Germline *de novo* mutation rate of the highly heterozygous amphioxus genome

Jing Xue[1,2*], Lei Tao[1,*],Junwei Cao[3], Guang Li[3,#], Cai Li[1,#]

[1]State Key Laboratory of Biocontrol, School of Life Sciences, Guangdong Provincial Key Laboratory for Aquatic Economic Animals, Sun Yat-sen University, Guangzhou, Guangdong, China

[2]Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, Guangdong, China

[3]School of Life Sciences, Xiamen University, Xiamen, Fujian, China

[*]contributed equally

[#]corresponding authors: licai@mail.sysu.edu.cn; guangli@xmu.edu.cn

Abstract

Germline *de novo* mutations (DNMs) are the ultimate source of heritable variation, yet their patterns in highly heterozygous genomes remain poorly understood. Amphioxus, an early-branching chordate with exceptionally high genomic heterozygosity (3.2~4.2% in sequenced species), offers a unique model to explore mutational dynamics in such contexts. It is unclear whether high heterozygosity in amphioxus is due to a large effective population size, an increased mutation rate, or both. Here, we perform deep short-read whole genome sequencing of a two-generation pedigree of the amphioxus *Branchiostoma floridae* comprising two parents and 104 offspring, and develop a framework based on allele-aware parental assemblies as the reference to accurately identify DNMs. We detect 205 high-confidence DNMs, yielding a genome-wide mutation rate of 5.10×10^{-9} per base per generation, which is comparable to that of vertebrates. Combining this estimate with observed nucleotide diversity, we obtain an effective population size of ~1.9 million, indicating that the elevated heterozygosity mainly results from a large effective population size. We observe a maternal-origin bias when considering all DNMs but a paternal-origin bias for early-occurring ones. Amphioxus harbors a much smaller fraction of CpG>TpG DNMs relative to vertebrates, attributable to its low methylation levels. We also investigate putative post-zygotic mutations in the offspring, revealing an unexpected paternal-origin bias. These suggest some distinct mutational mechanisms in amphioxus. Our study not only provides the first DNM measurement for amphioxus but also offers a generalizable strategy for studying DNMs in highly heterozygous genomes, facilitating mutation rate studies across chordates and other lineages.

Introduction

Germline *de novo* mutations (DNMs) serve as the ultimate source of genetic variation and are critical for biological evolution. These mutations arise as a consequence of replication errors or mis-repaired/un-repaired DNA damage. The *de novo* mutation rate, the frequency at which DNMs occur, is an important parameter for many genetic and evolutionary analyses. Because most mutations tend to be deleterious, organisms incur considerable adaptive costs to minimize such errors and damage, thereby maintaining mutation rates at an extremely low level, ranging from 10^{-11} to 10^{-7} per base per generation in eukaryotes (Lynch, et al. 2023; Wang and Obbard 2023). Thanks to the advances in sequencing and computational methods, in the past decade, direct detection of DNMs by sequencing parent-offspring trios has been available for many species, particularly for vertebrates (Besenbacher, et al. 2019; Bergeron, et al. 2021; Wang, et al. 2022; Zhang, et al. 2022; Bergeron, et al. 2023; Lin, et al. 2023; Sendell-Price, et al. 2023; Zhang, et al. 2023; Liang, et al. 2024; Porubsky, et al. 2025; Prentout, et al. 2025; Versoza, et al. 2025; Wang, et al. 2025; Zhang, et al. 2025), providing insight into inter-species mutation rate variation and underlying causes.

Mutation rates also vary across genomic positions, developmental stages, parental origins, and mutation types. For example, in mammals the C>T mutation rate at CpG sites is about one order of magnitude higher than the genome-wide average, which is mainly caused by spontaneous deamination of methylated cytosines (Hodgkinson and Eyre-Walker 2011). A recent study in amniotes found that while both sexes exhibit similar mutation rates early in development, the male germline accumulates more mutations after sexual differentiation (de Manuel, et al. 2022), highlighting the influence of gametogenesis on mutation accumulation. Despite the progress, many details of underlying mutational mechanisms remain uncovered.

Amphioxus (cephalochordates) represents the earliest-branching lineage of extant chordates, diverging from the common ancestor of vertebrates and tunicates over 500 million years ago, serving as a key model for understanding vertebrate origins (Holland and Holland 2017). Although amphioxus shares a basic body plan with vertebrates, it is anatomically simpler, and its genome has undergone fewer lineage-specific rearrangements compared to tunicates (Delsuc, et al. 2006; Putnam, et al. 2008). Notably, amphioxus displays an exceptionally high level of genomic heterozygosity

(3.2~4.2% in sequenced species), among the highest observed in chordates (Bi, et al. 2020). Under the neutral theory of molecular evolution, the equilibrium level of heterozygosity is proportional to both mutation rate and effective population size (Kimura 1983). In the absence of accurate mutation rate measurement, it is unclear whether the high heterozygosity in amphioxus is due to a large effective population size, an increased mutation rate, or both.

High heterozygosity is not unique to amphioxus—many species across diverse clades, such as marine invertebrates (Small, et al. 2007; Ketchum, et al. 2020; Nong, et al. 2020; Shingate, et al. 2020; Penaloza, et al. 2021; Wooldridge, et al. 2025), insects (Mackintosh, et al. 2019; Kim, et al. 2021; Bastide, et al. 2022; Garcia-Berro, et al. 2023; Deng, et al. 2024; Jiang, et al. 2025), and plants (Bredeson, et al. 2016; Tang, et al. 2016; Shen, et al. 2018; Wang, et al. 2020; Xian, et al. 2025), also exhibit elevated genomic diversity. Interestingly, multiple sequenced tunicate species such as sea squirts (Small, et al. 2007; Wei, et al. 2020) and larvaceans (Denoeud, et al. 2010; Bliznina, et al. 2021), which are close relatives of vertebrates, also possess genomes with high heterozygosity. Yet, for these species, direct estimates of the germline mutation rate remain unavailable, largely due to technical challenges in working with highly heterozygous genomes (Garimella, et al. 2020). The high heterozygosity presents significant challenges to key genomic analyses such as genome assembly (Kajitani, et al. 2014; Zhang, et al. 2020), read alignment (Landan and Graur 2009), and variant detection (Gan, et al. 2011; Veeckman, et al. 2019). Indeed, previous research in macaques has shown that even gold-standard pipelines for DNM detection exhibit elevated error rates in highly heterozygous genomes (Bergeron, et al. 2022). As a result, the evolutionary forces shaping high genomic diversity in these lineages remain poorly understood.

Amphioxus, with its exceptionally high heterozygosity and relatively small genomes (~500 Mb) (Huang, et al. 2023), represents an ideal model system for addressing these questions. Direct measurement of DNMs in the amphioxus genome is not only critical for estimating its mutation rate and studying mutational mechanisms, but also highly informative for a broader set of species with extreme genomic diversity.

In this study, we generated deep short-read sequencing data for a two-generation pedigree (two parents and 104 F1 offspring) of *Branchiostoma floridae*, to

systematically study DNMs in the amphioxus. To overcome the challenges in alignment and variant detection, we developed a workflow leveraging the high heterozygosity to assemble the contig-level haplotypes of parental genomes, thereby allowing mapping offspring reads to the parental genomes for identifying DNMs. This new and cost-effective strategy can be generalized to other species with high heterozygosity. With the detected DNMs and estimated mutation rate (5.10×10^{-9} per base per generation), we concluded that the high heterozygosity in amphioxus is in fact caused by the large effective population size. We further investigated multiple aspects of DNMs as well as postzygotic mutations (PZMs) in amphioxus, providing insights into mutational mechanisms in this basal chordate lineage.

Results

An analysis strategy based on parental genome assembly for detecting DNMs

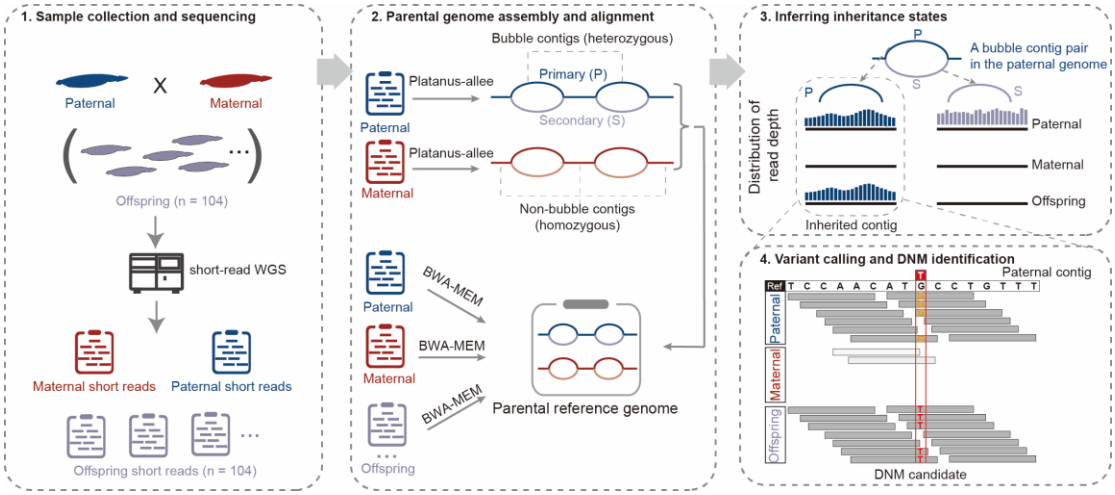
To identify DNMs, we performed deep short-read whole genome sequencing (WGS) for two parental individuals and 104 F1 offspring of a *B. floridae* family, yielding an average coverage of 83× per sample (ranging from 49× to 141×) (**Supplementary Figure 1 and Supplementary Table 1**).

k-mer analysis revealed that the heterozygosity of the parental genomes was approximately 3% (**Supplementary Figure 2**). The high level of heterozygosity in the *B. floridae* genome poses a challenge for conventional DNM detection pipelines based on aligning reads of parents and offspring against a public reference genome (Bergeron, et al. 2022). Because of low sequence similarity between our sequencing data and the public reference genome (Huang, et al. 2023), many reads in our sequencing data cannot be correctly mapped to the reference genome by common short-read aligners (e.g. BWA), leading to unreliable variant calling results. By simulating sequencing reads with ART (Huang, et al. 2012), we estimated that, when the sequence difference between the query genome and the reference genome is 4%, >2% of reads are mapped to wrong genomic regions (**Supplementary Table 2**) and such problematic alignments are difficult to distinguish in real datasets when calling variants.

However, we reasoned that the high heterozygosity of *B. floridae* can facilitate allele-aware assembly of the parental genomes, as many divergent regions in the

143 parental genomes can be assembled into separated contigs. The high sequencing depths
144 of our data enabled us to do contig-level *de novo* assembly for two parents. Using the
145 parental genome assembly as a custom reference, we should be able to align most
146 offspring and parental reads to the parental genomes, thereby obtaining high-quality
147 genotype information of offspring and parents for DNM detection (**Figure 1**).

148 Therefore, we constructed the genome assembly for each parent using the allele-aware
149 assembler Platanus-allee (v2.2.2) (Kajitani, et al. 2019). We obtained a maternal
150 genome assembly of 733.8 Mb and a paternal genome assembly of 763.7 Mb
151 (**Supplementary Table 3**), approximately 1.6 times the length of the previously
152 published haploid genome (490 Mb) (Huang, et al. 2023), indicating that a large
153 fraction of the diploid parental genome was assembled into phased alleles. By the
154 definition of Platanus-allee, heterozygous regions are assembled as "bubble" contigs
155 with two allelic sequences, named as "primary" and "secondary" respectively.
156 Homozygous regions with identical or nearly identical sequences in an individual are
157 assembled as "non-bubble" contigs (**Figure 1**).

158 

159 **Figure 1 Schematic of the workflow for detecting DNMs.** We collected DNA
160 samples from a *B. floridae* family of two parents and 104 F1 offspring and performed
161 deep short-read sequencing (>50×) for all individuals. Using the parental WGS data,
162 we assembled allele-aware parental genomes with Platanus-allee (Kajitani, et al. 2019)
163 and combined the resulting assemblies to form a parental reference genome for
164 downstream analyses. Short reads from both parents and offspring were aligned to this
165 custom reference using BWA-MEM2 (Vasimuddin, et al. 2019). We then computed the
166 read depth distribution for each contig in both parents and offspring. An example is
167 shown for a pair of paternal bubble contigs. Based on the read depth distributions in
168 parents and an offspring, we identified the inherited contig from the father in the
169 offspring. The inheritance states of all parental contigs and read alignments were further

170 used to identify DNMs in each offspring, with an exemplar DNM shown in the bottom
171 right panel.

172

173 By merging the two parental genome assemblies, we generated a composite parental
174 reference genome, with detailed assembly statistics provided in **Supplementary Table**
175 **3**. The parental reference genome had a total length of approximately 1.4 Gb, consisting
176 of 263,358 bubble contigs covering ~856 Mb and 758,976 non-bubble contigs spanning
177 ~780 Mb. The contig N50 metrics of maternal and paternal assemblies were 6.6 kb and
178 7.6 kb, respectively. >91% of sequencing reads of the two parents were aligned to this
179 parental reference genome with proper insert sizes (**Supplementary Figure 1**),
180 indicating that the vast majority of genomic segments were successfully assembled and
181 available for downstream analysis.

182 Next, we aligned sequencing reads from both offspring and parents to this parental
183 reference genome and detected variants using BCFtools (v1.21) (Danecek, et al. 2021)
184 and freebayes (v1.3.10) (Garrison and Marth 2012). A custom filtering pipeline,
185 accounting for factors such as read depth and variant quality, was employed to obtain
186 high-confidence DNMs (see Methods). After applying the filtering criteria, 321
187 candidate DNMs were retained. Among these DNMs, putative false positives were
188 further ruled out by manual check with IGV visualization (see Methods). After careful
189 check, a total of 205 high-confidence DNMs (**Supplementary Table 4**) were retained
190 across 104 offspring, corresponding to an average of 1.97 DNMs per individual. 22
191 DNMs from 9 individuals were selected for orthogonal validation using independent
192 amplicon-seq data. For each selected DNM, 150~200 bp regions encompassing the
193 DNM were amplified and sequenced for both the offspring and parents. The amplicon-
194 seq data confirmed 20 of the 22 selected DNMs as true DNMs (**Supplementary Table**
195 **5-7**), while the remaining two (in a same amplicon) yielded inconclusive results due to
196 amplification noises, indicating a low false positive rate in our DNMs. With an average
197 callable genome size of 398.6 Mb and a false negative rate of 3% estimated by
198 simulation (Methods), the mean genome-wide mutation rate was calculated to be
199 5.10×10^{-9} mutations per base per generation, with a 95% confidence interval (CI) of
200 $4.46\text{-}5.83\times10^{-9}$. The mutation rate of *B. floridae* is comparable to those reported in fish
201 and other vertebrates (Bergeron, et al. 2023) **(Figure 2A)**.

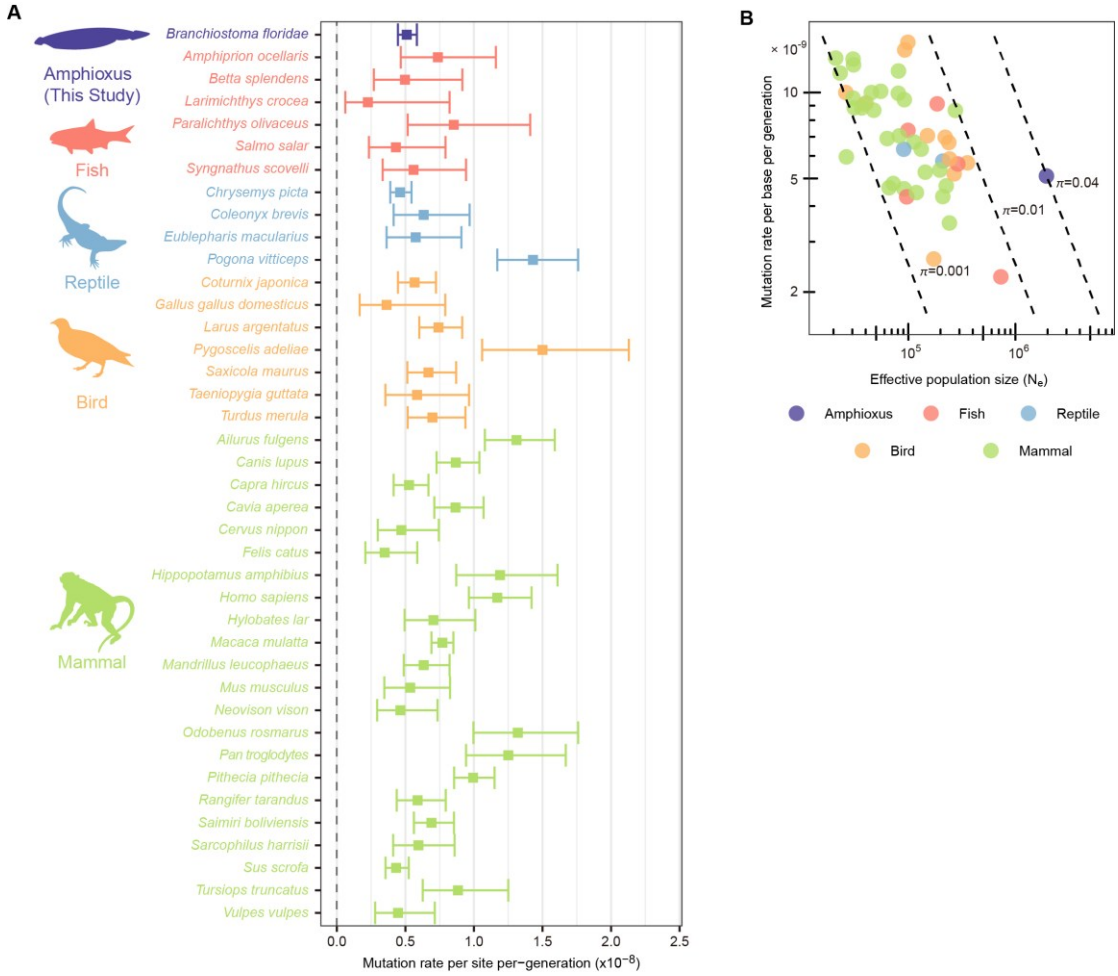


Figure 2 Estimates of mutation rate and effective population size. (**A**) Per base per generation *de novo* mutation rate (μ) estimates from this study and previous studies (Bergeron, et al. 2023). Error bars denote 95% confidence intervals. (**B**) Comparison of mutation rates and effective population sizes across diverse taxonomic groups, using data for other taxa from Bergeron et al. (2023). The dashed lines from left to right represent π values of 0.001, 0.01, and 0.04, respectively. The π values, labeled near each dashed line, were calculated according to $\pi = 4N_e\mu$.

High heterozygosity of *B. floridae* is mainly attributed to its large effective population size

It has long been recognized that the genomic heterozygosity of *B. floridae* is much higher than in other species (Putnam, et al. 2008). Yet, since the mutation rate in *B. floridae* has not been directly quantified, the cause(s) of its high heterozygosity remains uncertain. The nucleotide diversity, expressed as π, can be represented by $\pi = 4N_e\mu$ under neutral evolution (Nei and Tajima 1981; Kimura 1983), where N_e represents the effective population size and μ the germline mutation rate. Therefore, the effective

219 population size can be estimated by the equation $N_e = \pi/(4\mu)$. Based on synonymous
220 substitutions in the population genomic data from Huang et al. (2023), we obtained the
221 π estimate of 0.0397 for *B. floridae* (Methods). For comparison, the π estimates of most
222 vertebrate species from Bergeron et al. (2023) fall into the range of 0.001 to 0.01, with
223 none above 0.01. The effective population size, calculated with our estimated mutation
224 rate (μ) and π, was 1,946,770. We placed *B. floridae*'s N_e and μ within the context of
225 N_e and μ values derived from Bergeron et al. (Bergeron, et al. 2023) to show its unique
226 position (**Figure 2B; Supplementary Table 8**), indicating that *B. floridae*'s N_e is
227 largest among chordates with measurements. We note that, the sequenced population
228 in Huang et al. (2023) experienced some degree of inbreeding in the laboratory, which
229 could lead to lower nucleotide diversity than wild populations. Therefore, the actual N_e
230 in wild populations is likely larger than the current estimate. Collectively, we concluded
231 that the high heterozygosity of amphioxus is not due to an elevated mutation rate but a
232 very large effective population size.

233 **Characteristics of DNMs in *B. floridae*.**

234 Distinguishing the parental origins of DNMs is crucial for understanding mutational
235 dynamics. While previous studies could only assign parental origin for 30%~80% of
236 DNMs based on nearby polymorphic markers (Bergeron, et al. 2023; Wooldridge, et al.
237 2025), our DNM calling method inherently provided parental origin information for all
238 detected DNMs. Among the 205 DNMs, 105 originated from the mother and 100 from
239 the father (**Figure 3A**), yielding a male mutation bias coefficient (α) of 0.95.
240 Considering the difference in the callable genome sizes between the parents, we
241 obtained a corrected α of 0.88 (95% CI = 0.83-0.91), indicating a maternal bias. For a
242 cross-species comparison, we integrated several representative vertebrates with known
243 α values (**Figure 3B)**. The results showed that the α value of *B. floridae* is lower than
244 most vertebrates. Previous studies (Bergeron, et al. 2023; Zhang, et al. 2025) reported
245 that fish tend to have apparently lower α values than other vertebrates (mammals, birds
246 and reptiles). These studies suggested that differences in gametogenesis such as
247 seasonal breeding and synchronous ovulation could lead to the lower α values in fish.
248 The low α of amphioxus might be partly due to these reasons.

249 Due to the fragmented nature of the parental genome assemblies, direct genomic
250 annotation was challenging; therefore, we mapped the parental genome coordinates to

the published haploid reference genome from Huang et al. (Huang, et al. 2023) and used its annotation for downstream functional analyses. Of the total 490 Mb of the public haploid reference genome, 363 Mb (74%) was covered by the assembled contigs of parental genomes. The parental contigs that were mapped to the haploid reference genome were defined as mappable regions with annotation information. Based on the haploid genome annotation (Huang, et al. 2023), the genome was divided into three gene annotation contexts: exonic, intronic, and intergenic regions. We found no significant difference between the fractions of paternal and maternal DNMs for each gene annotation context (**Figure 3C**).

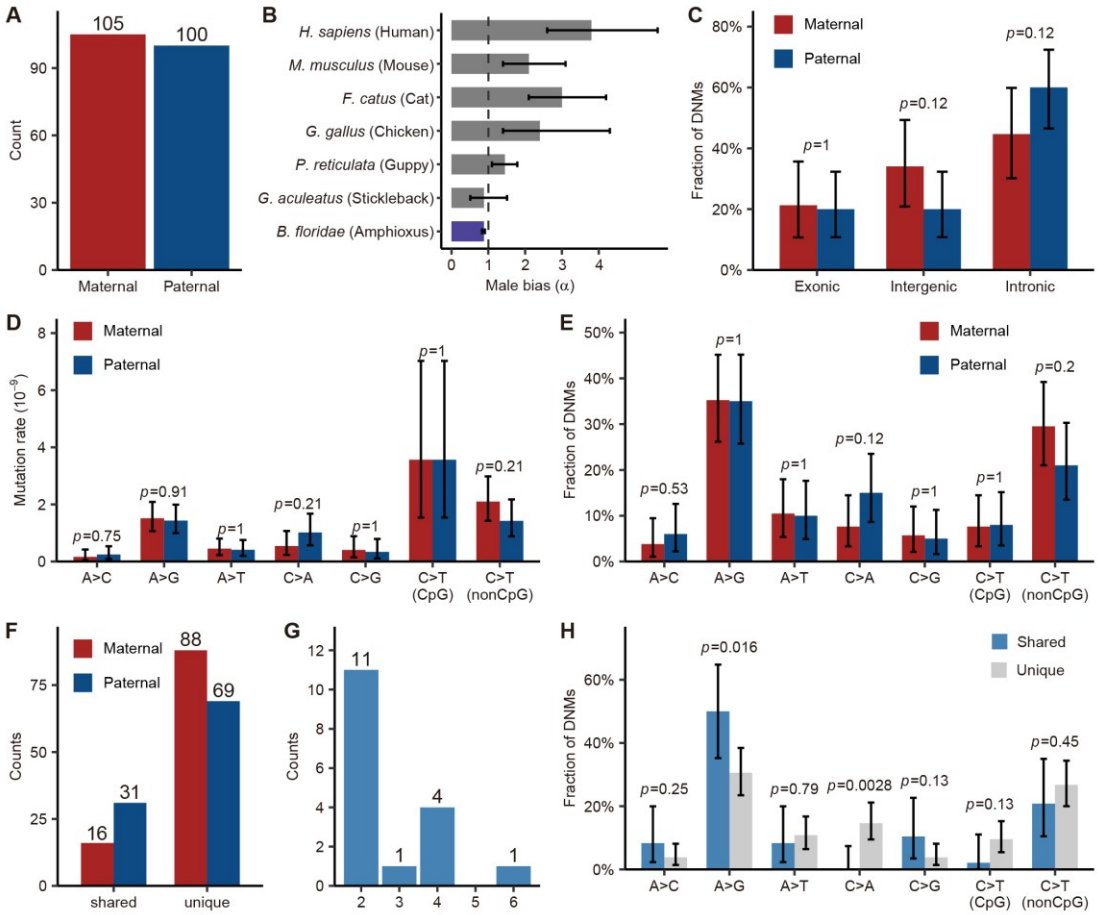


Figure 3 Characteristics of DNMs in *B. floridae*. (A) Parental origins of DNMs. **(B)** The male-to-female ratios (α) of DNMs for different chordates. **(C)** Proportions of DNMs in different gene annotation contexts, separating paternal (blue) and maternal (red) DNMs. Parental origin differences were assessed by Fisher's exact tests. **(D)** Mutation rates of different substitution types, separating paternal (blue) and maternal (red) DNMs. Parental origin differences assessed by χ^2 tests. The mutation spectrum was folded (e.g., A>C representing both A>C and T>G mutations). **(E)** Proportions of DNMs of different substitution types, separating paternal (blue) and maternal (red) DNMs. Parental origin differences were assessed by Fisher's exact tests. **(F)** Counts of

sibling-shared and unique DNMs, separating paternal (blue) and maternal (red) DNMs. **(G)** Counts of sibling-shared DNMs shared by different numbers of offspring. **(H)** Proportions of DNMs of different substitution types, separating shared and unique DNMs. Differences between shared and unique DNMs were assessed by Fisher's exact tests. Error bars denote 95% confidence intervals computed using binomial distributions. All p-values for Fisher's exact tests and χ^2 tests are given above the bars.

We further investigated DNM counts of different substitution types. The results indicated no significant difference between maternal and paternal mutation rates for all substitution types (**Figure 3D**), and no significant difference between parental origins when looking at proportions of DNMs of different substitution types (**Figure 3E**). In vertebrates, the mutation rate of C>T at CpG sites is usually much higher than that at non-CpG sites (e.g., >10 fold difference in humans) (Kong, et al. 2012; Francioli, et al. 2015). However, the mutation rate of C>T at CpG sites was only about twice that at non-CpG sites in *B. floridae* (**Figure 3D**). Cross-species comparison of mutational spectra (**Supplementary Figure 3**) revealed some unique patterns in *B. floridae*. CpG>TpG mutations only accounted for 8% of DNMs in *B. floridae*, significantly lower than those (16-20%) observed in vertebrates. On the other hand, A>G mutations accounted for 35% of DNMs in *B. floridae*, consistently higher than those (16-26%) in surveyed vertebrates. The lower mutation frequency at CpG sites was probably due to lower methylation levels at CpG sites in amphioxus (Marletaz, et al. 2018).

47 (23%) DNMs were shared among siblings (named ssDNMs, **Figure 3F**), which likely occurred at early developmental stages. A greater number of ssDNMs were identified as paternal in origin, with 31 ssDNMs derived from the father compared to only 16 from the mother. The male bias α of 1.80 (after accounting for callable size difference), was substantially higher than the α of 0.88 observed for all DNMs (Fisher's exact test, $p = 0.036$), suggesting that during earlier stages of germ cell development, male germline accumulates more mutations relative to female germline. Majority of ssDNMs were shared by only two offspring, with one shared by up to six offspring (**Figure 3G**). Comparison of mutation spectra indicated a significant increase in the proportion of A>G mutations in ssDNMs relative to unique DNMs, but a depletion of C>A mutations in ssDNMs, likely reflecting the differences in mutational processes between early and late gametogenesis (**Figure 3H**).

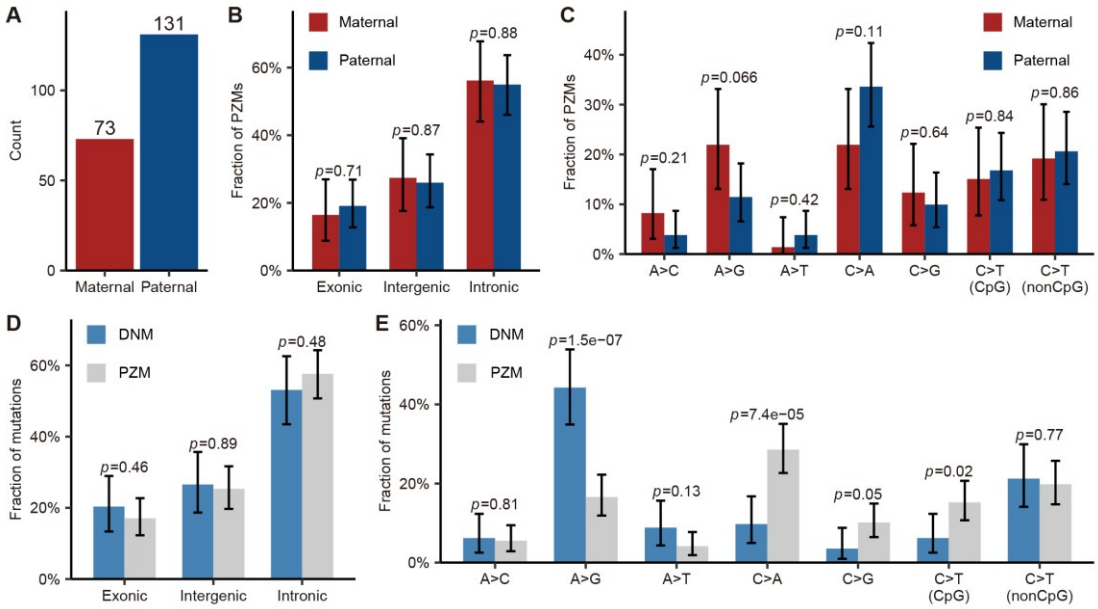


Figure 4 Characteristics of PZMs. (**A**) Parental origins of PZMs. (**B**) Proportions of PZMs in different gene annotation contexts, separating paternal (blue) and maternal (red) PZMs. Parental origin differences were assessed by Fisher's exact tests. (**C**) Proportions of PZMs of different substitution types, separating paternal (blue) and maternal (red) PZMs. (**D**) Proportions of mutations in different gene annotation contexts, separating DNMs (blue) and PZMs (grey). (**E**) Proportions of mutations of different substitution types, separating DNMs (blue) and PZMs (grey). Parental origin differences and differences between DNMs and PZMs were assessed by Fisher's exact tests, with p-values above the bars. Error bars denote 95% confidence intervals computed based on binomial distributions.

Characteristics of PZMs in *B. floridae*

Postzygotic mutations (PZMs) arise during mitotic cell divisions in the developing embryo or later in life, and thus typically affect only a subset of the body's cells. Some PZMs are present in the germ cells can be passed to the next generation. The high sequencing depths of our data also enabled us to detect PZMs in offspring (see Methods for details). We identified 204 putative PZMs across 104 offspring (**Figure 4A; Supplementary Table 9**), yielding an average of 1.96 PZMs per individual. The variant allele frequencies of PZMs ranged from 0.05 to 0.45 (**Supplementary Table 9**). As the detected PZMs had relatively high allele frequencies, they probably arose during early development.

Unexpectedly, PZMs in *B. floridae* exhibited a pronounced paternal origin bias (α =

327 1.67, corrected for callable size difference; binomial test, p = 6.9e-07; **Figure 4A**). The
328 sex-biased pattern of PZMs, which has not been observed in other species to our
329 knowledge, suggested different mutational processes operating on paternal and
330 maternal chromosomes during early development. In mammals, early divisions of
331 embryos erase epigenetic memory (primarily DNA methylation and histone
332 modifications) from sperm and egg and re-establish the epigenomic environment. Yet
333 in some vertebrates such as zebrafish and sea lamprey (Potok, et al. 2013; Angeloni, et
334 al. 2024), the methylation landscape in the paternal genome of early embryos is largely
335 maintained and the maternal methylation is reprogrammed to match the paternal.
336 Available methylome data suggested that similar patterns are likely to be present in sea
337 squirt and sea urchin (Xu, et al. 2019). While there was no study comparing methylomes
338 of gametes and early embryos in amphioxus, we suspected that amphioxus may employ
339 similar epigenetic reprograming like that in sea lamprey, given its phylogenetic position.
340 The epigenomic differences between paternal and maternal chromosomes in early
341 embryos might contribute to different frequencies of paternal and maternal PZMs.

342 We further examined differences in parental origin with respect to proportions of PZMs
343 across gene annotation contexts and across mutation types, but found no significant
344 difference (**Figure 4B-C**). Similarly, comparison between PZMs and DNMs in gene
345 annotation contexts revealed no notable difference (**Figure 4D**). However, analysis of
346 specific mutation types revealed that A>G mutations were significantly more frequent
347 in DNMs than in PZMs. Conversely, C>A mutations were markedly enriched in PZMs
348 relative to DNMs, and C>G and CpG>TpG mutations were also more prevalent in
349 PZMs (**Figure 4E**). Previous methylome analysis revealed that amphioxus embryos
350 exhibit more methylated CpGs relative to adult tissues, which may partly explain the
351 relatively higher frequency of CpG>TpG mutations in PZMs (Marletaz, et al. 2018).
352 While the parental origin ratio of PZMs was close to that of ssDNMs (PZM: 1.67;
353 ssDNM: 1.80), their mutation spectra were largely different (**Figure 3H** versus **Figure**
354 **4E**), suggesting different underlying mutational mechanisms.

355

356 **Discussion**

357 In this study, we generated high-depth short-read sequencing data for a pedigree
358 comprising both parents and 104 offspring to investigate DNMs in amphioxus. Due to

high heterozygosity of amphioxus, which results in many sequence differences between the public reference genome and re-sequenced individuals, conventional alignment and variant calling methods that rely on the public reference genome are prone to errors, thereby complicating DNM detection. However, high heterozygosity can also facilitate allele-aware diploid genome assembly without requiring long-read data. With this consideration, we assembled haplotype-resolved diploid genomes at the contig level for each parent and developed a custom read mapping and variant calling pipeline using the parental assemblies as the reference. This cost-effective strategy can help resolve the challenges of detecting DNMs in highly heterozygous species like *B. floridae*. It also allowed us to assign the parent-of-origin state for all detected DNMs. Furthermore, based on phased genotypes of parents and offspring, we were also able to comprehensively investigate the meiotic recombination landscape, which has been described in a companion study (Tao, et al. 2025).

With an average callable diploid genome size of 398.6 Mb, we identified 205 DNMs and estimated the germline mutation rate in *B. floridae* to be 5.10×10^{-9} (95% CI of $4.46\text{-}5.83 \times 10^{-9}$) per base per generation. Based on this mutation rate, the effective population size of *B. floridae* was estimated to be about two million, suggesting that the high heterozygosity resulted from a massive effective population size. Our estimated mutation rate for contemporary populations is close to a recently reported estimate (4.65×10^{-9}) based on divergence times and intra-cephalochordate synonymous substitution rates of single-copy genes (Ren, et al. 2025). Our study did not estimate the germline mutation rate for small insertions/deletions (INDELs), as generally poor alignments around these variants (usually associated with repeats) make it challenging to accurately detect *de novo* INDELs. In addition, repetitive or low-complexity regions, which usually have relatively high mutation rates, are difficult to assemble with only short reads and thus were under-represented in the callable regions of our study. Future studies incorporating long-read sequencing technologies (e.g., PacBio or Oxford Nanopore) can facilitate identifying *de novo* INDELs as well as other DNMs in complex genomic regions.

We further analyzed the parental origin, genomic distribution, mutation spectrum, and ssDNMs in *B. floridae*, revealing some distinctive features in this basal chordate. We observed a maternal bias of DNMs, with a male-to-female mutation ratio (α) of 0.88.

391 This ratio was lower than most reported values for vertebrates, likely associated with
392 seasonal breeding and synchronous ovulation (de Manuel, et al. 2022; Bergeron, et al.
393 2023). The overall mutation spectra showed no obvious difference between different
394 parental origins. The low frequency of C>T transitions at CpG sites was consistent with
395 previously reported low levels of CpG methylation in amphioxus (Marletaz, et al. 2018).
396 The presence of ssDNMs provided a valuable window into early germline mutational
397 events. We observed more ssDNMs in the paternal genome relative to the maternal, but
398 the underlying cause is unclear. Future DNM studies spanning different ages and
399 developmental stages of parents can help clarify the mechanisms of mutagenesis in
400 amphioxus. In addition, functional genomics studies focusing on DNA damage and
401 repair pathways are also essential to understanding the mutational mechanisms.

402 We identified 204 putativePZMs across 104 offspring, with a surprisingly strong
403 paternal-origin bias ($\alpha = 1.67$). As the bias was unlikely due to artifacts after careful
404 check, we speculated that the paternal bias of PZMs is likely linked to underlying
405 epigenomic asymmetries between parental genomes in early embryos—such as
406 incomplete or asymmetric reprogramming of DNA methylation—which may lead to
407 different susceptibility to mutations. Further functional genomic analysis is needed to
408 test this hypothesis. While PZMs and DNMs showed similar distributions across gene
409 annotation contexts, PZMs were significantly enriched for C>A and CpG>TpG
410 mutations, whereas germline DNMs more frequently involved A>G mutations. As most
411 PZMs are somatic mutations, the distinct mutation spectra between PZMs and DNMs
412 highlighted different biological mechanisms governing mutagenesis in somatic versus
413 germline lineages.

414 In summary, through high-depth pedigree sequencing and the development of a novel
415 analysis pipeline, we directly quantified the germline mutation rate in *B. floridae* for
416 the first time and resolved the long-standing question of what causes its high
417 heterozygosity. We also explored DNMs from multiple perspectives, including parental
418 origin, gene annotation context, mutation spectrum and shared DNMs among siblings.
419 Some of the unique DNM features observed in amphioxus may represent ancestral
420 characteristics of germline mutagenesis in chordates. The analysis framework based on
421 assembling parental genomes for read alignment and variant calling can facilitate
422 related research in other highly heterozygous species.

Materials and Methods

Amphioxus culture

Amphioxus (*B. floridae*) strains obtained from a stock maintained by Dr. Jr-Kai Yu originating from Tampa, Florida, were housed in a temperature-controlled facility at Xiamen University, which maintained a 14 h light and 10 h dark cycle and ambient seawater temperature at 19°C (Li, et al. 2012). A pair of parental individuals was selected for crossing. Fertilized eggs were subsequently cultured following the cultivation strategy described previously until they reached the designated sampling stage. A total of 104 F1 offspring were generated through this process. To capture distinct developmental time points, the offspring were snap-frozen in liquid nitrogen at two specific post-growth intervals: 6 months and 1 year. The first batch, comprising 35 individuals, was preserved at the 6-month time point, and the second batch, consisting of 71 individuals, was preserved at the 1-year time point.

Library construction and sequencing

We extracted high molecular weight genomic DNAs from the muscle tissues of a single individual using the DNeasy Blood & Tissue Kit (QIAGEN, Valencia, CA), and inspected the DNA quality by Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA) and 2100 Agilent Bioanalyzer (Agilent). After sample QC, gDNA was fragmented by ultrasound on Covaris E220 (Covaris, Brighton, UK), then selected to 300-500 bp using magnetic beads size selection. The selected DNA fragments were then end-repaired and A-tailed. The indexed adaptors were ligated to both ends of the DNA fragments. The ligation product was amplified by PCR, and circularized to get single-stranded circular (ssCir) library. The ssCir library was then amplified through rolling circle amplification (RCA) to obtain DNA nanoball (DNB). The DNB was then loaded to flowcell, and sequenced by DNBSEQ Platform (MGI, China).

Parental genome assembly

Genome heterozygosity of parental individuals was estimated via *k*-mer analysis of Illumina short-read data using GenomeScope (v2.0) (Ranallo-Benavidez, et al. 2020). Raw short-read sequences were quality-processed with fastp (v0.24.0) (Chen 2023), implementing: (1) elimination of reads with >10% ambiguous bases ('N'), and (2)

454 retention of reads exhibiting a minimum mean Phred quality score of 20 (Q20). Adapter
455 sequences were concurrently trimmed. These processed reads were subjected to allele-
456 aware diploid genome assembly using Platanus-allee (v2.2.2) (Kajitani, et al. 2019)
457 through a two-stage workflow: First, the *de novo* assembly phase was executed with
458 the command "assemble -m 200 -t 20"; second, haplotype resolution was performed
459 via the phase module using "phase -i 3 -t 20". Contigs shorter than 150 bp in length
460 were subsequently filtered from both parental phased assemblies. Initial assembly
461 quality metrics, including contig N50, total length, and GC content, were
462 comprehensively evaluated with QUAST (v5.3.0) (Gurevich, et al. 2013) and SeqKit
463 (v2.9.0) (Shen, et al. 2016).

464 Following the phased assembly procedure, contigs underwent automated classification
465 by Platanus-allee. In the contig-level assemblies generated by Platanus-allee, contigs
466 were classified into two distinct categories: (1) bubble contigs harboring allelic variants,
467 which were further stratified as primary contigs (longer allelic sequences)
468 and secondary contigs (shorter counterparts), and (2) non-bubble contigs designated by
469 the assembler as lacking allelic diversity. However, through comprehensive whole-
470 genome alignment analyses, we discovered that a subset of non-bubble contigs
471 exhibited unexpected sequence overlaps with other contigs across the assembly,
472 suggesting residual heterozygous regions unresolved by the phasing algorithm.

473 Finally, the filtered phased diploid contig sets derived from paternal and maternal
474 genomes were concatenated to construct a custom reference genome (hereafter termed
475 parental reference genome), which were used in many downstream analyses.

476 **Alignment of offspring and parental reads to the parental reference genome and**
477 **variant calling**

478 Processed short reads from two parents and offspring were aligned to the parental
479 reference genome assembly using BWA-MEM2 (v2.2.1) (Vasimuddin, et al. 2019)
480 with parameters "-M -k 19". Post-alignment processing entailed coordinate-based
481 sorting and removal of PCR duplicates via SAMtools (v1.21) (Li, et al. 2009) and
482 Picard Tools in GATK4 (v4.6.1.0) (Van der Auwera and O'Connor 2020) with default
483 parameters. Per-base depths across the whole reference genome and individual-level
484 average sequencing depth metrics were subsequently computed using mosdepth
485 (v0.3.10) (Pedersen and Quinlan 2018) with default parameters.

Based on the read alignments above, we used BCFtools (v1.21) (Danecek, et al. 2021) and freebayes (v1.3.10) (Garrison and Marth 2012) to identify small variants. First, we generated "pileup" files for the set of all samples by running "bcftools mpileup" with parameters "-annotate FORMAT/AD, FORMAT/ADF, FORMAT/ADR, FORMAT/DP, FORMAT/SP, INFO/AD, INFO/ADF, INFO/ADR" on all input bam files. With the "pileup" files as input, we ran "bcftools call -mv" to generate VCF files, which contain SNVs and short insertions/deletions, for 104 offspring and two parents. We additionally used freebayes (v1.3.10) (Garrison and Marth 2012) to identify variants within similar sequences between the two parents with parameters "-m 0 --legacy-gls." In regions exhibiting high sequence similarity between the two parents, only the freebayes calls were retained; in all other regions, variant calls from both tools were used.

Inferring inheritance states of parental contigs in offspring genomes based on read alignments against the parental reference genome

According to the principles of Mendelian inheritance, genomic reads in the offspring are derived from the gene pool of two parental genomes. As described in previous sections, we obtained the paternal and maternal diploid genome assemblies as sets of haplotype-resolved contigs. By mapping sequencing reads of an offspring back to the parental reference genome (combining paternal and maternal genome assemblies), we can determine which allele of a specific locus (bubble or non-bubble contigs) of a given parent is inherited by the offspring. More specifically, to determine whether a particular contig from a given parent is inherited by a specific offspring, we compared read depth distribution profiles of that contig between the offspring and each parent. For a contig with specific parental origin, the depth distribution of the offspring's reads aligned to this contig is expected to more closely resemble the depth distribution of the reads from the parent-of-origin rather than that from the opposite parent.

The workflow for inferring inheritance states is described below and illustrated in **Supplementary Figure 4**. To quantify the depth distribution along a contig, each contig was divided into 50 bp bins, discarding any terminal bin shorter than 50 bp. The mean depth within each bin was calculated for each individual, generating a depth distribution vector per individual across the contig. The similarity between offspring and parental vectors was quantified using a weighted measure that combines (1) Cosine

518 similarity for bins with non-zero depth in both individuals being compared, and (2)
519 Jaccard similarity for bins with zero coverage in at least one of the two individuals
520 (accounting for regions of no aligned reads). Weights of different similarity scores were
521 allocated based on corresponding proportions of zero-value bins and non-zero-value
522 bins. A parental contig was considered inherited by an offspring if its depth distribution
523 vector in that offspring showed greater similarity to the vector of its specific parental
524 source than to the other parent. Parental contigs exhibiting a depth less than 1/4 of an
525 offspring's mean depth or a depth less than 10 were classified as low-depth and thus
526 non-inherited in the specific offspring.

527 However, the similarity of some contigs between the parental genomes, probably
528 arising from shared ancestry, could confound the inheritance calls derived from the
529 offspring-parent depth distribution similarity algorithm (**Supplementary Figure 5**).
530 Since direct pairwise genome sequence alignment of two parental genomes is difficult
531 for assessing parental contig similarities due to potential length disparities of
532 orthologous contigs in different parental assemblies, we used the similarity of read
533 depth distributions (calculated using the same algorithm employed for offspring-parent
534 comparisons) when mapping each parent's reads to the same contig as a robust proxy.
535 If the depth distribution similarity between the two parents for a specific contig is
536 greater than 0.5, the inheritance calls for that contig in the offspring are deemed
537 unreliable due to indistinguishable parental origins, necessitating further validation
538 with additional data.

539 Because a given diploid offspring can inherit only one allele from each parent at a locus,
540 determining the inheritance status of one contig within a pair of bubble contigs from a
541 parent inherently resolves the inheritance status of both contigs in the pair. In other
542 words, the inheritance of one contig in a bubble pair can be inferred based on the
543 inheritance of its counterpart. Therefore, we further took advantage of the very low read
544 depths in some loci to conclude that an allele was not inherited; under such
545 circumstances, the alternative allele with aligned reads must be inherited by the
546 offspring. In contrast, for non-bubble contigs, the inability to identify a second allele
547 precludes this method of assessing similarity between parental contigs, and therefore
548 these contigs were excluded from further analysis. Finally, all scenarios arising from
549 the combinations of inheritance and contig similarity, which lead to differences in depth

550 distributions in offspring, were comprehensively illustrated in **Supplementary Figure**
551 **5**.

552 Based on the inheritance states of parental contigs in each offspring, we in fact have
553 reconstructed the offspring genomes with phased alleles in the majority of regions. Such
554 phased offspring genomes can facilitate downstream analyses.

555 **Determining the callable genome**

556 Estimating mutation rates requires determining the length of genomic regions within
557 each offspring in which DNMs can be confidently identified, defined as the callable
558 genome size. In each offspring, the callable genome is primarily composed of the
559 inherited genomic segments defined in the previous section. However, the accurate
560 inference of the inheritance of non-bubble contigs with high similarity between the two
561 parents proved challenging. For non-bubble contigs lacking pre-existing allelic data in
562 the parental genomes, the allele-depth-distribution-based method for inferring contig
563 inheritance patterns is inapplicable. Consequently, determining the inheritance of
564 homologous contigs within non-bubble contigs becomes challenging. We therefore
565 implemented stricter filtering criteria, excluding non-bubble contigs with parental depth
566 distribution similarity ≥0.5 from the callable genome regions.

567 Regions with high sequence similarity between parental alleles impede the variant
568 detection and phasing in our pipeline. In this study, the parents were derived from an
569 inbred population, and thus their offspring could inherit highly similar or identical
570 alleles from parents in some regions. To address this, we developed a pipeline to
571 identify pairs of parental contigs with high sequence similarity. Briefly, we performed
572 pairwise genome alignments using Minimap2 (v2.24) (Li 2021) with parameters "-cx
573 asm20 --secondary=no", designating the paternal genome as the reference and the
574 maternal genome as the query, and *vice versa*. Reciprocal best-hit contig pairs were
575 taken as evidence of homology. For each pair, we further assessed the depth distribution
576 similarity based on alignments of parental reads to assembled contigs; a pair of contigs
577 with a similarity score >0.5 was classified as a highly similar allele pair. For each
578 offspring, if both contigs of a highly similar allele pair were inherited by the offspring,
579 they were excluded from the callable genome for that offspring, as DNMs in these
580 contigs could not be reliably phased and could be confounded by PZMs. In contrast,
581 when only one homologous contig from either parent of a highly similar allele pair was

582 inherited, the contig was retained in the callable genome with the freebayes-called
583 variants.

584 Additionally, we removed sites in repetitive regions, as they are prone to erroneous
585 alignments and often exhibit abnormally high alignment depths. Sites with too many
586 reads may represent problematic repetitive regions (Li 2014). Specifically, based on the
587 observed depth distribution in each offspring, any regions with a read depth exceeding
588 three times the individual's average were excluded. The terminal regions of contigs (100
589 bp at each end) were also excluded from the callable genome, because these regions
590 tended to enrich assembly and alignment errors.

591 Finally, by integrating these filters based on both inheritance and sequencing depth, we
592 obtained the callable genome size for each offspring ranging from 334.11 Mb to 463.06
593 Mb, accounting for approximately 34-47% of the diploid genome size (**Supplementary**
594 **Table 10**).

595 **Identifying the candidate DNMs**

596 For each parent-offspring trio, the variants in three VCF files of each trio were filtered
597 to a subset of single nucleotide variants (SNVs) by BCFtools (v1.21) (Danecek, et al.
598 2021) and freebayes (v1.3.10) (Garrison and Marth 2012).

599 Variants were retained only if they met all of the following criteria: (1) the variant was
600 located within the predefined callable genome regions; (2) the variant had homozygous
601 alternative allele status in the offspring's alignment against the contig with read
602 depth >10, and read depth of the inherited parent at the variant site of the contig was
603 required to be >5; (3) there were at most two alternative alleles at the variant site in
604 each parental BAM file; (4) the variant had a minimum QUAL score of 220, ensuring
605 high-confidence base calls (BCFtools only); (5) no proximity to indels (outside ±20 bp
606 of indels); (6) variants not at the sites experiencing putative gene conversions described
607 in a companion study (Tao, et al. 2025), as gene conversions lead to read alignments
608 resembling DNMs; (7) for a candidate DNM detected on a specific parental contig in
609 an offspring, all other offspring did not contain a heterozygous genotype at the same
610 site of the contig, as heterozygosity implies an inherited variant or alignment errors.

611 After applying these filters, the resulting set of candidate DNMs was further evaluated
612 by importing both parental and offspring BAM files into IGV (v2.19.2)

613 (Thorvaldsdottir, et al. 2013), and using IGVtools to generate screenshots for each
614 candidate DNM in both the offspring and the parents. Subsequent manual inspection
615 led to the removal of any candidate DNMs if they met one or more of the following
616 criteria: (1) structural variant/recombination breakpoint neighborhoods, and (2)
617 variants were found to arise from misassigned parental origin of contigs, upon manual
618 review. These steps yielded a DNM dataset for each individual (**Supplementary Table**
619 **4**). The genome-wide mutation rate was calculated by the number of all DNMs divided
620 by the sum of callable genome sizes of all offspring. The callable genome size was not
621 multiplied by two, since the reconstructed inherited genomic segments effectively
622 represent an approximately diploid genome for each offspring.

623 Because the coordinates of DNMs were located on the parental genomic contigs, the
624 parental origin of the contig harboring each DNM naturally revealed its source,
625 allowing for unequivocal determination of the DNM's parental origin without any
626 further processing. This same principle applied to the callable genome as well.

627 **Simulation for assessing alignment errors if using the public reference genome**

628 To assess alignment errors due to low sequence similarity between our sequencing data
629 and the public *B. floridae* reference genome, we performed a simulation analysis.
630 Artificial genomes were generated by introducing varying proportions of substitutions
631 into the published *B. floridae* reference. Paired-end 150 bp reads were then simulated
632 from these artificial genomes using ART (Huang, et al. 2012) with parameters "-p -m
633 450 -s 50 -f 10 -l 150 -ef -ss HS25". The simulated reads were aligned to the original
634 *B. floridae* reference using BWA-MEM2 (v2.2.1) (Vasimuddin, et al. 2019), and mis-
635 mapped reads were quantified with a custom script. The results were summarized in
636 **Supplementary Table 2**.

637 **Simulation for assessing the false negative rate of DNM detection**

638 For evaluating the false negative rate (FNR) in our DNM detection pipeline, we
639 generated simulated data based on callable genome regions of three selected offspring.
640 For each of the three individuals, we made an artificial genome based on the callable
641 genome regions of that individual and put a specific number of simulated DNMs within
642 the artificial genome. We next generated 80× simulated paired-end 150 bp reads for
643 each selected individual using ART (Huang, et al. 2012) and re-ran the full alignment

644 and DNM detection workflow. The detected DNMs were then compared with the
645 known positions of simulated DNMs to estimate the FNR (the equation below) for each
646 selected offspring independently. The final FNR value was reported as the average
647 across the three simulations (details given in **Supplementary Table 11**).

648 $$FNR = 1 - \frac{n_{Detected\ DNM}}{n_{Simulated\ DNM}} \quad (1)$$

649 ***De novo* mutation rate estimation**

650 The per-site-per-generation mutation rate (μ) was calculated as:

651 $$\mu = \frac{n_{DNM\ Candidate}}{Length_{Callable\ genome} \times (1 - FNR)} \quad (2)$$

652 The callable genome size ($Length_{Callable\ genome}$) was independently determined for each
653 offspring using the procedures described above. The FNR of the DNM detection
654 pipeline was estimated from the simulation described above. All candidate DNMs were
655 visually reviewed in IGV (v2.19.2) (Thorvaldsdottir, et al. 2013), and a subset was
656 further validated by Amplicon-seq. These quality-control steps ensured high confidence
657 in DNM identification. As all putative false positive DNMs were removed, we did not
658 apply false discovery rate (FDR) correction for mutation rate estimates.

659 **Analysis of nucleotide diversity in amphioxus populations**

660 To estimate nucleotide diversity in *B. floridae* populations, we re-analyzed previously
661 published population data. First, we downloaded whole-genome sequencing data of 35
662 *B. floridae* individuals (Huang, et al. 2023) from NCBI's SRA database. The short-read
663 alignment and variant calling largely followed the method of Huang et al. (Huang, et
664 al. 2023), with key adaptations. Specifically, the *B. floridae* genome assembly of that
665 study (Huang, et al. 2023) was utilized as the reference, with reads mapped using BWA-
666 MEM2 (v2.2.1) (Vasimuddin, et al. 2019) under default parameters optimized for
667 vertebrate genomes. We then generated gVCF files for each individual using GATK4
668 (v4.6.1.0) (Van der Auwera and O'Connor 2020) HaplotypeCaller, setting the "-ERC
669 BP_RESOLUTION" to include non-variant sites in the output, and adjusting
670 parameters to reflect the heterozygosity of amphioxus by parameters "--heterozygosity
671 to 0.03 and --indel-heterozygosity to 0.003". The sequence alignment and variant
672 detection workflows were integrated into an automated nf-core/sarek pipeline (v3.4.0)

673 (Garcia, et al. 2020) with key parameters fine-tuned, thereby maximizing the utilization
674 of parallel computing resources.

675 For variant sites, we performed initial filtering using GATK4 VariantFiltration on SNPs
676 and INDELs independently, excluding SNPs with QUAL<30, QD<2, FS>60, MQ<40,
677 MQRankSum<-12.5, Read PosRankSum<-8 or SOR>3 and excluding INDELs with
678 QUAL<30, QD<2, FS>200, ReadPosRankSum<-20, and SOR>10 (Mirchandani, et al.
679 2024). For invariant sites, we filtered based on quality ("QUAL>30") and the fraction
680 of missing genotypes at a site ("F_MISSING<0.25").

681 To estimate nucleotide diversity at synonymous sites (π_S), we used the script
682 dNdSpiNpiS (v1.0) (https://kimura.univ-montp2.fr/PopPhyl/index.php?section=tools)
683 with the default parameters. To extract phase information for the CDS features, we used
684 GenomeTools (gt gff3 -sort -tidy -retainids). Based on the VCF file, we then generated
685 the diploid fasta sequence for each individual using the FastaAlternateReferenceMaker
686 module from GATK (v4.6.1.0) (Van der Auwera and O'Connor 2020). We then
687 intersected these fasta sequences with the gene annotation GFF file to keep only CDS
688 with SNPs. For each CDS, individuals with missing data were excluded when
689 estimating π_S. As for π, these estimates were corrected to account for truly invariant
690 sites versus those resulting from insufficient sequencing depth. We used the invariant
691 sites generated above to estimate the expected percentage of bases called per CDS and
692 applied this measure as a correction factor for the π_S estimates. The mean diversities
693 π_S of the CDS were then calculated for the whole genome.

694 **Analysis of the effective population size**

695 The average long-term N_e of *B. floridae* was estimated using the formula $N_e = \pi/(4\mu)$
696 (Kimura 1983), where μ was the mutation rate estimated by our work and π_S was
697 estimated with data from a previous study (Huang, et al. 2023) (**Supplementary Table**
698 **8**). For comparison, μ and π data of vertebrates were from Bergeron et al. (Bergeron, et
699 al. 2023).

700 **Aligning parental genome assemblies to the published *B. floridae* genome**

701 Owing to inherent limitations of short-read sequencing, the assembly of parental
702 genomes consisted of primarily short contigs. Defining orthologous relationships
703 between contigs of two parents in a specific locus based solely on pairwise alignments

704 of contigs, as previously mentioned, can be complicated by the variability in the lengths
705 of assembled contigs derived from that locus. To address these issues, we aligned the
706 contigs of parental genomes to a published *B. floridae* haploid reference genome
707 (Huang, et al. 2023). After testing multiple alignment tools under various parameter
708 configurations, Minimap2 (v2.24) (Li 2021) was selected as the optimal aligner, with
709 empirically optimized parameters: "-x asm20 --secondary=no". Resultant alignments
710 were processed through sequential refinement: (1) coordinate sorting via SAMtools
711 (v1.21) (Li, et al. 2009) and (2) removal of multi-mapping regions using Sambamba
712 (v0.8.2) (Tarasov, et al. 2015). A custom script was employed to establish the mapping
713 relationship between each parental genome and the haploid reference genome. This
714 approach yielded a multiple genome alignment that aligned positions across the paternal,
715 maternal, and reference genomes. As a result, the majority of parental genomic
716 positions were successfully anchored onto the reference, thereby obtaining orthologous
717 genomic coordinates across the parental genomes and facilitating subsequent analyses
718 dependent on chromosomal coordinates.

719 **Comparative analysis of mutation rates in different contexts**

720 To compare mutation rates in different contexts (e.g., different annotation groups), we
721 utilized the coordinate mapping relationships described in the previous section. We
722 mapped the gene functional annotations from the published *B. floridae* reference
723 genome (Huang, et al. 2023) onto the parental genomes. In doing so, the coordinates
724 for exonic, intronic, and intergenic regions were converted to those of the parental
725 genomes. This allowed us to determine the lengths of each gene feature where DNMs
726 and callable genome regions were present, enabling the calculation of annotation-
727 specific mutation rates for each individual.

728 To generate the mutation spectrum, the DNMs were also stratified into six groups
729 according to their mutation types (A>C/T>G, A>T/T>A, A>G/T>C, C>A/G>T,
730 C>G/G>C and C>T/G>A), with a special group for C>T mutations at CpG sites. We
731 extracted CpG positions from the parental genomes, thereby allowing us to distinguish
732 between CpG and non-CpG sites.

733 **Identifying the candidate PZMs**

734 Raw VCF files of parent-offspring trios were obtained from the section "Alignment of

735 offspring and parental reads to the parental reference genome and variant calling".
736 PZMs were retained only if they met all the following criteria: (1) the variant was
737 located within the predefined callable genome regions; (2) the offspring harbored two
738 alleles for a specific site in a parental contig with a total read depth > 10; (3) there were
739 at most two alleles at the variant site in each parental BAM file; (4) the variant with no
740 proximity to indels (outside ±20 bp of indels); (5) the variant was not located in the
741 regions experiencing putative gene conversions; (6) excluding candidate PZMs where
742 the derived allele had ⩾2 reads in any other individual.

743 For each candidate PZM, IGV screenshots were generated for both offspring and
744 parents using the method outlined in the "Identifying the candidate DNMs" section.
745 Subsequent manual inspection led to the removal of any candidate PZMs if they met
746 one or more of the following criteria (1) structural variant/recombination breakpoint
747 neighborhoods; (2) variants were found to arise from misassigned parental origins of
748 contigs, upon manual review.

749 **Amplicon-seq validation**

750 To assess the reliability of DNMs across different data sources, we randomly selected
751 a subset (**Supplementary Table 7**) for validation by amplifying target regions with
752 specific primers, followed by amplicon sequencing for genotyping. The validation
753 workflow comprised three main steps: primer design, *in silico* PCR, and multiplex PCR
754 amplification coupled with amplicon sequencing.

755 In the primer design phase, due to the high sequence polymorphism observed in *B.*
756 *floridae* (~4%), it was challenging to simultaneously amplify parental alleles using a
757 single pair of primers targeting conserved regions; therefore, it became necessary to
758 design allele-specific primers for each allele from the two parents. Specifically, we
759 adopted the framework of the NGS-PrimerPlex (v1.3.4) (Kechin, et al. 2020) multiplex
760 PCR primer design process, with modifications tailored to our specific requirements.
761 Since we needed to amplify all relevant alleles independently in offspring and parents,
762 we chose the parental reference genome as the reference for primer design to maximize
763 the number of alleles amplified by a single primer pair. For each locus, we used the
764 mapping relationships between parental genomes, as described in the section "Aligning
765 parental genome assemblies to the published *B. floridae* genome". This allowed us to

766 extract the target alleles harboring the DNMs and the orthologous alleles from the other
767 parent. To avoid software misinterpretation of homologous regions as multiple
768 alignments, we removed non-target contigs from the reference genome. We then used
769 the processed parental reference genome for each individual together with the DNM
770 positions to run NGS-PrimerPlex (parameters given in the **Supplementary Table 5**).
771 After executing the pipeline, we selected the top five multiplex PCR primer sets with
772 the highest primer scores as candidate primer combinations for the subsequent steps
773 (**Supplementary Table 6**).

774 *In silico* PCR was employed to confirm that primers designed for multiplex
775 amplification can effectively amplify all parental alleles at the loci harboring selected
776 DNMs, using *in_silico_pcr.pl* (v0.6, https://github.com/egonozer/in_silico_pcr). To
777 optimize efficiency, only the contig harboring the DNM and orthologous contigs were
778 extracted as the reference genome for *in silico* PCR. If *in silico* PCR showed incomplete
779 allele amplification, we designed additional primers for the unamplified contig. These
780 primers were added to the multiplex primer set, and the process was repeated until all
781 target fragments could be amplified.

782 For multiplex PCR, we used the gDNA templates of offsprings and parents stored
783 previously and performed PCR amplification with a multiplex PCR kit (VAZYME, CN:
784 PM101-02) optimized for both sensitivity and specificity; the resulting products are
785 then subjected to agarose gel electrophoresis, and finally, DNA fragments ranging from
786 150 to 200 bp were excised from agarose gel following electrophoresis and
787 subsequently purified using gel extraction kit. Library quality was assessed by fragment
788 analysis using the Qsep-400 and quantified for concentration using the Qubit 3.0
789 Fluorometer (Thermo Fisher Scientific, Waltham, MA). Library construction was
790 performed using the VAHTS® Universal Plus DNA Library Prep Kit for Illumina
791 ND617, following the manufacturer's protocol. Libraries meeting the quality control
792 criteria were subsequently sequenced on the Illumina NovaSeq 6000 platform.

793

794 **Data Availability**

795 The raw sequence data reported in this paper have been deposited in the Genome
796 Sequence Archive (GSA: CRA027444) that are publicly accessible at

797 https://ngdc.cncb.ac.cn/gsa. The raw sequence data of previously sequenced
798 populations can be accessed under the BioProject with accession number:
799 PRJNA602496. Genome assemblies produced by Platanus-allee and the amplicon
800 sequencing data have been deposited in Science Data Bank (ScienceDB) repository
801 (https://doi.org/10.57760/sciencedb.27648). Custom scripts used are available online at
802 https://github.com/xuejinghub/Amphioxus_denovo_mutation_rate.

803

804 **Author contributions**

805 **Jing Xue**: Investigation, Methodology, Formal analysis, Writing - Original draft,
806 Writing - Review & Editing. **Lei Tao**: Investigation, Methodology, Formal analysis,
807 Writing - Review & Editing. **Junwei Cao**: Data Curation, Investigation. **Guang Li**:
808 Conceptualization, Data Curation, Resources, Writing - Review & Editing, Supervision.
809 **Cai Li**: Conceptualization, Resources, Methodology, Writing - Review & Editing,
810 Supervision.

811

812 **Conflict of interest**

813 The authors have declared no competing interests.

814

815 **Acknowledgments**

816 We thank Dr. Luohao Xu for his assistance in obtaining previous genomic data of
817 amphioxus populations. We thank Dr. Chung-I Wu and Dr. Luohao Xu for their
818 comments on the manuscript. This work was supported by National Natural Science
819 Foundation of China (32470690), Natural Science Foundation of Fujian Province of
820 China (2022J06004), State Key Laboratory of Biocontrol and Guangdong Provincial
821 Key Laboratory for Aquatic Economic Animals.

822

823 **References**

824 Angeloni A, Fissette S, Kaya D, Hammond JM, Gamaarachchi H, Deveson IW, Klose RJ, Li
825 W, Zhang X, Bogdanovic O. 2024. Extensive DNA methylome rearrangement during

826 early lamprey embryogenesis. Nat Commun 15:1977.
827 Bastide H, Lopez-Villavicencio M, Ogereau D, Lledo J, Dutrillaux AM, Debat V, Llaurens V.
828 2022. Genome assembly of 3 Amazonian Morpho butterfly species reveals Z-chromosome
829 rearrangements between closely related species living in sympatry. Gigascience
830 12:giad033.
831 Bergeron LA, Besenbacher S, Bakker J, Zheng J, Li P, Pacheco G, Sinding MS, Kamilari M,
832 Gilbert MTP, Schierup MH, et al. 2021. The germline mutational process in rhesus
833 macaque and its implications for phylogenetic dating. Gigascience 10:giab029.
834 Bergeron LA, Besenbacher S, Turner T, Versoza CJ, Wang RJ, Price AL, Armstrong E, Riera
835 M, Carlson J, Chen HY, et al. 2022. The Mutationathon highlights the importance of
836 reaching standardization in estimates of pedigree-based germline mutation rates. Elife
837 11:e73577.
838 Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, Hoffman JI, Li Z, St
839 Leger J, Shao C, et al. 2023. Evolution of the germline mutation rate across vertebrates.
840 Nature 615:285-291.
841 Besenbacher S, Hvilsom C, Marques-Bonet T, Mailund T, Schierup MH. 2019. Direct
842 estimation of mutations in great apes reconciles phylogenetic dating. Nat Ecol Evol 3:286-
843 292.
844 Bi C, Lu N, Han T, Huang Z, Chen JY, He C, Lu Z. 2020. Whole-genome resequencing of
845 twenty *Branchiostoma belcheri* individuals provides a brand-new variant dataset for
846 *Branchiostoma*. Biomed Res Int 2020:3697342.
847 Bliznina A, Masunaga A, Mansfield MJ, Tan Y, Liu AW, West C, Rustagi T, Chien HC, Kumar
848 S, Pichon J, et al. 2021. Telomere-to-telomere assembly of the genome of an individual
849 *Oikopleura dioica* from Okinawa using Nanopore-based sequencing. BMC Genomics
850 22:222.
851 Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, Grimwood J,
852 Schmutz J, Rabbi IY, Egesi C, et al. 2016. Sequencing wild and cultivated cassava and
853 related species reveals extensive interspecific hybridization and genetic diversity. Nat
854 Biotechnol 34:562-570.
855 Chen S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication
856 using fastp. Imeta 2:e107.
857 Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T,
858 McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools.
859 Gigascience 10:giab008.
860 de Manuel M, Wu FL, Przeworski M. 2022. A paternal bias in germline mutation is widespread
861 in amniotes and can arise independently of cell division numbers. Elife 11:e80008.

862 Delsuc F, Brinkmann H, Chourrout D, Philippe H. 2006. Tunicates and not cephalochordates
863 are the closest living relatives of vertebrates. Nature 439:965-968.
864 Deng Y, Ren S, Liu Q, Zhou D, Zhong C, Jin Y, Xie L, Gu J, Xiao C. 2024. A high
865 heterozygosity genome assembly of *Aedes albopictus* enables the discovery of the
866 association of PGANT3 with blood-feeding behavior. BMC Genomics 25:336.
867 Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, Mikhaleva J, Olsen
868 LC, Jubin C, Canestro C, et al. 2010. Plasticity of animal genome architecture unmasked
869 by rapid evolution of a pelagic tunicate. Science 330:1381-1385.
870 Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, Genome of the Netherlands
871 C, van Duijn CM, Swertz M, Wijmenga C, et al. 2015. Genome-wide patterns and
872 properties of *de novo* mutations in humans. Nat Genet 47:822-826.
873 Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ,
874 Osborne EJ, Sreedharan VT, et al. 2011. Multiple reference genomes and transcriptomes
875 for *Arabidopsis thaliana*. Nature 477:419-423.
876 Garcia-Berro A, Talla V, Vila R, Wai HK, Shipilina D, Chan KG, Pierce NE, Backstrom N,
877 Talavera G. 2023. Migratory behaviour is positively associated with genetic diversity in
878 butterflies. Mol Ecol 32:560-574.
879 Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, DiLorenzo S, Sandgren J, Diaz
880 De Stahl T, Ewels P, et al. 2020. Sarek: A portable workflow for whole-genome
881 sequencing analysis of germline and somatic variants. F1000Res 9:63.
882 Garimella KV, Iqbal Z, Krause MA, Campino S, Kekre M, Drury E, Kwiatkowski D, Sa JM,
883 Wellems TE, McVean G. 2020. Detection of simple and complex *de novo* mutations with
884 multiple reference sequences. Genome Res 30:1154-1169.
885 Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing.
886 arXiv preprint arXiv:1207.3907.
887 Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for
888 genome assemblies. Bioinformatics 29:1072-1075.
889 Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian
890 genomes. Nat Rev Genet 12:756-766.
891 Holland ND, Holland LZ. 2017. The ups and downs of amphioxus biology: a history. Int J Dev
892 Biol 61:575-583.
893 Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator.
894 Bioinformatics 28:593-594.
895 Huang Z, Xu L, Cai C, Zhou Y, Liu J, Xu Z, Zhu Z, Kang W, Cen W, Pei S, et al. 2023. Three
896 amphioxus reference genomes reveal gene and chromosome evolution of chordates. Proc
897 Natl Acad Sci U S A 120:e2201504120.

898 Jiang SD, Chen YA, Guo PY, Sun J, Feng HY, Li Y, Wang JJ, Wei DD. 2025. Chromosome-
899 level genome of the parthenogenetic booklouse *Liposcelis bostrychophila* reveals high
900 heterozygosity and a nonhomologous chromosome. Insect Sci 32:437-456.
901 Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M,
902 Nagayasu E, Maruyama H, et al. 2014. Efficient *de novo* assembly of highly heterozygous
903 genomes from whole-genome shotgun short reads. Genome Res 24:1384-1395.
904 Kajitani R, Yoshimura D, Okuno M, Minakuchi Y, Kagoshima H, Fujiyama A, Kubokawa K,
905 Kohara Y, Toyoda A, Itoh T. 2019. Platanus-allee is a *de novo* haplotype assembler
906 enabling a comprehensive access to divergent heterozygous regions. Nat Commun
907 10:1702.
908 Kechin A, Borobova V, Boyarskikh U, Khrapov E, Subbotin S, Filipenko M. 2020. NGS-
909 PrimerPlex: High-throughput primer design for multiplex polymerase chain reactions.
910 PLoS Comput Biol 16:e1008468.
911 Ketchum RN, Smith EG, DeBiasse MB, Vaughan GO, McParland D, Leach WB, Al-Mansoori
912 N, Ryan JF, Burt JA, Reitzel AM. 2020. Population genomic analyses of the sea urchin
913 *Echinometra* sp. EZ across an extreme environmental gradient. Genome Biol Evol
914 12:1819-1829.
915 Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D,
916 D'Agostino ERR, Pelaez J, et al. 2021. Highly contiguous assemblies of 101 drosophilid
917 genomes. Elife 10:e78579.
918 Kimura M. 1983. The neutral theory of molecular evolution. Cambridge Cambridgeshire ; New
919 York: Cambridge University Press.
920 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA,
921 Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of *de novo* mutations and the
922 importance of father's age to disease risk. Nature 488:471-475.
923 Landan G, Graur D. 2009. Characterization of pairwise and multiple sequence alignment errors.
924 Gene 441:141-147.
925 Li G, Yang X, Shu Z, Chen X, Wang Y. 2012. Consecutive spawnings of Chinese amphioxus,
926 *Branchiostoma belcheri*, in captivity. PLoS One 7:e50838.
927 Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37:4572-
928 4574.
929 Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage
930 samples. Bioinformatics 30:2843-2851.
931 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
932 Genome Project Data Processing S. 2009. The sequence alignment/map format and
933 SAMtools. Bioinformatics 25:2078-2079.

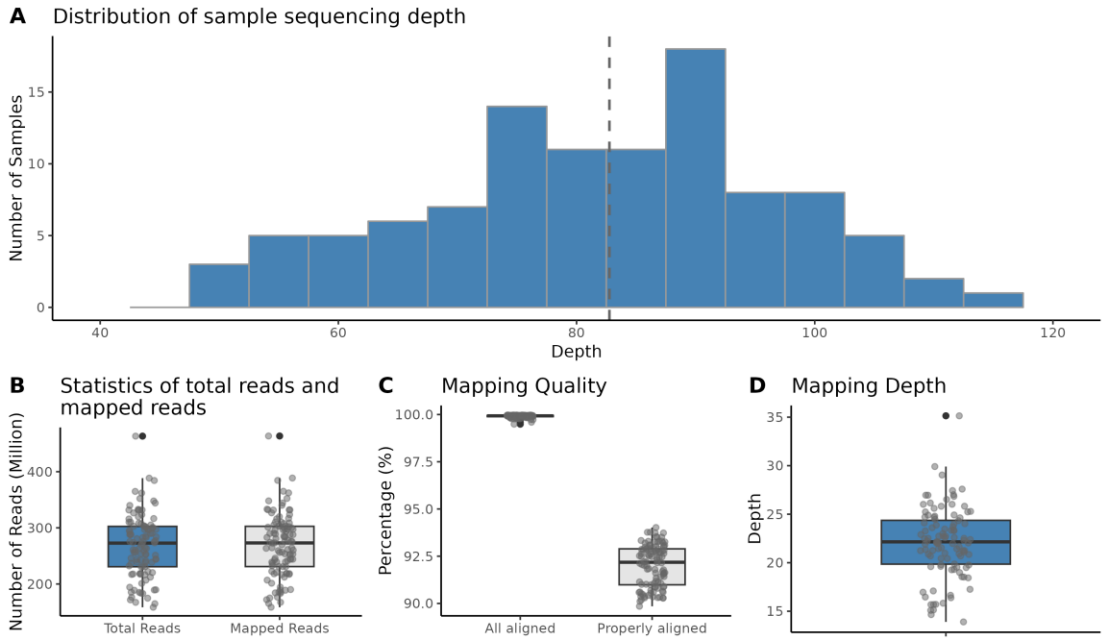
934 Liang X, Yang S, Wang D, Knief U. 2024. Characterization and distribution of *de novo*
935 mutations in the zebra finch. Commun Biol 7:1243.
936 Lin Y, Darolti I, van der Bijl W, Morris J, Mank JE. 2023. Extensive variation in germline *de*
937 *novo* mutations in *Poecilia reticulata*. Genome Res 33:1317-1324.
938 Lynch M, Ali F, Lin T, Wang Y, Ni J, Long H. 2023. The divergence of mutation rates and
939 spectra across the Tree of Life. EMBO Rep 24:e57561.
940 Mackintosh A, Laetsch DR, Hayward A, Charlesworth B, Waterfall M, Vila R, Lohse K. 2019.
941 The determinants of genetic diversity in butterflies. Nat Commun 10:3466.
942 Marletaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, Wyatt CDR, de la Calle-
943 Mustienes E, Bertrand S, Burguera D, et al. 2018. Amphioxus functional genomics and
944 the origins of vertebrate gene regulation. Nature 564:64-70.
945 Mirchandani CD, Shultz AJ, Thomas GWC, Smith SJ, Baylis M, Arnold B, Corbett-Detig R,
946 Enbody E, Sackton TB. 2024. A fast, reproducible, high-throughput variant calling
947 workflow for population genomics. Mol Biol Evol 41:e80008.
948 Nei M, Tajima F. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics
949 97:145-163.
950 Nong W, Cao J, Li Y, Qu Z, Sun J, Swale T, Yip HY, Qian PY, Qiu JW, Kwan HS, et al. 2020.
951 Jellyfish genomes reveal distinct homeobox gene clusters and conservation of small RNA
952 processing. Nat Commun 11:3051.
953 Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and
954 exomes. Bioinformatics 34:867-868.
955 Penaloza C, Gutierrez AP, Eory L, Wang S, Guo X, Archibald AL, Bean TP, Houston RD.
956 2021. A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*.
957 Gigascience 10:giab020.
958 Porubsky D, Dashnow H, Sasani TA, Logsdon GA, Hallast P, Noyes MD, Kronenberg ZN,
959 Mokveld T, Koundinya N, Nolan C, et al. 2025. Human *de novo* mutation rates from a
960 four-generation pedigree reference. Nature 643:427-436.
961 Potok ME, Nix DA, Parnell TJ, Cairns BR. 2013. Reprogramming the maternal zebrafish
962 genome after fertilization to match the paternal methylation pattern. Cell 153:759-772.
963 Prentout D, Bykova D, Hoge C, Hooper DM, McDiarmid CS, Wu F, Griffith SC, de Manuel
964 M, Przeworski M. 2025. Germline mutation rates and fine-scale recombination parameters
965 in zebra finch. PLoS Genet 21:e1011661.
966 Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi
967 M, Shoguchi E, Terry A, Yu JK, et al. 2008. The amphioxus genome and the evolution of
968 the chordate karyotype. Nature 453:1064-1071.
969 Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for

970 reference-free profiling of polyploid genomes. Nat Commun 11:1432.
971 Ren Y, Miao Z, Lin C, Yang L, Li H, Holland L, Cho S, Yu J, Yue J. 2025. The genome of the
972 early diverged amphioxus, *Asymmetron lucayanum*, illuminates the evolution of genome
973 architecture and gene repertoires in cephalochordates. bioRxiv:2025.2007.2021.665930.
974 Sendell-Price AT, Tulenko FJ, Pettersson M, Kang D, Montandon M, Winkler S, Kulb K,
975 Naylor GP, Phillippy A, Fedrigo O, et al. 2023. Low mutation rate in epaulette sharks is
976 consistent with a slow rate of evolution in sharks. Nat Commun 14:6628.
977 Shen Q, Zhang L, Liao Z, Wang S, Yan T, Shi P, Liu M, Fu X, Pan Q, Wang Y, et al. 2018.
978 The genome of *Artemisia annua* provides insight into the evolution of Asteraceae family
979 and artemisinin biosynthesis. Mol Plant 11:776-788.
980 Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q
981 file manipulation. PLoS One 11:e0163962.
982 Shingate P, Ravi V, Prasad A, Tay BH, Garg KM, Chattopadhyay B, Yap LM, Rheindt FE,
983 Venkatesh B. 2020. Chromosome-level assembly of the horseshoe crab genome provides
984 insights into its genome evolution. Nat Commun 11:2322.
985 Small KS, Brudno M, Hill MM, Sidow A. 2007. Extreme genomic variation in a natural
986 population. Proc Natl Acad Sci U S A 104:5698-5703.
987 Tang C, Yang M, Fang Y, Luo Y, Gao S, Xiao X, An Z, Zhou B, Zhang B, Tan X, et al. 2016.
988 The rubber tree genome reveals new insights into rubber production and species adaptation.
989 Nat Plants 2:16073.
990 Tao L, Xue J, Cao J, Li G, Li C. 2025. Direct detection of meiotic recombination events in the
991 highly heterozygous amphioxus genome. bioRxiv:2025.2007.2014.664623.
992 Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS
993 alignment formats. Bioinformatics 31:2032-2034.
994 Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-
995 performance genomics data visualization and exploration. Brief Bioinform 14:178-192.
996 Van der Auwera GA, O'Connor BD. 2020. Genomics in the Cloud: Using Docker, GATK, and
997 WDL in Terra: O'Reilly Media.
998 Vasimuddin M, Misra S, Li H, Aluru S editors. 2019 IEEE International Parallel and
999 Distributed Processing Symposium (IPDPS). 2019 20-24 May 2019.
1000 Veeckman E, Van Glabeke S, Haegeman A, Muylle H, van Parijs FRD, Byrne SL, Asp T,
1001 Studer B, Rohde A, Roldan-Ruiz I, et al. 2019. Overcoming challenges in variant calling:
1002 exploring sequence diversity in candidate genes for plant development in perennial
1003 ryegrass (*Lolium perenne*). DNA Res 26:1-12.
1004 Versoza CJ, Ehmke EE, Jensen JD, Pfeifer SP. 2025. Characterizing the Rates and Patterns of
1005 De Novo Germline Mutations in the Aye-Aye (Daubentonia madagascariensis). Mol Biol

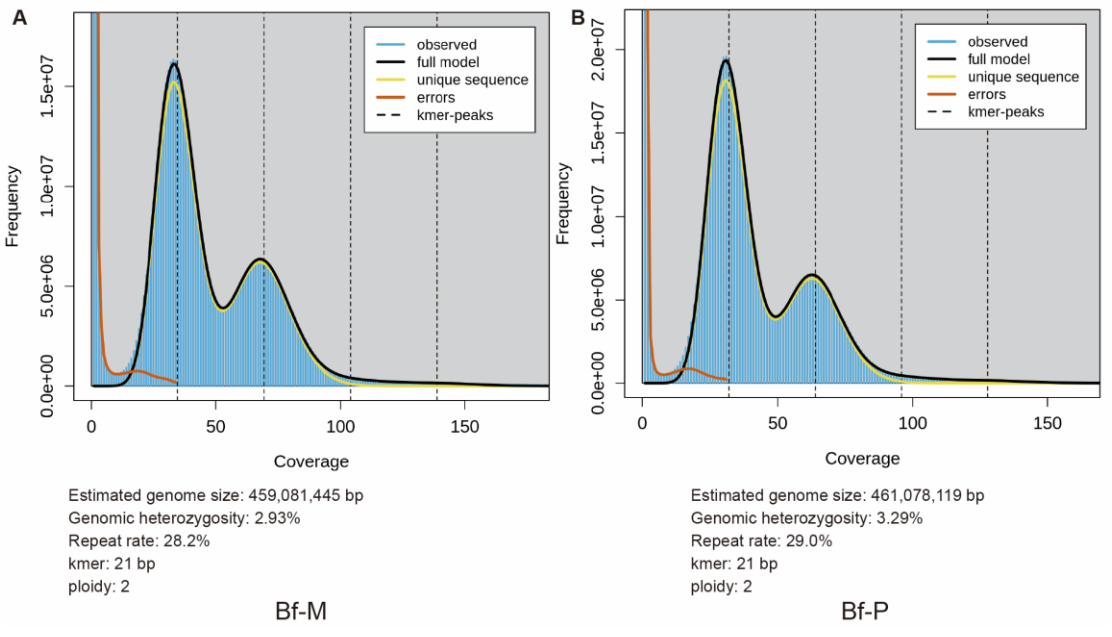
Evol 42.
Wang RJ, Pena-Garcia Y, Raveendran M, Harris RA, Nguyen TT, Gingras MC, Wu Y, Perez L, Yoder AD, Simmons JH, et al. 2025. Unprecedented female mutation bias in the aye-aye, a highly unusual lemur from Madagascar. PLoS Biol 23:e3003015.
Wang RJ, Raveendran M, Harris RA, Murphy WJ, Lyons LA, Rogers J, Hahn MW. 2022. *De novo* mutations in domestic cat are consistent with an effect of reproductive longevity on both the rate and spectrum of mutations. Mol Biol Evol 39:msac147.
Wang X, Feng H, Chang Y, Ma C, Wang L, Hao X, Li A, Cheng H, Wang L, Cui P, et al. 2020. Population sequencing enhances understanding of tea plant evolution. Nat Commun 11:4447.
Wang Y, Obbard DJ. 2023. Experimental estimates of germline mutation rate in eukaryotes: a phylogenetic meta-analysis. Evol Lett 7:216-226.
Wei J, Zhang J, Lu Q, Ren P, Guo X, Wang J, Li X, Chang Y, Duan S, Wang S, et al. 2020. Genomic basis of environmental adaptation in the leathery sea squirt (*Styela clava*). Mol Ecol Resour 20:1414-1431.
Wooldridge TB, Ford SM, Conwell HC, Hyde J, Harris K, Shapiro B. 2025. Direct measurement of the mutation rate and its evolutionary consequences in a critically endangered mollusk. Mol Biol Evol 42:msae266.
Xian W, Carbonell-Bejerano P, Rabanal FA, Bezrukov I, Reymond P, Weigel D. 2025. Minimizing detection bias of somatic mutations in a highly heterozygous oak genome. G3 (Bethesda):jkaf143.
Xu X, Li G, Li C, Zhang J, Wang Q, Simmons DK, Chen X, Wijesena N, Zhu W, Wang Z, et al. 2019. Evolutionary transition between invertebrates and vertebrates via methylation reprogramming in embryogenesis. Natl Sci Rev 6:993-1003.
Zhang C, Reid K, Sands AF, Fraimout A, Schierup MH, Merila J. 2023. *De novo* mutation rates in sticklebacks. Mol Biol Evol 40:msad192.
Zhang C, Reid K, Schierup MH, Wang H, Candolin U, Merila J. 2025. Rate of *de novo* mutations in the three-spined stickleback. Heredity (Edinb) 134:387-395.
Zhang M, Yang Q, Ai H, Huang L. 2022. Revisiting the evolutionary history of pigs via *de novo* mutation rate estimation in a three-generation pedigree. Genomics Proteomics Bioinformatics 20:1040-1052.
Zhang X, Wu R, Wang Y, Yu J, Tang H. 2020. Unzipping haplotypes in diploid and polyploid genomes. Comput Struct Biotechnol J 18:66-72.

Supplementary Tables and Figures

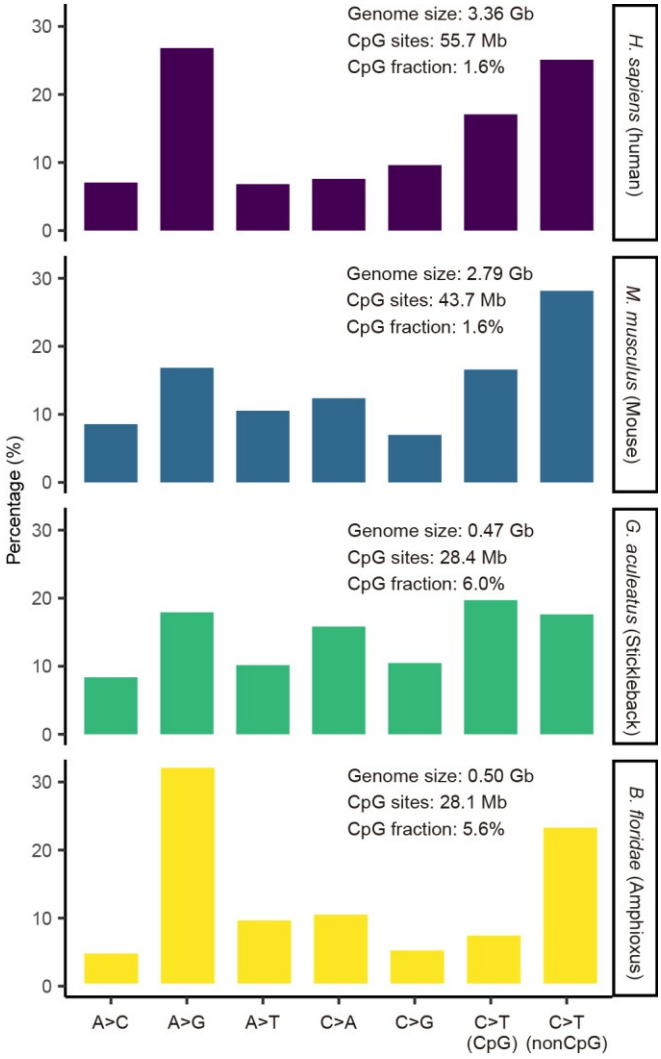
Supplementary Tables are provided in a separate Excel file, and Supplementary Figures are shown below.



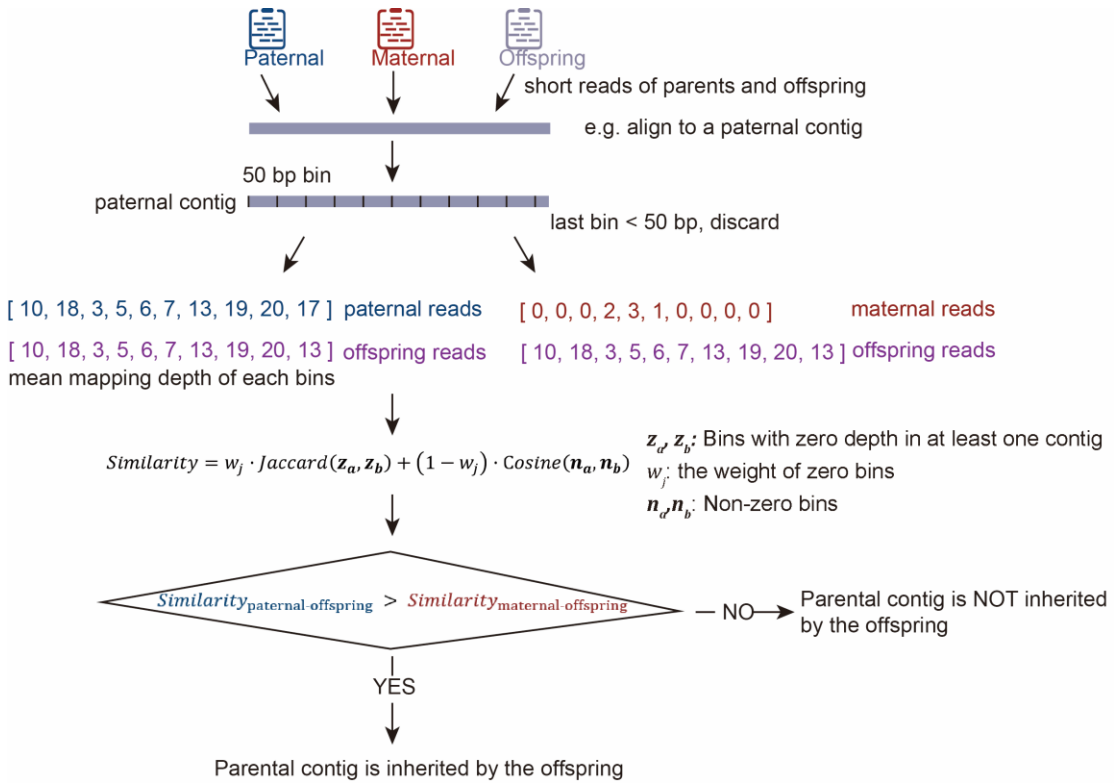
Supplementary Figure 1 Statistics of sequencing and mapping results. (A) Distribution of sequencing depths of all samples. (**B**) Boxplots showing the numbers of all reads and the numbers of mapped reads against the parental genomes. (**C**) Boxplots showing the percentages of all reads that were mapped to the parental genome, and the percentage of properly aligned reads. 'Properly aligned' means that paired-end reads were aligned to the same sequence with an insert size of ≤ 1,000 bp. (**D**) Boxplot showing average read depths of samples based on the alignments against the parental genome.



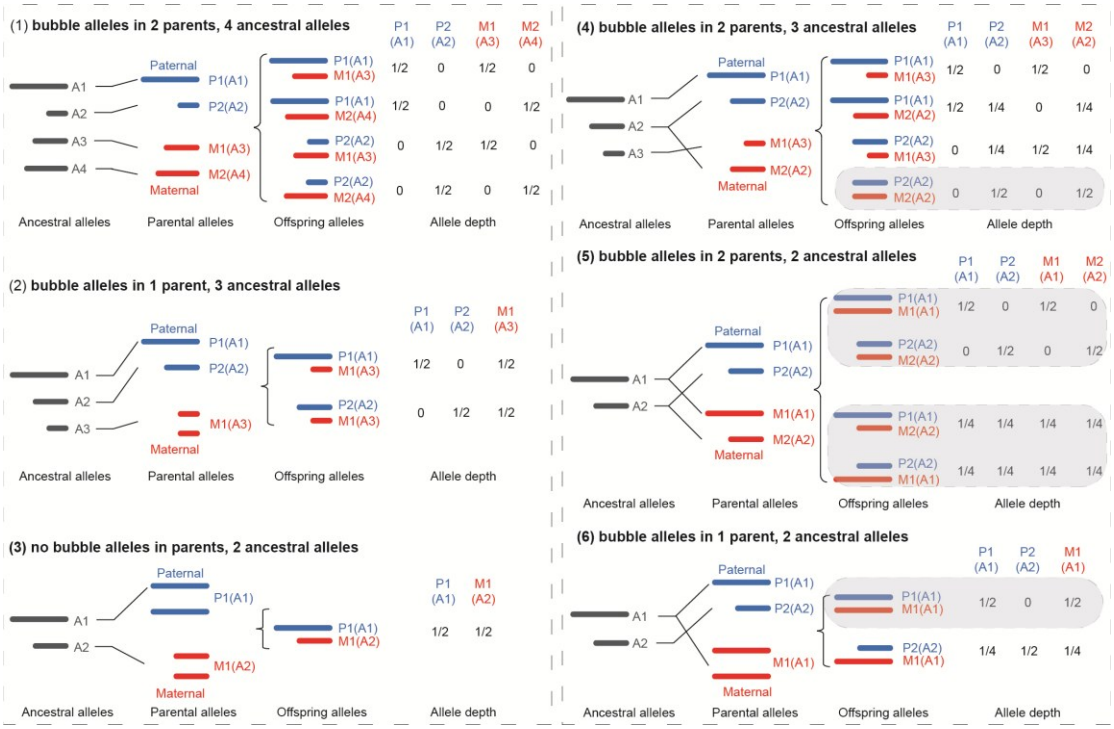
Supplementary Figure 2 Genome characteristics of *B. floridae* parental genomes based on 21-mer analysis. (**A**) Maternal genome. (**B**) Paternal genome. The bottom of each panel displays genome characteristics inferred from 21-mer analysis, including estimated genome size, genomic heterozygosity, repeat rate, *k*-mer size and ploidy.



Supplementary Figure 3 Mutation spectra of *B. floridae* and representative vertebrate species. *H. sapiens* (Human), purple; *M. musculus* (Mouse), blue; *G. aculeatus* (Stickleback), green; *B. floridae* (Amphioxus), yellow.



Supplementary Figure 4 Inferring inheritance states of parental contigs in offspring. First, we aligned parents' and offsprings' WGS reads to the parental reference genome. In the figure, we chose a paternal contig as an example. To quantify depth distribution, the paternal contig was divided into 50-bp bins, discarding any terminal bin shorter than 50 bp. The mean depth within each bin was calculated for the offspring, the mother and the father, respectively, generating a depth distribution vector per individual across the paternal contig. The similarity between offspring and parental depth vectors was quantified using a weighted measure combining: (1) Cosine similarity for bins with non-zero depths in both individuals being compared, and (2) Jaccard similarity for bins with zero coverage in at least one of the two individuals (accounting for regions with no aligned reads). The paternal contig was considered inherited by an offspring if the offspring's depth distribution vector showed greater similarity to that of the father than to the mother.



Supplementary Figure 5 Possible allele combinations in the paternal and maternal genomes at a locus and possible resulting read depths of alleles in the offspring. Distinct alleles (assembled contigs) are named A1, A2, etc. P1, P2, M1 and M2 are names of paternal and maternal alleles. The value of an allele depth represents the relative ratio of read depth of the specific allele to an offspring's average read depth. For scenarios (1) - (3), the two parents do not have shared (highly similar) alleles, the inheritance states of alleles can be determined and the offspring alleles can be phased. In contrast, for scenarios (4) - (6), when some alleles that are highly similar between parents are passed to an offspring (see shaded areas in the figure), the inheritance states of alleles cannot be reliably determined or offspring alleles cannot be phased. Therefore, the genomic regions in the offspring matching the conditions in shaded areas were not included in the callable genome.