Reply: We agree to the suggestion, and have removed that sentence "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."

Line 47: I would rephrase the sentence here in line 47. Even though the biology of the species may positively influence its invasive capacity, such characteristics are not exclusive of invasive mollusks. For that reason, I would exclude the "was due to" (line 47) which implies causality.

Reply: In the revised manuscript, we rephrased "was due to" to be "is closely related to".

Line 63: Please present and refer to the lower temperature the species can establish populations in.

Reply: We added a sentence "*P. canaliculata* has been reported to establish populations at temperatures ranged from 10 °C to 35 °C" in the new manuscript, as well as two reference papers (Seuffert ME, Burela S, Martín PR. Influence of water temperature on the activity of the freshwater snail *Pomacea canaliculata* (Caenogastropoda: Ampullariidae) at its southernmost limit (Southern Pampas, Argentina). *Journal of Thermal Biology*. 2010; 35:77-84; Matsukura K, Tsumuki H, Izumi Y, Wada T. Physiological response to low temperature in the freshwater apple snail, *Pomacea canaliculata* (Gastropoda: Ampullariidae). *J Exp Biol*. 2009;212:2558-63).

Line 95: I would cite here also the draft genome of the invasive *Limnoperna fortunei* mussel.

Reply: The golden mussel "*Limnoperna fortunei*" and the related article were added in the new manuscript.

Line 95: There is a new version of the Pearl oyster published. If analysis were performed with data cited in line 95, I would advise for updating the analysis with proteins from the new genome (Du X, Fan G, Jiao Y et al. The pearl oyster *Pinctada fucata martensii* genome and multi-omic analyses provide insights into biomineralization. *Gigascience* 2017;6(8):1-12).

Reply: We have replaced the proteins data of *Pinctada fucata* to the latest version, and updated all the analysis in the new manuscript.

Line 100-101: Rephrasing is necessary as cellular homeostasis, color and nutrient of the eggs are not species-specific invasive characteristics.

Reply: We revised "invasive characters" to "environmental adaptation characteristics".

Line 104-105: same argument as for lines 36-37. Some rephrasing starting from "interrupt transmission..." is necessary.

Reply: We agree with the suggestion, and weakened the mood. The sentence is modified to "and provide a basis for interrupting the transmission of pathogenetic nematode parasites".

Table S1: Table S1 would benefit of having 2 columns: one with (i) number of reads generated and (ii) total bp produced for each library, instead of having a column 'Data size' (and what G bp means?).

Reply: We have made 2 columns in Table S1 according to the suggestions. One column refers to number of sequenced reads, the other column refers to number of sequenced bases.

Line 122: The ratio of genome coverage by reads used as input in the assembly? Rephrase it together with the sentences in lines 126-127, please.

Reply: In this sentence "another important aspect for evaluating genome assembly is the ratio of genome coverage." (between line 132 and 133), we want to explain that the ratio of assembly coverage is important. In *P. canaliculata*, the genome size of 446 Mb was estimated by the distribution of k-mer frequency. In this assembly genome, ~98.6 % sequence has been assembled.

In the sentence "we mapped the Illumina shotgun reads to the assembled reference genome. Significantly, 97% and 95% of the genome-derived and transcriptome-derived reads, respectively, could be aligned to the reference genome," (between line 136 and 137), we want to confirm the accuracy and no obvious bias for sequencing and assembly.

Line 123-124 and line 403: Please estimate and present the levels of heterozygosity using the illumina reads.

Reply: We used K-mer with K-size 17 to estimate the genome heterozygosity based on algorithm from reference (Liu B, et al. *Quantitative Biology* 2013:arXiv:1308.2012 [q-bio.GN]). The estimated heterozygosity of *P. canaliculata* range from 1% to 2%. In addition, we also used FIndError (Gnerre S et al., 2011) in the Allpath-LG package to estimate the heterozygosity, the result is 1.75%, consistent with the first method. We have added it in revised manuscript. "With an estimated genome size of 446 Mb and genome heterozygosity between 1% and 2% based on the distribution of k-mer frequency."

Line 415-416: "Then, the protein-coding sequences were mapped by RNA-seq data." - please explain this sentence.

Reply: To determine whether the predicted genes are expressed or not, we used the transcriptome data to map to the CDS of genes. The gene models were retained if they had at least one supporting evidence from UniProt database, InterProScan domain and RNA-seq data.

To be more clear, we have revised this sentence in the new manuscript: "Then, these gene models were annotated by RNA-seq data, UniProt database and InterProScan software".

Line 163: Withdraw "and so on".

Reply : "and so on" is removed.

Lines 146-163: To start understanding if the genome composition itself - and not only regulation of gene expression - can play a major role in the success of invasive species, I would advise to compare gene family expansions and contractions between the genomes of two invasive mollusks, which is now possible once the draft genome of *L. fortunei* is available (GigaScience doi: 10.1093/gigascience/gix128.). Further discussion about the presence - or lack thereof - of common expansions and contractions of gene families would be a great contribution. Such gene families could be further investigated for their roles in the expression of phenotypes related to invasive ecology and behaviour. I would strongly suggest for a comparative analysis of *P. canaliculata* and *L. fortunei* protein sets leading to a new Figure S4 and brief discussion on the findings.

Reply: We agree with the reviewer's comments. In the revised version, we added the genome data of *L. fortunei* to re-construct the orthoFinder ortholog and paralog gene families. Then, we identified the common expanded gene families both in *P. canaliculata* and *L. fortunei*. The functions of these gene families are mainly enriched in signal transduction, replication and repair, Translation, glycan biosynthesis and metabolism, Lipid metabolism, endocrine, immune and nervous system. And we have revised the results in Figure S4.

Line 171-172: "interestingly, only the results of DNA transposons showed a unique peak at ~4% divergence rate for *P. canaliculata* and *C. gigas*" - rewrite this sentence.

Reply: We rephrased it as "Notably, the TE class of DNA transposons showed a specific peak at a divergence rate of ~4% divergence rate for *P. canaliculata* and *C. gigas*".

Line 249: Please indicate how many and which genes were highly induced to facilitate further investigation by other groups in the future.

Reply: Gene IDs "Pc06G011748, Pc06G011460, Pc06G011458, Pc06G011459, Pc04G006708, Pc04G006710 and Pc04G006707" were added.

Line 254 -257: This direct link between phenotype and molecular characteristics cannot be supported by your data. Please rephrase it.

Reply: We revised these sentences in the new manuscript:

"*P. canaliculata* has eggs characterized by abundant nutrients, reddish or pinkish colour, aerial oviposition and neurotoxicity due to the perivitelline Fluid (PVF), which fills the space between the eggshell and the embryo and consists of carbohydrates, lipids and proteins (Figure 5a)."

Line 264- 269: Please clarify what was performed here. In any case, blast alone is not the best tool to predict orthology. I would use RBBH methods.

Reply: In the revised manuscript, we used the reported 59 PVF protein fragments as query to identify the PVF genes from *P. canaliculata* reference genes by blastp (e-value 10⁻⁵), and further used the requirements of more than 85% sequence identity and over half alignment length for the query to get 36 best hits, corresponding to 28 candidate *P. canaliculata* PVF genes, and then confirmed 6 perivitellin genes which encode the subunits of PcOvo, PcPV2, and PcPV3, according to their high RNA expression in ovary and albumen gland tissue.

As OrthoFinder could analyze the orthology and paralogy of more than two species at the same time, we utilized the OrthoFinder results to investigate the ortholog and paralog relationships of these *P. canaliculata* PVF genes compared with other 8 sequenced mollusc species. Notably, 5 of the 6 perivitellin genes fall into single-gene families. In other words, it is hard to detect any homologs for most of these perivitellin genes in other sequenced mollusc species. One reason may be that the divergence time is too long (>200 Mya), another reason may be that these major PVF genes may have experienced fast evolution in the history, in order to adapt to the changing environment. At last, we used the RBBH method (reciprocal best hit) to identify ortholog genes between each species pair, and the result is consistent with that of OrthoFinder.

Line 327: Conclusion and discussion? At this point, only conclusions should be stated.

Please eliminate sentences from 346 to 359. [condense and add into introduction]
 Reply: We have condensed this paragraph, and moved it to the introduction part between line 376 and line 382. The revised paragraph is:
 “In this study, we report a fine reference genome of *P. canaliculata*, first chromosome-level Mollusca genome published. With its easy acquisition, rapid growth and efficient reproduction, *P. canaliculata* possesses the potential to be a model organism of Mollusca. As its the cellular complexity and conservation of pathways also make *P. canaliculata* a useful representative of Mollusca, the genome described in this study can be used to advance our understanding of the molecular mechanisms involved in various scientific questions regarding Mollusca.”

Line 395: Please indicate software used for trimming.
 Reply: We used an in-house software for trimming (clean_dapter, clean_lowqual, filter_unpaired_reads.pl), which is freely available at Github “https://github.com/fanagislab/common_use”.

Line 424-431: I would state the masking before stating the gene prediction. Rewrite.
 Reply: We have moved the repeat paragraph before the gene prediction paragraph.

Line 448: Any trimming performed for the transcriptome?
 Reply: Yes, we use the same method. This sentence is revised to:
 “Transcriptome reads were trimmed with the same method for genomic reads (https://github.com/fanagislab/common_use), and then mapped to the reference genome of *P. canaliculata* using TopHat (v. 2.1.0) with default settings”.

Line 591: Please make available a supplementary material with the IDs of all sequences presented in Figure 4b. Please explain the scale in the heat maps of figure 4b.
 Reply: The IDs are listed in Supplemental table S5.
 We explain the scale meaning in the legend of figure 4b. The scale for the left heat map represent FPKM value, showing by gradually changing colors; The scale for the right heat map represent fold change (FPKM-stimulus/FPKM-control), showing by gradually changing colors. To be more clear, we also add two marks “FPKM” and “Fold-change” alongside the scale on the figure.

--
 Please also take a moment to check our website at <https://giga.editorialmanager.com/l.asp?i=38889&l=8DS0D5CA> for any additional comments that were saved as attachments. Please note that as GigaScience has a policy of open peer review, you will be able to see the names of the reviewers.

Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes
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?	

<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 Standards Reporting Checklist?</p>	Yes

1 **The genome of the golden apple snail *Pomacea canaliculata* provides insight into**
2 **stress tolerance and invasive adaptation**

3 Conghui Liu^{1*}, Yan Zhang^{1*}, Yuwei Ren^{1*}, Hengchao Wang¹, Shuqu Li¹, Fan Jiang¹, Lijuan Yin¹,
4 Guojie Zhang², Wanqiang Qian^{1†}, Bo Liu^{1†}, Wei Fan^{1†}

5 ¹Agricultural Genomic Institute, Chinese Academy of Agricultural Sciences,
6 Shenzhen, Guangdong, 518120, China.

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

8 Conghui Liu: rapherlch@163.com; Yan Zhang: milrazhang@163.com; Yuwei Ren:
9 xiaoshudaxia@126.com; Hengchao Wang: wanghengchao000@qq.com; Shuqu Li:
10 lishuqu1234@163.com; Fan Jiang: greatjf@163.com; Lijuan Yin:
11 yinlijuan1005@163.com; Guojie Zhang: guojie.zhang@bio.ku.dk

12 *These authors contributed equally to this work.

13 †Correspondence should be addressed to Wanqiang Qian (qianwanqiang@caas.cn),
14 Bo Liu (lb_bobo@aliyun.com) or Wei Fan (fanwei@caas.cn).

15 **Abstract**

16 **Background:** The golden apple snail (*Pomacea canaliculata*) is a fresh water snail
17 listed among the top-100 worst invasive species, worldwide and a noted agricultural
18 and quarantine pest that causes great economic losses. It is characterized by fast
19 growth, strong stress tolerance, a high reproduction rate, and adaptation to a broad
20 range of environments.

21 **Results:** Here, we used long-read sequencing to produce a 440-Mb high-quality
22 chromosome-level assembly for the *P. canaliculata* genome. In total, 50 Mb (11.4%)

1 23 repeat sequences and 21,533 gene models were identified in the genome. The major
2
3 24 findings of this study include the recent explosion of DNA/hAT-Charlie transposable
4
5
6 25 elements (TEs), the expansion of the P450 gene family and the constitution of the
7
8
9 26 cellular homeostasis system, which contributes to ecological plasticity in stress
10
11
12 27 adaptation. In addition, the high transcriptional levels of perivitellin genes in the
13
14
15 28 ovary and albumen gland promote the function of nutrient supply and defence ability
16
17
18 29 in eggs. Furthermore, the gut metagenome also contains diverse genes for food
19
20
21 30 digestion and xenobiotic degradation.

22
23 31 **Conclusions:** These findings collectively provide novel insight into the molecular
24
25
26 32 mechanisms of the ecological plasticity and high invasiveness.

27
28 33 **Keywords:** golden apple snail, *Pomacea canaliculata*, genome, adaptive evolution,
29
30
31 34 stress tolerance, P450, reproduction, perivitelline, metagenome

35 **Background**

36
37 36 The golden apple snail *Pomacea canaliculata* (family Ampullariidae, order
38
39
40 37 Architaenioglossa) is a fresh water snail listed among the world's top 100 worst
41
42
43 38 invasive species [1] and is considered an agricultural and quarantine pest worldwide
44
45
46 39 [2]. Native to tropical and subtropical South America, *P. canaliculata* gradually
47
48
49 40 spread to non-indigenous regions, such as Southeast and East Asia [3], Africa [4],
50
51
52 41 North America [5], Oceania [6] and even Europe [7]. Its successful
53
54
55 42 biological invasion was closely related to its polyphagous feeding habits [8],
56
57
58 43 voracious appetite [9], broad environmental adaptability [10] and rapid growth and
59
60
61

1 44 high rate of reproduction [11]. In addition to its ecological impact, *P. canaliculata*
2
3 45 ravages a wide range of crops, including grains, fruits and vegetables [12], causing
4
5
6 46 severe economic losses each year as a result of yield loss, replanting cost and
7
8
9 47 expenditures on control (<https://www.cabi.org/isc/datasheet/68490>). More seriously, *P.*
10
11 48 *canaliculata* has been involved in the transmission of a fatal human disease,
12
13
14 49 eosinophilic meningitis, that first appeared in East Asia where people frequently
15
16
17 50 consume these snails [13]. During this pathophoresis, *P. canaliculata* acts as an
18
19
20 51 important intermediate host of the pathogenic parasite *Angiostrongylus cantonensis*,
21
22
23 52 and the range of infected regions is still expanding, creating a great challenge in terms
24
25
26 53 of human health [14, 15].

27
28 54 Molluscs are a highly diverse group, second only to arthropods in species number [16],
29
30
31 55 and their high biodiversity makes them an excellent model to address issues such as
32
33
34 56 biogeography, adaptability and evolutionary processes [17]. The worldwide invasive
35
36
37 57 species *P. canaliculata* provides valuable potential in these fields [18]. As a primitive
38
39
40 58 circumtropical species, *P. canaliculata* possesses strong ecological plasticity with
41
42
43 59 many advantages, including low-temperature resistance [19] and drought tolerance
44
45
46 60 [20], which has contributed to its competitive success in resource acquisition. *P.*
47
48
49 61 *canaliculata* has been reported to establish populations at temperatures ranging from
50
51
52 62 10 °C to 35 °C [19, 21]. Additionally, *P. canaliculata* tolerates heavy metal
53
54
55 63 contamination. When living in contaminated water, the gill is enriched with a high
56
57
58 64 concentration of heavy metals, and histopathological changes in the digestive tract are
59
60
61 65 detected; however, an extremely low mortality rate is observed [22]. The conspicuous

1 66 colouration and neurotoxic lectin could confer a survival advantage on the eggs,
2
3 67 defending the embryos against potential predators [23]. Moreover, an
4
5
6 68 immune-neuroendocrine system can also be detected in *P. canaliculata*, as
7
8
9 69 demonstrated by the existence of a specific immune memory after bacterial challenge
10
11
12 70 [24, 25], broadening the study of invertebrate immunology.

13
14 71 The rich phenotypic and genetic diversity of molluscs makes them an excellent
15
16
17 72 species group for addressing many important issues in evolution, ecology and
18
19
20 73 function. However, the genomic resources on Mollusca are still insufficient compared
21
22
23 74 with those of other close phyla, such as Arthropoda and Nematoda, and few molluscs
24
25
26 75 can be employed as model organisms. *P. canaliculata*, however, possesses the
27
28
29 76 potential to be a model organism among molluscs because of several inherent
30
31
32 77 characteristics. For example, *P. canaliculata* is easy to acquire because it has a broad
33
34
35 78 global distribution originating from a primarily circumtropical environment.
36
37
38 79 Moreover, its high adaptability, rapid growth and efficient reproduction facilitate the
39
40
41 80 cultivation of *P. canaliculata* in the laboratory.

42 81 In recent years, the genomic features of *P. canaliculata* have been increasingly studied.
43
44
45 82 After the discovery of 14 pachytene bivalents in the karyotype [26], molecular
46
47
48 83 markers were identified to investigate the genetic diversity of the *P. canaliculata*
49
50
51 84 population, including 369 amplified fragment length polymorphism (AFLP) loci [27],
52
53
54 85 16,717 simple sequence repeats (SSR) [28, 29] and 15,412 single-nucleotide
55
56
57 86 polymorphisms SNPs [30]. In addition, multiple transcriptome analyses have been
58
59
60 87 performed to investigate the adaptation, invasion and immune mechanisms of *P.*

1 88 *canaliculata*. For instance, Sun et al. reported 128,436 unigenes based on a de novo
2
3 89 assembly of Illumina reads [30]; transcriptome changes in response to heat stress and
4
5
6 90 starving incubation were used to characterize its invasive and adaptive abilities [31,
7
8
9 91 32]; a transcriptome analysis comparing invasive *P. canaliculata* and indigenous
10
11 92 *Cipangopaludina cathayensis* provided insights into biological invasion [29]; and 402
12
13
14 93 immune-related differentially expressed genes (DEGs) in response to
15
16
17 94 lipopolysaccharide (LPyS) challenge were used to explore the mechanisms of defence
18
19
20 95 against pathogens [33]. Furthermore, proteomics tools such as isobaric tags for
21
22
23 96 relative and absolute quantitation (iTRAQ), and liquid chromatography-tandem mass
24
25
26 97 spectrometry (LC-MS/MS) were also applied in the study of protein expression
27
28
29 98 during estivation and oviposition [34,35], together providing plentiful omics- data for
30
31 99 the functional analysis of *P. canaliculata*.

32
33
34 100 However, research at the whole-genome level in *P. canaliculata* still lags far behind
35
36 101 that in other mollusc species due to the lack of a high-quality reference genome.
37
38
39 102 Multiple draft genomes of molluscs have been published, including the genomes of
40
41
42 103 the California sea hare [36], Pacific oyster [37], pearl oyster [38], owl limpet [39],
43
44
45 104 California two-spot octopus [40], golden mussel [41], and *Biomphalaria* snails [42],
46
47
48 105 greatly promoting research on mollusc genomics. In this study, we present a
49
50
51 106 chromosome-level genome assembly of *P. canaliculata* with high-quality gene
52
53 107 annotation, transcriptome data from several tissues and under various conditions, and
54
55
56 108 metagenomic data from the intestinal tracts, all of which were then applied to study
57
58
59 109 the species-specific environmental adaptation characteristics, such as the cellular

1 110 homeostasis system underlying strong stress and the colour and nutrient contents of
2
3 111 the eggs. Our data will not only strengthen the understanding of the evolutionary
4
5
6 112 mechanisms of molluscs and the molecular basis of biological invasion but also foster
7
8
9 113 the development of approaches to control the invasion of *P. canaliculata* and provide
10
11
12 114 a basis for interrupting the transmission of pathogenetic nematode parasites.
13
14
15

16 115 **RESULTS**

17 18 19 20 116 **Complete genome assembly at the chromosome level**

21
22
23
24 117 We generated 26.6 Gb (60.1 X) of PacBio SMRT raw reads with an average read
25
26
27 118 length of 10.1 kb, and 291 Gb (652.4 X) of Illumina HiSeq paired-end reads with an
28
29
30 119 average read length of 150-250 bp using DNA extracted from a single adult *P.*
31
32 120 *canaliculata* (Table S1). The 24.4 Gb (55.4 X) of clean PacBio SMRT reads that
33
34
35 121 passed quality filtering were assembled by smartdenovo
36
37
38 122 (<https://github.com/ruanjue/smartdenovo>), resulting in an assembly of 1,234 raw
39
40
41 123 contigs with a total length of 473.6 Mb and an N50 length of 1.0 Mb. After filtering of
42
43
44 124 alternatively heterozygous contigs, the 745 resulting contigs with a total length of
45
46
47 125 440.1 Mb and an N50 length of 1.1 Mb were taken as the final contigs. Previous
48
49
50 126 karyotype research has shown that the haploid *P. canaliculata* genome consists of 14
51
52 127 chromosomes [26]. Based on the Hi-C data, 439.5 Mb (99.9%) of final contigs were
53
54
55 128 anchored and oriented into 14 large scaffolds, each corresponding to a natural
56
57
58 129 chromosome (Figure 1a and Figure 1b), with the longest 45.4 Mb and the shortest
59
60 130 27.2 Mb. This assembly quality is much better than that of the other molluscan
61
62
63
64
65

1 131 genomes published thus far (Table 1). In addition to the length and continuity of the
2
3 132 assembled sequences, another important aspect for evaluating genome assembly is the
4
5
6 133 ratio of genome coverage. With an estimated genome size of 446 Mb and genome
7
8
9 134 heterozygosity between 1% and 2% based on the distribution of k-mer frequency [43]
10
11 135 (Figure S1), ~98.6 % of the *P. canaliculata* genome has been assembled. To further
12
13
14 136 confirm the accuracy and completeness of the assembly, we mapped the Illumina
15
16
17 137 shotgun reads to the assembled reference genome. Significantly, 97% and 95% of the
18
19
20 138 genome-derived and transcriptome-derived reads, respectively, could be aligned to the
21
22
23 139 reference genome, suggesting no obvious bias in sequencing and assembly.
24
25 140 Additionally, the mitochondrial genome of *P. canaliculata* was assembled as a single
26
27
28 141 contig 15,707 bp in length, which has 99.9% sequence identity to the published
29
30
31 142 mitochondrial genome (GenBank: KJ739609.1) (Figure S2). This high-quality
32
33
34 143 reference genome provides a good foundation for gene annotation.
35
36 144 The protein-coding genes were predicted on the reference genome by EVM,
37
38
39 145 integrating evidence from *de novo* prediction, transcriptome and homology data. In
40
41
42 146 total, 21,533 gene models were predicted as the reference gene set, with coding
43
44
45 147 regions spanning ~32.2 Mb (7.3 %) of the genome (Table 1 and Table S2). The
46
47
48 148 distribution of CDS length in *P. canaliculata* is similar to that in closely related
49
50
51 149 species (Figure 1c). Overall, 97.5% of the reference genes were supported by
52
53
54 150 transcriptome data, and 98.0% of eukaryote core genes from OrthoDB
55
56 151 (<http://www.orthodb.org/>) were identified in the reference gene set by BUSCO. These
57
58
59 152 results were comparable to those in other published molluscan genomes (Table 1). In
60
61
62
63
64
65

1 153 functional annotation, a total of 19,815 (91.9 %) reference genes were annotated by at
2
3
4 154 least one functional database. Specifically, 15,662 (72.7%), 13,769 (63.4%), 17,081
5
6 155 (79.3%), 18,847 (87.5%) and 17,003 (79.9%) reference genes were annotated with the
7
8
9 156 eggNOG, KEGG, NR, InterPro and UniProt databases, respectively (Figure S3).

10 11 12 157 **Signs of adaptive evolution in *P. canaliculata* genome**

13
14
15
16 158 To gain insight into the evolutionary perspective of *P. canaliculata*, a phylogenetic
17
18
19 159 tree was built based on 306 high-confidence single-copy orthologous genes from nine
20
21
22 160 related species (*P. canaliculata*, *Lottia gigantea*, *Aplysia californica*, *Biomphalaria*
23
24
25 161 *glabrata*, *Crassostrea gigas*, *Octopus bimaculoides*, *Pintada fucata*, *Lingula anatina*
26
27
28 162 and *Limnoperna fortunei*) by PhyML [44] and the divergence time was estimated
29
30
31 163 using MCMCTree [45]. The results show that *P. canaliculata* diverged from the
32
33
34 164 ancestor of *B. glabrata* and *A. californica* 372 million years ago (Mya) and from *L.*
35
36 165 *gigantea* 491 Mya (Figure 2a).

37
38 166 Then, the molluscan orthologous genes were investigated for adaptive evolution.
39
40
41 167 Utilizing pairwise protein sequence similarities, gene family clustering was conducted
42
43
44 168 by orthfinder [46]. A total of 239,541 reference genes from the nine species were
45
46
47 169 clustered into 69,582 orthologous groups, among which 14,766 orthologous groups
48
49
50 170 contained at least two genes each. We identified 66 orthologue groups that underwent
51
52
53 171 common expansion in both *P. canaliculata* and *L. fortunei* but not the other seven
54
55
56 172 species. The functions of these orthologous groups are mainly related to signal
57
58 173 transduction; replication and repair; translation, glycan biosynthesis and metabolism;

1 174 lipid metabolism; and the endocrine, immune and nervous systems (Figure S4). These
2
3 175 relations suggests that the gene families that underwent expansion may play important
4
5
6 176 roles in adaptation to the environment as invasive species.
7
8
9 177 The high-coverage genome assembly enables a comprehensive analysis of the
10
11 178 transposable elements (TEs), which play multiple roles in driving genome evolution
12
13
14 179 in eukaryotes [47]. In total, we identified 49.6 Mb TE sequences in the assembled *P.*
15
16
17 180 *canaliculata* genome (Table 1), including 3.4 Mb long terminal repeats (LTRs), 27.2
18
19
20 181 Mb long interspersed elements (LINEs), 17.5 Mb DNA transposons and 1.5 Mb short
21
22
23 182 interspersed elements (SINEs). Next, we analysed the divergence rate of each class of
24
25
26 183 TEs among the available sequenced mollusc genomes. Notably, the TE class of DNA
27
28
29 184 transposons showed a specific peak at a divergence rate of ~4% divergence rate for *P.*
30
31 185 *canaliculata* and *C. gigas* (Figure 2b), indicating a recent explosion of DNA
32
33
34 186 transposons in these two species. We analysed the expression of 709 genes, including
35
36
37 187 DNA elements restricted to the 4% peak inside the gene region, compared with that of
38
39
40 188 the other genes outside the 4% peak (Figure S5). DEGs were defined here by P-values
41
42
43 189 smaller than 0.05 for comparison of the treatment (heat, cold, heavy metal and air
44
45
46 190 exposure) and control data. The percentages of DEGs in the 4% peak were higher than
47
48
49 191 those of genes outside the peak (10.2% higher for heat, 8.6% higher for cold, 8.6%
50
51
52 192 higher for heavy metal, and 7.3% higher for air exposure). Among the DEGs in the 4%
53
54
55 193 peak, approximately half were up-regulated, and the other half were down-regulated.
56
57
58 194 Moreover, the DEGs in the 4% peak were mainly enriched in cellular metabolic
59
60
61 195 process, response to stimulus, localization and signaling according to GO annotation.

1 196 These results indicated that genes in the 4% peak were likely to be more active in the
2
3 197 response to stimulus, promoting potential plasticity in stress adaptation. TEs are
4
5
6 198 powerful facilitators of evolution that generate “evolutionary potential” to introduce
7
8
9 199 small adaptive changes within a lineage, and the importance of TEs in stress
10
11
12 200 responses and adaptation has been reported in numerous studies [48,49]. The recent
13
14
15 201 explosion of DNA TEs in *P. canaliculata* could also play an important role in
16
17
18 202 promoting the potential plasticity in stress adaptation.

21 203 **Investigation of cellular homeostasis system underlying strong stress adaptation**

22
23
24 204 The homeostasis system plays a crucial role in stress adaptability, providing the
25
26
27 205 molecular basis for re-establishing dynamic equilibrium after challenges by various
28
29
30 206 environmental stressors, including temperature, air exposure, anthropogenic pollution
31
32
33 207 and pathogens [50]. In this study, we addressed three constituent parts of the cellular
34
35
36 208 homeostasis system, which contributes to the successful ecological plasticity of *P.*
37
38
39 209 *canaliculata* (Figure 3). The transcriptomes of the hemocytes after different stimuli
40
41
42 210 (cold, heat, heavy metal and air exposure) were also sequenced and analysed to
43
44
45 211 address the potential roles of these genes in the cellular homeostasis system.

46
47 212 The unfolded protein response (UPR) system is the central component of protein
48
49
50 213 homeostasis [51]. Heat shock proteins (HSPs) act as molecular chaperones to
51
52
53 214 maintain correct folding, and heat shock transcription factor 1 (HSF1) is responsible
54
55
56 215 for the transcriptional induction of HSPs [52]. In the *P. canaliculata* genome, 13
57
58
59 216 HSP70s, 6 HSP90s, 7 HSP40s and 11 HSFs were identified (Table S3), and the

1 217 expression of HSP90s and HSFs was highly induced in response to heat, cold, heavy
2
3 218 metal and air exposure (Table S4 and Figure S6). Inositol-requiring protein 1 (IRE1),
4
5
6 219 protein kinase RNA-like ER kinase (PERK), and activating transcription factor 6
7
8
9 220 (ATF6) are three mediators recruited by the endoplasmic reticulum (ER) to regulate
10
11 221 the UPR [53]. We found putative coding genes of the three core mediators, their
12
13
14 222 respective downstream transcription factors, and the corresponding recognition
15
16
17 223 chaperones in the *P. canaliculata* genome (Table S3).

18
19
20 224 The xenobiotic biotransformation system helps the molluscs adapt to toxicants,
21
22
23 225 especially pesticides in aquatic environments [54]. Manual annotation of this genome
24
25
26 226 identified 157 cytochrome P450s (CYP450s), 15 flavin-containing monooxygenases
27
28
29 227 (FMOs), 53 glutathione S-transferases (GSTs) and 105 ATP binding cassette (ABC)
30
31 228 transporters, most of which showed up-regulated expression under stress (Table S3
32
33
34 229 and Table S4). These proteins have been shown to function in contaminant detection,
35
36
37 230 conjugative modification and expulsion for xenobiotic detoxification [55-57].

38
39 231 The massive production of reactive oxygen species (ROS) and reactive oxygen
40
41
42 232 intermediates (ROIs) induced by stress leads to many pathological conditions, and
43
44
45 233 antioxidant systems protect the organism from superoxide [58]. Four main antioxidant
46
47
48 234 enzyme classes, namely, superoxide dismutase (SOD), catalase (CAT), peroxidase
49
50
51 235 (Prx), and glutathione peroxidase (GPX), were found in *P. canaliculata* and showed
52
53
54 236 elevated global expression in response to stress (Table S3 and Table S4).

55
56 237 Apoptosis is a process of cell death when sensing stress and the regulation of
57
58
59 238 apoptosis maintains the dynamic homeostasis of the internal environment. In *P.*

1 239 *canaliculata*, we propose the existence of both intrinsic and extrinsic apoptotic
2
3 240 signaling pathways, evidenced by the presence of homologous genes involved in both
4
5
6 241 pathways. These two pathways could be activated by cytochrome C and tumour
7
8
9 242 necrosis factor receptor (TNFR), respectively (Table S3). Inhibitors of apoptosis, such
10
11
12 243 as XIAP, Bcl2 and Bak, are also detected and show increased expression in response
13
14
15 244 to stress (Table S4), which is expected to delay the process of apoptosis and cell death
16
17
18 245 in the stress response.

21 246 **The expansion of the P450 gene family contribute to stress tolerance**

22
23
24 247 Cytochrome P450 (CYP) enzymes are a monooxygenase family with highly diverse
25
26
27 248 structures and functions that have been widely identified in all kingdoms of life [59].
28
29
30 249 P450s catalyse the reductive scission of molecular oxygen and are responsible for the
31
32
33 250 synthesis and metabolism of various molecules, including drugs, hormones,
34
35
36 251 antibiotics, pesticides, carcinogens and toxins [60]. The hormones they synthesize,
37
38
39 252 such as glucocorticoids, mineralocorticoids, progestins, and sex hormones, are critical
40
41
42 253 to stress response, growth and reproduction, and the endogenous and exogenous
43
44
45 254 chemical metabolism participate in combatting toxic compounds [61].

46
47 255 We found that the *P. canaliculata* CYP gene family had undergone an expansion
48
49
50 256 compared to that in the other molluscs. We identified 157 genes in the genome of *P.*
51
52
53 257 *canaliculata* and 128, 102, 135, 78, 52 and 94 genes in *A. californica*, *B. glabrata*, *C.*
54
55
56 258 *gigas*, *L. gigantea*, *O. bimaculoides* and *P. fucata* respectively, using the same standard
57
58
59 259 (Figure 4a). An expansive trend was also observed in comparison with other model

1 260 species, such as *Homo sapiens* (57), *Mus musculus* (102), *Danio rerio* (94) and
2
3
4 261 *Drosophila melanogaster* (94) [62]. Gene expansion was mainly found in the CYP2U
5
6 262 and CYP3A sub-families, whereas fewer genes were expanded in CYP4F. In
7
8
9 263 mammals, CYP2U participates in the metabolism of fatty acids to generate bioactive
10
11
12 264 eicosanoid derivatives, potentially regulating the development of immune function
13
14
15 265 [63]. In *P. canaliculata*, 40 genes formed the CYP2U clade, mainly expressed in the
16
17 266 hepatopancreas (Figure 4b and Table S5_a, Table S5_b). CYP3A is a versatile enzyme
18
19
20 267 that metabolizes a wide range of xenobiotics, and its production promotes the growth
21
22
23 268 of various cell types [64]. The 56 CYP3A genes are comprehensively expressed in the
24
25
26 269 hepatopancreas, gill and kidney (Figure 4b and Table S5_a, Table S5_b). CYP4F
27
28
29 270 possesses epoxygenase activity, metabolizing fatty acids to epoxides to suppress
30
31
32 271 hypertension, pain perception and inflammation [65]. Twenty genes were identified in
33
34 272 CYP4F, and Pc06G011748, Pc06G011460, Pc06G011458, Pc06G011459,
35
36 273 Pc04G006708, Pc04G006710 and Pc04G006707 exhibited highly induced expression
37
38
39 274 levels under cold, heat, heavy metal and air exposure stress, indicating their critical
40
41
42 275 roles in the stress tolerance (Figure 4b, Table S5_a and Table S5_b).

43
44
45 276 **The identification of perivitellin genes and their high transcriptional levels in the**
46
47 277 **ovary and albumen gland**

48
49
50
51
52 278 *P. canaliculata* has eggs characterized by abundant nutrients, reddish or pinkish
53
54
55 279 colour, aerial oviposition and neurotoxicity [23, 66] due to the perivitelline Fluid
56
57
58 280 (PVF), which fills the space between the eggshell and the embryo and consists of
59
60 281 carbohydrates, lipids and proteins (Figure 5a). The PVF proteins in *P. canaliculata*,

1 282 include three major components, PcOvo, PcPV2, and PcPV3 [67], collectively named
2
3 283 perivitellins, which make up 90% of the total proteins, whereas most of the other
4
5
6 284 dozens of low-abundance components each account for less than 1% of the total
7
8
9 285 proteins [35]. The perivitellins are not only responsible for the major supply of
10
11
12 286 materials and energy during embryogenesis but also provide warning pigments and
13
14
15 287 deadly toxicants against predators [23, 68, 69].

16
17 288 We identified 28 candidate PVF genes in *P. canaliculata* by mapping each of the 59
18
19
20 289 fragmental PVF protein sequences derived from a previous proteomics study by Sun
21
22
23 290 [35] to its best hit in the reference gene set of *P. canaliculata*, using BLASTP with
24
25
26 291 requirements of over 85% identity and at least 50% alignment length (Table S6). Then,
27
28
29 292 the functional annotation of those fragmental proteins was also transferred to our
30
31
32 293 identified PVF genes. The transcriptome data show that 22 (79%) of the 28 candidate
33
34
35 294 PVF genes exhibit their highest expression in the ovary and albumen gland (PVF
36
37
38 295 protein synthesis factory) among all 7 tissues (Figure 5b and Table S7), confirming
39
40
41 296 that most of them are genuine functional PVF genes. Six of these 28 candidate PVF
42
43
44 297 genes are perivitellin genes, including two PcOvo genes, Pc09G015543 (PcOvo2) and
45
46
47 298 Pc09G015548 (PcOvo3); two PcPV2 genes, Pc07G012572 (PcPV2-31) and
48
49
50 299 Pc07G012571 (PcPV2-67); and two possible PcPV3 genes, Pc09G015546 and
51
52
53 300 Pc09G015547. The expression levels of these 6 genes in the ovary and albumen gland
54
55
56 301 are much higher than those of the other 22 candidate PVF genes.

57
58
59 302 By analysing the orthoFinder gene families that include orthologous and paralogous
60
61
62 303 genes from *P. canaliculata* and 8 other sequenced mollusc species, we found that

1 304 these 28 candidate PVF genes were classified into 20 multiple-gene families (≥ 2
2
3 305 genes) and 7 single-gene families (only one gene) (Table S8). Notably, 5 of the 6
4
5
6 306 perivitellin genes were classified into single-gene families, except for Pc07G012571
7
8
9 307 (PcPV2-67), which not only has homologous genes in other mollusc species but also
10
11 308 has three paralogous genes in *P. canaliculata* itself. However, none of these three
12
13 309 PcPV2-67 paralogous genes in *P. canaliculata* showed higher expression in the ovary
14
15 310 and albumen gland than in other tissues, indicating that they are likely not
16
17 311 PVF-related genes, i.e., only Pc07G012571 plays a role in PVF. The nearly unique
18
19 312 and single-copy nature of the 6 perivitellin genes in *P. canaliculata*, may be explained
20
21 313 by the long evolutionary distance, over 200 Mya for *P. canaliculata* and its most
22
23 314 closely related species, *A. californica*, as well as numerous differences in their living
24
25 315 characteristics and egg structures. Another possible explanation is that these 6 major
26
27 316 PVF genes may have experienced rapid evolution in their history to adapt to the
28
29 317 changing environment.
30
31
32
33
34
35
36
37
38
39

40 318 **The gut microbiome plays important roles in stress resistance and food digestion**

41
42
43 319 The gut microbiome is well known as the second genome of animals and plays
44
45 320 important roles in food digestion, immune defence, and other processes that are
46
47 321 essential to the animal host. To investigate whether the gut microbiome influences the
48
49 322 invasive lifestyle, we collected gut digesta samples from 70 *P. canaliculata* snails and
50
51 323 generated 31 Gb of high-quality metagenomic data on the Illumina HiSeqX10
52
53 324 platform. To our knowledge, this study is the first in-depth sequencing of the snail gut
54
55
56
57
58
59
60
61
62
63
64
65

1 325 microbiome. A total of 1,142,095 non-redundant genes were obtained with an average
2
3 326 open reading frame (ORF) length of 604 bp (Table S9). The taxonomic composition
4
5
6 327 analysis showed that, at the phylum level, Proteobacteria was predominant, followed
7
8
9 328 by Verrucomicrobia, Bacteroidetes, Firmicutes, Spirochaetes, Actinobacteria, etc.
10
11 329 (Table S10_a). At the genus level, the most abundant genera included *Aeromonas*,
12
13 330 *Enterobacter*, *Desulfovibrio*, *Citrobacter*, *Comamonas*, *Klebsiella* and *Pseudomonas*
14
15 331 (Table S10_b), most of which were also present in *Achatina fulica* [70,71].
16
17 332 Interestingly, some of the most abundant genera, such as *Desulfovibrio*, *Citrobacter*
18
19 333 and *Pseudomonas*, were reported as having strong abilities to remove heavy metals by
20
21 334 bioprecipitation and bioabsorption [72-74]. For example, the sulfur-reducing bacteria
22
23 335 *Desulfovibrio* produces H₂S, which precipitates metals and therefore reduces the toxic
24
25 336 effects of dissolved metals [72]. Based on the KEGG pathway database, the complete
26
27 337 sulfate reduction metabolism pathway was identified in the *P. canaliculata* gut
28
29 338 microbiome. We suggested that these gut microbes might help *P. canaliculata* survive
30
31 339 the environmental stress of heavy metals in harsh conditions. In addition, a large
32
33 340 number of genes in xenobiotic biodegradation and metabolism pathways were
34
35 341 annotated, corresponding to 288 KEGG orthologous groups (KOs) and 21 pathways
36
37 342 (Table S11). As many of the pathways, such as benzoate degradation, toluene
38
39 343 degradation, xylene degradation and steroid degradation, could not be identified in the
40
41 344 host genome through KO analysis, we suggested that microbial detoxification abilities
42
43 345 may contribute to the ability *P. canaliculata* to resist stresses caused by xenobiotics
44
45 346 such as pesticides and environmental pollutants.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 347 In digestion, the gut microbes are directly involved in the breakdown of the cellulose
2
3 348 portion of the diet, and previous studies have isolated cellulolytic bacteria and
4
5
6 349 evaluated the cellulolytic enzyme activities [75]. Our work found a broader range of
7
8
9 350 carbohydrate active enzymes (CAZymes). Of the 208 annotated CAZyme families, 99
10
11
12 351 were glycoside hydrolase (GH) families (Table S12). Enzymes that could be classified
13
14 352 as cellulases, endohemicelluloses, debranching enzymes, and
15
16
17 353 oligosaccharide-degrading enzymes were all identified. These findings indicate that
18
19
20 354 the gut microbiome provides assistance in digesting a broad range of food sources,
21
22
23 355 enabling *P. canaliculata* to grow rapidly and adapt to an invasive lifestyle.
24
25
26

27 356 **Conclusion and discussion**

28
29
30

31
32 357 Given its environmental invasiveness, broad stress adaptability and rapid reproduction,
33
34 358 the golden apple snail *P. canaliculata* has received a vast amount of attention
35
36
37 359 worldwide. However, the underlying genetic mechanisms of these properties have not
38
39
40 360 been comprehensively uncovered. The chromosome-level genome of *P. canaliculata*
41
42
43 361 presented in this study sheds the first light on into the genomic basis of its ecological
44
45
46 362 plasticity in response to various stressors. The major findings of this study include the
47
48
49 363 recent explosion of DNA/hAT-Charlie TEs, the expansion of the P450 gene family
50
51
52 364 and the constitution of the cellular homeostasis system, all of which contribute to the
53
54
55 365 plasticity of the organism in stress adaptation. Although the function of the recently
56
57
58 366 originated TEs could not be confirmed, TEs are considered powerful facilitators in
59
60
61 367 adaptive evolution, suggesting that their increased number plays an important role in
62
63
64
65

1 368 the stress resistance of *P. canaliculata*. The UPR system, xenobiotic biotransformation
2
3 369 system and ROS system are all major components of the cellular homeostasis system,
4
5
6 370 and the P450s in particular underwent expansion with specific functions. In addition,
7
8
9 371 exclusive perivitellin genes were identified in the *P. canaliculata* genome, and they
10
11
12 372 are believed to contribute to the high reproductive rate and the expansion of habitats.
13
14 373 Furthermore, the gut metagenome contains diverse genes for food digestion and
15
16
17 374 xenobiotic degradation. These findings collectively provide novel insight into the
18
19
20 375 molecular mechanisms of ecological plasticity and high invasiveness.
21
22
23 376 In this study, we report a fine reference genome of *P. canaliculata*, first
24
25
26 377 chromosome-level Mollusca genome published. With its easy acquisition, rapid
27
28
29 378 growth and efficient reproduction, *P. canaliculata* possesses the potential to be a
30
31
32 379 model organism of Mollusca. As its cellular complexity and conservation of pathways
33
34
35 380 also make *P. canaliculata* a useful representative of Mollusca, the genome described
36
37
38 381 in this study can be used to advance our understanding of the molecular mechanisms
39
40
41 382 involved in various scientific questions regarding Mollusca.

42 43 44 383 **Methods**

45 46 47 48 384 **Samples collection and sequencing**

49
50
51
52 385 Adults of *P. canaliculata* were collected from a local paddy field in Shenzhen,
53
54
55 386 Guangdong province, China, and maintained in aerated freshwater at 15 ± 2 °C for a
56
57
58 387 week before processing. Genomic DNA was extracted from the foot muscles of a
59
60
61 388 single *P. canaliculata* for constructing PCR free Illumina 350-bp insert libraries and

1 389 PacBio 20-kb insert library, and sequenced on Illumina HiSeq 2500 and PacBio
2
3 390 SMRT platforms, respectively. The Hi-C library was prepared using the muscle tissue
4
5
6 391 of another single *P. canaliculata* by following methods: Nuclear DNA was
7
8
9 392 cross-linked in situ, extracted, and then digested with a restriction enzyme. The sticky
10
11
12 393 ends of the digested fragments were biotinylated, diluted, and then ligated to each
13
14
15 394 other randomly. Biotinylated DNA fragments were enriched and sheared again for
16
17
18 395 preparing the sequencing library, which was then sequenced on a HiSeq X Ten
19
20 396 platform (Illumina).

21
22 397 Seven tissues including embryos (2 days post fertilization), gill, hemocytes,
23
24
25 398 hepatopancreas, kidney, ovary and albumen gland and testis from six animals were
26
27
28 399 collected as parallel samples. Next, animals were cultivated in 37 °C and 10 °C for 24
29
30
31 400 hours heat and cold tolerance, in Cr³⁺(2mg L⁻¹), Cu²⁺(0.2mg L⁻¹) and Pb²⁺(1mg L⁻¹)
32
33
34 401 for 24 hours heavy metal tolerance, and in waterless tank for 7 days air exposure.

35
36 402 Then the hemocytes were harvested and stored, with three replicates for each group.
37
38
39 403 In final, total RNAs were extracted from the stored tissues of *P. canaliculata*
40
41
42 404 materials, and then mRNAs were pulled out by beads with poly-T for constructing
43
44
45 405 cDNA libraries (insert 350-bp), and sequenced on an Illumina HiSeq 2500 sequencer.

46
47 406 The intestinal digesta from 70 adult snails of *P. canaliculata* were collected, pooled
48
49
50 407 into 6 samples and stored at -20 °C until microbial DNA was extracted. A
51
52
53 408 combination of cell lysis treatments was applied, including five freeze-thaw cycles
54
55
56 409 (alternating between 65 °C and liquid nitrogen for 5 min), repeated beads-beating in
57
58
59 410 ASL buffer (cat. no. 19082; Qiagen Inc.), and incubated at 95 °C for 15 min. DNA

1 411 was isolated following the protocol reported protocol [76]. Paired-end libraries of
2
3 412 metagenomic DNA were prepared with an insert size of 350 base pairs (bp) following
4
5
6 413 the manufacture's protocol (cat. no. E7645L; New England Biolabs). Sequencing was
7
8
9 414 performed on Illumina HiSeq X10.

10 11 12 415 **Genome assembly and annotation**

13
14
15
16 416 The Illumina raw reads were filtered by trimming the adapter sequence and
17
18
19 417 low-quality regions (https://github.com/fanagislab/common_use), resulting in clean
20
21
22 418 and high-quality reads with an average error rate < 0.001. For the PacBio raw data,
23
24
25 419 the short subreads (< 2 kb) and low-quality (error rate > 0.2) subreads were filtered
26
27
28 420 out, and only one representative subread was retained for each PacBio read. The clean
29
30 421 PacBio reads were assembled by the software smartdenovo
31
32 422 (<https://github.com/ruanjue/smartdenovo>), after which Illumina reads were aligned to
33
34
35 423 the contigs by BWA-MEM, and single base errors in the contigs were corrected by
36
37
38 424 Pilon (v1.16) with the parameters “-fix bases, -nonpf, -minqual 20”. The *P.*
39
40
41 425 *canaliculata* genome is highly heterozygous, as illustrated by the double peaks on the
42
43
44 426 distribution curve of k-mer frequency, and the current assembly algorithm tends to
45
46
47 427 collapse homozygous regions and report heterozygous regions in alternative contigs.
48
49
50 428 To obtain a haploid reference contigs, we employed a whole-genome alignment
51
52 429 (WGA) strategy with MUMmer v3.23 to recognize and selectively remove alternative
53
54
55 430 heterozygous contigs, which were characterized by shorter length (less than 200 kb)
56
57
58 431 and the ability of most regions (more than 50%) to be aligned to another larger contig
59
60
61
62
63
64
65

1 432 with confident identity (higher than 80%). Next, Hi-C sequencing data were aligned
2
3 433 to the haploid reference contigs by BWA-MEM, and then these contigs were clustered
4
5
6 434 into chromosomes with LACH-ESIS (<http://shendurelab.github.io/LACHESIS/>).
7
8
9 435 A de novo repeat library for *P. canaliculata* was constructed by RepeatModeler
10
11 436 (v1.0.4; <http://www.repeatmasker.org/RepeatModeler.html>). TEs in the *P. canaliculata*
12
13
14 437 genome were also identified by RepeatMasker (v4.0.6; <http://www.repeatmasker.org/>)
15
16
17 438 using both the Repbase library and the de novo library. Tandem repeats in the *P.*
18
19
20 439 *canaliculata* genome were predicted using Tandem Repeats Finder v4.07b [77]. The
21
22
23 440 divergence rates of TEs were calculated between the identified TE elements in the
24
25
26 441 genome and their consensus sequence at the TE family level.
27
28 442 The gene models in the *P. canaliculata* genome were predicted by EVIDENCE Modeler
29
30
31 443 v1.1.1 [78], integrating evidence from ab initio predictions, homology-based searches
32
33
34 444 and RNA-seq alignments. Then, these gene models were annotated by RNA-seq data,
35
36
37 445 UniProt database and InterProScan software [79]. Finally, the gene models were
38
39
40 446 retained if they had at least one piece of supporting evidence from the UniProt
41
42
43 447 database, InterProScan domain and RNA-seq data. Gene functional annotation was
44
45
46 448 performed by aligning the protein sequences to the NCBI NR, UniProt, COG and
47
48
49 449 KEGG databases with BLASTP v2.3.0+ under an E-value cutoff of 10^{-5} and choosing
50
51
52 450 the best hit. Pathway analysis and functional classification were conducted based on
53
54
55 451 the KEGG database [80]. InterProScan was used to assign preliminary GO terms,
56
57
58 452 Pfam domains and IPR domains to the gene models.
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

453 **Evolutionary analysis**

454 Orthologous and paralogous groups were assigned from seven species (*P.*
455 *canaliculata*, *Lottia giganta*, *Aplysia californica*, *Biomphalaria glabrata*, *Crassostrea*
456 *gigas*, *Octopus bimaculoides*, *Pintada fucata*, *Limnoperna fortunei* and *Lingula*
457 *anatina*) by OrthoFinder [46] with default parameters. Orthologous groups that
458 contained only one gene for each species were selected to construct the phylogenetic
459 tree. The protein sequences of each gene family were independently aligned by
460 muscle v3.8.31 [81] and then concatenated into one super-sequence. The phylogenetic
461 tree was constructed by maximum likelihood (ML) using PhyML v3.0 [44] with the
462 best-fit model (LG+I+G) estimated by ProtTest3 [82]. The Bayesian relaxed
463 molecular clock (BRMC) approach was adopted to estimate the neutral evolutionary
464 rate and species divergence time using the program MCMCTree, implemented in the
465 PAML v4.9 package [45]. The tree was calibrated with the following time frames to
466 constrain the age of the nodes between the species: minimum = 260 Ma and
467 maximum = 290 Ma for *P. fucata* and *C. gigas* [83]; minimum = 450 Ma and
468 maximum = 480 Ma for *A. californica* (or *B. glabrata*) and *L. giganta* [84]. The
469 calibration time (fossil record time) interval (550-610 Mya) of *O. bimaculoides* was
470 adopted from previous results [85].

471 **Transcriptome data analysis**

472 Transcriptome reads were trimmed with the same method for genomic reads
473 (https://github.com/fanagislab/common_use), and then mapped to the reference

1 474 genome of *P. canaliculata* using TopHat (v. 2.1.0) with default settings. The
2
3 475 expression level of each reference gene in terms of FPKM was computed by cufflinks
4
5
6 476 v2.2.1. A gene was considered to be expressed if its FPKM >0. Differential gene
7
8
9 477 expression analysis was conducted using cuffdiff v2.2.1.

10 11 12 478 **Metagenome data analysis**

13
14
15
16 479 Raw reads were cleaned to exclude adapter sequences, low-quality sequences, and
17
18
19 480 contaminated DNA. The adapter sequence was identified and trimmed from the reads
20
21
22 481 by an ungapped dynamic programming algorithm; the low-quality part (head or tail)
23
24
25 482 of the reads was trimmed off to ensure that the average error rate of the remaining
26
27
28 483 reads was lower than 0.001; the reads that were mapped to contaminated DNA by
29
30
31 484 BWA-MEM [86] were filtered out; and finally, shorter reads (length < 75 bp) and
32
33
34 485 unpaired reads were excluded to form a set of clean reads. The BWA database built
35
36
37 486 for cleaning contamination included genomes of 10 species: the *P. canaliculata*
38
39 487 genome, the *Brassica rapa* genome, the *Oryza sativa* genome, 2 *Angiostrongylus*
40
41 488 *cantonensis* genomes, the *Caenorhabditis elegans* genome, the *Schistosoma mansoni*
42
43
44 489 genome, the *Clonorchis sinensis* genome, the *Fasciola hepatica* genome, the *Danio*
45
46
47 490 *rerio* genome, and the *human hg38* genome.

48
49 491 The clean reads were assembled by metaSPAdes (v3.11.1) [87] in paired-end mode
50
51
52 492 for each sample. Then, gene prediction was performed on contigs longer than 500 bp
53
54
55 493 by Prodigal (v2.6.3) [88] with the parameter “-p meta”, and gene models with cds
56
57
58 494 length less than 102 bp were filtered out. A non-redundant (NR) gene set (539,344

1 495 genes) was constructed using the gene models predicted from each sample by
2
3 496 cd-hit-est (v4.6.6) [89] with the parameter “-c 0.95 -n 10 -G 0 -a S 0.9”, which adopts
4
5
6 497 a greedy incremental clustering algorithm and the criteria of identity > 95% and
7
8
9 498 overlap > 90% of the shorter genes. Then, the clean reads were mapped onto this NR
10
11
12 499 gene set by BWA-MEM with the criteria of alignment length \geq 50bp and identity >
13
14 500 95%. The unmapped reads from all samples were assembled together, and the genes
15
16
17 501 were predicted again. The newly predicted genes were combined with the previous
18
19
20 502 gene set by cd-hit-est to obtain a new NR gene set (1,147,339 genes). After the
21
22
23 503 taxonomic assignments to the new NR gene set, 5244 genes classified as Eukaryota
24
25
26 504 but not fungi were removed, and the final NR gene set (1,142,095 genes) was
27
28
29 505 obtained.

30
31 506 The taxonomic assignments of the final NR genes were made on the basis of
32
33
34 507 DIAMOND [90] protein alignment against the NCBI -NR database by CARMA3 [91].
35
36
37 508 Functional annotation was performed by aligning all the protein sequences to the
38
39
40 509 KEGG [92] database (release 79) using DIAMOND and taking the best hit with the
41
42
43 510 criteria of E-value < 1e-5. CAZymes were annotated with dbCAN (release 5.0) [93]
44
45
46 511 using HMMER (v3.0) hmmscan [94] by taking the best hit with an E-value < 1e-18
47
48
49 512 and coverage > 0.35.

50
51 513 The clean reads from each sample were aligned against the gene catalogue (1,142,095
52
53
54 514 genes) by BWA-MEM with the criteria of alignment length \geq 50bp and identity >
55
56
57 515 95%. Sequence-based gene abundance profiling was performed as previously
58
59
60 516 described [95]. The taxonomic profiles of the samples were calculated by summing

1 517 the gene abundance according to the taxonomic assignment result.
2
3
4

5 518 **Abbreviations**
6
7
8
9

10 519 *A. californica*, *Aplysia californica*; *B. glabrata*, *Biomphalaria glabrata*; *C. gigas*,

11
12 520 *Crassostrea gigas*; *O. bimaculoides*, *Octopus bimaculoides*; *L. anatina*, *Lingula*

13
14
15 521 *anatina*; *L. fortune*, *Limnoperna fortune*; *L. giganta*, *Lottia giganta*; *P. canaliculata*,

16
17
18 522 *Pomacea canaliculata*; *P. fucata*, *Pinctada fucata*; Hem, hemocyte; Te, testis; Ov,

19
20
21 523 ovary and albumen gland; Kn, kidney; GI, gill; Hp, hepatopancreas, Em, embryo;

22
23 524 SSR, simple sequence repeats; mya, million years ago; *BLAST*, *basic local*

24
25
26 525 *alignment search tool*; SNP, single nucleotide polymorphism; PVF, Pervitelline Fluid;

27
28
29 526 Ovo, ovorubin; AFLP, amplified fragment length polymorphism; DEGs, differentially

30
31
32 527 expressed genes; LPyS, Lipopolysaccharide; iTRAQ, Isobaric Tags For Relative,

33
34
35 528 Absolute Quantitation; LC-MS/MS, Liquid Chromatography-tandem Mass

36
37 529 Spectrometry; TEs, transposable elements; LTR, long terminal repeats; LINE, long

38
39
40 530 interspersed elements; SINE, short interspersed elements; UPR, Unfolded protein

41
42
43 531 response; HSPs, heat shock proteins; HSF1, heat shock transcription factor 1; PERK,

44
45
46 532 protein kinase RNA-like ER kinase; ATF6, activating transcription factor 6; ER,

47
48
49 533 endoplasmic reticulum; CYP450s, cytochrome P450s; FMOs, flavin-containing

50
51
52 534 monooxygenases; GSTs, glutathione S-transferases; ABC, ATP binding cassette; ROS,

53
54
55 535 reactive oxygen species; ROI, reactive oxygen intermediates; SOD, superoxide

56
57
58 536 dismutase; CAT, catalase; Prx, peroxidase; GPX, glutathione peroxidase; TNFR,

59
60
61 537 tumor necrosis factor receptor; NR, non-redundant genes; ORF, open reading frame;

1 538 Kos, orthologous groups; CAZymes, carbohydrate active enzymes; GH, Glycoside
2
3 539 Hydrolase.
4
5
6
7

8 **540 Availability of data and materials**
9

10
11
12 541 Tables S1 to S12 and Figures S1 to S6 are available in the supplementary information
13
14
15 542 file. The raw sequencing data has been deposited in DDBJ/EMBL/GenBank under
16
17
18 543 project accession PRJNA427478, SRR6425828 for genomic Illumina_PE125
19
20
21 544 sequencing data, SRR6425829 for genomic Illumina_PE150 sequencing data,
22
23
24 545 SRR6425827 for genomic PacBio sequencing data, SRR6429132~SRR6429164 for
25
26
27 546 transcriptome sequencing data, and SRR6472920~SRR6472925 for gut microbiome
28
29
30 547 data. All the analysis data have also been released for public use and can be freely
31
32 548 accessed at AGIS
33
34 549 ftpsite: ftp://ftp.agis.org.cn/~fanwei/Pomacea_canaliculata_Genome/ .
35
36
37
38

39 **550 Authors' contributions**
40
41
42

43 551 WF and WQ conceived the study and designed the experiments. CL and YZ
44
45
46 552 performed the genome sequencing and assembly, BL performed annotation and
47
48
49 553 evolutionary analysis. CL performed the stress tolerance analysis, YR performed the
50
51
52 554 reproduction analysis, YZ performed the metagenome analysis. HW, SL, FJ, LY
53
54
55 555 provide suggestions and help checking. WF, CL, BL, YR, YZ wrote the manuscript,
56
57
58 556 and GZ help revise the manuscript. All authors read and approved the final
59
60 557 manuscript.
61
62
63
64
65

1 558 **Competing interests**

2
3
4
5 559 The authors declare that they have no competing interests.
6
7
8
9

10 560 **Acknowledgements**

11
12
13
14 561 This project is supported by the National key research and development program of
15
16
17 562 China (2016YFC1200600), Shenzhen science and technology program
18
19
20 563 (JCYJ20150630165133395), Fund of Key Laboratory of Shenzhen
21
22
23 564 (ZDSYS20141118170111640), and The Agricultural Science and Technology
24
25
26 565 Innovation Program (ASTIP) of Chinese Academy of Agricultural Sciences(CAAS) &
27
28
29 566 Elite Youth Program of Chinese Academy of Agricultural Sciences. We thank
30
31
32 567 Fanghao Wan, Jue Ruan, Yutao Xiao for providing constructive suggestions to this
33
34 568 project.
35
36
37

38 569 **Legends of tables and figures**

39
40
41
42 570 **Tables**

43
44
45 571 **Table 1. Summary of assembly and annotation of mollusc genomes**

Genome feature	<i>P. canaliculata</i>	<i>L. gigantea</i>	<i>A. californica</i>	<i>B. glabrata</i>	<i>C. gigas</i>	<i>O. bimaculoides</i>
Assembled sequences (bp)	440,071,717	359,505,668	927,310,431	916,377,450	557,735,934	23,381,887,882
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

572 **Figures**

573 **Figure 1. The genome characteristics of *P. canaliculata*.** (a) Circos plot showing the
574 genomic features. Track 1: 14 linkage groups of the genome; Track 2: distribution of
575 transposon elements in chromosomes; Track 3: protein-coding genes located on
576 chromosomes; Track 4: distribution of GC contents. (b) A genome-wide contact
577 matrix from Hi-C data between each pair of the 14 chromosomes using a 100 kb
578 window size. The colour value indicates the base 2 logarithm of the number of valid
579 reads ($\log_2(\text{valid reads})$). (c) Distribution of CDS length in six closely related species.

580 **Figure 2. Evolutionary genomic analysis of *P. canaliculata*.** (a) Phylogenetic
581 placement of *P. canaliculata* within the dated tree of molluscs. The estimated
582 divergence time is shown at each branching point, and *P. canaliculata* is shown in red.
583 (b) Distribution of divergence rate for the class of DNA transposons in molluscs
584 genomes. The divergence rate was calculated by comparing all TE sequences
585 identified in the genome to the corresponding consensus sequence in each TE
586 subfamily. The red arrow indicates that *P. canaliculata* and *C. gigas* had a recent
587 explosion of TEs at a divergence rate of ~4%.

588 **Figure 3. The cellular homeostasis system in *P. canaliculata*.** The unfolded protein
589 response (UPR) system includes HSPs and HSF in the heat shock response and CNX,
590 NEF, GRP94, BIP, HSP40, ATF6, IRE1, PERK, COP2, XBP, ATF4, TRAM and
591 Derlin in the endoplasmic reticulum unfolded protein response (UPR-ERAD).
592 Apoptotic pathways include XIAPs, Bcl2, caspases, TNFR, and FADD. The
593 antioxidant systems include PRX, SOD, CAT and GPX. The xenobiotic
594 biotransformation system includes EPHX3, P450, FMO and ABC transporter. The
595 colours of the boxes for gene families represent the degree of upregulation

1 596 (FPKM-stimulus/FPKM-control) as an overall result of stress, including heat, cold,
2
3 597 heavy metal and air exposure. Pathways and genes were obtained based on KEGG
4
5
6 598 annotation.

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

10
11 600 Phylogenetic tree demonstrating orthologous and paralogous relationships of all P450
12
13 601 genes from 7 species including *P. canaliculata*, *A. californica*, *B. glabrata*, *C. gigas*, *L.*
14
15 602 *giganta*, *O. bimaculoides* and *P. fucata*. P450 genes from seven species were obtained
16
17 603 based on Pfam annotation (Interpro) with an E-value of 10^{-5} . Clades are labelled by
18
19
20
21 604 P450 subfamily names. The tree was constructed using the maximum likelihood
22
23
24
25 605 method in MEGA7, and the branch length scale indicates the average number of
26
27
28 606 residue substitutions per site. (b) Phylogenetic tree of P450 genes in *P. canaliculata*,
29
30
31 607 which is a subset of the phylogenetic tree for the 7 species, and their heat map of
32
33
34 608 expression (FPKM) in seven tissues (Hem, hemocyte; Te, testis; Ov, ovary and
35
36 609 albumen gland; Kn, kidney; Gl, gill; Hp, hepatopancreas; Em, embryo) and heat map
37
38
39 610 of induced expression (FPKM-stimulus/FPKM-control) under stress (Con: control;
40
41
42 611 heat; cold; Hm: heavy metal; Exp: air exposure).

45 **Figure 5. The composition and expression of the *P. canaliculata* perivitellins in**

46
47 613 **different tissues.** (a) Perivitelline fluid (PVF) lies under the eggshell and surrounds
48
49
50 614 the embryo. It contains carbohydrates, lipids, and proteins. The proteins are also
51
52
53 615 known as perivitellins and are classified into three categories, PcOvo, PcPV2, and
54
55
56 616 PcPV3. (b) The displayed expression value of PVF proteins is the base 10 logarithm
57
58
59 617 of FPKM (\log_{10} FPKM). The genes marked in red encode perivitellins. The 7 tissues

1 618 examined are abbreviated as follows: Hem, hemocyte; Te, testis; Ov, ovary and
2
3
4 619 albumen gland; Kn, kidney; Gl, gill; Hp, hepatopancreas; Em, embryo.
5
6
7

8 **620 References**
9

- 10
11
12 621 1. Lowe S, Browne M, Boudjelas S, de Poorter M. 100 of the World's Worst Invasive Alien
13
14
15 622 Species: A selection from the Global Invasive Species Database. Auckland, New Zealand:
16
17
18 623 World Conservation Union (IUCN); 2000.
19
20
21 624 2. Ranamukhaarachchi SL, Wickramasinghe S. Golden apple snails in the world:
22
23 625 introduction, impact, and control measures. Global advances in ecology and management
24
25
26 626 of golden apple snails. 2006:133-52.
27
28
29 627 3. Naylor R. Invasions in Agriculture: Assessing the Cost of the Golden Apple Snail in Asia.
30
31
32 628 Royal Swedish Academy of Sciences. 1996;25:443-8.
33
34
35 629 4. Berthold T. Vergleichende Anatomie, Phylogenie und historische Biogeographie der
36
37 630 Ampullariidae: (Mollusca, Gastropoda). 1991.
38
39
40 631 5. Howells RG, Burlakova LE, Karatayev AY, Marfurt RK, Burks RL. Native and
41
42
43 632 introduced Ampullariidae in North America: History, status, and ecology. 2006:73-112.
44
45
46 633 6. Halwart M, Bartley DM. International mechanisms for the control and responsible use of
47
48 634 alien species in aquatic ecosystems, with special reference to the golden apple snail. Los
49
50
51 635 Baños, Philippines: Philippine Rice Research Institute (PhilRice); 2006.
52
53
54 636 7. López MA, Altaba CR, Andree KB, López V. First invasion of the Apple snail *Pomacea*
55
56 637 *insularum* in Europe. *Tentacle*. 2010;18:26-8.
57
58
59 638 8. Estebenet AL, Martín PR. *Pomacea canaliculata* (Gastropoda: Ampullariidae): life-history
60
61
62
63
64
65

1 639 traits and their plasticity. *Biocell* 2002;26:83-9.

2

3 640 9. Lach L. The spread of the introduced freshwater apple snail *Pomacea canaliculata*

4

5

6 641 (Lamarck) (Gastropoda Ampullariidae) on Oahu, Hawaii. *Bishop Museum Occasional*

7

8

9 642 Papers. 1999;58:66-71.

10

11 643 10. Yusa Y, Sugiura N, Wada T. Predatory Potential of Freshwater Animals on an Invasive

12

13 644 Agricultural Pest, the Apple Snail *Pomacea canaliculata* (Gastropoda: Ampullariidae), in

14

15 645 Southern Japan. *Biol Invasions*. 2006;8:137-47.

16

17

18 646 11. Lach L, Britton DK, Rundell RJ, Cowie RH. Food Preference and Reproductive Plasticity

19

20 647 in an Invasive Freshwater Snail. *Biol Invasions*. 2000;2:279-88.

21

22

23 648 12. Mochida O. Spread of freshwater *Pomacea* snails (Pilidae, Mollusca) from Argentina to

24

25 649 Asia. *Micronesica*. 1991;3 51-62.

26

27

28 650 13. Shan L, Zhang Y, Steinmann P, Zhou X. Emerging Angiostrongyliasis in Mainland China.

29

30 651 *Emerg Infect Dis*. 2008;14:161-4.

31

32

33 652 14. Caldeira RL, Mendonca CL, Goveia CO, Lenzi HL, Graeff-TeixeiraC Lima WS, et al.

34

35 653 First record of molluscs naturally infected with *Angiostrongylus cantonensis* (Chen, 1935)

36

37 654 (Nematoda: Metastrongylidae) in Brazil. *Memórias do Instituto Oswaldo Cruz*.

38

39 655 2007;102:887-9.

40

41

42 656 15. McMichael AJ, Beaglehole R. The changing global context of public health. *Lancet*

43

44 657 (London, England). 2000;356:495-9.

45

46

47 658 16. Chapman A. Numbers of Living Species in Australia and the World. *Australian Biological*

48

49 659 *Resources Study*; 2009.

50

51

52 660 17. Lindberg DR, Ponder WF, Haszprunar G. The Mollusca: relationships and patterns from

53

54

55

56

57

58

59

60

61

62

63

64

65

1 661 their first half-billion years. Oxford University Press, Oxford; 2004.

2

3 662 18. Hayes KA, Cowie RH, Thiengo SC. A global phylogeny of apple snails: Gondwanan

4

5

6 663 origin, generic relationships, and the influence of outgroup choice (Caenogastropoda:

7

8

9 664 Ampullariidae). Biol J Linn Soc Lond. 2009;98:61-76.

10

11 665 19. Matsukura K, Tsumuki H, Izumi Y, Wada T. Physiological response to low temperature in

12

13

14 666 the freshwater apple snail, *Pomacea canaliculata* (Gastropoda: Ampullariidae). J Exp

15

16

17 667 Biol. 2009;212:2558-63.

18

19

20 668 20. Yusa Y, Wada T, Takahashi S. Effects of dormant duration, body size, self-burial and

21

22

23 669 water condition on the long-term survival of the apple snail, *Pomacea canaliculata*

24

25

26 670 (Gastropoda: Ampullariidae). Appl Entomol Zool. 2006;41:627-32.

27

28 671 21. Seuffert ME, Burela S, Martín PR. Influence of water temperature on the activity of the

29

30

31 672 freshwater snail *Pomacea canaliculata* (Caenogastropoda: Ampullariidae) at its

32

33

34 673 southernmost limit (Southern Pampas, Argentina). Journal of Thermal Biology. 2010;

35

36

37 674 35:77-84.

38

39 675 22. Kruatrachue M, Sumritdee C, Pokethitiyook P, Singhakaew S. Histopathological effects

40

41

42 676 of contaminated sediments on golden apple snail (*Pomacea canaliculata*, Lamarck 1822).

43

44

45 677 Bull Environ Contam Toxicol. 2011;86:610-4.

46

47

48 678 23. Dreon MS, Frassa MV, Ceolín M, Ituarte S, Qiu JW, Sun J, et al. Novel animal defenses

49

50

51 679 against predation: a snail egg neurotoxin combining lectin and pore-forming chains that

52

53

54 680 resembles plant defense and bacteria attack toxins. PLoS One. 2013;8:e63782.

55

56 681 doi:10.1371/journal.pone.0063782.

57

58

59 682 24. Ottaviani E, Caselgrandi E, Fontanili P, Franceschi C. Evolution, immune responses and

60

61

62

63

64

65

1 683 stress: studies on molluscan cells. Acta Biol Hung. 1992;43:293-8.
2
3 684 25. Ottaviani E, Accorsi A, Rigillo G, Malagoli D, Blom JM, Tascetta F. Epigenetic
4
5
6 685 modification in neurons of the mollusc *Pomacea canaliculata* after immune challenge.
7
8
9 686 Brain Res. 2013;1537:18-26.
10
11 687 26. Mercado Laczkó AC, Lopretto EC. Estudio cromosómico y cariotípico de *pomacea*
12
13 688 *canaliculata* (Lamarck, 1801) (Gastropoda, Ampullariidae). Revista del Museo Argentino
14
15 689 de Ciencias Naturales "Bernardino Rivadavia" Hidrobiología. 1998;8:15-20.
16
17
18
19 690 27. Xu J, Han X, Li N, Yu J, Qian C, Bao Z. Analysis of genetic diversity of three geographic
20
21 691 populations of *Pomacea canaliculata* by AFLP. Acta Ecol Sin. 2009;29:4119- 26.
22
23
24
25 692 28. Chen L, Xu H, Li H, Wu J, Ding H, Liu Y. Isolation and characterization of sixteen
26
27 693 polymorphic microsatellite loci in the golden apple snail *Pomacea canaliculata*. Int J Mol
28
29 694 Sci. 2011;12:5993-8.
30
31
32
33 695 29. Mu X, Hou G, Song H, Xu P, Luo D, Gu D, et al. Transcriptome analysis between
34
35 696 invasive *Pomacea canaliculata* and indigenous *Cipangopaludina cahayensis* reveals
36
37 697 genomic divergence and diagnostic microsatellite/SSR markers. BMC Genet. 2015;16:12.
38
39
40
41 698 30. Sun J, Wang M, Wang H, Zhang H, Zhang X, Thiyagarajan V, et al. De novo assembly of
42
43 699 the transcriptome of an invasive snail and its multiple ecological applications. Mol Ecol
44
45 700 Resour. 2012;12:1133-44.
46
47
48
49 701 31. Mu H, Sun J , Fang L, Luan T, Williams GA, Cheung SG, et al. Genetic Basis of
50
51 702 Differential Heat Resistance between Two Species of Congeneric Freshwater Snails:
52
53 703 Insights from Quantitative Proteomics and Base Substitution Rate Analysis. J Proteome
54
55 704 Res. 2015;14:4296-308.
56
57
58
59
60
61
62
63
64
65

- 1 705 32. Yang L, Cheng TY, Zhao FY. Comparative profiling of hepatopancreas transcriptomes in
2
3 706 satiated and starving *Pomacea canaliculata*. BMC Genet. 2017;18:18.
4
5
6 707 33. Xiong YM, Yan ZH, Zhang JE, Li HY. Analysis of albumen gland proteins suggests
7
8 708 survival strategies of developing embryos of *Pomacea canaliculata*. Molluscan Res.
9
10 709 2017:1-6.
11
12
13 710 34. Sun J, Mu H, Zhang H, Chandramouli KH, Qian PY, Wong CK, et al. Understanding the
14
15 711 regulation of estivation in a freshwater snail through iTRAQ-based comparative
16
17 712 proteomics. J Proteome res. 2013;12:5271-80.
18
19
20 713 35. Sun J, Zhang H, Wang H, Heras H, Dreon MS, Ituarte S, et al. First proteome of the egg
21
22 714 perivitelline fluid of a freshwater gastropod with aerial oviposition. J Proteome Res.
23
24 715 2012;11:4240-8.
25
26
27 716 36. Aplysia Genome Project. Broad Institute. Vertebrate Biology Group. 2009.
28
29 717 <https://www.broadinstitute.org/aplysia/aplysia-genome-project>
30
31
32 718 37. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress
33
34 719 adaptation and complexity of shell formation. Nature. 2012;490:49-54.
35
36
37 720 38. Du X, Fan G, Jiao Y, Zhang H, Guo X, Huang R, et al. The pearl oyster *Pinctada fucata*
38
39 721 *martensii* genome and multi-omic analyses provide insights into biomineralization.
40
41 722 Gigascience. 2017;6:1-12.
42
43
44 723 39. Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, et al.
45
46 724 Insights into bilaterian evolution from three spiralian genomes. Nature. 2013;493:526-31.
47
48
49 725 40. Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The
50
51 726 octopus genome and the evolution of cephalopod neural and morphological novelties.
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 727 Nature. 2015;524:220-4.
2
3 728 41. Uliano-Silva M, Dondero F, Dan Otto T, Costa I, Lima NCB, Americo JA, et al. A
4
5
6 729 hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel
7
8
9 730 *Limnoperna fortunei*. *Gigascience*. 2017. doi: 10.1093/gigascience/gix128
10
11 731 42. Adema CM, Hillier LW, Jones CS, Loker ES, Knight M, Minx P, et al. Corrigendum:
12
13 732 Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nat Commun*.
14
15 733 2017;8:16153.
16
17
18
19 734 43. Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics
20
21 735 by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology*
22
23 736 2013:arXiv:1308.2012 [q-bio.GN].
24
25
26
27
28 737 44. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms
29
30 738 and methods to estimate maximum-likelihood phylogenies: assessing the performance of
31
32 739 PhyML 3.0. *Syst Biol* 2010;59:307-21. doi:10.1093/sysbio/syq010.
33
34
35
36 740 45. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*.
37
38 741 2007;24:1586-91. doi:10.1093/molbev/msm088.
39
40
41
42 742 46. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
43
44 743 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol*.
45
46 744 2015;16:157. doi:10.1186/s13059-015-0721-2.
47
48
49
50 745 47. Feschotte C, Wessler SR. Mariner-like transposases are widespread and diverse in
51
52 746 flowering plants. *Proc Natl Acad Sci U S A* 2002;99:280-5.
53
54
55 747 48. Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. The struggle for life of the
56
57 748 genome's selfish architects. *Biol Direct*. 2011;6:19.
58
59
60
61
62
63
64
65

1 749 49. Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc
2
3 750 Natl Acad Sci U S A. 2011;108:10863-70.
4
5
6 751 50. Chrousos GP. Stress and disorders of the stress system. Nat Rev Endocrinol.
7
8 752 2009;5:374-81.
9
10
11 753 51. Vabulas RM, Raychaudhuri S, Hayer-Hartl M. Protein folding in the cytoplasm and the
12
13 754 heat shock response. Cold Spring Harbor perspectives in biology. 2010;2:a004390.
14
15
16 755 52. Chen B, Retzlaff M, Roos T, Frydman J. Cellular Strategies of Protein Quality Control.
17
18 756 Cold Spring Harbor Perspectives in Biology. 2011;3:a004374.
19
20
21 757 53. Korennykh A and Walter P. Structural basis of the unfolded protein response. Annu Rev
22
23 758 Cell Dev Biol. 2012;28:251-77.
24
25
26 759 54. Chambers JE and Yarbrough JD. Xenobiotic biotransformation systems in fishes. Comp
27
28 760 Biochem Physiol C. 1976;55:77-84.
29
30
31 761 55. Mello DF, de Oliveira ES, Vieira RC, Simoes E, Trevisan R, Dafre AL, et al. Cellular and
32
33 762 Transcriptional Responses of *Crassostrea gigas* Hemocytes Exposed in Vitro to
34
35 763 Brevetoxin (PbTx-2) Mar Drugs. 2012;10: 583-97.
36
37
38 764 56. Boutet I, Tanguy A, Moraga D. Characterisation and expression of four mRNA sequences
39
40 765 encoding glutathione S-transferases pi, mu, omega and sigma classes in the Pacific oyster
41
42 766 *Crassostrea gigas* exposed to hydrocarbons and pesticides. Mar Biol 2004;146:53-64.
43
44
45 767 57. Deeley RG, Westlake C, Cole SP. Transmembrane transport of endo- and xenobiotics by
46
47 768 mammalian ATP-binding cassette multidrug resistance proteins. Physiol Rev.
48
49 769 2006;86:849-99.
50
51
52 770 58. Liu C, Zhang T, Wang L, Wang M, Wang W, Jia Z, et al. The modulation of extracellular
53
54
55
56
57
58
59
60
61
62
63
64
65

1 771 superoxide dismutase in the specifically enhanced cellular immune response against
2
3 772 secondary challenge of *Vibrio splendidus* in Pacific oyster (*Crassostrea gigas*). *Dev*
4
5
6 773 *Comp Immunol.* 2016;63:163-70.
7
8
9 774 59. Lamb DC, Lei L, Warrilow AG, Lepesheva GI, Mullins JG, Waterman MR, et al. The first
10
11 775 virally encoded cytochrome p450. *J Virol.* 2009;83:8266-9.
12
13
14 776 60. Urlacher VB, Girhard M. Cytochrome P450 monooxygenases: an update on perspectives
15
16 777 for synthetic application. *Trends Biotechnol.* 2012;30:26-36.
17
18
19 778 61. Sanderson T, van den Berg M. Topic 3.1: Interactions of xenobiotics with the steroid
20
21 779 hormone biosynthesis pathway. *Pure Appl Chem.* 2003;75:1957-71.
22
23
24
25 780 62. Goldstone JV, McArthur AG, Kubota A, Zanette J, Parente T, Jönsson ME, et al.
26
27 781 Identification and developmental expression of the full complement of Cytochrome P450
28
29 782 genes in Zebrafish. *BMC Genomics.* 2010;11:643.
30
31
32
33 783 63. Chuang SS, Helvig C, Taimi M, Ramshaw HA, Collop AH, Amad M, et al. CYP2U1, a
34
35 784 novel human thymus- and brain-specific cytochrome P450, catalyzes omega- and
36
37 785 (omega-1)-hydroxylation of fatty acids. *J Biol Chem.* 2004;279:6305-14.
38
39
40
41 786 64. Fleming I. The pharmacology of the cytochrome P450 epoxygenase/soluble epoxide
42
43 787 hydrolase axis in the vasculature and cardiovascular disease. *Pharmacol Rev.*
44
45 788 2014;66:1106-40.
46
47
48
49 789 65. Zhang G, Kodani S, Hammock BD. Stabilized epoxygenated fatty acids regulate
50
51 790 inflammation, pain, angiogenesis and cancer. *Prog Lipid Res.* 2014;53:108-23.
52
53
54
55 791 66. de Jong-Brink M, Boer HH, Joosse J. Mollusca. In: Adiyodi, K.G., Adiyodi,
56
57 792 R.G. (Eds.), *Reproductive Biology of invertebrates. Oogenesis oviposition and*

1 793 oosorption, vol. 1. John Wiley & Sons Ltd., New York, 1983; pp. 297-355.

2

3 794 67. Garin CF, Heras H, Pollero RJ. Lipoproteins of the egg perivitelline fluid of *Pomacea*

4

5

6 795 *canaliculata* snails (Mollusca: Gastropoda). J Exp Zool. 1996;276:307-14.

7

8

9 796 68. Dreon MS, Schinella G, Heras H, Pollero RJ. Antioxidant defense system in the apple

10

11 797 snail eggs, the role of ovorubin. Arch Biochem Biophys. 2004;422:1-8.

12

13

14 798 69. Dreon MS, Ituarte S, Heras H. The role of the proteinase inhibitor ovorubin in apple snail

15

16 799 eggs resembles plant embryo defense against predation. PLoS One. 2010;5:e15059.

17

18 800 doi:10.1371/journal.pone.0015059.

19

20

21

22 801 70. Cardoso AM, Cavalcante JJV, Vieira RP, Lima JL, Grieco MAB, Clementino MM, et al.

23

24 802 Gut Bacterial Communities in the Giant Land Snail *Achatina fulica* and Their

25

26 803 Modification by Sugarcane-Based Diet. Plos One. 2012;7 doi:ARTN

27

28 804 e3344010.1371/journal.pone.0033440.

29

30

31 805 71. Cardoso AM, Cavalcante JJV, Cantão ME, Thompson CE, Flatschart RB, Glogauer A, et

32

33 806 al. Metagenomic Analysis of the Microbiota from the Crop of an Invasive Snail Reveals a

34

35 807 Rich Reservoir of Novel Genes. Plos One. 2012;7 doi:ARTN

36

37 808 e4850510.1371/journal.pone.0048505.

38

39

40 809 72. Cabrera G, Pérez R, Gómez JM, Ábalos A, Cantero D. Toxic effects of dissolved heavy

41

42 810 metals on *Desulfovibrio vulgaris* and *Desulfovibrio* sp strains. J Hazard Mater

43

44 811 2006;135:40-6. doi:10.1016/j.jhazmat.2005.11.058.

45

46

47 812 73. Finlay JA, Allan VJ, Conner A, Callow ME, Basnakova G, Macaskie LE. Phosphate

48

49 813 release and heavy metal accumulation by biofilm-immobilized and chemically-coupled

50

51 814 cells of a *Citrobacter* sp. pre-grown in continuous culture. Biotechnol Bioeng.

52

53

54

55

56

57

58

59

60

61

62

63

64

65

1 815 1999;63:87-97.
2
3 816 74. Valls M, de Lorenzo V, Gonzalez-Duarte R, Atrian S. Engineering outer-membrane
4
5
6 817 proteins in *Pseudomonas putida* for enhanced heavy-metal bioadsorption. J Inorg
7
8
9 818 Biochem. 2000;79:219-23.
10
11 819 75. Pinheiro GL, Correa RF, Cunha RS, Cardoso AM, Chaia C, Clementino MM, et al.
12
13
14 820 Isolation of aerobic cultivable cellulolytic bacteria from different regions of the
15
16
17 821 gastrointestinal tract of giant land snail *Achatina fulica*. Front Microbiol. 2015;6 doi:Artn
18
19
20 822 86010.3389/Fmicb.2015.00860.
21
22 823 76. Zoetendal EG, Heilig HG, Klaassens ES, Booiijink CC, Kleerebezem M, Smidt H, et al.
23
24
25 824 Isolation of DNA from bacterial samples of the human gastrointestinal tract. Nature
26
27
28 825 protocols 2006, 1(2): 870-873.
29
30
31 826 77. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids
32
33
34 827 Res. 1999;27:573-80.
35
36 828 78. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic
37
38
39 829 gene structure annotation using EVidenceModeler and the Program to Assemble Spliced
40
41
42 830 Alignments. Genome Biol. 2008;9:R7.
43
44 831 79. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan:
45
46
47 832 protein domains identifier. Nucleic Acids Res. 2005;33:W116-20.
48
49
50 833 80. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and
51
52
53 834 interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109-D14.
54
55 835 81. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
56
57
58 836 throughput. Nucleic Acids Res. 2004;32:1792-7.

1 837 82. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models
2
3 838 of protein evolution. *Bioinformatics*. 2011;27:1164-5. doi:10.1093/bioinformatics/btr088.
4
5
6 839 83. Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea
7
8 840 chemosynthetic environments as revealed by mussel genomes. *Nature Ecology &*
9
10
11 841 *Evolution*. 2017; 1: 121.
12
13
14 842 84. Benton MJ, Donoghue PCJ, Asher RJ. in *The Timetree of Life:Calibrating and*
15
16
17 843 *Constraining Molecular Clocks* (eds Hedges, S. B. & Kumar, S.)35–86 (Oxford Univ.
18
19
20 844 Press, 2009.
21
22
23 845 85. Zapata F, Wilson NG, Howison M, Andrade SC, Jörger KM, Schrödl M, et al.
24
25 846 Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. *Proc Biol*
26
27
28 847 *Sci*. 2014;281(1794):20141739. doi: 10.1098/rspb.2014.1739.
29
30
31 848 86. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
32
33
34 849 transform. *Bioinformatics*. 2009;25:1754-60.
35
36
37 850 87. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile
38
39 851 metagenomic assembler. *Genome Res*. 2017;27:824-34.
40
41
42 852 88. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site
43
44
45 853 prediction in metagenomic sequences. *Bioinformatics*. 2012;28:2223-30.
46
47
48 854 89. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
49
50 855 sequencing data. *Bioinformatics*. 2012;28:3150-2.
51
52
53 856 90. Buchfink B, Chao X, Huson DH. Fast and sensitive protein alignment using DIAMOND.
54
55
56 857 *Nat Methods*. 2015;12:59-60.
57
58
59 858 91. Gerlach W and Stoye J. Taxonomic classification of metagenomic shotgun sequences
60
61
62
63
64
65

1 859 with CARMA3. *Nucleic Acids Res.* 2011;39 doi:Artn E9110.1093/Nar/Gkr225.
2
3 860 92. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for
4
5
6 861 deciphering the genome. *Nucleic Acids Res.* 2004;32:D277-80.
7
8
9 862 93. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated
10
11 863 carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 2012;40:W445-51.
12
13
14 864 94. Eddy SR. Accelerated Profile HMM Searches. *Plos Comput Biol.* 2011;7 doi:ARTN
15
16 865 e100219510.1371/journal.pcbi.1002195.
17
18
19 866 95. Qin JJ, Li YR, Cai ZM, Li SH, Zhu JF, Zhang F, et al. A metagenome-wide association
20
21 867 study of gut microbiota in type 2 diabetes. *Nature.* 2012;490:55-60.
22
23
24 868
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65



1
2
3
4
5
6
7 1 **The genome of the golden apple snail *Pomacea canaliculata* provides insight into**
8
9 2 **stress tolerance and invasive adaptation**

10
11 3 ~~Conghui Liu^{1*}, Bo Liu^{1*}, Yuwei Ren^{1*}, Yan Zhang^{1*}, Hengchao Wang¹, Shuqu Li¹;~~

12
13 4 ~~Fan Jiang¹, Lijuan Yin¹, Guojie Zhang², Wanqiang Qian^{1†} and Wei Fan^{1†}~~

14
15 5 ~~Conghui Liu^{1*}, Yan Zhang^{1*}, Yuwei Ren^{1*}, Hengchao Wang¹, Shuqu Li¹, Fan Jiang¹, Lijuan Yin¹,~~

16
17 6 ~~Guojie Zhang², Wanqiang Qian^{1†}, Bo Liu^{1†}, Wei Fan^{1†}~~

18
19 7 ¹Agricultural Genomic Institute, Chinese Academy of Agricultural Sciences,
20
21 Shenzhen, Guangdong, 518120, China.

22
23 8 ²BGI-Shenzhen, Shenzhen, Guangdong, 518083, China

24
25 9
26
27 10
28 11 Conghui Liu: rapherlch@163.com; Yan Zhang: milrazhang@163.com; Bo Liu:

29
30 12 ~~lb_bobo@aliyun.com;~~ Yuwei Ren: xiaoshudaxia@126.com; ~~Yan Zhang:~~

31
32 13 ~~milrazhang@163.com;~~ Hengchao Wang: wanghengchao000@qq.com; Shuqu Li:

33
34 14 lishuqu1234@163.com; Fan Jiang: greatjf@163.com; Lijuan Yin:

35
36 15 yinlijuan1005@163.com; Guojie Zhang: guojie.zhang@bio.ku.dk

37
38 16 *These authors contributed equally to this work.

39
40 17 †Correspondence should be addressed to Wanqiang Qian (qianwanqiang@caas.cn),

41
42 18 [Bo Liu \(lb_bobo@aliyun.com\)](mailto:Bo Liu (lb_bobo@aliyun.com) or Wei Fan (fanwei@caas.cn).) or Wei Fan (fanwei@caas.cn).

43
44
45
46 19
47 20 **Abstract**

48
49 21 **Background:** The golden apple snail (*Pomacea canaliculata*) is a ~~worldwide~~-fresh
50
51 22 water snail listed ~~amongst~~ the top-100 worst invasive species, ~~worldwide~~ and a noted

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

23 agricultural and quarantine pest ~~that causes causing huge~~ great economic losses. ~~It~~
24 is characterized ~~by with~~ fast growth, strong stress tolerance, a high reproduction rate,
25 and adaptation to a broad range of environments.

26 **Results:** Here, we used long-read sequencing to produce a 440-Mb high-quality
27 chromosome-level assembly for the *P. canaliculata* genome. In total, 50 Mb (11.4%)
28 repeat sequences and 21,533 gene models were identified in the genome. The major
29 ~~Major~~ findings of this study include the recent explosion of DNA/hAT-Charlie
30 transposable elements (TEs), the expansion of the P450 gene family and the
31 constitution of the cellular homeostasis system, which ~~contributes~~ to ~~the~~ ecological
32 plasticity in ~~the~~ stress adaptation. In addition, the ~~perivitellin gene expansion and~~ high
33 transcriptional levels of perivitellin genes in the ovary and albumen gland promote the
34 function of nutrients supplying and defense-defence ability in ~~the~~ eggs. Furthermore,
35 the gut metagenome also contains diverse encodes a rich array of genes for food
36 digestion and xenobiotics degradation.

37 **Conclusions:** These findings collectively provide novel insight into the molecular
38 mechanisms of the ecological plasticity and high invasiveness. ~~Our results not only~~
39 ~~strengthen the understanding of molluscs genomics and biological invasion, but also~~
40 ~~benefit preventing the invasion of apple snail and transmission of pathogenetic~~
41 ~~parasites.~~

42 **Keywords:** golden apple snail, *Pomacea canaliculata*, genome, adaptive evolution,
43 stress tolerance, P450, reproduction, perivitelline, metagenome

44 **Background**

45 The golden apple snail *Pomacea canaliculata* (family Ampullariidae, Order
46 Architaenioglossa) is a fresh water snail listed ~~in~~among the world's top 100 of the
47 ~~world's~~ worst invasive species [1], and is considered ~~as a noted an~~ agricultural and
48 quarantine pest worldwide [2]. Native to ~~the~~tropical and subtropical South America,
49 ~~the~~*P. canaliculata* gradually spread to ~~the~~ non-indigenous regions, such as Southeast
50 and East Asia [3], Africa [4], North America [5], Oceania [6] and even Europe [7],
51 ~~and the~~Its successful biological invasion ~~was due was~~ closely related to its
52 polyphagous feeding habits [8], voracious appetite [9], broad environmental
53 adaptability [10] and rapid growth and high rate of reproduction [11]. In addition to its
54 ~~Besides the~~ ecological impact, ~~the~~*P. canaliculata* ravaged a wide range of crops,
55 including grains, fruits and vegetables [12], causing severe economic losses each year
56 as a result of yield loss, replanting cost and ~~the funds of~~ expenditures on control
57 (<https://www.cabi.org/isc/datasheet/68490>). More seriously, *P. canaliculata* has been
58 involved in the transmission of a ~~human~~ fatal human disease, ~~e~~Eosinophilic
59 meningitis, that ~~firstly~~ appeared in East Asia where people ~~take them as food~~
60 frequently consume these snails [13]. During this pathophoresis, *P. canaliculata* acts
61 as an important intermediate host of the pathogenic parasite *Angiostrongylus*
62 *cantonensis*, and the range of ~~infectious~~ infected regions is still expanding, creating a
63 ~~causing~~ great challenge in terms of ~~to~~ human health [14, 15].
64 Molluscs are a highly diverse group, ~~and~~ second only to arthropods in species

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

65 number [16], and their high biodiversity makes ~~them molluscs~~ an excellent model to
66 address ~~the~~ issues such as biogeography, adaptability and evolutionary processes [17].
67 ~~And~~ the worldwide invasive species *P. canaliculata* provides valuable potential in
68 these fields [18]. As a primitive circumtropical species, *P. canaliculata* possesses
69 strong ecological plasticity with many advantages to hold advantage on plenty of
70 aspects, including low-temperature resistance [19], and drought tolerance [20], which
71 has contributed to its competitive success ~~succeed~~ in resource acquisition ~~over the~~
72 ~~competitive species~~. It was reported *P. canaliculata* has been reported to establish
73 populations ~~could set population in the distribution of at~~ temperatures ranging from
74 10 °C to 35 °C [19, 21]. Additionally, *P. canaliculata* ~~is tolerate~~ tolerant with heavy
75 metal contamination. When living in contaminated water, ~~the~~ gill is enriched with
76 a high concentration of heavy metals and histopathological changes in the digestive
77 tract ~~is~~ are detected; however, an ~~with~~ extremely low mortality rate is observed [21, 22].
78 ~~For protection of embryos,~~ The conspicuous coloration and neurotoxic lectin could
79 confer ~~the eggs~~ a survival advantage on the eggs, and ~~defend~~ the embryos
80 against ~~the~~ potential predators [22, 23]. Moreover, ~~the~~ immune-neuroendocrine
81 system can also be detected in *P. canaliculata*, as demonstrated by the existence of a
82 specific immune memory after ~~the~~ bacterial challenge [24, 25], broadening the
83 studies of invertebrate immunology.

84 The rich phenotypic and genetic diversity of molluscs makes them an excellent
85 species group for addressing many important issues in evolution, ecology and
86 function. However, the genomic resources of Mollusca are still insufficient

Formatted: Not Highlight

Formatted: Not Highlight

Formatted: Indent: First line: 0"

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

87 ~~compared with those of other close phyla~~ms, such as Arthropoda and Nematoda, and
88 ~~few molluscs can~~be employed as model organisms. *P. canaliculata*, however,
89 ~~possesses the potential to be a model organism among~~of molluscs because of several
90 ~~inherent characteristic~~s. For example, *P. canaliculata* is easy to acquire ~~for~~because
91 ~~it has a broad global distribution originating~~ed from a primarily circumtropical
92 ~~environment. Due to the~~Moreover, its high adaptability, rapid growth and efficient
93 ~~reproduction~~, facilitate the cultivation of *P. canaliculata* ~~also facilitate the cultivation~~
94 ~~in the laboratory.~~
95 ~~In recent~~During the past years, the genomic features of *P. canaliculata* have been
96 increasingly studied. After the discovery of 14 pachytene bivalents in the karyotype
97 ~~[2526]~~, molecular markers were identified to investigate the genetic diversity of ~~the~~ *P.*
98 *canaliculata* population, including 369 amplified fragment length polymorphism
99 (AFLP) loci ~~[2627]~~, 16,717 simple sequence repeats (SSR) ~~[27, 2828, 29]~~ and
100 15,412 single-nucleotide polymorphisms SNPs ~~[2930]~~. In addition, multiple
101 transcriptome analyses have been performed to investigate the adaptation, invasion
102 and immune mechanisms ~~of~~ *P. canaliculata*. For instance, Sun et al. reported 128,436
103 unigenes based on a de novo assembly of Illumina reads ~~[2930]~~; transcriptome
104 changes in response to heat stress and starving incubation ~~were~~was used to
105 characterize ~~its~~ invasive and adaptive abilities ~~[30, 3131, 32]~~; a transcriptome
106 analysis ~~between~~comparing invasive *P. canaliculata* and indigenous
107 *Cipangopaludina ca*hayensis provideds insights into biological invasion ~~[2829]~~; and
108 402 immune-related differentially expressed genes (DEGs) ~~in response to~~ by

Formatted: Font: Not Italic

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

109 Lipopolysaccharide (LPyS) challenge were used to explore the mechanisms of
110 defence against pathogens [3332]. Furthermore, proteomics tools such as isobaric
111 ~~Tags tags For for Relative relative, and a Absolute Quantitation quantitation~~ (iTRAQ),
112 and ~~Liquid liquid Chromatography chromatography~~-tandem ~~Mass mass Spectrometry~~
113 ~~spectrometry~~ (LC-MS/MS) were also applied in the study of protein expression
114 ~~during for the~~ estivation and oviposition [34, 3533, 34], together providing plentiful
115 omics_ data for the functional analysis of *P. canaliculata*.
116 However, researches at the whole_ genome level in *P. canaliculata* still lags far behind
117 ~~that in~~ other mollusc_ species, due to the lack of a high-quality reference genome. ~~By~~
118 ~~far, M~~ultiple draft genomes of molluscs have been published, ~~such as including the~~
119 ~~genomes of the~~ California sea hare [3635], Pacific oyster [3736], ~~Pearl oyster [37]~~
120 ~~pearl oyster [38]~~, owl limpet [3938], California two-spot octopus [3940], ~~deep sea~~
121 ~~golden mussel [4140]@,~~ and *Biomphalaria* snails [4142], greatly promoting the
122 research onf mollusc_ genomics. In this study, we present a chromosome-level
123 genome assembly of *P. canaliculata* with high-quality gene annotation, transcriptome
124 data from several tissues and under various conditions, ~~as well as the and~~
125 metagenomic data from the intestinal tracts, all of which were then applied to study
126 the species-specific ~~environmental adaptation~~~~invasive~~ characteristics, such as the
127 cellular homeostasis system underlying strong stress, and the colour and nutrient
128 contents of the eggs. Our data will not only strengthen the understanding of the
129 evolutionary mechanisms of molluscs and the molecular basis of biological invasion,
130 but also foster the developments of approaches to control the invasion of *P.*

- Formatted: Not Highlight
- Formatted: Not Highlight
- Formatted: Not Highlight
- Formatted: Not Highlight
- Formatted: Not Highlight
- Formatted: Not Highlight

1
2
3
4
5
6
7 131 *canaliculata* and ~~provide a basis for interrupting~~~~interrupt~~ the transmission of
8
9 132 pathogenetic nematode parasites.
10
11
12

13 133 RESULTS

14 134 Complete genome assembly at the chromosome level

15
16 135 We generated 26.6 Gb (60.1 X) of PacBio SMRT raw reads with an average read
17
18 136 length of 10.1 ~~Kb~~kb, and 291 Gb (652.4 X) of Illumina HiSeq paired-end reads with
19
20 137 an average read length of 150-250 bp, using DNA extracted from ~~gene~~ single adult *P.*
21
22 138 ~~*canaliculate-canaliculata*~~ (Table S1). The 24.4 Gb (55.4 X) of clean PacBio SMRT
23
24 139 reads that passed quality filtering were assembled by smartdenovo
25
26 140 (<https://github.com/ruanjue/smartdenovo>), ~~giving rise to resulting in~~ an assembly of
27
28 141 1,234 raw contigs with a total length of 473.6 Mb and an N50 length of 1.0 Mb. After
29
30 142 filtering of alternatively heterozygous contigs, the 745 resulting contigs with a total
31
32 143 length of 440.1 Mb and an N50 length of 1.1 Mb were taken as the final contigs.
33
34 144 Previous karyotype research has shown that the haploid *P.* ~~*canaliculate-canaliculata*~~
35
36 145 genome consists of 14 chromosomes [~~2526~~]. Based on the Hi-C data, 439.5 Mb
37
38 146 (99.9%) of final contigs were anchored and oriented into 14 large scaffolds, each
39
40 147 corresponding to a natural chromosome (Figure 1a and Figure 1b), with the longest
41
42 148 45.4 Mb and the shortest 27.2 Mb. This assembly quality is much better than that of
43
44 149 the other ~~published~~ molluscan genomes ~~published thus~~ far (Table 1). ~~Besides In~~
45
46 150 addition to the length and continuity of the assembled sequences, another important
47
48 151 aspect for evaluating genome assembly is the ratio of genome coverage. With an
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7 152 estimated genome size of 446 Mb and genome heterozygosity between 1% and 2%
8
9 153 based on the distribution of k-mer frequency [4243] (Figure S1), ~98.6 % of the P.
10
11 154 canaliculata genome has been assembled ~~in *P. canaliculata*~~. To further confirm the
12
13 155 accuracy and completeness of the assembly, we mapped the Illumina shotgun reads to
14
15 156 the assembled reference genome. Significantly, 97% and 95% of the genome-derived
16
17 157 and transcriptome-derived reads, respectively, could be aligned to the reference
18
19 158 genome, ~~respectively~~, suggesting no obvious bias ~~for~~in sequencing and assembly.
20
21 159 Additionally, the mitochondrial genome of *P. canaliculata* was ~~also~~ assembled as a
22
23 160 single contig ~~with~~ 15,707 bp in length, which has 99.9-% sequence identity to the
24
25 161 published mitochondrial genome (GenBank: KJ739609.1) (Figure S2). These
26
27 162 high-quality reference genome provides a good foundation for gene annotation.
28
29 163 The protein-coding genes were predicted on the reference genome by EVM,
30
31 164 integrating evidences from *de novo* prediction, transcriptome and homology data. In
32
33 165 total, 21,533 gene models were predicted as the reference gene set, with coding
34
35 166 regions spanning ~32.2 Mb (7.3 %) of the genome (Table 1 and Table S2). The
36
37 167 distribution of CDS length in *P. canaliculata* is similar to that in the closely related
38
39 168 species (Figure 1c). Overall, 97.5-% of the reference genes were supported by
40
41 169 transcriptome data, and 98.0-% of eukaryote core genes from OrthoDB
42
43 170 (<http://www.orthodb.org/>) were identified in the reference gene set by BUSCO. These
44
45 171 results were comparable to those in the other published molluscan genomes (Table 1).
46
47 172 ~~For the~~ In functional annotation, a total of 19,815 (91.9 %) reference genes were
48
49 173 annotated by at least one functional database. Specifically, 15,662 (72.7-%), 13,769
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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

174 (63.4%), 17,081 (79.3%), 18,847 (87.5%) and 17,003 (79.9%) reference genes
175 were annotated with [the](#) eggNOG, KEGG, NR, InterPro and UniProt databases,
176 respectively (Figure S3).

177 **Signs of ~~Adaptive adaptive Evolution evolution~~ in *P. canaliculata* ~~Genomegenome~~**

178 To gain insight into [the](#) evolutionary perspective of *P. canaliculata*, ~~the a~~ phylogenetic
179 tree was built based on ~~306474~~ high-confidence single-copy orthologous genes from
180 ~~nineseven~~ related species (*P. canaliculata*, ~~*Lottia*-*gigantea*~~, ~~*Aplysia*-*californica*~~,
181 ~~*Biomphalaria*-*glabrata*~~, ~~*Crassostrea*-*gigas*~~, ~~*Octopus*-*bimaculoides*~~, ~~*Pintada*-*fucata*~~,
182 ~~*Lingula*-*anatina* and *Limnoperna*-*fortunei* and *L. anatina*~~) by PhyML ~~ml~~ [4344] and
183 the divergence time was estimated using ~~MCMCTree~~ ~~metree~~ [4445]. The results
184 shows that *P. canaliculata* diverged from the ancestor of *B. glabrata* and *A.*
185 ~~*californica*~~ ~~372290~~ million years ago (Mya), and from *L. gigantea* ~~491415~~ Mya
186 (Figure 2a).

187 Then, the molluscan orthologous genes were investigated for adaptive evolution.
188 Utilizing pairwise protein sequence similarities, ~~the~~ gene family clustering was
189 conducted by orthfinder [4546]. A total of ~~239,541,152,878~~ reference genes from the
190 ~~seven-nine~~ species were clustered into ~~69,582,68,942~~ orthologous groups, amongst
191 which ~~14,766,13,805~~ orthologous groups ~~with-contained~~ at least two genes each.
192 ~~Compared to other seven species, w~~We identified 66 orthologue groups ~~undergone that~~
193 ~~underwent -common expansion in both in~~ *P. canaliculata* and *L. fortunei* ~~fortune but~~
194 ~~not the other seven species.~~ ~~In *P. canaliculata*, we identified 9,626 ortholog groups,~~

Formatted: Font: Italic

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

195 ~~amongst which 117 and 5,462 ortholog groups undergone species-specific expansion,~~

196 ~~thus may play important roles in adaptation to the environment as an invasive species.~~

197 The functions of these orthologous groups are mainly related to signal transduction;

198 replication and repair; translation, glycan biosynthesis; and metabolism; Lipid

199 metabolism; and the endocrine, immune and nervous systems ~~digestive, endocrine,~~

200 ~~signal transduction, immune, or carbohydrate metabolism and so on~~ (Figure S4).

201 ~~These relations~~ ~~It suggested that the expansion~~ gene families that underwent

202 expansion may play important roles in adaptation to the environment as invasive

203 species.

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

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

206 in eukaryotes [4647]. In total, we identified 49.6 Mb TE sequences in the assembled *P.*

207 *canaliculata* genome (Table 1), including 3.4 Mb long terminal repeats (LTRs), 27.2

208 Mb long interspersed elements (LINEs), 17.5 Mb DNA transposons and 1.5 Mb short

209 interspersed elements (SINEs). Next, we ~~analyzed~~ analysed the divergence rate of ~~TEs~~

210 ~~for~~ each class of TEs among the available sequenced mollusk genomes.

211 ~~Notably worthy, the TE class of DNA transposons showed a specific, interestingly,~~

212 ~~only the results of DNA transposons showed a unique~~ peak at a divergence rate of ~4%

213 divergence rate for *P. canaliculata* and *C. gigas* (Figure 2b), indicating a recent

214 explosion of DNA transposons in these two species. We analyzed the expression of

215 709 genes, including DNA elements that ~~restricted to the 4% peak inside the gene~~

216 region, compared with that of the other genes that ~~outside the 4% peak~~ (Figure S5).

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

217 ~~Differentially expressed genes (DEGs)~~ were defined here by P-values smaller than
218 0.05 for comparison of ~~the~~ treatments (heat, cold, heavy metal and air exposure) and
219 control data. The percentages of DEGs in the 4% peak were higher than those of
220 genes outside the peak (10.2% higher for heat, 8.6% higher for cold, 8.6% higher for
221 heavy metal, and 7.3% higher for air exposure). Among the DEGs in the 4% peak,
222 approximately ~~about~~ half ~~are~~ were up-regulated, and the other half ~~were~~ ~~are~~
223 down-regulated. Moreover, the DEGs in the 4% peak were mainly enriched in cellular
224 metabolic process, response to stimulus, localization and ~~signaling~~ ~~signaling~~
225 according to ~~by~~ GO annotation. These results indicated that genes in the 4% peak
226 were likely to be more active in the response to ~~of~~ stimulus, promoting ~~the~~ potential
227 plasticity in ~~the~~ stress adaptation. More than half of the DNA transposons belong to the
228 DNA/hAT Charlie TE family, which is ~~22.7%~~ of total DNA/hAT Charlie TEs in the
229 genome. TEs are powerful facilitators of evolution ~~that generate~~ ~~by~~ ~~generating~~
230 “evolutionary potential” to introduce small adaptive changes within a lineage, and the
231 importance of TEs ~~in~~ ~~to~~ stress responses and adaptation has been reported in
232 numerous ~~studies~~ ~~researches~~ [48, 49, 47, 48]. The recent explosion of DNA/hAT Charlie
233 TEs in *P. canaliculata* could also play ~~an~~ important roles ~~in~~ ~~to~~ promoting ~~the~~ potential
234 plasticity in ~~the~~ stress adaptation.

235 Investigation of ~~Cellular~~ ~~cellular~~ homeostasis system underlying strong stress 236 adaptation

237 ~~The~~ ~~h~~Homeostasis system plays a crucial role in ~~the~~ stress adaptability, providing the
238 molecular basis ~~for~~ ~~in~~ re-establishing ~~the~~ dynamic equilibrium after ~~the~~ challenges ~~by~~

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

239 ~~of~~ various environmental stressors, including temperature, air exposure,
240 anthropogenic pollution and pathogens [5049]. In ~~this the present~~ study, we addressed
241 three constituent parts of the cellular homeostasis system, which contributes to the
242 successful ecological plasticity of *P. canaliculata* (Figure 3). ~~The~~ transcriptomes ~~data~~
243 of the hemocytes after ~~different~~ stimuli~~s~~ (cold, heat, heavy ~~metal~~ and air exposure)
244 ~~were~~ also sequenced and ~~analyzed~~ ~~analysed~~ to address the potential roles of ~~these~~
245 genes in ~~the c~~ellular homeostasis system.
246 ~~The u~~nfolded protein response (UPR) system ~~is makes~~ the central ~~component part~~ of
247 protein homeostasis [5150]. Heat shock proteins (HSPs) ~~acts~~ as molecular chaperones
248 to maintain ~~the~~ correct folding, and heat shock transcription factor 1 (HSF1) ~~are is~~
249 responsible for the transcriptional induction of HSPs [5254]. In ~~the~~ *P. canaliculata*
250 genome, 13 HSP70s, ~~7~~ 6 HSP90s, 7 HSP40s and 11 HSFs were identified (Table S3),
251 and the expression of HSP90s and HSFs ~~were was~~ highly induced in response to ~~the~~
252 ~~stress of~~ heat, cold, heavy metal and air exposure (Table S4 ~~and Figure S6~~).
253 Inositol-requiring protein 1 (IRE1), protein kinase RNA-like ER kinase (PERK), and
254 activating transcription factor 6 (ATF6) are three mediators recruited by ~~the~~
255 endoplasmic reticulum (ER) to ~~regulated~~ the UPR [5352]. We found putative coding
256 genes of the three core mediators, their respective downstream transcription factors,
257 and the corresponding recognition chaperones in ~~the~~ *P. canaliculata* genome (Table
258 S3).
259 ~~The x~~enobiotic biotransformation system helps the molluscs adapt to toxicants,
260 especially ~~the~~ pesticides in aquatic environments [5453]. Manual annotation ~~on of~~ this

1
2
3
4
5
6
7 261 genome identified 157 cytochrome P450s (CYP450s), 15 flavin-containing
8
9 262 monooxygenases (FMOs), 53 glutathione S-transferases (GSTs) and 105 ATP binding
10
11 263 cassette (ABC) transporters, most of which showed ~~an up-regulation-regulated~~
12
13 264 expression under stress (Table S3 ~~and~~ Table S4). These proteins ~~are evidenced~~ have
14
15 265 been shown to function in contaminant ~~detecting~~ detection, conjugative modification
16
17 266 and expulsion for xenobiotic detoxification [55-57] ~~54-56~~.
18
19 267 ~~The m~~Massive production of reactive oxygen species (ROS) and reactive oxygen
20
21 268 intermediates (ROI_s) induced by stress leads to many pathological conditions, and
22
23 269 antioxidant systems protect the organism from superoxide [58] ~~57~~. Four main
24
25 270 antioxidant enzyme classes, namely, superoxide dismutase (SOD), catalase (CAT),
26
27 271 peroxidase (Prx), and glutathione peroxidase (GPX), were found in ~~the~~ *P. canaliculata*
28
29 272 ~~and showed with an elevating~~ elevated global expression in response to stress (Table
30
31 273 S3 ~~and~~ Table S4).
32
33 274 Apoptosis is a process of cell death when sensing stress and the regulation of
34
35 275 apoptosis maintains the dynamic homeostasis of the internal environment. In *P.*
36
37 276 *canaliculata*, we propose the existence of both intrinsic and extrinsic apoptotic
38
39 277 ~~signaling signaling~~ pathways, evidenced by the presence of homologous genes
40
41 278 involved in both pathways. ~~It seems t~~These two pathways could be activated by
42
43 279 cytochrome C and tumou~~r~~ necrosis factor receptor (TNFR), respectively (Table S3).
44
45 280 ~~The i~~nhibitors of apoptosis, such as XIAP, Bcl2 and Bak, are also detected and show
46
47 281 ~~with an~~ increased expression in response to ~~the~~ stress (Table S4), which ~~are~~ is
48
49 282 expected to delay the process of apoptosis ~~process~~ and ~~the~~ cell death in the stress
50
51
52
53

1
2
3
4
5
6
7 283 response.

8
9
10 284 **The expansion of the P450 gene family contribute to stress tolerance**

11
12
13 285 Cytochromes P450 (CYP) enzymes are a monooxygenase family with highly diverse
14
15 286 structures and functions; that have been widely broadly identified in all kingdoms of
16
17 287 life [5859]. P450s catalyze-catalyse the reductive scission of molecular oxygen, and
18
19 288 are responsible for the synthesis and metabolism of various molecules, including
20
21 289 drugs, hormones, antibiotics, pesticides, carcinogens and toxins [5960]. The hormones
22
23 290 they synthesized-hormones, such as glucocorticoids, mineralocorticoids, progestins,
24
25 291 and sex hormones, are critical to stress response, growth and reproduction, and the
26
27 292 endogenous and exogenous chemical metabolism participate in helps-the host
28
29 293 combatting with the toxic compounds [6061].

30
31
32 294 We found that the *P. canaliculata* CYP gene family had greater level of undergone an
33
34 295 expansion compared to that in the other molluscs. We identified 157 genes in the
35
36 296 genome of *P. canaliculata*, and 128, 102, 135, 78, 52 and 94 genes in from-A.
37
38 297 Californiacalifornica, *B. glabrata*, *C. gigas*, *L. gigantea*, *O. bimaculoides* and *P.*
39
40 298 *fucata* respectively, using under the same standard (Figure 4a). An The expansive
41
42 299 trend was also observed; in comparison compared with other the model species, such
43
44 300 as *Homo sapiens* (57), *Mus musculus* (102), *Danio rerio* (94) and *Drosophila*
45
46 301 *melanogaster* (94) [6264]. The gene expansion was mainly found in the CYP2U and
47
48 302 CYP3A sub-families, whereas and fewer genes were expanded in CYP4F. In
49
50 303 mammals, CYP2U participates plays-a role in the metabolism of fatty acids to

1
2
3
4
5
6
7 304 generate bioactive eicosanoid derivatives, potentially regulating the development of
8
9 305 immune function [6362]. In *P. canaliculata*, 40 genes ~~formed forged into~~ the CYP2U
10
11 306 clade, mainly ~~expressing expressed~~ in ~~the~~ hepatopancreas (Figure 4b and Table S5_a,
12
13 307 Table S5_b). CYP3A ~~is acts as~~ a versatile enzyme ~~metabolizing that metabolizes~~ a
14
15 308 wide range of xenobiotics, and ~~its production the productions~~ promotes the growth of
16
17 309 various cell types [6463]. The 56 CYP3A genes ~~are have~~ comprehensively ~~expression~~
18
19 310 ~~expressed~~ in ~~the~~ hepatopancreas, gill and kidney (Figure 4b and Table S5_a, Table
20
21 311 S5_b). CYP4F possesses epoxygenase activity, metabolizing fatty acids to epoxides to
22
23 312 suppress hypertension, pain perception and inflammation [6564]. ~~Twenty20~~ genes
24
25 313 were identified in CYP4F, and ~~Pc06G011748, Pc06G011460, Pc06G011458,~~
26
27 314 ~~Pc06G011459, Pc04G006708, Pc04G006710 and Pc04G006707~~ ~~several CYP4F genes~~
28
29 315 ~~exhibited present~~ highly induced expression levels under ~~the stress of~~ cold, heat,
30
31 316 heavy metal and air exposure ~~stress~~, indicating their critical roles in the stress
32
33 317 tolerance (Figure 4b, ~~and~~ Table S5_a, ~~and~~ Table S5_b).

34
35
36
37
38 318 **The identification of perivitellin genes and their high transcriptional levels in the**
39
40 319 **ovary and albumen gland perivitellin gene expansion and high transcriptional**
41
42 320 **level in ovary enhance reproduction**

43
44 321 ~~To adapt to the fast invasion life, besides the strong ability to stress tolerance, t~~ ~~The P.~~
45
46 322 ~~canaliculata possesses a high reproductive rate, and one important contributor is their~~
47
48 323 ~~distinct has~~ eggs characterized ~~with by~~ abundant nutrients, reddish or pinkish colour,
49
50 324 aerial oviposition and neurotoxicity [2223, 3466]. ~~In most gastropod eggs, and these~~
51
52 325 ~~characteristics are contributed by due to the perivitelline~~ ~~Pervitelline~~ Fluid (PVF),
53
54
55
56
57
58
59
60
61
62
63
64
65

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

326 ~~with large amounts of nutrients which fills filled in the~~ space between the eggshell
327 and the embryo, ~~and consists which is composed~~ of carbohydrates, lipids and proteins
328 ~~(Figure 5a). -termed perivitellins, which is not only responsible for the major supply~~
329 ~~of material and energy during embryogenesis, but also provide warning pigment and~~
330 ~~deadly toxicant against the predators [65]. ~~In *P. canaliculata*, The PVF proteins iIn *P.*~~~~
331 ~~*canaliculata*, include- three major Perivitellins components, of PcOvo, PcPV2, and~~
332 ~~PcPV3 [67], collectively named perivitellins, which mtake up 90% of the total~~
333 ~~proteins, whereas most of the other dozens of low-abundance components each only~~
334 ~~account for less than 1% of the total proteins [35]. The perivitellins ~~isare~~ not only~~
335 ~~responsible for the major supply of materials and energy during embryogenesis, but~~
336 ~~also provide warning pigments and deadly toxicants against the predators [23, 68, 69].~~
337 ~~of *P. canaliculata* (Pe) have been verified by proteomics approach and was further~~
338 ~~divided into three categories called Pe Ovorubin (PeOvo), PePV2, PePV3, which are~~
339 ~~all high density lipoprotein (HDL) [66] (Figure 5a).-~~
340 ~~We identified 28 candidate PVF genes in *P. canaliculata*, by mapping each of the 59~~
341 ~~fragmental PVF protein sequences derived from a previous proteomics study by Sun~~
342 ~~[35] to ~~their~~its best hit in the reference gene set of *P. canaliculata*, using~~
343 ~~BLASTPblastp- with requirements of over 85% identity and at least 50% alignment~~
344 ~~length (Table S6). Then, the functional annotation of those fragmental proteins was~~
345 ~~also transferred to our identified PVF genes. We totally identified 18 perivitellin genes~~
346 ~~from the *P. canaliculata* genome, compared to 2 and 1 perivitellin genes from *A.*~~
347 ~~*californica* and *P. fucata* respectively, by aligning the seven reference perivitellin gene~~

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

~~sequences (NCBI accession AFQ23940.1, AFQ23939.1, AFQ23938.1, AFQ23945.1, AFQ23937.1, P0C8G7.2, P0C8G6.2) to each genome sequences with the same method (blastn e-value 10^{-20}). It is apparent that the copy number of perivitellin genes was expanded in *P. canaliculata*, and our orthologous and paralogous gene family data by orthoFinder confirmed this. Among the 20 perivitellin genes in *P. canaliculata*, there are 2 PcOvo, 13 PcPV2, and 3 unclassified PVFs (Figure 5b and Table S6). The PcOvo carotenoprotein is responsible for the red coloration of the eggs and antioxidant to protect against sun radiation and desiccation [67, 68], while PcPV2 is reported to be neurotoxin implying lethal effect on rodents [22]. The expansion of these genes may enhance the underlying functions of nutrition and protection, offering the eggs an advantage of survival and improve the reproduction rate.~~

The transcriptome data shows that 22 (79%) of the 28 candidate PVF genes exhibit their highest expression in the ovary and albumen gland (PVF protein synthesis factory) among all the 7 tissues (Figure 5b and Table S7), confirming that most of them are genuine real-functional PVF genes. Six of In these 28 candidate PVF genes are, there are 6 perivitellin genes, including two PcOvo genes, Pc09G015543 (PcOvo2) and Pc09G015548 (PcOvo3); two PcPV2 genes, Pc07G012572 (PcPV2-31) and Pc07G012571 (PcPV2-67); and two possible PcPV3 genes, Pc09G015546 and Pc09G015547. The expression levels of these 6 genes in the ovary and albumen gland for these 6 genes are much higher than those of the other 22 candidate PVF genes.

By analyzing the orthoFinder gene families that include orthologous and paralogous genes from *P. canaliculata* and 8 other sequenced mollusc species, we found that

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

370 ~~these 28 candidate PVF genes were classified into 20 multiple-gene families (≥ 2~~
371 ~~genes) and 7 single-gene families (only one gene) (Table S8). Notably, 5 of~~
372 ~~the 6 perivitellin genes were classified fall into single-gene families, except for~~
373 ~~Pc07G012571 (PcPV2-67), which that not only has homologous genes in other~~
374 ~~mollusc species but also has three paralogous genes in *P. canaliculata* itself. However,~~
375 ~~none of these three PcPV2-67 paralogous genes of in *P. canaliculata* showed higher~~
376 ~~expressed higher by in the ovary and albumen gland than in other tissues,~~
377 ~~indicating that they are were likely not PVF-related genes, i.e., only Pc07G012571~~
378 ~~plays a role in PVF. The nearly unique and single-copy nature of the 6 perivitellin~~
379 ~~genes in *P. canaliculata*, may be explained by the long evolutionary distance, over~~
380 ~~200 Mya for *P. canaliculata* and its most closely related species, *A. californica*,~~
381 ~~as well as numerous plenty of differences in their living characteristics and egg~~
382 ~~structures. s. On the other hand, it also indicates Another possible explanation is that~~
383 ~~these 6 major PVF genes may have experienced fast rapid evolution in their the history,~~
384 ~~in order to adapt to the changing environment.~~
385 ~~The expression of 18 *P. canaliculata* perivitelline genes were detected in 7 tissues,~~
386 ~~including embryo, testis, ovary, kidney, gill, hepatopancreas and hemocyte. The~~
387 ~~highest expression of each gene concentrated in embryo and two sexual gland testis~~
388 ~~and ovary, especially in the ovary (Figure 5b and Table S7), suggesting that their~~
389 ~~decoding proteins might be of importance in germ cell production and embryo~~
390 ~~development. Taken together, *P. canaliculata* distinguish its embryo development~~
391 ~~from other seven species on the preponderance of perivitellin gene number and high~~

Formatted: Font: Not Italic

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

392 ~~expression level, that further promotes corresponding function of nutrients supplying~~
393 ~~and defense ability and eventually contribute to reproduction.~~

394 **The gut microbiome plays important roles in stress resistance and food**
395 **digestion**

396 The gut microbiome is well known as the second genome of animals, ~~which and~~ plays
397 ~~important~~key roles in food digestion, immune ~~defense~~defence, ~~etc and other processes~~
398 that are essential to the animal ~~hosts~~. To investigate whether the gut microbiome ~~has~~
399 ~~influence on influences~~ the invasive life-style, we collected gut digesta samples from
400 70 ~~adults of~~*P. canaliculata*, ~~adults~~snails and generated 31 Gb ~~of~~ high-quality
401 metagenomic data on ~~the~~ Illumina HiseqX10 platform. To our knowledge, this ~~study~~
402 is the first ~~in~~high-depth sequencing of ~~the~~ snail gut microbiome. A total of 1,142,095
403 non-redundant genes were obtained, with an average open reading frame (ORF)
404 length of 604 bp (Table ~~S8S9~~). The taxonomic composition analysis showed that, at
405 the phylum level, Proteobacteria was ~~the~~ predominant, followed by Verrucomicrobia,
406 Bacteroidetes, Firmicutes, Spirochaetes, Actinobacteria, etc. (Table ~~S9S10~~ a). At the
407 genus level, the most abundant genera included ~~d~~*Aeromonas*, *Enterobacter*,
408 *Desulfovibrio*, *Citrobacter*, *Comamonas*, *Klebsiella* and *Pseudomonas*. (Table
409 ~~S9S10~~ b), most of which were also presented in ~~the snails of~~*Achatina fulica*
410 [70,71~~69~~,70].

411 ~~It is interesting that~~ Interestingly, some of the most abundant genera, such as
412 *Desulfovibrio*, *Citrobacter* and *Pseudomonas*, were reported ~~as having to have~~ strong
413 abilities ~~to of removing remove~~ heavy metals, ~~by~~ mechanisms of bioprecipitation

Formatted: Font: Not Italic

1
2
3
4
5
6
7 414 and bioabsorption [72-7471-73]. For example, the sulfur-reducing bacteria
8
9 415 *Desulfovibrio* produces ~~ed~~ H₂S, ~~which that~~ precipitates ~~s~~ metals, and therefore ~~reduced~~
10
11 416 ~~reduces~~ the toxic effects of ~~dissolving-dissolved~~ metals [7172]. Based on the KEGG
12
13 417 pathway database, the complete sulfate reduction metabolism pathway was identified
14
15 418 in the *P. canaliculata* gut microbiome. We suggested that ~~these~~ gut microbes might
16
17 419 help *P. canaliculata* ~~to confront with~~ ~~survive~~ the environmental stress of heavy metals
18
19 420 in ~~harsh~~ conditions. In addition, a large number of genes in ~~pathways-related-to-of~~
20
21 421 xenobiotics biodegradation and metabolism ~~pathways~~ were annotated, corresponding
22
23 422 to 288 KEGG orthologous groups (KOs) and 21 pathways (Table ~~S10S11~~). As many
24
25 423 of the pathways, such as benzoate degradation, toluene degradation, xylene
26
27 424 degradation and steroid degradation, could not be identified in the host genome
28
29 425 through KO analysis, we suggested that ~~the~~-microbial detoxification abilities may
30
31 426 contribute ~~to~~ the ~~ability~~ *P. canaliculata* to resist stresses caused by xenobiotics such as
32
33 427 pesticides and environmental pollutants.

34
35
36
37 428 In ~~view of dietary~~ digestion, the gut microbes ~~are were~~ directly involved in ~~the~~
38
39 429 breakdown of the cellulose portion ~~of the diet~~, and previous studies have isolated
40
41 430 ~~some~~-cellulolytic bacteria and evaluated the cellulolytic enzyme activities [7574]. ~~On~~
42
43 431 ~~our work~~ ~~found~~, a broader range of carbohydrate active enzymes (CAZymes) ~~were~~
44
45 432 ~~found~~. Of the 208 annotated CAZyme families, 99 were ~~Glycoside-glycoside~~
46
47 433 ~~Hydrolase-hydrolase~~ (GH) families (Table ~~S11S12~~). Enzymes that could be classified
48
49 434 as cellulases, endohemicelluloses, debranching enzymes, ~~and~~
50
51 435 oligosaccharide-degrading enzymes were all ~~-identified~~~~presented~~. These findings
52
53

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 indicate that the gut microbiome ~~provides give~~ assistance ~~to in~~ digesting a broad range
437 of food sources, ~~enabling making~~ *P. canaliculata* ~~to grow~~ rapidly ~~and in order fast to~~
438 adapt to an invasive life-style.

439 **Conclusion and discussion**

440 Given its environmental invasiveness, broad stress adaptability and rapid reproduction,
441 the golden apple snail *P. canaliculata* has received a vast amount of attention
442 worldwide. However, the underlying genetic mechanisms of these properties ~~haves~~
443 not been comprehensively uncovered. The chromosome-level genome of *P.*
444 *canaliculata* presented in this study sheds the first lights on into the genomic basis of
445 ~~its the~~ ecological plasticity in response to various stressors. ~~M~~The major findings of
446 this study include the recent explosion of DNA/hAT-Charlie TEs, the expansion of the
447 P450 gene family and the constitution of the cellular homeostasis system, all of
448 ~~which contributing contribute~~ to the plasticity ~~in of the organism in the~~ stress
449 adaptation. Although the ~~defined~~ function of the recently originated TEs could not be
450 confirmed, ~~the explosion of~~ TEs are considered is deemed as powerful facilitators in
451 adaptive evolution, suggesting that indicating its their increased number plays an
452 important role in the stress resistance of *P. canaliculata*'s stress resistance. The UPR
453 system, ~~Xenobiotic xenobiotic~~ biotransformation system and ROS system are all
454 major components of the ~~Cellular cellular~~ homeostasis system, and the especially
455 P450s in particular underwent expansion expands with specific functions. In addition,
456 exclusive perivitellin genes ~~are were identified in characterized from~~ the *P.*

1
2
3
4
5
6
7 457 *canaliculata* genome, ~~and they are believed to contributing~~ contribute to the high
8
9 458 reproductive rate and the expansion of habitats. Furthermore, the gut metagenome
10
11 459 ~~encodes contains diverse a rich array of~~ genes for food digestion and xenobiotics
12
13 460 degradation. These findings collectively provide novel insight into the molecular
14
15 461 mechanisms of ~~the~~ ecological plasticity and high invasiveness.
16
17 462 ~~The rich phenotypic and genetic diversity of molluscs make them an excellent species~~
18
19 463 ~~group to address many valuable issues about evolution, ecology and function.~~
20
21 464 ~~However, the genomic resource of Mollusca is still insufficient compared with other~~
22
23 465 ~~close phylums, such as Arthropoda and Nematoda, and few molluscs could be~~
24
25 466 ~~employed as model organism. *P. canaliculata* possesses potential to be a model~~
26
27 467 ~~organism of molluscs because of several inherent characters. For example, *P.*~~
28
29 468 ~~*canaliculata* is easy to acquire, for it has a broad global distribution originated from a~~
30
31 469 ~~primarily circumtropical environment. Due to the high adaptability, rapid growth and~~
32
33 470 ~~efficient reproduction, *P. canaliculata* also facilitate the cultivation in laboratory. In~~
34
35 471 ~~this study, w~~We report a fine reference genome of *P. canaliculata* ~~in the present study,~~
36
37 472 ~~which is the~~ first chromosome-level Mollusca genome published ~~in Mollusca.~~
38
39 473 ~~Together with the~~ With its easy acquisition, rapid growth and efficient
40
41 474 reproduction, *P. canaliculata* possesses the potential to be a model organism of
42
43 475 Mollusca. As ~~its the~~ cellular complexity and ~~the~~ conservation of pathways also make,
44
45 476 *P. canaliculata* ~~could be a~~ useful representative of Mollusca, ~~so~~ the genome described
46
47 477 in this study can be used to advance our understanding of the molecular mechanisms
48
49 478 involved in ~~for~~ various scientific questions regarding issues in Mollusca.
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

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

479

480 **Methods**

481 **Samples collection and sequencing**

482 Adults of *P. canaliculata* were collected from a local paddy field in Shenzhen,
483 Guangdong province, China, and maintained in aerated freshwater at 15 ± 2 °C for a
484 week before processing. Genomic DNA was extracted from the foot muscles of a
485 single *P. canaliculata* for constructing PCR free Illumina 350-bp insert libraries and
486 PacBio 20-kb insert library, and sequenced on Illumina HiSeq 2500 and PacBio
487 SMRT platforms, respectively. The Hi-C library was prepared using the muscle tissue
488 of another single *P. ~~canaliculata~~ canaliculata* by following methods: Nuclear DNA
489 was cross-linked in situ, extracted, and then digested with a restriction enzyme. The
490 sticky ends of the digested fragments were biotinylated, diluted, and then ligated to
491 each other randomly. Biotinylated DNA fragments were enriched and sheared again
492 for preparing the sequencing library, which was then sequenced on a HiSeq X Ten
493 platform (Illumina).

494 Seven tissues including embryos (2 days post fertilization), gill, hemocytes,
495 hepatopancreas, kidney, ovary and albumen gland and testis from six animals were
496 collected as parallel samples. Next, animals were cultivated in 37 °C and 10 °C for 24
497 hours heat and cold tolerance, in $Cr^{3+}(2mg L^{-1})$, $Cu^{2+}(0.2mg L^{-1})$ and $Pb^{2+}(1mg L^{-1})$
498 for 24 hours heavy metal tolerance, and in waterless tank for 7 days air exposure.
499 Then the hemocytes were harvested and stored, with three replicates for each group.

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

500 In final, ~~total RNAs were extracted from the stored tissues of *P. canaliculata*~~
501 ~~materials, and then mRNAs were pulled out by beads with poly-T for constructing~~
502 ~~cDNA libraries~~ ~~total messenger RNAs (mRNA) were extracted from the stored tissues~~
503 ~~of *P. canaliculata* materials for constructing cDNA libraries~~ (insert 350-bp), and
504 sequenced on an Illumina HiSeq 2500 sequencer.

505 The intestinal digesta from 70 adult snails of *P. canaliculata* were collected, pooled
506 into 6 samples and stored at -20 °C until microbial DNA was extracted. A
507 combination of cell lysis treatments was applied, including five freeze-thaw cycles
508 (alternating between 65 °C and liquid nitrogen for 5 min), repeated beads-beating in
509 ASL buffer (cat. no. 19082; Qiagen Inc.), and incubated at 95 °C for 15 min. DNA
510 was isolated following the protocol reported protocol [7675]. Paired-end libraries of
511 metagenomic DNA were prepared with an insert size of 350 base pairs (bp) following
512 the manufacture's protocol (cat. no. E7645L; New England Biolabs). Sequencing was
513 performed on Illumina HiSeq X10.

514
515 **Genome assembly and annotation**

516 The Illumina raw reads were filtered by trimming the adapter sequence and
517 low-quality- ~~regions~~ ~~part~~ (https://github.com/fanagislab/common_use), resulting in a
518 clean and high-quality reads ~~data~~ with an average error rate < 0.001. For the PacBio
519 raw data, the short subreads (< 2 kb) and low-quality (error rate > 0.2) subreads were
520 filtered out, and only one representative subread was retained for each PacBio read.

Formatted: Font: Not Bold

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

521 The clean PacBio reads were assembled by the software ~~smrt~~smartdenovo
522 (<https://github.com/ruanjue/smartdenovo>), ~~then~~after which Illumina reads were
523 aligned to the contigs by BWA-MEM, and single base errors in the contigs were
524 corrected by Pilon (v1.16) ~~with~~ the parameters “-fix bases, -nonpf, -minqual 20”.
525 The *P. canaliculata* genome is highly heterozygous, as illustrated by the double peaks
526 on the distribution curve of ~~k~~k-mer frequency, and the current assembly algorithm
527 tends to collapse homozygous regions and report heterozygous regions in alternative
528 contigs. To ~~obtain~~get a haploid reference contigs, we employed a whole-genome
529 alignment (WGA) strategy ~~with~~by MUMmer v3.23 to recognize and selectively
530 remove alternative heterozygous contigs, which were characterized by shorter length
531 (less than 200 kb) and the ability of most regions (~~more~~larger than 50%) ~~to~~can be
532 aligned to another larger contig with confident identity (higher than 80%). Next, Hi-C
533 sequencing data were aligned to the haploid reference contigs by BWA-MEM, and
534 then these contigs were clustered into chromosomes with LACH-ESIS
535 (<http://shendurelab.github.io/LACHESIS/>~~http://shendurelab.github.io/LACHESIS/~~).
536 A de novo repeat library for *P. canaliculata* was constructed by RepeatModeler
537 (v1.0.4; <http://www.repeatmasker.org/RepeatModeler.html>). TEs in the *P. canaliculata*
538 genome were also identified by RepeatMasker (v4.0.6; <http://www.repeatmasker.org/>)
539 using both the Repbase library and the de novo library. Tandem repeats in the *P.*
540 *canaliculata* genome were predicted using Tandem Repeats Finder v4.07b [77]. The
541 divergence rates of TEs were calculated between the identified TE elements in the
542 genome and their consensus sequence at the TE family level.

Formatted: Font: Italic

Field Code Changed

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

543

544 The gene models in the *P. canaliculata* genome were predicted by Evidence Modeler
545 v1.1.1 [7876], integrating evidences from ab initio predictions, homology-based
546 searches and RNA-seq alignments. Then, these gene models were annotated by
547 RNA-seq data, UniProt database and InterProScan software ~~the protein-coding~~
548 ~~sequences were mapped by RNA-seq data and functionally annotated using UniProt~~
549 ~~and InterProScan (5.16-55.0) databases~~ [7779]. Finally, the gene models were retained
550 if they had at least one piece of supporting evidence from the UniProt database,
551 InterProScan domain and RNA-seq data. Gene functional annotation was performed
552 by aligning the protein sequences to the NCBI NR, UniProt, COG and KEGG
553 databases with BLASTP v2.3.0+ under an E-value cutoff of 10^{-5} and choosing the best
554 hit. ~~The p~~Pathway analysis and functional classification were conducted based on the
555 KEGG database [8078]. InterProScan was used to assign preliminary GO terms, Pfam
556 domains and IPR domains to the gene models.

557 ~~A de novo repeat library for *P. canaliculata* was constructed by RepeatModeler~~
558 ~~(v1.0.4; <http://www.repeatmasker.org/RepeatModeler.html>). TEs in the *P. canaliculata*~~
559 ~~genome were also identified by RepeatMasker (v4.0.6; <http://www.repeatmasker.org/>)~~
560 ~~using both Replibase library and the de novo library. Tandem repeats in the *P.*~~
561 ~~*canaliculata* genome were predicted using Tandem Repeats Finder v4.07b [79]. The~~
562 ~~divergence rates of TEs were calculated between the identified TE elements in the~~
563 ~~genome and their consensus sequence at the TE family level.~~

564

Formatted: Font: (Default) Times New Roman, Font color: Black

1
2
3
4
5
6
7 565 **Evolutionary analysis**

8
9
10 566 Orthologous and paralogous groups were assigned from seven species (*P.*
11
12 567 *canaliculata*, *Lottia gigantea*, *Aplysia californica*, *Biomphalaria glabrata*,
13
14 568 *Crassostrea gigas*, *Octopus bimaculoides*, *Pintada fucata*, *Limnoperna fortunei* and
15
16 569 *Lingula anatina*) by OrthoFinder [4546] with default parameters. Orthologous groups
17
18 570 that contained only one gene for each species were selected to construct the
19
20 571 phylogenetic tree. The protein sequences of each gene family ~~were was~~
22
23 572 independently aligned by muscle v3.8.31 [8081] and then concatenated into one
24
25 573 super-sequence. The phylogenetic tree was constructed by maximum likelihood (ML)
26
27 574 using PhyML v3.0 [4443] with the
28
29 575 best-fit model (LG+I+G) ~~that was~~ estimated by ProtTest3 [8281]. The Bayesian
30
31 576 ~~Relaxed-relaxed Molecular-molecular Clock-clock~~ (BRMC) approach was adopted to
32
33 577 estimate the neutral evolutionary rate and species divergence time using the program
34
35 578 MCMCTree, implemented in the PAML v4.9 package [4544]. ~~The calibration time~~
36
37 579 ~~(fossil record time) interval (173-398 Mya) of Octopus bimaculoides was adopted~~
38
39
40 580 ~~from previous results.~~
41
42 581 The tree was calibrated with the following time frames to constrain the age of the
43
44 582 nodes between the species: minimum = 260 Ma and maximum = 290 Ma for *P. fucata*
45
46 583 and *C. gigas* [83]; minimum = 450 Ma and maximum = 480 Ma for *A. californica* (or
47
48 584 *B. glabrata*) and *L. gigantea* [84]. The calibration time (fossil record time) interval
49
50 585 (550-610 Mya) of *O. bimaculoides* was adopted from previous results [85].

51
52
53 586

Formatted: Font: Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

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

587 **Transcriptome data analysis**

588 ~~Transcriptome reads were mapped to the reference genome of *P. canaliculata* using~~
589 ~~TopHat (v. 2.1.0) with default settings. Transcriptome reads were trimmed with the~~
590 ~~same method for genomic reads (https://github.com/fanagislab/common_use), and~~
591 ~~then mapped to the reference genome of *P. canaliculata* using TopHat (v. 2.1.0) with~~
592 ~~default settings.~~ The expression level of each reference gene in terms of FPKM was
593 computed by cufflinks v2.2.1. A gene was considered to be expressed if its FPKM >0.
594 Differential gene expression analysis was conducted using cuffdiff v2.2.1.

596 **Metagenome data analysis**

597 Raw reads were cleaned to exclude adapter sequences, low-quality sequences, ~~and as~~
598 ~~well as~~ contaminated DNA. The adapter sequence ~~was identified and trimmed from~~
599 ~~the reads in reads were identified and trimmed~~ by an ungapped dynamic programming
600 algorithm; the low-quality part (head or tail) of ~~the reads~~ ~~were-was~~ trimmed off to
601 ensure that the average error rate of the ~~left-remaining~~ reads ~~is-was~~ lower than 0.001;
602 the reads that ~~were~~ mapped to ~~the~~ contaminated DNA by BWA-MEM [8286] were
603 filtered out; ~~and~~ finally, shorter reads (length < 75-~~_~~bp) and unpaired reads were
604 excluded to form a ~~set of~~ clean reads-~~data~~. The BWA database built for cleaning
605 contamination included genomes of 10 species: ~~the~~ *P. canaliculata* genome, ~~the~~
606 *Brassica rapa* genome, ~~the~~ *Oryza sativa* genome, 2 *Angiostrongylus cantonensis*
607 genomes, ~~the~~ *Caenorhabditis elegans* genome, ~~the~~ *Schistosoma mansoni* genome, ~~the~~

Formatted: Font: Not Bold

Formatted: Font: Not Italic

1
2
3
4
5
6
7 608 *Celonorchis sinensis* genome, the *Ffasciola hepatica* genome, the *Danio rerio*
8
9 609 genome, and the human hg38 genome.
10
11 610 The clean reads were assembled by metaSPAdes (v3.11.1) [~~8387~~] ~~in under-paired-end~~
12
13 611 mode for each sample. ~~T~~hen, gene prediction was performed on contigs longer than
14
15 612 500 bp by Prodigal (v2.6.3) [~~8488~~] with the parameter “-p meta”, and gene models
16
17 613 with cds length less than 102 bp were filtered out. A non-redundant (NR) gene set
18
19 614 (539,344 genes) was constructed using the gene models predicted from each samples
20
21 615 by cd-hit-est (v4.6.6) [~~8589~~] with the parameter “-c 0.95 -n 10 -G 0 -a S 0.9”, which
22
23 616 adopts a greedy incremental clustering algorithm and the criteria of identity > 95%
24
25 617 and overlap > 90% of the shorter genes. Then, the clean reads were mapped onto this
26
27 618 NR gene set by BWA-MEM with the criteria of alignment length \geq 50bp and
28
29 619 identity > 95%. The unmapped reads from all samples were assembled together, and
30
31 620 the genes were predicted again. The newly predicted genes were combined with the
32
33 621 previous gene set by cd-hit-est to ~~obtain get~~ a new NR gene set (1,147,339 genes).
34
35 622 After the taxonomic assignments to the new NR gene set, 5244 genes classified as
36
37 623 Eukaryota but not fungi were removed, and the final NR gene set (1,142,095 genes)
38
39 624 was obtained.
40
41 625 The taxonomic assignments ~~for of~~ the final NR genes were made on the basis of
42
43 626 DIAMOND [~~8690~~] protein alignment against the NCBI_NR database by CARMA3
44
45 627 [~~8791~~]. Functional annotation was performed by aligning all the protein sequences to
46
47 628 the KEGG [~~8892~~] database (release 79) using DIAMOND and taking the best hit with
48
49 629 the criteria of E-value < 1e-5. CAZymes were annotated with dbCAN (release 5.0)
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

630 [8993] using HMMER (v3.0) hmmscan [9094] by taking the best hit with an E-value
631 < 1e-18 and coverage > 0.35.

632 The clean reads from each sample were aligned against the gene catalogue (1,142,095
633 genes) by BWA-MEM with the criteria of alignment length \geq 50bp and identity >
634 95%. Sequence-based gene abundance profiling was performed as previously
635 described [9195]. The taxonomic profiles of the samples were calculated by summing
636 ~~adding~~ the gene abundance ~~together~~ according to the taxonomic assignment result.

637

638

639 Abbreviations

640 ~~*P. Canaliculata*, *Pomacea canaliculata*, *L. gigantean*, *Lottia gigantean*;~~
641 *A. californica*, *Aplysia californica*; *B. glabrata*, *Biomphalaria glabrata*; *C.*
642 *gigas*, *Crassostrea gigas*; *O. bimaculoides*, *Octopus bimaculoides*; ~~*L. anatinae*,~~
643 *Lingula anatinae*; *L. fortune*, *Limnoperna fortune*; *L. gigantea*, *Lottia gigantea*; *P.*
644 ~~*canaliculata*, *Pomacea canaliculata*~~; *P. fucata*, *Pinctada fucata*; Hem, hemocyte; Te,
645 testis; Ov, ovary and albumen gland; Kn, kidney; GI, gill; Hp, hepatopancreas, Em,
646 embryo; SSR, simple sequence repeats; mya, million years ago; *BLAST*, *basic local*
647 *alignment search tool*; SNP, single nucleotide polymorphism; PVF, Pervitelline Fluid;
648 Ovo, ovarubin; AFLP, amplified fragment length polymorphism; DEGs, differentially
649 expressed genes; LPyS, Lipopolysaccharide; iTRAQ, Isobaric Tags For Relative,
650 Absolute Quantitation; LC-MS/MS, Liquid Chromatography-tandem Mass

Formatted: Font: (Default) Times New Roman, 12 pt, Italic

1
2
3
4
5
6
7 651 Spectrometry; TEs, transposable elements; LTR, long terminal repeats; LINE, long
8
9 652 interspersed elements; SINE, short interspersed elements; UPR, Unfolded protein
10
11 653 response; HSPs, heat shock proteins; HSF1, heat shock transcription factor 1; PERK,
12
13 654 protein kinase RNA-like ER kinase; ATF6,activating transcription factor 6; ER,
14
15 655 endoplasmic reticulum; CYP450s, cytochrome P450s; FMOs, flavin-containing
16
17 656 monooxygenases; GSTs, glutathione S-transferases; ABC, ATP binding cassette; ROS,
18
19 657 reactive oxygen species; ROI, reactive oxygen intermediates; SOD, superoxide
20
21 658 dismutase; CAT, catalase; Prx, peroxidase; GPX, glutathione peroxidase; TNFR,
22
23 659 tumor necrosis factor receptor; NR, non-redundant genes; ORF, open reading frame;
24
25 660 Kos, orthologous groups; CAZymes, carbohydrate active enzymes; GH, Glycoside
26
27 661 Hydrolase.
28
29
30
31
32
33

34 663 **Availability of data and materials**

35
36
37
38 664 Tables S1 to ~~S11-S12~~ and Figures S1 to ~~S4-S6~~ are available in the supplementary
39
40 665 information file. The raw sequencing data has been deposited in
41
42 666 DDBJ/EMBL/GenBank under project accession PRJNA427478, SRR6425828 for
43
44 667 genomic Illumina_PE125 sequencing data, SRR6425829 for genomic
45
46 668 Illumina_PE150 sequencing data, SRR6425827 for genomic ~~Paebio~~ PacBio
47
48 669 sequencing data, SRR6429132~SRR6429164 for transcriptome sequencing data, and
49
50 670 SRR6472920~SRR6472925 for gut microbiome data. All the analysis data have also
51
52 671 been released for public use and can be freely accessed at AGIS
53
54
55
56
57
58
59
60
61
62
63
64
65

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

672 ftp://ftp.agis.org.cn/~fanwei/Pomacea_canaliculata_Genome/ .

673 **Authors' contributions**

674 WF and WQ conceived the study and designed the experiments. CL and YZ
675 performed the genome sequencing and assembly, BL performed annotation and
676 evolutionary analysis. CL performed the stress tolerance analysis, YR performed the
677 reproduction analysis, YZ performed the metagenome analysis. HW, SL, FJ, LY
678 provide suggestions and help checking. WF, CL, BL, YR, YZ wrote the manuscript,
679 and GZ help revise the manuscript. All authors read and approved the final
680 manuscript.

682 **Competing interests**

683 The authors declare that they have no competing interests.

684 **Acknowledgements**

685 This project is supported by the National key research and development program of
686 China (2016YFC1200600), Shenzhen science and technology program
687 (JCYJ20150630165133395), Fund of Key Laboratory of Shenzhen
688 (ZDSYS20141118170111640), and The Agricultural Science and Technology
689 Innovation Program (ASTIP) of Chinese Academy of Agricultural Sciences(CAAS) &

690 Elite Youth Program of Chinese Academy of Agricultural Sciences. We thank
 691 Fanghao Wan, Jue Ruan, Yutao Xiao for providing constructive suggestions to this
 692 project.

693
 694
 695

696 Legends of **t**Tables and **f**Figures

697 Tables

698 Table 1. Summary of assembly and annotation of mollusk genomes

Genome feature	<i>P. canaliculata</i>	<i>L. gigantea</i>	<i>A. californica</i>	<i>B. glabrata</i>	<i>C. gigas</i>	<i>O. bimaculoides</i>
Assembled sequences (bp)	440,071,717	359,505,668	927,310,431	916,377,450	557,735,934	2,338,887,882
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

699

700 Figures

701 **Figure 1. The genome characteristics of *P. canaliculata*.** (a) Circos plot showing the
 702 genomic features. Track 1: 14 linkage groups of the genome; Track 2: distribution of
 703 transposon elements in chromosomes; Track 3: protein-coding genes located on
 704 chromosomes; Track 4: distribution of GC contents. (b) A genome-wide contacting
 705 matrix from Hi-C data between each pair of the 14 chromosomes, using a 100 kb

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

706 window size. The colour value ~~indicates means~~ the base 2 logarithm of the number of
707 valid reads ~~to base 2~~ ($\log_2(\text{valid reads})$). (c) Distribution of CDS length in six closely
708 related species.

709
710 **Figure 2. Evolutionary genomic analysis of between *P. canaliculata* ~~and other~~**
711 **molluscs.** (a) Phylogenetic placement of *P. canaliculata* within the ~~molluscs~~ dated
712 tree of molluscs. The estimated divergence time ~~were is~~ shown ~~on at~~ each branching
713 point, and *P. canaliculata* is shown ~~the species marked with in~~ red color was *P.*
714 *canaliculata*. (b) Distribution of divergence rate for the class of DNA transposons in
715 molluscs genomes. The divergence rate was calculated by comparing all TE
716 sequences identified in the genome to ~~the its~~ corresponding consensus sequence in
717 each TE subfamily. The red arrow indicates ~~the that~~ *P. canaliculata* and *C. gigas* had a
718 recent explosion of TEs at a divergence rate of ~4% ~~divergence rate~~.

Formatted: Font: Italic
Formatted: Font: Italic

719
720 **Figure 3. The cellular homeostasis system in *P. canaliculata*.** The unfolded
721 protein response (UPR) system ~~included includes~~ HSPs and HSF in the heat shock
722 response and CNX, NEF, GRP94, BIP, HSP40, ATF6, IRE1, PERK, COP2, XBP,
723 ATF4, TRAM and Derlin in the endoplasmic reticulum unfolded protein response
724 (UPR-ERAD). Apoptotic pathways ~~included~~ XIAPs, Bcl2, caspases, TNFR, and
725 FADD. The antioxidant systems ~~included~~ PRX, SOD, CAT and GPX. The xenobiotic
726 biotransformation system ~~included includes~~ EPHX3, P450, FMO and ABC transporter.
727 The colours of the Gene boxes for gene families ~~with the filled colors~~ represent the
728 degree of upregulation (FPKM-stimulus/FPKM-control) ~~by as~~ an overall result of
729 stress, including heat, cold, heavy metal and air exposure. Pathways and genes were
730 obtained based on KEGG annotation.

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

731
732 **Figure 4. The expansion of the P450 gene family in *P. canaliculata*.** (a)
733 Phylogenetic tree demonstrating orthologous and paralogous relationships of all P450
734 genes from 7 species including *P. canaliculata*, *A. californica*, *B. glabrata*, *C. gigas*, *L.*
735 *gigantea*, *O. bimaculoides* and *P. fucata*. P450 genes from seven species were
736 obtained based on Pfam annotation (Interpro) with an the E-value of 10^{-5} . Clades are
737 labelled by P450 subfamily names. The tree was constructed using the Maximum
738 maximum likelihood method in MEGA7, and the branch length scale indicates the
739 average number of residue substitutions per site. (b) Phylogenetic tree of P450 genes
740 in *P. canaliculata*, which is a subset of the phylogenetic tree for the 7 species, and
741 their heat map of expression (FPKM) in seven tissues (Hem, hemocyte; Te, testis; Ov,
742 Ovary ovary and albumen gland; Kn, kidney; Gl, gill; Hp, hepatopancreas; Em,
743 Embryo embryo); and heat map of induced expression
744 (FPKM-stimulus/FPKM-control) under stress (Con: control; heat; cold; Hm: heavy
745 metal; Exp: air exposure).

746
747 **Figure 5. The composition and expression of the *P. canaliculata* perivitellins**
748 **composition and expression in different tissues.** (a) Perivitelline Fluid-fluid (PVF)
749 is-lies under the eggshell and surrounds the embryo. It contains carbohydrates, lipids,
750 and proteins. Tand the proteins is-are also known as perivitellins and are classified
751 into three categories, of PcOvo, PcPV2, and PcPV3. (b) The displayed shown
752 expression value of PVF proteins is the base 10 logarithm of FPKM to base 2

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

753 (~~log₂FPKM~~log₁₀FPKM). The ~~first 3 letters in each gene ID~~ genes marked in red
754 encode refer to three classes of perivitellins, ~~uPV means unclassified perivitellins,~~
755 ~~PV2 means PePV2, Ovo means PeOvo. Abbreviations were used for~~ The 7 tissues
756 examined are abbreviated as follows: ~~(Hem, hemocyte; Te, testis; Ov, Ovary~~ ovary and
757 albumen gland; Kn, kidney; Gl, gill; Hp, hepatopancreas; Em, ~~Embryo~~ embryo).

758

759 **References**

760 1. Lowe S, Browne M, Boudjelas S, de Poorter M. 100 of the World's Worst Invasive Alien
761 Species: A selection from the Global Invasive Species Database. Auckland, New Zealand:
762 World Conservation Union (IUCN); 2000.

763 2. Ranamukhaarachchi SL, Wickramasinghe S. Golden apple snails in the world:
764 introduction, impact, and control measures. Global advances in ecology and
765 management of golden apple snails. 2006:133-52.

766 3. Naylor R. Invasions in Agriculture: Assessing the Cost of the Golden Apple Snail in Asia.
767 Royal Swedish Academy of Sciences. 1996;25:443-8.

768 4. Berthold T. Vergleichende Anatomie, Phylogenie und historische Biogeographie der
769 Ampullariidae: (Mollusca, Gastropoda). 1991.

770 5. Howells RG, Burlakova LE, Karatayev AY, Marfurt RK, Burks RL. Native and
771 introduced Ampullariidae in North America: History, status, and ecology. 2006:73-112.

772 6. Halwart M, Bartley DM. International mechanisms for the control and responsible use of
773 alien species in aquatic ecosystems, with special reference to the golden apple snail. Los

1
2
3
4
5
6
7 774 Baños, Philippines: Philippine Rice Research Institute (PhilRice); 2006.
8
9 775 7. López MA, Altaba CR, Andree KB, López V. First invasion of the Apple snail *Pomacea*
10
11 776 *insularum* in Europe. *Tentacle*. 2010;18:26-8.
12
13 777 8. Estebenet AL, Martín PR. *Pomacea canaliculata* (Gastropoda: Ampullariidae): life-history
14
15 778 traits and their plasticity. *Biocell* 2002;26:83-9.
16
17 779 9. Lach L. The spread of the introduced freshwater apple snail *Pomacea canaliculata*
18
19 780 (Lamarck) (Gastropoda Ampullariidae) on Oahu, Hawaii. *Bishop Museum Occasional*
20
21 781 *Papers*. 1999;58:66-71.
22
23 782 10. Yusa Y, Sugiura N, Wada T. Predatory Potential of Freshwater Animals on an Invasive
24
25 783 Agricultural Pest, the Apple Snail *Pomacea canaliculata* (Gastropoda: Ampullariidae), in
26
27 784 Southern Japan. *Biol Invasions*. 2006;8:137-47.
28
29 785 11. Lach L, Britton DK, Rundell RJ, Cowie RH. Food Preference and Reproductive Plasticity
30
31 786 in an Invasive Freshwater Snail. *Biol Invasions*. 2000;2:279-88.
32
33 787 12. Mochida O. Spread of freshwater *Pomacea* snails (Pilidae, Mollusca) from Argentina to
34
35 788 Asia. *Micronesica*. 1991;3 51-62.
36
37 789 13. Shan L, Zhang Y, Steinmann P, Zhou X. Emerging Angiostrongyliasis in Mainland China.
38
39 790 *Emerg Infect Dis*. 2008;14:161-4.
40
41 791 14. Caldeira RL, Mendonca CL, Goveia CO, Lenzi HL, Graeff-TeixeiraC Lima WS, et al.
42
43 792 First record of molluscs naturally infected with *Angiostrongylus cantonensis* (Chen, 1935)
44
45 793 (Nematoda: Metastrongylidae) in Brazil. *Memórias do Instituto Oswaldo Cruz*.
46
47 794 2007;102:887-9.
48
49 795 15. McMichael AJ, Beaglehole R. The changing global context of public health. *Lancet*
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7 796 (London, England). 2000;356:495-9.
8
9
10 797 16. Chapman A. Numbers of Living Species in Australia and the World. Australian Biological
11 798 Resources Study; 2009.
12
13 799 17. Lindberg DR, Ponder WF, Haszprunar G. The Mollusca: relationships and patterns from
14 800 their first half-billion years. Oxford University Press, Oxford; 2004.
15
16 801 18. Hayes KA, Cowie RH, Thiengo SC. A global phylogeny of apple snails: Gondwanan
17 802 origin, generic relationships, and the influence of outgroup choice (Caenogastropoda:
18 803 Ampullariidae). Biol J Linn Soc Lond. 2009;98:61-76.
19
20 804 19. Matsukura K, Tsumuki H, Izumi Y, Wada T. Physiological response to low temperature in
21 805 the freshwater apple snail, *Pomacea canaliculata* (Gastropoda: Ampullariidae). J Exp
22 806 Biol. 2009;212:2558-63.
23
24 807 20. Yusa Y, Wada T, Takahashi S. Effects of dormant duration, body size, self-burial and
25 808 water condition on the long-term survival of the apple snail, *Pomacea canaliculata*
26 809 (Gastropoda: Ampullariidae). Appl Entomol Zool. 2006;41:627-32.
27
28 810 21. Seuffert ME, Burela S, Martín PR. Influence of water temperature on the activity of the
29 freshwater snail *Pomacea canaliculata* (Caenogastropoda: Ampullariidae) at its
30 southernmost limit (Southern Pampas, Argentina). Journal of Thermal Biology. 2010;
31 35:77-84.
32
33 814 ~~2122~~. Kruatrachue M, Sumritdee C, Pokethitoyook P, Singhakaew S. Histopathological effects
34 815 of contaminated sediments on golden apple snail (*Pomacea canaliculata*, Lamarck 1822).
35 816 Bull Environ Contam Toxicol. 2011;86:610-4.
36
37 817 ~~2223~~. Dreon MS, Frassa MV, Ceolín M, Ituarte S, Qiu JW, Sun J, et al. Novel animal defenses

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

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

818 against predation: a snail egg neurotoxin combining lectin and pore-forming chains that
819 resembles plant defense and bacteria attack toxins. PLoS One. 2013;8:e63782.
820 doi:10.1371/journal.pone.0063782.

821 [2324](#). Ottaviani E, Caselgrandi E, Fontanili P, Franceschi C. Evolution, immune responses and
822 stress: studies on molluscan cells. Acta Biol Hung. 1992;43:293-8.

823 [2425](#). Ottaviani E, Accorsi A, Rigillo G, Malagoli D, Blom JM, Tascetta F. Epigenetic
824 modification in neurons of the mollusc *Pomacea canaliculata* after immune challenge.
825 Brain Res. 2013;1537:18-26.

826 [2526](#). Mercado Laczkó AC, Lopretto EC. Estudio cromosómico y cariotípico de *pomacea*
827 *canaliculata* (Lamarck, 1801) (Gastropoda, Ampullariidae). Revista del Museo Argentino
828 de Ciencias Naturales "Bernardino Rivadavia" Hidrobiología. 1998;8:15-20.

829 [2627](#). Xu J, Han X, Li N, Yu J, Qian C, Bao Z. Analysis of genetic diversity of three geographic
830 populations of *Pomacea canaliculata* by AFLP. Acta Ecol Sin. 2009;29:4119- 26.

831 [2728](#). Chen L, Xu H, Li H, Wu J, Ding H, Liu Y. Isolation and characterization of sixteen
832 polymorphic microsatellite loci in the golden apple snail *Pomacea canaliculata*. Int J Mol
833 Sci. 2011;12:5993-8.

834 [2829](#). Mu X, Hou G, Song H, Xu P, Luo D, Gu D, et al. Transcriptome analysis between
835 invasive *Pomacea canaliculata* and indigenous *Cipangopaludina cahayensis* reveals
836 genomic divergence and diagnostic microsatellite/SSR markers. BMC Genet. 2015;16:12.

837 [2930](#). Sun J, Wang M, Wang H, Zhang H, Zhang X, Thiyagarajan V, et al. De novo assembly of
838 the transcriptome of an invasive snail and its multiple ecological applications. Mol Ecol
839 Resour. 2012;12:1133-44.

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

840 ~~3031~~. Mu H, Sun J , Fang L, Luan T, Williams GA, Cheung SG, et al. Genetic Basis of
841 Differential Heat Resistance between Two Species of Congeneric Freshwater Snails:
842 Insights from Quantitative Proteomics and Base Substitution Rate Analysis. J Proteome
843 Res. 2015;14:4296-308.

844 ~~3432~~. Yang L, Cheng TY, Zhao FY. Comparative profiling of hepatopancreas transcriptomes in
845 satiated and starving *Pomacea canaliculata*. BMC Genet. 2017;18:18.

846 ~~3233~~. Xiong YM, Yan ZH, Zhang JE, Li HY. Analysis of albumen gland proteins suggests
847 survival strategies of developing embryos of *Pomacea canaliculata*. Molluscan Res.
848 2017:1-6.

849 ~~3334~~. Sun J, Mu H , Zhang H, Chandramouli KH, Qian PY, Wong CK, et al. Understanding the
850 regulation of estivation in a freshwater snail through iTRAQ-based comparative
851 proteomics. J Proteome res. 2013;12:5271-80.

852 ~~3435~~. Sun J, Zhang H, Wang H, Heras H, Dreon MS, Ituarte S, et al. First proteome of the egg
853 perivitelline fluid of a freshwater gastropod with aerial oviposition. J Proteome Res.
854 2012;11:4240-8.

855 ~~3536~~. Aplysia Genome Project. Broad Institute. Vertebrate Biology Group. 2009.
856 <https://www.broadinstitute.org/aplysia/aplysia-genome-project>

857 ~~3637~~. Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, et al. The oyster genome reveals stress
858 adaptation and complexity of shell formation. Nature. 2012;490:49-54.

859 ~~3738~~. Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, et al. Draft genome
860 of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology. DNA
861 Res. 2012;19:117-30.

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

862 [Du X, Fan G, Jiao Y, Zhang H, Guo X, Huang R, et al. The pearl oyster *Pinctada fucata martensii*](#)
863 [genome and multi-omic analyses provide insights into biomineralization. *Gigascience*.](#)
864 [2017;6:1-12.](#)

865 [3839.](#) Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, et al.
866 Insights into bilaterian evolution from three spiralian genomes. *Nature*. 2013;493:526-31.

867 [3940.](#) Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsinger-Gonzales E, et al. The
868 octopus genome and the evolution of cephalopod neural and morphological novelties.
869 *Nature*. 2015;524:220-4.

870 [4041.](#) ~~Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea~~
871 ~~chemosynthetic environments as revealed by mussel genomes. *Nat Ecol Evol*. 2017;1:121.~~
872 ~~doi:10.1038/s41559-017-0121.~~

873 [Uliano-Silva M, Dondero F, Dan Otto T, Costa I, Lima NCB, Americo JA, et al. A](#)
874 [hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel](#)
875 [Limnoperna fortunei. *Gigascience*. 2017. doi: 10.1093/gigascience/gix128](#)

876 [4142.](#) Adema CM, Hillier LW, Jones CS, Loker ES, Knight M, Minx P, et al. Corrigendum:
877 Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nat Commun*.
878 2017;8:16153.

879 [4243.](#) Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, et al. Estimation of genomic characteristics
880 by analyzing k-mer frequency in de novo genome projects. *Quantitative Biology*
881 2013;arXiv:1308.2012 [q-bio.GN].

882 [4344.](#) Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms
883 and methods to estimate maximum-likelihood phylogenies: assessing the performance of

Formatted: Font: (Default) Times New Roman, 10.5 pt

Formatted

Formatted

Formatted

Formatted

Formatted

Formatted: Font: (Default) Times New Roman, 10.5 pt

Formatted: Font: (Default) Times New Roman, 10.5 pt

Formatted: Indent: First line: 0", Left 3.5 ch

Formatted

Formatted

Formatted

Formatted

Formatted

Formatted

Formatted: Font: (Default) Times New Roman, 10.5 pt

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

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

885 ~~4445~~. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol.
886 2007;24:1586-91. doi:10.1093/molbev/msm088.

887 ~~4546~~. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
888 comparisons dramatically improves orthogroup inference accuracy. Genome Biol.
889 2015;16:157. doi:10.1186/s13059-015-0721-2.

890 ~~4647~~. Feschotte C, Wessler SR. Mariner-like transposases are widespread and diverse in
891 flowering plants. Proc Natl Acad Sci U S A 2002;99:280-5.

892 ~~4748~~. Hua-Van A, Le Rouzic A, Boutin TS, Filée J, Capy P. The struggle for life of the
893 genome's selfish architects. Biol Direct. 2011;6:19.

894 ~~4849~~. Werren JH. Selfish genetic elements, genetic conflict, and evolutionary innovation. Proc
895 Natl Acad Sci U S A. 2011;108-Suppl 2:10863-70.

896 ~~4950~~. Chrousos GP. Stress and disorders of the stress system. Nat Rev Endocrinol.
897 2009;5:374-81.

898 ~~5051~~. Vabulas RM, Raychaudhuri S, Hayer-Hartl M. Protein folding in the cytoplasm and the
899 heat shock response. Cold Spring Harbor perspectives in biology. 2010;2:a004390.

900 ~~5152~~. Chen B, Retzlaff M, Roos T, Frydman J. Cellular Strategies of Protein Quality Control.
901 Cold Spring Harbor Perspectives in Biology. 2011;3:a004374.

902 ~~5253~~. Korennykh A and Walter P. Structural basis of the unfolded protein response. Annu Rev
903 Cell Dev Biol. 2012;28:251-77.

904 ~~5354~~. Chambers JE and Yarbrough JD. Xenobiotic biotransformation systems in fishes. Comp
905 Biochem Physiol C. 1976;55:77-84.

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

906 ~~5455~~. Mello DF, de Oliveira ES, Vieira RC, Simoes E, Trevisan R, Dafre AL, et al. Cellular and
907 Transcriptional Responses of *Crassostrea gigas* Hemocytes Exposed in Vitro to
908 Brevetoxin (PbTx-2) Mar Drugs. 2012;10: 583-97.

909 ~~5556~~. Boutet I, Tanguy A, Moraga D. Characterisation and expression of four mRNA sequences
910 encoding glutathione S-transferases pi, mu, omega and sigma classes in the Pacific oyster
911 *Crassostrea gigas* exposed to hydrocarbons and pesticides. Mar Biol 2004;146:53-64.

912 ~~5657~~. Deeley RG, Westlake C, Cole SP. Transmembrane transport of endo- and xenobiotics by
913 mammalian ATP-binding cassette multidrug resistance proteins. Physiol Rev.
914 2006;86:849-99.

915 ~~5758~~. Liu C, Zhang T, Wang L, Wang M, Wang W, Jia Z, et al. The modulation of extracellular
916 superoxide dismutase in the specifically enhanced cellular immune response against
917 secondary challenge of *Vibrio splendidus* in Pacific oyster (*Crassostrea gigas*). Dev
918 Comp Immunol. 2016;63:163-70.

919 ~~5859~~. Lamb DC, Lei L, Warrilow AG, Lepesheva GI, Mullins JG, Waterman MR, et al. The first
920 virally encoded cytochrome p450. J Virol. 2009;83:8266-9.

921 ~~5960~~. Urlacher VB, Girhard M. Cytochrome P450 monooxygenases: an update on perspectives
922 for synthetic application. Trends Biotechnol. 2012;30:26-36.

923 ~~6061~~. Sanderson T, van den Berg M. Topic 3.1: Interactions of xenobiotics with the steroid
924 hormone biosynthesis pathway. Pure Appl Chem. 2003;75:1957-71.

925 ~~6162~~. Goldstone JV, McArthur AG, Kubota A, Zanette J, Parente T, Jönsson ME, et al.
926 Identification and developmental expression of the full complement of Cytochrome P450
927 genes in Zebrafish. BMC Genomics. 2010;11:643.

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

928 [6263](#). Chuang SS, Helvig C, Taimi M, Ramshaw HA, Collop AH, Amad M, et al. CYP2U1, a
929 novel human thymus- and brain-specific cytochrome P450, catalyzes omega- and
930 (omega-1)-hydroxylation of fatty acids. *J Biol Chem.* 2004;279:6305-14.

931 [6364](#). Fleming I. The pharmacology of the cytochrome P450 epoxygenase/soluble epoxide
932 hydrolase axis in the vasculature and cardiovascular disease. *Pharmacol Rev.*
933 2014;66:1106-40.

934 [6465](#). Zhang G, Kodani S, Hammock BD. Stabilized epoxygenated fatty acids regulate
935 inflammation, pain, angiogenesis and cancer. *Prog Lipid Res.* 2014;53:108-23.

936 [6566](#). de Jong-Brink M, Boer HH, Joosse J. Mollusca. In: Adiyodi, K.G., Adiyodi,
937 R.G. (Eds.), *Reproductive Biology of invertebrates. Oogenesis oviposition and*
938 *oosorption*, vol. 1. John Wiley & Sons Ltd., New York, 1983; pp. 297-355.

939 [6667](#). Garin CF, Heras H, Pollero RJ. Lipoproteins of the egg perivitelline fluid of *Pomacea*
940 *canaliculata* snails (Mollusca: Gastropoda). *J Exp Zool.* 1996;276:307-14.

941 [6768](#). Dreon MS, Schinella G, Heras H, Pollero RJ. Antioxidant defense system in the apple
942 snail eggs, the role of ovorubin. *Arch Biochem Biophys.* 2004;422:1-8.

943 [6869](#). Dreon MS, Ituarte S, Heras H. The role of the proteinase inhibitor ovorubin in apple snail
944 eggs resembles plant embryo defense against predation. *PLoS One.* 2010;5:e15059.
945 doi:10.1371/journal.pone.0015059.

946 [6970](#). Cardoso AM, Cavalcante JJV, Vieira RP, Lima JL, Grieco MAB, Clementino MM, et al.
947 Gut Bacterial Communities in the Giant Land Snail *Achatina fulica* and Their
948 Modification by Sugarcane-Based Diet. *Plos One.* 2012;7 doi:ARTN
949 e3344010.1371/journal.pone.0033440.

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

950 ~~7071~~. Cardoso AM, Cavalcante JJV, Cantão ME, Thompson CE, Flatschart RB, Glogauer A, et
951 al. Metagenomic Analysis of the Microbiota from the Crop of an Invasive Snail Reveals a
952 Rich Reservoir of Novel Genes. Plos One. 2012;7 doi:ARTN
953 e4850510.1371/journal.pone.0048505.

954 ~~7472~~. Cabrera G, Pérez R, Gómez JM, Ábalos A, Cantero D. Toxic effects of dissolved heavy
955 metals on *Desulfovibrio vulgaris* and *Desulfovibrio* sp strains. J Hazard Mater
956 2006;135:40-6. doi:10.1016/j.jhazmat.2005.11.058.

957 ~~7273~~. Finlay JA, Allan VJ, Conner A, Callow ME, Basnakova G, Macaskie LE. Phosphate
958 release and heavy metal accumulation by biofilm-immobilized and chemically-coupled
959 cells of a *Citrobacter* sp. pre-grown in continuous culture. Biotechnol Bioeng.
960 1999;63:87-97.

961 ~~7374~~. Valls M, de Lorenzo V, Gonzalez-Duarte R, Atrian S. Engineering outer-membrane
962 proteins in *Pseudomonas putida* for enhanced heavy-metal bioadsorption. J Inorg
963 Biochem. 2000;79:219-23.

964 ~~7475~~. Pinheiro GL, Correa RF, Cunha RS, Cardoso AM, Chaia C, Clementino MM, et al.
965 Isolation of aerobic cultivable cellulolytic bacteria from different regions of the
966 gastrointestinal tract of giant land snail *Achatina fulica*. Front Microbiol. 2015;6
967 doi:Artn 86010.3389/Fmicb.2015.00860.

968 ~~7576~~. Zoetendal EG, Heilig HG, Klaassens ES, Booiijink CC, Kleerebezem M, Smidt H, et al.
969 Isolation of DNA from bacterial samples of the human gastrointestinal tract. Nature
970 protocols 2006, 1(2): 870-873.

971 [77. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids](#)

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

972 [Res. 1999;27:573-80.](#)

973 [78.](#) Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic
974 gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced
975 Alignments. Genome Biol. 2008;9:R7.

976 [7779.](#) Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan:
977 protein domains identifier. Nucleic Acids Res. 2005;33:W116-20.

978 [7880.](#) Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and
979 interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012;40:D109-D14.

980 [7980.](#) ~~Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids~~
981 ~~Res. 1999;27:573-80.~~

982 [8081.](#) Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
983 throughput. Nucleic Acids Res. 2004;32:1792-7.

984 [8182.](#) Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models
985 of protein evolution. Bioinformatics. 2011;27:1164-5. doi:10.1093/bioinformatics/btr088.

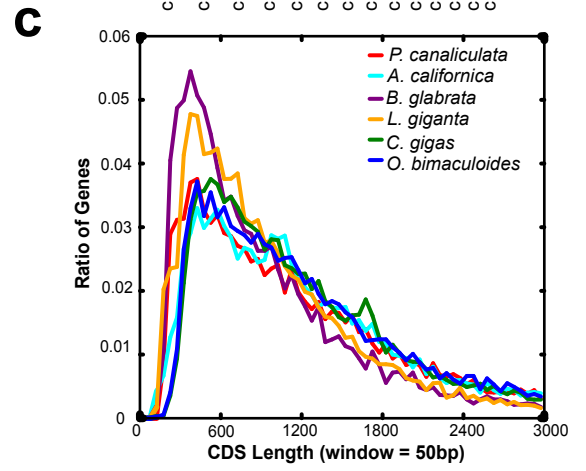
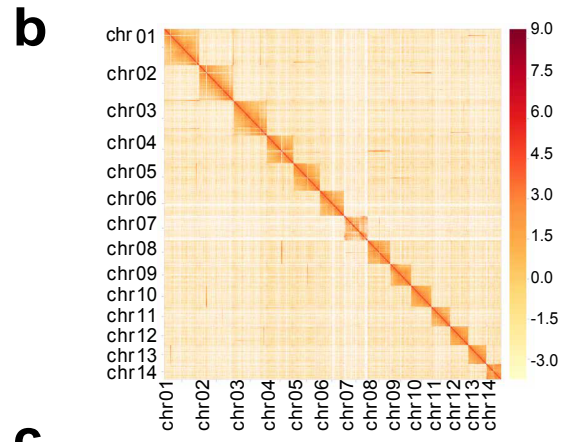
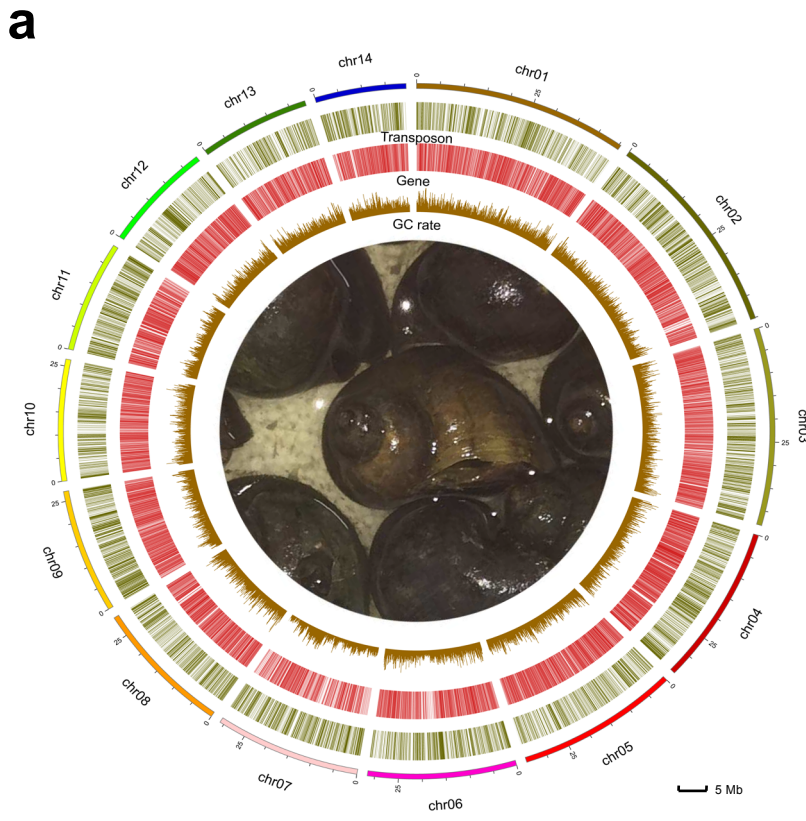
986 [83.](#) ~~Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, et al. Adaptation to deep-sea~~
987 ~~chemosynthetic environments as revealed by mussel genomes. Nature Ecology &~~
988 ~~Evolution. 2017; 1;121.~~

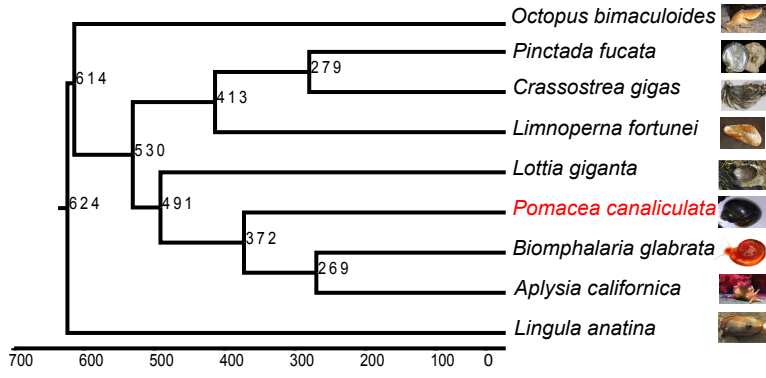
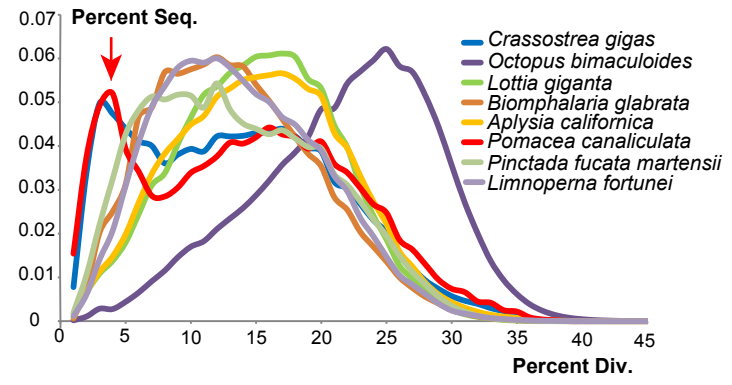
989 [84.](#) ~~Benton MJ, Donoghue PCJ, Asher, RJ. in The Timetree of Life:Calibrating and~~
990 ~~Constraining Molecular Clocks (eds Hedges, S. B. & Kumar, S.)35-86 (Oxford Univ.~~
991 ~~Press, 2009).~~

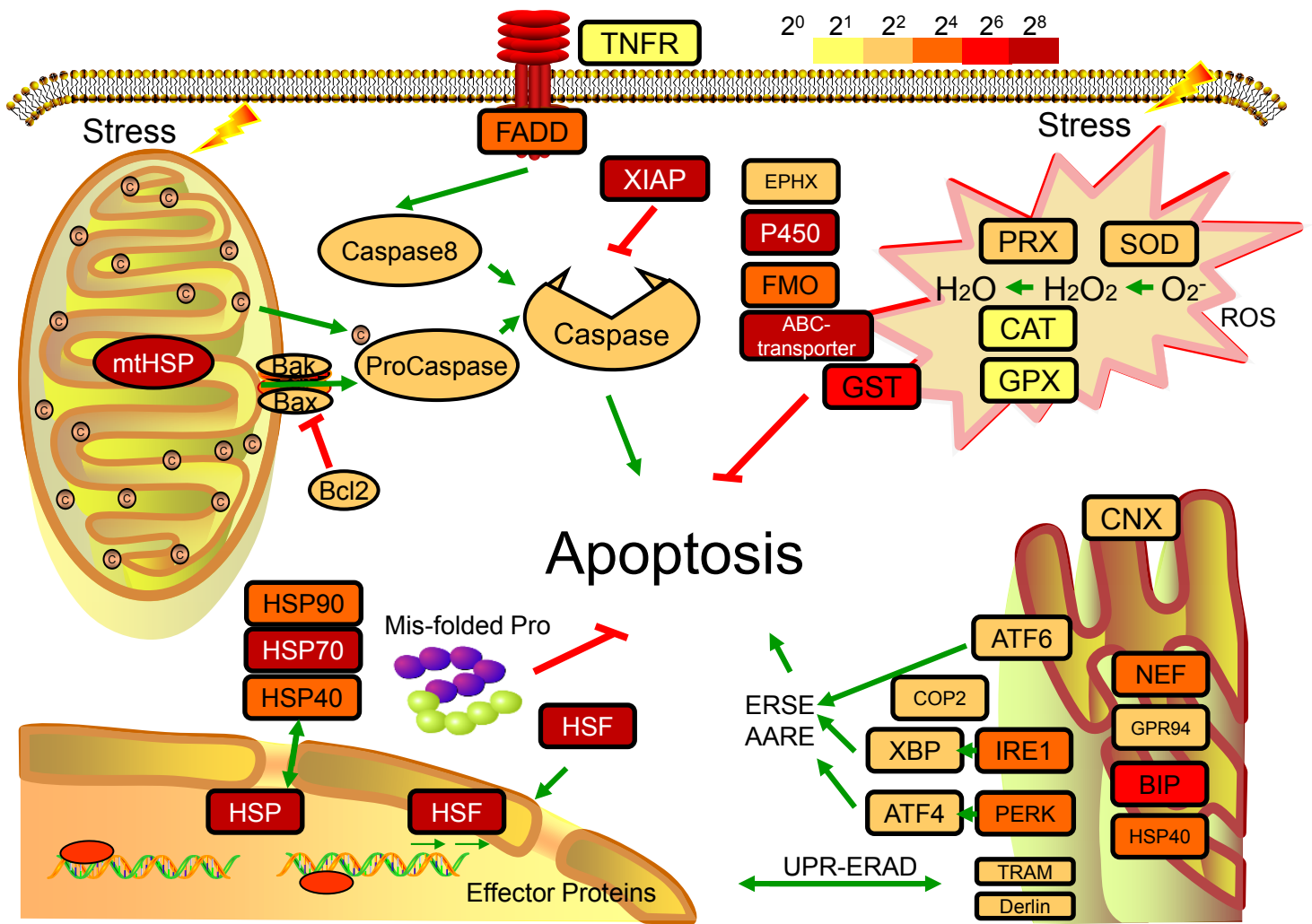
992 [85.](#) ~~Zapata F, Wilson NG, Howison M, Andrade SC, Jörger KM, Schrödl M, et al.~~
993 ~~Phylogenomic analyses of deep gastropod relationships reject Orthogastropoda. Proc Biol~~

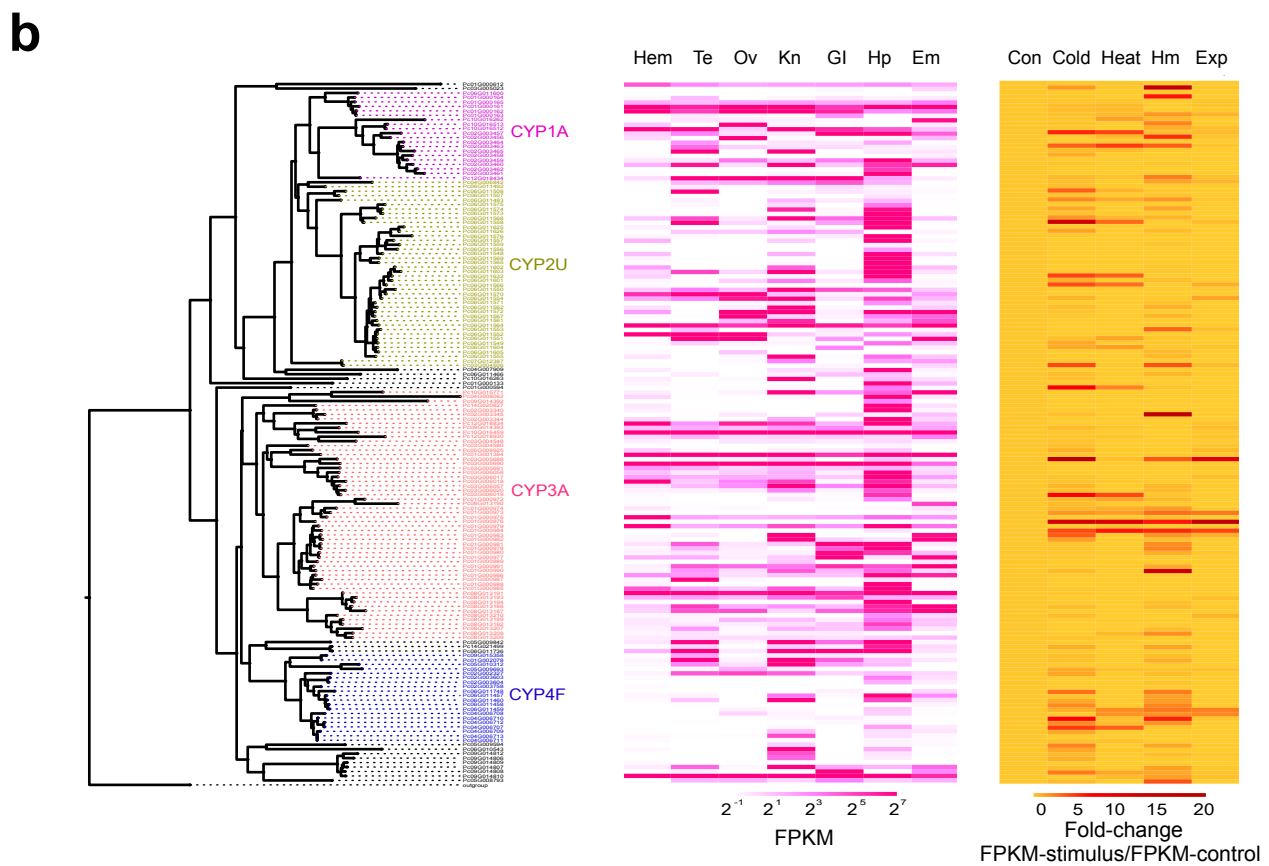
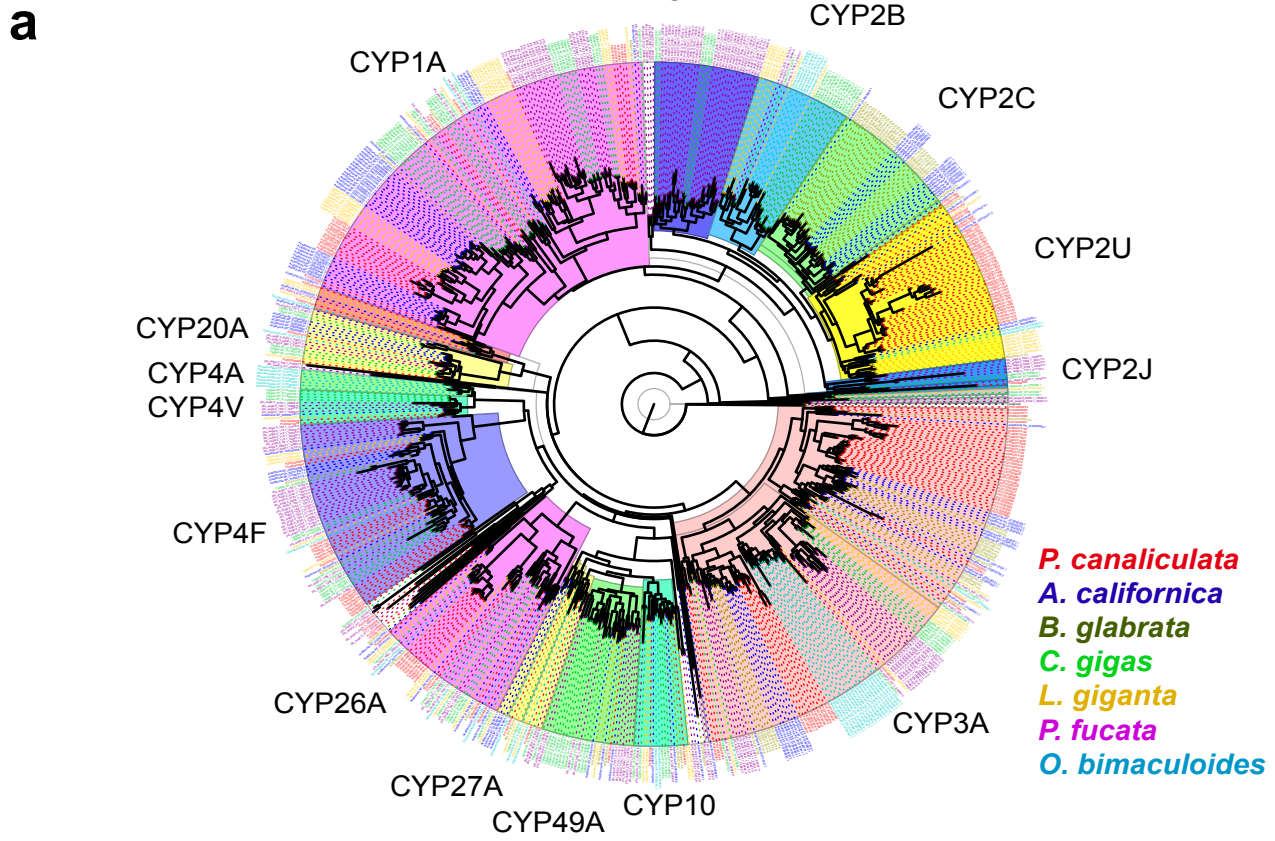
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight
- Formatted: Font: 10.5 pt, Not Highlight

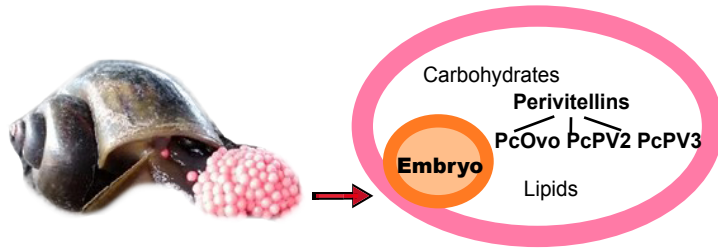
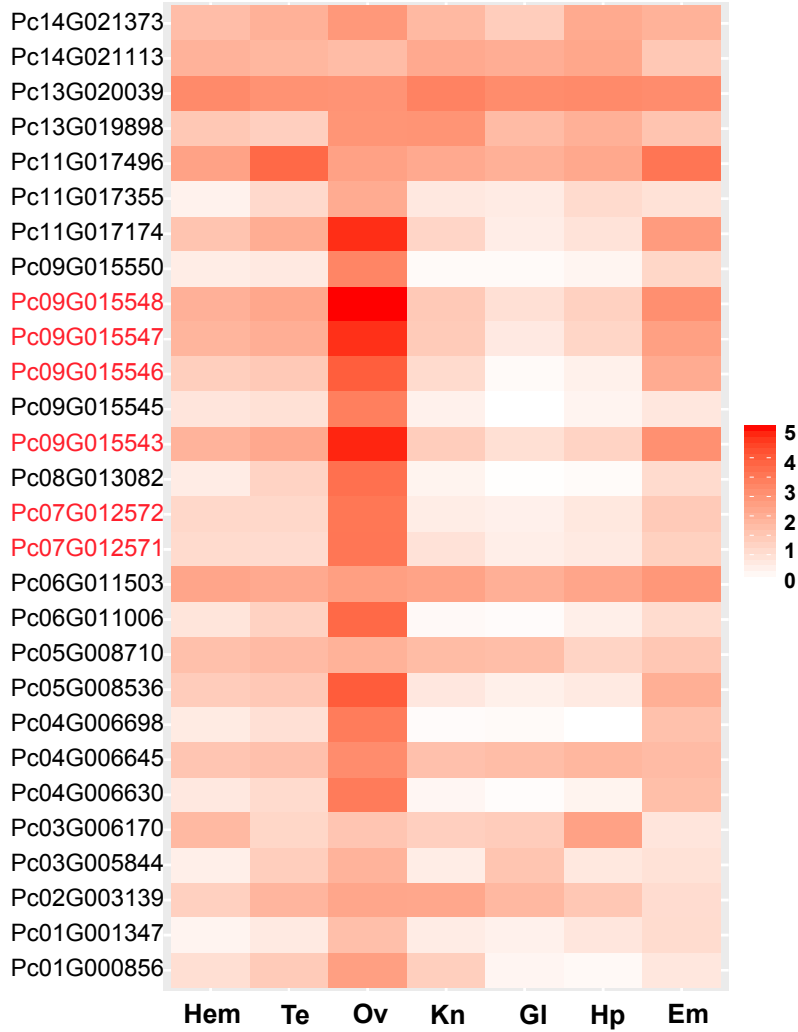
1
2
3
4
5
6
7 | 994 [Sci. 2014;281\(1794\):20141739. doi: 10.1098/rspb.2014.1739.](#)
8
9 | 995 [8286](#). Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler
10 | transform. *Bioinformatics*. 2009;25:1754-60.
11 | 996
12 |
13 | 997 [8387](#). Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile
14 | metagenomic assembler. *Genome Res*. 2017;27:824-34.
15 | 998
16 |
17 | 999 [8488](#). Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. Gene and translation initiation site
18 | prediction in metagenomic sequences. *Bioinformatics*. 2012;28:2223-30.
19 | 1000
20 |
21 | 1001 [8589](#). Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
22 | sequencing data. *Bioinformatics*. 2012;28:3150-2.
23 | 1002
24 |
25 | 1003 [8690](#). Buchfink B, Chao X, Huson DH. Fast and sensitive protein alignment using DIAMOND.
26 | *Nat Methods*. 2015;12:59-60.
27 | 1004
28 |
29 | 1005 [8791](#). Gerlach W and Stoye J. Taxonomic classification of metagenomic shotgun sequences
30 | with CARMA3. *Nucleic Acids Res*. 2011;39 doi:Artn E9110.1093/Nar/Gkr225.
31 | 1006
32 |
33 | 1007 [8892](#). Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for
34 | deciphering the genome. *Nucleic Acids Res*. 2004;32:D277-80.
35 | 1008
36 |
37 | 1009 [8993](#). Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated
38 | carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2012;40:W445-51.
39 | 1010
40 |
41 | 1011 [9094](#). Eddy SR. Accelerated Profile HMM Searches. *Plos Comput Biol*. 2011;7 doi:ARTN
42 | e100219510.1371/journal.pcbi.1002195.
43 | 1012
44 |
45 | 1013 [9195](#). Qin JJ, Li YR, Cai ZM, Li SH, Zhu JF, Zhang F, et al. A metagenome-wide association
46 | study of gut microbiota in type 2 diabetes. *Nature*. 2012;490:55-60.
47 | 1014
48 |
49 | 1015



a**b**





a*P. canaliculata***b**



Click here to access/download
Supplementary Material
Supplemental_Information-final.doc





Click here to access/download

Supplementary Material

Supplemental_Information_with-tracks.doc

