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 occurrence in highly heterozygous genomes remains 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 the 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 256 high-confidence DNMs, yielding a genome-wide mutation rate of 4.41×10^{-9} per base per generation, which is comparable to that of vertebrates. By combining this estimate with observed genomic heterozygosity of *B. floridae*, we obtain an effective population size of ~2.2 million, indicating that its extreme heterozygosity mainly results from a large effective population size. Furthermore, we observed a maternal mutation bias and increased mutation rates in coding regions, suggesting unique mutational mechanisms in amphioxus. Our study not only provides the first mutation rate estimate for amphioxus but also offers a scalable strategy for studying DNMs in highly heterozygous genomes, shedding light on the evolution of mutation rates across chordates.

Main text

Introduction

Germline *de novo* mutations (DNMs) differ from somatic mutations in that they can be transmitted across generations, serving as the ultimate source of genetic variation and fueling biological evolution. These mutations arise as a consequence of DNA damage or replication errors, and because most mutations are relatively deleterious to the individual, organisms incur considerable adaptive costs to minimize such errors and damage, thereby maintaining the number of DNMs at an extremely low level, approximately 10^{-8} to 10^{-9} per site per generation (Lynch 2010).

The *de novo* mutation rate measures how often DNMs occur. Its variation across genomic positions, developmental stages, parental origins, and mutation types are closely linked to the underlying mutational mechanisms. 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 comparative study in amniotes showed that while both sexes exhibit similar mutation rates during early development, the male germline exhibits a higher mutation rate following sexual differentiation (de Manuel et al. 2022), indicating influence of different developmental stages and gametogenesis of different sexes on mutational occurrence.

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. Amphioxus possesses a body plan that is similar to vertebrates but markedly more simplified, and its chromosomal and genomic sequences have undergone far fewer lineage-specific changes compared with those of tunicates (Putnam et al. 2008).

Intriguingly, amphioxus exhibits an extremely high level of genomic heterozygosity (3.2~4.2% in sequenced species), one of the highest among sequenced 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. In the absence of accurate mutation rate data, it is unclear whether the high heterozygosity in amphioxus is due to a large effective population size, an increased mutation rate, or

both.

Therefore, measuring the magnitude and distribution of DNMs in the amphioxus genome is crucial for precisely estimating the mutation rate and understanding underlying mutational mechanisms. With an accurate estimate of the mutation rate, we can also uncover the reason why the amphioxus genome has such an astonishingly high level of heterozygosity.

To date, no study has investigated DNMs in amphioxus. A major obstacle is its high heterozygosity, which complicates key genomic analysis steps such as genome assembly, read alignment, and variant detection. Studies in macaques have shown that even gold-standard pipelines for DNM detection produce elevated error rates in highly heterozygous species (Bergeron et al. 2022).

In this study, we sequenced a two-generation pedigree (two parents and 104 F1 offspring) of *Branchiostoma floridae*, comprising 104 trios, to systematically study DNMs in the amphioxus. To address 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. Using this approach, we successfully identified 256 high-confidence DNMs in amphioxus. We estimated its mutation rate to be 4.41×10^{-9} per base per generation, which is comparable to those reported in fish and other vertebrates. With the mutation rate, we estimated the effective population size of *B. floridae* populations to be 2,199,546, among the largest in surveyed chordates. Therefore, the high heterozygosity in amphioxus is mainly caused by the large effective population size. Furthermore, we discovered that multiple aspects of DNMs in amphioxus differ from those observed in vertebrates, implying some unique mutational mechanisms in this lineage.

Results

A 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 81× per sample (ranging from 50× to 144×) and high sequencing quality (**Supplementary Figure 1**). The high sequencing depth enabled us to do contig-level *de novo* assembly (Figure 1 A).

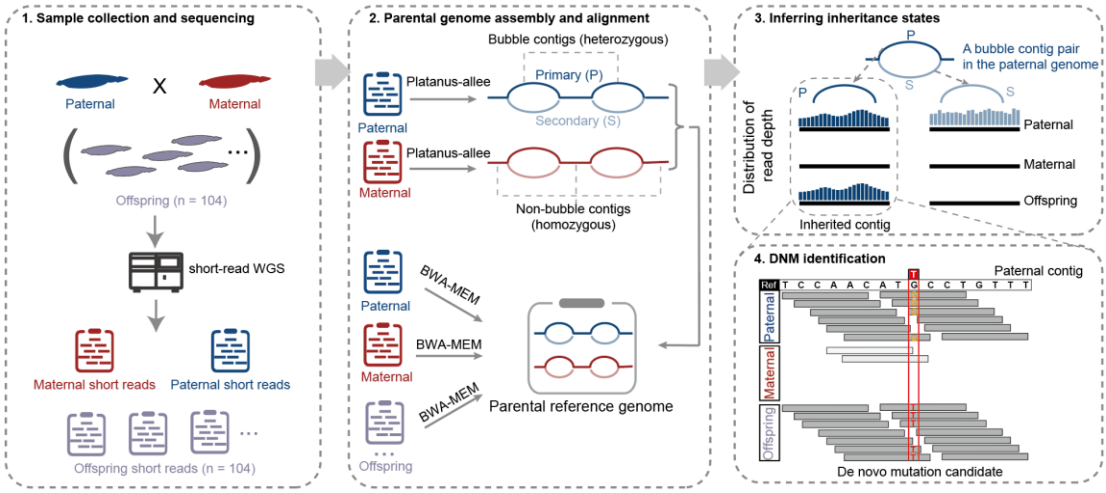

Figure 1 Schematic of the workflow for detecting DNMs. First, we collected DNA samples from a *B. floridae* family of two parents and 104 F1 offspring. We performed deep short-read sequencing (>50X) for all individuals. Using the parental WGS data, we assembled allele-aware parental genomes with Platanus-allee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

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 existing DNM detection pipelines (Bergeron et al. 2022), which struggle to accurately genotype both offspring and parents simultaneously. Because of low sequence similarity between our sequencing data and the public reference genome, many reads in our sequencing data cannot be correctly

129 mapped to the reference genome by common short-read aligners (e.g. BWA), leading
130 to unreliable variant calling results.

131 We reasoned that the heterozygosity of *B. floridae* can facilitate allele-aware assembly
132 of the parental genomes, as many divergent regions in the parental genomes can be
133 assembled into separated contigs. Using the parental genome assembly as a custom
134 reference, we should be able to align most offspring and parental reads to the parental
135 genomes, thereby obtaining high-quality genotype information of offspring and parents
136 for DNM detection (**Figure 1**).

137 Therefore, we constructed the genome assembly for each parent using the allele-aware
138 assembler Platanus_allee (v2.2.2) (Kajitani et al. 2019). We obtained a maternal
139 genome assembly of 773.8 Mb and a paternal genome assembly of 763.7 Mb,
140 approximately 1.6 times the length of the previously published haploid genome (489
141 Mb), indicating that a large fraction of the diploid parental genomes was assembled into
142 phased alleles. In Platanus-allee, heterozygous regions are assembled as "bubble" con-
143 tigs with two allelic sequences, named as "primary" and "secondary" respectively.
144 Homozygous regions with identical sequences in an individual are assembled as "non-
145 bubble" contigs (**Figure 1**).

146 By merging the two parental genome assemblies, we generated a composite parental
147 reference genome, with detailed assembly statistics provided in **Supplementary Table**
148 **1**. The parental reference genome has a total length of approximately 1.4 Gb, consisting
149 of 263,358 bubble contigs covering around 856 Mb and 758,976 non-bubble contigs
150 spanning approximately 780 Mb. The contig N50 metrics of bubble contigs and
151 non-bubble contigs are 6.5 kb and 7.6 kb, respectively. Most sequencing reads of the
152 two parents can be aligned to this parental reference genome (**Supplementary Figure**
153 **1**), indicating that the vast majority of genomic segments have been successfully
154 assembled and are available for downstream analysis.

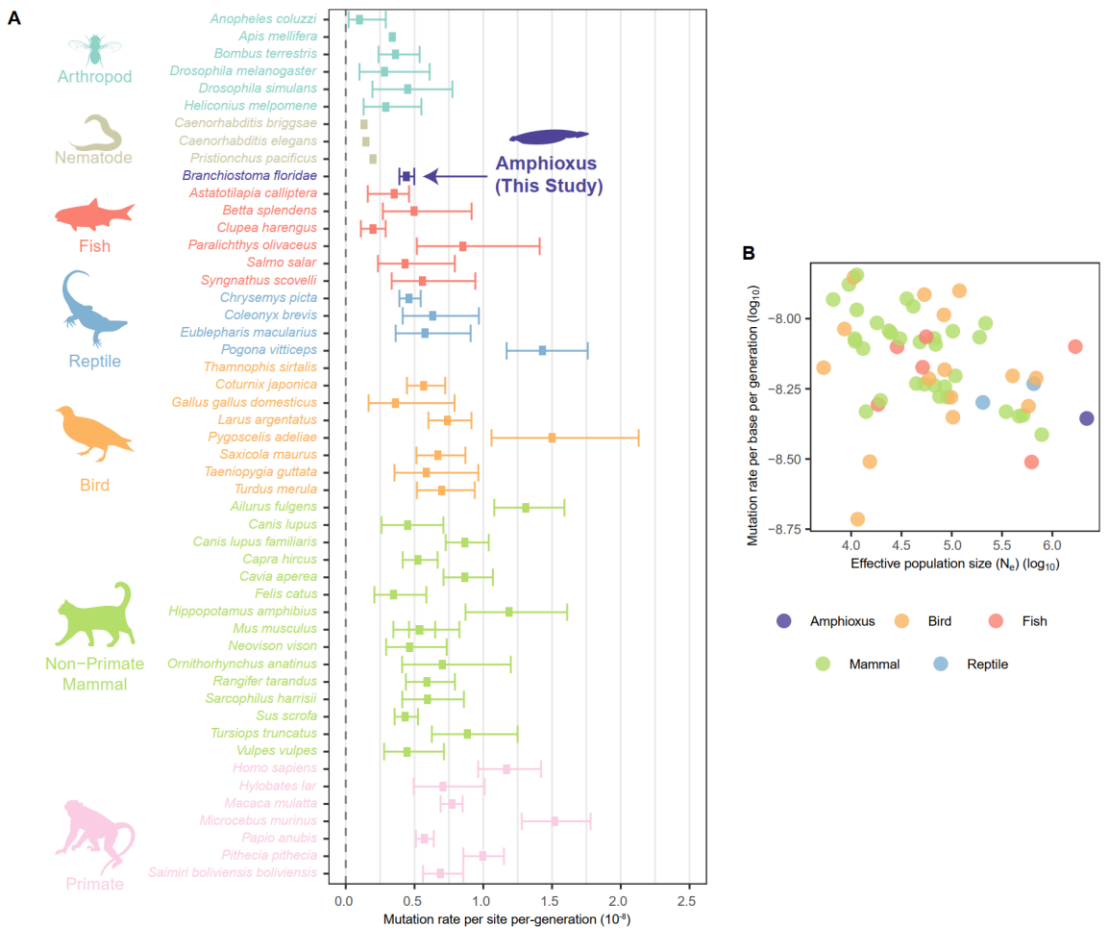

Figure 2 Estimates of mutation rate and effective population size. (**A**) Per site per generation *de novo* mutation rate (μ) estimates from this study and previous studies (Bergeron et al. 2023). Error bars denote 95% confidence intervals. (**B**) Comparison of mutation rates and effective population sizes across diverse taxonomic groups, using data for other taxa from Bergeron et al. (2023).

Next, we aligned sequencing reads from both offspring and parents to this parental reference genome and detected variants using bcftools (Danecek et al. 2021). A custom filtering pipeline, accounting for factors such as read depth and variant quality (**Supplementary Table 2**), was employed to ensure the accuracy of DNMs (see Methods). After applying the filtering criteria, 1,102 candidate DNMs were retained. Among these DNMs, some false positives that could only be ruled out through visualization of IGV identified by manual check (see Methods). After careful check, a total of 256 DNMs were confirmed across 104 offspring, corresponding to an average of 2.46 DNMs per individual. With an average callable genome size of 558.4 Mb, the mean mutation rate across the genome was calculated to be 4.4×10^{-9}mutations per base per generation, with a 95% confidence interval (CI) of $3.90\text{-}4.98\times10^{-9}$, as shown in

Figure 2A. The mutation rate of *B. floridae* is comparable to those reported in fish and other vertebrates (Bergeron et al. 2023).

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

It has long been recognized that the heterozygosity of *B. floridae* is much higher than in other species (Putnam et al. 2008). Yet, since the mutation rate in *B. floridae* was not accurately quantified, the cause(s) of its high heterozygosity was uncertain. The expected heterozygosity per nucleotide, expressed as π, can be calculated by $\pi = 4N_e\mu$ under neutral evolution (Kimura 1985), where N_e represents the effective population size and μ the germline mutation rate. The effective population size ($\pi/4\mu$) calculated with our mutation rate and the π estimate (3.88%) from Huang et al. (2023) is approximately 2,199,546. According to Bergeron et al. (Bergeron et al. 2023), the harmonic mean of N_e over multiple time points estimated by PSMC correlated very well with N_e estimated by $\pi/4\mu$ ($r^2 = 0.81$). Therefore, we placed *B. floridae*'s N_e and μ within the context of N_e and μ derived from Bergeron et al. to show its unique position (**Figure 2B**), indicating that *B. floridae*'s N_e is largest among chordates with measurements. Collectively, we concluded that the high heterozygosity of amphioxus is not due to an elevated mutation rate but a very large effective population size.

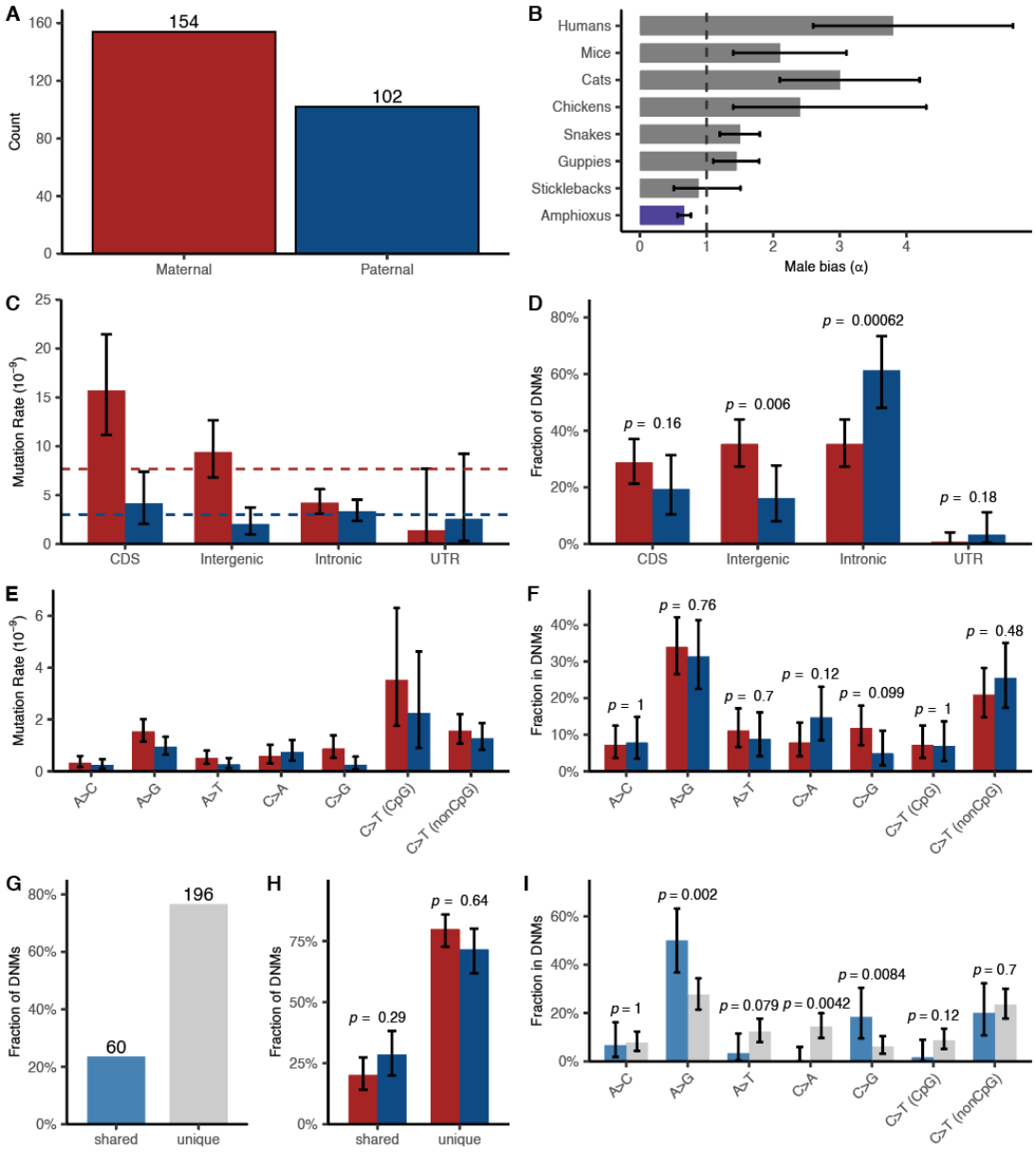

Figure 3 Characteristics of germline DNMs in *B. floridae*. (A) Parental origins of DNMs. **(B)** The male-to-female ratios (α) of DNMs for different groups of chordates. **(C)** Mutation rates in different gene annotation contexts, separating paternal (blue) and maternal (red) DNMs. Blue and red dashed lines indicate the average paternal and maternal mutation rates, respectively. **(D)** Proportions of DNMs in different gene annotation contexts, separating paternal (blue) and maternal (red) DNMs. Parental origin differences assessed by χ^2 tests. Parental origin effects in mutation spectrum: **(E)** Mutation rates of different substitution types, separating paternal (blue) and maternal (red) DNMs. **(F)** Proportions of DNMs of different substitution types, separating paternal (blue) and maternal (red) DNMs. Parental origin differences assessed by χ^2 tests. **(G)** Counts of sibling-shared and unique DNMs. **(H)** Proportion of DNMs across shared and unique DNM. Parental origin differences assessed by χ^2 tests. **(I)** Proportions of DNMs of different substitution types, separating shared and unique DNMs. Shared and unique DNMs differences for each substitution type assessed by χ^2 tests. All error bars are represented with 95% confidence intervals computed using binomial variance, and all p-values are labeled above the bars.

Characteristics of germline DNMs in *B. floridae*.

Distinguishing the parental origins of DNMs is crucial for understanding mutational mechanisms; while previous studies could only assign parental origin for 30%–80% of DNMs based on nearby polymorphic markers (Bergeron et al. 2023; Wooldridge et al. 2025), our method inherently provides parental origin information for all detected DNMs in *B. floridae*.

Among the 256 DNMs, 154 originated from the mother and 102 from the father (**Figure 3A**), yielding a male mutation bias coefficient (α) of 0.66 (i.e., a maternal bias). For a cross-species comparison, we integrated several representative vertebrates with known α values (**Figure 3B**). The results show that the α value of *B. floridae* is lower than most vertebrates. Bergeron et al. (2023) reported that on average fish tend to have apparently lower α values (mean $\alpha = 0.8$, 95% CI = 0.5–1.4) than other vertebrates (mammals, birds and reptiles). They suggested that differences in gametogenesis such as seasonal breeding and synchronous ovulation could lead to the lower α values in fish. The low α of amphioxus might be partly due to these reasons.

Due to the fragmented nature of the parental genome assemblies, direct genome 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 489 Mb of the public haploid reference genome, 363.21 Mb (74.27%) was covered by the contigs of the assembled parental contigs. The parental contigs that were mapped to the haploid reference genome were defined as mappable regions. Among the DNMs located in these regions, 198 were identified, accounting for 77.34% of all DNMs.

Based on the haploid genome annotation (Huang et al. 2023), the genome was divided into four functional regions: coding sequences (CDS), untranslated regions (UTR), intronic regions, and intergenic regions. The results show that the mutation rate in CDS regions is significantly higher than in other regions for both paternal and maternal DNMs (**Figure 3C**). In terms of DNM proportions, the intergenic region exhibits a significantly higher proportion of maternal DNMs compared to paternal DNMs, while the opposite trend tends to be in intronic regions (**Figure 3D**). These findings demonstrate that both parental origin and functional genomic context exert complex influences on the accumulation of DNMs, likely due to species-specific differences in

242 reproductive strategies or DNA damage/repair processes in *B. floridae*.

243 We further investigated DNM counts of different substitution types. The results indicate
244 that maternal mutation rates of all substitution types are consistently higher than
245 paternal ones (**Figure 3E**), and there is no significant difference between sexes when
246 looking at proportions of DNMs of different substitution types (**Figure 3F**). Cross-
247 species comparisons of mutational spectra (**Supplementary Figure 3**) revealed that *B.*
248 *floridae* exhibits a unique pattern, with a CpG>TpG mutation fraction of only 5%,
249 significantly lower than the 16–20% observed in typical vertebrates, whereas its A>G
250 mutation frequency of 31% is consistently higher than the 16-26% seen in other
251 chordate groups; the lower mutation frequency at CpG sites is probably due to lower
252 methylation levels at CpG sites in amphioxus (Marletaz et al. 2018).

253 60 (23%) of DNMs are shared among siblings (named ssDNMs, **Figure 3G**), which
254 likely occurred at early developmental stages. The proportions of shared DNMs in the
255 two parents are similar (**Fig. 3H**). The mutational spectrum shows a significant increase
256 in the proportions of A>G and C>G mutations in ssDNMs (**Fig. 3I**).

257 **Discussion**

258 In this study, we generated a pedigree comprising both parents and 104 offspring and
259 conducted high-depth short-read sequencing for all individuals. Standard approaches
260 for identifying DNMs typically rely on detecting heterozygous genotypes in offspring
261 that are homozygous in both parents. However, the high heterozygosity of the *B.*
262 *floridae* genome complicates such identification. The large difference between the
263 commonly used reference genome and re-sequenced individuals increases the chance
264 of incorrect alignment.

265 However, high heterozygosity can also facilitate allele-aware diploid genome assembly
266 without requiring long-read or Hi-C data. We assembled haplotype-resolved diploid
267 genomes at the contig level for each parent and developed a custom mapping and
268 variant calling pipeline using the parental assemblies as reference. This strategy
269 resolved the challenges of detecting DNMs in highly heterozygous species like *B.*
270 *floridae*. It also allowed us to assign the parent-of-origin for all detected DNMs.
271 Furthermore, based on the built parental assemblies and read alignments of offsprings,
272 we were also able to comprehensively investigate the meiotic recombination landscape,

273 which has been described in a companion study (Tao et al. 2025).

274 With an average callable genome size of 558 Mb, we identified 256 DNMs and
275 estimated the germline mutation rate in *B. floridae* to be 4.41×10^{-9} (95% CI of 3.90-
276 4.98×10^{-9}) per base per generation. Based on this mutation rate, the effective population
277 size of *B. floridae* was estimated to be over two million, suggesting that the high
278 heterozygosity resulted from a large effective population size. Our study did not
279 estimate the germline mutation rate for small insertions/deletions (INDELs), as
280 generally poor alignments around these variants make it challenging to accurately
281 detect *de novo* INDELs. Future studies incorporating long-read sequencing
282 technologies may facilitate estimating INDEL mutation rates.

283 We further analyzed the parental origin, genomic distribution, mutation spectrum, and
284 ssDNMs in *B. floridae*, revealing distinctive features in this basal chordate. We
285 observed a strong maternal bias in DNM origin, with a male-to-female mutation ratio
286 (α) of 0.66. This is lower than most reported values for vertebrates (de Manuel et al.
287 2022; Bergeron et al. 2023). We also found that coding sequences have higher mutation
288 rates than intronic and intergenic regions. The overall mutation spectrum showed no
289 obvious difference between sexes. The low frequency of C>T transitions at CpG sites
290 is consistent with previously reported low levels of CpG methylation in amphioxus
291 (Marletaz et al. 2018). The presence of ssDNMs provided a valuable window into early
292 germline mutational events. The relative proportions of different parental origins in
293 ssDNMs did not differ significantly from that of non-shared DNMs, a pattern distinct
294 from that in humans. Taken together, it suggests that mutational mechanisms involved
295 in early embryo development and later gametogenesis shaping parental biases in
296 amphioxus could be different from other chordate lineages. Future DNM studies
297 spanning different ages and developmental stages of parents could further clarify the
298 mechanisms of mutagenesis in amphioxus. In addition, functional genomics studies
299 focusing on DNA damage and repair pathways will be essential to understanding the
300 mutational mechanisms.

301 In summary, through high-depth sequencing and the development of a novel analysis
302 pipeline, we quantified the germline mutation rate in *B. floridae* for the first time and
303 resolved the long-standing question of what causes its high heterozygosity. We also
304 explored DNMs from multiple perspectives, including parental origin, gene annotation

305 contexts, mutation spectrum and shared DNMs among siblings. As amphioxus
306 represents the most basal living chordate, some of the unique DNM features observed
307 in this study may reflect ancestral characteristics of germline mutagenic processes. The
308 DNM analysis framework based on assembling parental genomes for read alignment
309 and variant calling is cost effective and can be generalized to other highly heterozygous
310 species.

311

312 **Materials and Methods**

313 **Amphioxus Culture**

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

325 **Sample collection and library construction**

326 We extracted high molecular weight genomic DNAs from the muscle tissues of a single
327 individual using the DNeasy Blood & Tissue Kit (QIAGEN, Valencia, CA), and
328 inspected the DNA quality by Qubit 2.0 Fluorometer (Thermo Fisher Scientific,
329 Waltham, MA) and 2100 Agilent Bioanalyzer (Agilent). After Sample QC, gDNA was
330 fragmented by ultrasound on Covaris E220 (Covaris, Brighton, UK), then selected to
331 300 bp - 500 bp using magnet beads size selection. The selected DNA fragments were
332 then end-repaired and A-tailed. The indexed adaptors were ligated to both ends of the
333 DNA fragments. The ligation product was amplified by PCR, and circularized to get
334 single-stranded circular (ssCir) library. The ssCir library was then amplified through
335 rolling circle amplification (RCA) to obtain DNA nanoball (DNB). The DNB was then

336 loaded to flowcell, and sequenced by DNBSEQ Platform.

337 **Parental genome assembly**

338 Genome heterozygosity of parental individuals was estimated via k-mer analysis of
339 Illumina short-read data using GenomeScope (v2.0) (Ranallo-Benavidez et al. 2020).
340 Raw short-read sequences were quality-processed with Fastp (v0.24.0) (Chen 2023),
341 implementing: (1) elimination of reads with >10% ambiguous bases ('N'), and (2)
342 retention of reads exhibiting a minimum mean Phred quality score of 20 (Q20). Adapter
343 sequences were concurrently trimmed. These processed reads were subjected to allele-
344 aware diploid genome assembly using Platanus_allee (v2.2.2) (Kajitani et al. 2019)
345 through a two-stage workflow: First, the *de novo* assembly phase was executed with
346 the command "assemble -m 200 -t 20"; second, haplotype resolution was performed
347 via the phase module using "phase -i 3 -t 20". Contigs shorter than 150 bp in length
348 were subsequently filtered from both parental phased assemblies. Initial assembly
349 quality metrics, including contig N50, total length, and GC content, were
350 comprehensively evaluated with QUAST (v5.3.0) (Gurevich et al. 2013) and seqkit
351 (v2.9.0) (Shen et al. 2016).

352 Following the phased assembly procedure, contigs underwent automated classification
353 by Platanus_allee. In the contig-level assemblies generated by Platanus_allee, contigs
354 were classified into two distinct categories: (1) *bubble contigs* harboring allelic variants,
355 which were further stratified as *primary contigs* (longer allelic sequences)
356 and *secondary contigs* (shorter counterparts), and (2) *non-bubble contigs* designated by
357 the assembler as lacking allelic diversity. However, through comprehensive whole-
358 genome alignment analyses, we discovered that a subset of non-bubble contigs
359 exhibited unexpected sequence overlaps with other contigs across the assembly,
360 indicating residual heterozygous regions or potential mis-assemblies unresolved by the
361 phasing algorithm.

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

365 **Alignment of offspring and parental reads to parental reference genome and**
366 **variant calling**

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

375 Based on the read alignments above, we used bcftools (v1.21) (Danecek et al. 2021) to
376 identify small variants. First, we generated "pileup" files for the set of all samples by
377 running "bcftools mpileup" with parameters "–annotate FORMAT/AD,
378 FORMAT/ADF, FORMAT/ADR, FORMAT/DP, FORMAT/SP, INFO/AD,
379 INFO/ADF, INFO/ADR" on all input bam files. With the "pileup" files as input, we
380 ran "bcftools call -mv" to generate VCF files, which contain SNVs and short
381 insertions/deletions, for 104 offspring and two parents.

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

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

396 To quantify depth distribution, each contig was divided into 50-bp bins, discarding any
397 terminal bin shorter than 50 bp. The mean depth within each bin was calculated for each
398 individual, generating a depth distribution vector per individual across the contig. The

similarity between offspring and parental vectors was quantified using a weighted measure combining: (1) Cosine similarity for bins with non-zero depth 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 of no aligned reads). A parental contig was considered inherited by an offspring if its depth distribution vector in that offspring showed greater similarity to the vector of its specific parental source than to the other parent. Parental contigs exhibiting a depth less than 1/4 of an offspring's mean depth or a depth less than 10 were classified as low-depth and thus non-inherited in the specific offspring.

However, the similarity of some contigs between the parental genomes, probably arising from shared ancestry, could confound the inheritance calls derived from the offspring-parent depth distribution similarity algorithm (**Supplementary Figure 4**). Since direct pairwise genome sequence alignment of two parental genomes is difficult for assessing parental contig similarities due to potential length disparities of orthologous contigs in different parental assemblies, we used the similarity of read depth distributions (calculated using the same algorithm employed for offspring-parent comparisons) when mapping each parent's reads *to the same contig* as a robust proxy. If the depth distribution similarity between the two parents for a specific contig exceeds 0.9, the inheritance calls for that contig in the offspring are deemed unreliable due to indistinguishable parental origins, necessitating further validation with additional data.

Because a given diploid offspring can inherit only one allele from each parent at a locus, determining the inheritance status of one contig within a pair of bubble contigs from a parent inherently resolves the inheritance status of both contigs in the pair. In other words, the inheritance of one contig in a bubble pair can be inferred based on the inheritance of its counterpart. Therefore, we further took advantage of the very low read depths in some loci to conclude that an allele has not been inherited; under such circumstances, the alternative allele with aligned reads must be inherited by the offspring. In contrast, for non-bubble contigs, the inability to identify a second allele precludes this method of assessing similarity between parental contigs, and therefore these contigs are excluded from further analysis. Finally, all scenarios arising from the interplay between inheritance and contig similarity, which lead to differences in depth distributions in offspring, are comprehensively visualized in Supplementary Figure 4.

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

434 **Determining the callable genome**

435 Estimating mutation rates requires determining the length of genomic regions within
436 each offspring in which DNMs can be confidently identified, defined as the callable
437 genome size. In each offspring, the callable genome is primarily composed of the
438 inherited genomic segments defined in the previous section. However, the accurate
439 inference of the inheritance of non-bubble contigs with high similarity between the two
440 parents proved challenging. For non-bubble contigs lacking pre-existing allelic data in
441 the parental genomes, the allele-depth-distribution-based method for inferring contig
442 inheritance patterns is inapplicable. Consequently, determining the inheritance of
443 homologous contigs within non-bubble contigs becomes challenging. We therefore
444 implemented stricter filtering criteria, excluding both non-bubble contigs with depth
445 distribution similarity ≥0.5 from the callable genome regions.

446 Additionally, we removed sites in repetitive regions, as they are prone to erroneous
447 alignments and often exhibit abnormally high alignment depths. Sites with too many
448 reads may represent problematic repetitive regions (Li 2014). Specifically, based on the
449 observed depth distribution in each offspring, any regions with a read depth exceeding
450 three times the individual's average were excluded.

451 Finally, by integrating these filters based on both inheritance and sequencing depth, we
452 obtained the callable genome size for each offspring ranging from 521.70 Mb to 580.88
453 Mb, accounting for approximately 53–59% of the diploid genome size.

454 **Identifying the candidate DNMs**

455 For each parent-offspring trio, the variants in three VCF files of each trio were filtered
456 to a subset of single nucleotide variants (SNVs) by bcftools (v1.21) (Danecek et al.
457 2021).

458 Variants were retained only if they met all computational criteria: (1) homozygous
459 alternative allele status in the offspring with read depth >10×; (2) the variant was
460 located exclusively within the predefined callable genome regions; (3) there were no

more than two alternative allele observations in each parental BAM file, as verified by our custom Python allelic-depth quantification pipeline; and (4) the variant had a minimum QUAL score 220, ensuring high-confidence base calls; (5) no proximity to indels (±10 bp), (6) not located in terminal contig regions (within 200 bp of contig ends); (7) variants not at the sites experiencing putative gene conversions, as gene conversions lead to read alignments resembling DNMs.

After applying these filters, the resulting set of candidate DNMs was further evaluated by importing both parental and offspring BAM files into IGV (v2.19.2) (Thorvaldsdottir et al. 2013), and using IGVtools to generate screenshots for each candidate DNM in both the offspring and the parents. Subsequent manual inspection led to the removal of any candidate DNMs if they met one or more of the following criteria (1) structural variant/recombination breakpoint neighborhoods, and (2) complex repetitive domains in short contigs (<5 kb) with ambiguous alignments. All filtered variants were categorized by exclusion criteria, with counts of discarded sites per filter tier provided in Figure S2. These steps yielded a DNM dataset for each individual. Individual mutation rates were calculated by dividing the number of DNMs detected in each offspring by the corresponding callable genome size, and the final genome-wide mutation rate was calculated by the number of all DNMs divided by sum of callable genome sizes in offspring. Notably, the callable genome size was not multiplied by two, since the reconstructed inherited genomic segments effectively represent an approximately diploid genome for each offspring.

Because the coordinates of DNMs are located on the parental genomic contigs, the parental origin of the contig harboring each DNM naturally reveals its source, allowing for unequivocal determination of the DNM's parental origin without any further processing. This same principle applies to the callable genome as well.

Analysis of the effective population size

The average long-term N_e was estimated using the formula $N_e = \pi / 4\mu$ (Kimura 1985), where μ is the mutation rate estimated by our work (4.41×10^{-9} per site per generation) and π from a previous study (Huang et al. 2023). For comparison, the harmonic mean of historical effective population size in vertebrates calculated using PSMC (Li and Durbin 2011) was obtained from data published by Bergeron et al. (Bergeron et al. 2023). The rationale for choosing this dataset is that the authors found a strong

correlation (r^2 = 0.83) between effective population sizes computed via PSMC and long-term effective population sizes, demonstrating that this dataset can serve as a robust proxy for long-term effective population size.

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

Owing to inherent limitations of short-read sequence, the assembly of parental genomes could not achieve chromosome-scale or high continuity, posing challenges for some downstream analyses. Furthermore, defining orthologous relationships between contigs of two parents in a specific locus based solely on pairwise alignments of contigs, as previously mentioned, was complicated by the variability in the lengths of assembled contigs derived from that locus. To address these issues, we aligned the contigs of parental genomes to a published *B. floridae* haploid reference genome (Huang et al. 2023). After testing multiple alignment tools under various parameter configurations, Minimap2 (v2.24) (Li 2021) was selected as the optimal aligner, with empirically optimized parameters: "-x asm20 --secondary=no". Resultant alignments were processed through sequential refinement: coordinate sorting via samtools (v1.21) (Li et al. 2009), (2) removal of multi-mapping regions using Sambamba (v0.8.2) (Tarasov et al. 2015). A custom script was employed to establish the mapping relationship between each parental genome and the haploid reference genome. This approach yielded a multiple genome alignment that aligned positions across the paternal, maternal, and reference genomes. As a result, the majority of parental genomic positions were successfully anchored onto the reference, thereby obtaining orthologous genomic coordinates across the parental genomes and facilitating subsequent analyses.

Comparative analysis of mutation rates in different contexts

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

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

Data Availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (GSA: CRA027444) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. Genome assemblies produced by Platanus-allee have been deposited in Science Data Bank (ScienceDB) repository (https://www.scidb.cn/en/anonymous/SmppTUpm).

Author contribution

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

Conflict of interest

The authors have declared no competing interests.

Acknowledgments

We thank Dr. Luohao Xu for his assistance in obtaining previous genomic data of amphioxus populations. This work was supported by National Natural Science Foundation of China (32470690), Natural Science Foundation of Fujian Province of China (2022J06004), State Key Laboratory of Biocontrol and Guangdong Provincial Key Laboratory for Aquatic Economic Animals.

References

Bergeron LA, Besenbacher S, Turner T, Versoza CJ, Wang RJ, Price AL, Armstrong E, Riera M, Carlson J, Chen HY et al. 2022. The Mutationathon highlights the importance of reaching standardization in estimates of pedigree-based germline mutation rates. *Elife* **11**.

Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, Hoffman JI, Li Z, St Leger J, Shao C et al. 2023. Evolution of the germline mutation rate across vertebrates. *Nature* **615**: 285-291.

Bi C, Lu N, Han T, Huang Z, Chen JY, He C, Lu Z. 2020. Whole-Genome Resequencing of Twenty Branchiostoma belcheri Individuals Provides a Brand-New Variant Dataset for Branchiostoma. *Biomed Res Int* **2020**: 3697342.

Chen S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**: e107.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**.

de Manuel M, Wu FL, Przeworski M. 2022. A paternal bias in germline mutation is widespread in amniotes and can arise independently of cell division numbers. *Elife* **11**.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072-1075.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* **12**: 756-766.

Huang Z, Xu L, Cai C, Zhou Y, Liu J, Xu Z, Zhu Z, Kang W, Cen W, Pei S et al. 2023. Three amphioxus reference genomes reveal gene and chromosome evolution of chordates. *Proc Natl Acad Sci U S A* **120**: e2201504120.

Kajitani R, Yoshimura D, Okuno M, Minakuchi Y, Kagoshima H, Fujiyama A, Kubokawa K, Kohara Y, Toyoda A, Itoh T. 2019. Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nat Commun* **10**: 1702.

Kimura M. 1985. *The neutral theory of molecular evolution*. Cambridge university press.

Li G, Yang X, Shu Z, Chen X, Wang Y. 2012. Consecutive spawnings of Chinese amphioxus, Branchiostoma belcheri, in captivity. *PLoS One* **7**: e50838.

Li H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**: 2843-2851.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**: 4572-4574.

Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493-496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lynch M. 2010. Evolution of the mutation rate. *Trends Genet* **26**: 345-352.

Marletaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, Wyatt CDR, de la Calle-Mustienes E, Bertrand S, Burguera D et al. 2018. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**: 64-70.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**: 867-868.

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064-1071.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**: 1432.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**: e0163962.

Tao L, Xue J, Cao JW, Li G, Li C. 2025. Direct detection of meiotic recombination events in the highly heterozygous amphioxus genome. *bioRxiv*.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**: 2032-2034.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-

performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178-192.
Van der Auwera GA, O'Connor BD. 2020. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media.
Vasimuddin M, Misra S, Li H, Aluru S. 2019. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, doi:10.1109/IPDPS.2019.00041, pp. 314-324.
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**.
Zhang C, Reid K, Sands AF, Fraimout A, Schierup MH, Merila J. 2023. De Novo Mutation Rates in Sticklebacks. *Mol Biol Evol* **40**.