

Additional file 5

Supplementary results - haplotypes

Methods for haplotype phasing and scaffolding are described in Additional file 4. 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

The final Medaka-polished Flye assembly was subjected to haplotype phasing and diploid-aware reassembly using HapDup v0.12 [1-3]. A total of 2,595,267 variants were identified, of which 1,781,824 (68.7%) were successfully phased into 87,691 phase blocks, with a block N50 of 12,990 bp and an average length of 5,424 bp. Of the 17,071,244 high-accuracy (HAC) basecalled reads (mean length ~3.26 kb), 787,351 were haplotagged based on high-confidence alignments overlapping phased variants: 335,766 reads were assigned to haplotype 1 (H1), 318,685 to haplotype 2 (H2), and 132,900 remained unphased (H0).

Haplotype-aware polishing with Flye produced two dual haplotype assemblies, with total lengths of 1,660,158,632 bp and 1,659,294,964 bp, respectively (Table HR1). Each assembly comprised ~17,700 contigs, with N50 values of 326,244 bp (haplotype 1) and 326,851 bp (haplotype 2). HapDup's structural variant detection step, which identifies heterozygous breakpoints in long-read alignments and applies corresponding changes to the polished haplotypes, identified four unbalanced breakpoints, only one with majority read support, but no large-scale inversions.

By contrast, the phase-only assemblies were substantially more fragmented, containing 95,520 and 95,493 contigs, respectively, with an N50 of ~27.6 kb and total lengths of ~1.66 Gb, comparable to the full haplotypes. The dual haplotype assemblies were therefore used as input for reference-guided scaffolding.

Table HR1. Assembly statistics for dual haplotype-resolved *Pristimantis euphronides* assemblies before and after scaffolding. Dual haplotype assemblies were generated using HapDup v0.12 [1-3] from the Medaka-polished Flye assembly (Workflow 1). Scaffolding was performed using RagTag v2.1.0 [7] against the soft-masked *Eleutherodactylus coqui* haplotype 2 genome (GCA_035609135.1). Assembly statistics were computed using SeqKit v2.10.0 [8] and RagTag internal metrics. Bp: base pairs.

Metric	Haplotype 1 (Contigs)	Haplotype 2 (Contigs)	Haplotype 1 (Scaffolds)	Haplotype 2 (Scaffolds)
File	hapdup_dual_1.fas ta	hapdup_dual_2.fas ta	hapdup_dual_1.ragt ag.scaffold.fasta	hapdup_dual_2.ragt ag.scaffold.fasta
Number of sequences	17,787	17,686	9,852	9,753
Total Length (bp)	1,660,158,632	1,659,294,964	1,765,936,622	1,766,937,936
Average Length (bp)	93,336	93,820	179,247	181,169
Max Contig/Scaffold (bp)	3,441,356	3,444,968	257,746,147	258,527,921
N50 (bp)	326,244	326,851	125,610,897	125,516,806
Unplaced Sequences	—	—	9,824	9,726
Unplaced Bases (bp)	—	—	191,727,293	190,276,223
Gap Sequences	—	—	7,935	7,933
Gap Bases (bp)	—	—	105,777,990	107,642,972
Gap Percentage (%)	—	—	5.99%	6.09%
GC Content (%)	42.90%	42.91%	40.33%	40.29%

Haplotype alignment and structural comparison

To assess structural consistency and divergence between the two haplotype-resolved assemblies a whole-genome alignment was conducted using D-Genies v1.5.0 [5] with embedded Minimap2 v2.28 [4], with haplotype 2 as query (vertical axis) and haplotype 1 as target (horizontal axis). While visual inspection of the dotplot revealed a consistent diagonal alignment, limited off-diagonal signals were detected (Figure HR1). No large-scale inversions or rearrangements were identified. Total assembly lengths of the two dual haplotype assemblies were comparable (1.660 Gb and 1.659 Gb), and alignments spanned nearly the entire genome for both assemblies.

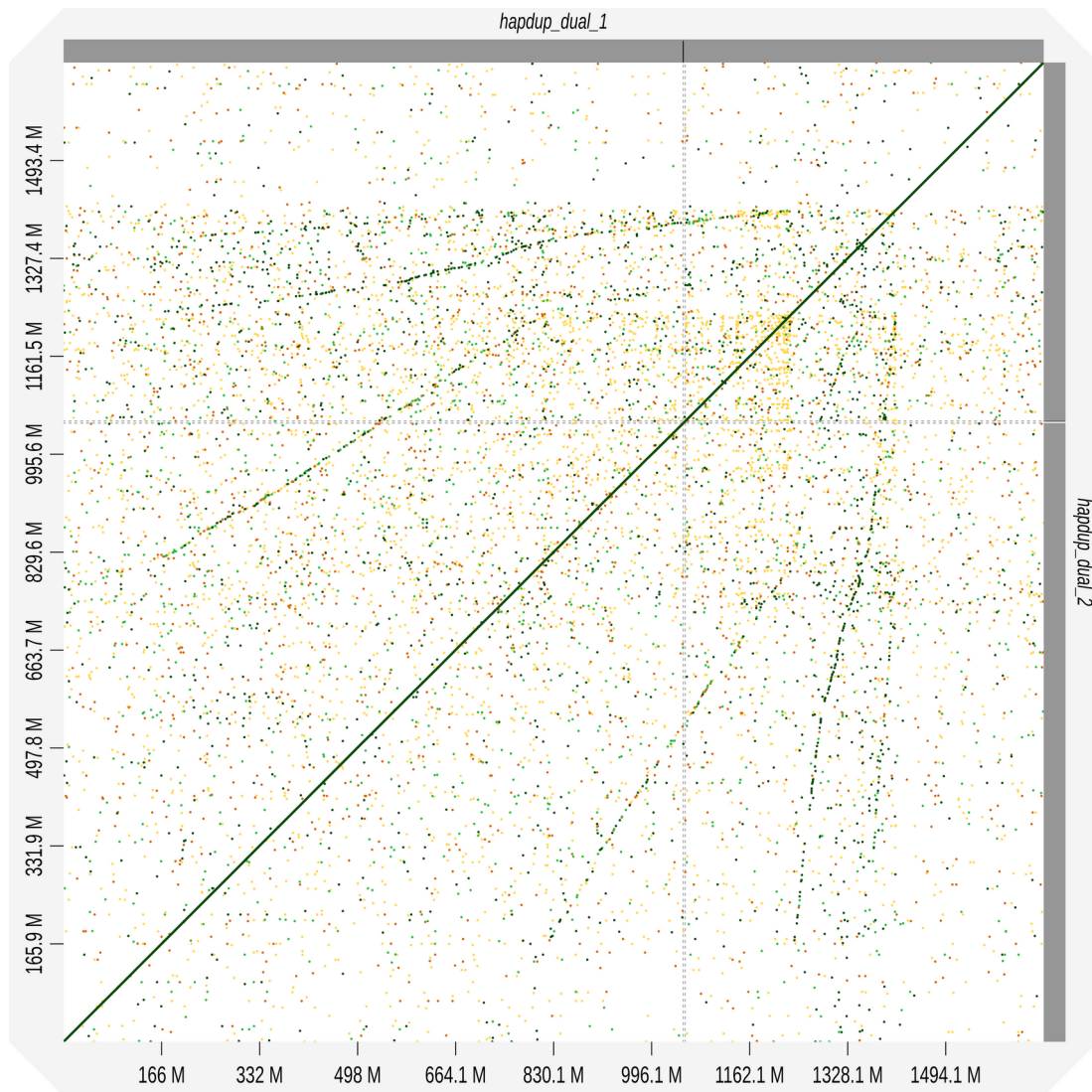


Figure HR1. Pre-scaffolding alignment of the two haplotype-resolved *Pristimantis euphronides* assemblies. D-Genies v1.5.0 [5] dotplot of the Medaka-polished Flye (Workflow 1) assembly resolved into two haplotypes using HapDup v0.12 [1-3]. Horizontal axis: haplotype 1. Vertical axis: haplotype 2. Color reflects nucleotide identity in percentage (%) using the default D-Genies scale: white (0%), yellow (1-24%), orange (25-49%), light green (50-74%), and dark green (75-100%).

As summarized in Table HR2, D-Genies aligned 1,639,161,794 bp of haplotype 2 to haplotype 1 (98.8% of total haplotype 2 length) across 33,502 alignment blocks, with a mean block length of 38,658 bp. The weighted mean per-base divergence between haplotypes was 0.28%.

Table HR2. Structural alignment comparison of dual haplotype-resolved *Pristimantis euphronides* assemblies before and after scaffolding. Dual haplotype assemblies were generated using HapDup v0.12 [1-3] from the Medaka-polished Flye assembly (Workflow 1). Scaffolding was performed independently using RagTag v2.1.0 [7] against the soft-masked *Eleutherodactylus coqui* haplotype 2 genome (GCA_035609135.1). Pairwise whole-genome alignments were computed using D-Genies v1.5.0 [5] with embedded Minimap2 v2.28 [4], with haplotype 2 as query and haplotype 1 as target. Alignment span is reported for the query (haplotype 2) as D-Genies evaluates each query contig against the full target. bp, base pairs.

Metric	Pre-scaffolding (HapDup)	Post-scaffolding (RagTag)
Alignment input assemblies	hapdup_dual_1.fasta vs hapdup_dual_2.fasta	hapdup_dual_1.ragtag.scaffold.fasta vs hapdup_dual_2.ragtag.scaffold.fasta
Haplotype 2 bases aligned (bp)	1,639,161,794 of 1,659,294,964 (98.8%)	1,753,171,120 of 1,766,937,936 (99.2%)
Number of alignment blocks	33,502	25,949
Total matching base pairs (bp)	1,295,121,497	1,290,591,950
Mean alignment block length (bp)	38,658	49,736
Weighted mean per-base divergence (%)	0.28	0.27

Reference genome similarity assessment

Pairwise genomic distances computed with Mash v2.3 [6] revealed that among the 21 reference assemblies analyzed, *Eleutherodactylus coqui* haplotype 2 (GCA_035609135.1) exhibited the lowest sequence divergence to both haplotype-resolved assemblies (Additional file 3: Figure SR3 and detailed in Additional file 6: Table). The estimated Mash distances were 0.1542 for haplotype 1 and 0.1519 for

haplotype 2 assembly. In contrast, the NCBI RefSeq assembly, *E. coqui* haplotype 1 (GCF_035609145.1), showed slightly higher Mash distances of 0.1591 and 0.1566, respectively. Based on these results, GCA_035609135.1 was identified as the most similar reference genome available and was selected for downstream scaffolding of both haplotype-resolved assemblies.

Reference-guided scaffolding of haplotype assemblies

The dual haplotype-resolved assemblies were scaffolded using RagTag v2.1.0 [7] with the soft-masked *Eleutherodactylus coqui* haplotype 2 genome (GCA_035609135.1) as reference. Assembly statistics were collected before and after scaffolding (Table HR1). RagTag reduced the number of sequences from 17,787 to 9,852 in haplotype 1 and from 17,686 to 9,753 in haplotype 2. Total assembly lengths increased from ~1.66 Gb to ~1.77 Gb in both cases, primarily due to gap (N) insertions introduced during scaffolding. The N50 increased from 326,244 bp to 125.6 Mb in haplotype 1 and from 326,851 bp to 125.5 Mb in haplotype 2. The maximum scaffold length exceeded 257 Mb in both assemblies. Scaffolding anchored 7,963 sequences (1.47 Gb) in haplotype 1 and 7,960 sequences (1.47 Gb) in haplotype 2, while ~190 Mb of sequence remained unplaced in each assembly. Gap sequences accounted for 105.8 Mb (5.99%) and 107.6 Mb (6.09%) of the total scaffolded genome length in haplotypes 1 and 2, respectively. GC content shifted from ~42.9% to ~40.3% due to the inclusion of gap regions.

Scaffolded haplotype alignment and structural comparison

After independent RagTag scaffolding of the two haplotype assemblies using the *Eleutherodactylus coqui* haplotype 2 genome (GCA_035609135.1) as reference, the resulting assemblies were compared using pairwise whole-genome alignment. Alignments were computed using D-Genies with embedded Minimap2. The D-Genies dotplot displayed a compact and continuous diagonal alignment with a considerable reduction in off-diagonal signals (Figure HR2). Visual inspection did not indicate large-scale inversions or rearrangements.

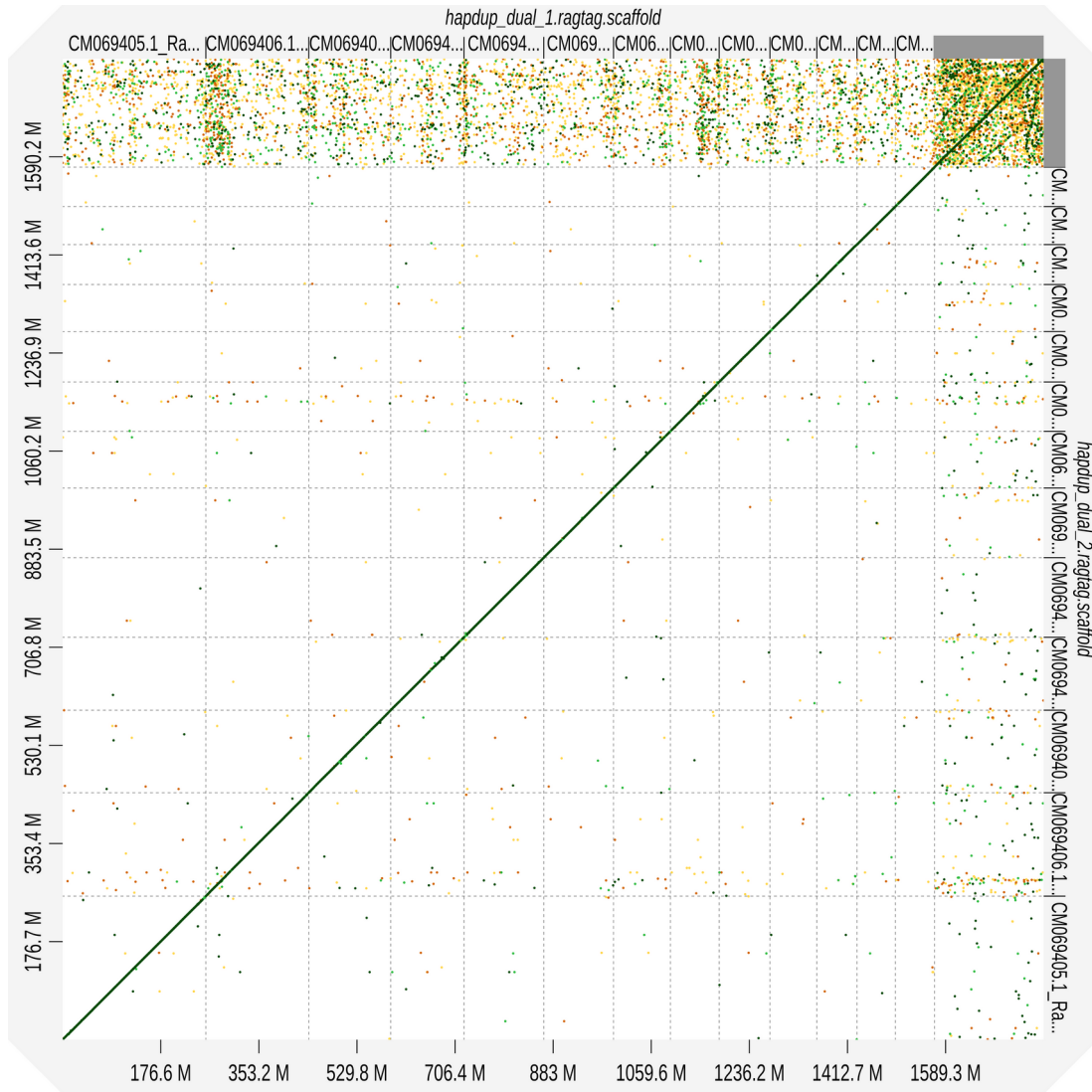


Figure HR2. Post-scaffolding alignment of the two haplotype-resolved *Pristimantis euphronides* assemblies. D-Genies v1.5.0 [5] dotplot of the Medaka-polished Flye (Workflow 1) assembly resolved into two haplotypes using HapDup v0.12 [1-3] scaffolded each against the *Eleutherodactylus coqui* haplotype 2 reference genome (GCA_035609135.1) using RagTag v2.1.0 [7]. Horizontal axis: scaffolded haplotype 1. Vertical axis: scaffolded haplotype 2. Color reflects nucleotide identity in percentage (%) using the default D-Genies scale: white (0%), yellow (1-24%), orange (25-49%), light green (50-74%), and dark green (75-100%).

As listed in Table HR2, D-Genies aligned 1,753,171,120 bp of haplotype 2 to haplotype 1 post-scaffolding (99.2% of total length), compared to 98.8% pre-scaffolding. The number of alignment blocks decreased from 33,502 to 25,949 and mean block length increased from 38,658 bp to 49,736 bp (+29%), reflecting improved contiguity from scaffolding. The weighted mean per-base divergence remained stable (0.28% pre-scaffolding, 0.27% post-scaffolding).

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. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ.* 2018 Jun 4;6:e4958.
6. 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.
7. 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.
8. Shen W, Sipos B, Zhao L. SeqKit2: A Swiss army knife for sequence and alignment processing. *Imeta.* 2024;3(3):e191.