

Additional file 2

Supplementary materials and methods

Sample collection and genome project overview

The genome of *Pristimantis euphronides*, a critically endangered frog endemic to Grenada, was sequenced from a single female individual designated “GrenadaFrog144.” A female specimen was selected due to its larger body size and the putatively associated higher tissue and genomic DNA yield. The individual was sampled on 19 September 2024 near Grand Etang Forest, Grenada (12.09388, -61.69583), during the evening after sunset when frogs are most active. At the time of sampling, the frog weighed 4.65 g and had a snout-vent length of 35 mm.

All sample collection procedures were approved by the Institutional Animal Care and Use Committee (IACUC) of the School of Veterinary Medicine, St. George’s University, Grenada (protocol IACUC-24010-R, approved 18 September 2024). Sampling was conducted using non-lethal sampling procedures, in accordance with established conservation guidelines for critically endangered anuran species [1].

The frog was captured by hand using new, non-latex disposable gloves and placed in a clean, disposable plastic bag for weighing and visual examination, including assessment of body condition and general demeanor to confirm good health. Weighing was performed while the animal remained in the bag using a mechanical tubular spring scale. The frog was then removed from the bag and manually restrained. Two buccal swabs were collected by gently rubbing sterile synthetic-tipped swabs along the inner surfaces of the oral cavity, avoiding contact with the tongue and minimizing handling time.

The toes selected for clipping were disinfected using a cotton-tip applicator soaked in diluted povidone-iodine solution. One large toe from the hind limb and one toe from the fore limb were excised using sterile, sharp scissors. A small volume of blood was collected opportunistically from each toe-clipping site and absorbed onto a sterile swab, after which each surgical site was disinfected again with diluted

povidone-iodine solution and checked to confirm hemostasis and normal responsiveness prior to release. The frog was released at the exact location of capture.

Sampling consisted of two toe clips, two buccal swabs, and two blood samples, with the latter collected opportunistically when mild hemostasis was required following toe clipping. Each buccal swab and clipped toe was immediately placed in previously prepared screw cap O-ring 2.0 ml microtubes (Cat. No. I2020.0500, Genaxxon, Ulm, Germany) in 100 µl Monarch® gDNA Cell Lysis Buffer (Cat. No. T3012L, New England Biolabs, Ipswich, MA, USA). Similarly, each blood swab was immersed in 100 µl Monarch® gDNA Blood Lysis Buffer (Cat. No. T3013L, New England Biolabs, Ipswich, MA, USA). In the field and during transport to the laboratory, all samples were kept cool in a new styrofoam box on ice packs for less than 1.5 hours before being stored at -80 °C until processing.

DNA extraction and quality control

Genomic DNA of the female “GrenadaFrog144” was extracted separately from each of the six collected samples using a modified phenol-chloroform protocol. Given the small size of the individual and the limited amount of biological material available per sample, as well as the species’ critically endangered status, each extraction was performed independently to ensure quality control and preserve redundancy in case of downstream processing failure.

Sample lysis was initiated by immediately adding 600 µl of pre-warmed (56 °C) Tris-HCl-EDTA lysis buffer (10 mM Tris-HCl, 1 mM EDTA, 1% SDS, pH 8.0) to the frozen sample-containing tubes upon removal from the -80 °C ultrafreezer. The swab or tissue and entire ~700 µl lysis volume (100 µl original storage buffer + 600 µl added lysis buffer) were transferred into Rotaprep™ Lysing Matrix D 2.0 ml microtubes (Rotaprep Inc., Tustin, CA, USA) containing approximately ten 1.4 mm zirconium beads. Samples were homogenized using a Rotaprep™ MonoLyzer™ (Rotaprep Inc., Tustin, CA, USA) five times for 5 seconds at 4,000 rpm. Protein digestion was performed by adding 20 µl of Proteinase K (Cat. No. P8200AAVIAL, New England Biolabs, Ipswich, MA, USA), followed by incubation for ~3 h at 56 °C on a rotating/rocking heat block. Samples were briefly vortexed at least once every hour and again at the end of incubation.

Following digestion, plastic swab sticks were removed from the oral and blood swab samples, and the lysate was transferred to fresh 2.0 ml microtubes. Residual bead matrix and swab fibers were compressed using a micropipette tip, and the tubes were centrifuged at 14,800 rpm for 3 min. This process was repeated until no additional liquid could be recovered. The combined supernatant for each aliquot was centrifuged again at 14,800 rpm until clear, and the clarified lysate was transferred to 2.0 ml DNA LoBind® tubes (Cat. No. 022431048, Eppendorf, Hamburg, Germany). Toe clip samples were processed by centrifugation at 14,800 rpm for 10 min, and the resulting clear supernatant was likewise transferred to DNA LoBind® tubes.

RNA was removed by adding 6 µl Monarch® RNase A (Cat. No. T3018-2, New England Biolabs, Ipswich, MA, USA) and incubating for 5 min at 56 °C. After a 3 min spin at 14,800 rpm, the supernatant was transferred to a fresh DNA LoBind® tube. DNA was purified by phase separation: 700 µl of phenol:chloroform:isoamyl alcohol v/v/v/ 25:24:1 (Cat. No. 516726 2025, Merck KGaA, Darmstadt, Germany) was added, mixed gently for 1 min, and centrifuged for 10 min at room temperature. The upper aqueous phase was transferred to a new DNA LoBind® tube, followed by an equal-volume wash with chloroform:isoamyl alcohol (24:1, mixed freshly at 0.96:0.04 ratio), gently inverted for 1 min, and centrifuged again for 10 min.

The final aqueous phase was transferred to a clean DNA LoBind® tube. DNA was precipitated by adding 2 volumes of absolute ethanol (≥99.5%, Cat. No. 1.07017, Merck KGaA, Darmstadt, Germany), 0.1 volume of 3 M sodium acetate (Cat. No. 567418-M, Merck KGaA, Darmstadt, Germany), and 2 µl glycogen (20 mg/ml, Cat. No. M6015.0250, Genaxxon, Ulm, Germany). Samples were incubated at -80 °C overnight and centrifuged at 14,800 rpm for 30 min. After removing the supernatant, pellets were washed with freshly prepared 70% ethanol, gently inverted, and centrifuged again at 14,800 rpm for 5 min. Ethanol was carefully removed, followed by a 1 min spin and final removal of residual ethanol using a 200 µl micropipette. Pellets were air-dried briefly (~1 min) to avoid over-drying. DNA was eluted in 60 µl of Monarch® Elution Buffer (preheated to 56 °C) and fully dissolved by incubating for 2 h at 56 °C on a rocking heat block.

DNA concentration was measured using a Qubit™ 4.0 Fluorometer (Cat. No. Q33238, ThermoFisher Inc., Waltham, MA, USA) with the dsDNA HS Assay Kit (Cat. No. Q32854, ThermoFisher Inc.,

Waltham, MA, USA), and purity was assessed with a NanoDrop™ 2000 spectrophotometer (Cat. No. ND2000USCAN, ThermoFisher Inc., Waltham, MA, USA).

Library preparation and genome sequencing

A total of six DNA extracts - two each from toe clips, blood swabs, and oral swabs - were prepared independently, as described above. Sequencing libraries were constructed for 12 individual Oxford Nanopore Technologies, Oxford, UK (ONT) MinION flow cells with 1 µg of input DNA each (Table SM1). Extracts yielding sufficient DNA were used to prepare multiple libraries, while extracts with lower yields were selectively pooled with residual DNA from other sample types from the single sampled individual to meet input thresholds. One toe clip sample was used in two libraries; the other was partially combined with remaining DNA from a blood sample for one library and with combined DNA extractions from oral and blood samples for another library. One blood sample contributed to two libraries, and the other to three. One oral swab sample was used in three libraries, while the second was pooled with residual blood and toe clip DNA for one library, as mentioned before.

Table SM1. Summary of sequencing libraries and read statistics for high-accuracy basecalled data. Composition of sequencing libraries generated from the single individual *Pristimantis euphronides* (“GrenadaFrog144”). ENA Run ID: European Nucleotide Archive [2] run accession. Sample type: tissue source of the DNA extract. Library ID: internal identifier assigned to each library preparation. Number of reads: total number of basecalled reads passing quality filters. Total bp: total number of base pairs sequenced. Average read length and maximum read length are shown in base pairs (bp). N50: length for which the collection of all reads of that length or longer contains at least 50% of total bases. Q20% and Q30%: percentage of bases with Phred quality score $Q \geq 20$ or $Q \geq 30$, respectively.

ENA Run ID	Sample Type	Library ID	Number of reads	Total bp	Average read length	Maximum read length	N50 (bp)	Q20%	Q30%
------------	-------------	------------	-----------------	----------	---------------------	---------------------	----------	------	------

ERR14924614	Blood	GrenadaFrog144_blood1_run1	826,758	3,838,422,088	4642.7	673,750	8,440	79.49	63.36	
ERR14936814	Blood	GrenadaFrog144_blood1_partlib2_run1	1,986,199	7,538,269,886	3795.3	314,687	7,038	85.86	72.79	
ERR14927616	Blood	GrenadaFrog144_blood2_run1_libpart_1	1,961,799	6,095,143,972	3106.9	262,895	3,881	84.60	70.81	
ERR14936801	Blood	GrenadaFrog144_blood2_run1_libpart_2	707,039	2,441,621,935	3453.3	168,111	4,088	83.43	68.91	
ERR14936815	Blood	GrenadaFrog144_blood2_partlib2_run1	2,369,568	7,735,283,298	3264.4	314,906	3,953	85.63	72.40	
ERR14936802	Oral	GrenadaFrog144_oral1_run1	1,393,540	4,357,823,292	3127.2	254,370	6,371	82.88	68.28	
ERR14936803	Oral	GrenadaFrog144_oral1_run1_2	887,402	2,349,233,601	2647.3	275,131	3,605	78.65	62.22	
ERR14936816	Oral	GrenadaFrog144_oral1_partlib2_run1	1,549,135	6,183,350,360	3991.5	411,486	6,504	84.38	70.48	
ERR14936804	Toe	GrenadaFrog144_toe2_run1_libpart_1	1,124,420	3,103,737,003	2760.3	283,059	4,705	81.83	66.65	
ERR14936786	Toe	GrenadaFrog144_toe2_run1_libPart02	2,390,681	6,570,179,769	2748.2	365,487	4,613	80.47	64.58	
ERR14936785	Toe & Blood	GrenadaFrog144ToeBlood_01	1,946	1,880,199	966.2	347,661	155,440	67.69	43.79	
ERR14936787	Oral & Blood & Toe	GrenadaFrog144_oral_blood_toe_mix_lib1	1,872,757	5,380,231,212	2872.9	424,872	3,711	81.40	66.12	
Total			GrenadaFrog144_ONT_HAC_all	17,071,244	55,595,176,615	3256.7	673,750	4,842	83.14	68.68

Library preparation was conducted using the Ligation Sequencing Kit V14 (SQK-LSK114; ONT) following the manufacturer's protocol [3]. Briefly, for each library, 1.0 µg of genomic DNA underwent end-repair using the NEBNext® Ultra™ II End Repair/dA-Tailing Module (Cat. No. E7546; New England Biolabs (NEB), Ipswich, MA, USA), followed by purification with AMPure XP beads (Cat. No. A63880; Beckman Coulter, Brea, CA, USA). Sequencing adapters were ligated using the NEBNext® Quick Ligation Module (Cat. No. E6056; NEB), with a subsequent AMPure XP bead cleanup. The final library was eluted in the kit's elution buffer, quantified using the Qubit™ dsDNA HS Assay Kit (Cat. No. Q32854; Thermo Fisher Scientific, Waltham, MA, USA), and loaded onto a R10.4.1 flow cell primed as per the Ligation Sequencing Kit V14 protocol. Details of sample

composition per library, European Nucleotide Archive (ENA) [2] run accession numbers, and sequencing metrics are provided in Table SM1.

Following library preparation, sequencing was performed using two ONT MinION Mk1B devices (Cat. No. MIN-101B, ONT) loaded with R10.4.1 flow cells and operated with Kit 14 chemistry according to the manufacturer's protocol. Each device was controlled via MinKNOW software version 24.06.16 [4] running on a dedicated high-performance laptop (Captiva, Dachau, Germany) connected via USB 3.0. Each laptop was configured with Ubuntu 22.04.5 LTS, a 16-core CPU, 64 GB RAM, a 2 TB SSD, and an NVIDIA GeForce RTX 3060 graphics processing unit (GPU; 6 GB video random-access memory (VRAM); driver version 535.183.01; Compute Unified Device Architecture (CUDA) 12.2). Real-time basecalling during sequencing was performed on-GPU using Dorado Basecall Server v7.4.12 [5] in FAST mode with the model `dna_r10.4.1_e8.2_400bps_fast@v4.3.0`, generating signal-level .pod5 files and .fastq raw read files (FASTQ format, [6]). Adapter sequences and ONT-specific artificial sequences introduced during ligation-based library preparation were automatically trimmed by MinKNOW during both real-time FAST and following high-accuracy (HAC) basecalling. Reads with an average ONT quality score below Q 8 (FAST) or below Q 9 (HAC) were filtered out automatically by MinKNOW based on model-specific default thresholds [7].

After completion of sequencing, HAC basecalling was performed offline on-GPU on the same laptops using the model `dna_r10.4.1_e8.2_400bps_hac@v4.3.0`. The GPU performance during both real-time and offline basecalling was monitored using `nvidia-smi` (v535.183.01). Only the HAC .fastq read files were used for the assembly workflow and downstream analysis. Genome sequencing and assembly details are summarized in Table SM2. The raw sequencing data and study metadata were deposited in the European Nucleotide Archive (ENA) [2] under the project number PRJEB89028.

Table SM2. Genome sequencing and assembly metadata. Sequencing library types, platform, coverage, assembly tools, and associated database accessions for the genome of the individual “GrenadaFrog144” (*Pristimantis euphronides*). NCBI: National Center for Biotechnology Information. ENA: European Nucleotide Archive.

Property	Term
Finishing quality	High-quality scaffold-level draft genome
Libraries used	Oxford Nanopore Technologies: single end ligation library
Sequencing platforms	Oxford Nanopore Technologies MinION Mk1B
Fold coverage	31.8× ("Definition of coverage metrics": <i>(ii) bases per assembly length</i>)
Assembly and Scaffolding Tools	Workflow 1: Flye v2.9.5, Medaka v2.0.1, (HapDup v0.12,) and RagTag v2.1.0 Workflow 2: modified CSA v2.6 (Wtdbg v1.0, Ragout v2.3)
Gene calling method	HANNO v0.4
ENA/GenBank Accession	GCA_965278355.2
Date of Release	May 12, 2025 (NCBI), May 10, 2025 (ENA)
ENA Bioproject ID	PRJEB89028

Remote computational resources

All computational analyses apart from real-time and offline basecalling (see "Library preparation and genome sequencing") and assembly polishing with Medaka (see "Workflow 1" in "Genome assembly") were performed on a dedicated remote physical server (Scaleway, Paris, France) running Ubuntu 24.04 LTS (Noble Numbat). The system was equipped with an AMD EPYC™ 7272 12-core processor (24 threads total), 256 GB RAM, and two 1.9 TB NVMe SSDs configured in a RAID 1 mirrored array, yielding ~1.8 TB of redundant storage. A 1 Gbps Ethernet connection provided consistent network

throughput, and private networking (RPNv2) was enabled for data isolation. The root and boot partitions were configured with RAID 1 protection, while large-scale data processing was performed on the /data mount, also mirrored. The server was accessed remotely from Grenada via Secure Shell (SSH).

General software environment and data availability

All analyses on the remote server were performed under Python v3.11.8 [8], with Biopython v1.84 [9], NumPy [10], SciPy [11], pandas [12], and matplotlib [13] and R v4.3.3 [14] with ggplot2 [15]. Bash pipelines were executed under GNU Bash v5.2.21 [16] with GNU parallel 20231122 ('Grindavík') [17] for task parallelisation. All analytical steps not attributed to a specific published tool in the sections below were implemented in custom Python, R, or Bash scripts. All custom scripts, input data, intermediate results, and final output files are archived in Zenodo (doi: 10.5281/zenodo.15298547) and all scripts are additionally available at GitHub (<https://github.com/koppk/Grenada-Frog-144>).

Read quality analysis

Summary statistics for HAC basecalled reads from the 12 sequencing runs (Table SM1), which were merged into a combined dataset for downstream analysis, were computed using SeqKit v2.10.0 [18]. As SeqKit operates at the base level ("per-base Q"), it was used to calculate the proportion of sequenced bases with Phred quality scores ($Q = -10 \log_{10} p_{\text{error}}$) [19], [20] exceeding fixed thresholds (Q20, Q30). To compare overall read length and quality characteristics between real-time (FAST) and post-run (HAC) Dorado basecalling modes (orchestrated in MinKNOW), NanoComp 1.24.2 from the NanoPack suite [21] was applied to the respective datasets. Unlike SeqKit, NanoComp summarizes per-read metrics ("mean read Q") and reports the number and proportion of reads whose mean quality exceeds specified thresholds (Q10, Q15, Q20). While both quality assessment tools use Phred-scaled Q values written by the Dorado basecaller, SeqKit captures base-wise quality distributions, whereas NanoComp summarizes read-wise quality. These two complementary approaches were selected to separately capture base-wise quality distributions and read-wise quality profiles. As a third metric, the

per-read arithmetic mean of Phred quality scores was computed for the HAC dataset. For each read, all per-base Phred scores were summed and divided by the number of bases; these per-read means were then averaged across all reads (per-read mean). In addition, all per-base Phred scores across all reads were summed and divided by the total number of bases (read-length-weighted mean). These are the same calculations applied per contig in the assembly quality assessment (see below).

Definition of coverage metrics

Three complementary coverage metrics were used in this study: (i) *expected depth* = total HAC bases divided by the GenomeScope 2.0 [22] genome-size estimate (see Materials and Methods section “Genome Size and Complexity Estimation”); (ii) *bases per assembly length* = total HAC bases divided by the final scaffolded assembly size (see “Reference-guided scaffolding of genome assemblies”); and (iii) *mapped mean depth* = the genome-wide mean of mosdepth v0.3.11 [23] per-window (10 kb) coverage after aligning HAC reads to the scaffolded assembly (see “Read mapping and genome coverage analysis”). Unless explicitly stated otherwise, the term “coverage” refers to metric (ii), *bases per assembly length*.

Genome size and complexity estimation

Reference-free analysis of quality-filtered reads was performed using k-mer frequency profiling (k=21) with Jellyfish v2.3.1 [24], followed by GenomeScope 2.0 modeling with a diploid (2p) model to estimate genome size, heterozygosity, repeat content, and sequencing error rate. Genome coverage (“Definition of coverage metrics”: (i) *expected depth*) was estimated by dividing the total number of high-accuracy basecalled bases by the genome size predicted using GenomeScope 2.0. Read-based G+C content was estimated by calculating the proportion of guanine and cytosine bases relative to the total number of sequenced bases in the HAC read set.

Taxonomic classification of reads

To detect potential non-host sequences such as environmental or pathogenic microorganisms and to obtain an initial indication of taxonomic composition, FAST basecalled Nanopore reads were classified using Kraken2 (v2.1.2) [25]. A custom Kraken2 database was built using the `kraken2-build` command by combining two datasets: 15 chromosome-level Hyloidea genome assemblies representing nine species across five families (Bufonidae, Dendrobatidae, Eleutherodactylidae, Hylidae, and Leptodactylidae), and the preconfigured Kraken2 PlusPF reference database (release date: Dec 28, 2024, [26]), which included archaeal, bacterial, viral, fungal, protozoan, and plasmid genomes along with UniVec_Core for vector contamination screening. The 15 Hyloidea assemblies were downloaded from the NCBI Genome database as of December 12, 2024 [27] and comprised *Eleutherodactylus coqui* (GCF_035609145.1, GCA_035609135.1, GCA_019857665.1), *Engystomops pustulosus* (GCA_040894005.1, GCA_040894015.1, GCA_019512145.1), *Leptodactylus fuscus* (GCA_031893025.1, GCA_031893055.1), *Dendropsophus ebraccatus* (GCF_027789765.1), *Hyla sarda* (GCF_029499605.1), *Bufo bufo* (GCF_905171765.1), *Bufo gargarizans* (GCF_014858855.1), *Bufotes viridis* (GCA_033119425.1, GCA_037900795.1), and *Ranitomeya imitator* (GCF_032444005.1). Classification was carried out with a confidence threshold of 0.2.

Genome assembly

The quality-filtered and adapter-trimmed HAC basecalled Nanopore reads were assembled without prior contaminant filtering and scaffolded using two independent workflows to assess assembly quality in the absence of a chromosome-level reference genome for the genus *Pristimantis* or the family Strabomantidae. Only Workflow 1 was used for detailed follow-up analyses, while Workflow 2 served as an independent comparison point for scaffold structure and annotation outcomes.

Workflow 1 involved *de novo* assembly using Flye v2.9.5 [28], a long-read genome assembler optimized for high-error long reads, repeat-rich regions, and high heterozygosity. The assembly was run with the `--nano-hq` flag to reflect the input of HAC basecalled Nanopore reads, using 24 CPU threads (`--threads 24`). The number of iterations was set to 3 (`--iterations 3`) to increase consensus accuracy.

The Flye assembly was polished using Medaka v2.0.1 [29], a neural network-based consensus tool designed for Nanopore assemblies. The polishing run utilized the HAC model r1041_e82_400bps_hac_v4.3.0, corresponding to the read chemistry used in sequencing. A batch size of 32 was set with -b 32 to balance GPU memory usage, and 16 CPU threads were allocated (-t 16). Because Medaka requires GPU acceleration, polishing was performed on one of the two GPU-equipped laptops (see “Library preparation and genome sequencing”), using the same NVIDIA GeForce RTX 3060 GPU employed for basecalling. Medaka’s -x flag enabled automatic selection of available GPU resources.

Workflow 2 followed a modified version of the Chromosome-Scale Assembly (CSA v2.6) pipeline [30], utilizing Wtdbg v1.0 [31] for *de novo* assembly and Ragout v2.3 [32], [33] for reference-guided scaffolding, using the NCBI RefSeq [34] *Eleutherodactylus coqui* haplotype 1 genome (RefSeq ID: GCF_035609145.1) as reference. Results from this workflow were not included in all primary assembly statistics but were mainly retained for structural comparison and gene annotation benchmarking at later stages. Summary assembly statistics including contig and scaffold counts, total length, N50, and GC content were computed for each assembly and its corresponding scaffolded version with SeqKit (stats -a).

Assembly quality assessment

Consensus quality and structural integrity of the Medaka-polished Flye assembly (Workflow 1) of *P. euphronides* were assessed prior to scaffolding using two independent approaches. Inspector v1.2 [35] was run in reference-free mode with --datatype nanopore, --min_contig_length 3000, and --min_contig_length_assemblyerror 50000, using HAC reads in FASTQ format with quality scores preserved. Merqury v1.3 [36] estimated consensus quality from a 21-mer database constructed from the same read set using meryl v1.4.1 [37], run in non-trio haploid mode with automatic ploidy-depth estimation.

All HAC reads were mapped to the pre-scaffolding assembly using Minimap2 v2.28-r1209 [38] (map-ont preset), sorted and indexed with SAMtools v1.19.2 [39]. Global mapping statistics (total reads

mapped, bases mapped, error rate) were computed with SAMtools stats. Mean read-to-assembly alignment identity was computed with NanoComp. Per-contig mean coverage was computed with mosdepth (--fast-mode, --no-per-base). Per-contig GC content was computed with SeqKit (fx2tab -gc).

Per-contig quality metrics were computed from the same alignment in five categories: coverage uniformity, mapping quality, read base quality, variant density, and alignment characteristics. Coverage uniformity was quantified as the coefficient of variation (CV; standard deviation/mean) of mosdepth coverage in non-overlapping 1 kb windows per contig. Per-contig mean mapping quality was computed from the MAPQ field (SAM column 5) of all alignments to each contig using SAMtools view, together with the fraction of alignments with MAPQ = 0 and MAPQ $\geq$ 60. Per-contig read base quality was computed from the Phred-scaled quality scores retained in the BAM (SAM column 11). For each alignment, the arithmetic mean of all per-base Phred scores was calculated; these per-read values were then averaged per contig as a simple mean across reads and as a base-weighted mean. This arithmetic mean differs from the error-probability-weighted mean reported by NanoComp (see “Read quality analysis”), which converts each base Phred score to an error probability before averaging.

Per-contig variant density was estimated by calling variants with BCFtools v1.19 [39] mpileup (minimum mapping quality 10, minimum base quality 20, maximum pileup depth 200) and BCFtools call (-mv), parallelized across 24 contig subsets using GNU parallel. Variants were classified as heterozygous (0/1) or homozygous-alternate and counted per contig to obtain variant density (variants per kb) and heterozygous variant density (het per kb). Soft-clipping rate was computed per contig as the fraction of query-consuming bases that were soft-clipped, parsed from the CIGAR string of each alignment using SAMtools view. Supplementary and secondary alignment rates were computed from SAM flags 2048 and 256, respectively. Contigs meeting two or more of four criteria (CV > 1, mean MAPQ < 10, soft-clip fraction > 0.3, variant density > 20 per kb) were flagged for cross-metric quality assessment.

Gene-space completeness of both pre-scaffolding assemblies (Workflow 1 and Workflow 2) was evaluated using the HANNO pipeline v0.4 [40] with identical evidence inputs. External protein evidence was provided through a curated dataset of Hyloidea protein sequences. This dataset was generated by concatenating predicted protein sets from seven high-quality Hyloidea frog assemblies available (as of April 2025) in NCBI RefSeq that had complete genome assemblies and BUSCO-

evaluated annotations: GCF_014858855.1 (*Bufo gargarizans*), GCF_040894005.1 (*Engystomops pustulosus*), GCF_905171765.1 (*Bufo bufo*), GCF_032444005.1 (*Ranitomeya imitator*), GCF_035609145.1 (*Eleutherodactylus coqui*), GCF_027789765.1 (*Dendropsophus ebraccatus*), and GCF_029499605.1 (*Hyla sarda*). Transcript evidence was provided through a corresponding dataset of mRNA sequences from the same seven RefSeq genome annotations.

The combined protein dataset was mapped to the assembly using minimap v0.13-r248 [41] to guide gene model prediction. Transcript sequences were aligned to the assembly using Minimap2 in splice-aware mode (-x splice) with splice junction guidance (--junc-bed). Transcript assemblies were generated and merged using StringTie v2.2.3 [42] and TACO v0.6.2 [43]. Open reading frames (ORFs) were predicted with TransDecoder v5.7.1 [44] and scored based on LAST v1595 [45] alignments to the Hyloidea input protein set. Gene model completeness was assessed using BUSCO v3.0.2 [46] against the tetrapoda_odb10 dataset (n = 5,310 orthologs) [47].

To assess structural concordance between the two independent assemblies, the Workflow 2 pre-scaffolding assembly (Wtdbg) was aligned as query against the Workflow 1 pre-scaffolding assembly (Flye/Medaka) as target using D-Genies v1.5.0 [48] with embedded Minimap2 and default parameters. The alignment was repeated with query and target reversed to assess directionality effects on unmatched sequence counts. D-Genies output files including whole-genome dotplots, contig-to-contig association tables, and unaligned sequence lists were retained for both comparisons.

Reference genome selection via k-mer-based distance estimation

To identify an appropriate reference genome for scaffolding, 21 publicly available Hyloidea genome assemblies (16 chromosome-level and five scaffold-level, including haplotype-specific and primary genomes) were downloaded from NCBI Genome (accessed April 2025). The set covered 14 species across six of the 22 families comprising the superfamily Hyloidea: Aromobatidae, Bufonidae, Dendrobatidae, Eleutherodactylidae, Hylidae, and Leptodactylidae. The assemblies comprised *Anomaloglossus baeobatrachus* (GCA_048569485.1, GCA_048569465.1), *Bufo bufo* (GCF_905171765.1), *Bufo gargarizans* (GCF_014858855.1), *Bufo viridis* (GCA_033119425.1, GCA_037900795.1), *Eleutherodactylus coqui* (GCF_035609145.1, GCA_035609135.1,

GCA_019857665.1), *Engystomops pustulosus* (GCF_040894005.1, GCA_040894015.1, GCA_019512145.1), *Hyla sarda* (GCF_029499605.1), *Leptodactylus fallax* (GCA_947044405.1), *Leptodactylus fuscus* (GCA_031893025.1, GCF_031893055.1), *Leptodactylus rhodomystax* (GCA_046270665.1), *Leptodactylus rugosus* (GCA_046271005.1), *Osteocephalus oophagus* (GCA_046270745.1), *Ranitomeya imitator* (GCF_032444005.1), and *Scinax boesemani* (GCA_046270805.1).

Genomic distances between each candidate reference and the primary Medaka-polished Flye assembly as well as the phased assemblies (see “Haplotype Phasing and Dual Assembly” in Additional file 4) of the “GrenadaFrog144” sequencing data were estimated using Mash v2.3 [49]. Sketches were generated using default parameters ($k = 21$, sketch size = 1,000), and pairwise distances were computed to quantify sequence divergence. The resulting Mash distance matrix was used to identify the reference genomes with the lowest estimated divergence to the primary and both haplotype assemblies. These genomes were then selected as scaffold references in subsequent masking and alignment steps.

Reference genome selection via phylogenetic proximity

To complement the k-mer-based distance estimation, phylogenetic proximity between *P. euphronides* and anuran species with chromosome-level genome assemblies was assessed using the time-calibrated phylogeny of Portik et al. (2023) [50]. The species-level dated tree (5,242 tips) was obtained from the associated data repository [51]. A list of 46 anuran species with chromosome-level genome assemblies, representing 24 families, was compiled from NCBI Genome (accessed January 16, 2026).

Each species was matched to the Portik et al. (2023) [50] phylogeny by species name and its divergence time to *P. euphronides* was calculated as the age of the most recent common ancestor (tMRCA) in million years (Ma) using the time-calibrated branch lengths. In parallel, topological distance was quantified as the number of edges (nodes) along the shortest path connecting *P. euphronides* to each other tip in the tree [52]. Phylogeny traversal and MRCA extraction were performed using Biopython.

Species with chromosome-level genomes were ranked by phylogenetic proximity to *P. euphronides* based on tMRCA. The phylogenetic gap between the closest available reference and the closest relative lacking a chromosome-level genome was calculated to quantify the data limitation for reference-guided scaffolding.

Reference-guided scaffolding of genome assemblies

To scaffold the primary Medaka-polished Flye-assembly and the phased assemblies of the “GrenadaFrog144” sequencing data to near-chromosome scale, RagTag v2.1.0 [53] was used. For scaffolding the primary genome assembly of *P. euphronides*, the soft-masked chromosome-level genome of *Eleutherodactylus coqui* haplotype 1 (RefSeq ID: GCF_035609145.1) was selected as the reference based on sequence homology and availability of curated annotation from the NCBI RefSeq database. Each of the two haplotype-resolved assemblies were scaffolded individually against the soft-masked chromosome-level genome of *Eleutherodactylus coqui* haplotype 2 (GenBank ID: GCA_035609135.1) based on the lowest sequence divergence found by using Mash (“Reference-Guided Scaffolding of Haplotype Assemblies” in Additional file 4). Repeat masking of these reference genomes was performed using WindowMasker v1.0.0 [54] with default parameters prior to scaffolding to prevent anchoring of contigs to repetitive regions during alignment.

The primary assembly and both haplotype-resolved assemblies were scaffolded independently using RagTag with default alignment parameters (-x asm5) via the embedded Minimap2 aligner. RagTag output included the respective scaffolded genome FASTA [55] file, AGP [56] file, gap coordinates BED [57] file, and placement summary statistics file. RagTag scaffold names, derived from the *E. coqui* reference sequence identifiers (NC_ and NW_ accessions), were renamed to scaffold_1 through scaffold_31 following *E. coqui* chromosome size order. Contigs shorter than 200 bp (n = 12) were removed from the final assembly.

Structural summary statistics (e.g., scaffold counts, N50, gap content) were extracted using SeqKit with the -a flag, the Seqtk v1.4-r122 [58] comp command, and RagTag’s internal metrics. The final scaffolded assemblies were evaluated and compared to their pre-scaffold counterparts based on standard metrics including N50, total length, maximum scaffold length, and gap content.

To assess the structural composition of the final RagTag-scaffolded primary assembly, contigs were classified by their AGP-file assignment into three groups: the 13 longest scaffolds, any shorter scaffolds, and unplaced contigs. Per-group contig counts, total lengths, mean lengths, and percentages of the total ungapped scaffolded assembly length were computed for all three groups. Length distribution statistics and a log-transformed boxplot comparing contigs in the 13 longest scaffolds with contigs in any shorter scaffolds were generated from the placed groups. Final genome coverage was calculated by dividing the total number of HAC basecalled bases by the total length of the scaffolded genome assembly, including gap sequences introduced during scaffolding ("Definition of coverage metrics": (ii) *bases per assembly length*).

Read mapping and genome coverage analysis

To evaluate sequencing depth and genome-wide read support, HAC Nanopore reads were aligned to the primary scaffolded *P. euphronides* genome using Minimap2 with the map-on preset optimized for Nanopore data. The resulting alignments were sorted and indexed using SAMtools. Coverage analysis was performed using mosdepth. We computed average depth in fixed, non-overlapping 10 kb windows across the genome using the --by 10000 parameter. The output BED file provided window-specific depth values which were used to derive summary statistics: mean genome-wide coverage ("Definition of coverage metrics": (iii) mapped mean depth), percentage of 10 kb windows with $\geq 20\times$ coverage, and percentage of 10 kb windows with depth below $10\times$. These statistics were computed both across all windows in the assembly (genome-wide) and restricted to the 13 longest scaffolds. To assess scaffold-specific support, coverage metrics were additionally aggregated per scaffold.

For each of the longest 13 scaffolds by length, we calculated the mean coverage from all 10 kb windows assigned to that scaffold, enabling comparison of depth ranges across near-chromosomal-scale scaffolds. Mosdepth-derived read depth profiles were integrated with scaffold layout information from the RagTag AGP file.

Absolute midpoint coordinates were calculated for each 10 kb window based on cumulative scaffold offsets. To distinguish between local fluctuations and broader coverage patterns, two levels of

smoothing were applied to the coverage values using uniform filters with window sizes of 2 and 1000 coverage bins, corresponding to smoothing windows of approximately 20 kb and 10 Mb, respectively.

Scaffold-specific mean coverage was computed from the high-resolution smoothed profiles. These values were used to generate a genome-aligned coverage plot across scaffolds 1-13, as well as individual scaffold plots using consistent coordinate scaling. Coverage anomalies were identified by applying a uniform filter of 1000 bins (10 Mb) to mosdepth-derived 10 kb coverage estimates.

For each scaffold, deviations from the smoothed profile were classified as high or low if the smoothed depth exceeded or fell below the scaffold-wide mean by more than 5 $\times$, respectively. Contiguous regions ≥ 300 kb with consistent deviations were retained and reported. Anomalous regions were extracted, and all positional coordinates were assigned relative to absolute genome positions.

Evaluation of reference-guided scaffolding

To evaluate the *E. coqui* reference-guided RagTag scaffolding of the Medaka-polished Flye assembly of *P. euphronides*, a three-tier validation was applied: (i) alignment of the pre-scaffolding, reference-naïve assembly contigs to *E. coqui*, (ii) four-species pairwise synteny analysis of scaffold composition, and (iii) contig allegiance analysis to classify individual breakpoints within compound scaffolds as predicted fusions or predicted scaffolding artifacts.

For tier (i), the pre-scaffolding Medaka-polished Flye assembly (34,106 contigs; contig N50 = 302 kb) was aligned to the *E. coqui* haplotype 1 reference genome (GCF_035609145.1) using D-Genies with embedded Minimap2 and default parameters. As a matched control, the scaffolded assembly was aligned against the same target using identical parameters. D-Genies output files including whole-genome dotplots, per-chromosome alignment plots, scaffold-to-chromosome association tables, and unaligned sequence lists were retained for both comparisons. RagTag scaffold composition was assessed from the AGP file. The number of contigs placed per scaffold was extracted.

For tier (ii), six pairwise whole-genome alignments were performed among the scaffolded *P. euphronides* assembly ($n = 16$; [59]) and three hyloid frogs with chromosome-level genome assemblies: *Eleutherodactylus coqui* haplotype 1 ($n = 13$; GCF_035609145.1), *Engystomops*

pustulosus maternal assembly (n = 11; GCF_040894005.1), and *Dendropsophus ebraccatus* paternal assembly (n = 15; GCF_027789765.1). All pairwise comparisons were performed using D-Genies with default Minimap2 parameters. The pre-scaffolding *P. euphronides* assembly was additionally aligned to both *E. coqui* and *D. ebraccatus*. The resulting Pairwise mApping Format (PAF) alignment files were parsed to quantify aligned bases between all chromosome pairs across each comparison (minimum alignment length ≥ 1 Mb). For each *P. euphronides* scaffold, the number of *D. ebraccatus* chromosomes contributing ≥ 5 Mb of aligned sequence was determined. Scaffolds receiving substantial alignments from two or more *D. ebraccatus* chromosomes were classified as compound. Cross-validation was performed by testing whether the same *D. ebraccatus* chromosome combinations were detected in the *P. euphronides* vs. *E. pustulosus* comparison.

For tier (iii), the pre-scaffolding Medaka-polished Flye assembly was aligned to the *D. ebraccatus* genome using Minimap2 with the same D-Genies parameters (minimum alignment length ≥ 5 kb, minimum mapping quality ≥ 10). Each contig placed within a compound scaffold (as determined from the RagTag AGP file) was assigned a primary *D. ebraccatus* chromosome allegiance based on its alignment profile. For each breakpoint identified in the four-species analysis, contigs within a ± 5 Mb zone around the predicted breakpoint position were examined for their *D. ebraccatus* chromosome allegiance. A mixing score was computed as the proportion of the minority *D. ebraccatus* chromosome in the zone (0.0 = pure single-chromosome zone; 0.5 = equal representation of both *D. ebraccatus* chromosomes). Breakpoints with a mixing score > 0.2 in the ± 5 Mb zone were classified as predicted fusions; breakpoints with a mixing score < 0.1 were classified as predicted scaffolding artifacts. A wider ± 15 Mb window and transition sharpness analysis were applied as supporting criteria.

To assess whether scaffold-level concordance between the two workflows was maintained after independent reference-guided scaffolding, the Workflow 2 post-scaffolding assembly (Ragout) was aligned as query against the Workflow 1 post-scaffolding assembly (RagTag) as target using D-Genies with embedded Minimap2 and default parameters.

Alignment with previously published *Pristimantis euphronides* and *Eleutherodactylus johnstonei* sequences

As of March 20, 2025, only three nucleotide sequences for *P. euphronides* were available in GenBank, originally published by [60] based on morphological and molecular identification. These sequences, derived from Sanger-sequenced PCR amplicons, included two nuclear loci and one mitochondrial fragment. The nuclear loci were a 578 bp partial coding sequence of recombination activating gene 1, *RAG-1*, GenBank accession EF493427.1, and a 493 bp fragment spanning exon 1 and partial coding sequence of the Tyrosinase precursor gene, *Tyr*, GenBank accession EF493489.1. The mitochondrial fragment, 2,570 bp in total, GenBank accession EF493527.1, contained partial tRNA-Phe, complete 12S rRNA, complete tRNA-Val, and partial 16S rRNA.

These three sequences, sampled from a different individual (“voucher BWMC6918”) were used as molecular references to assess sequence similarity with the assembled genome of the frog individual (“GrenadaFrog144”) of this study. Given the absence of additional genomic or barcoding data for *P. euphronides* in public repositories, and the presence of morphologically similar, co-occurring species such as *Eleutherodactylus johnstonei*, these reference sequences represented the only available markers for comparative alignment.

Each of the three reference sequences was aligned to the Medaka-polished Flye assembly contigs using Minimap2 with the -x map-ont preset and default parameters, optimized for Nanopore long-read data. Contigs with significant alignments were subsequently queried using Basic Local Alignment Search Tool, nucleotide-nucleotide (BLASTn v2.14.1) [61] against the NCBI Nucleotide database [62], accessed on March 28, 2025, to assess concordance with the published *P. euphronides* sequences.

To provide a differential comparison against the only other morphologically similar anuran species present on Grenada, the same three assembly contigs identified by Minimap2 were additionally queried via BLASTn against the NCBI Nucleotide database with the organism field restricted to *Eleutherodactylus johnstonei* (taxid:350008). For each contig, the top-scoring alignment was retained.

Annotation of scaffolded genome assembly

Genome annotation of the Medaka-polished, RagTag-scaffolded primary Flye assembly of *P. euphronides* was performed using the HANNO v0.4 pipeline with the same Hyloidea protein and mRNA evidence datasets and identical parameters as described under "Assembly quality assessment". Annotation outputs included Gene Transfer Format (GTF) files and corresponding Coding Sequence (CDS), mRNA, and protein FASTA sequences for downstream analyses. Gene models were counted per scaffold group by filtering the HANNO BESTMODELS GFF3 on sequence identifier, distinguishing the 13 longest scaffolds, any shorter scaffolds, and unplaced contigs.

A genome-wide summary of structural and functional annotation categories was computed from the HANNO BESTMODELS annotation in bedDB format, a 12-column BED file variant used by HANNO to store gene models with exon block coordinates and CDS boundaries. Coordinate-identical duplicate entries, defined as records sharing the same chromosome, start, end, and strand, were first excluded. Structural coordinates and functional annotation fields were then parsed to extract feature counts, compute genomic spans in base pairs, and calculate coverage relative to the total scaffolded genome size of 1,748,533,034 bp, including RagTag-inserted 100 bp gaps.

All categories were processed independently from raw coordinates or annotation tags, with no reuse of values across categories. For protein-coding genes, the total count and genomic span were obtained from mRNA start and end coordinates per gene model. The corresponding CDS features were computed by intersecting exon blocks, parsed from the bedDB blockSizes and blockStarts fields, with the annotated CDS boundaries, CDSstart and CDSend, for each gene model. Exons were counted using the blockCount field, and their lengths were derived by summing the blockSizes entries per gene. Mean exons per gene, mean CDS length, and mean mRNA length were computed by dividing the respective totals by the number of gene models. Functional annotation categories were assessed using fields generated by eggNOG-mapper v2.1.13 [63], integrated as a component of the HANNO pipeline. Genes were classified as functionally annotated if they had a non-hypothetical product description or any valid assignment from the Protein families database (Pfam) [64], Clusters of Orthologous Genes (COG) [65], Kyoto Encyclopedia of Genes and Genomes (KEGG) [66], or Gene Ontology (GO) [67]. Each functional category was counted independently across all gene models.

For submission to public repositories, the HANNO bedDB annotation was converted to Generic Feature Format Version 3 (GFF3) to generate a gene-mRNA-exon-CDS hierarchy with locus tags,

protein identifiers, and product names, applying the same duplicate exclusion. Companion FASTA files for protein, CDS, and mRNA sequences were filtered to the retained gene set. Both conversion and filtering were executed via a wrapper pipeline that ensures consistent duplicate removal and verifies identical gene counts across all output files.

To provide the annotation in a format directly comparable to the NCBI genome annotation package for GCA_965278355.2, the GFF3 and companion FASTA files were additionally reformatted with INSDC accessions as sequence identifiers, derived from the NCBI assembly report, and sequential locus tags (prefix PRIEUP) numbered in genomic order. A scaffold-to-accession mapping is provided to link the INSDC identifiers to the scaffold and contig names used in this manuscript.

Identification of candidate sex chromosome-linked regions

Schmid et al. (2002) [59] characterized the species under the former designation *Eleutherodactylus euphronides* and *Eleutherodactylus shrevei* (according to current taxonomy *Pristimantis euphronides* and *Pristimantis shrevei*, respectively). A schematic ideogram of the ZW sex chromosome pair was reconstructed from their C-banding and fluorochrome staining data. As a proxy for *P. euphronides*, for which Schmid et al. [59] performed karyotyping but not flow cytometry, genome size and the size difference between the W and the Z chromosome (W-Z) were calculated from the *P. shrevei* DNA flow cytometry data. The male haploid (1C) genome size (autosomes + Z chromosome) was derived by halving the male diploid (2C) nuclear DNA content and converted to megabases using $1 \text{ pg} \approx 978 \text{ Mb}$ [68]. The W-Z difference was obtained by subtracting the male 2C from the female 2C nuclear DNA content and converting to megabases. The female haploid (1C) genome size was calculated as the sum of the male 1C genome size plus the W-Z difference ((autosomes + Z chromosome) + (W-Z) = autosomes + W chromosome).

Placed and unplaced contig name lists were derived from the RagTag AGP file of the scaffolded primary assembly. Contigs listed as sequence components assigned to the *E. coqui* reference genome (RefSeq ID: GCF_035609145.1) were classified as placed. All remaining contigs in the pre-scaffolding Medaka-polished Flye primary assembly were classified as unplaced. All contigs in both sets retained their original identifiers (contig_*) from the Medaka-polished Flye assembly. Both contig sets were

extracted as separate FASTA files from the pre-scaffolding assembly using SeqKit (grep -f). A filtered subset of the unplaced contigs retaining only those with a sequence length of 500 bp or more was generated for RepeatMasker v4.2.3 [69] analysis. Per-contig sequence length and GC content were computed for both sets with SeqKit (fx2tab -nlg). Summary assembly statistics including contig count, total length, N50, and GC content were computed per set with SeqKit (stats -a). Shared k-mer frequency counts for WindowMasker were built from the full pre-scaffolding primary assembly using windowmasker -mk_counts. This shared count table was used in all subsequent WindowMasker analyses on both scaffolds and contig sets to maintain a consistent genome-wide definition of repetitive sequence.

To identify candidate Z-linked regions in the scaffolded assembly, a multi-criterion analysis was applied to the 13 longest scaffolds of the RagTag-scaffolded primary assembly of *P. euphronides*. Four genomic properties were assessed per scaffold: read depth, inter-haplotype assembly concordance, repeat density, and gene density. All statistical comparisons were performed both genome-wide (Z-candidate bins versus all autosomal bins across the 13 scaffolds) and within-scaffold (Z-candidate bins versus flanking autosomal bins on the same scaffold) to control for chromosome-level baseline differences.

Scaffold-level coverage screening was performed on the mosdepth 10 kb windowed coverage data by aggregating windows into 1 Mb bins per scaffold. For this analysis, the genome-wide reference coverage was defined as the median of per-scaffold mean coverages. For each scaffold, the number of half-coverage bins (mean depth $< 0.625 \times$ the reference) was counted and tested for significant enrichment relative to the genome-wide background rate using a one-sided binomial test with Bonferroni correction for 13 tests ($\alpha = 0.05$). For flagged scaffolds, boundaries of the half-coverage block were determined by optimal binary segmentation. Coverage values were first capped at twice the per-scaffold median to reduce the influence of repeat-driven outlier bins. Two-segment and three-segment models were fitted by exhaustive search over all possible breakpoints, minimizing the within-segment sum of squared deviations, and selected by the Bayesian Information Criterion (BIC). The core block was then extended outward into adjacent bins falling more than two standard deviations below the flanking-segment mean coverage, corresponding to a one-tailed threshold of $p < 0.023$ under a normal approximation. Boundaries are reported as inclusive Mb coordinates.

Inter-haplotype assembly concordance was assessed as an independent line of evidence for hemizygosity. For each scaffold, and separately for Z-candidate and flanking autosomal sub-regions on flagged scaffolds, the corresponding sequences were extracted from the two independently RagTag-scaffolded haplotype assemblies of *P. euphronides* (Additional files 4 and 5 under “Reference-Guided Scaffolding of Haplotype Assemblies”) using SAMtools faidx and aligned with Minimap2 (preset asm5). Mean pairwise identity was calculated from the PAF output as the per-alignment ratio of matching bases to alignment block length, averaged across all alignments per region. Significance was assessed by an exact one-tailed Mann-Whitney U test (H_1 : Z-candidate identity < autosomal identity). Because the exact test enumerates all possible assignments, no distributional assumptions were required.

Repeat content of the 13 scaffolds was characterized using WindowMasker and RepeatMasker, with rmblastn v2.14.1 [61] and the Dfam 3.9 database [70] (FamDB partitions 0 and 12, taxon Anura). Scaffolds were extracted from the scaffolded assembly with SAMtools faidx. WindowMasker was run in -ustat mode using the shared k-mer frequency counts described above. RepeatMasker was run with -species Anura, -xsmall, and --uncurated to include uncurated Dfam families, using six parallel search processes (-pa 6). Masked intervals were aggregated into 1 Mb bins. RepeatMasker annotations were classified into seven broad categories: SINE, LINE, LTR (including Retroposon), DNA transposon, satellite, simple repeat (including low complexity), and other/unclassified. Per-bin masked fractions were compared between Z-candidate and autosomal bins using one-tailed Mann-Whitney U tests with normal approximation and continuity correction (H_1 : Z-candidate > autosomal).

Gene density was assessed from the HANNO BESTMODELS annotation of the RagTag-scaffolded primary assembly, generated with seven Hyloidea protein and mRNA evidence datasets (see "Genome annotation of scaffolded assembly"). Each annotated gene was assigned to the 1 Mb bin containing its coordinate midpoint. Per-bin gene counts were compared between Z-candidate and autosomal bins by one-tailed Mann-Whitney U tests with normal approximation (H_1 : Z-candidate < autosomal). Gene body length (genomic span from start to end of each gene model) was additionally compared between Z-candidate and autosomal genes (H_1 : Z-candidate > autosomal).

Per-gene mean coverage was computed from the same annotation by supplying gene body coordinates as a BED region file to mosdepth with the --fast-mode flag, using the HAC-read BAM aligned to the

scaffolded assembly. The genome-wide reference coverage was taken from the mosdepth summary total_region field. Each gene was assigned to one of four coverage categories based on its mean depth relative to this reference: Hemi_0.5× (0.375-0.625× reference), Auto_1.0× (0.75-1.25× reference), High_Coverage/Repeat (>1.25× reference), and Low_Coverage/Other (all remaining values).

The per-contig mean coverage computed from the alignment to the pre-scaffolding assembly (see "Assembly quality assessment") was used to assign each contig to a coverage class. The bp-weighted mean coverage of placed contigs served as the autosomal reference baseline. Each contig in both sets was classified into the same four coverage categories described above. Contigs with no mapped reads were assigned to a fifth class (Zero). Sequence length and GC content of zero-coverage contigs were assessed with SeqKit, and repeat content with WindowMasker.

WindowMasker was applied to the placed and unplaced contig FASTA files separately in -ustat mode using the shared k-mer frequency counts, with output in interval format (-outfmt interval). Per-contig masked fractions were computed from the interval output for each set. RepeatMasker was run separately on the placed contig FASTA and the length-filtered (≥ 500 bp) unplaced contig FASTA using the same parameters described above for the scaffold-level analysis. Per-contig repeat content was parsed from the RepeatMasker .out files and classified into nine categories (SINE, LINE, LTR, DNA transposon, RC/Helitron, satellite, simple repeat and low complexity, unknown, other). Per-contig masked base pairs and percentage were computed for each category. Results were aggregated at three levels: by contig set (placed, unplaced), for unplaced contigs by coverage class (Hemi_0.5×, Auto_1.0×, High_Coverage/Repeat, Low_Coverage/Other, Zero), and for placed contigs by scaffold zone (Z-candidate, autosomal). The scaffold zone assignment used contig-to-scaffold coordinate mapping through the AGP to identify placed contigs falling within the Z-candidate regions determined by the half-coverage bin enrichment test and boundary segmentation described above. The per-contig coverage uniformity, mapping quality, variant density, and soft-clipping and supplementary alignment rates described under "Assembly quality assessment" were computed from the same HAC-read alignment to the pre-scaffolding assembly and retained for incorporation into the multi-evidence contig table described below.

To obtain gene models at contig-level resolution, the full pre-scaffolding primary assembly was annotated using HANNO with the same seven-species Hyloidea protein (-p) and mRNA (-r) evidence

datasets described under "Genome annotation of scaffolded assembly". Functional annotation was performed via eggNOG-mapper; gene names were taken from the eggNOG Preferred_name field and converted to uppercase. The resulting BESTMODELS annotation was split post-hoc into placed and unplaced gene sets by matching each gene's contig name against the contig name lists derived from the AGP. For placed genes, scaffold coordinates and Z-candidate region flags were assigned through the AGP. As an independent annotation layer, miniprot was run separately on both contig FASTA files using reviewed Anuran protein sequences from UniProt (taxonomy ID 8292; accessed February 2026), with gene names extracted from the UniProt GN= field.

Candidate Z-W gametolog pairs were identified by cross-referencing named genes found on hemizygous (Hemi_0.5×) placed contigs with named genes found on hemizygous unplaced contigs. For each gene name, the following was compiled: the number of hits on placed contigs, the number on unplaced contigs, the number of hemizygous hits on each side, and whether placed hemizygous hits fell within the Z-candidate regions defined by the half-coverage bin enrichment test and boundary segmentation described above. Each named gene was classified into one of seven tiers based on the distribution of its hemizygous hits across placed and unplaced contigs. Tier 1: exactly one hemizygous hit on a placed contig within a Z-candidate region, exactly one hemizygous hit on an unplaced contig, no additional copies on either side. Tier 2: same as Tier 1 but placed hemizygous hit on any scaffold. Tier 3: one hemizygous hit on a placed Z-candidate contig and one on an unplaced contig, but with additional copies on placed or unplaced contigs. Tier 4: same as Tier 3 but placed hemizygous hit on any scaffold. Tier 5: more than one hemizygous copy on either or both sides. Tier 6: hemizygous hit on one side only (placed or unplaced). Tier 7: no hemizygous hits.

For genes classified as Tier 1, the contig pairs carrying them were further sub-classified. Tier 1a comprised contig pairs that shared two or more Tier 1 genes. Tier 1b comprised contig pairs that shared a single Tier 1 gene. Tier 1c comprised contigs that could not be unambiguously paired because they shared Tier 1 genes with more than one partner contig. For Tier 1-5 gametolog candidates, the placed and unplaced contigs carrying each gene were aligned using BLASTn with default parameters. Percent nucleotide identity and query coverage were recorded; pairs producing no significant alignment were recorded as no-hit.

Compositional banding plots of scaffolds and individual contigs were generated from GC content and repeat density in sliding windows (2 kb window, 400 bp step) using the WindowMasker masked intervals from the full primary assembly. Each window was classified as AT-rich repetitive (GC < 50% and masked), GC-rich repetitive (GC ≥ 50% and masked), or non-repetitive (unmasked). Per-contig fractions of AT-rich repetitive, GC-rich repetitive, and non-repetitive windows were computed from the same classification. For scaffold-level figures, contig-level WindowMasker intervals were translated to scaffold coordinates via the RagTag AGP file. Compositional segments along scaffolds were identified by walking the rendered bin array and recording positions where the classified color changed; segments shorter than 1 Mb were merged into their longer neighbor. Annotated genes were assigned to compositional zones by scaffold coordinate midpoint, and per-gene coverage classes were tabulated per zone. For gametolog contig pairs, the placed contig was drawn at its scaffold coordinates and the unplaced contig was drawn adjacent to it along its contig length. Annotated gene positions were overlaid as rectangles spanning full gene body length on both contigs.

All per-contig evidence was joined into a master table with one row per contig, covering all contigs in the primary assembly. Each row contained contig name, set membership, sequence length, and GC content. Coverage was recorded as mean depth and coverage class. For placed contigs, scaffold coordinates and the Z-candidate flag were derived from the AGP. Repeat content included the WindowMasker masked fraction and RepeatMasker total masked percentage with per-class breakdown (SINE, LINE, LTR, DNA transposon, RC/Helitron, satellite, simple repeat and low complexity, unknown, other). The BAM quality metrics from “Assembly quality assessment” comprised coverage CV, mean MAPQ with fractions at MAPQ = 0 and MAPQ ≥ 60, variant density, heterozygous variant density, soft-clip fraction, and supplementary alignment fraction. Gene content was recorded as HANNO total and named gene counts and miniprot gene count.

Each unplaced contig was classified into one of six categories based on its coverage class, gene content, and repeat content. Contigs classified as Auto_1.0× were assigned to the category autosomal-un scaffolded. Contigs classified as Hemi_0.5× that carried at least one annotated gene (HANNO or miniprot) were assigned to W-genic. Within the W-genic category, contigs carrying Tier 1a, 1b, or 1c gametolog genes were sub-ranked by their highest tier. Contigs not classified as Auto_1.0× that carried at least one gene but were not Hemi_0.5× were assigned to W-genic-weak. Contigs carrying no

annotated genes that had a WindowMasker masked fraction ≥ 0.6 or a RepeatMasker total masked percentage $\geq 60\%$ were assigned to W-heterochromatin. Contigs with zero coverage were assigned to zero-coverage. All remaining contigs were assigned to unclassified. Contigs in the categories W-genic, W-genic-weak, and W-heterochromatin were flagged as W-linked candidates.

To assess whether a Dm-W-like [71] sex determination mechanism could be present, the DM domain of the *Xenopus laevis* Dm-W protein (Universal Protein Resource (UniProt) [72] accession Q1L8R5, residues 20–86) was used as query in a tBLASTn v2.14.1 [61] search against the full pre-scaffolding primary assembly, with -seg no, -word_size 2, and e-value thresholds of 0.01 and 10. Any Doublesex and Mab-3 Related Transcription factor (DMRT) family gene models identified in the HANNO BESTMODELS annotation were additionally used as tBLASTn queries. Scaffold positions and coverage classes of all DMRT family hits were recorded.

To identify candidate W-exclusive genes, named genes on W-genic and W-genic-weak contigs were cross-referenced by gene name against all named genes on placed scaffolds. Genes with a name match were classified as gametologs. For the remaining genes without a name match, together with all unnamed W-genic genes, nucleotide sequences were extracted from the pre-scaffolding assembly using the annotated gene coordinates and queried by BLASTn (e-value $1e-5$) against a database consisting of placed scaffold sequences only. Genes producing no significant alignment were classified as W-exclusive candidates.

Unnamed W-exclusive genes were screened for protein domains associated with sex determination or differentiation, including DM, HMG-box, Forkhead, steroid hormone receptor, cytochrome P450, Tudor, Piwi, and Homeobox domains, using the Pfam annotations from the HANNO BESTMODELS output. As a final step, all unnamed W-exclusive gene sequences were queried by BLASTn (e-value $1e-5$, max_target_seqs 3) against a local copy of the NCBI Nucleotide database [62], and hits were screened for sex-related functional annotations.

Phylogenetic analysis using nuclear gene orthologs

To assess the phylogenetic placement of the sequenced *P. euphronides* individual within the anuran superfamily Hyloidea, a set of 19 protein-coding nuclear genes commonly used in amphibian phylogenetics [50] was targeted for multi-gene analysis. The HANNO BESTMODELS annotation (Workflow 1, see “Annotation of scaffolded genome assembly”) was queried by gene name in column 26 to locate gene models for each of the 19 target genes. To accommodate naming differences between the HANNO annotation and the phylogenetic literature, a synonym mapping was applied (e.g., MYC for CMYC, NTF3 for NT3, SLC8A1 for NCX1). For each identified gene, the corresponding coding sequence was extracted from the BESTMODELS CDS file using SeqKit and renamed to a species-specific header format (e.g., >PriEup_RAG1). Genes for which no match was found in the annotation were excluded from subsequent steps.

For each of the extracted *P. euphronides* nuclear genes, homologous sequences from Hyloidea species were retrieved from NCBI GenBank using Entrez queries with gene-specific search terms and synonyms. Gene-specific FASTA files containing CDS or partial CDS records were downloaded and compiled per gene. A single orthologous sequence per gene from the *Rana temporaria* reference genome (GCF_905171775.1) was extracted as non-Hyloidea outgroup, selecting the longest CDS isoform per gene using SeqKit.

For each gene, the *P. euphronides*, Hyloidea, and *R. temporaria* sequences were concatenated into a single multi-species FASTA file. Preliminary multiple sequence alignments were performed using MAFFT v7.475 [73] with default settings, followed by removal of gap-containing terminal alignment regions. Genes in which the *P. euphronides* sequence was reduced entirely to alignment gaps after terminal trimming, indicating insufficient positional overlap between the *P. euphronides* CDS and the majority of Hyloidea sequences, were excluded at this stage. Remaining trimmed alignments were re-aligned with MAFFT using the high-accuracy setting (--globalpair --maxiterate 1000) and manually curated in Jalview v2.11.4.1 [74] to assess alignment quality. Genes were excluded if their alignments showed excessive internal fragmentation of the *P. euphronides* sequence, poor positional overlap between sequences across taxa, or representation from only a single Hyloidea genus, as these would not permit meaningful multi-family phylogenetic inference. Sequences containing ambiguous bases or

internal gaps were removed from retained gene alignments, and retained sequences were re-aligned with MAFFT to generate final gene alignments. The taxonomic identity of each retained GenBank sequence was recorded at species, genus, and family level using TaxonKit v0.19.0 [75] with the NCBI Taxonomy database (taxdump, downloaded 17 April 2025). Two concatenated supermatrices were constructed: one including all retained genes but restricted to species with sequences available for every gene, and one based on a reduced gene subset permitting broader taxonomic sampling. Both supermatrices were manually curated in Jalview to remove gap-only columns, and partition files were updated accordingly.

Phylogenetic inference was performed using IQ-TREE v2.2.0.3 [76]. For concatenated supermatrices, gene-partitioned models were selected automatically by ModelFinder using the -m MFP+MERGE setting. For individual gene alignments, best-fit substitution models were selected with -m MFP. All analyses used 1,000 ultrafast bootstrap replicates, 1,000 SH-aLRT tests, and the --bnni option.

Phylogenetic trees were visualized and annotated in iTOL v6 [77]. Branches were colored by taxonomic family, and for genes with large numbers of taxa, clades with average branch length distance below 0.05 were collapsed for legibility. Trees were exported as PNG images for figures and as Newick format for archiving.

Divergence times and biogeographic context

The divergence times and topological distances extracted from the time-calibrated phylogeny in [50], as described under "Reference genome selection via phylogenetic proximity," were additionally used to assess the phylogenetic relationships between *P. euphronides* and co-occurring and geographically proximate anuran species. Species occurrence data were obtained from an AmphibiaWeb [78] export containing ISO 3166-1 alpha-2 country codes for each species and matched to all tip labels in the phylogeny. Anuran species co-occurring on Grenada were identified by filtering for ISO country code "GD." Country centroid coordinates [79] were used to compute minimum great-circle distances to Grenada using the haversine formula [80]. Each species was classified as Caribbean island or American mainland based on its occurrence codes.

For the closest mainland *Pristimantis* congeners of *P. euphronides* identified by topological proximity at the next phylogenetic node beyond *P. shrevei*, per-species type-locality coordinates were compiled to provide a finer geographic resolution than country centroids allow. Type-locality strings were retrieved from Amphibian Species of the World v6.2 [81], the curated taxonomic database that reproduces the original published type-locality text from each species description. Where the type-locality string itself contained explicit coordinates from the original taxonomic description, expressed either in degrees-minutes-seconds or decimal-degree form, those coordinates were extracted directly. Where no explicit coordinates were given, the most specific named place in the type-locality string was located on Google Maps and the displayed decimal-degree coordinates recorded. The minimum great-circle distance from each type locality to Pointe Salines (12.00°N, 61.79°W), the southwesternmost point of Grenada and the closest single point of the island to the South American mainland, was then computed using the haversine formula.

References

1. IUCN SSC Amphibian Specialist Group. 2021. *Pristimantis euphronides*. The IUCN Red List of Threatened Species 2021: e.T56593A3043096.
<https://www.iucnredlist.org/en/species/56593/3043096#population>. Accessed 2 Apr 2026.
2. O'Cathail C, Ahamed A, Burgin J, Cummins C, Devaraj R, Gueye K, et al. The European Nucleotide Archive in 2024. *Nucleic Acids Res.* 2025;53(D1):D49–55.
3. Oxford Nanopore Technologies. Genomic DNA by Ligation (SQK-LSK114).
<https://nanoporetech.com/document/genomic-dna-by-ligation-sqk-lsk114>. Accessed 1 Oct 2024.
4. Oxford Nanopore Technologies (ONT). MinKNOW. Version 24.06.16.
<https://community.nanoporetech.com/downloads>. Accessed 1 Sep 2024.
5. Oxford Nanopore Technologies (ONT). Dorado basecall server. Version 7.4.12.
<https://github.com/nanoporetech/dorado>. Accessed 3 Apr 2025.
6. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 2010;38(6):1767-1771.
7. Oxford Nanopore Technologies. Quality scores and filtering. Technical support pages.
<https://nanoporetech.com/document/experiment-companion-minknow>. Accessed 5 Oct 2024.

8. Python Software Foundation. Python: Version 3.11.8 [software]. Wilmington (DE): Python Software Foundation; 2024. <https://www.python.org>. Accessed 3 Apr 2026.
9. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009 Jun 1;25(11):1422-3.
10. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020 Sep;585(7825):357-362.
11. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020 Mar;17(3):261-272.
12. McKinney W. Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*. 2010:56-61.
13. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90-95.
14. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2024. <https://www.R-project.org/>. Accessed 3 Apr 2026.
15. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag; 2016. <https://ggplot2.tidyverse.org>. Accessed 3 Apr 2026.
16. Ramey C, Fox B. *Bash Reference Manual*. Edition 5.2, for Bash version 5.2. Free Software Foundation; 2022. <https://www.gnu.org/software/bash/manual/>. Accessed 3 Apr 2026.
17. Tange, O. 2023, Nov 22. GNU Parallel 20231122 ('Grindavík'). Zenodo. <https://doi.org/10.5281/zenodo.10199085>. Accessed 3 Apr 2026.
18. Shen W, Sipos B, Zhao L. SeqKit2: A Swiss army knife for sequence and alignment processing. *Imeta*. 2024;3(3):e191.
19. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*. 1998 Mar;8(3):175-85.
20. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 1998 Mar;8(3):186-94.
21. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*. 2018;34(15):2666-2669.
22. Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020 Mar 18;11(1):1432.
23. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 2018;34(5):867-868.
24. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764-770.

25. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 2019;20:257.
26. Langmead B, Salzberg SL. Kraken 2 indexes. PlusPF database, release 2024-12-28. <https://benlangmead.github.io/aws-indexes/k2>. Accessed 30 Dec 2024.
27. NCBI Genome. National Center for Biotechnology Information. <https://www.ncbi.nlm.nih.gov/datasets/genome>. Accessed 15 Oct 2025.
28. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol.* 2019 May;37(5):540-546.
29. Oxford Nanopore Technologies (ONT). medaka: sequence correction provided by ONT Research. GitHub. Version 2.0.1. <https://github.com/nanoporetech/medaka/releases/tag/v2.0.1>. Accessed 1 Mar 2025.
30. Kuhl H, Li L, Wuertz S, Stöck M, Liang XF, Klopp C. CSA: A high-throughput chromosome-scale assembly pipeline for vertebrate genomes. *Gigascience.* 2020;9(5):giaa034.
31. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;17(2):155-158.
32. Kolmogorov M, Raney B, Paten B, Pham S. Ragout - a reference-assisted assembly tool for bacterial genomes. *Bioinformatics.* 2014;30(12):i302-i309.
33. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, et al. Chromosome assembly of large and complex genomes using multiple references. *Genome Res.* 2018;28(11):1720-1732.
34. Goldfarb T, Kodali VK, Pujar S, Brover V, Robbertse B, Farrell CM, et al. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res.* 2025;53(D1):D243-D257.
35. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* 2021;22:312.
36. Rhie A, Walenz BP, Koren S, Phillippy AM. Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* 2020;21:245.
37. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics.* 2008 Dec 15;24(24):2818-24.
38. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018 Sep 15;34(18):3094-3100.
39. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021;10(2):giab008.
40. Kuhl H. 2024. HANNO: efficient High-throughput ANNOtation of protein coding genes in eukaryote genomes. doi:10.5281/zenodo.11532370. Accessed 10 Apr 2025.
41. Li H. Protein-to-genome alignment with minimap2. *Bioinformatics.* 2023;39(1):btad014.

42. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Nat Biotechnol.* 2019;37(8):907-910.
43. Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Bioinformatics.* 2017;33(4):606-608.
44. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 2013;8(8):1494–512.
45. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Res.* 2011 Mar;21(3):487-93.
46. Manni M, Berkeley MR, Seppey M, Simão 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.
47. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* 2019;47(D1):D807–11.
48. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ.* 2018 Jun 4;6:e4958.
49. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 2016;17(1):132.
50. Portik DM, Streicher JW, Wiens JJ. Frog phylogeny: A time-calibrated, species-level tree based on hundreds of loci and 5,242 species. *Mol Phylogenet Evol.* 2023 Nov;188:107907.
51. Portik DM, Streicher JW. Frog phylogeny: supplementary data for Portik et al. (2023). GitHub repository. <https://github.com/nhm-herpetology/frog-phylogeny>. Accessed 1 Jan 2026.
52. Steel MA, Penny, D. Distributions of Tree Comparison Metrics- Some New Results. *Sys Biol.* 1993;42: 126-141.
53. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* 2022;23(1):258.
54. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics.* 2006;22(2):134-141.
55. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A.* 1988 Apr;85(8):2444-8.
56. National Center for Biotechnology Information (NCBI). AGP Specification v2.1. Bethesda (MD): NCBI; 2024. https://www.ncbi.nlm.nih.gov/genbank/genome_agp_specification. Accessed 3 Apr 2026.

57. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun;12(6):996-1006.
58. Li H. Seqtk: a fast and lightweight tool for processing sequences in the FASTA/Q format. Version 1.4-r122. GitHub repository. <https://github.com/lh3/seqtk>. Accessed 10 Apr 2025.
59. Schmid M, Feichtinger W, Steinlein C, Rupprecht A, Haff T, Kaiser H. Chromosomal banding in Amphibia. XXIII. Giant W sex chromosomes and extremely small genomes in *Eleutherodactylus euphronides* and *Eleutherodactylus shrevei* (Anura, Leptodactylidae). *Cytogenet Genome Res.* 2002;97(1-2):81-94.
60. Heinicke MP, Duellman WE, Hedges SB. Major Caribbean and Central American frog faunas originated by ancient oceanic dispersal. *Proc Natl Acad Sci U S A.* 2007 Jun 12;104(24):10092-7.
61. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10:421.
62. NCBI Nucleotide. National Center for Biotechnology Information. <https://www.ncbi.nlm.nih.gov/nucleotide/>. Accessed 22 Mar 2025.
63. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol.* 2021 Dec 9;38(12):5825-5829.
64. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. *Nucleic Acids Res.* 2021;49(D1):D412–9.
65. Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D274-D281.
66. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017 Jan 4;45(D1):D353-D361.
67. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D325-D334.
68. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Nuclear DNA content and genome size of trout and human. *Cytometry A.* 2003 Feb;51(2):127-8.
69. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015. <http://www.repeatmasker.org>. Accessed 10 Dec 2025.
70. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA.* 2021 Jan 12;12(1):2.
71. Yoshimoto S, Okada E, Umemoto H, Tamura K, Uno Y, Nishida-Umehara C, et al. A W-linked DM-domain gene, DM-W, participates in primary ovary development in *Xenopus laevis*. *Proc Natl Acad Sci U S A.* 2008;105(7):2469–74.

72. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* 2025 Jan 6;53(D1):D609-D617.
73. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013 Apr;30(4):772-80.
74. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics.* 2009 May 1;25(9):1189-91.
75. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet Genomics.* 2021 Sep 20;48(9):844-850.
76. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol.* 2020 May 1;37(5):1530-1534.
77. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* 2024 Jul 5;52(W1):W78-W82.
78. AmphibiaWeb. 2026. https://amphibiaweb.org/amphib_names.txt. University of California, Berkeley, CA, USA. Accessed 16 Jan 2026.
79. Google. 2026. <https://raw.githubusercontent.com/google/dspl/master/samples/google/canonical/countries.csv>. Accessed 16 Jan 2026.
80. Veness C. Calculate distance, bearing and more between Latitude/Longitude points. Movable Type Scripts. <https://www.movable-type.co.uk/scripts/latlong.html>. Accessed 02 Apr 2026.
81. Frost DR. Amphibian Species of the World: an Online Reference. Version 6.2. American Museum of Natural History, New York. <https://amphibiansoftheworld.amnh.org>. Accessed 26 Apr 2026.