Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A minimal pangenome graph for *Salminus brasiliensis*,a freshwater migratory Neotropical flagship species: it takes two to tango

Gabriel M. Yazbeck

dna@ufsj.edu.br

LARGE-DEZOO-CTAN Universidade Federal de São João Del Rei https://orcid.org/0000-0003-1729-348X

Data Note

Keywords: cryptic species, dorado, dourado, migratory species, structural variants, genomics

Posted Date: October 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11381812/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Two chromosome-level, reference-guided consensus genomes of *Salminus brasiliensis* are presented for an Upper Paraná Basin specimen (Brazil), representing a potentially distinct lineage from the Vertebrate Genomes Project reference specimen, from Argentina. These resources allow for a genomic variation framework, including the minimal reference pangenome graph. Short reads mapped against the VGP reference assemblies yielded two ~ 1.07–1.08 Gb pseudohaplotypes (N_{50}=38.9–40.8 Mb), with union k-mer completeness of 94%, Q_V ~33 each. BUSCO completeness scores are 97.8% for the principal genome and 97.4% for the lifted annotation proteome. The reference pangenome graph for *S. brasiliensis* was obtained from four chromosome-level haploid sets, using minigraph. It is 1.1 Gb, comprising 288,788 segments (average length 3,826 bp), 404,621 links (809,242 arcs), with average node degree 1.4 (maximum 3). Additional read-to-graph alignments (GAM/GAF) and other accompanying genomic datasets, such as gene annotation, are also provided. The bioinformatics datasets made available for this model Neotropical migratory freshwater species—relevant for conservation, fisheries and aquaculture—can be applied in fields such as ecology, evolution and biotechnology.

Background & Summary

A pangenome provides a unified data presentation framework for studying a collection of genomes. It comprehensively encapsulates the genetic variation within and among individuals[1], and has recently leveraged efficient methods for approaching the fundamental level of molecular biodiversity. Pangenomes offer a coherent, comprehensive pathway to untangle population genetic structure[2], the pattern in which polymorphisms are distributed throughout a given species' range, at an arbitrary scale[3]. Among the alternative representation forms available for recording pangenomes (*e.g.* linear, mutation-annotation trees), a notable current trend is to use graphs (systems modeled by nodes connected by edges) to depict them[4]. A particularly straightforward way of initially presenting a pangenome graph is the reference pangenome graph (rGFA) format, introduced by Heng Li and collaborators[5]. It records the topological relationships of a set of haploid genomes, by using the linear reference genome's coordinate system, and subsequently building upon it. According to such a scheme, nodes carry sequences, while edges represent connections between them. This makes such simple graphs convenient for genome-wide comparison, particularly for annotating large structural variants (SV). More elaborate (and much more demanding) approaches have since expanded upon it, allowing for base-level resolution, such as the minigraph-cactus pipeline[6]. Among many other inherent advantages, pangenome graphs can reduce reference bias (for resequencing and mapping of low-cost short reads) and support haplotype phasing validation and advanced variant calling procedures[7].

The Neotropical freshwater "piracema" (*i.e.* migratory) fish, locally known as "dourado" among other common names, *Salminus brasiliensis* (Cuvier, 1816) is a notable member of the family Bryconidae[8], highly prized in angling[9]. French naturalist Georges Cuvier (1769–1832) coined the genus name

Salminus in reference to its salmonid counterparts, despite the phylogenetic distance (orders Characiformes and Salmoniformes, respectively).

As a voracious predator, *S. brasiliensis* is capable of influencing biotic community structure[10], and has often been singled out as a natural flagship species[11–13]. It is well recognized by the public, culturally important, and arguably displays a unique beauty, with an exquisite, vivid and iridescent-golden colouration. This fish is originally distributed across the broad, megadiverse La Plata River Basin (LPRB), spanning five South American countries, encompassing an area of ~ 3 x 10^6 km^2, serving more than 110 million people[14]. It has now been introduced in other basins[15] and even countries, posing a substantial risk of consequential biological invasion.

Notably, *S. brasiliensis* is one of the largest freshwater fishes in the Neotropics, surpassed in size only by *Arapaima gigas* and a few Amazonian catfishes. Due to its sheer size, exceptionally reaching 1 m in length and 30 kg in weight, it is a potential source of high-quality protein, capable of contributing to food security. It holds considerable promise for aquaculture[16] and, thus, for promoting sustainable economic growth and alleviating fishing pressure on wild populations.

Most importantly, molecular and meristic studies have cast serious doubt on its current taxonomic status[17,18]. The presence of cryptic species within *S. brasiliensis*, from different subbasins of the LPRB, has therefore been proposed: a distinct genetic lineage was detected in the Upper Paraná River Subbasin[18,19]. It differs from the lineages found in the Uruguay Subbasin[19] (Pampas biome, Argentina) and the upper reaches of the Paraguay Subbasin[20]. The significance of this vicariant cryptic diversity and the question of precise species delineation remains largely unresolved. It potentially challenges the present International Union for Conservation of Nature (IUCN) classification for *S. brasiliensis* as “least concern”[21].

The Convention on the Conservation of Migratory Species of Wild Animals is an international agreement established in 1979, also known as CMS or the Bonn Convention, currently with more than 130 signatory parties. Its Conference of Parties 15 (COP15) was held in late March 2026 in Brazil. The COP15 also highlighted the challenges of conserving transboundary migratory species and, alongside the launch of a Global Assessment of Migratory Freshwater Fishes, specifically identified large migratory characids, including *Salminus*, as a high-priority group undergoing pronounced declines across the LPRB.

Graciano and colleagues[12] made available the first *de novo* genomic assembly[22] (*i.e.* not reference-guided), for *S. brasiliensis*, along with molecular data such as partial gene annotation[23] and a genome-wide microsatellite panel[24]. It is a scaffold-level assembly totalling 0.92 Gb (gigabases) in a heavily discontiguous genome, with N_{50} of ~ 11 kbp. It was built from a high-quality next-generation sequencing dataset, comprising ~ 16 Gb of 90 bp paired-end reads[25]. Together, these resources provide comprehensive mapping information—positional and depth—in sequence alignment/map (SAM) format. These short reads were obtained from a sample[26] (“isolate”) from the uppermost northeastern region of

the LPRB, at the Upper Grande River, MG, Brazil, in the Upper Paraná Subbasin. This short-read dataset was used directly to generate the genomic data presented here.

Then, in 2023, the first chromosome-level genome assembly was made publicly available under the Vertebrate Genomes Project[27] (VGP). It currently stands as the linear reference genome[28] for *S. brasiliensis*, in the Genome database of the National Center for Biotechnology Information (NCBI). It was assembled *de novo*, relying on multiple combined methods, using DNA sequencing data[29] which included long reads (10–25 kb PacBio HiFi), and chromosome conformation capture assays (Arima Hi-C and Bionano optical maps), as well as full-length transcripts (PacBio ISO-seq) and short-read (Illumina NovaSeq) transcriptome sequencing data[30]. This genome assembly likewise serves as the primary reference backbone for the new data presented here.

Finally, in 2025 another scaffold-level genome[31] was deposited in NCBI, from a Uruguayan specimen, although the sequencing reads were not made available. It has 100x coverage and N_{50} = 2.4 Mbp and was assembled with Supernova[32]. Sequences were obtained through the single-tube long fragment read (stLFR) technology[33]. It was delivered under the umbrella of the 10,000 Fish Genomes Project (Fish10K)[34].

Here, I present a pair of chromosome-level, reference-guided consensus pseudohaplotypes for a *Salminus brasiliensis* specimen from the Upper Paraná River Basin. Together with the two VGP phased haplotypes from a geographically distant individual, they allow for initial screening of genome-wide variants for this fish species. I also present a minimal chromosome-level reference pangenome graph.

The two haploid datasets (LARGE's Hap1 and Hap2) are 1.078 and 1.067 Gb, respectively, and the minimal reference pangenome graph versions stem from the simultaneous combination of four chromosome-level genomes: the two phased VGP haplotypes (Argentine specimen) and the two LARGE consensus pseudohaplotypes (Brazilian specimen). The full graph version is 1.1 Gb long, including unplaced scaffolds, comprising 288,788 segments, 404,621 links (809,242 directed arcs), with an average node degree of 1.4 (maximum of 3). The reduced graph version (25 chromosomes only) is also ~ 1.1 Gb, with 277,993 segments and 389,619 links.

These genomic resources are shared under the FAIR guiding principle[35], accompanied by sufficient methodological, technical and machine-readable metadata for broader computational reuse, with direct applicability in ecology, evolution and biotechnology.

Methods

Ethics statement

This work did not involve any new animal specimen collection. DNA reads were reused from a previously published sample under its original permits[12].

Bioinformatics

For all major bioinformatics steps below, default parameters were applied, unless otherwise noted. I applied previously available paired-end NGS sequencing data[25], from a single size-selected whole-genome shotgun library (500 bp inserts) used to resolve the chromosome-level genome for a specimen from the Upper Paraná Basin (*Laboratório de Recursos Genéticos* voucher LARGE130547, also known as LARGE13054, henceforth LARGE). It consists of ~ 89 million Illumina HiSeq 2000 read pairs (90 bases), previously quality- and duplicate-filtered from the NCBI's BioSample SAMN24583349 subject[26]. Reads are Phred + 64-encoded, with average Q of 36.6 (R1) and Q36.0 (R2) and per-position mean quality Q30.7–39.0 (seqtk[36], version 1.4).

Initially, the short reads from the subject sample were individually aligned to both VGP's *S. brasiliensis* phased nuclear haplotypes[28] (specimen from Argentina): the principal reference assembly, haplotype 2, containing 25 chromosomes and 78 unplaced scaffolds (RefSeq accession GCF_030463535.1); and independently to the alternate one, haplotype 1, which has 157 unplaced scaffolds (Genome accession GCA_030448965.1). For this, alignment was performed with BWA-MEM (Burrows-Wheeler Aligner[37] - version 0.7.17), penalizing soft-clipping with parameter '-L 30' (rewarding end-to-end alignments) to mitigate reference bias, particularly for indels[38]. The alignments were coordinate-sorted with SAMtools (version 1.13), without additional duplicate marking, and subsequent BAM handling and statistics used SAMtools (version 1.19.2); variant calling used BCFtools (version 1.22)[39].

Subsequently, I applied BCFtools mpileup to the BAM files from the previous step, passing the output to BCFtools call ('-mv', default diploid ploidy), invoking its Bayesian multiallelic model for short insertions/deletions (indels) and single nucleotide variants (SNVs), and outputting only variant sites to each resulting VCF file. This allowed me to generate two chromosome-level pseudohaplotypes for the LARGE specimen using BCFtools consensus. For Hap2, the '--iupac-codes' flag was applied: homozygous ALT alleles were written as unambiguous bases, and the LARGE specimen's heterozygous SNVs (GT 0/1, one allele matching the reference; GT 1/2, two non-reference alleles) were recorded as the IUPAC ambiguity code representing its two alleles. For Hap1, BCFtools consensus was run without sample-aware options, applying the first ALT allele at every site where the LARGE specimen carries a non-reference allele, regardless of genotype; its own heterozygous sites are therefore represented by their ALT allele only, yielding a sequence free of IUPAC ambiguity codes.

Quality control of resulting genome datasets relied primarily on *a posteriori* BAM alignments, QUAST[40] (version 5.2.0) and BUSCO (version 6.1.0) metrics, using the Actinopterygii OrthoDB (version 12 - *actinopterygii_odb12* - 'eukaryota, 2025-07-01')[41], which is based on 7,207 markers and 75 fish genomes. These checks are reported for Hap2 only. I also performed k-mer distribution spectrum analysis (k = 21), using Merqury[42] (version 1.4.1; meryl 1.4.1), applied to both genome sets Hap1 and Hap2.

Genome features of the 25 LARGE Hap2 chromosomes were visualized as Circos plots with pyCirclize[43] (version 1.10.1). Gene density was computed from genomic annotation (details below) and GC content was calculated with bedtools[44] (version 2.31.1), both in 1 Mb windows. Microsatellites were annotated with RepeatMasker[45] (version 4.2.3; RMBlast 2.14.1+; -species Actinopterygii), from which (CG)n dinucleotide and general tri- to hexanucleotide motifs were extracted in 100 kb windows. Miniature inverted-repeat transposable element (MITE) candidates were identified with MiteFinderII[46] (GitHub commit 868e05e), using the program's own scoring profile (threshold 0.5). Inter-chromosomal paralogous gene pairs were defined from the annotation as protein-coding genes sharing a gene symbol differing only by an a/b suffix, each present once on the 25 chromosomes (n = 2,563). Following checks, the Upper Paraná Basin *S. brasiliensis* whole-genome shotgun projects (one for each pseudohaplotype set) were then submitted to NCBI's Genome database for screening and formal deposit. For each haploid complement, related BAM files were produced and submitted to NCBI's Sequence Read Archive (SRA).

Gene annotation was performed on LARGE's principal Hap2 with the aid of LiftOn[47] (version 1.0.14), which invokes minimap2[48] (version 2.31) and miniprot[49] (version 0.18), using NCBI's RefSeq general feature file (GFF) for VGP's *S. brasiliensis* principal genome (Hap2). I used gffread[50] (version 0.12.7) for extracting polypeptide sequences from the annotation, which were assessed with BUSCO, for quality control of the new annotated proteome data ('-m proteins' - actinopterygii_odb12, n = 7,207). Data files with polypeptide sequences (as FASTA) and genome annotation as general feature file (GFF3) were deposited in Figshare.

Using Minigraph[5] (version 0.21), I generated two versions of the minimal chromosome-level reference pangenome graph: i) a reduced representation consisting only of the 25 chromosomes, and ii) a full version which also includes all the remaining unplaced scaffolds. Both used the VGP's Hap2 assembly as the reference backbone and were constructed with '-cxggs' (CIGAR/base-alignment and preset incremental mode). The augmentation input order for both instances was VGP Hap1, LARGE Hap2 and finally LARGE Hap1. I then aligned the short reads to the reduced graph using minigraph with '-xsr' (short-read preset), resulting in a GAF file. The latter and the graph files were deposited in Figshare. For short-read alignment, the full version of the reference pangenome graph was converted to vg format with nodes split at 1,024 bp (vg convert; vg mod --chop 1024), and a GBZ index was built from its embedded reference paths (vg gbwt --index-paths; --gbz-format). Because vg recovers paths only for the rank-0 (VGP Hap2) sequences of the reference pangenome graph, this index comprises the backbone subgraph only. Paired-end reads were aligned to it with vg giraffe[51] (version 1.75.1) using default parameters, with the distance and minimizer indexes (k = 29, w = 11) generated by giraffe, yielding a Graph Alignment/Map (GAM) file. This file was deposited at Zenodo, given its size, and the index was deposited at Figshare. Pangenome graph visualization was conducted with Bandage[52] (version 0.9.0).

Data Records

The main genomic resources for the *Salminus brasiliensis* fish species introduced here are available in two main dataset families: I) A pair of chromosome-level pseudohaplotype genomes; and II) a minimal reference pangenome.

The genomes (I) are under NCBI's BioProjects PRJNA1433447 (LARGE's pseudohaplotypes 1 or Hap1) and PRJNA1433448 (LARGE's pseudohaplotype set 2 or Hap2), mirroring their respective VGP references used for reference-guided mapping. Thus, LARGE's Hap2 is the principal genome, consisting of 25 defined chromosomes, alongside 78 unplaced scaffolds, four of which are unlocalized scaffolds of chromosome 20. LARGE's Hap1 is the alternative haploid genome. It consists of 25 chromosomes and 157 unplaced scaffolds, with chromosomes and unplaced scaffolds listed as scaffolds in the assembly, without distinction. Short-read alignments for the LARGE specimen are provided as three BAM file sets per haplotype at NCBI's SRA, under accessions SRP684677[53] (Hap1) and SRP684629[54] (Hap2). These comprise reads mapped onto: i) the full LARGE genomes, including unplaced scaffolds; ii) LARGE's 25 main chromosome sets only; and iii) the main 25 chromosomes version from the original VGP deposits.

These new whole-genome shotgun projects have been deposited at DDBJ/ENA/GenBank under the accession JCAFYL000000000 (Hap1) and JCAFYM000000000 (Hap2). The versions described in this paper are JCAFYL010000000[55] and JCAFYM010000000[56], respectively. The gene annotation for LARGE's Hap2 is available as a general feature file[57] (GFF3) at Figshare (https://doi.org/10.6084/m9.figshare.34023231) and its corresponding predicted proteome is deposited separately in FASTA format[58] (https://doi.org/10.6084/m9.figshare.34027611).

The second dataset family (II) comprises the minimal reference pangenome graph for *S. brasiliensis*, provided in two versions at Figshare: i) a full version[59] including all unplaced scaffolds (https://doi.org/10.6084/m9.figshare.32678319) and ii) a reduced version[60] comprising the 25 chromosome sets only (https://doi.org/10.6084/m9.figshare.34005651). The GAM file[61] 29.1 GB (gigabytes), contains the alignment of the LARGE specimen's short reads to the backbone subgraph (VGP Hap2 reference paths) of the full graph version (i); it is available at Zenodo (https://doi.org/10.5281/zenodo.20706937). Short-read alignments to the reduced version (ii) are provided in GAF format[62] (https://doi.org/10.6084/m9.figshare.34005663). Additionally, the GBZ index[63] of the backbone subgraph, required to interpret the node identifiers of the GAM file, is available at Figshare (https://doi.org/10.6084/m9.figshare.34335735).

Data Overview

The two *S. brasiliensis* genomes presented here (Table 1) inherit their chromosome-level structure from the VGP's reference haplotypes and are expected to incorporate reference bias, particularly considering long insertions and tandem repeat expansions. The LARGE Hap2 pseudohaplotype encodes the specimen's heterozygous SNVs as IUPAC ambiguity codes (R, Y, W, S, K and M), each representing the two alleles carried by the LARGE fish at that site. The LARGE Hap1 genome, in turn, contains no IUPAC

codes; the first ALT allele is applied at every variant site, regardless of the specimen's genotype (heterozygous 0/1 or 1/2, or homozygous 1/1). The LARGE Hap2 genome gene annotation accounts for 33,984 genes, including 24,084 protein-coding and immune loci (24,007 protein-coding genes plus 77 immune gene segments — V_segment and C_region types). The reduced and full minimal reference pangenome graph versions differ only in the inclusion or omission of unplaced scaffolds (54.1 Mb or 2.52% across both LARGE pseudohaplotypes). The accompanying alignments map the Upper Paraná River Basin specimen onto the reduced pangenome graph (GAF) and onto the backbone subgraph of the full graph (GAM).

Table 1
Genome statistics for *Salminus brasiliensis* from the Upper Paraná River Basin (sample LARGE130547). SeqIDs: number of sequences (chromosomes plus unplaced scaffolds). N_{50}: sequence length at which ≥ 50% of the assembly is contained in sequences of that length or longer; L_{50}: number of such sequences; N are positionally defined, indeterminate bases; and IUPAC are standard DNA base ambiguity symbols.

Metric	Hap1 (alternate)	Hap2 (principal)
Length (bp)	1,078,312,478	1,066,501,764
SeqIDs	182	103
Chromosomes	25	25
Unplaced scaffolds	157	78
Largest sequence	101,951,486 bp	99,396,630 bp
N_{50}	38,932,466 bp	40,765,212 bp
L_{50}	11	11
N	1.88%	0.85%
IUPAC	0% (0)	0.28% (3,023,889)
Runs of N (gaps)	641	726

Technical Validation

Quality metrics support the reliability of the genomic datasets. Despite inherent expected reference bias from mapping and consensus calling from VGP's genomes, the new LARGE pseudohaplotypes retain their own genetic signatures. This is exemplified by verifying chromosome length variation (Δ) of up to 0.011% for LARGE Hap2 (Table 2). The short reads from specimen LARGE130547 map back to LARGE Hap2 at a rate of 98.99%, with 93.9% being properly paired. Only 1% of reads did not map back to it, partly because VGP's references do not include the mitochondrial chromosome (the LARGE fish

specimen's mitogenome is under GenBank accession KY825190.1[12]). Notably, the fraction of reads with 0 mapping quality (MQ_0 = 17.9%) is consistent with the expected repeat content of teleost genomes and does not necessarily reflect systematic poor mapping.

In Hap2, multiallelic SNV calls (5,337, or 0.18% of heterozygous SNVs), from positions with limited pileup depth, constitute a negligible low-quality fraction. Rebuilding Hap2 from the BCF file with BCFtools consensus reproduces the deposited sequence's 3,023,889 IUPAC-coded bases: 2,885,301 at heterozygous SNVs and 138,588 at heterozygous indel records. At these records, BCFtools applies the ALT allele and encodes base-wise mismatches across the overlapping REF and ALT strings, mostly where an indel shifts a tandem repeat out of frame; these codes do not represent single-nucleotide heterozygosity. The two LARGE pseudohaplotypes therefore differ mainly in their respective VGP backbones and the encoding of heterozygous sites. Figures 1–3 show examples of Hap2's genomic features.

Table 2

Chromosome (Chr.) sizes of the VGP *Salminus brasiliensis* principal reference assembly (Hap2) and the corresponding LARGE Hap2 consensus pseudohaplotype. Δ = LARGE − VGP. GC content was identical for every chromosome pair.

Chr.	VGP RefSeq accession	VGP size (bp)	LARGE GenBank accession	LARGE size (bp)	Δ (bp)	GC (%)
1	NC_132878.1	99,388,286	CM188175.1	99,396,630	+ 8,344	40.5
2	NC_132879.1	60,494,693	CM188176.1	60,492,526	−2,167	41.0
3	NC_132880.1	52,544,285	CM188177.1	52,546,278	+ 1,993	41.5
4	NC_132881.1	50,283,246	CM188178.1	50,285,850	+ 2,604	41.0
5	NC_132882.1	46,574,413	CM188179.1	46,574,155	−258	41.0
6	NC_132883.1	46,419,030	CM188180.1	46,421,178	+ 2,148	41.0
7	NC_132884.1	44,467,116	CM188181.1	44,469,449	+ 2,333	41.0
8	NC_132885.1	43,347,643	CM188182.1	43,348,752	+ 1,109	41.5
9	NC_132886.1	42,620,587	CM188183.1	42,624,741	+ 4,154	41.5
10	NC_132887.1	41,637,409	CM188184.1	41,635,246	−2,163	41.5
11	NC_132888.1	40,765,202	CM188185.1	40,765,212	+ 10	41.0
12	NC_132889.1	40,068,765	CM188186.1	40,069,717	+ 952	42.0
13	NC_132890.1	38,865,264	CM188187.1	38,865,047	−217	41.5
14	NC_132891.1	37,349,020	CM188188.1	37,346,743	−2,277	41.5
15	NC_132892.1	36,871,833	CM188189.1	36,875,773	+ 3,940	41.5
16	NC_132893.1	36,840,974	CM188190.1	36,840,508	−466	41.5
17	NC_132894.1	36,758,898	CM188191.1	36,759,448	+ 550	41.5
18	NC_132895.1	36,616,962	CM188192.1	36,613,648	−3,314	41.0
19	NC_132896.1	36,202,799	CM188193.1	36,206,046	+ 3,247	41.0
20†	NC_132897.1	33,995,598	CM188194.1	33,996,257	+ 659	40.5
21	NC_132898.1	34,019,765	CM188195.1	34,020,798	+ 1,033	41.5
22	NC_132899.1	33,960,001	CM188196.1	33,958,819	−1,182	41.5
23	NC_132900.1	32,802,579	CM188197.1	32,802,673	+ 94	41.5
24	NC_132901.1	29,456,005	CM188198.1	29,454,567	−1,438	41.5
25	NC_132902.1	28,888,514	CM188199.1	28,887,726	−788	42.0

Chr.	VGP RefSeq accession	VGP size (bp)	LARGE GenBank accession	LARGE size (bp)	Δ (bp)	GC (%)
Total		**1,061,238,887**		**1,061,257,787**	**+ 18,900**	

†Chromosome 20 has four unlocalized scaffolds in both assemblies. Together with 74 unplaced scaffolds, these add ~ 5.2 Mb to each assembly, beyond the chromosome totals above.

QUAST flagged four possible misassemblies in LARGE Hap2: two flanking a ~ 20 kb segment at approximately 20.16 Mb on chromosome 12. The other two are short offsets (1–3 kb) at the ends of unplaced scaffolds 131 and 77 (JCAFYM010000053.1 and JCAFYM010000096.1). The total length of sequences flagged as problematic (40.2 Mb) is explained by the length of chromosome 12 itself (40.07 Mb), since the metric counts entire sequences rather than misassembled segments (all four are relocations). The remaining ~ 153 kb are accounted for by the unplaced scaffolds. Independently, per-sequence Merqury Q_V (Fig. 4) identified chromosome 14 (Q_V 29.5), chromosome 15 (QV 30.9) and, more dramatically, scaffold143 (JCAFYM010000064.1; Q_V 17.1) in Hap2 as outliers relative to its genome-wide value (Q_V 33.0). These regions warrant caution in downstream analyses. Union (LARGE's Hap1 and Hap2) k-mer completeness of 94% is reported by Merqury, while BUSCO screening found 97.8% complete BUSCOs (actinopterygii_odb12, n = 7,207) for the Hap2 genome.

Genome annotation transferred 96.9% of reference genes (33,984/35,057) and 35.9% of pseudogene *loci* (2,734/7,620). Of the lifted mRNAs, 92.6% showed protein identity ≥ 0.99 against the VGP reference coding DNA sequence models (Table 3). A minor fraction (~ 0.3%) showed identity below 0.50, more likely reflecting frameshifting indels or lift-over boundary errors than sequence divergence. Conversely, the $0.99 \leq I < 1$ class (54.2%) almost entirely reflects IUPAC-coded heterozygous sites within coding sequences, which translate as ambiguous residues (X): 54.1% of lifted polypeptides contain at least one X. The LARGE Hap2 gene annotation accounts for 33,984 genes across eight biotypes: 24,007 protein-coding genes; 7,103 tRNAs; 1,359 snRNAs; 917 lncRNAs; 267 snoRNAs; 245 rRNAs; 77 immune gene segments (V_segment and C_region); and 9 ncRNAs. Together, the 24,084 protein-coding and immune loci carry 41,435 mRNAs (including isoforms). An additional 2,734 pseudogenes were transferred separately. Protein-mode BUSCO on the lifted annotation proteome (41,435 polypeptide sequences) yielded 97.4% completeness with 32.7% duplication. These duplicated BUSCOs reflect isoform inclusion, rather than true gene-level duplication, consistent with the low genome-mode duplication (~ 0.4%).

The full minimal reference pangenome graph (Fig. 5a) records segment provenance from the rank tags in relation to the backbone (VGP Hap2), which accounts for 96.5% of the graph's sequence, while VGP Hap1 adds 33.2 Mb (78,405 segments). LARGE Hap2, in turn, adds 357 segments (24.7 kb) and LARGE Hap1 adds another 8,346 segments (5.2 Mb). In chromosome 1 alone, for example, 93% of LARGE Hap1-derived segments (≥ 100 bp) also occur in VGP Hap1, indicating that its contribution largely re-captures VGP Hap1, plus segments not previously incorporated during incremental augmentation. As expected for reference-guided consensus pseudohaplotypes, which mostly differ from their backbones by SNVs and

short indels (below minigraph's default minimum variant length of 50 bp), graph topology mainly reflects divergence between the two phased VGP haplotypes. The LARGE specimen's variation is mainly captured in the reference pangenome graph framework through its read alignments to the reduced graph (GAF), complemented by an alignment to the backbone subgraph of the full graph (GAM).

The reduced graph version mirrors the full graph's topology, with an average node degree of 1.4 and maximum of 3 (Fig. 5b), consistent with a predominantly linear structure punctuated by simple bubbles, as expected for a minimal pangenome graph from the minigraph building approach. The GAM file of short-read alignment to the backbone subgraph of the full graph (including scaffolds) yielded a mapping rate of 98.67%, with 94.3% properly paired and 66.5% perfectly aligned. This alignment is consistent with the one for the linear Hap2 genome (98.99%) and with the GAF alignment to the reduced version of the *S. brasiliensis* reference pangenome graph (98.94%).

Table 3
Proportion of *Salminus brasiliensis* mRNAs (including isoforms) by protein identity, from genome annotation lift-over for LARGE Hap2.

Protein identity (I)	**mRNAs**	**% of annotated polypeptides (n = 41,435)**
I = 1	15,909	38.4%
0.99 ≤ I < 1	22,442	54.2%
≥ 0.99 (subtotal)	38,351	92.6%
≥ 0.95 (subtotal)	40,623	98.0%
0.50 ≤ I < 0.95	692	1.7%
I < 0.50	120	0.3%

Together, these metrics indicate that both graph versions incorporate the sequence content of the four input haplotypes with limited expansion: the full graph (1.105 Gb) exceeds its backbone by only 3.6%, consistent with the segment provenance. Short-read alignment to the backbone subgraph with vg giraffe yielded a mean mapping quality of 46.1 (median 60); because giraffe and BWA-MEM calibrate mapping qualities differently, this is not directly comparable with the MQ_0 fraction of the linear mapping (17.9%).

Usage Notes

The datasets provide sequences from a potentially divergent lineage of S. brasiliensis (Upper Paraná River Basin) [17–20] . About 3.6 million heterozygous sites (~ 2.9 million SNVs and ~ 0.7 million indels), or roughly 3.4 per kb, were called against VGP Hap2. The heterozygous SNVs appear across Hap2 as IUPAC ambiguity codes (R = A or G; Y = C or T; W = A or T; S = C or G; K = G or T and M = A or C). A further 138,588 codes, mostly in runs within tandem repeats, arise from the encoding of heterozygous indels, which Hap2 carries with their ALT allele, and should not be interpreted as SNVs. Variant calls were not quality-filtered; users requiring high-confidence sites should filter on QUAL, depth and allelic support. The

pairing of these new Brazilian LARGE pseudohaplotypes with the phased VGP haplotypes from the Argentine specimen provides a direct observation of chromosome-level variation among genomes for *S. brasiliensis*.

The Argentine and Brazilian specimens' combined reference pangenome records SV information as graph topology, with each node carrying a stable sequence name and offset (SN/SO tags), from its linear reference genome backbone coordinates (VGP Hap2). It can be used for incremental sequence alignment/mapping. The reference pangenome graph can be employed for downstream applications such as variant-aware mapping (reference bias mitigation), genotyping and SV annotation. It also serves as a sketch map for comparison and guidance for eventual base-level pangenome graph construction with the minigraph-cactus approach. Graph topology currently mostly reflects structural variation between the two VGP haplotypes; LARGE variation is represented at the SNV/short-indel level and through read alignments. The GAF file uses the rGFA segment names directly, whereas the GAM file, aligned to the backbone subgraph, provides a baseline for comparing graph-based and linear mapping of the same reads; its node identifiers are interpreted with the deposited GBZ index[63].

Declarations

Data Availability

The two chromosome-level pseudohaplotype genome assemblies for *Salminus brasiliensis* (LARGE specimen, Upper Paraná Basin) are available at NCBI GenBank and can be accessed with accessions JCAFYL000000000 (Hap1[55] - https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_059933045.1) and JCAFYM000000000 (Hap2[56] - https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_059958755.1).

Associated short-read alignment files (BAM) are available at NCBI SRA under accessions SRP684677 (Hap1[53] - https://www.ncbi.nlm.nih.gov/sra/?term=SRP684677) and SRP684629 (Hap2[54] - https://www.ncbi.nlm.nih.gov/sra/?term=SRP684629).

The gene annotation[57] (GFF3) and predicted proteome[58] (FASTA) for LARGE's Hap2 are available at Figshare (https://doi.org/10.6084/m9.figshare.34023231 and https://doi.org/10.6084/m9.figshare.34027611, respectively).

The minimal reference pangenome graph for *S. brasiliensis*, in full[59] and reduced[60] (25 chromosomes only) versions, is available at Figshare (https://doi.org/10.6084/m9.figshare.32678319 and https://doi.org/10.6084/m9.figshare.34005651, respectively).

The short-read alignments to the backbone subgraph of the full graph in GAM[61] format are available at Zenodo (https://doi.org/10.5281/zenodo.20706937), and the corresponding GBZ index[63] at Figshare (https://doi.org/10.6084/m9.figshare.34335735).

Short-read alignments to the reduced graph (GAF)[62] are available at Figshare (https://doi.org/10.6084/m9.figshare.34005663).

Funding

This work was funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant Number: 303023/2023-6.

Code Availability

No custom code was used to generate or process the datasets. All analyses used publicly available software; versions and non-default parameters are given in Methods.

Author Contributions

G.M. Yazbeck conceived, planned and conducted this study, performed all bioinformatics analyses, managed data deposits, wrote and edited the manuscript.

Competing Interests

I declare no competing interests in reporting this study.

Acknowledgments

I thank Dr. Raquel Coelho Loures Fontes and Míriam Aparecida de Castro, MSc, at Programa Peixe Vivo (CEMIG), and Dr. Rosiane de Paula Santos for facilitating access to high-performance computing resources. I also thank CNPq and Claude (Anthropic, Inc.) for subsidized access to its Claude Team Plan for Scientists program.

References

1. Gong Y, Li Y, Liu X, Ma Y, Jiang L (2023) A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? J Anim Sci Biotechnol 14:73. https://doi.org/10.1186/s40104-023-00860-1
2. Jeon JY, Allen NM, Black AN, DeWoody JA (2026) Short-Read Pangenomes and Their Potential Utility in Population and Conservation Genomics. *Mol. Ecol. Resour.* 26, e70060 https://doi.org/10.1111/1755-0998.70060
3. Yazbeck GM et al (2026) A fractal pattern of hierarchical genetic population structure in mixed stocks across fish segregated by dams revealed by genomic resources for curimba *Prochilodus lineatus*. J Fish Biol 108:1029–1046. https://doi.org/10.1111/jfb.70278
4. Paten B, Novak AM, Eizenga JM, Garrison E (2017) Genome graphs and the evolution of genome inference. Genome Res 27:665–676. https://doi.org/10.1101/gr.214155.116

5. Li H, Feng X, Chu C (2020) The design and construction of reference pangenome graphs with minigraph. Genome Biol 21:265. https://doi.org/10.1186/s13059-020-02168-z
6. Hickey G et al (2024) Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol 42:663–673. https://doi.org/10.1038/s41587-023-01793-w
7. Cudby J, Bonfield J, Zhou C, Durbin R, Strelchuk S (2026) Pangenome-guided sequence assembly via binary optimization. Brief Bioinform 27:bbag084. https://doi.org/10.1093/bib/bbag084
8. Abe KT, Mariguela TC, Avelino GS, Foresti F, Oliveira C (2014) Systematic and historical biogeography of the Bryconidae (Ostariophysi: Characiformes) suggesting a new rearrangement of its genera and an old origin of Mesoamerican ichthyofauna. BMC Evol Biol 14:152. https://doi.org/10.1186/1471-2148-14-152
9. de Dantas ES O., Diniz PLJ (2020) Ethnoichthyological Study of Biological Characteristics and Ecology (*Salminus brasiliensis*: Teleostei/Characiform): Importance, Status and Conservation Challenges in the Region of Mato Grosso, Brazil. Asian J Fish Aquat Res 10:9–34. https://doi.org/10.9734/ajfar/2020/v10i430187
10. Boersma KS, Bogan MT, Henrichs BA, Lytle DA (2014) Top predator removals have consistent effects on large species despite high environmental variability. Oikos 123:807–816. https://doi.org/10.1111/oik.00925
11. Cao Y-L et al (2016) Microsatellite records for 8, issue 3. Conserv Genet Resour 8:359–370. https://doi.org/10.1007/s12686-016-0581-4
12. Graciano RCD, Oliveira RS, Santos IM, Yazbeck GM (2022) Genomic Resources for *Salminus brasiliensis*. Front Genet 13:855718. https://doi.org/10.3389/fgene.2022.855718
13. Dias RM et al (2022) Importance of protection strategies in the conservation of the flagship species dourado *Salminus brasiliensis* (Characiformes: Bryconidae). *Neotropical Ichthyol.* 20, e220046 https://doi.org/10.1590/1982-0224-2022-0046
14. García-Alonso J, Rodríguez-Yaniero A, Lercari D (2025) Indicators of scientific production show limited and fragmented information for integrated aquatic quality on La Plata River transboundary basin. Environ Sustain Indic 26:100651. https://doi.org/10.1016/j.indic.2025.100651
15. Geller IV, Vitule JRS, Ferraz JD, Pereira AD, Orsi ML (2025) Current and future invasion of a predator with potential for impact negative in a region of high neotropical endemism. Neotropical Ichthyol 23:e250056. https://doi.org/10.1590/1982-0224-2025-0056
16. Braun N, de Lima RL, Baldisserotto B, Dafre AL (2010) Oliveira Nuñer, A. P. Growth, biochemical and physiological responses of *Salminus brasiliensis* with different stocking densities and handling. Aquaculture 301:22–30. https://doi.org/10.1016/j.aquaculture.2010.01.022. de
17. Rosso JJ, Mabragaña E, Castro MG, de Astarloa JM (2012) D. DNA barcoding Neotropical fishes: recent advances from the Pampa Plain, Argentina. Mol Ecol Resour 12:999–1011. https://doi.org/10.1111/1755-0998.12010
18. Machado CDB, Ishizuka TK, Freitas PDD, Valiati VH, Galetti Jr. (2017) P. M. DNA barcoding reveals taxonomic uncertainty in *Salminus* (Characiformes). Syst Biodivers 15:372–382.

https://doi.org/10.1080/14772000.2016.1254390

19. Rosso JJ et al (2018) Basin-scale distribution and haplotype partitioning in different genetic lineages of the Neotropical migratory fish *Salminus brasiliensis*. Aquat Conserv Mar Freshw Ecosyst 28:444–456. https://doi.org/10.1002/aqc.2830
20. Mondin LA, de Machado C, Resende CB, de Marques EK, D. K. S., Galetti PM (2018) Genetic Pattern and Demographic History of *Salminus brasiliensis*: Population Expansion in the Pantanal Region during the Pleistocene. Front Genet 9:1. https://doi.org/10.3389/fgene.2018.00001
21. Frederico RG (2023) IUCN Red List of Threatened Species: Salminus brasiliensis. IUCN Red List Threat. Species https://doi.org/10.2305/IUCN.UK.2023-1.RLTS.T186400A1812492.en
22. Graciano RCD, Oliveira RS, Santos IM, Yazbeck GM (2022) Salminus brasiliensis de novo short-reads genome assembly k-mer 47. https://doi.org/10.6084/m9.figshare.18131495.v1. Figshare
23. Graciano RCD, Oliveira RS, Santos IM, Yazbeck GM (2022) Genes characterized in the Salminus brasiliensis genome. https://doi.org/10.6084/m9.figshare.19169756. Figshare
24. Graciano RCD, Oliveira RS, Santos IM, Yazbeck GM (2022) Genome-wide microsatellite panel for Salminus brasiliensis. https://doi.org/10.6084/m9.figshare.17754188.v1. Figshare
25. NCBI Sequence Read (2022) Archive https://identifiers.org/ncbi/insdc.sra:SRP353262
26. Graciano RCD, Oliveira RS, Santos IM, Yazbeck GM (2022) Salminus brasiliensis isolate LARGE130547. NCBI BioSample https://identifiers.org/biosample:SAMN24583349
27. Rhie A et al (2021) Towards complete and error-free genome assemblies of all vertebrate species. Nature 592:737–746. https://doi.org/10.1038/s41586-021-03451-0
28. Komoroske L, Fedrigo O, Jarvis ED, GenBank (2023) https://identifiers.org/ncbi/insdc:JASKUN000000000
29. NCBI Sequence Read (2023) Archive https://identifiers.org/ncbi/insdc.sra:SRP527408
30. NCBI Sequence Read (2023) Archive https://identifiers.org/ncbi/insdc.sra:SRP518386
31. Song Y, GenBank (2025) https://identifiers.org/ncbi/insdc:JBNTTH000000000
32. De Visendi P (2022) Novo Assembly of Linked Reads Using Supernova 2.0. Methods Mol Biol 2443:233–243. https://doi.org/10.1007/978-1-0716-2067-0_12
33. Wang O et al (2019) Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res 29:798–808. https://doi.org/10.1101/gr.245126.118
34. Fan G et al (2020) Initial data release and announcement of the 10,000 Fish Genomes Project (Fish10K). *GigaScience* 9, giaa080 https://doi.org/10.1093/gigascience/giaa080
35. Wilkinson MD et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. https://doi.org/10.1038/sdata.2016.18
36. Li H (2023) seqtk, version 1.4. GitHub https://github.com/lh3/seqtk
37. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324

38. Lin M-J, Iyer S, Chen N-C, Langmead B (2024) Measuring, visualizing, and diagnosing reference bias with biastools. Genome Biol 25:101. https://doi.org/10.1186/s13059-024-03240-8
39. Danecek P et al (2021) Twelve years of SAMtools and BCFtools. *GigaScience* 10, giab008 https://doi.org/10.1093/gigascience/giab008
40. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086
41. Tegenfeldt F et al (2025) OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. Nucleic Acids Res 53:D516–D522. https://doi.org/10.1093/nar/gkae987
42. Rhie A, Walenz BP, Koren S, Phillippy AM (2020) Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21:245. https://doi.org/10.1186/s13059-020-02134-9
43. Shimoyama Y, pyCirclize (2022) Circular visualization in Python, version 1.10.1. https://github.com/moshi4/pyCirclize
44. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033
45. Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr Protoc Bioinforma 25. 4.10.1-4.10.14
46. Hu J, Zheng Y, Shang X (2018) MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. BMC Med Genomics 11:101. https://doi.org/10.1186/s12920-018-0418-y
47. Chao K-H et al (2025) Combining DNA and protein alignments to improve genome annotation with LiftOn. Genome Res 35:311–325. https://doi.org/10.1101/gr.279620.124
48. Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191
49. Li H (2023) Protein-to-genome alignment with miniprot. Bioinformatics 39:btad014. https://doi.org/10.1093/bioinformatics/btad014
50. Pertea G, Pertea MGFF, Utilities (2020) GffRead and GffCompare. *F1000Research* 9, ISCB Comm J-304 https://doi.org/10.12688/f1000research.23297.2
51. Sirén J et al (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science 374. https://doi.org/10.1126/science.abg8871. abg8871
52. Wick RR, Schultz MB, Zobel J, Holt KE (2015) Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352. https://doi.org/10.1093/bioinformatics/btv383
53. NCBI Sequence Read (2026) Archive https://identifiers.org/ncbi/insdc.sra:SRP684677
54. NCBI Sequence Read (2026) Archive https://identifiers.org/ncbi/insdc.sra:SRP684629
55. Yazbeck GM, GenBank (2026) https://identifiers.org/ncbi/insdc:JCAFYL000000000
56. Yazbeck GM, GenBank (2026) https://identifiers.org/ncbi/insdc:JCAFYM000000000

57. Yazbeck GM (2026) Genome annotation for Salminus brasiliensis specimen from the Upper Paraná Basin (LARGE Hap2), from LiftOn of RefSeq's VGP reference. https://doi.org/10.6084/m9.figshare.34023231.v1. figshare
58. Yazbeck GM (2026) Salminus brasiliensis LiftOn predicted proteome for a specimen from the Upper Paraná River Basin (LARGE Hap2). https://doi.org/10.6084/m9.figshare.34027611.v1. figshare
59. Yazbeck GM (2026) Salminus brasiliensis minimal reference pangenome graph, with four full chromosome-level haplotypes (two individuals: ARG-BRA), including unplaced scaffolds. https://doi.org/10.6084/m9.figshare.32678319.v1. figshare
60. Yazbeck GM (2026) Salminus brasiliensis minimal reference pangenome graph (reduced version), with four chromosome-level haplotypes (two individuals: ARG-BRA), Only 25 chromosome sets (without unplaced scaffolds). figshare https://doi.org/10.6084/m9.figshare.34005651.v1
61. Yazbeck GM (2026) The minimal pangenome graph for Salminus brasiliensis a Neotropical flagship species, or it takes two to tango. https://doi.org/10.5281/zenodo.20706937. Zenodo
62. Yazbeck GM (2026) Reads to Graph alignments of short-reads to the reduced reference pangenome graph for Salminus brasiliensis, 25 chromosomes only version (without unplaced scaffolds) - GAF. https://doi.org/10.6084/m9.figshare.34005663.v1. figshare
63. Yazbeck GM (2026) Backbone GBZ index of the Salminus brasiliensis minimal reference pangenome graph used for vg giraffe short-read alignment. https://doi.org/10.6084/m9.figshare.34335735. figshare

Figures

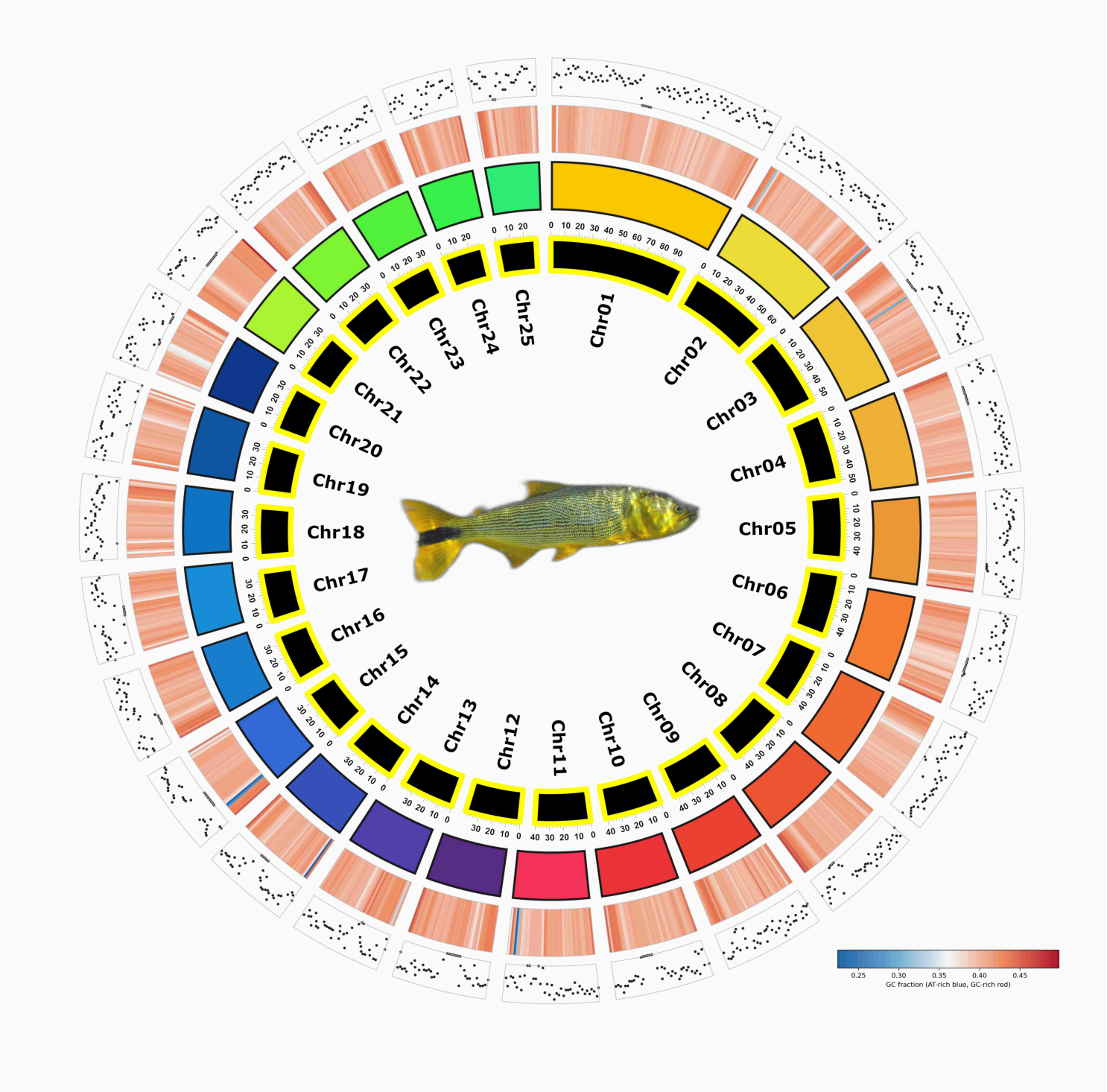


Figure 1

Genome features of *Salminus brasiliensis* LARGE Hap2 (Upper Paraná Basin, Brazil). Innermost ring: chromosome positions (Mb). Second ring: chromosome ideogram. Third ring: GC content (%). Outermost ring: gene density per 1 Mb window.

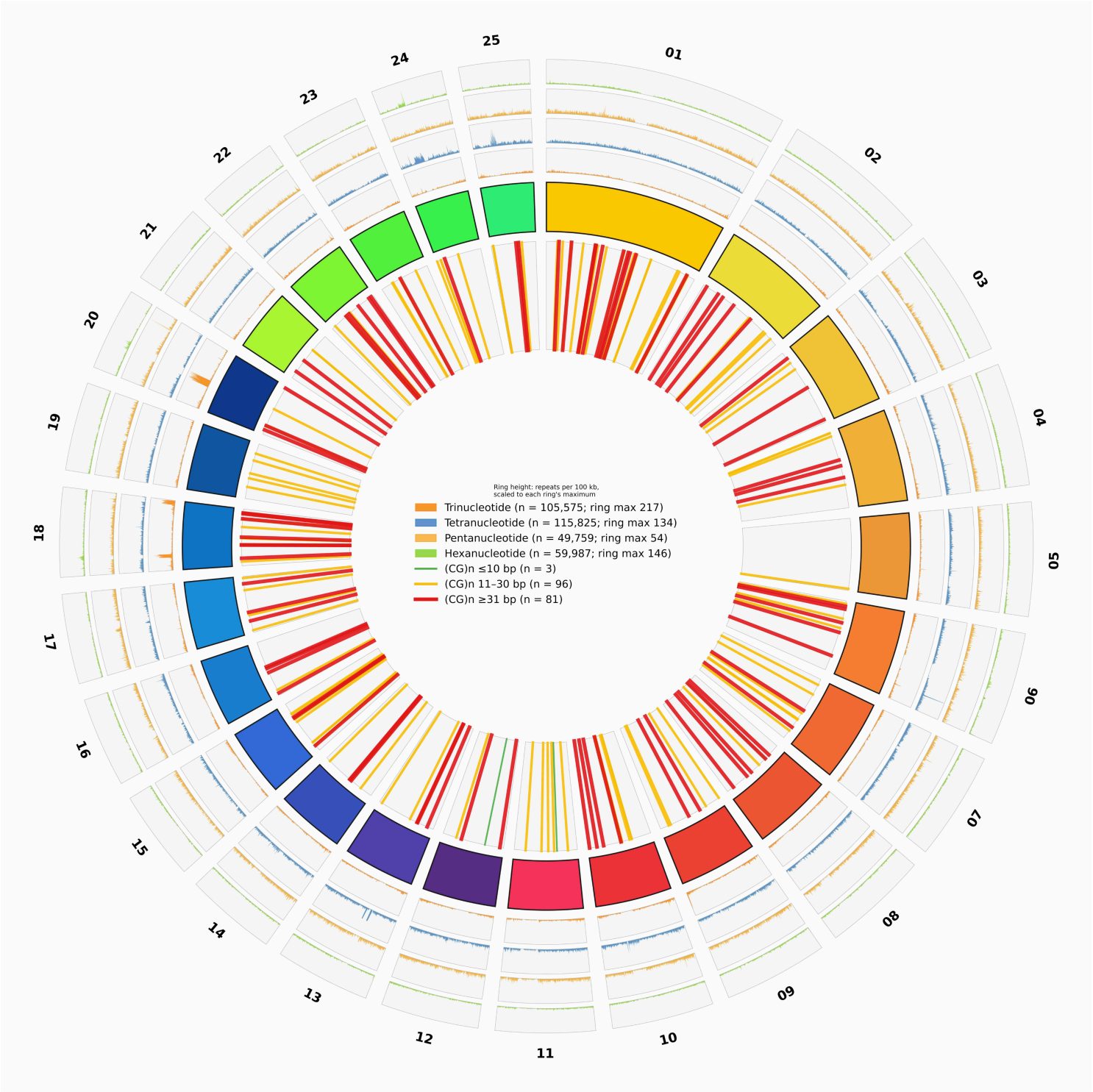


Figure 2

Genomic microsatellite distribution for *Salminus brasiliensis* (100 kb windows). Innermost ring: (CG)n dinucleotide locations, where red-marked regions may contain CpG islands. Second ring: chromosome ideogram. The following outward rings show general simple sequence repeat motif density (per window).

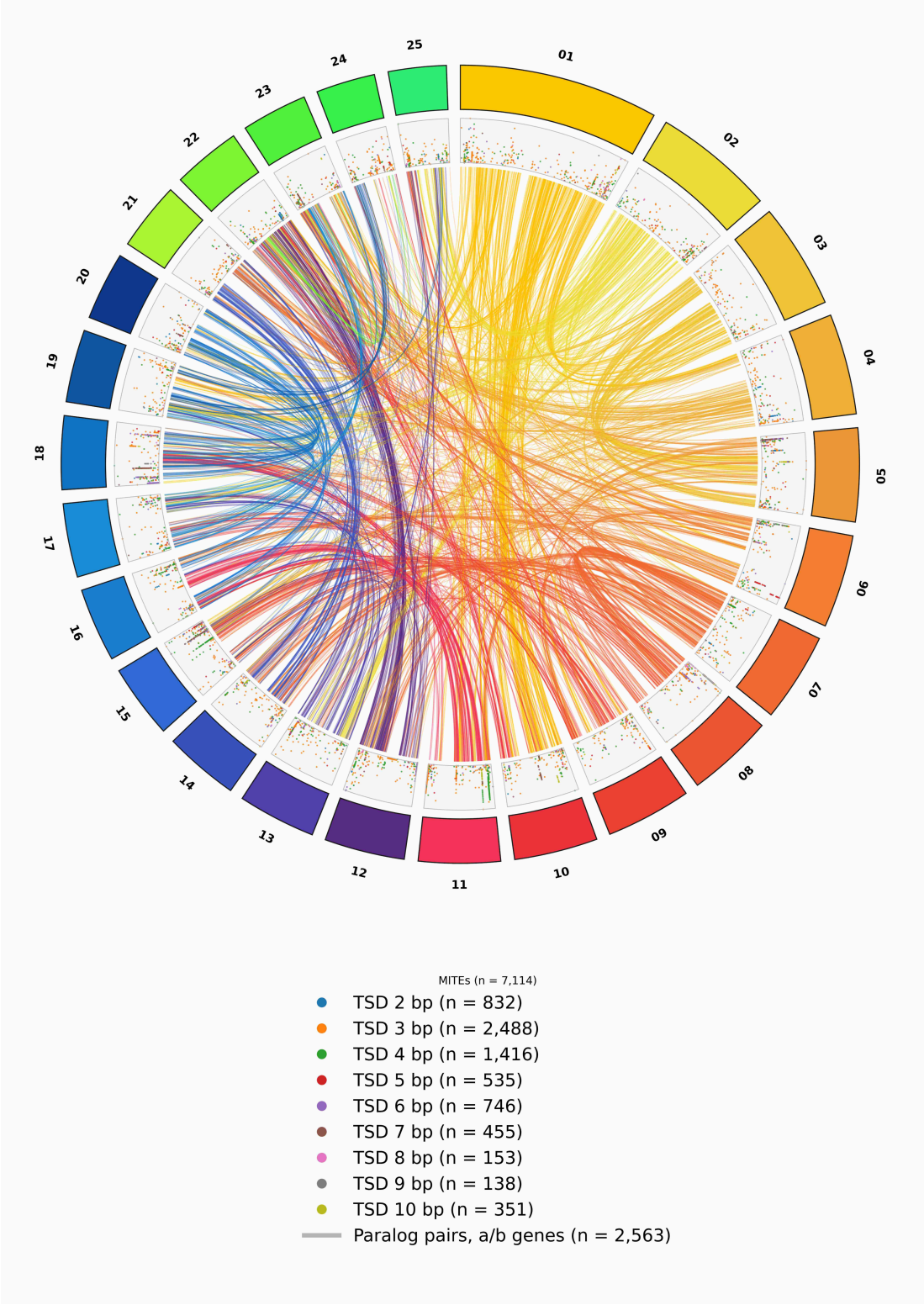


Figure 3

Internal links: intragenomic protein-coding paralogous gene pairs (n = 2,563), coloured by the lower-numbered chromosome of each pair. Innermost ring: Miniature Inverted-repeat Transposable Elements (MITE) candidate locations (n = 7,114) placed by position, MiteFinderII score (increasing outward along the radial axis) and coloured by target-site duplication (TSD) length (2–10 bp). Outermost ring: chromosome ideogram.

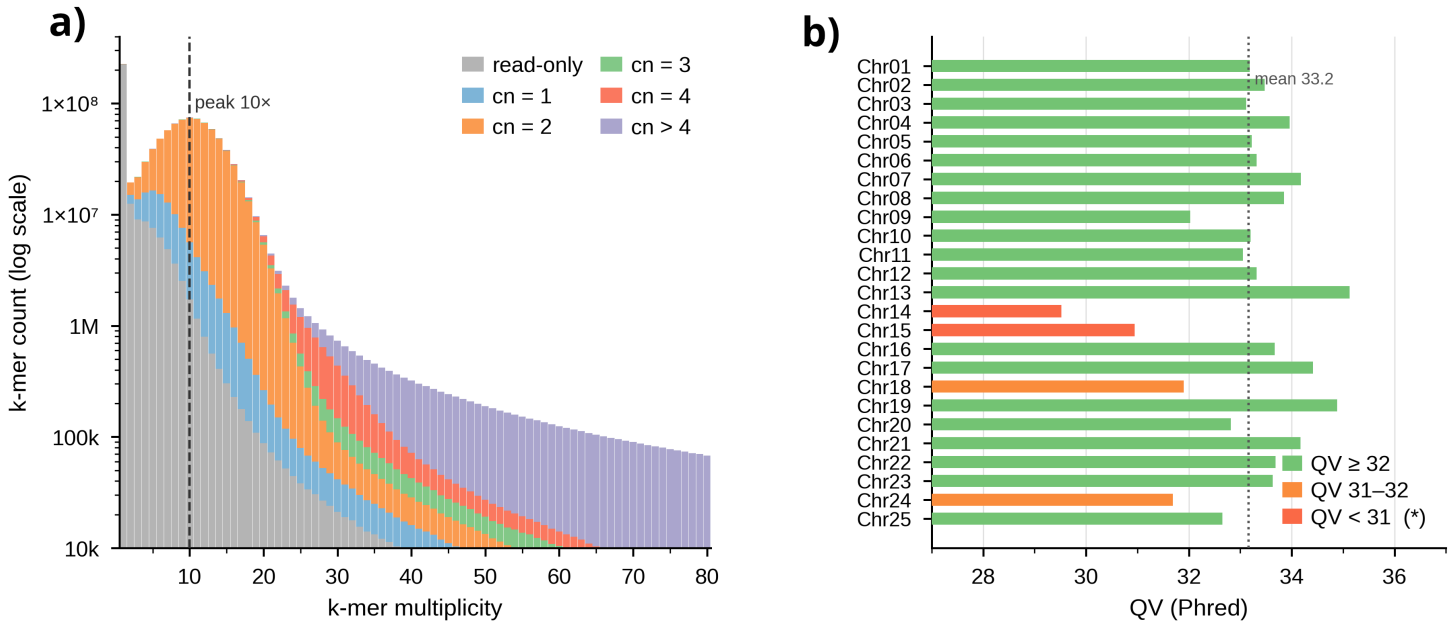


Figure 4

Quality values (Q_V), estimated from k-mer distribution spectra, across the two LARGE pseudohaplotypes (Hap1 and Hap2) for *Salminus brasiliensis*. a) Log-transformed k-mer count as a function of k-mer frequency and copy number (cn). b) Quality values for the main 25 chromosomes, flagging below-average values, particularly for chromosomes 14 and 15.

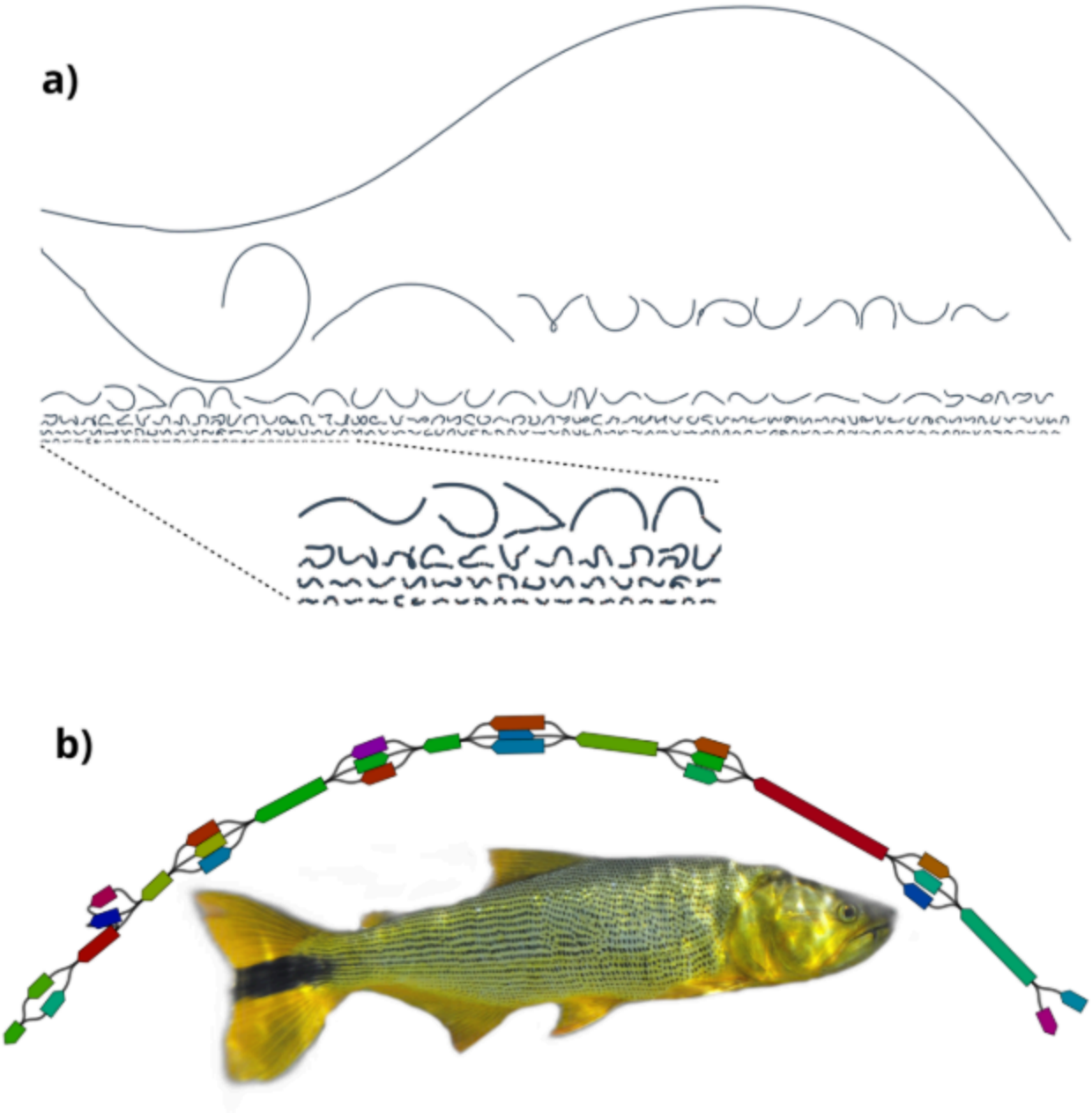


Figure 5

Visualization examples of a minimal reference pangenome graph for *Salminus brasiliensis,* with four haploid genome sets. **a)** Overview of the full version, including unplaced scaffolds, with detail inset. Lines represent genome segments (graph nodes), proportional to their lengths; arrows represent orientation; and graph edges represent connections among nodes. **b)** A subgraph section from the reduced version, between nodes s97017–s97025, showcasing third-degree traversal depth. Coloured arrow-shaped nodes represent sequence segment lengths. Edges (dark lines) represent node connections. Bubbles along the common backbone correspond to divergent haplotype-specific regions.