

Additional file 4

Supplementary materials and methods - haplotypes

Data and scripts are available on Zenodo (DOI: 10.5281/zenodo.15298547). Scripts are also available on GitHub (<https://github.com/koppk/Grenada-Frog-144>).

Haplotype phasing and dual assembly

To generate a haplotype-resolved diploid assembly, HapDup v0.12 [1-3] was applied to the Medaka-polished Flye assembly of Workflow 1 (Additional file 2 under "Genome assembly"). The high-accuracy (HAC) basecalled reads were mapped to the assembly using Minimap2 v2.28 (-ax map-ont) [4], sorted and indexed with SAMtools v1.19.2 [5], producing the input Binary Alignment Map (BAM) required by HapDup. HapDup filtered the alignment to remove reads likely originating from unassembled regions of the genome.

Variant calling was performed with PEPPER v0.8 [2] to call single nucleotide polymorphisms (SNPs) and insertions/deletions (indels) across the reference assembly. Phasing was performed using Margin v2.3.1 [2], which statistically phased variants based on haplotype co-occurrence across long Nanopore reads. The phased variants were output in both Variant Call Format (VCF) and Browser Extensible Data (BED) formats, with phasing information encoded using phase set tags. Reads were haplotagged using the phased variants and the resulting haplotagged BAM was used for haplotype-aware polishing.

Two haplotype-specific assemblies were independently polished using Flye's haplotype polishing mode, producing the final dual assemblies for Haplotype 1 and Haplotype 2. These represent two full-length, phase-informed assemblies constructed through haplotype-aware polishing. They include all genomic regions regardless of phasing confidence. Additionally, phased-only assemblies were generated using only regions confidently phased by Margin and are expected to be more fragmented. HapDup does not assign biological parental origin to the haplotypes. For downstream analyses and

scaffolding, only the structurally resolved dual assemblies were used due to their higher contiguity and genome coverage.

Haplotype alignment and structural comparison

To assess structural similarity and continuity between the two haplotype assemblies generated by HapDup from the Workflow 1 assembly, a pairwise alignment using Minimap2 with default settings for long-read assemblies was performed and the output was visualized using D-Genies v1.5.0 [6]. Dotplot rendering was configured to reflect nucleotide-level identity using a green-to-yellow color gradient to indicate similarity strength. No filters were applied to alignment length or percent identity. The resulting plot enables visual inspection of global collinearity and potential structural divergence between haplotypes prior to reference-guided scaffolding.

Reference genome similarity assessment

Genomic distances between each candidate reference and the two haplotype assemblies were estimated using Mash v2.3 [7] as described in Additional file 2 under "Reference genome selection via k-mer-based distance estimation".

Reference-guided scaffolding of haplotype assemblies

To scaffold the two haplotype assemblies of the “GrenadaFrog144” sequencing data to near-chromosome scale, RagTag v2.1.0 [8] was used in reference-guided mode. The soft-masked chromosome-level genome of *Eleutherodactylus coqui* haplotype 2 (*GCA_035609135.1*) was used as the reference. Repeat masking was performed using WindowMasker v1.0.0 [9] with default parameters prior to scaffolding. Both haplotype assemblies were scaffolded independently using RagTag with default alignment parameters (-x asm5) via the embedded Minimap2 aligner. The reference genome was soft-masked to prevent anchoring of contigs to repetitive regions during alignment. RagTag output included the respective scaffolded genome, AGP file, gap coordinates, and placement summary.

Structural summary statistics (e.g., scaffold counts, N50, gap content) were extracted using SeqKit v2.10.0 [10] with the -a flag, 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.

Scaffolded haplotype alignment and structural comparison

To assess large-scale structural consistency between the scaffolded haplotype assemblies, they were aligned using D-Genies in default whole-genome alignment mode. Resulting alignments were visualized as a dotplot to inspect global synteny and potential structural variations between haplotypes.

References

1. 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.
2. Shafin K, Pesout T, Chang PC, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat Methods.* 2021 Nov;18(11):1322-1332.
3. Kolmogorov M, Billingsley KJ, Mastoras M, Meredith M, Monlong J, Lorig-Roach R, et al. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nat Methods.* 2023 Oct;20(10):1483-1492.
4. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018 Sep 15;34(18):3094-3100.
5. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021 Feb 16;10(2):giab008.
6. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ.* 2018 Jun 4;6:e4958.
7. 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.

8. 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.
9. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics.* 2006 Jan 15;22(2):134-41.
10. Shen W, Sipos B, Zhao L. SeqKit2: A Swiss army knife for sequence and alignment processing. *Imeta.* 2024;3(3):e191.