Repeat-Rich Regions Cause False-Positive Detection of NUMTs: A Case Study in Amphibians Using an Improved Cane Toad Reference Genome

Kelton Cheung [1,2], Lee Ann Rollins [1], Jillian M. Hammond [3,4], Kirston Barton [5], James M. Ferguson [4], Harrison J. F. Eyck [6], Richard Shine [7], Richard J. Edwards [2,8,*]

[1]Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of New South Wales, Sydney, Australia

[2]Evolution & Ecology Research Centre, School of Biotechnology & Biomolecular Sciences, University of New South Wales, Sydney, Australia

[3]Genomics and Inherited Disease Program, Garvan Institute of Medical Research, Sydney, New South Wales, Australia

[4]Centre for Population Genomics, Garvan Institute of Medical Research and Murdoch Children's Research Institute, Darlinghurst, New South Wales, Australia

[5]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney, New South Wales, Australia

[6]National Collections and Marine Infrastructure, CSIRO, Canberra, Australian Capital Territory, Australia

[7]School of Natural Sciences, Macquarie University, Sydney, New South Wales, Australia

[8]Minderoo OceanOmics Centre at UWA, Oceans Institute, The University of Western Australia, Western Australia, Australia

*Corresponding author: E-mail: rich.edwards@uwa.edu.au.

Accepted: November 04, 2024

Abstract

Mitochondrial DNA (mtDNA) has been widely used in genetics research for decades. Contamination from nuclear DNA of mitochondrial origin (NUMTs) can confound studies of phylogenetic relationships and mtDNA heteroplasmy. Homology searches with mtDNA are widely used to detect NUMTs in the nuclear genome. Nevertheless, false-positive detection of NUMTs is common when handling repeat-rich sequences, while fragmented genomes might result in missing true NUMTs. In this study, we investigated different NUMT detection methods and how the quality of the genome assembly affects them. We presented an improved nuclear genome assembly (aRhiMar1.3) of the invasive cane toad (*Rhinella marina*) with additional long-read Nanopore and 10× linked-read sequencing. The final assembly was 3.47 Gb in length with 91.3% of tetrapod universal single-copy orthologs ($n = 5{,}310$), indicating the gene-containing regions were well assembled. We used 3 complementary methods (NUMTFinder, *dinumt*, and *PALMER*) to study the NUMT landscape of the cane toad genome. All 3 methods yielded consistent results, showing very few NUMTs in the cane toad genome. Furthermore, we expanded NUMT detection analyses to other amphibians and confirmed a weak relationship between genome size and the number of NUMTs present in the nuclear genome. Amphibians are repeat-rich, and we show that the number of NUMTs found in highly repetitive genomes is prone to inflation when using homology-based detection without filters. Together, this study provides an exemplar of how to robustly identify NUMTs in complex genomes when confounding effects on mtDNA analyses are a concern.

Key words: cane toad, *Rhinella marina*, NUMT, mitochondrial DNA, genome assembly, amphibians.

Significance

This study uses an updated cane toad nuclear genome assembly and multiple nuclear DNA of mitochondrial origin (NUMT) detection methods to confirm a lack of NUMTs that might confound the use of mitochondrial DNA as a population genetic marker in the cane toad. We provide an exemplar study for NUMT detection accounting for genome assembly quality and composition and highlight the risks of using BLASTN-based approaches in highly repetitive nuclear genomes.

Introduction

Mitochondrial DNA (mtDNA) has been utilized in phylogenetics, DNA barcoding, and population genetics studies for decades. High copy numbers per cell make mitochondrial DNA easier to retrieve from samples of low quantity or quality. Additionally, the haplotype nature, mode of inheritance, and abundant knowledge of mitochondrial genomes contribute to their usefulness in genetics research. Studies of mitochondrial DNA have yielded rich bodies of research on the evolution, population structure, and phylogenetics of a broad range of taxa (Ballard and Rand 2005; Gray 2012; Kivisild 2015).

One issue that requires consideration when analyzing mtDNA data is the presence of nuclear DNA of mitochondrial origin (NUMTs). Numtogenesis is the transferal of mtDNA into the nuclear genome, creating NUMTs (Singh et al. 2017). NUMTs can become integrated into the nuclear genome when double-strand breaks are repaired by nonhomologous end joining (NHEJ) (Hazkani-Covo and Covo 2008) or microhomology-mediated end joining (Wei et al. 2022). A large-scale human study of tumor and germline cells revealed an enrichment in the tumor cells of NUMTs from the noncoding D-loop, in relation to the transcription start site and origin of replication (Wei et al. 2022). NUMTs have been reported across almost every taxonomic group in eukaryotes, and a positive relationship between the nuclear genome size and total NUMT content in the genome has been reported (Bensasson et al. 2001; Richly and Leister 2004; Hazkani-Covo et al. 2010). However, this relationship is not always consistent. For example, the zebrafish nuclear genome (1.4 to 1.7 Gb in length) does not contain NUMTs (Hazkani-Covo et al. 2010) while avian species with similarly small genomes (0.91 to 1.3 Gb) (Zhang et al. 2014) show a wide range of NUMT content, from as few as 4 to >600 NUMTs (Liang et al. 2018). This illustrates that even with similar genome sizes, some species can have vastly different NUMT content, indicating that the relationship between genome size and NUMT abundance is not universally applicable across all taxa. The variation in abundance of NUMTs across different taxonomic groups could result from differences in the rate of NUMT insertion, the rate of NUMT duplication, and the rate of NUMT removal (Hazkani-Covo et al. 2010).

NUMT sequences are expected to diverge from their mtDNA counterparts over time, as the nuclear copy will not be under functional constraint. However, at their creation, NUMTs are exact copies of their mitochondrial sequence of origin and can remain similar for extended periods of time due to the low mutation rate of nuclear DNA. As a consequence, NUMTs may be amplified by mtDNA PCR primers, leading to incorrect interpretation of data. Because the age of NUMTs is unknown a priori, it is important to investigate the potential for NUMT contamination of mtDNA data. When undetected, NUMTs can cause inaccuracies in estimates derived from mtDNA data. For example, NUMTs can lead to an overestimation of the number of novel mutations on a population scale, potentially resulting in inaccurate inferences related to allelic frequency, population structure, demography, phylogenetic relationships, or misidentification of heteroplasmy (Song et al. 2008; Maude et al. 2019; Schultz and Hebert 2022). Methods to detect NUMTs include bioinformatic prediction (Dayama et al. 2014; Zhou et al. 2020; Hebert et al. 2023) and laboratory-based detection (PCR amplifications) (Machida and Lin 2017; Kuprina et al. 2023), but their implementation is often restricted by cost and efficiency.

Previously, we investigated mitochondrial population genomics of the notorious cane toad (*Rhinella marina*) invasion from its native range in French Guiana to Hawai'i and subsequently to Australia (Cheung et al. 2024). This invasion is a well-known case study of rapid evolution following introduction and has a broad collection of supporting evidence (Rollins et al. 2015; Hudson et al. 2020; Shine and Baeckens 2023). Although cane toads were only introduced to Australia in the 1930s, we detected a large number of mitochondrial genetic variants private to the Australian introduced range (Cheung et al. 2024). The haplotype network formed a star-shaped topology, indicating that these variants arose recently. Such a result could indeed arise from the incorporation of NUMTs, but our explicit search for mitochondrial insertions in the nuclear genome did not identify any (Cheung et al. 2024). However, because genome quality is known to be important to the detection of NUMTs (Triant and Pearson 2022), that study may have been limited by the use of a relatively incomplete draft nuclear genome, hereon referred to as aRhiMar1.2 (Edwards et al. 2018). Furthermore, only a single approach, NUMTFinder (Edwards et al. 2021), was used to detect NUMTs. NUMTFinder is a genome assembly curation tool, designed to identify and differentiate genuine NUMTs from misincorporation of mtDNA into nuclear

genome assemblies. It uses BLASTN to query the nuclear genome for mitochondrial sequences and identify putative NUMT fragments. The fragments are combined into NUMT blocks based on the proximity. Although the use of homology searching is the most widely used approach to detect NUMTs (e.g. Richly and Leister 2004; Hazkani-Covo et al. 2010; Hebert et al. 2023), it has several pitfalls. This method is susceptible to false positives arising from incorrect alignments, especially when dealing with sequences rich in repeats, even when stringent *E*-values are applied. Moreover, in cases where the genome assembly is highly fragmented, NUMTs might have failed to assemble correctly, and searches may fail to detect NUMTs located at assembly gaps. An alternative approach involves mapping raw short-read or long-read sequencing data for the detection process. While harder to implement, this approach offers an advantage by inferring NUMTs within the raw reads themselves based on alignment properties (e.g. orientation or mapping to different chromosomes). This bypasses potential issues related to genome assembly quality or heterozygosity. For example, *dinumt* (Dayama et al. 2014) utilizes paired-end short reads in which the split read pairs map to the nuclear genome and mitochondrial genome respectively. *dinumt* identifies candidate NUMTs by clustering the mapped paired reads based on the mapping positions and orientation. *PALMER* (Zhou et al. 2020) uses long-reads to detect nonreference mobile element insertions (MEI) and, with modification, can be expanded to detect other categories of nonreference insertion (e.g. NUMTs). Combining several of these methods could help to reduce false positives from homology methods and false negatives from read-based methods, potentially leading to more accurate detection of NUMTs than any single approach.

In this study, we aimed to clarify the NUMT landscape in the cane toad genome. Specifically, we significantly improved the quality of the cane toad nuclear genome (aRhiMar1.3) to examine how this affects the detection of NUMTs. Additionally, we compared 2 alternative NUMT detection methods (*dinumt* and *PALMER*) with NUMTFinder. Because little is known about NUMTs in amphibians, we also assessed whether the lack of NUMTs we found in the cane toad genome is representative of other amphibians. This robust approach to NUMT detection furthers our understanding of evolutionary processes in the introduced cane toad population in Australia but also provides a model for other mitogenome studies aiming to remove the potential confound of NUMT inclusion.

Results

Updated Cane Toad Nuclear Genome Assembly

We assembled an updated cane toad genome (aRhiMar1.3) using a combination of newly generated Nanopore [Oxford Nanopore Technologies (ONT)] long reads and 10X Genomics linked-reads, along with the previously published PacBio CLR and Illumina PE data from the same individual. We generated 7,245,603 ONT long reads with 44 Gb of sequences, averaging 5.7 kb in length and an N50 read length of 11,549 bp. Only reads longer than 500 bps were retained, resulting in 43.7 Gb of sequences. These were co-assembled with the original PacBio data using Flye v2.7b (Kolmogorov et al. 2019), yielding an initial assembly of 34,696 contigs with contig N50 of 451 kb and Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness at 80.3%. Additionally, we generated 1,179,094,866 paired 10× linked-reads with 349.41 Gb of sequences for scaffolding and error correction. Multiple rounds of polishing, scaffolding, gap-filling, and tidying (supplementary table S1, Supplementary Material online) generated the final assembled genome with a length of 3.47 Gb on 7,124 scaffolds (13,172 contigs) (Table 1). The final contig N50 was 859 kb, a 5-fold increase compared with the previous draft genome. The scaffold N50 reached 2.5 Mb. DepthSizer analysis of both versions of the genome assembly estimated that the genome size of the cane toad is 3.4 Gb to 3.5 Gb (IndelRatio mode), consistent with the final assembly size.

The quality of the final assembly was assessed with different methods. BUSCO evaluation against the tetrapoda_odb10 dataset ($n = 5310$) revealed that aRhiMar1.3 assembly included 91.3% of the complete conserved single-copy genes (Fig. 1). This is comparable to the 16 other amphibian genomes analyzed, with only *Xenopus* species having notably fewer missing BUSCO genes (supplementary fig. S1, Supplementary Material online). The completeness and quality value (QV) from Merqury analyses of the cane toad genome assembly yielded 95.4 and 32.4, respectively, indicating high completeness and over 99.99% accuracy in nucleotide level (error rate = $5.81e^{-4}$) of the genome assembly.

For DepthKopy analysis (Fig. 2), the mean copy number (CN) of completed BUSCO genes in both versions was close to 1, which indicated true single copies in the assembly. In aRhiMar1.2, the copy number of duplicated BUSCO genes showed a bimodal distribution with peaks at around 0.5 CN and 1 CN, with the former indicating that some false duplications remain in the assembly. In contrast, 709 (2.25%) contigs/scaffolds had a copy number greater than 2, suggesting possible collapsed repeats while 223 (0.71%) had a copy number of 0, suggesting low-quality sequences (Fig. 2 "sequences"). In aRhiMar1.3, the bimodal distribution in "duplicated" was skewed toward 1 CN, suggesting that fewer false duplications appeared in the assembly than in the draft assembly. The copy number analysis of the 100 kb window scan showed a mean value closer to 1 and fewer windows having CN greater than 1, suggesting that most collapsed repeats were also resolved through the

Table 1 Genome assembly statistics and the BUSCO score of the draft genome assembly (aRhiMar1.2) and the updated *Rhinella marina* assembly, aRhiMar1.3

…	aRhiMar1.2	aRhiMar1.3[a]
Total length (bp)	2,551,760,159	3,473,313,001
Number of scaffolds	NA	7,124
Longest scaffold (bp)	NA	15,571,941
Mean scaffold length (bp)	81,287	487,551
Median scaffold length (bp)	38,590	30,657
Scaffold N50 (bp)	NA	2,502,259
Scaffold L50 (bp)	NA	377
Number of contigs	31,192	13,172
Contig N50 (bp)	167,498	859,429
Contig L50 (bp)	3,373	1,052
Gap (N) length (bp)	…	8,571,305 (0.25%)
GC content (%)	43.23	43.75
BUSCO completeness[b]	…	…
Lineage: Tetrapoda (genome mode)	C:85.8% (S: 83.7%, D: 2.1%), F: 4.7%, M: 9.5%, $n = 5,310$	C: 91.3% (S: 89.6%, D: 1.7%), F: 2.7%, M: 6.0%, $n = 5,310$
Lineage: Tetrapoda (proteome mode)	…	C: 93.6% (S: 91.5%, D: 2.1%), F: 1.8%, M: 4.6%

[a]Statistics included mtDNA. [b]BUSCO scores were shown in the following notation: C, complete (S, single, D, duplicated); F, fragmented; M, missing; *n*, gene number.

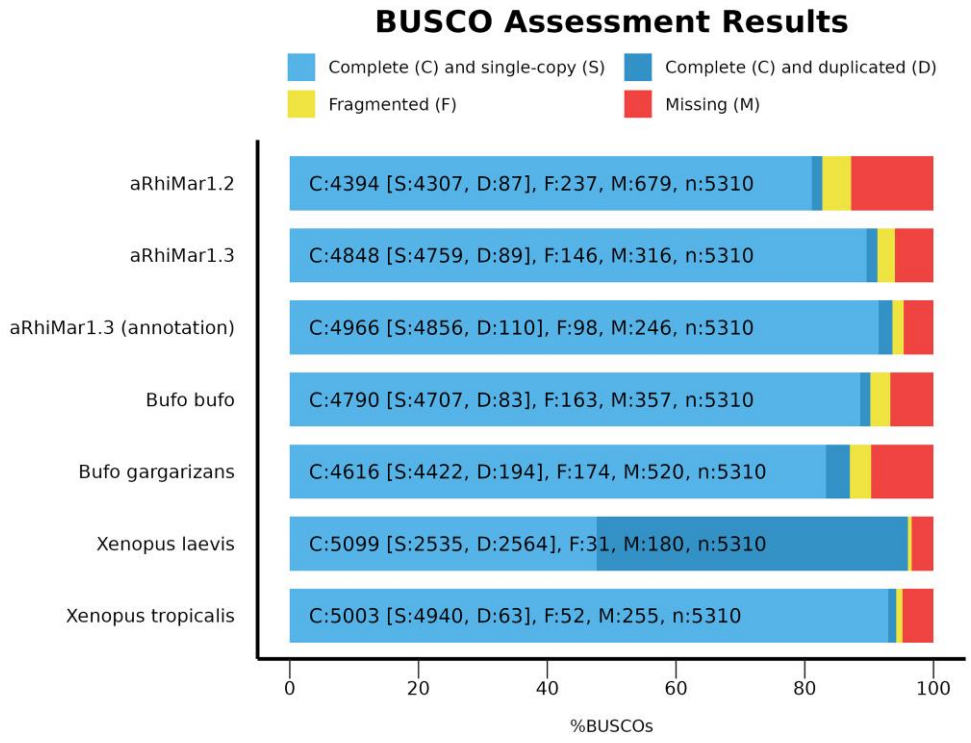

Fig. 1. BUSCO assessment of the draft assembly (aRhiMar1.2), updated (aRhiMar1.3) cane toad (*Rhinella marina*), and 4 other amphibians genome assemblies.

addition of ONT reads. There is still a peak of contigs/scaffolds (Fig. 2 "sequences") with a CN of approximately 0.5, indicating a number of false duplicates or haplotype-specific sequences. These could be the consequence of conservative duplicate removal with Diploidocus but could also represent some sequences specific to sex chromosomes, because a heterogametic ZW female individual was sequenced.

Nuclear Genome Annotation and Evaluation

A total of 39,121 protein-coding genes with a median transcript length of 2,535 bp were predicted in the cane toad genome assembly by GeMoMa across 2,879 scaffolds. BUSCO assessment of the longest protein per gene from the predicted proteome with 5,310 orthologous proteins in Tetrapoda classified 93.6% as "complete," in which

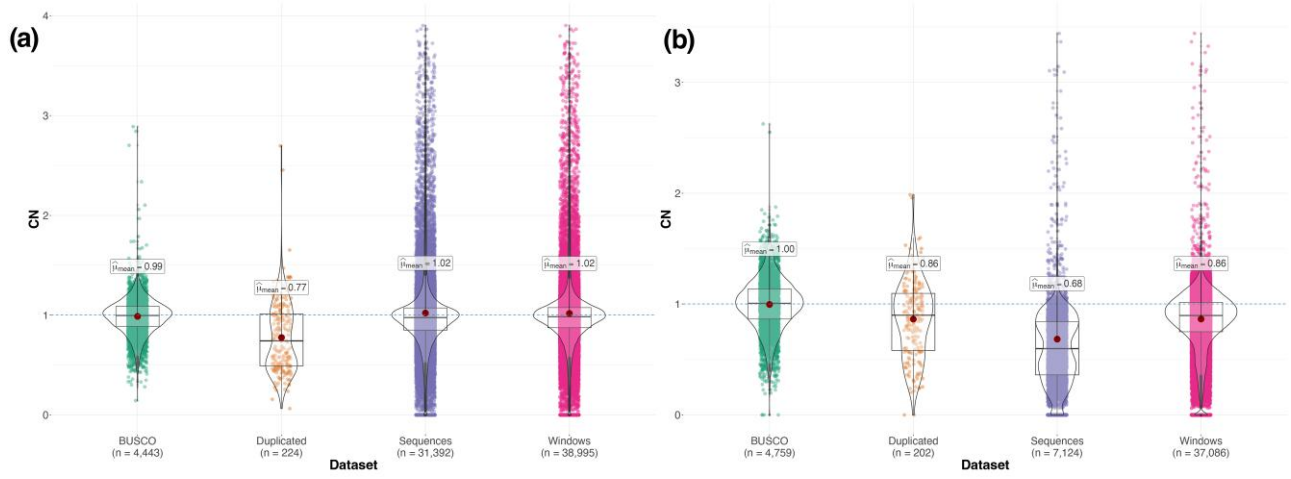

Fig. 2. Copy number analysis of a) the original *Rhinella marina* draft genome assembly (aRhiMar1.2) and b) the updated genome assembly (aRhiMar1.3) based on the BUSCO results by DepthKopy. BUSCO, single-copy BUSCO genes; duplicated, duplicated BUSCO genes; sequences, contigs or scaffolds; windows: 100,000 bps window.

Table 2 Repeat element summary of the draft (aRhiMar1.2) and the updated *Rhinella marina* genome assembly, aRhiMar1.3

Repeats	aRhiMar1.2 (Mb)	aRhiMar1.3[a] (Mb)
Class I TEs—retroelements	247.31	322.56
Class II TEs—DNA transposons	507.37	620.75
Rolling circles	NA	0.22
Unclassified	826	1,468.84
Small RNA	14.69	77.9
Satellites	11.3	4.39
Simple repeats	55.1	20.49
Low complexity	7.3	2.42
Total	1.63 Gb	2.51 Gb

[a]Statistics included mtDNA.

Table 3 NUMTFinder results on the draft (aRhiMar1.2) and updated *Rhinella marina* genome assembly, aRhiMar1.3

…	aRhiMar1.2	aRhiMar1.3
Genome size (Gb)	2.55	3.47
No. of NUMTs (before quality control)	42	64
No. of NUMTs (after quality control—default repeat-masking (interspersed repeats, low-complexity DNA)	0	1

2.1% were "duplicated." Nearly 5% of the proteins were "missing" in the annotation (Fig. 1).

Repeat annotation revealed that 72.48% of the genome sequences are repetitive elements (Table 2). More than half of the repeats were unclassified repetitive elements (42.29%) which is higher than the average percentage (34.22%) reported in published amphibian genomes (Zuo et al. 2023). Of the classified repeats, Class II transposable elements (TEs) (DNA transposons) were the most abundant (17.87%), dominated by hobo-Activator (13.16%). Only 9.29% were classified as Class I TEs (retroelements), dominated by long interspersed nuclear elements (LINEs). A total of 1,338 high-confidence tRNAs were predicted.

NUMT Detection and Validation in the Cane Toad

We chose 3 different approaches to detect NUMTs in the cane toad genome. Using NUMTFinder, a total of 64 putative NUMTs were found in the aRhiMar1.3 (Table 3). These were identified in 63 different scaffolds, and all but 1 hit aligned to the control region of the mitogenome (supplementary table S2, Supplementary Material online). Hit lengths ranged from 38 to 141 base pairs. There was 1 hit on scaffold 1637 that was 114 bp long, which corresponded to a partial *COX1* gene. When the repeats in the assembly were masked using RepeatMasker (default parameters), a total of 8 putative NUMTs remained detected by NUMTFinder where 1 came from a partial *COX1* gene and the remaining 7 came from the control region (supplementary table S3, Supplementary Material online). After visual examination of the sequence composition of the putative NUMTs (supplementary fig. S2, Supplementary Material online), the 7 hits from the control region had clusters of AT microsatellite and extremely low GC content (0% to 7.14%), indicating these sequences were low-complexity regions of the genome assembly. Therefore, we decided to consider these 7 hits from the control region as artifactual hits, along with all the control region hits from the unmasked genome.

Two methods of raw sequencing read analysis to detect NUMTs yielded different results. No NUMTs were detected using *dinumt*. Our analysis using *PALMER* with PacBio and Nanopore sequencing reads yielded nonoverlapping results from the different sequencing technologies (supplementary table S4, Supplementary Material online). In total, 49

(PacBio) and 45 (Nanopore) putative NUMTs were detected, each on a different scaffold. No overlapping putative NUMT was detected in both PacBio and Nanopore data sets. For each of the 94 putative NUMTs detected by *PALMER*, we manually inspected the regions in Integrative Genomic Viewer (IGV). In all cases, the corresponding regions of aRhiMar1.3 were well-supported by numerous spanning reads, showing that the assembly was correctly assembled. In contrast, each putative NUMT candidate was only supported by a single read, with each read partially mapping to a different part of the aRhiMar1.3 assembly. This indicated that the putative NUMTs were likely to be the result of low-frequency events (e.g. chimeric reads generated by library preparation or sequencing steps). As no *PALMER* NUMT candidates were supported by multiple independent reads, we excluded all candidates as false-positive signals.

NUMT Detection in Amphibians

In order to gain insight into the NUMT landscape across amphibians, we ran NUMTFinder on 16 amphibian genomes. Predicted NUMT counts across amphibians ranged from 0 (*Xenopus tropicalis*) to 20,496 (*Leptobrachium leishanense*) (Table 4). We observed similar patterns as in the cane toad genome assemblies, with an abundance of putative NUMTs matching low-complexity control regions. Therefore, we adopted the same filtering regime used in the cane toad and removed putative NUMTs corresponding to short hits to the mitogenome control region. This reduced the number of putative NUMTs by 2 orders of magnitude (0 to 356 NUMTs). Several species showed minimal reduction while, in an extreme case, all 20,496 putative NUMTs of *Leptobrachium leishanense* were filtered. NUMT fragments spanned across the mtDNA genome, from tRNA to protein-coding genes and rRNA (supplementary table S5, Supplementary Material online). Visual inspection of the distribution of NUMT numbers after filtering revealed a dominant cluster of 11 species (65%, including cane toad) with 0 to 4 NUMTs, and the remaining 6 (35%) with 10+. To test the previous relationship of genome size and NUMT count, we therefore divided species into low (<5) and high (≥10) NUMT counts. There was a significant genome size difference between the low and high genomes after filtering (Wilcoxon test, $P = 0.007$), which was not seen with the unfiltered NUMTFinder results (Fig. 3).

Discussion

The presence of NUMTs can significantly confound the results of mtDNA analyses. In the cane toad mitochondrial population genomics study, we did not identify any NUMT insertions (Cheung et al. 2024). However, the results may have been limited by a relatively incomplete draft genome (aRhiMar1.2) and the use of only 1 detection method

Table 4 NUMTFinder (BLASTN) results for 16 other amphibians and *Rhinella marina* (bottom) before and after filtering

Species	Genome accession number	mtDNA accession number	References	Genome size (Gb)	No. of NUMTs	No. of NUMTs (after filtering)
Bombina bombina	GCA_027579735.1	NC_042501.1	(Pabijan et al. 2008; Rhie et al. 2021)	10.02	124	123
Bufo bufo	GCA_905171765.1	LR991678.1	(Streicher et al. 2021b)	5.04	1	1
Bufo gargarizans	GCA_014858855.1	NC_008410.1	(Cao et al. 2006; Lu et al. 2021)	4.55	10	4
Dendropsophus ebraccatus	GCA_027789765.1	CM051072.1	(Rhie et al. 2021)	2.21	87	0
Discoglossus pictus	GCA_027410445.1	CM050223.1	(Rhie et al. 2021)	3.87	356	112
Geotrypetes seraphini	GCA_902459505.2	NC_020155.1	(San Mauro et al. 2009; Ovchinnikov et al. 2023)	3.78	36	33
Hymenochirus boettgeri	GCA_019447015.1	NC_015615.1	(Irisarri et al. 2011; Bredeson et al. 2024)	3.21	5	1
Leptobrachium ailaonicum	GCA_018994145.1	MZ394043.1	(Li et al. 2019b)	3.54	170	2
Leptobrachium leishanense	GCA_009667805.1	NC_031411.1	(Liang et al. 2016; Li et al. 2019a)	3.55	20,496	0
Microcaecilia unicolor	GCA_901765095.2	NC_023515.1	(San Mauro et al. 2014; Ovchinnikov et al. 2023)	4.69	34	33
Pyxicephalus adspersus	GCA_004786255.1	NC_044480.1	(Denton et al. 2018; Cai et al. 2019)	1.56	1	1
Rana temporaria	GCA_905171775.1	NC_042226.1	(Chen 2018; Streicher et al. 2021a)	4.11	27	26
Rhinatrema bivittatum	GCA_901001135.2	NC_006303.1	(San Mauro et al. 2004; Ovchinnikov et al. 2023)	5.32	15	14
Xenopus borealis	GCA_024363595.1	NC_018776.1	(Lloyd et al. 2012; Evans et al. 2022)	2.75	2	1
Xenopus laevis	GCA_017654675.1	NC_001573.1	(Roe et al. 1985; Session et al. 2016)	2.74	19	1
Xenopus tropicalis	GCA_000004195.4	NC_006839.1	(Bredeson et al. 2024)	1.45	0	0
Rhinella marina	GCA_900303285.2	NC_066225.1	This paper, (Cheung et al. 2024)	3.47	64	1

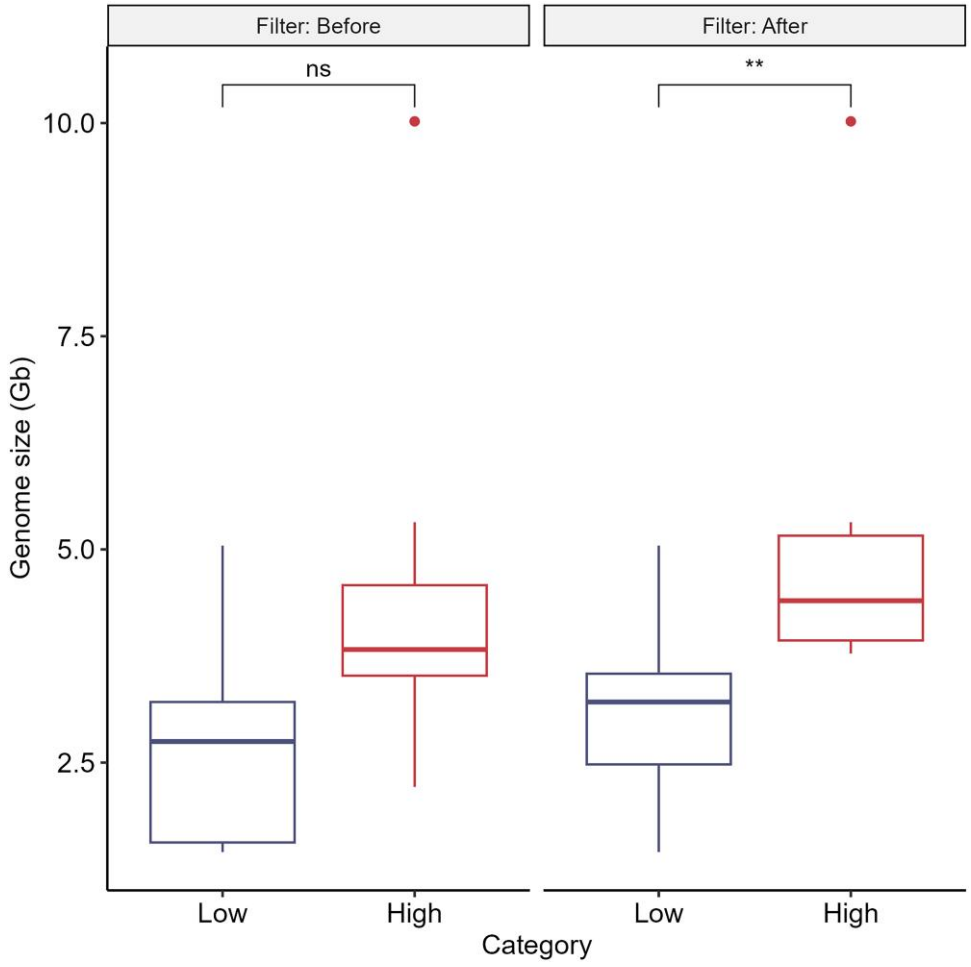

Fig. 3. Relationship between the genome size and the number of NUMT 16 other amphibians and *Rhinella marina* found by NUMTFinder before and after custom filtering. Species with <5 NUMTs are categorized into "low" and those with >5 NUMTs are categorized into "high."

(NUMTFinder). To address these potential limitations, we improved the nuclear genome assembly and incorporated 2 additional NUMT detection methods (*dinumt* and *PALMER*). These new data suggested 1 newly found NUMT in the cane toad genome. Nevertheless, this NUMT appears to be old and divergent, and did not contribute to the haplotypes identified in the invasive cane toad population. Furthermore, we investigated how NUMT detection was affected by repeat content. Masking with RepeatMasker substantially reduced the number of putative NUMT found in several amphibian genomes, consistent with inflation due to the low-complexity regions of the genome. These findings highlight the importance of accounting for repeat regions within the mtDNA when using a homology search approach.

Revision to Cane Toad Nuclear Reference Genome

Due to its repetitive nature, aRhiMar1.2 (Edwards et al. 2018) appears to be highly collapsed, resulting in significant discrepancies in genome size estimates obtained from flow cytometry and densitometry (2.55 Gb vs. 3.98 to 5.65 Gb). To address this issue, we assembled a more contiguous reference genome by leveraging multiple sequencing technologies, including PacBio long reads, Nanopore long reads, linked-read sets (10X Genomics), and short-read Illumina sequences. This combined data set yielded a 3.47 Gb haploid genome assembly. This new assembly has a substantial 0.9 Gb increase in genome size compared to aRhiMar1.2, increasing the previous 167 kb contig N50 to 859 kb and 2.5 Mb for contigs and scaffolds, respectively. This addition in the improved assembly coincides with the identification of an additional 0.88 Gb of repeats (unclassified, 643 Mb; TEs, 189 Mb; small RNA, 63 Mb) (Table 2), suggesting that a large portion of collapsed repeats has been resolved. This observation is supported by DepthKopy analysis, which indicated a reduction in collapsed repeats across different scaffolds (Fig. 2). In addition, the contiguous assembly has enhanced the discovery and identification of repeat elements boundaries and

relationships (e.g. fewer satellites and simple and low-complexity repeats). Detailed characterization of the repeat landscape of the cane toad is beyond the scope of this study. The assembly size fell within the DepthSizer estimation range (3.4 to 3.5 Gb) and was much closer to the estimates obtained from flow cytometry and densitometry. It is notable that the margin of error in genome size estimation by flow cytometry could be significant, and we have previously found it to overestimate genome sizes in other species (Chen et al. 2022a, 2022b).

Beyond contiguity, the completeness and quality of the assembly also improved significantly with BUSCO completeness reaching 91.3%. This marks a substantial increase from 85.8% of aRhiMar1.2 with a 5.9% increase in single-copy BUSCO genes and a 3.5% decrease in missing BUSCO genes (Fig. 1). This indicated that the additional ONT sequencing reads and use of polishing software have facilitated the assembly of more conserved protein orthologs ($n = 452$). Despite the assembly of the improved cane toad genome still being at the scaffold level, the BUSCO score is already comparable to other amphibian genome assemblies at the chromosome level. Specifically, the mean BUSCO of chromosomal-level assembly stands at $87.45 \pm 0.09\%$ while scaffold-level assembly averages $81.56 \pm 0.17\%$ (Zuo et al. 2023). The relatively low BUSCO completeness of amphibian genomes could reflect particular challenges with assembling these repeat-rich species, or it could reflect taxonomic differences in gene content that are not currently captured by the BUSCO datasets. Similarly, the majority of duplicated BUSCO genes in aRhiMar1.3 appear to be genuine duplicates based on read depth (Fig. 2b).

NUMT Detection

Our previous mitochondrial study of the cane toad carefully examined the validity of the nuclear-derived sequencing reads for constructing the mitochondrial genome (Cheung et al. 2024). Here, we aimed to clarify the NUMT landscape by producing an updated genome assembly with higher contiguity and completeness. Using the same method (NUMTFinder) as was used in Cheung et al. (2024), where no NUMTs were detected, with the improved genome, we detected 64 putative NUMTs. After repeat-masking, only 1 from the *COX1* gene remained suggesting that mostly these putative NUMTs from the control region of the mtDNA were highly likely to be false positives. None of the haplotypes identified in introduced populations arose from this *COX1* NUMT. To make sure that the homology search approach was not missing NUMTs, we employed 2 alternative NUMT detection methods, *dinumt*, and *PALMER*. These methods directly search for signatures of NUMT insertions, utilizing raw sequencing reads. *dinumt* yielded no results while all *PALMER* candidates were supported by only a single read and therefore rated as false-positive results. Together, these results indicate that NUMTs are likely absent from the cane toad nuclear genome. Alternatively, it is possible that NUMTs exist at such low frequencies within the mosaic of cells and mitochondria that they could not be detected at the sequencing depth used in our study. Of most importance, there is no evidence to suggest the presence of recent (high identity) NUMT fragments that might confuse the analysis of real mtDNA sequencing. This finding further supports our previous mtDNA analyses of the cane toad and indicates that the new haplotypes identified in introduced populations are likely to be genuine, rather than NUMTs.

Accounting for Repeats in Homology-based NUMT Detection

Repeat sequences pose challenges for homology-based searches. This issue was previously recognized and masking repetitive and low-complexity regions in the nuclear genome prior to NUMT detection has been recommended to address this problem (Kielbasa et al. 2011; Tsuji et al. 2012). However, overzealous masking of repeats could result in missing real NUMTs and may not be necessary where the mtDNA lacks repetitive regions. Amphibian genomes are known to be repeat-rich, ranging from 33.69% to 74.91% in the nuclear genomes (Zuo et al. 2023), and 50% of amphibians have repeats in the control region of the mitochondrial genome (Formenti et al. 2021). So, although care must be taken when masking repeats, it is also expected that random alignment to repeat-rich sequences will be observed when using NUMTFinder or similar techniques. Here, we demonstrate that performing the NUMT search with and without repeat-masking can provide a more nuanced view of the impact of repeats in homology-based NUMT detection.

We extrapolate our experience in detecting NUMT in the cane toad nuclear genome to other amphibian nuclear genomes. Among the 16 amphibian genomes tested, we observed that 10 of the genomes (62.5%) had putative NUMTs originating from the control region of the mitochondrial genome. This result suggests that false-positive signals of NUMTs are common in repeat-rich nuclear genomes, rather than being specific to the cane toad genome. Based on our observations in a taxonomic group with a high-repeat nuclear genome, we advise future research to carefully inspect the results from BLASTN (the most common way to detect NUMTs) or apply multiple methods to validate results. It is technically challenging to distinguish between genuine NUMTs transferred from the control region of the mitochondrial genome and false-positive NUMTs generated by the homology-based search software. For this work, we operate on the assumption genuine NUMTs will also include regions outside the repeat regions.

Nevertheless, there is the risk that genuine NUMTs originating from the control region may inadvertently be removed by repeat-masking, and the strategy employed should consider the relative risks of false positives and false negatives. Considering the trade-off between accuracy and feasibility, more research should focus on developing alternative and less computation-intensive software.

NUMT Observation in Amphibians

There has not been a genome-wide NUMT analysis in class Amphibia to date. The majority of the studies in amphibians have focused on a single mitochondrial gene, such as cytochrome b or cytochrome oxidase I, because of the utility of these genes in phylogenetics and phylogeography (Meng et al. 2014; Vences et al. 2014). Here, we found that the number of NUMTs across amphibians varies, ranging from none to a few hundred, similar to what has been found in avian genomes (Liang et al. 2018; Baltazar-Soares et al. 2023). In general agreement with previous work (Hazkani-Covo et al. 2010), we found a weak positive relationship between genome size and the NUMT content. However, it was clearly nonlinear, and the majority of amphibian genomes had very few NUMTs regardless of genome size (Table 4). These findings suggested each species and taxon should be investigated independently to understand more about the NUMT evolution across the tree of life.

For *Xenopus tropicalis*, the only amphibian with previously documented NUMTs (Hazkani-Covo et al. 2010), we found no NUMTs. The lack of details about how these NUMTs were identified hinders our ability to explain these divergent results. The previous analysis might have been looking for more ancient/degraded NUMTs, while in our study, we used a different methodology and identified NUMTs with relatively high similarity to the mtDNA, which limits results to more recently transferred NUMTs. In addition, different *Xenopus tropicalis* genome assembly versions (v4.0 or earlier vs. v10.0) were used across studies. It could also be that the previously identified NUMTs were assembly artifacts and were resolved during the updated assembly.

It is worth noting that the genome-wide NUMT search in this study, similar to that in avian genomes (Liang et al. 2018), only considers 1 individual of each species for the presence/absence of NUMTs. Large-scale human studies have already shown that the composition of the NUMT landscape varies across different individuals (Wei et al. 2022). In order to understand more about NUMTs in non-model species, multiple individuals should be included in future studies. As the number of high-quality reference genomes continues to grow, it will also be interesting to confirm and then explain apparent differences in NUMT abundance between taxonomic groups.

Conclusion

To validate the absence of NUMTs in the cane toad nuclear genome, we significantly improved the genome assembly in terms of contiguity and completeness. We found 1 NUMT in this assembly across 3 different NUMT prediction methods. This NUMT was short and too divergent from the mitogenome to contribute false haplotypes in population genetics studies. Additionally, we investigated the NUMT landscape in a range of amphibian species, supporting previous observations of a weak relationship between genome size and number/content of NUMTs. We show how repetitiveness of the nuclear genome may confound the number of NUMTs reported in some species, highlighting a need to be clear and consistent in how NUMTs are to be defined if comparison across studies is to be made. Future studies should check for potential false-positive results, a common problem given the nature of these methods.

Materials and Methods

DNA Sequencing

Raw Illumina 2×150 bp paired-end and PacBio continuous long-read genomic sequencing data from the original cane toad genome (Edwards et al. 2018) were downloaded from the National Center for Biotechnology Information (NCBI) BioProject PRJEB24695. The same high-molecular-weight genomic DNA (gDNA) [see Edwards et al. (2018) for details] was used for an additional 10× linked-read and Oxford Nanopore Technologies (ONT) sequencing. A 10X Genomics (Pleasanton, United States) linked-read library was prepared using the Chromium Genome Reagent Kit (v2 chemistry) and sequenced (2× 150 bp paired-end) on the Illumina HiSeq X Ten platform. For ONT sequencing, a combination of 5 libraries was sequenced on PromethION (1) and GridION (4). For PromethION sequencing, 1 μg gDNA was prepared using the Genomic DNA by ligation protocol (SQK-LSK109) according to the manufacturer's instructions. The final library prep was loaded onto a FLO-PRO002 PromethION flow cell (pore version 9.4.1). For GridION sequencing, 4 libraries were made by preparing 400 ng gDNA using the rapid barcoding protocol (SQK-RBK004) according to the standard protocol. Each library was loaded onto a FLO-MIN106 GridION flow cell (R9.4.1 chemistry). The sequencing was run for 72 h. GPU-enabled guppy (v3.3.0) (Wick et al. 2019) base-calling was performed after sequencing (PromethION high-accuracy flip-flop model; config "dna_r9.4.1_450bps_hac.cfg" config). The resulting FASTQ files were combined and filtered with a custom script.

Genome Assembly

The assembly workflow involved assembling a draft long-read assembly, hybrid polishing with long and short reads,

and gap-filling with long reads. Computational tasks were carried out on the computational cluster Katana supported by Research Technology Services at UNSW Sydney (PVC Research Infrastructure, 2010).

Due to the discrepancy in the genome size estimation among cell-based [flow cytometry: 3.98 to 4.90 Gb (MacCulloch et al. 1996; Chipman et al. 2001); densitometry: 4.06 to 5.65 Gb (Goin et al. 1968; Bachmann 1970)], *k*-mer-based, and qPCR approaches [1.77 to 2.3 Gb; 2.38 Gb (Edwards et al. 2018)], we employed an alternative method, DepthSizer v1.8.0 (Chen et al. 2022b), utilizing long-read sequencing depth profiles of single-copy BUSCO genes to estimate the genome size. This estimated the size at approximately 3.5 Gbp, which was used for guiding the subsequent assembly.

The first step was to use Flye v2.7b (Kolmogorov et al. 2019) to assemble both PacBio long-reads and Nanopore long-reads. The initial assembly was tidied with Diploidocus v0.15.0 (Chen et al. 2022b) (default dipcycle mode). Hybrid polishing was performed using HyPo v1.0.3 (Ritu et al. 2019) with 10× chromium reads and mixed long-read data mapped onto the genome using LongRanger v2.2.2 (Marks et al. 2019) and Minimap2 v2.17 (Li 2018), respectively. The HyPo-polished assembly was scaffolded with 10× chromium reads using ARCS v1.1.1 (Yeo et al. 2018) and gap-filled with both PacBio long-reads and Nanopore long-reads using SSPACE-LongRead (March 2021) (Boetzer and Pirovano 2014) and GapFinisher v20190917 (Kammonen et al. 2019). The assembly was corrected with a second round of HyPo hybrid polishing, followed by a final purging of false duplicates and low-quality sequences with Diploidocus. Contigs corresponding to pure mtDNA were identified with NUMTFinder v0.5.1 (Edwards et al. 2021) and removed from the assembly to produce the final nuclear reference genome. The published high-quality mtDNA reference genome (Cheung et al. 2024) was then added back to the genome (aRhiMar1.3).

Genome Annotation and Evaluation

The genome was annotated using the homology-based gene prediction program GeMoMa v1.7beta (Keilwagen et al. 2019) and 10 protein annotations from 10 species {*Anolis carolinensis* [GCA_000090745.2 (Alfoldi et al. 2011)], *Canis lupus familiaris* [GCA_014441545.1], *Danio rerio* [GCA_000002035.4 (Howe et al. 2013)], *Gallus gallus* [GCA_016699485.1], *Homo sapiens* [GCA_000001405.29], *Mus musculus* [GCA_000001635.9 (Church et al. 2011)], *Naja naja* [GCA_009733165.1 (Suryamohan et al. 2020)], *Taeniopygia guttata* [GCA_003957565.2 (Rhie et al. 2021)], *Takifugu rubripes* [GCA_901000725.2], *Xenopus tropicalis* [GCA_000004195.4 (Hellsten et al. 2010)]} available on Ensembl release 103 (Yates et al. 2020). The GeMoMaPipeline function was executed with a maximum intron size of 200 kb and "GeMoMa.Score=ReAlign AnnotationFinalizer.r=SIMPLE AnnotationFinalizer.p=RMARP pc=true o=true" parameters. The longest isoform per gene was extracted from the annotation and assessed using BUSCO v5.4.3 in proteome mode.

Transposable elements (TEs) were discovered using RECON, RepeatScout, and LtrHarvest implemented in RepeatModeler v2.0.3 (Flynn et al. 2020) with RMBlast as the search engine and default parameters to build a de novo repeat library. TEs and repetitive elements were searched using the de novo repeat library using RepeatMasker v4.1.2 with default parameters for protein-coding gene annotation. The annotation table was made using the buildSummary.pl script from RepeatMasker. The tRNAs were annotated using tRNAScan v2.0.11 (Chan et al. 2021) with default parameters. The script "EukHighConfidenceFilter" in tRNAscan was used to filter low-confidence tRNAs.

We assessed the completeness of the genome assembly using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.4.3 (Manni et al. 2021), implementing BLAST+ v2.11.0, HMMer v3.3.2, Metaeuk v6-a5d39d9, and SEPP v4.5.1 with genome mode and the lineage "tetrapoda_odb10" set of 5,310 core genes. Assembly quality (QV) was estimated using a *k*-mer analysis of 10× reads by Merqury v1.3 (Rhie et al. 2020). We also employed DepthKopy v1.3.0 (Chen et al. 2022b) to identify possible over-assembly (collapsed regions) and under-assembly (false duplications or low-quality regions) in the genome assembly.

"Complete" BUSCO genes were compiled across the cane toad and 16 other amphibian genomes (*Bombina bombina*, *Bufo bufo*, *Bufo gargarizans*, *Dendropsophus ebraccatus*, *Discoglossus pictus*, *Geotrypetes seraphini*, *Hymenochirus boettgeri*, *Leptobrachium ailaonicum*, *Leptobrachium leishanense*, *Microcaecilia unicolor*, *Pyxicephalus adspersus*, *Rana temporaria*, *Rhinatrema bivittatum*, *Xenopus borealis*, *Xenopus laevis*, *Xenopus tropicalis*) (Table 4).

NUMT Detection and Validation in the Cane Toad Nuclear Genome

NUMT detection was compared between aRhiMar1.2 (Edwards et al. 2018) and aRhiMar1.3 to determine whether genome assembly quality affects the detection of NUMTs. NUMTFinder (Edwards et al. 2021) is an open-source tool to identify putative NUMTs through BLASTN and collapse nearby NUMT fragments into NUMT blocks. We constructed a species-specific repeat library using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/) and masked aRhiMar1.3 using RepeatMasker with default parameters (Tarailo-Graovac and Chen 2009). We used NUMTFinder v0.5.4 to search for fragments of the high-quality

mitogenome in both full and masked versions of aRhiMar1.3 as described in Cheung et al. (2024). We used *dinumt* package (Dayama et al. 2014) to identify NUMTs with the customized setup "–include_mask –ensembl –mask_filename =refNUMTs.bed." For long-read sequencing reads, we utilized reads from both PacBio and Nanopore technologies, generated for the cane toad genome assembly. We adapted the package *PALMER* (Zhou et al. 2020) which detects nonreference mobile element insertion events with long-read sequencing data. The detection could be expanded to the detection of the NUMT insertion with a customized setup "–ref_ver other –type CUSTOMIZED –-mode raw –TSD_finding FALSE –custom_seq custom_seq.fasta."

To validate the putative NUMTs insertions found by raw sequence-based approaches, we manually inspected the read-to-genome alignment files (BAM files) using Integrative Genomic Viewer (IGV) (Robinson et al. 2011; Thorvaldsdottir et al. 2013). Only insertions supported by at least 10% of the sequencing depth (PacBio: 22×, 3+ reads; ONT: 12.5×, 2+ reads) were considered genuine insertions to minimize false-positive detection.

NUMT Detection in Amphibian Species

To investigate whether the low number of NUMTs in the cane toad genome is a species-specific event or a common occurrence across the family, we analyzed a diverse range of amphibian species with chromosome-level genome assemblies and complete mitochondrial genomes available in NCBI (assessed on 2023-02-01) (Table 4). For each species, we obtained both the nuclear genome and mitochondrial genome using the NCBI Datasets command line tools. Given the significant computational resources required for applying both *dinumt* and *PALMER* to the entire dataset of amphibian raw sequencing data, we opted for the NUMTFinder approach for this analysis. The relationship between genome size and NUMT count was statistically tested with a Wilcoxon rank-sum test using *ggpubr* package (Kassambara 2023) in R (Team 2023).

Supplementary Material

Supplementary material is available at *Genome Biology and Evolution* online.

Acknowledgments

We thank the Ramaciotti Centre for Genomics, University of New South Wales (UNSW), for performing the sequencing in this study. This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney. This study was supported by the Australian Research Council (grant LP180100721 to R.J.E., grant DP160102991 to R.S. and L.A.R.), the UNSW Scientia Program (to L.A.R.), the UNSW Scientia PhD Scholarship (to K.C.), and the Minderoo Foundation (to R.J.E.).

Data Availability

The data underlying this article are available in the European Nucleotide Archive with the study accession PRJEB24695 and assembly accession GCA_900303285.2. Supplementary materials that support the findings of this study are available in the Open Science Framework (OSF) at (https://osf.io/n2dsm/).

Literature Cited

Alfoldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD, et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. Nature. 2011:477(7366):587–591. https://doi.org/10.1038/nature10390.

Bachmann K. Specific nuclear DNA amounts in toads of the genus *Bufo*. Chromosoma. 1970:29(3):365–374. https://doi.org/10.1007/BF00325949.

Ballard JWO, Rand DM. The population biology of mitochondrial DNA and its phylogenetic implications. Annual review of ecology. Evol Syst. 2005:36(1):621–642. https://doi.org/10.1146/annurev.ecolsys.36.091704.175513.

Baltazar-Soares M, Karell P, Wright D, Nilsson JA, Brommer JE. Bringing to light nuclear-mitochondrial insertions in the genomes of nocturnal predatory birds. Mol Phylogenet Evol. 2023:181:107722. https://doi.org/10.1016/j.ympev.2023.107722.

Bensasson D, Zhang D, Hartl DL, Hewitt GM. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 2001:16(6):314–321. https://doi.org/10.1016/s0169-5347(01)02151-6.

Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 2014:15(1):211. https://doi.org/10.1186/1471-2105-15-211.

Bredeson JV, Mudd AB, Medina-Ruiz S, Mitros T, Smith OK, Miller KE, Lyons JB, Batra SS, Park J, Berkoff KC, et al. Conserved chromatin and repetitive patterns reveal slow genome evolution in frogs. Nat Commun. 2024:15(1):579. https://doi.org/10.1038/s41467-023-43012-9.

Cai YY, Shen S-Q, Lu L-X, Storey KB, Yu D-N, Zhang J-Y. The complete mitochondrial genome of *Pyxicephalus adspersus*: high gene rearrangement and phylogenetics of one of the world's largest frogs. PeerJ. 2019:7:e7532. https://doi.org/10.7717/peerj.7532.

Cao SY, Wu X-B, Yan P, Hu Y-L, Su X, Jiang Z-G. Complete nucleotide sequences and gene organization of mitochondrial genome of *Bufo gargarizans*. Mitochondrion. 2006:6(4):186–193. https://doi.org/10.1016/j.mito.2006.07.003.

Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res. 2021:49(16):9077–9096. https://doi.org/10.1093/nar/gkab688.

Chen JJ. The complete mitochondrial genome of common terrestrial frog (*Rana temporaria*). Mitochondrial DNA B Resour. 2018:3(2):978–979. https://doi.org/10.1080/23802359.2018.1507649.

Chen SH, Martino AM, Luo Z, Schwessinger B, Jones A, Tolessa T, Bragg JG, Tobias PA, Edwards RJ. A high-quality pseudo-phased genome for *Melaleuca quinquenervia* shows allelic diversity of NLR-type resistance genes. Gigascience. 2022a:12:giad102. https://doi.org/10.1093/gigascience/giad102.

Chen SH, Rossetto M, van der Merwe M, Lu-Irving P, Yap J-YS, Sauquet H, Bourke G, Amos TG, Bragg JG, Edwards RJ, et al. Chromosome-level de novo genome assembly of *Telopea speciosissima* (New South Wales waratah) using long-reads, linked-reads and Hi-C. Mol Ecol Resour. 2022b:22(5):1836–1854. https://doi.org/10.1111/1755-0998.13574.

Cheung K, Amos TG, Shine R, DeVore JL, Ducatez S, Edwards RJ, Rollins LA. Whole-mitogenome analysis unveils previously undescribed genetic diversity in cane toads across their invasion trajectory. Ecol Evol. 2024:14(3):e11115. https://doi.org/10.1002/ece3.11115.

Chipman AD, Khaner O, Haas A, Tchernov E. The evolution of genome size: what can be learned from anuran development? J Exp Zool. 2001:291(4):365–374. https://doi.org/10.1002/jez.1135.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen H-C, Agarwala R, McLaren WM, Ritchie GRS, et al. Modernizing reference genome assemblies. PLoS Biol. 2011:9(7):e1001091. https://doi.org/10.1371/journal.pbio.1001091.

Dayama G, Emery SB, Kidd JM, Mills RE. The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res. 2014:42(20):12640–12649. https://doi.org/10.1093/nar/gku1038.

Dayama G, Zhou W, Prado-Martinez J, Marques-Bonet T, Mills RE. Characterization of nuclear mitochondrial insertions in the whole genomes of primates. NAR Genom Bioinform. 2020:2(4):lqaa089. 10.1093/nargab/lqaa089.

Denton RD, Kudra RS, Malcom JW, Preez LD, Malone JH. The African bullfrog (*Pyxicephalus adspersus*) genome unites the two ancestral ingredients for making vertebrate sex chromosomes. bioRxiv 329847. https://doi.org/10.1101/329847, 20 November 2018, preprint: not peer reviewed.

Edwards RJ, Field MA, Ferguson JM, Dudchenko O, Keilwagen J, Rosen BD, Johnson GS, Rice ES, Hillier LD, Hammond JM, et al. Chromosome-length genome assembly and structural variations of the primal Basenji dog (*Canis lupus familiaris*) genome. BMC Genomics. 2021:22(1):188. https://doi.org/10.1186/s12864-021-07493-6.

Edwards RJ, Tuipulotu DE, Amos TG, O'Meally D, Richardson MF, Russell TL, Vallinoto M, Carneiro M, Ferrand N, Wilkins MR, et al. Draft genome assembly of the invasive cane toad, *Rhinella marina*. Gigascience. 2018:7(9):giy095. https://doi.org/10.1093/gigascience/giy095.

Evans BJ, Mudd AB, Bredeson JV, Furman BLS, Wasonga DV, Lyons JB, Harland RM, Rokhsar DS. New insights into *Xenopus* sex chromosome genomics from the Marsabit clawed frog *X. borealis*. J Evol Biol. 2022:35(12):1777–1790. https://doi.org/10.1111/jeb.14078.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020:117(17):9451–9457. https://doi.org/10.1073/pnas.1921046117.

Formenti G, Rhie A, Balacco J, Haase B, Mountcastle J, Fedrigo O, Brown S, Capodiferro MR, Al-Ajli FO, Ambrosini R, et al. Complete vertebrate mitogenomes reveal widespread repeats and gene duplications. Genome Biol. 2021:22(1):120. https://doi.org/10.1186/s13059-021-02336-9.

Goin OB, Goin CJ, Bachmann K. DNA and amphibian life history. Copeia. 1968:1968(3):532. https://doi.org/10.2307/1442021.

Gray MW. Mitochondrial evolution. Cold Spring Harb Perspect Biol. 2012:4(9):a011403. https://doi.org/10.1101/cshperspect.a011403.

Hazkani-Covo E, Covo S. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet. 2008:4(10):e1000237. https://doi.org/10.1371/journal.pgen.1000237.

Hazkani-Covo E, Zeller RM, Martin W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 2010:6(2):e1000834. https://doi.org/10.1371/journal.pgen.1000834.

Hebert PDN, Bock DG, Prosser SWJ. Interrogating 1000 insect genomes for NUMTs: a risk assessment for estimates of species richness. PLoS One. 2023:18(6):e0286620. https://doi.org/10.1371/journal.pone.0286620.

Hellsten U, Harland RM, Gilchrist MJ, Hendrix D, Jurka J, Kapitonov V, Ovcharenko I, Putnam NH, Shu S, Taher L, et al. The genome of the western clawed frog *Xenopus tropicalis*. Science. 2010:328(5978):633–636. https://doi.org/10.1126/science.1183670.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013:496(7446):498–503. https://doi.org/10.1038/nature12111.

Hudson CM, Vidal-Garcia M, Murray TG, Shine R. The accelerating anuran: evolution of locomotor performance in cane toads (*Rhinella marina*, Bufonidae) at an invasion front. Proc Biol Sci. 2020:287(1938):20201964. https://doi.org/10.1098/rspb.2020.1964.

Irisarri I, Vences M, San Mauro D, Glaw F, Zardoya R. Reversal to air-driven sound production revealed by a molecular phylogeny of tongueless frogs, family Pipidae. BMC Evol Biol. 2011:11(1):114. https://doi.org/10.1186/1471-2148-11-114.

Kammonen JI, Smolander O-P, Paulin L, Pereira PAB, Laine P, Koskinen P, Jernvall J, Auvinen P. gapFinisher: a reliable gap filling pipeline for SSPACE-LongRead scaffolder output. PLoS One. 2019:14(9):e0216885. https://doi.org/10.1371/journal.pone.0216885.

Kassambara A. ggpubr: 'ggplot2' based publication ready plots. R package version 0.6.0. https://rpkgs.datanovia.com/ggpubr/2023.

Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. Methods Mol Biol. 2019:1962:161–177. https://doi.org/10.1007/978-1-4939-9173-0_9.

Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011:21(3):487–493. https://doi.org/10.1101/gr.113985.110.

Kivisild T. Maternal ancestry and population history from whole mitochondrial genomes. Investig Genet. 2015:6(1):3. https://doi.org/10.1186/s13323-015-0022-2.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019:37(5):540–546. https://doi.org/10.1038/s41587-019-0072-8.

Kuprina K, Smorkatcheva A, Rudyk A, Galkina S. Numerous insertions of mitochondrial DNA in the genome of the northern mole vole, *Ellobius talpinus*. Mol Biol Rep. 2023:51(1):36. https://doi.org/10.1007/s11033-023-08913-4.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018:34(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li J, Yu H, Wang W, Fu C, Zhang W, Han F, Wu H. Genomic and transcriptomic insights into molecular basis of sexually dimorphic nuptial spines in *Leptobrachium leishanense*. Nat Commun. 2019a:10(1):5551. https://doi.org/10.1038/s41467-019-13531-5.

Li Y, Ren Y, Zhang D, Jiang H, Wang Z, Li X, Rao D. Chromosome-level assembly of the mustache toad genome using third-generation DNA sequencing and Hi-C analysis. Gigascience. 2019b:8(9):giz114. https://doi.org/10.1093/gigascience/giz114.

Liang B, Wang N, Li N, Kimball RT, Braun EL. Comparative genomics reveals a burst of homoplasy-free numt insertions. Mol Biol Evol. 2018:35(8):2060–2064. https://doi.org/10.1093/molbev/msy112.

Liang XX, Shu G-C, Wang B, Jiang J-P, Li C, Xie F. Complete mitochondrial genome of the Leishan moustache toad, *Vibrissaphora*

leishanensis (Anura: Megophryidae). Mitochondrial DNA B Resour. 2016:1(1):275–276. https://doi.org/10.1080/23802359.2016.1159937.

Lloyd RE, Foster PG, Guille M, Littlewood DT. Next generation sequencing and comparative analyses of *Xenopus* mitogenomes. BMC Genomics. 2012:13(1):496. https://doi.org/10.1186/1471-2164-13-496.

Lu B, Jiang J, Wu H, Chen X, Song X, Liao W, Fu J. A large genome with chromosome-scale assembly sheds light on the evolutionary success of a true toad (*Bufo gargarizans*). Mol Ecol Resour. 2021:21(4):1256–1273. https://doi.org/10.1111/1755-0998.13319.

MacCulloch RD, Upton DE, Murphy RW. Trends in nuclear DNA content among amphibians and reptiles. Comp Biochem Physiol Part B Biochem Mol Biol. 1996:113(3):601–605. https://doi.org/10.1016/0305-0491(95)02033-0.

Machida RJ, Lin YY. Occurrence of mitochondrial CO1 pseudogenes in *Neocalanus plumchrus* (Crustacea: Copepoda): hybridization indicated by recombined nuclear mitochondrial pseudogenes. PLoS One. 2017:12(2):e0172710. https://doi.org/10.1371/journal.pone.0172710.

Manni M, Berkeley MR, Seppey M, Simao FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021:38(10):4647–4654. https://doi.org/10.1093/molbev/msab199.

Marks P, Garcia S, Barrio AM, Belhocine K, Bernate J, Bharadwaj R, Bjornson K, Catalanotti C, Delaney J, Fehr A, et al. Resolving the full spectrum of human genome variation using linked-reads. Genome Res. 2019:29(4):635–645. https://doi.org/10.1101/gr.234443.118.

Maude H, Davidson M, Charitakis N, Diaz L, Bowers WHT, Gradovich E, Andrew T, Huntley D. NUMT confounding biases mitochondrial heteroplasmy calls in favor of the reference allele. Front Cell Dev Biol. 2019:7:201. https://doi.org/10.3389/fcell.2019.00201.

Meng H, Li X, Qiao P. Population structure, historical biogeography and demographic history of the alpine toad *Scutiger ningshanensis* in the Tsinling Mountains of Central China. PLoS One. 2014:9(6):e100729. https://doi.org/10.1371/journal.pone.0100729.

Ovchinnikov V, Uliano-Silva M, Wilkinson M, Wood J, Smith M, Oliver K, Sims Y, Torrance J, Suh A, McCarthy SA, et al. Caecilian genomes reveal the molecular basis of adaptation and convergent evolution of limblessness in snakes and caecilians. Mol Biol Evol. 2023:40(5):msad102. https://doi.org/10.1093/molbev/msad102.

Pabijan M, Spolsky C, Uzzell T, Szymura JM. Comparative analysis of mitochondrial genomes in *Bombina* (Anura; Bombinatoridae). J Mol Evol. 2008:67(3):246–256. https://doi.org/10.1007/s00239-008-9123-3.

PVC Research Infrastructure. 2010. Katana. https://doi.org/10.26190/669x-a286

R Core Team. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. https://www.R-project.org/.2023.

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021:592(7856):737–746. https://doi.org/10.1038/s41586-021-03451-0.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020:21(1):245. https://doi.org/10.1186/s13059-020-02134-9.

Richly E, Leister D. NUMTs in sequenced eukaryotic genomes. Mol Biol Evol. 2004:21(6):1081–1084. https://doi.org/10.1093/molbev/msh110.

Ritu K, Joshua C, Wing-Kin S 2019. HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. bioRxiv 882506. https://doi.org/10.1101/2019.12.19.882506, 20 December 2019, preprint: not peer reviewed.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol. 2011:29(1):24–26. https://doi.org/10.1038/nbt.1754.

Roe BA, Ma DP, Wilson RK, Wong JF. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J Biol Chem. 1985:260(17):9759–9774. https://doi.org/10.1016/S0021-9258(17)39303-1.

Rollins LA, Richardson MF, Shine R. A genetic perspective on rapid evolution in cane toads (*Rhinella marina*). Mol Ecol. 2015:24(9):2264–2276. https://doi.org/10.1111/mec.13184.

San Mauro D, Gower DJ, Massingham T, Wilkinson M, Zardoya R, Cotton JA. Experimental design in caecilian systematics: phylogenetic information of mitochondrial genomes and nuclear rag1. Syst Biol. 2009:58(4):425–438. https://doi.org/10.1093/sysbio/syp043.

San Mauro D, Gower DJ, Müller H, Loader SP, Zardoya R, Nussbaum RA, Wilkinson M. Life-history evolution and mitogenomic phylogeny of caecilian amphibians. Mol Phylogenet Evol. 2014:73:177–189. https://doi.org/10.1016/j.ympev.2014.01.009

San Mauro D, Gower DJ, Oommen OV, Wilkinson M, Zardoya R. Phylogeny of caecilian amphibians (Gymnophiona) based on complete mitochondrial genomes and nuclear RAG1. Mol Phylogenet Evol. 2004:33(2):413–427. https://doi.org/10.1016/j.ympev.2004.05.014.

Schultz JA, Hebert PDN. Do pseudogenes pose a problem for metabarcoding marine animal communities? Mol Ecol Resour. 2022:22(8):2897–2914. https://doi.org/10.1111/1755-0998.13667.

Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, Fukui A, Hikosaka A, Suzuki A, Kondo M, et al. Genome evolution in the allotetraploid frog *Xenopus laevis*. Nature. 2016:538(7625):336–343. https://doi.org/10.1038/nature19840.

Shine R, Baeckens S. Rapidly evolved traits enable new conservation tools: perspectives from the cane toad invasion of Australia. Evolution. 2023:77(8):1744–1755. https://doi.org/10.1093/evolut/qpad102.

Singh KK, Choudhury AR, Tiwari HK. Numtogenesis as a mechanism for development of cancer. Semin Cancer Biol. 2017:47:101–109. https://doi.org/10.1016/j.semcancer.2017.05.003.

Song H, Buhay JE, Whiting MF, Crandall KA. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc Natl Acad Sci U S A. 2008:105(36):13486–13491. https://doi.org/10.1073/pnas.0803076105.

Streicher JW, Wellcome Sanger Institute Tree of Life p, Wellcome Sanger Institute Scientific Operations DNAPc, Tree of Life Core Informatics c, Darwin Tree of Life C. The genome sequence of the common frog, *Rana temporaria* Linnaeus 1758. Wellcome Open Res. 2021a:6:286. https://doi.org/10.12688/wellcomeopenres.17296.1.

Streicher JW, Wellcome Sanger Institute Tree of Life p, Wellcome Sanger Institute Scientific Operations DNAPc, Tree of Life Core Informatics c, Darwin Tree of Life C. The genome sequence of the common toad, *Bufo bufo* (Linnaeus, 1758). Wellcome Open Res. 2021b:6:281. https://doi.org/10.12688/wellcomeopenres.17298.1.

Suryamohan K, Krishnankutty SP, Guillory J, Jevit M, Schröder MS, Wu M, Kuriakose B, Mathew OK, Perumal RC, Koludarov I, et al. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. Nat Genet. 2020:52(1):106–117. https://doi.org/10.1038/s41588-019-0559-8.

Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009: 25:4 10 11–14 10 14. https://doi.org/10.1002/0471250953.bi0410s25.

Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013:14(2):178–192. https://doi.org/10.1093/bib/bbs017.

Triant DA, Pearson WR. Comparison of detection methods and genome quality when quantifying nuclear mitochondrial insertions in vertebrate genomes. Front Genet. 2022:13:984513. https://doi.org/10.3389/fgene.2022.984513.

Tsuji J, Frith MC, Tomii K, Horton P. Mammalian NUMT insertion is non-random. Nucleic Acids Res. 2012:40(18):9073–9088. https://doi.org/10.1093/nar/gks424.

Vences M, de Pous P, Nicolas V, Díaz-Rodríguez J, Donaire D, Hugemann K, Hauswaldt JS, Amat F, Barnestein JAM, Bogaerts S, et al. New insights on phylogeography and distribution of painted frogs (*Discoglossus*) in Northern Africa and the Iberian Peninsula. Amphib Reptil. 2014:35(3):305–320. https://doi.org/10.1163/15685381-00002954.

Wei W, Schon KR, Elgar G, Orioli A, Tanguy M, Giess A, Tischkowitz M, Caulfield MJ, Chinnery PF. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. Nature. 2022:611(7934): 105–114. https://doi.org/10.1038/s41586-022-05288-7.

Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019:20(1): 129. https://doi.org/10.1186/s13059-019-1727-y.

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, et al. Ensembl 2020. Nucleic Acids Res. 2020:48(D1):D682–D688. https://doi.org/10.1093/nar/gkz966.

Yeo S, Coombe L, Warren RL, Chu J, Birol I. ARCS: scaffolding genome drafts with linked reads. Bioinformatics. 2018:34(5):725–731. https://doi.org/10.1093/bioinformatics/btx675.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. Comparative genomics reveals insights into avian genome evolution and adaptation. Science. 2014:346(6215):1311–1320. https://doi.org/10.1126/science.1251385.

Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, Moran JV, Mills RE. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res. 2020:48(3):1146–1163. https://doi.org/10.1093/nar/gkz1173.

Zuo B, Nneji LM, Sun YB. Comparative genomics reveals insights into anuran genome size evolution. BMC Genomics. 2023:24(1):379. https://doi.org/10.1186/s12864-023-09499-8.

Associate editor: Toni Gossmann