Evolution, 2025, **0**(**0**), 1–17
https://doi.org/10.1093/evolut/qpaf188
Advance access publication 23 September 2025
Original Article

A pangenomic approach reveals the sources of genetic variation fueling the rapid radiation of Capuchino Seedeaters

María Recuerda[1], Simón Kraemer[2], Jonas R. R. Rosoni[2], Márcio Repenning[3], Melanie Browne[2], Juan Francisco Cataudela[2], Adrián S. Di Giacomo[2], Cecilia Kopuchian[2], Leonardo Campagna[1,4]

[1]Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, Ithaca, United States
[2]Laboratorio de Biología de la Conservación, Centro de Ecología Aplicada del Litoral (CECOAL), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Corrientes, Argentina
[3]Universidade Federal do Rio Grande (FURG), Laboratório de Aves Aquáticas e Tartarugas Marinhas (LAATM), Rio Grande do Sul, Brazil
[4]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, United States

Corresponding author. Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, Ithaca, 14850, New York, United States. Email: mjr432@cornell.edu

Abstract

The search for the genetic basis of phenotypes has primarily focused on single nucleotide polymorphisms, often overlooking structural variants (SVs). SVs can significantly affect gene function, but detecting and characterizing them is challenging, even with long-read sequencing. Moreover, traditional single-reference methods can fail to capture many genetic variants. Using long reads, we generated a Capuchino Seedeater (*Sporophila*) pangenome, including 16 individuals from 7 species, to investigate how SVs contribute to species and coloration differences. Leveraging this pangenome, we mapped short-read data from 127 individuals, genotyped variants identified in the pangenome graph, and subsequently performed F_{ST} scans and genome-wide association studies. Species divergence primarily arises from SNPs and indels (< 50 bp) in non-coding regions of melanin-related genes, as larger SVs rarely overlap with divergence peaks. One exception was a 55 bp deletion near the *OCA2* and *HERC2* genes, associated with feather pheomelanin content. These findings support the hypothesis that the reshuffling of small regulatory alleles, rather than larger species-specific mutations, accelerated plumage evolution leading to prezygotic isolation in Capuchinos.

Keywords: Capuchino Seedeaters, F_{ST} outlier scan, genome-wide association study (GWAS), *Sporophila*, structural variants, transposable elements

Introduction

Genetic variation, the raw material on which evolutionary forces act, exists in different forms with unique properties. These include single and multi-nucleotide polymorphisms (SNPs and MNPs), small insertion/deletions (indels), and different types of structural variants (SVs)—which encompass insertions, deletions, duplications, inversions, or translocations generally larger than 50 bp. These variant types differ in aspects such as their frequency in the genome, their overall size (i.e., the number of nucleotide bases involved in the variant), and their potential evolutionary impact (Mérot et al., 2020). For instance, because recombination within inversions is mostly suppressed, multiple genes can co-evolve as a unit, forming what is known as a supergene, which can shape complex phenotypes (Schwander et al., 2014). In contrast, the influence of multiple SNPs on a given phenotype can be broken apart by recombination, hindering the ability of such variants to collectively shape a trait unless recombination and/or gene flow are suppressed. Nevertheless, key questions related to the evolutionary significance of SVs, such as whether larger variants tend to drive more complex evolutionary changes, remain unanswered (Mérot et al., 2020).

The different types of genetic variants available for evolution are also shaped by their genomic context and the underlying mechanisms by which they are formed. For example, a copy number mutation in a repetitive region of the genome may be more likely (through replication slippage) than a point mutation (Pumpernik et al., 2008). Transposable elements (TEs), mobile segments of DNA that can copy themselves and integrate in different parts of the genome, contribute to generating mutations and shaping genome evolution (Bourque et al., 2018). TE-derived mutations are not necessarily random, as TEs can preferentially integrate in certain genomic regions and be more prevalent in certain genomes versus others (Wells & Feschotte, 2020; Zhang et al., 2020). Therefore, the rate of genetic change may depend on the type of genetic variants involved, and the availability of genetic variants will partially determine the pace of the evolutionary process. Rapid speciation may be fueled by the availability of novel genetic variation, a process that can be further accelerated by gene flow and recombination. Like mutations, these mechanisms can also introduce genetic variants into different genomic backgrounds, providing new genomic variation for evolution to act on (Marques et al., 2019). It is generally unknown whether certain types of mutations can more rapidly lead to the evolution of new traits and species, although there is a growing literature on the evolutionary importance of chromosomal inversions (Mérot et al., 2020; Wellenreuther & Bernatchez, 2018).

Received April 5, 2025; revisions received September 4, 2025; accepted September 17, 2025
Associate Editor: Claire Mérot; Handling Editor: Tim Connallon

Despite this diversity of variant types, SNPs remain the most commonly used genetic markers in genomic studies of non-model organisms, largely due to technological and analytical limitations (Campagna & Toews, 2022; Mérot et al., 2020). Some of these limitations stem from challenges in reference genome construction. The prevalence of short-read sequencing technologies complicates the assembly of complex repetitive regions of the genome, as these reads typically do not span such regions, leading to incorrect assemblies (Kellogg, 2015). Consequently, repetitive areas like those rich in TEs are either incorrectly assembled or split into many small scaffolds, resulting in these types of genomic regions (and genetic changes) being underrepresented in genomic studies. Additionally, limitations also arise from the common reliance on a single reference genome, either from the focal study species or a closely related one, to which population level whole-genome re-sequencing data of a larger number of individuals are mapped (da Fonseca et al., 2016). This process can introduce what is known as reference bias, missing SVs that are absent from the reference genome, as variants in a given population or species that are not represented in this reference (for example, an insertion) will be lost in the alignment process (Recuerda & Campagna, 2024). Finally, although small indel mutations can still be recovered from short-read sequencing data, most of the bioinformatics pipelines and population genomic software for downstream analyses are designed for SNP data, on which many researchers tend to focus (Pool et al., 2010). Thus, the growing number of genomic studies on non-model organisms have primarily focused on SNPs in non-repetitive regions of the genome to assess patterns of genetic variation. Nevertheless, it is increasingly possible to detect SVs in resequencing datasets using a combination of methods (reviewed in Mahmoud et al., 2019). However, it remains technically challenging to detect long or complex SVs and those embedded within repetitive regions (Mahmoud et al., 2019). Therefore, our understanding of how genetic changes other than SNPs contribute to evolution remains limited, particularly in non-model systems.

The use of long-read sequencing technologies can produce higher quality reference genomes and improve SV discovery by spanning repetitive regions and thus improving genomic assemblies (Kellogg, 2015). Moreover, the use of these technologies with approaches that leverage the combination of several reference genomes into a pangenome can capture a more complete representation of the genomic variation in a species or population (Wang et al., 2022). Ideally, a pangenome represents the full spectrum of genetic variation present in an individual or the entire sample under study, such as a population or multiple closely related species. The use of pangenomes has the potential to help mitigate the bias against genetic variants that have been traditionally harder to detect in evolutionary studies of non-model organisms, offering a more comprehensive view of genetic variation, including rare and population-specific SVs.

In this study, we focus on a rapid radiation of 12 bird species in the genus *Sporophila* known as the Capuchino Seedeaters, which originated during the Pleistocene, roughly within the last million years (Campagna et al., 2012, 2013; Lijtmaer et al., 2004). Capuchinos differ primarily in adult male vocalizations and plumage, traits that in these species mediate assortative mating, yet show low genome-wide genetic differentiation between species (F_{ST} ~0.008; Campagna et al., 2017). Song evolution is a mostly cultural process in songbirds (but see Wheatcroft & Qvarnström, 2017), while coloration differences are inherited genetically. Male coloration differences between Capuchino species follow a modular pattern, with distinct patches (e.g., throat, belly, cap) consistently varying in a series of colors (e.g., black, cinnamon, white). For example, the Dark-throated Seedeater (*Sporophila ruficollis*), the Tawny-bellied Seedeater (*Sporophila hypoxantha*), and the Marsh Seedeater (*Sporophila palustris*) differ by having black, cinnamon, or white throats, respectively. Despite the overall genomic homogeneity, previous studies have identified a small number of narrow genomic regions with elevated differentiation, many of which are near genes involved in the melanogenesis pathway (Campagna et al., 2017; Estalles et al., 2022; Turbek et al., 2021), and have undergone selective sweeps (Hejase et al., 2020). Genetic changes in these regions containing melanogenesis genes are strongly associated with variation in the composition of melanin pigment types and their deposition across different body parts in the Capuchinos (Estalles et al., 2022).

The genetic variants that are candidates for controlling plumage coloration are predominantly non-coding SNPs near otherwise conserved pigmentation genes (Campagna et al., 2017; Estalles et al., 2022). These non-coding regions are in some cases conserved across more distantly related species, suggesting they could serve important regulatory functions (Campagna et al., 2017). The outlier regions are repeatedly involved in the divergence between different Capuchinos and generally do not contain species-specific variants but rather have shared haplotypes among species in unique combinations across the different divergence peaks (Campagna et al., 2017; Turbek et al., 2021). For example, the Iberá Seedeater (*Sporophila iberaensis*) and *S. ruficollis*, both with black throats, share genotypes near the *TYRP1* gene, yet differ in a genomic region close to the *HERC2* and *OCA2* genes, which is in turn also shared between *S. iberaensis* and other Capuchinos (Turbek et al., 2021). The unique combinations of genotypes across multiple outlier regions may underlie the emergence of novel coloration phenotypes (Marques et al., 2019; Turbek et al., 2021). Taken together, these findings suggest that the sharing and reshuffling of regulatory alleles at pigmentation genes (e.g., Wallbank et al., 2016) may have been the engine behind the generation of novel plumage patterns. These phenotypic differences function in mate recognition, leading to the establishment and maintenance of species boundaries early in the speciation process (Turbek et al., 2021). Additionally, the Z sex chromosome plays a disproportionate role in species differences (Campagna et al., 2017), potentially contributing to rapid evolution, as has been described in other systems (Irwin et al., 2018).

However, these findings are based on genomic studies that employed a single reference genome from a *S. hypoxantha* male sampled in the Esteros del Iberá, Argentina, which was primarily assembled using short-read sequences (Campagna et al., 2017). Moreover, the F_{ST} outlier scans and genome-wide association studies (GWAS) were conducted exclusively using SNPs. It is therefore possible that the variation in non-coding SNPs near melanogenesis genes in the Capuchinos is accompanied by other, yet undetected genetic changes, such as species-specific SVs (perhaps generated by TE activity) absent in the *S. hypoxantha* individual used to assemble the reference genome. A preliminary analysis using long-read sequences to compare a pool of three Pearly-bellied Seedeaters (*Sporophila pileata*) to the *S. hypoxantha* reference genome found ~500 SVs between these two individuals, four of which were small inversions (~450 bp) located within the much larger areas of genomic divergence (with an average length of ~243 kb) (Campagna et al., 2017). This result shows that SVs may be present in at least some divergence peaks, but their prevalence and level of differentiation across species remain unknown.

Here, we aim to assess the relative contribution of different types of mutations to the evolution of Capuchinos, with the goal of achieving a better understanding of the genomic changes promoting rapid speciation. To this end, we assembled a pangenome from 16 individual reference genomes generated de novo through high-coverage Pacific Biosciences long-read sequencing. This Capuchino pangenome combined information from males and females of the seven species present in the area showing the highest sympatry in this group, the Esteros del Iberá in the Province of Corrientes, Argentina (Campagna et al., 2017; Turbek et al., 2021). We subsequently combined this pangenome with information from previously published and new short-read whole genome resequencing data for all Capuchinos, obtaining genotypes for these individuals for SNPs, indels, and SVs. We used this information in F_{ST} outlier scans and coloration GWAS to ask how the different types of markers contribute to species divergence and coloration differences. We find that the differences in the previously identified divergence peaks among Capuchinos are primarily shaped by SNPs and small indels (< 50 bp). Although we can detect larger SVs, these tend to segregate at low frequencies and generally do not associate with divergence peaks. Our study strengthens the hypothesis that the shuffling of regulatory alleles between Capuchinos has promoted the rapid evolution of plumage traits, which leads to prezygotic reproductive isolation early in the speciation process.

Materials and methods

Sampling and sequencing

Long-read sequencing for pangenome construction

We generated a pangenome by selecting 16 individuals, including between one and four individuals from seven highly sympatric species of Capuchino Seedeaters: *S. palustris* (3), *S. ruficollis* (2), *S. pileata* (1), *S. iberaensis* (4), *S. cinnamomea* (2), *S. hypoxantha* (3), and *S. hypochroma* (1) (Table S1). We extracted high molecular weight DNA using the Zymo Research Quick-DNA HMW MagBead Kit and sequenced all individuals using one PacBio HiFi Revio SMRT Cell per individual at the Novogene (Sacramento, CA) and Cornell Weill (New York, NY) sequencing centers. The average sequencing yield was 72 Gb per sample (range 36–89 Gb) with a mean read length of 15,596 bp (range of 11,376 to 19,637 bp; Table S1). Further details on DNA extraction, library preparation, and sequencing platforms are provided in the Supplementary Methods.

Short-read (Illumina) sequencing for population-level genotyping

We also used previous whole-genome resequencing data from 121 individuals of 10 species (Table S2), and generated new data for 41 individuals (Table S3). Sequencing was performed in two batches using an Illumina NovaSeq X—paired end x 150 bp lane from Novogene and one from the Biotechnology Resource Center at the Cornell Genomics Facility (Table S3). The two sequencing batches yielded an average of 90.4 and 214.6 million raw reads, respectively (Table S3). Details on DNA extraction and library preparation are provided in the Supplementary Methods.

Genome assemblies and annotations

De novo genome assemblies from PacBio HiFi reads

The PacBio HiFi reads were used for de novo genome assemblies. We produced primary and alternate assemblies with hifiasm v0.19.9 (Cheng et al., 2021), followed by the removal of haplotigs using purge_dups v1.2.6 (Guan et al., 2020). Genome size and heterozygosity were estimated using Jellyfish v2.3.0 (Marçais & Kingsford, 2011) and GenomeScope v2.0 (Vurture et al., 2017). GenomeScope predicted a similar heterozygosity and genome length across all assemblies, with an estimated heterozygosity of ~1.2% and an initial estimated genome size of ~0.99 Gb (Table S4, Figure S1). We assessed assembly metrics using assembly-stats v.1.0.1 (https://github.com/sanger-pathogens/assembly-stats) (Table S5), Merqury plots (Figure S2), and QV scores (Quality Value; Table S5) obtained with Merqury v1.3. The quality and completeness of the assemblies were further evaluated with the Benchmarking Universal Single-Copy Orthologs (BUSCO v5.5.0) pipeline (Simão et al., 2015) using the Aves database (aves_odb10; Table S6). Further details can be found in the Supplementary Methods.Table S2

Reference genome selection

We used the genome HYPOXB009684 as the reference for subsequent analyses requiring a single reference and for anchoring the pangenome to a coordinate system. We selected this genome because it belongs to the same species (*S. hypoxantha*) as the original reference genome (Genbank Assembly GCA_002167245.1) described by Campagna et al. (2017), and because it has slightly higher contiguity among the three available assemblies from this species.

Repeat masking and gene annotation

We built a custom repeat library for the Capuchino Seedeaters using RepeatModeler v2.0.1 (Smit et al., 2019). To complement this custom library, avian repeat families were retrieved from the Dfam 3.8 database (Hubley et al., 2016) and merged with the custom Capuchino repeat library. Finally, the combined repeat library was applied to soft-mask repetitive regions of each Capuchino genome assembly using RepeatMasker v4.0.7 with the RMBlast engine (Smit et al., 2015).

We predicted genes in all primary assemblies using BRAKER3 (Gabriel et al., 2024), with a custom protein database combining OrthoDB and the Zebra finch proteome (UniProt Consortium, 2019). To refine gene models, we used annotations from the chicken and the Zebra finch (GCF_000002315.6_GRCg6a and GCF_003957565.2_bTaeGut1.4) and processed them with GeMoMa v1.9 (Keilwagen et al., 2019), initially predicting around 51,380 ± 2,078 (SD) genes. We then used the GeMoMa GAF tool to merge predictions and apply filtering

(Table S7). Details on gene prediction, GeMoMa usage, and filtering are provided in the Supplementary Methods.

Synteny among assemblies and gene PAV analysis

We used the GENESPACE v1.3.1 (Lovell et al., 2022) R package in R version 4.2.3 (R Core Team, 2017) to infer and visualize synteny blocks among the thirty longest scaffolds from all the primary assemblies. We note that because our genome assemblies vary in quality, it is hard to distinguish large-scale structural changes like translocations from assembly artifacts.

We assessed gene-level variation using two complementary approaches. First, we compared gene lists from each GFF file, identifying 13,573 common genes. Second, we used Pangene v1.1 (Li et al., 2024), which aligns protein-coding exons with miniprot, to construct a pangenome graph and identified 11,760 shared genes. Unique genes per individual were compared across methods, and the overlapping genes between these methods were considered robust presence absence variation (PAV) candidates and further validated using BLAST v2.16.0 (Altschul et al., 1990). Visualization was done using ggVennDiagram v1.5.2 (Gao et al., 2024) and UpSetR v1.4.0 (Conway et al., 2017). We combined the individual lists of genes per species for the Venn diagram plots. Both methods found a similar number of unique genes per species, though overlap across methods was limited, highlighting challenges in annotation (Figures S3, S4). Only 13 genes were detected to be uniquely present in certain species by both methods but then were recovered using BLAST in the rest of the species (Table S8). Therefore, we do not have strong evidence for genes that are present or absent in certain species, although not all genes are represented in every individual annotation. Further details are provided in the Supplementary Methods.

Identifying and genotyping SVs using long read data

We performed direct SV calling from long reads as a complementary strategy to the pangenomic approach (see below). We used our PacBio HiFi reads to characterize SVs longer than 50 bp (commonly considered the lower size limit for SVs and the size range the SV callers are optimized for), employing three SV calling methods: PBSV v2.6.2 (Pacific Biosciences, 2021), Sniffles v2.2 (Sedlazeck et al., 2018), and SVIM-asm v.1.0.3 (Heller & Vingron, 2020). Depending on the strategy, either reads or assemblies were aligned to the HYPOXB009684 reference genome, and SVs were called per sample. PBSV detected the most SVs (~244K), about 2–2.5 times more than the other callers (Table S9). We merged SVs from all three tools using SURVIVOR v1.0.7 (Jeffares et al., 2017) retaining only the shared calls (within 1 kb). This conservative approach (De Coster et al., 2021) recovered ~55K SVs/sample and was robust to parameter changes. Merged SVs per individual were further combined into a final dataset of variants > 50 bp, merged within 1 kb and classified by type and size. See Supplementary Methods for alignment parameters, merging options, and classification details.

Pangenome graph construction and variant decomposition

We built the Capuchino pangenome using the Cactus Pangenome pipeline v2.8.0 (Hickey et al., 2024), starting with a GFA graph generated by minigraph v0.20 from 32 haplotypes (16 individuals). Assemblies were re-mapped and processed with Cactus to generate a multi-format pangenome graph. We calculated pangenome metrics, including core and accessory genome lengths, using Panacus v0.2.3 (Parmigiani et al., 2024). Variant types were annotated from VCF files using vcf-annotate from VCFtools v0.1.16 (Danecek et al., 2011) and classified as SNPs or MNPs if all alternate alleles matched those types, as insertions or deletions, or as complex if they mixed types or were labeled complex by vcf-annotate. While a direct comparison between SVs derived from the long-read data and the pangenome is challenging due to differences in variant representation, both approaches identified a similar number of SVs > 50 bp (~182 and ~231 thousand with long reads and the pangenome, respectively) with ~70% overlap. See Supplementary Methods for graph construction, variant annotation, and comparison of SV workflows.

Mapping short-read data to the pangenome

We used the vg toolkit v1.53.0 (Garrison et al., 2018) for pangenome-based variant calling and genotyping. Short-read data from 161 individuals were mapped to the pangenome using vg giraffe (Sirén et al., 2021). Read support was computed with vg pack (quality threshold: -Q 5), and genotypes were called using vg call (Hickey et al., 2020) to produce VCF files per individual. We note that we ran vg call without adding new paths to the pangenome graph from the short read data but rather only found support for known SVs, sometimes adding new alleles due to SNPs or indels embedded within the SVs. In highly variable regions this can result in sites with large numbers of alleles that may partially derive from methodological artifacts, even though the graph was originally built from 32 haplotypes. Detailed steps are provided in the Supplementary Methods.

Filtering and genotyping quality control

We retained only individuals with > 4X average depth of coverage, resulting in 127 individuals in this dataset (Table S10). We tested for coverage-related bias by correlating heterozygosity and coverage, finding no statistically significant association (Pearson's $r = 0.079$, $p = .38$). VCFs were indexed and merged using BCFtools v1.20 (Danecek et al., 2021), then filtered to retain sites with 4–50X depth, < 80% missing data, and a non-reference allele count ≥ 4. Additional details can be found in the Supplementary Methods.

Generation of genetic variant datasets

The resulting VCF file was divided to generate five datasets: (1) SNPs and MNPs (referred to as the SNP dataset); (2) short SVs (SVs < 50 bp) also referred to as indels; (3) SVs longer than 50 bp; and (4–5) all SNPs and SVs combined from categories 1–3, reclassified based on whether they fall within (4) or outside (5) annotated repetitive elements and TEs identified with RepeatModeler, referred to as "repeat-associated" and "non-repeat-associated," respectively. Then the SNP dataset was further filtered by a minor allele count of 4. For the GWAS analysis we generated a dataset with all SVs (merging datasets 2 and 3) and colored the resulting plots according to the variant length (greater or smaller than 50 bp).

Allele frequency and SV summary statistics

We calculated variant length as the average length of all alternative alleles at each site, defined as any allele that differs from the reference genome used in the variant calling step. Allele frequencies and allele counts were computed using VCFtools (Danecek et al., 2011) and visualized as histograms. For SVs with multiple alternative alleles, we tested different allele frequency calculations, which yielded similar distributions, and present results using the average among all alternative alleles. Relationships between SV length, frequency, and count were visualized with 2D hexbin plots in ggplot2 v3.5.1 (Wickham, 2016) after log-transforming SV lengths. Additional details are in the Supplementary Methods.

Assessing mapping quality and differences in coverage across the genome

To investigate the lower number of variants recovered from short-read mapping to the pangenome, we analyzed mapping quality. We aimed to determine if lower mapping quality of short-read data in repetitive or divergent regions explained discrepancies in the number of variants and to assess biases in variant detection with these data. Using vg surject, we generated BAM files from GAM alignments of five individuals (Table S10), then extracted per-site mapping quality with SAMtools v1.20 (Danecek et al., 2021). Regions were grouped by coverage (< 1 vs. ≥ 1 per 50 kb window), and the larger group was downsampled for comparison. A Wilcoxon rank-sum test was used to compare the distribution of mapping quality between the two groups. Additional details are in the Supplementary Methods.

Detection of large inversions using local PCA

We scanned for large inversions using local PCA with the R package lostruct (Li & Ralph, 2019) and short-read data mapped to the pangenome. The method computes PCAs across SNP windows and uses multidimensional scaling to detect outlier regions. Analyses were run on scaffolds > 1 Mb with 1,000-SNP windows, and PCAs were visualized using SNPRelate v1.36.1 (Zheng et al., 2012). Additional details are in the Supplementary Methods.

GWAS using SNPs and SVs

We performed GWAS using PLINK v2 (Chang et al., 2015) on 127 individuals from the 10 southern Capuchino species (Table S10), combining all SVs regardless of size (datasets described in the *Generation of genetic variant datasets* section). *Sporophila minuta* and *S. castaneiventris* were excluded due to their comparatively higher divergence (Campagna et al., 2012; Lijtmaer et al., 2004). SVs with > 254 alleles were removed, as PLINK cannot process sites exceeding this limit of alternative alleles (excluding 3,444 sites for the SVs dataset and 1,103 and 2,341 for the datasets including SNPs and SVs that are repeat-associated and non-repeat-associated, respectively) affecting primarily SVs > 50 bp (99.7%) (Table S11). These hypervariable variants were still included in F_{ST} scans but did not produce values exceeding 0.75, which was the threshold we designated for considering a variant as an outlier. However, our current data limit the ability to draw robust conclusions about the relevance—or lack thereof—of such regions to species differentiation. Complex variants represented as multiallelic SVs will have reduced statistical power relative to biallelic SNPs in our GWAS and F_{ST} analyses due to their low allele frequencies, limiting direct comparisons between variant types. Phenotypes were species-level mean eumelanin and pheomelanin concentrations across six plumage patches (Estalles et al., 2022), resulting in 12 GWAS. We accounted for population structure using the first 10 principal components from a PCA including all samples and applied a Bonferroni-corrected significance threshold of $p \leq 2.65 \times 10^{-9}$. Outliers were clustered into peaks (< 50 kb apart), with isolated hits reported in Table S12. We obtained similar results analyzing SNPs and SVs separately and all variant types jointly, opting to present those from the former strategy as this was our initial workflow. Further details are in the Supplementary Methods.

F_{ST} scans using SNPs and SVs

We performed F_{ST} scans with VCFtools (Danecek et al., 2011) on all five variant datasets (see *Generation of genetic variant datasets* section) per-site and in 10 kb windows (Table S11). Analyses included 95 individuals from six species and 15 pairwise comparisons (*S. cinnamomea*, *S. iberaensis*, *S. hypoxantha*, *S. hypochroma*, *S. melanogaster*, and *S. ruficollis*), excluding the Copper Seedeater (*S. bouvreuil*) due to its higher overall divergence (Campagna et al., 2013). While methods combining multiple populations could reduce the number of comparisons, our pairwise approach allows us to assess genetic differentiation associated with phenotypic divergence at focal plumage patches (e.g., throat). Outlier windows were defined as those in the top 0.1% of weighted F_{ST} and containing at least one variant with $F_{ST} > 0.75$. Consecutive outliers were merged into peaks, and additional outlier windows are listed in Table S13. We note that complex variants represented as multiallelic SVs, with generally lower allele frequencies, will produce lower F_{ST} values than biallelic SNPs, limiting direct comparisons between marker types. See Supplementary Methods for more details.

TE and SV content in outlier peaks

We compared TE and SV content per kb in outlier peaks to the genome-wide distribution using permutation tests based on 1 kb windows from the 30 longest scaffolds (excluding the terminal 50 kb). Observed values did not deviate significantly from the genome-wide distribution (outside the top/bottom 2.5%). To visualize linkage disequilibrium (LD), we used LDBlockShow v1.40 (Dong et al., 2021) to generate D'-based LD plots for peak variants with $F_{ST} \geq 0.75$ and no missing data. See Supplementary Methods for more details.

Results

High similarity among reference genome assemblies from seven Capuchino species

We recovered two genome assemblies per diploid individual: a higher-quality primary haplotype and an alternate one. Primary assemblies were longer (1.14 Gb vs. 1.09 Gb on average), more contiguous (315 vs. 1,075 scaffolds), and had a tenfold higher N50 (31 Mb vs. 2.9 Mb) and an eightfold smaller L50 (13 vs. 112 scaffolds) than alternate assemblies (Figure 1A). These metrics were consistent across species, and the primary assemblies showed high synteny

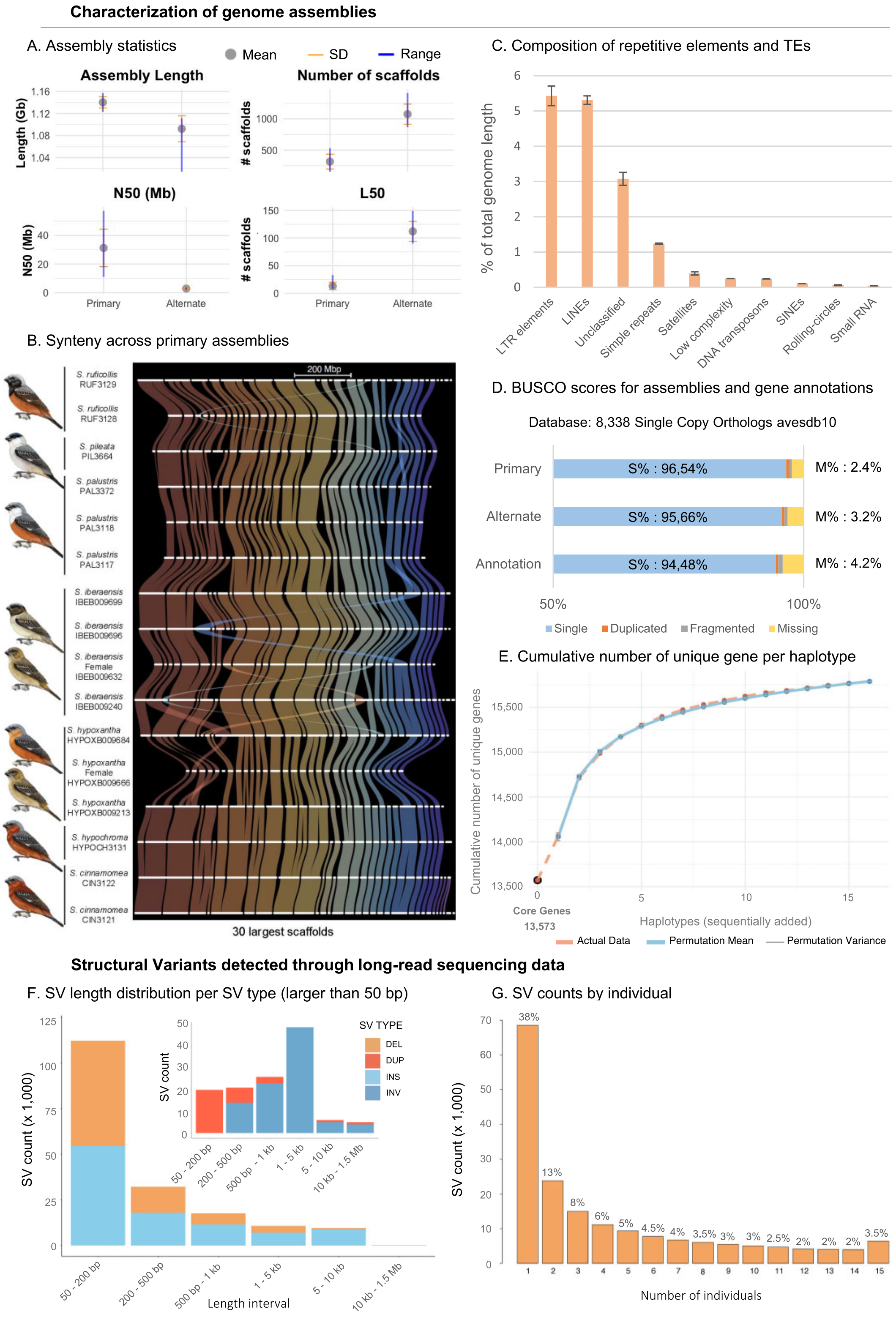

Figure 1. Genome assemblies and statistics for structural variants (SVs) called from long-read data. (A) Length and contiguity statistics for the primary and alternate assemblies. (B) Synteny representation of the 30 longest scaffolds across all primary assemblies generated using GENESPACE. (C) Composition of transposable elements (TEs) in the primary assemblies. (D) Evaluation of gene annotations through BUSCO analyses. Mean BUSCO scores for all primary and alternate assemblies, as well as the annotations for primary assemblies. (E) Cumulative unique gene count per haplotype using the annotations, showing actual data (dashed line) and the expected curve based on 1,000 permutations (solid line), with variation depicted by the bars. The circle at haplotype 0 represents the number of core genes shared by all haplotypes. (F and G) SV statistics from long-read sequencing data. (F) SV length distribution per SV type, based on results from Sniffles2 supported by three SV callers (Sniffles2, PBSV, and SVIM-asm). (G) SV counts per individual, showing the number of structural variants present in 1–15 individuals (the reference genome HYPOXB009684 is not included), with most SVs found in only a single individual.

among Capuchinos (Figure 1B). Repetitive elements and TEs accounted for ~16% of each genome, ~11% of which were retrotransposons, with similar composition across species (Table S14, Figure 1C). Gene content showed an average of 14,666 ± 46 (SD) genes per assembly, with little variation across individuals and species (Table S7). Gene completeness was high across all assemblies and annotations, yet slightly higher in primary assemblies (96.5% single-copy orthologs; Tables S6, S15; Figure 1D). A total of 13,573 core genes were shared across all primary assemblies, and the cumulative gene discovery from adding genomes sequentially to our analysis plateaued at 15,788 unique genes (Figure 1E). Initial analyses of gene PAV suggested there are several genes present uniquely in each species (e.g., 25 in *S. pileata* and 100 in *S. iberaensis*; Figures S3, S4). However, BLAST searches recovered fragments or complete gene sequences for these putatively missing genes in the other assemblies, indicating that their apparent absence likely reflects assembly or annotation limitations (Table S8).

The landscape of structural variation and the Capuchino Seedeater pangenome

We identified an average of ~55 K structural variants (SVs > 50 bp) per individual from long-read data, supported by three SV callers, totaling 182,213 SVs across all individuals. Insertions (54.6%) and deletions (45.3%) dominated, while inversions and duplications were comparatively rare (≤ 0.05%) (Figure 1F). Most SVs were small; however, inversions were more common in the 200 bp–5 kb range and absent from the smallest size class (Figure 1F). About 38% of SVs were private to single individuals, while only 3.5% were shared by all (Figure 1G). For example, just one of 95 inversions was shared across all individuals. Species with more representatives contributed more unique SVs, though these were often found in single individuals within those species (Figure S5). To detect shared SV patterns, we built a pangenome using all assemblies.

The pangenome spanned 1.5 Gb, measured in the total number of unique base pairs recovered from all individuals (i.e., all alternative paths), with 66% forming the core genome and 48% of nodes shared across all individuals (Figure 2A, S6). The pangenome contained 59.2 million variants, with 7.6 times more SNPs than SVs (Figure 2B). Although less frequent, SVs covered 184.4 Mb compared to the 52.3 Mb spanned by SNPs/MNPs. Both types of genetic variants are represented (by the reference or alternative alleles) in a similar number of samples, with an average of 15.7 samples for SNPs and 15.3 for SVs. Nearly half of the variants (45%) were rare, appearing only once as the alternative allele among the 16 individuals. Most SVs (96.5%) were indels (< 50 bp) (Figure 2B). To understand the patterns of differentiation of these markers among the different Capuchino species, we used short-read data to genotype 127 individuals across 10 species using the pangenome as a reference, enabling GWAS and F_{ST} analyses using both SNPs and SVs.

From this larger dataset, we recovered ~35.5 million variants, which after filtering was roughly half of the original pangenome set (31.8M). Regardless, the proportions between SNPs and SVs remained similar, with 8 times more SNPs than SVs (28.3M vs. 3.5M), and only 2.7% of variants > 50 bp (Figure 2C). We identified ~5.9M variants that were absent from the pangenome, likely due to increased sampling. However, ~29.6M pangenome variants were lost in the short-read dataset due to mapping limitations in complex, repetitive regions. These regions showed significantly lower mean mapping quality (3.86 ± 6.53 vs. 55.3 ± 7.64; Wilcoxon $W = 127, p < 2.2e{-}16$) and were often near scaffold ends enriched for TEs and showing lower coverage even in the pangenome dataset, suggesting they are inherently hard to resolve across sequencing platforms (Figure S7). Despite their lower prevalence, SVs still spanned a greater portion of the genome than SNPs (~31.8 Mb vs. 28.3 Mb), with similar mean coverage (9.5X SVs, 9.3X SNPs). In this dataset, the difference in bases covered by SVs and SNPs is less pronounced than in the pangenome alone, likely due to the loss of long SVs.

We filtered SVs to retain only those present in ≥ 80% of individuals (discarding variants in fewer than 102 individuals) for GWAS and F_{ST} scans. Both the SNP and SV alternative allele frequencies were characterized by a higher abundance of loci with low-frequency alleles, with a gradual decline in abundance of loci toward intermediate and high frequencies of alternative alleles (Figures 2D, 2E, S8A). The lowest allele frequency bin was underrepresented due to missing data and allele count filters (Figure S8B,C). SV density decreased with length, and longer SVs had lower allele frequencies and counts (Figure 2E, 2F). Most low-frequency variants were short SVs (Figure 2F). The alternative allele frequency distribution shows that low-frequency variants are prevalent across all SV lengths, particularly at shorter lengths (Figure 2E). Allele count distributions across variant types in both the pangenome and short-read datasets showed that SNPs and indels were predominantly biallelic, whereas SVs were more frequently multiallelic, with 75% of SV sites having more than two alleles (Figure S9). The same pattern emerged within species: while SNPs and indels remained largely biallelic, SVs showed a greater proportion of multiallelic sites, although biallelic SVs still represented the single most common class (Figure S10).

Outlier genomic regions associated with plumage coloration

We conducted GWAS across six plumage patches using eumelanin and pheomelanin concentrations as phenotypes and partitioned the dataset by variant type (e.g., SNPs, SVs). We identified seven strong outlier peaks repeatedly associated with pigment concentration across body parts (Figure 3, Figures S11–S16). These peaks were consistently observed in different combinations depending on the plumage patch and pigment type. Five peaks were shared across SNP and SV datasets and included melanogenesis genes (*OCA2/HERC2*, *ASIP*, *TYRP1*, *SLC45A2*) and genes involved in amino acid metabolism (*AHCY*, *GPT2*). The remaining two peaks were exclusive to the SNP dataset and lacked annotated genes (Table 1, Table S16). We did not observe strong associations with eumelanin in the head and pheomelanin in the throat (Figures S13, S16). Most peaks were identified through SNPs, with only a few associated with indels (SVs < 50 bp). A single larger SV—a 55 bp deletion—was found within a peak associated with pheomelanin concentration in the belly plumage patch (Figure 3). Repeat-associated genetic variants recovered fewer peaks compared to non-repeat-associated variants (Figure 3, Figures S11–S16). The GWAS results also

Table 1. Genome-wide association studies (GWAS) and F_{ST} outlier peaks.

	Peak 1*	Peak 2	Peak 3*	Peak 4*	Peak 5*	Peak 6	Peak 7	Peak 8*	Peak 9	Peak 10
Scaffold	**10**	14	**21**	7	9	41	**7**	**3**	7	7
Chromosome	10	11	20	Z	6	Z	Z	4	Z	Z
Start coordinates (bp)	6,190,001	13,950,001	13,545,422	25,260,001	15,010,001	108,180	5,260,234	9,950,001	5,560,001	5,730,001
End coordinates (bp)	6,250,000	13,990,000	13,700,000	25,320,000	15,030,000	112,141	5,272,125	9,990,000	5,580,000	5,750,000
Peak length (bp)	59,999	39,999	154,578	59,999	19,999	3,961	11,891	39,999	19,999	19,999
GWAS/F_{ST}	GWAS/F_{ST}	GWAS/F_{ST}	GWAS/F_{ST}	GWAS/F_{ST}	GWAS/F_{ST}	GWAS	GWAS	F_{ST}	F_{ST}	F_{ST}
Gene1	*HERC2*	*GPT2*	*ASIP*	*TYRP1*			*SLC45A2*	*ALB*		*SPEF2*
Gene2	*OCA2*#	*CDCA9*	*AHCY*							*LOC112530520*
SNPs#	1,375	1,161	2,683	1,402	335	104	133	269	180	269
SVs#	181	149	342	161	49	15	20	31	19	32
Repeat-associated SNPs#	147	19	984	263	23	52	7	36	37	12
Repeat-associated SVs#	19	4	153	40	3	10	1	12	4	6
Repeat# per peak	25	8	122	39	8	4	5	31	12	8
Length TEs	5,149	602	54,332	10,665	973	1,543	611	4,665	3,869	610
Length repeat-assoc. SVs	217.55	6	3,594.4	196.6	17	25	1	79	415	78.8
SNPs coding	47	23	12	12	0	0	9	8	0	7
SNPs coding $F_{ST} > 0.75$	1	1	0	2	0	0	0	4	0	0
SVs coding	0	1 (1 bp INS CDCA9)	0	0	0	0	0	0	0	0
#SVs > 50 bp	3	1	4	1	0	0	0	0	1	0
#SVs > 50 bp $F_{ST} > 0.75$	1 (DEL 55 bp)	0	0	0	0	0	0	0	0	0
Max length SVs peak (bp)	146	64	7398§	97	15	10	9	43	389	42
Mean length SVs peak (bp)	11.9	10.6	126.7	9.1	2.8	2.8	2.7	4.3	24.0	8.9
Total length SVs (bp)	951.2	507.2	4839.9	571.5	139.5	41.0	57.0	160.5	502.0	140.6

Details for the 10 main outlier regions identified using both strategies, including peak coordinates (scaffold, start, end, and length), the chromosomal location according to the Zebra finch genome, whether the peak was identified as a GWAS and/or F_{ST} outlier, the number and type of variants [single-nucleotide polymorphisms (SNPs) and structural variants (SVs)] and their overlap with transposable elements (TEs) and coding regions, and the genes within each peak. Additionally, the table provides the maximum, mean, and total length of the SVs within each peak. Peaks detected with both SNPs and SVs are highlighted in bold, while those detected only with SNPs are in regular font. The peaks marked with an asterisk (*) are those that were also recovered by the repeat-associated variants. (#) In peak 1, we refer to the *OCA2/HERC2* gene pair, which is involved in melanogenesis, yet the specific gene in the peak is *HERC2*. In peak 3, the total length covered by SVs is shorter than the maximum length (§), because the longest variant is not fully contained within the peak—3,544 bp extend beyond its boundaries.

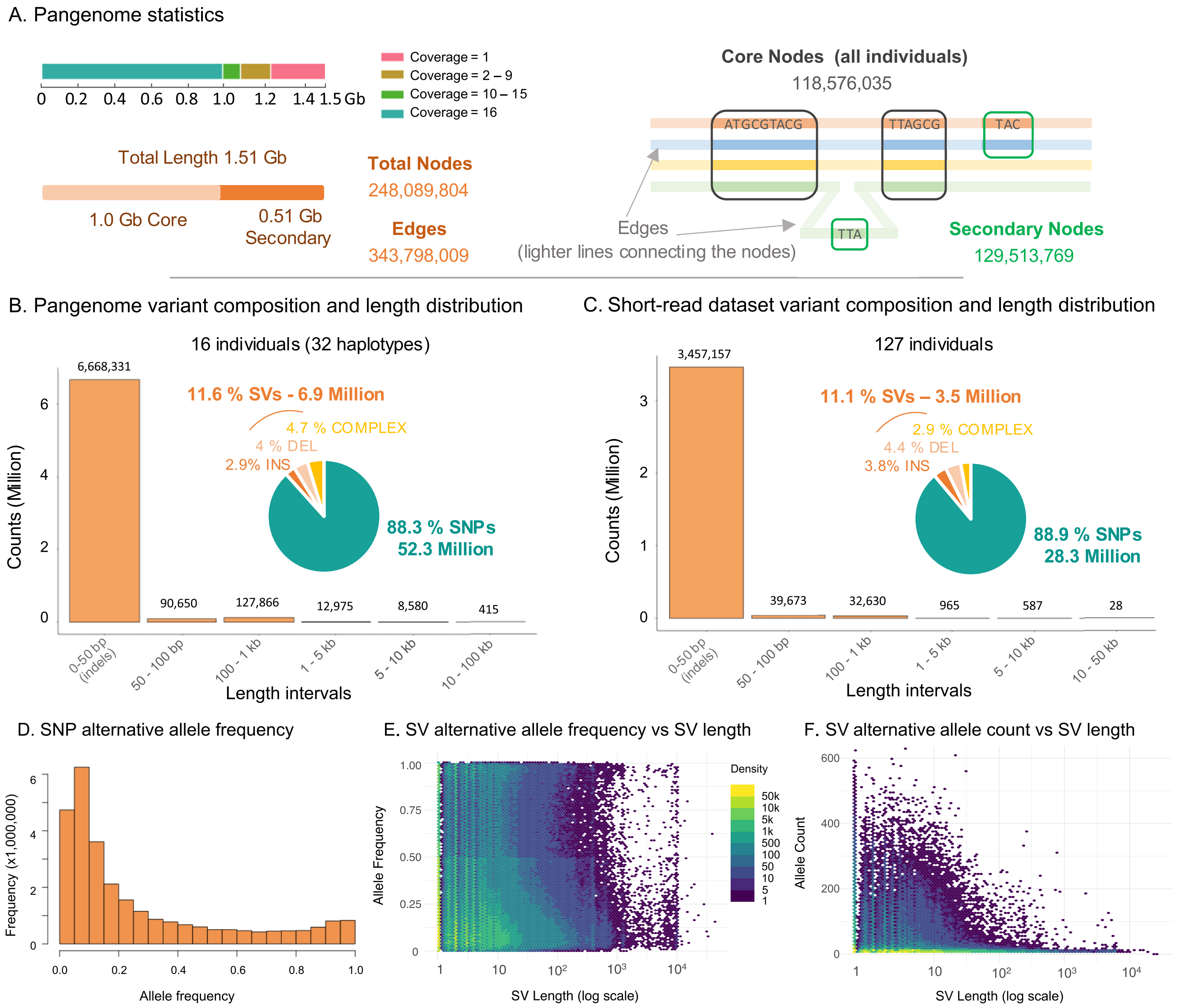

Figure 2. The Capuchino pangenome and variant genotyping using short-read data. (A) Overview of statistics describing the pangenome, including the shared sequence length (Gb) across varying numbers of haplotypes as computed with Panacus, total length, number of nodes and edges, and the lengths and counts of core and secondary genome nodes. (B) Length distribution and composition of structural variants (SVs) in the pangenome, highlighting total numbers and percentages of single-nucleotide polymorphism (SNPs) and SVs. SVs are further categorized into insertions, deletions, and complex rearrangements. (C) Length distribution and composition of variants genotyped for 127 individuals from short-read data mapped to the pangenome, showing total numbers and percentages of SNPs and different types of SVs. (D) Alternative allele frequency distribution for SNPs identified from short-read data mapped to the pangenome (see Figure S8A for the equivalent plot for SVs). (E) Relationship between SV length (in log scale) and alternative allele frequency, and (F) alternative allele count. In both E and F, the density of data points is represented by a color gradient, with darker and lighter shading indicating lower and higher densities, respectively.

identified 217 isolated SNPs and SVs (representing 24.5% of all significant GWAS hits) not included in the more prominent outlier peaks (Table S12).

Outlier genomic regions associated with species differences

Similar to the GWAS, F_{ST} scans revealed eight recurrent differentiation peaks across species comparisons (Figure 4, Figures S17–S30). Four peaks were shared between SNP and SV datasets: three included melanogenesis genes (as in the GWAS, except for the absence of the peak containing *SLC45A2*), and one contained the gene *ALB* (Table 1, S16). The remaining four peaks were SNP-specific, and two contained annotated genes (*SPEF2*, *GPT2*, and *CDCA9*). As in the GWAS, we observed only a few strong outlier peaks per pairwise comparison (Figure 4, Figures S17–S30). Two comparisons (the Rufous-rumped Seedeater, *Sporophila hypochroma*, vs. *S. hypoxantha* and *S. hypochroma* vs. the Chestnut Seedeater, *Sporophila cinnamomea*) lacked outlier windows or peaks, suggesting subtler differentiation patterns (Figures S18, S19). As with GWAS, SNPs and non-repeat-associated variants accounted for most peaks (Table S17). The 55 bp deletion on scaffold 10 detected in the GWAS was the only SV > 50 bp found within a peak, present in the *S. hypoxantha* vs. *S. iberaensis* comparison (Figure 4). Additionally, we identified 85 isolated F_{ST} outlier windows outside major peaks

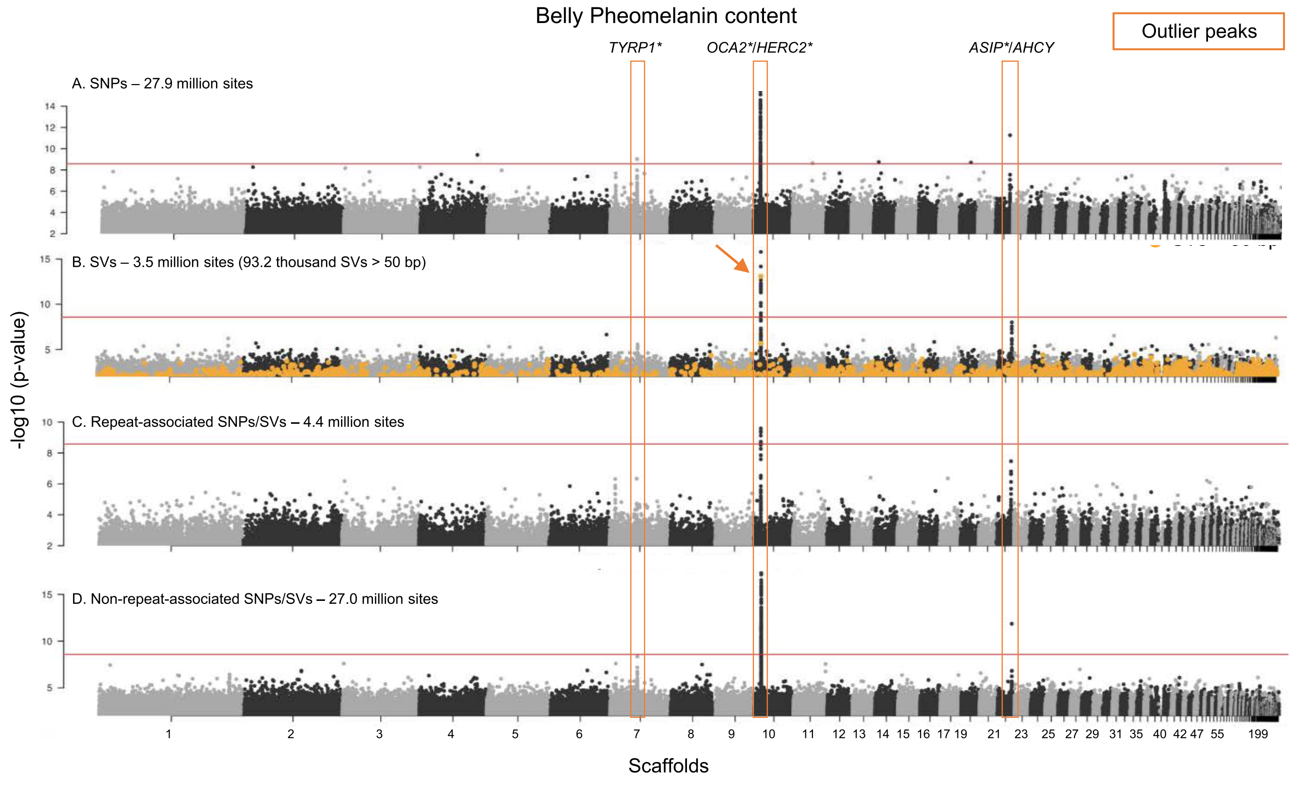

Figure 3. Genome-wide association study for the pheomelanin content in the belly plumage patch. The analysis includes four datasets, displayed from top to bottom: (A) single-nucleotide polymorphism (SNPs), (B) structural variants including indels (< 50 bp) and long variants (> 50 bp in orange), (C) repeat-associated variants [SNPs and structural variants (SVs)], and (D) non-repeat-associated variants (SNPs and SVs). The *y*-axis represents the −log10(*p*-value) obtained in the genome-wide association studies (GWAS), and the horizontal line is the Bonferroni-corrected threshold of statistical significance (for all comparisons and variants), corresponding to a *p*-value of 2.65 ×10^{-9}. Scaffolds are ordered by decreasing size and represented in alternating shading. Peaks are highlighted with rectangles, and known genes are labeled above the peaks. The genes marked with an asterisk (*) belong to the melanogenesis pathway. The peaks associated with pheomelanin content in the belly include *TRYP1* on scaffold 7, which is detected only by the SNPs dataset; *OCA2/HERC2* on scaffold 10, which is detected by all datasets, including the long SVs; and the peak containing the *ASIP* and *AHCY* genes on scaffold 21, which is detected by the SNPs and the non-repeat-associated variants datasets. The single large SV with a statistically significant association is marked with an arrow.

(12.9% of all outlier windows; Table S13), showing that there are areas of the genome with more subtle patterns of differentiation.

Multiple sources of genetic variation in Capuchino Seedeaters

Across both the GWAS and F_{ST} outlier strategies, we identified 10 outlier peaks averaging 43 kb in length (range: ∼4–155 kb; Table 1, S16). Except for peak 10 (Table 1), these were previously reported using SNPs and a single reference genome (Campagna et al., 2017; Estalles et al., 2022; Turbek et al., 2021). Five peaks were shared across GWAS and F_{ST} scans, representing our strongest candidates. Of these, four (peaks 1–4) were detected in both SNP and SV (< 50 bp) datasets and included the melanogenesis genes *OCA2/HERC2*, *ASIP*, and *TYRP1* (Table 1, S16), while peak 5 was SNP-specific and lacked annotated genes. Three peaks did not contain genes but may harbor regulatory loci influencing the expression of nearby genes (Table S16). Outlier detection was largely driven by SNPs and indels outside of repetitive regions or TEs. SVs > 50 bp were mostly absent from peaks, except for the 55 bp deletion on scaffold 10. We also found eight additional SVs with $F_{ST} > 0.75$ outside the peaks (Table S18), including two insertions overlapping introns of *SDHB* (207 bp, scaffold 34) and *TPM4* (73 bp, scaffold 38), both genes linked to reproductive traits in chickens (Kramer et al., 2025; Zhang et al., 2017). The remaining SVs were located 280 bp to ∼90 kb from the nearest gene (Table S18). Lastly, a windowed PCA analysis detected a possible large inversion on the Z chromosome (Figure S31), but it was not associated with species differences or outlier peaks.

Genetic variant and TE composition within peaks

Within outlier peaks, SNPs were more frequent and covered a greater proportion of bases than SVs, with mostly non-coding variants (Figure S32A, Table 1). Repetitive elements and TEs accounted for 1.4% to 37% of peak regions, with no consistent pattern in their overlap with variant types (Figure S32A, Table 1). TE composition within peaks mirrored the genome-wide distribution (Figure 1C), dominated by retrotransposons, particularly LINEs and LTRs (Figure S32B). While SV and TE levels in peaks were not extreme relative to the rest of the genome, TE composition varied more across peaks than SV content (Figures S32C, S32D). Notably, half of the peaks were located on the Z chromosome (Table 1), consistent with patterns observed previously in this (Estalles et al., 2022) and other systems (Bourgeois et al., 2020).

Only 1.5% (118 SNPs) of SNPs (irrespective of their level of differentiation) were coding variants, suggesting a minor

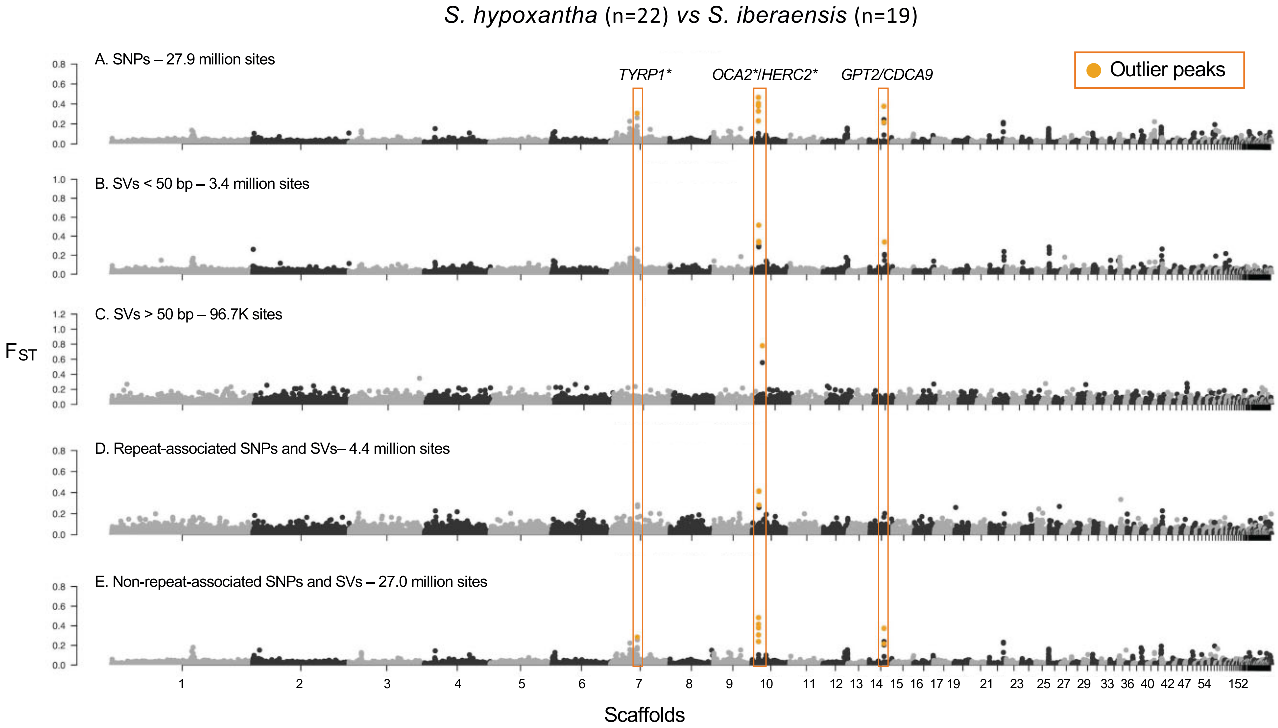

Figure 4. F_{ST} scan in 10 kb windows for the comparison between *S. hypoxantha* and *S. iberaensis*. The analysis includes five datasets, displayed from top to bottom: (A) single-nucleotide polymorphisms (SNPs), (B) structural variants < 50 bp, (C) structural variants > 50 bp, (D) repeat-associated variants [SNPs and structural variants (SVs)], and (E) non-repeat-associated variants (SNPs and SVs). Orange dots mark outlier windows within identified differentiation peaks, defined as the top 0.1% of the F_{ST} distribution and containing at least one variant with $F_{ST} > 0.75$. Scaffolds are ordered by decreasing size and represented in alternating shading. Peaks are marked with rectangles, and the known genes are labeled on top of the peaks. The genes marked with an asterisk (*) belong to the melanogenesis pathway. There are three peaks in this comparison: one containing *TYRP1* on scaffold 7, detected only by the SNP dataset; another containing *OCA2/HERC2* on scaffold 10, detected by all datasets; and the third containing *GPT2* and *CDCA9* on scaffold 14, detected by all datasets except the long SVs and the repeat-associated variants.

role for coding differences overall. However, among the coding variants, eight SNPs with F_{ST} values above 0.75 (four of which were found in multiple comparisons) likely play an important role in driving color/species differentiation (Table S19). The Black-bellied Seedeater (*Sporophila melanogaster*) is involved in 9 out of the 13 comparisons, and the genes affected were *TYRP1, ALB, GPT2*, and *HERC2* (Table S19). We detected a single coding indel (a 1 bp insertion in *CDCA9*), but it was not highly differentiated among species (Table 1). SVs accounted for 11% of variants within peaks, with only 10 longer than 50 bp, of which 1 (the 55 bp non-coding deletion on scaffold 10 located 12.8 kb from the *HERC2* gene) had an F_{ST} value above 0.75 in the comparison between *S. hypoxantha* and *S. iberaensis* (Table 1, Figure S33A and S33B). In this species pair, most individuals from *S. iberaensis* are homozygous for the deletion (1/1), with some cases of heterozygosity, whereas *S. hypoxantha* individuals are predominantly homozygous without the deletion (0/0), although there are two 1/1 individuals (Figure S33C). This variant is not species-specific, as *S. palustris*, the Black-and-tawny Seedeater (*Sporophila nigrorufa*), and *S. hypochroma* exhibit genotypes similar to *S. hypoxantha* at this site, while the remaining five species share the deletion with *S. iberaensis*. The deletion is in high LD with SNPs and indels within the peak, showing how different types of variants share their genomic signal (Figure 5). We do, however, observe species-specific patterns when assessing variation across all peaks combined. The genotypes of variants (SNPs, indels, and the 55 bp deletion) with $F_{ST} > 0.75$ within the peaks show clear genetic differentiation among species, with distinct clusters reflecting species-specific allele combinations in these highly differentiated regions (Figure 5). Among these regions, the two comparisons without significant peaks—*S. cinnamomea* vs. *S. hypochroma* and *S. hypoxantha* vs. *S. hypochroma*—emerge as the least differentiated species overall. However, some comparatively more subtle differences are still present in certain regions, such as peaks 1, 3, and 8, where *S. hypoxantha* and *S. cinnamomea* predominantly exhibit (0/0) and (1/1) genotypes, respectively. Additionally, each comparison includes other more subtly differentiated regions that do not meet our peak thresholds. For example, the peak on scaffold 7, which contains the *SLC45A2* gene, includes 19 and 23 variants with $F_{ST} > 0.5$ in the *S. hypoxantha* vs. *S. hypochroma* and *S. hypochroma* vs. *S. cinnamomea* comparisons, respectively.

Discussion

Leveraging a pangenome to study rapid speciation

Our study provides the most comprehensive view to date into the genetic changes associated with rapid speciation in Capuchino Seedeaters, using a pangenome built from 32 de novo genome assemblies. By integrating this resource with short-read whole-genome sequencing data from ten species (including three previously underrepresented species: *S. bouvreuil*, *S. cinnamomea*, and *S. hypochroma*), we refined the

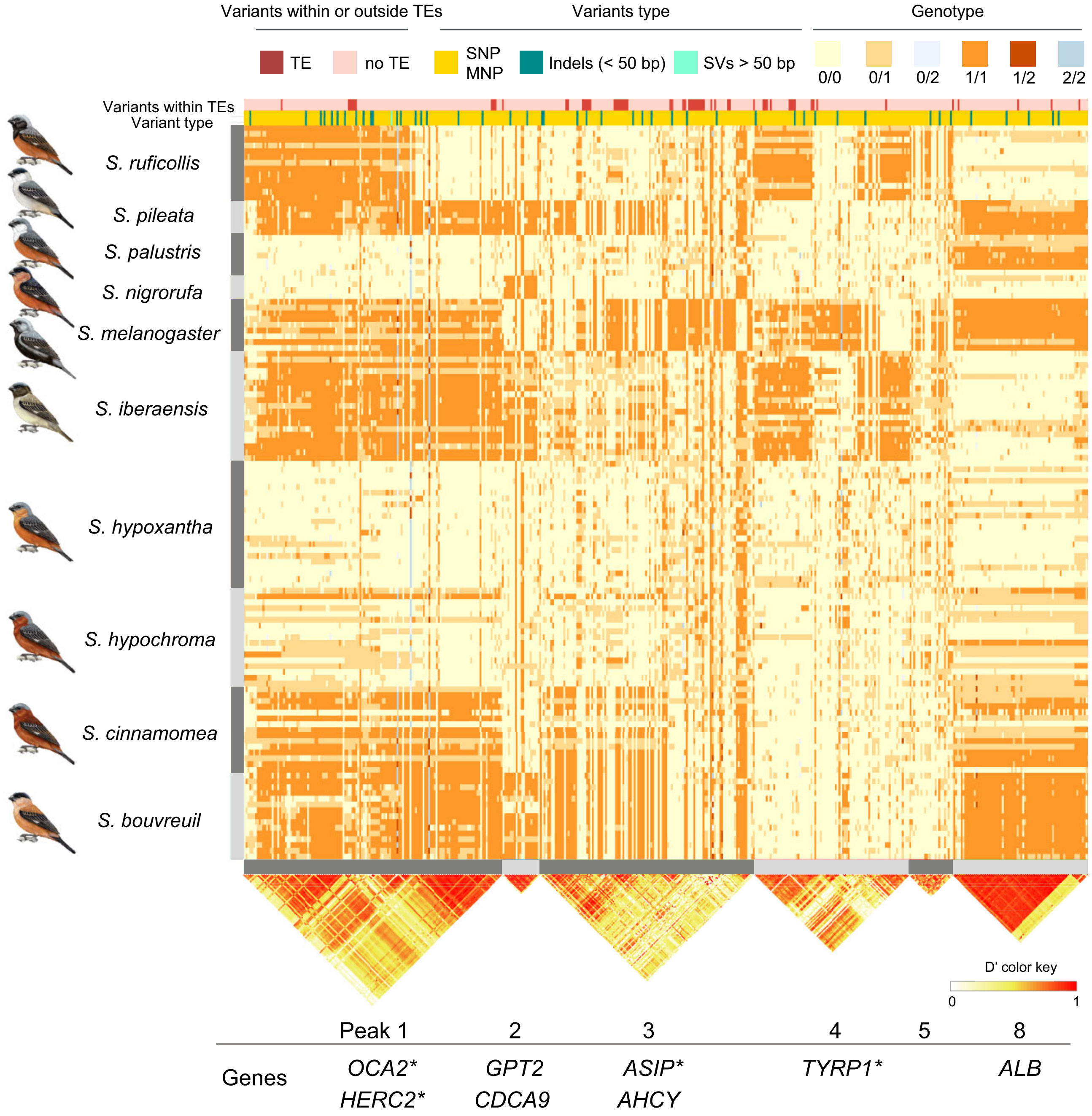

Figure 5. Capuchinos show species-specific patterns of genetic variation when comparing across all divergence peaks. Genotypes for 456 variants with no missing data (for visual simplicity) within peaks 1–5 and 8 with F_{ST} values > 0.75 across populations, including the 55 bp deletion (marked in the "Variant type" track, see Figure S33C for details), obtained using short-read sequencing data mapped to the pangenome. Peaks 6 and 7 were excluded as they were only detected in the genome-wide association studies (GWAS) and do not have variants with F_{ST} > 0.75, while peaks 9 and 10 were excluded due to having very few variants, which unnecessarily complicate the plot. The peak IDs correspond to those in Table 1. Rows represent individuals from different populations, while columns correspond to genomic sites, grouped by peak. Variants overlapping repetitive elements and transposable elements (TEs) are indicated at the top of the plot, with darker shading denoting repeat-associated sites and lighter shading indicating non-repeat associated sites. Variant types are also indicated, with single-nucleotide polymorphisms (SNPs) and multi-nucleotide polymorphisms (MNPs) in yellow, indels (< 50 bp) in teal, and the 55 bp deletion (light blue). The genotypes are color-coded: the homozygous reference (0/0) is represented in yellow, the heterozygous genotypes (0/1, 0/2, 1/2) are represented in light orange, light blue, and dark red, respectively; and the homozygous alternate genotypes (1/1, 2/2) are represented in orange and blue, respectively. The presence of multiple alternate genotypes arises from indels and MNPs, where more than one alternative allele is present. Below each peak, the linkage disequilibrium (LD) pattern is displayed based on the D' method, color-coded from yellow to red (0 to 1), with red indicating strong LD. Overall, most peaks exhibit high LD, suggesting that the variants within them are inherited in blocks. However, some peaks (e.g., 1, 3, and 4) show distinct LD blocks, indicating potential recombinant haplotypes. Additionally, the genes within each peak are listed, with those marked by an asterisk (*) indicating genes that are part of the melanogenesis pathway.

characterization of genetic variants driving species differentiation, expanding the analysis beyond SNPs to include small insertions/deletions and other SVs omitted in past research. Previous studies on Capuchino Seedeaters have proposed that the reshuffling of small regulatory alleles drives rapid plumage evolution, which coupled with male song differences, promotes prezygotic isolation and speciation (Turbek et al., 2021). These insights were based on SNP markers and a single (*S. hypoxantha*) reference genome. An alternative hypothesis proposes the existence of species-specific SVs near these outlier genes. These variants, potentially linked to TE activity, may drive differentiation but could have been overlooked due to methodological limitations. In this study, we leverage our pangenome to distinguish these hypotheses.

Improvements over previous genomic studies

Genome assemblies and gene content

The high-quality genome assemblies improved on previous studies in various ways. Combining several genomes produced a higher-quality reference capturing greater sequence diversity (approximately 1.5 Gb vs. 1.17 Gb, Campagna et al., 2017), including alternate haplotypes and structurally variable regions that are not present in every individual. This approach also captured a larger number of genes (15,788 vs. 14,667, Campagna et al., 2017). Assemblies were broadly similar across species, but distinguishing between genes that are truly missing from an individual from those that are absent due to limitations in the genome assembly process remains challenging. We therefore recommend using various approaches to assess gene presence/absence. Overall, our pangenome recovered genomic regions, genes, and variant types missed in earlier work, offering a methodological advance that is likely to also benefit other systems undergoing similar comparative analyses.

Relative prevalence of SNPs and SVs

The pangenome reveals SNPs are nearly eight times more common than SVs, with most SVs being < 50 bp (indels), as seen in other systems (e.g., Lecomte et al., 2024). Insertions and deletions dominated, while inversions and duplications were rare and typically found at low frequencies in few individuals. Two other studies (on the House Finch and *Aphelocoma* jays) reported 3–4 times more SVs > 50 bp than in our data (Edwards et al., 2025; Fang & Edwards, 2024), likely due to the more recent diversification and higher genomic similarity among Capuchinos (Campagna et al., 2017; Turbek et al., 2021). For example, a large inversion shared by three Haemorhous species dates to ∼10 million years ago (Fang & Edwards, 2024), far older than Capuchino divergence (Campagna et al., 2013).

Transposable elements

Improved assembly of complex/repetitive regions enabled the annotation of TEs, which showed limited contribution to species differentiation. About 16% of the genome consists of repeats, mainly retrotransposons. This TE content is slightly higher than previously reported for most birds [4.1–9.8% (Kapusta & Suh, 2017)], except for woodpeckers [which reach up to 31% (Manthey et al., 2018)]. However, TE detection nearly doubled in sparrows using long-read vs. short-read data (Benham et al., 2024), and this may also be the reason why we find higher TE content.

Limitations in recovering the landscape of structural variation

Combining the pangenome with short-read data failed to recover ∼29.6 million variants present in the pangenome. This reduction is probably due to the inherent limitations of short-read data in properly mapping to regions containing larger SVs and repetitive sequences (Mahmoud et al., 2019), resulting in variant loss at scaffold ends and other repetitive regions. Although the pangenome now includes these regions, long-read genotyping of more individuals could help resolve possible species differences in these challenging areas. However, these complex genomic regions will remain difficult to work with, as graph-based genotyping can generate multiallelic calls that may reflect real biological variation but that can also arise from technical artifacts, potentially erroneously inflating the number of alleles. Although many multiallelic loci are included in our F_{ST} and GWAS analyses, capturing the full complexity of these variants in a VCF file remains challenging (Edwards et al., 2025). Breaking such variants down into many independent low-frequency alleles could mask meaningful relationships among alleles within a locus and reduce statistical power relative to biallelic SNPs in outlier scans. These issues are likely more prominent in highly variable regions and complicate the direct comparisons of the relative relevance of SNPs and SVs in shaping phenotypes. Such regions and variants may be better explored in the future using emerging multiallelic-aware methods (Saitou et al., 2022). While our pangenome represents an improvement over previous studies, some key differences relevant to species divergence may still be missed, and these technical and biological constraints should be considered when applying these methods in other systems.

Species differentiation linked to possible regulatory changes in pigmentation loci

Despite variation in sample sizes and species, GWAS for pigment concentrations and F_{ST} scans consistently identified the strongest outlier regions containing melanogenesis genes (*HERC2/OCA2*, *ASIP*, *TYRP1*, and *SLC45A2*), confirming previous findings (Campagna et al., 2017; Estalles et al., 2022; Turbek et al., 2021). This study also confirmed patterns seen in the Capuchino radiation, such as the predominance of non-coding differences and enrichment of outlier peaks on the Z chromosome. As seen before (Turbek et al., 2021), species exhibit unique genotype combinations across key outlier peaks, though other regions with subtler differentiation exist genome-wide. Moreover, differentiation levels vary across species, suggesting there are differences in gene flow and/or the number and extent of genomic regions driving phenotypic traits. Differentiated regions mainly feature SNPs and small SVs, with large SVs largely absent. Outlier regions did not show significantly elevated SV or TE content compared to the rest of the genome. Variants within outlier peaks (SNPs, indels) are generally in high LD, likely acting together to influence phenotype. Integrating chromatin accessibility and interaction data (e.g., ATAC-seq, Hi-C) with variant analysis will help uncover regulatory networks, especially in non-coding regions where over 90% of enhancers—key gene expression regulators—are located (Liang et al., 2024). Importantly, we find that differentiation in outlier regions is largely driven by SNPs and small indels rather than large species-specific SVs, and we conclude that the reshuffling of regulatory alleles remains the most likely mechanism driving the rapid speciation of the Capuchinos.

Broader implications of the pangenomic approach

While SVs do not seem to have strongly shaped Capuchino Seedeater evolution, our results highlight the power of the pangenome framework for studying phenotypic evolution and speciation. SVs may play a more prominent role in systems with longer divergence times, but their evolutionary relevance may currently be overlooked due to methodological limitations. Pangenomes are beginning to emerge in a wide range of organisms beyond plants and bacteria, as advances in sequencing technologies make these comprehensive genomic studies more feasible. However, these approaches also bring technical challenges (discussed above), which will require the development of new methods for future analyses. Moreover, our study points to the importance of combining pangenomes with population-level long-read sequencing to overcome the limitations of mapping short-read data, which can fail to detect structural variation captured in the pangenome. These developments provide the opportunity to integrate all forms of genetic variation into the search for the genetic basis of phenotypes in non-model organisms (e.g., Chen et al., 2023; Edwards et al., 2025; Fang & Edwards, 2024; Wang et al., 2024; Wei et al., 2024), and promise to uncover previously hidden genetic contributions to phenotypic traits and adaptive evolution.

Supplementary material

Supplementary material is available online at *Evolution*.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Genomic data have been archived in GenBank (BioProject ID PRJNA382416). The genome assemblies, generated from PacBio HiFi data, are available under BioProjects PRJNA1223491–PRJNA1223508 and PRJNA1223510–PRJNA1223523. All the accession numbers are provided in the Supplementary Material.

Author contributions

Conceptualization: L.C., M.R.; Methodology: M.R., L.C.; Investigation: M.R., L.C., S.K., J.R.R.R., M.R., M.B., J.F.C., A.S.D.G., C.K.; Visualization: M.R.; Writing—original draft: M.R., L.C.; Writing—review & editing: M.R., L.C.

Funding

This study was funded by the National Science Foundation—NSF DEB-2232929 to L.C. and by a Genomics Scholar fellowship from the Center for Vertebrate Genomics (Cornell University) to L.C. and M.R. Fieldwork was funded by BirdLife International—Small Grants Program for Grassland Conservation (2022, 2023) and ANPCyT FONCYT (PICT 2019-4057, 2018-3407).

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

We thank the members of the Fuller Evolutionary Biology laboratory group (particularly B. Butcher and I. Lovette) for lab and computational support. We thank the volunteers for essential help with fieldwork. We thank J. Grenier and the BRC Genomics Facility (RRID:SCR_021727). We also thank the government agencies of Brazil (ICMBio, MMA, and SISBIO) and Argentina (SAyDS, Dirección de Recursos Naturales Corrientes, Dirección de Parques y Reservas de la Provincia de Corrientes, Administración de Parques Nacionales) for providing the necessary research authorizations and collection permits that facilitated our work (63556-12, Brazil, and IF-2018-53280705-APN-DRNEA#APNAC, IF-2022-133544133-APN-DRNEA#APNAC, Argentina). Samples for this study were exported through permits CE-2023-49830382-APN-DNBI#MAD, and CE-2024-83310820-APN-SSAM#JGM.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Benham, P. M., Cicero, C., Escalona, M., Beraut, E., Fairbairn, C., Marimuthu, M. P. A., Nguyen, O., Sahasrabudhe, R., King, B. L., Thomas, W. K., Kovach, A. I., Nachman, M. W., & Bowie, R. C. K. (2024). Remarkably high repeat content in the genomes of sparrows: The importance of genome assembly completeness for transposable element discovery. *Genome Biology and Evolution*, *16*(4), evae067. https://doi.org/10.1093/gbe/evae067

Bourgeois, Y. X. C., Bertrand, J. A. M., Delahaie, B., Holota, H., Thébaud, C., & Milá, B. (2020). Differential divergence in autosomes and sex chromosomes is associated with intra-island diversification at a very small spatial scale in a songbird lineage. *Molecular Ecology*, *29*(6), 1137–1153. https://doi.org/10.1111/mec.15396

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*, 1–12. https://doi.org/10.1186/s13059-018-1577-z

Campagna, L., Benites, P., Lougheed, S. C., Lijtmaer, D. A., Di Giacomo, A. S., Eaton, M. D., & Tubaro, P. L. (2012). Rapid phenotypic evolution during incipient speciation in a continental avian radiation. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1734), 1847–1856. https://doi.org/10.1098/rspb.2011.2170

Campagna, L., Repenning, M., Silveira, L. F., Fontana, C. S., Tubaro, P. L., & Lovette, I. J. (2017). Repeated divergent selection on pigmentation genes in a rapid finch radiation. *Science Advances*, *3*(5), e1602404. https://doi.org/10.1126/sciadv.1602404

Campagna, L., Silveira, L. F., Tubaro, P. L., & Lougheed, S. C. (2013). Identifying the sister species to the rapid capuchino seedeater radiation (Passeriformes: Sporophila). *The Auk*, *130*(4), 645–655. https://doi.org/10.1525/auk.2013.13064

Campagna, L., & Toews, D. P. L. (2022). The genomics of adaptation in birds. *Current Biology*, *32*(20), R1173–R1186. https://doi.org/10.1016/j.cub.2022.07.076

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742–s13015. https://doi.org/10.1186/s13742-015-0047-8

Chen, J., Liu, Y., Liu, M., Guo, W., Wang, Y., He, Q., Chen, W., Liao, Y. i, Zhang, W., Gao, Y., Dong, K., Ren, R., Yang, T., Zhang, L., Qi, M., Li, Z., Zhao, M., Wang, H., Wang, J., … Diao, X. (2023). Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nature Genetics*, *55*(12), 2243–2254. https://doi.org/10.1038/s41588-023-01571-z

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. https://doi.org/10.1093/bioinformatics/btx364

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., Zepeda-Mendoza, M. L., Campos, P. F., Heller, R., & Pereira, R. J. (2016). Next-generation biology: Sequencing and data analysis approaches for non-model organisms. *Marine Genomics*, *30*, 3–13. https://doi.org/10.1016/j.margen.2016.04.012

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

De Coster, W., Weissensteiner, M. H., & Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews Genetics*, *22*(9), 572–587. https://doi.org/10.1038/s41576-021-00367-3

Dong, S.–S., He, W.–M., Ji, J.–J., Zhang, C., Guo, Y., & Yang, T.–L. (2021). LDBlockShow: A fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Briefings in Bioinformatics*, *22*(4), bbaa227. https://doi.org/10.1093/bib/bbaa227

Edwards, S. V., Fang, B., Khost, D., Kolyfetis, G. E., Cheek, R. G., Deraad, D., Chen, N., Fitzpatrick, J. W., McCormack, J. E., & Funk, W. C. (2025). Comparative population pangenomes reveal unexpected complexity and fitness effects of structural variants. *bioRxiv*, 2022–2025.

Estalles, C., Turbek, S. P., José Rodríguez-Cajarville, M., Silveira, L. F., Wakamatsu, K., Ito, S., Lovette, I. J., Tubaro, P. L., Lijtmaer, D. A., & Campagna, L. (2022). Concerted variation in melanogenesis genes underlies emergent patterning of plumage in capuchino seedeaters. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1966), 20212277. https://doi.org/10.1098/rspb.2021.2277

Fang, B., & Edwards, S. V. (2024). Fitness consequences of structural variation inferred from a House Finch pangenome. *Proceedings of the National Academy of Sciences*, *121*(47), e2409943121. https://doi.org/10.1073/pnas.2409943121

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research*,*34*, 769–777. https://doi.org/10.1101/gr.278090.123

Gao, C., Chen, C., Akyol, T., Dusa, A., Yu, G., Cao, B., & Cai, P. (2024). ggVennDiagram: Intuitive Venn diagram software extended. *Imeta*, *3*(1), e177. https://doi.org/10.1002/imt2.177

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–879. https://doi.org/10.1038/nbt.4227

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

Hejase, H. A., Salman-Minkov, A., Campagna, L., Hubisz, M. J., Lovette, I. J., Gronau, I., & Siepel, A. (2020). Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. *Proceedings of the National Academy of Sciences*, *117*(48), 30554–30565. https://doi.org/10.1073/pnas.2015987117

Heller, D., & Vingron, M. (2020). SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics*, *36*(22–23), 5519–5521.

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., Dawson, E. T., Garrison, E., Novak, A. M., & Paten, B. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, *21*, 1–17. https://doi.org/10.1186/s13059-020-1941-7

Hickey, G., Monlong, J., Ebler, J., Novak, A. M., Eizenga, J. M., Gao, Y., Abel, H. J., Antonacci-Fulton, L. L., Asri, M., Baid, G., Baker, C. A., Belyaeva, A., Billis, K., Bourque, G., Buonaiuto, S., Carroll, A., Chaisson, M. J. P., Chang, P. i-C., Chang, X. H., … Paten, B. (2024). Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nature Biotechnology*, *42*(4), 663–673. https://doi.org/10.1038/s41587-023-01793-w

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., Smit, A. F. A., & Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. *Nucleic Acids Research*, *44*(D1), D81–D89. https://doi.org/10.1093/nar/gkv1272

Irwin, D. E., Milá, B., Toews, D. P. L., Brelsford, A., Kenyon, H. L., Porter, A. N., Grossen, C., Delmore, K. E., Alcaide, M., & Irwin, J. H. (2018). A comparison of genomic islands of differentiation across three young avian species pairs. *Molecular Ecology*, *27*(23), 4839–4855. https://doi.org/10.1111/mec.14858

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., & Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*(1), 14061. https://doi.org/10.1038/ncomms14061

Kapusta, A., & Suh, A. (2017). Evolution of bird genomes—A transposon's-eye view. *Annals of the New York Academy of Sciences*, *1389*(1), 164–185. https://doi.org/10.1111/nyas.13295

Keilwagen, J., Hartung, F., & Grau, J. (2019). GeMoMa: Homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Gene Prediction: Methods and Protocols*, *1962*, 161–177.

Kellogg, E. A. (2015). Genome sequencing: Long reads for a short plant. *Nature Plants*, *1*(12), 1–2. https://doi.org/10.1038/nplants.2015.169

Kramer, A. E., Ellwood, K. M., Guarino, N., Li, C.–J., & Dutta, A. (2025). Transcriptomic data reveals MYC as an upstream regulator in laying hen follicular recruitment. *Poultry Science*, *104*(1), 104547. https://doi.org/10.1016/j.psj.2024.104547

Lecomte, L., Árnyasi, M., Ferchaud, A., Kent, M., Lien, S., Stenløkk, K., Sylvestre, F., Bernatchez, L., & Mérot, C. (2024). Investigating structural variant, indel and single nucleotide polymorphism differentiation between locally adapted Atlantic salmon populations. *Evolutionary Applications*, *17*(3), e13653. https://doi.org/10.1111/eva.13653

Li, H., Marin, M., & Farhat, M. R. (2024). Exploring gene content with pangene graphs. *Bioinformatics*, *40*(7), btae456. https://doi.org/10.1093/bioinformatics/btae456

Li, H., & Ralph, P. (2019). Local PCA shows how the effect of population structure differs along the genome. *Genetics*, *211*(1), 289–304. https://doi.org/10.1534/genetics.118.301747

Liang, Y., Abedini, S., Farbehi, N., & Alinejad-Rokny, H. (2024) ' How chromatin interactions shed light on interpreting non-coding genomic variants: Opportunities and future directions', arXiv, arXiv:2411.17956, preprint: not peer reviewed.

Lijtmaer, D. A., Sharpe, N. M. M., Tubaro, P. L., & Lougheed, S. C. (2004). Molecular phylogenetics and diversification of the genus Sporophila (Aves: Passeriformes). *Molecular Phylogenetics and Evolution*, *33*(3), 562–579. https://doi.org/10.1016/j.ympev.2004.07.011

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M., & Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *Elife*, *11*, e78526. https://doi.org/10.7554/eLife.78526

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology*, *20*, 1–14. https://doi.org/10.1186/s13059-019-1828-7

Manthey, J. D., Moyle, R. G., & Boissinot, S. (2018). Multiple and independent phases of transposable element amplification in the genomes of piciformes (woodpeckers and allies). *Genome Biology and Evolution*, *10*(6), 1445–1456. https://doi.org/10.1093/gbe/evy105

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, *27*(6), 764–770.

Marques, D. A., Meier, J. I., & Seehausen, O. (2019). A combinatorial view on speciation and adaptive radiation. *Trends in Ecology & Evolution*, *34*(6), 531–544. https://doi.org/10.1016/j.tree.2019.02.008

Mérot, C., Oomen, R. A., Tigano, A., & Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends in Ecology & Evolution*, *35*(7), 561–572.

Pacific Biosciences. (2021). *pbsv-PacBio structural variant (SV) calling and analysis tools*. *GitHub*. https://github.com/PacificBiosciences/pbsv. Date acccesed November 18, 2024.

Parmigiani, L., Garrison, E., Stoye, J., Marschall, T., & Doerr, D. (2024). Panacus: Fast and exact pangenome growth and core size estimation. *Bioinformatics*, *40*, btae720. https://doi.org/10.1093/bioinformatics/btae720

Pool, J. E., Hellmann, I., Jensen, J. D., & Nielsen, R. (2010). Population genetic inference from genomic sequence variation. *Genome Research*, *20*(3), 291–300. https://doi.org/10.1101/gr.079509.108

Pumpernik, D., Oblak, B., & Borštnik, B. (2008). Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Molecular Genetics and Genomics*, *279*, 53–61. https://doi.org/10.1007/s00438-007-0294-1

R Core Team. (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Recuerda, M., & Campagna, L. (2024). How structural variants shape avian phenotypes: Lessons from model systems. *Molecular Ecology*, *33*(11), e17364. https://doi.org/10.1111/mec.17364

Saitou, M., Masuda, N., & Gokcumen, O. (2022). Similarity-based analysis of allele frequency distribution among multiple populations identifies adaptive genomic structural variants. *Molecular Biology and Evolution*, *39*(3), msab313. https://doi.org/10.1093/molbev/msab313

Schwander, T., Libbrecht, R., & Keller, L. (2014). Supergenes and complex phenotypes. *Current Biology*, *24*(7), R288–R294. https://doi.org/10.1016/j.cub.2014.01.056

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P. C., & Carroll, A. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, *374*(6574), abg8871.

Smit, A. F. A., Hubley, R., & Green, P. (2015). *RepeatMasker open-4.0 (2013–2015)*. Institute for Systems Biology. http://www.repeatmasker.org. Date accessed June 3, 2024.

Smit, A. F. A., Hubley, R., & Green, P. (2019). *RepeatModeler open-1.0 (2008–2015)*. http://www.repeatmasker.org/RepeatModeler. Date accessed June 3, 2024.

Turbek, S. P., Browne, M., Di Giacomo, A. S., Kopuchian, C., Hochachka, W. M., Estalles, C., Lijtmaer, D. A., Tubaro, P. L., Silveira, L. F., Lovette, I. J., Safran, R. J., Taylor, S. A., & Campagna, L. (2021). Rapid speciation via the evolution of pre-mating isolation in the Iberá Seedeater. *Science*, *371*(6536), eabc0256. https://doi.org/10.1126/science.abc0256

UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wallbank, R. W. R., Baxter, S. W., Pardo-Diaz, C., Hanly, J. J., Martin, S. H., Mallet, J., Dasmahapatra, K. K., Salazar, C., Joron, M., Nadeau, N., McMillan, W. O., & Jiggins, C. D. (2016). Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biology*, *14*(1), e1002353. https://doi.org/10.1371/journal.pbio.1002353

Wang, K., Hua, G., Li, J., Yang, Y. u, Zhang, C., Yang, L., Hu, X., Scheben, A., Wu, Y., Gong, P., Zhang, S., Fan, Y., Zeng, T., Lu, L., Gong, Y., Jiang, R., Sun, G., Tian, Y., Kang, X., … Li, W. (2024). Duck pan-genome reveals two transposon insertions caused bodyweight enlarging and white plumage phenotype formation during evolution. *Imeta*, *3*(1), e154. https://doi.org/10.1002/imt2.154

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. ' A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., … Haussler, D. (2022). The Human Pangenome Project: A global resource to map genomic diversity. *Nature*, *604*(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

Wei, H., Wang, X., Zhang, Z., Yang, L., Zhang, Q., Li, Y., He, H., Chen, D., Zhang, B., Zheng, C., Leng, Y., Cao, X., Cui, Y., Shi, C., Liu, Y., Lv, Y., Ma, J., He, W., Liu, X., … Shang, L. (2024). Uncovering key salt-tolerant regulators through a combined eQTL and GWAS analysis using the super pan-genome in rice. *National Science Review*, *11*(4), nwae043. https://doi.org/10.1093/nsr/nwae043

Wellenreuther, M., & Bernatchez, L. (2018). Eco-evolutionary genomics of chromosomal inversions. *Trends in Ecology & Evolution*, *33*(6), 427–440. https://doi.org/10.1016/j.tree.2018.04.002

Wells, J. N., & Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annual Review of Genetics*, *54*(1), 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Wheatcroft, D., & Qvarnström, A. (2017). Genetic divergence of early song discrimination between two young songbird species. *Nature Ecology & Evolution*, *1*(7), 0192. https://doi.org/10.1038/s41559-017-0192

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Zhang, H., Yu, J.–Q., Yang, L.–L., Kramer, L. M., Zhang, X.–Y., Na, W., Reecy, J. M., & Li, H. (2017). Identification of genome-wide SNP-SNP interactions associated with important traits in chicken. *BMC*

Genomics [Electronic Resource], *18*, 1–10. https://doi.org/10.1186/s12864-017-4252-y

Zhang, X., Zhao, M., McCarty, D. R., & Lisch, D. (2020). Transposable elements employ distinct integration strategies with respect to transcriptional landscapes in eukaryotic genomes. *Nucleic Acids Research*, *48*(12), 6685–6698. https://doi.org/10.1093/nar/gkaa370

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, *28*(24), 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Received April 5, 2025; revisions received September 4, 2025; accepted September 17, 2025
Associate Editor: Claire Mérot; Handling Editor: Tim Connallon