Journal of Heredity, 2026, **117**, 469–477
https://doi.org/10.1093/jhered/esaf076
Advance access publication 7 October 2025
Genome Resources

American Genetic Association
OXFORD

Genome Resources

Reference genome of an irruptive migrant, the pine siskin (*Spinus pinus*)

Jair Cortez[1], Joel W.G. Slade[2], Tricia A. Van Laar[1,*]

[1]Department of Biological Sciences, College of Science, California State University, Stanislaus, Turlock, CA, United States, [2]Department of Biology, College of Science and Mathematics, California State University, Fresno, Fresno, CA, United States

*Corresponding author: Department of Biological Sciences, College of Science, California State University, Stanislaus, Turlock, CA, United States.
Email: tvanlaar@csustan.edu

Abstract

Pine siskins (*Spinus pinus*) are irruptive migratory songbirds of biological interest in studies of endocrine regulation, immune function, and behavioral flexibility. Here, we present a chromosome-level reference genome from a female pine siskin, assembled de novo using long-read sequencing and scaffolded with a reference-guided approach. Synteny analyses also showed that our assembly can reliably reveal genomic rearrangements relative to other finches. Using reference-based annotation, we identified thousands of protein-coding genes, including loci relevant to metabolism and immune function that demonstrate the utility of this assembly for downstream studies. We also found evidence of gene duplications and pseudogenization in immune loci, showing the utility of our assembly for immunogenetic studies. Our analysis provides the first genome-wide view of transposable element (TE) activity in *Spinus*, revealing multiple bursts of long terminal repeat (LTR) retrotransposon expansion, including a recent one that coincides with the estimated diversification of North and South American siskins approximately 2.7 million years ago. We also detected putative lineage-specific LTR sequences, suggesting recent or ongoing TE diversification. This assembly fills a critical gap in passerine genomic resources and provides a resource for comparative, transcriptomic, and population-level studies across species with diverse migratory strategies.

Key words: comparative genomics, immunogenetics, long-read sequencing, repetitive elements

Introduction

Pine siskins (*S. pinus*) are small, gregarious finches known for their unpredictable and long-distance irruptive migrations, which differ from the more predictable movements of their relatives within the same genus (Newton 2006; DeSimone et al. 2023). These migrations have made pine siskins an effective comparative model for studying behavioral ecology, social communication, and physiological adaptations (DeSimone et al. 2021; Robart and Watts 2023; Tir and Watts 2025). For example, pine siskins show similar physiological and behavioral changes as obligate migrants, but the underlying endocrine changes do not always match, which suggests that pine siskins have unique molecular mechanisms that help prepare them for their migration (Vernasco et al. 2021; Becker and Watts 2025).

The evolutionary relationships within the pine siskin complex have also generated interest, as some researchers have suggested that a currently recognized subspecies of pine siskin (*S. p. perplexus*) may represent a distinct species (Alvarez et al. 2016). This uncertainty is similar to broader challenges in resolving relationships within *Spinus*, where even genus-level analyses have struggled to place Caribbean taxa and the Eurasian siskin using mtDNA alone (Beckman and Witt 2015).

More recently, the involvement of pine siskins in *Salmonella* outbreaks that impact wildlife and human populations has generated interest in the molecular basis of their immunity (Fig. 1). A 2021 human salmonellosis outbreak across multiple states in the United States was traced back to pine siskins (Patel et al. 2023). Follow-up studies showed that a pine siskin sampled during this outbreak carried multidrug-resistant *Salmonella* (Van Laar et al. 2022). In addition, there is evidence that pine siskins differ from other Cardueline finches in their ability to clear Salmonella with their innate immune systems (Biehler et al. 2025).

Despite their usefulness as a model for a range of research topics, molecular resources for pine siskins remain limited. The only available genome (Bodt et al. 2024) is based on short-read sequencing and may lack the resolution to properly explore structural variation and structurally complex genomic regions (reviewed in Ho et al. 2020). To address this gap, we present a highly complete and contiguous chromosome-level genome of a female pine siskin generated using long-read sequencing. This resource will support future studies

Received on 26 June 2025; revised on 30 September 2025; accepted on 1 October 2025

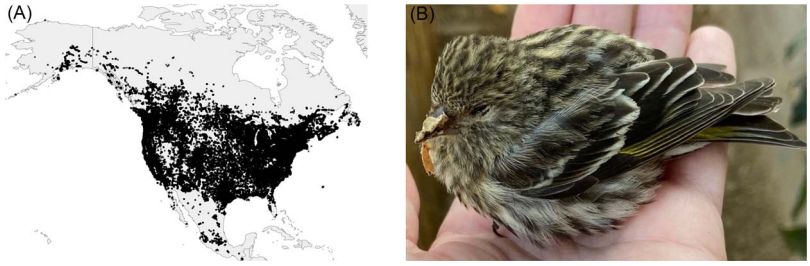

Fig. 1. (A) Pine siskins are found throughout North America, though their range can vary significantly year-to-year due to irruptive migrations. This map shows pine siskin observations from September 2020 to June 2021, during the most recent irruptive migration (eBird 2021). (B) A pine siskin infected with *Salmonella* is shown. Photographed by Tricia Van Laar in Fresno, CA in January 2021.

into the evolution, physiology, and immunogenetics of this species.

Methods

Blood sample collection

This individual was found at Scout Island Outdoor Education Center (36°51′34.35″N, 119°50′34.88″W) in Fresno, California, USA, on 29 April 2024. At the time of capture, the individual had the following measurements: wing chord, 77 mm; tail length, 50 mm; tarsus, 16.1 mm; fat score, 3; and mass, 14.1 g. We collected 70 μL of whole blood via brachial venipuncture. Before release, the bird was banded with an aluminum United States Geological Survey (USGS) bird banding lab band, etched with the unique ID code 2970-08367. This individual was confirmed to be female by polymerase chain reaction amplification of the chromo helicase DNA-binding gene (*CHD*), located on the sex chromosomes, using previously primers CHD1-i16 forward (5′-GTC CTG ATT TTC TCA CAG ATG G-3′) and CHD1-i16 reverse (5′-ATG ATC CAG TGC TTG TTT CC-3′) as previously described (Suh et al. 2011). To preserve cellular integrity, we stored the blood sample with 1 μL of 500 mM K_2-EDTA buffer at −80 °C before sending to UC Davis for DNA extraction. All work was performed under the following permits: USGS Bird Banding Lab Permit 24215; California Scientific Collection Permit S-191350004-19165-001; Fresno State IACUC 1607.

Library preparation and sequencing

High-molecular-weight (HMW) genomic DNA was extracted by UC Davis DNA Technologies and Expression Analysis Core using the sodium dodecyl sulfate-photorefractive keratectomy method from the whole blood of the pine siskin sampled. The extraction yielded 41 μg of DNA at a concentration of 249 ng/μL in a final volume of 166 μL. The HiFi SMRTbell library was constructed using the SMRTbell prep kit 3.0 (Pacific Biosciences, Menlo Park, CA; Cat. #102-182-700) according to the manufacturer's instructions. HMW gDNA was sheared to a target DNA size distribution between 15 and 18 kb using Diagenode's Megaruptor 3 system (Diagenode, Belgium; cat. B06010003). The sheared gDNA was concentrated using 1X of SMRTbell cleanup beads provided in the SMRTbell prep kit 3.0 for the repair and a-tailing incubation at 37 °C for 30 min and 65 °C for 5 min, followed by ligation of overhang adapters at 20 °C for 30 min, clean-up using 1X SMRTbell cleanup beads, and nuclease treatment at 37 °C for 15 min. The SMRTbell library was size selected using 3.1X of 35% v/v diluted AMPure PB beads (Pacific Biosciences, Menlo Park, CA; Cat. #100-265-900) to progressively remove SMRTbell templates < 5 kb. The 15 to 25 kb average HiFi SMRTbell library was sequenced at the UC Davis DNA Technologies Core (Davis, CA) using one 25 M SMRT cell (Pacific Biosciences, Menlo Park, CA; Cat #102-202-200), Revio SPRQ sequencing chemistry, and 30-h movies on a PacBio Revio sequencer.

Nuclear genome assembly

All software used for genome processing and analysis are summarized in Table 1. We trimmed raw reads using Cutadapt v3.5 (Martin 2011) with a 0.1 error rate, 35 bp minimum overlap, and the –discard-trimmed flag to retain only fully trimmed reads. k-mer histograms were generated with Meryl v1.4.1 (Rhie et al. 2020) using $k = 31$ and analyzed with GenomeScope v2.0 (Ranallo-Benavidez et al. 2020) to estimate genome size, heterozygosity, and repeat content. Assembly was performed with Hifiasm v0.25.0 (Cheng et al. 2021) using default parameters. We scaffolded both primary and alternate assemblies with RagTag v2.1.0 (Alonge et al. 2019), using the Azores bullfinch (*Pyrrhula murina*) as a reference due to its karyotype ($n = 41$), which matches that of other siskins (Malinovskaya et al. 2022; Grishko et al. 2023). In contrast, the house finch (*Haemorhous mexicanus*), which has a haploid chromosome count of 40, was less suitable (Fang and Edwards 2024). Before scaffolding, we removed 11 unplaced scaffolds from the *P. murina* reference. We evaluated contiguity with QUAST v5.3.0 (Gurevich et al. 2013), completeness with BUSCO v5.8.2 (Simão et al. 2015) using the passeriformes_odb12 dataset, and screened for contamination using NCBI's Foreign Contamination Screen v0.5.5 (Astashyn et al. 2024).

Reference-guided scaffolding

We scaffolded both primary and alternate assemblies with RagTag v2.1.0 (Alonge et al. 2019), using the Azores bullfinch (*P. murina*) as a reference (GCA_965183895.1) because its haploid chromosome count ($n = 41$; $2n = 82$) matches that of other siskins (Malinovskaya et al. 2022; Grishko et al. 2023). Although the house finch (*H. mexicanus*) is phylogenetically closer than the Azores bullfinch to the pine siskin (*S. pinus*), it has a derived karyotype of $2n = 80$ (Fang and Edwards 2024), whereas $2n = 82$ is likely ancestral for Carduelinae (Malinovskaya et al. 2022). The pine siskin's karyotype is unknown, so we cannot confirm the exact chromosome number, but its close relative, Eurasian siskin (*Spinus spinus*), has $2n = 82$. The *P. murina* reference yielded the highest total placed sequence, with the gain in placed bases (9,307,218 bp) far exceeding the small increase in gap sequence (9,800 bp). In both *P. murina* and *H. mexicanus*, chromosomes are ordered by size

Table 1. Summary of workflow and associated software.

Pipeline step	Software tool	Version
Convert BAM to FASTQ	SAMtools	1.21
Adapter removal from long reads	Cutadapt	5.0
k-mer counting ($k = 31$)	Meryl	1.4.1
k-mer histogram visualization	GenomeScope	2.0
Genome assembly	Hifiasm	0.25.0
Sequencing statistics	Seqkit	2.1.0
Assembly quality assessment	QUAST	5.3.0
Genome completeness	BUSCO	5.8.3
Quality and completeness visualization	BlobToolKit	4.4.6
Reference-guided scaffolding	RagTag	2.1.0
Adapter and contaminant screening	NCBI FCS	0.5.5
Synteny analysis	D-GENIES; minimap	1.5.0; 2.28
Dotplot visualization	R (pafr, dplyr, ggplot2, png)	4.4.3
De novo repeat family identification	RepeatModeler	2.0.6
Repeat library curation	MCHelper	1.7.1
Repeat masking (soft-masking)	RepeatMasker	4.1.8
Comparison of repeat libraries	BEDTools	2.31.1
Repeat content visualizations	R (tidyverse, reshape2, viridis, ggplot2, dplyr, tidyr, forcats)	4.4.3
Gene annotation transfer	Liftoff	1.6.3
Gene annotation summary statistics	AGAT	1.4.1
Coding sequence extraction	gffread	0.12.7

rather than by homology to model species, so chromosome numbers are not directly comparable. For example, zebra finch chromosome 16 is not equivalent to chromosome 16 in either assembly. The domestic canary (*Serinus canaria* forma *domestica*) also has $2n = 82$, but its public genome assembly is not at chromosome level. Before scaffolding, we removed 11 unplaced scaffolds from the *P. murina* reference, following the approach of Thorburn et al. (2023) and Hiller et al. (2021), where unmapped sequences were removed from reference genomes prior to scaffolding.

Assessing assembly quality

We evaluated contiguity of both raw and scaffolded assemblies with QUAST v5.3.0 (Gurevich et al. 2013), completeness with BUSCO v5.8.2 (Simão et al. 2015) using the passeriformes_odb12 dataset, and screened for contamination using NCBI's Foreign Contamination Screen v0.5.5 (Astashyn et al. 2024).

Synteny analysis

We performed synteny analyses to compare our pine siskin long-read genome assembly (Hap1) with the published short-read genome assembly (Bodt et al. 2024), using the Azores bullfinch (*P. murina*) genome as a common reference. Pairwise alignments were generated with D-GENIES (v1.5.0), which internally uses minimap2 (v2.28) under default parameters to produce PAF alignment files. The resulting PAF files were downloaded and imported into R v4.1.2, where we used the pafr package to generate customized dot plots. This approach allowed us to directly compare the long-read and short-read assemblies against the same reference and to visualize that RagTag scaffolding preserved local structural variations present in the raw Hap1 assembly.

Repeat content annotation

We generated an initial consensus library of repetitive elements using RepeatModeler2 v2.0.6 (Flynn et al. 2020) with the -LTRStruct option, using the primary assembly as input as it tends to be more complete than the alternate haplotype (Cheng et al. 2021). The library was curated using MCHelper (Orozco-Arias et al. 2024), an automated pipeline that clusters redundant elements, refines consensus sequences, and evaluates structural integrity. We used the curated repeat library to soft-mask both the primary and alternate assemblies with RepeatMasker v4.1.8 (Smit et al. 2013). For downstream analyses, we excluded elements flagged as incomplete or unconfirmed by MCHelper. Kimura divergence values were calculated using the calcDivergenceFromAlign.pl script provided by RepeatMasker, and repeat landscapes were visualized in R v4.1.2 with custom plotting scripts, revealing peaks consistent with historical bursts of transposable element (TE) activity. Finally, to quantify the similarity between repeat libraries produced by RepeatMasker and MCHelper, we used the jaccard function in BEDTools v2.31.0 (Quinlan and Hall 2010).

Gene annotation

We annotated the repeat-masked primary and alternate genomes using Liftoff v1.6.3 (Shumate and Salzberg 2021), transferring gene models from the house finch (*H. mexicanus*) reference genome (annotation GCF_027477595.1-RS_2023_09) to our assemblies. We used a minimum alignment coverage (-a) and sequence identity (-s) threshold of 85% to ensure high-confidence gene mappings. We included the –flank 0.2 option to extend alignments beyond annotated gene boundaries, improving model accuracy, and enabled both the –copies and –polish options to detect duplicated loci and refine gene models with incomplete open reading frames (ORFs).

Results

Sequencing statistics

PacBio HiFi sequencing produced 3,208,032 reads totaling 53.90 Gb. The reads had lengths ranging from 211 bp to 55.38 kb and an average length of 16.8 kb, which fall within

Table 2. Summary of primary and alternate pseudo-haplotype statistics.

Category	Metric	Hap1	Hap2
Basic statistics	# Scaffolds (≥ 0 bp)	208	109
	# Scaffolds (≥ 50,000 bp)	151	101
	Total length	1,253,257,106	1,097,225,515
	Largest contig	161,097,886	156,528,920
	GC (%)	44.18	43.65
	# N's per 100 kbp	3.10	1.08
Contiguity (Contig-level)	N50	19,100,671	23,035,499
	L50	16	14
Contiguity (Scaffold-level)	N50	63,847,955	64,938,551
	N90	11,963,686	8,200,734
	auN	69,404,653.9	69,369,973.1
	L50	7	6
	L90	26	26
BUSCO completeness (n = 6,684)	Complete (C)	98.4% (6,577)	95.5% (6,385)
	Single-copy (S)	98.2% (6,567)	95.4% (6,375)
	Duplicated (D)	0.1% (10)	0.1% (10)
	Fragmented (F)	0.3% (23)	0.3% (23)
	Missing (M)	1.3% (84)	4.1% (276)

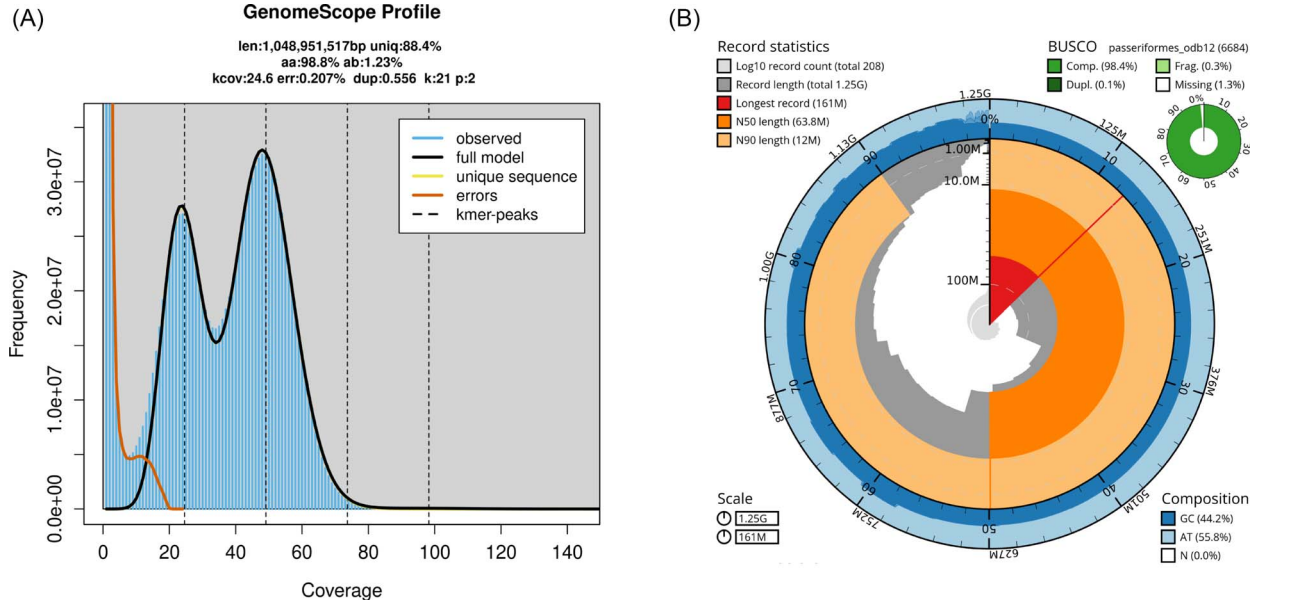

Fig. 2. (A) The histogram shows observed k-mer frequencies (columns) from raw sequencing reads, with an overlaid fitted diploid model (solid line). From left to right, the distinct peaks correspond to heterozygous and homozygous k-mers. The estimated genome size is 1.05 Gb with a low sequencing error rate (0.207%) and moderate heterozygosity (1.23%). (B) The final primary assembly spans 1.25 Gb, slightly exceeding the k-mer-based estimate. It has a GC content of 44.2% and high completeness, with 98.4% of 6684 conserved single-copy orthologs identified as complete.

the expected range for PacBio HiFi sequencing (10 to 25 kb; Hon et al. 2020). Read trimming did not remove or clip any sequences, so these statistics apply equally to both the raw and trimmed datasets.

Genome assembly

We assembled the genome of a female pine siskin (*S. pinus*) using Hifiasm, which generated two pseudo-haplotype assemblies, hereafter referred to as Hap1 (primary) and Hap2 (alternate) (Table 2). RagTag placed 430 of 596 contigs, corresponding to 97.7% of the genome, indicating strong structural similarity between *S. pinus* and *P. murina*. The total assembly sizes, 1.25 Gb for Hap1 and 1.10 Gb for Hap2, slightly exceeded the genome size estimated from k-mer analysis with Meryl and GenomeScope2 (Fig. 2). Given the 53.90 Gb of sequenced bases, the coverage values for Hap1 and Hap2 are 43.1× and 49.0×, respectively. Hap1 consists of 208 scaffolds and Hap2 of 109. Both assemblies show similar GC content (44.18% for Hap1 and 43.65% for Hap2) and low levels of ambiguous bases (3.10 and 1.08 N's per 100 kbp, respectively). Hap1 contains the longest scaffold at 161.1 Mb, though Hap2 is slightly more contiguous, with a higher N50 (64.9 Mb vs. 63.8 Mb), lower L50 (6 vs. 7), and nearly identical auN (69.4 Mb for both). Even at the contig-level before reference-guided scaffolding, Hap1 and Hap2 assemblies showed high continuity, with N50 values of 19.1 and 23.0 Mb and L50 values of 16 and 14, respectively. After scaffolding, continuity improved notably, with N50 values of about 63.8 and 64.9 Mb and L50 values reduced to 7 and 6, respectively. BUSCO analysis using the passeriformes_odb12

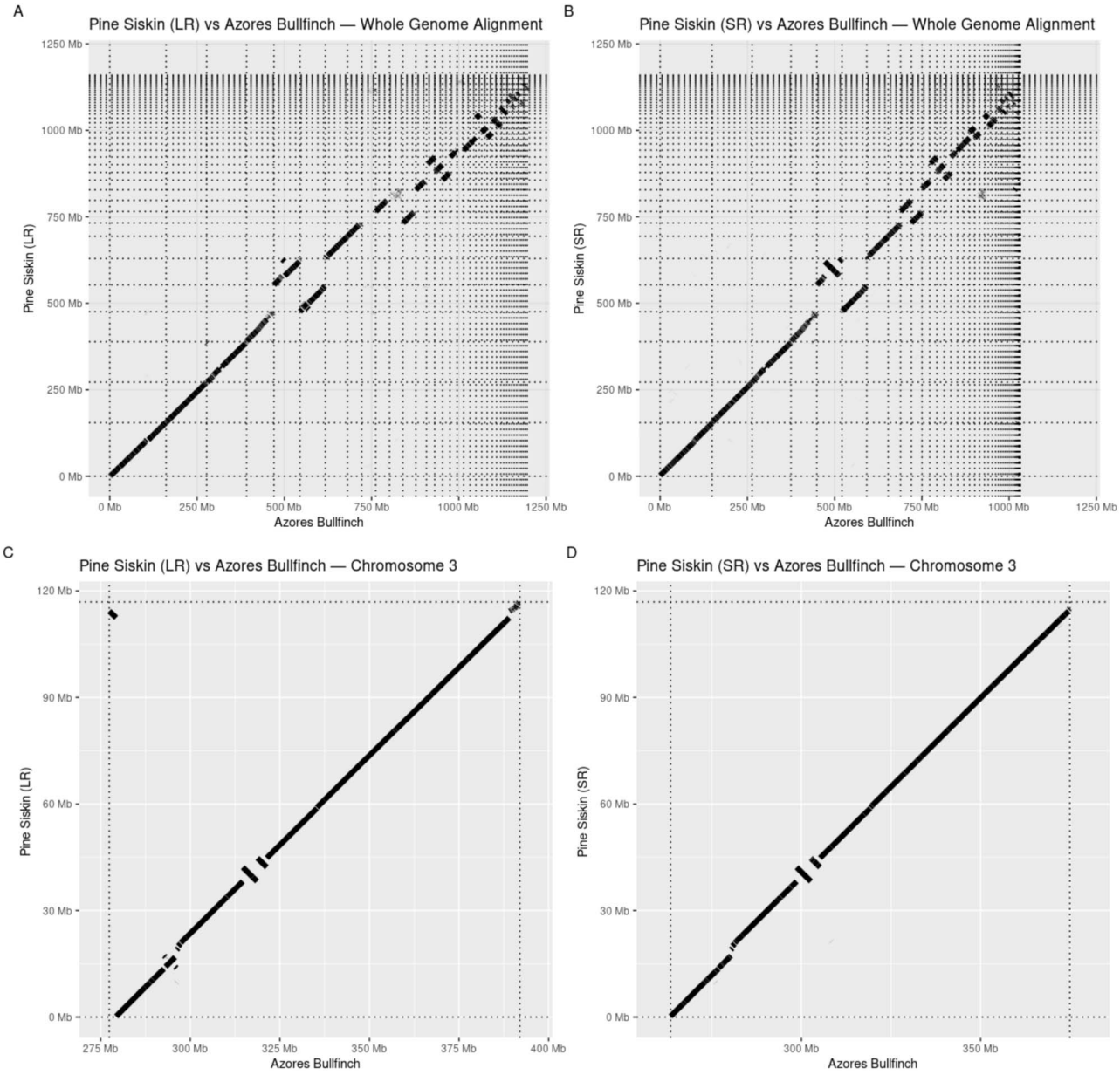

Fig. 3. Dot plot synteny comparisons of pine siskin assemblies aligned to the Azores bullfinch reference genome. (A) Whole-genome alignment of our pine siskin hap1 long-read (LR) assembly against the Azores bullfinch. (B) Whole-genome alignment of the published pine siskin short-read (SR) assembly (Bodt et al. 2024) against the Azores bullfinch. (C) Zoomed-in alignment of chromosome 3 from our LR assembly, showing two inversions relative to the Azores bullfinch. (D) Zoomed-in alignment of chromosome 3 from the published SR assembly, also showing two inversions, consistent with the LR assembly.

dataset ($n = 6{,}684$) showed that Hap1 recovers a higher percentage of complete orthologs (98.4% vs. 95.5%), whereas both assemblies have low duplication (0.1%) and fragmentation (0.3%) rates.

Synteny analysis

Whole-genome dot plots of the long-read Hap1 assembly and an independently assembled short-read genome (Bodt et al. 2024) against the *P. murina* reference revealed overall collinearity, consistent with conserved synteny between *S. pinus* and *P. murina* (Fig. 3). Both assemblies nonetheless showed localized differences relative to the bullfinch genome. Specifically, we observed two putative inversions on chromosome 3 that appeared in both assemblies, supporting their interpretation as genuine rearrangements in the *S. pinus* genome rather than artefacts of a single assembly strategy. However, the two assemblies did not show identical breakpoint structures or shapes in all regions. These discrepancies could reflect biological variation among individuals or technical differences associated with sequencing platforms and assembly methods. Additional genomes and validation will be required to confirm the prevalence and evolutionary context of these inversions in *S. pinus*.

Repetitive element analysis

We soft-masked the genome using both the raw output from RepeatModeler2 and a curated repeat library generated with MCHelper. The RepeatModeler2 library masked 23.22% of

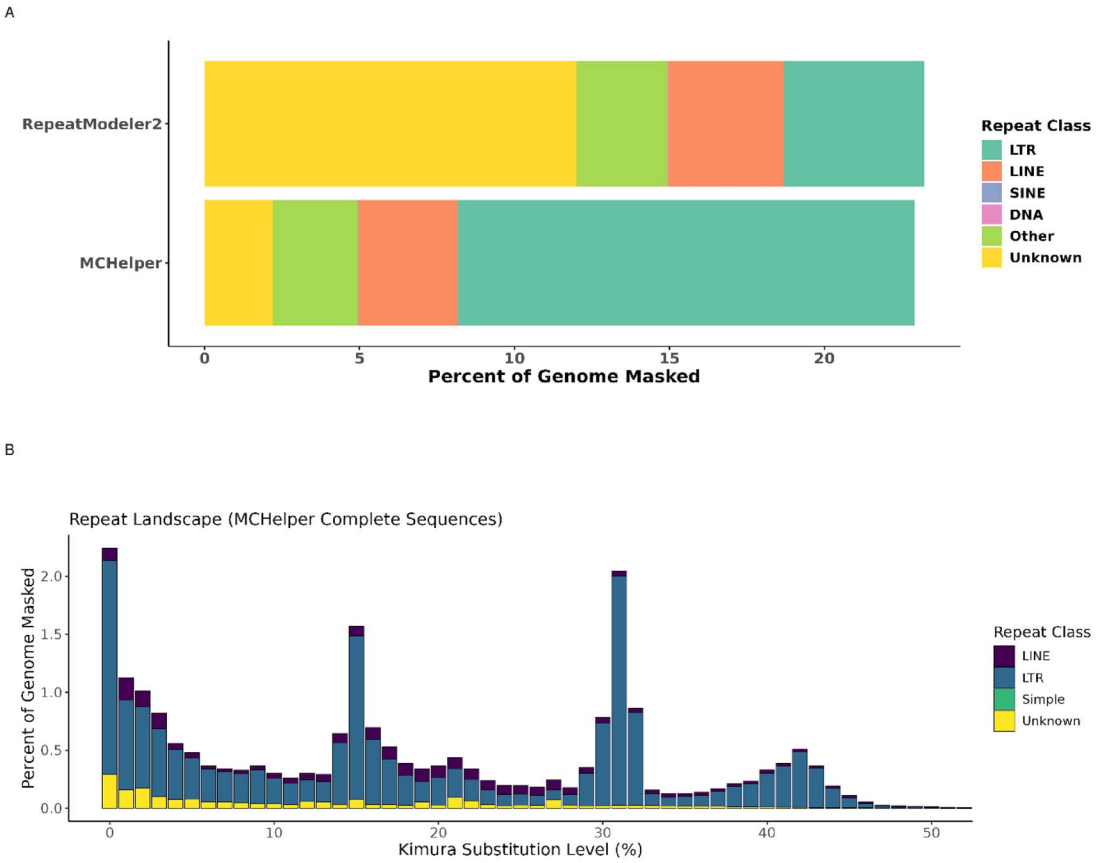

Fig. 4. (A) Curation of raw RepeatModeler2 outputs greatly reduced the proportion of unidentified repeats, reclassifying many as LTRs. Both raw and curated consensus libraries masked a similar proportion of the genome. (B) Repeat landscape for high-confidence, curated repeats from MCHelper. There have been several bursts of LTR activity, indicated by the three peaks in the landscape, including a recent burst within the past 2.2 million years ago.

the genome, whereas the MCHelper-curated library masked 22.90%. These outputs overlapped by 91%, indicating agreement in overall repeat content. Despite similar masking proportions, MCHelper improved classification by reassigning many previously unclassified elements in the RepeatModeler2 library as LTR retrotransposons (Fig. 4A). LTRs represented the most abundant class of TEs, followed by LINEs. The repeat landscape showed a prominent peak near 0% Kimura divergence, indicating a recent expansion of TEs (Fig. 4B). Assuming a neutral mutation rate of 2.3×10^{-9} substitutions/site/year, estimated from germline data in another passerine (Smeds et al. 2016), these elements were probably active within the last 2.2 million years. In addition, as a preliminary analysis, we extracted LTR sequences with canonical TG and CA boundaries and performed BLAST searches using the NCBI nucleotide database. Most of the top hits showed high sequence identity to the Eurasian siskin (*S. spinus*), suggesting that these LTR elements may be siskin-specific or recently active within the *Spinus* genus.

Gene annotation

We identified 15,730 putative protein-coding genes in the primary assembly (Hap1), compared with the 16,679 in the house finch reference annotation. Overall, this number is similar to that of other Cardueline birds such as the common canary (*S. canaria*) which has 15,457 protein-coding genes annotated (annotation GCF_022539315.1-RS_2023_02). The W-linked gene *CHD1W*, commonly used in avian sexing, was annotated in our primary assembly, confirming that the sampled individual was female. We also recovered several endocrine genes that have previously been studied in pine siskins, including *NR3C1* and *NR3C2* (glucocorticoid and mineralocorticoid receptors), *HSD11B1* and *HSD11B2* (enzymes involved in cortisol activation and inactivation), *CYP19A1* (aromatase), and *ESR1* (estrogen receptor alpha) (Vernasco et al. 2021).

To show the utility of our assembly in immunogenetic studies, we also examined immune genes. Most Toll-like receptor (TLR) genes had valid ORFs and high sequence identity to reference sequences (Table 3). TLR2 and TLR7 were both retained in duplicate, which is consistent with prior research (Alcaide and Edwards 2011). TLR2 was present as two dispersed paralogs (the passerine-specific TLR2A and TLR2B), each with high sequence identity (>91%) and intact ORFs, consistent with functional maintenance. TLR7 was represented by two tandem copies, also with high identity (>90%) and intact ORFs, indicating a retained duplication. TLR1 was annotated at two loci. One copy showed high sequence identity (92%) with valid ORFs, whereas the second showed lower sequence identity (74%) and was flagged as low confidence due to annotation errors in Liftoff. Manual inspection of the extracted CDS suggested that an intact ORF may be present, but the evidence is inconclusive without further validation. We therefore consider the status of this second TLR1 locus as ambiguous. TLR1 was annotated at two loci: one copy with high identity (92%) and valid ORFs, and a second with lower identity (74%), no valid ORF due to a missing stop codon, and flagged as low confidence. However,

Table 3. Summary of Toll-like receptors (TLRs) identified.

Gene	Copies	Status	Notes
TLR1	2	Both putatively functional	One copy had lower identity but intact ORF.
TLR2	2	Both functional	Dispersed paralogs (TLR2A vs. TLR2B).
TLR3	1	Functional	One copy with intact ORF.
TLR4	1	Functional	One copy with intact ORF.
TLR5	1	Likely pseudogene	Internal stop codons.
TLR7	2	Both functional	Tandem duplicates, intact ORFs.

the extracted and translated CDS showed an intact ORF for the gene. TLR3 and TLR4 were recovered as single-copy genes with intact ORFs. In contrast, TLR5 was annotated as a single locus with low sequence identity (75%) and no valid ORF.

Discussion

We present a high-quality reference genome for pine siskins based on long-read sequencing. This is the second publicly available long-read genome for the genus *Spinus*, following that of the Eurasian siskin (*S. spinus*). The pseudo-haplotype assemblies (1.3 Gb for Hap1 and 1.1 Gb for Hap2) are comparable in size to other cardueline genomes, including the long-read assemblies of *H. mexicanus* (1.1 Gb) and *S. spinus* (1.3 Gb) (accession GCF_027477595.1 and GCA_034780795.1, respectively). GenomeScope estimated a haploid genome size of ~1.05 Gb, close to the Hap2 assembly size (1.10 Gb) and smaller than Hap1 (1.25 Gb). Hap1 had slightly higher BUSCO completeness (98.4% vs. 95.5%) but also contained more ambiguous bases and was inflated compared with the k-mer estimate, suggesting retention of repeat-rich or partially redundant sequence. By contrast, Hap2 matched the estimated haploid size more closely, with fewer ambiguous bases and similar contiguity. Both haplotypes showed very low BUSCO duplication (0.1%), which may reflect aggressive haplotig purging by Hifiasm, potentially removing true paralogs.

We demonstrate the possible uses of this genome for metabolic and immunogenetic studies by annotating key genes involved in endocrine function and innate immunity. These annotated loci can serve as references for future transcriptomic analyses and for the design of species-specific primers, allowing investigations of gene expression and regulation (Burri et al. 2014; Vernasco et al. 2021). Because pine siskins have been compared with the white-crowned sparrow (*Zonotrichia leucophrys*) due to their different migration strategies, comparative genomics using our assemblies and that of *Z. leucophrys* could be insightful (Vernasco et al. 2021; Wu et al. 2024). In addition, we identified several immune loci that mapped to multiple locations, suggesting potential duplication and a pseudogenized copy of TLR5, consistent with previous findings in other Fringillidae species (Bainová et al. 2014). Because we used conservative alignment parameters for Liftoff to ensure high-confidence gene models, our annotation likely underestimates the full repertoire of highly divergent genes. More permissive annotation strategies and transcriptomic data will therefore be needed to fully capture the immune gene landscape in the pine siskin.

This study also represents, to our knowledge, the first genome-wide analysis of TEs in *Spinus* birds. The repeat landscape we recovered is consistent with patterns observed in other passerines, dominated by LTR retrotransposons and LINEs (Benham et al. 2024). In addition to detecting shared TE classes, we found evidence of putative lineage-specific LTR elements, supported by high sequence similarity to the Eurasian siskin. The repeat divergence landscape included a component of relatively low-divergence elements, consistent with TE activity in the more recent evolutionary history of the genus. Although precise timing cannot be inferred from divergence plots alone, these patterns suggest that TE activity has continued since the diversification of North and South American siskins (Beckman and Witt 2015). In other vertebrate clades such as rodents, bats, and woodpeckers, episodes of TE amplification have coincided with periods of rapid diversification (reviewed in Almojil et al. 2021). Because our analysis is based on a single genome, it remains uncertain whether these TE dynamics reflect a species-wide pattern. Thus, broader sampling of *Spinus* genomes will be needed to assess whether this observation can be generalized.

In addition, our genome assembly represents *S. p. pinus*, and future sequencing of an *S. p. perplexus* individual could help further clarify the relationship between these two currently recognized subspecies. Although future assemblies of *S. pinus* or the addition of complementary data (e.g., Hi-C, optical maps) may refine chromosome structure, the underlying contig sequences and gene annotations presented here can be a valuable resource for comparative genomics, functional studies, and subsequent scaffolding efforts. Together with recently published assemblies for the Eurasian (*S. spinus)*, thick-billed (*Spinus crassirostris*), and olivaceous *(Spinus olivaceus*) siskins (Bodt et al. 2024), our genome provides a framework for future evolutionary studies.

Author contributions

Jair Cortez (Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing—original draft, Writing—review & editing), Joel W. G. Slade (Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Writing—review & editing), and Tricia Van Laar (Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing—review & editing)

Funding

The research described in this manuscript was funded by the National Science Foundation (Award 2233633; JWGS and TAV), California State University Biotechnology Network (CSUBIOTECH; JC and TAV), and California State University, Stanislaus, Research, Scholarly, and Creative Activity Award Program (RSCA; TAV).

Data availability

All code used can be found at https://github.com/vanlaarlab/pine-siskin-genome. The assemblies can be found on NCBI with the BioProject accession numbers PRJNA1281197 (primary) and PRJNA1281196 (alternate). The raw reads can be accessed with the BioSample number

SAMN49551328. The unscaffolded genome assemblies and curated repeat libraries are available at Dryad: https://doi.org/10.5061/dryad.xsj3tx9tc.

References

Alcaide M, Edwards SV. Molecular evolution of the Toll-like receptor multigene family in birds. *Mol Biol Evol*. 2011;**28**:1703–1715. https://doi.org/10.1093/molbev/msq351

Almojil D, Bourgeois Y, Falis M, Hariyani I, Wilcox J, Boissinot S. The structural, functional and evolutionary impact of transposable elements in eukaryotes. *Genes*. 2021;**12**:918. https://doi.org/10.3390/genes12060918

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol*. 2019;**20**:224. https://doi.org/10.1186/s13059-019-1829-6

Alvarez S, Salter JF, McCormack JE, Milá B. Speciation in mountain refugia: phylogeography and demographic history of the pine siskin and black-capped siskin complex. *J Avian Biol*. 2016;**47**:335–345. https://doi.org/10.1111/jav.00814

Astashyn A, Tvedte ES, Sweeney D, Sapojnikov V, Bouk N, Joukov V, Mozes E, Strope PK, Sylla PM, Wagner L, et al. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol*. 2024;**25**:60. https://doi.org/10.1186/s13059-024-03198-7

Bainová H, Králová T, Bryjová A, Albrecht T, Bryja J, Vinkler M. First evidence of independent pseudogenization of toll-like receptor 5 in passerine birds. *Dev Comp Immunol*. 2014;**45**:151–155. https://doi.org/10.1016/j.dci.2014.02.010

Becker AJ, Watts HE. The role of testosterone in spring nomadic migration of male pine siskins (*Spinus pinus*). *Horm Behav*. 2025;**169**:105694. https://doi.org/10.1016/j.yhbeh.2025.105694

Beckman EJ, Witt CC. Phylogeny and biogeography of the New World siskins and goldfinches: rapid, recent diversification in the Central Andes. *Mol Phylogenet Evol*. 2015;**87**:28–45. https://doi.org/10.1016/j.ympev.2015.03.005

Benham PM, Cicero C, Escalona M, Beraut E, Fairbairn C, Marimuthu MPA, Nguyen O, Sahasrabudhe R, King BL, Thomas WK, et al. Remarkably high repeat content in the genomes of sparrows: the importance of genome assembly completeness for transposable element discovery. *Genome Biol Evol*. 2024;**16**:evae067. https://doi.org/10.1093/gbe/evae067

Biehler LG, Van Laar TA, Slade JWG. Cardueline finches vary in constitutive innate immunity against a harmful strain of avian *Salmonella*. *Integr Comp Biol*. 2025;**65**:119–126. https://doi.org/10.1093/icb/icaf005

Bodt L, Hains T, Coldren D, Pirro S, Bates J, Hackett S. The complete genome sequences of 23 finch species (Fringillidae, Passeriformes). *Biodivers Genomes*. 2024;**2024**:1–3. https://doi.org/10.56179/001c.117726

Burri R, Promerová M, Goebel J, Fumagalli L. PCR-based isolation of multigene families: lessons from the avian MHC class IIB. *Mol Ecol Resour*. 2014;**14**:778–788. https://doi.org/10.1111/1755-0998.12234

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 2021;**18**:170–175. https://doi.org/10.1038/s41592-020-01056-5

DeSimone JG, Tobalske BW, Breuner CW. Physiology and behavior under food limitation support an escape, not preparative, response in the nomadic pine siskin (*Spinus pinus*). *J Exp Biol*. 2021;**224**:jeb238774. https://doi.org/10.1242/jeb.238774

DeSimone JG, Domschot BS, Fylling MA, Blake WM, Breuner CW. Body mass and triglycerides predict departure of free-living nomadic pine siskins. *Funct Ecol*. 2023;**37**:372–382. https://doi.org/10.1111/1365-2435.14238

eBird. *eBird: an online database of bird distribution and abundance [web application]*. Ithaca, New York: eBird, Cornell Lab of Ornithology; 2021: Available: http://www.ebird.org (accessed 2025 June 1).

Fang B, Edwards SV. Fitness consequences of structural variation inferred from a house finch pangenome. *Proc Natl Acad Sci U S A*. 2024;**121**:e2409943121. https://doi.org/10.1073/pnas.2409943121

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020;**117**:9451–9457. https://doi.org/10.1073/pnas.1921046117

Grishko E, Malinovskaya L, Slobodchikova A, Kotelnikov A, Torgasheva A, Borodin P. Cytological analysis of crossover frequency and distribution in male meiosis of cardueline finches (Fringillidae, Aves). *Animals*. 2023;**13**:1–14. https://doi.org/10.3390/ani13233624

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinf (Oxf)*. 2013;**29**:1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hiller AE, Brumfield RT, Faircloth BC. A reference genome for the nectar-robbing black-throated Flowerpiercer (*Diglossa brunneiventris*). *G3 (Bethesda)*. 2021;**11**:jkab271. https://doi.org/10.1093/g3journal/jkab271

Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet*. 2020;**21**:171–189. https://doi.org/10.1038/s41576-019-0180-9

Hon T, Mars K, Young G, Tsai Y-C, Karalius JW, Landolin JM, Maurer N, Kudrna D, Hardigan MA, Steiner CC, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data*. 2020;**7**:399. https://doi.org/10.1038/s41597-020-00743-4

Malinovskaya LP, Slobodchikova AY, Grishko EO, Pristyazhnyuk IE, Torgasheva AA, Borodin PM. Germline-restricted chromosomes and autosomal variants revealed by pachytene karyotyping of 17 avian species. *Cytogenet Genome Res*. 2022;**162**:148–160. https://doi.org/10.1159/000524681

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J*. 2011;**17**:10–12. https://doi.org/10.14806/ej.17.1.200

Newton I. Advances in the study of irruptive migration. *Ardea*. 2006;**94**:433

Orozco-Arias S, Sierra P, Durbin R, González J. MCHelper automatically curates transposable element libraries across eukaryotic species. *Genome Res*. 2024;**34**:2256–2268. https://doi.org/10.1101/gr.278821.123

Patel K, Stapleton GS, Trevejo RT, Tellier WT, Higa J, Adams JK, Hernandez SM, Sanchez S, Nemeth NM, Debess EE, et al. Human salmonellosis outbreak linked to salmonella typhimurium epidemic in wild songbirds, United States, 2020–2021. *Emerg Infect Dis*. 2023;**29**:2298–2306. https://doi.org/10.3201/eid2911.230332

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;**26**:841–842. https://doi.org/10.1093/bioinformatics/btq033

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;**11**:1432. https://doi.org/10.1038/s41467-020-14998-3

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;**21**:245. https://doi.org/10.1186/s13059-020-02134-9

Robart AR, Watts HE. Integration of social and temperature cues alters facultative migratory response to declining food availability. *Anim Behav*. 2023;**198**:153–164. https://doi.org/10.1016/j.anbehav.2023.01.019

Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics*. 2021;**37**:1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness

with single-copy orthologs. *Bioinformatics*. 2015;**31**:3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smeds L, Qvarnström A, Ellegren H. Direct estimate of the rate of germline mutation in a bird. *Genome Res*. 2016;**26**:1211–1218. https://doi.org/10.1101/gr.204669.116

Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [computer software]. 2013. http://www.repeatmasker.org.

Suh A, Kriegs JO, Brosius J, Schmitz J. Retroposon insertions and the chronology of avian sex chromosome evolution. *Mol Biol Evol*. 2011;**28**:2993–2997. https://doi.org/10.1093/molbev/msr147

Thorburn DMJ, Sagonas K, Binzer-Panchal M, Chain FJJ, Feulner PGD, Bornberg-Bauer E, Reusch TBH, Samonte-Padilla IE, Milinski M, Lenz TL, et al. Origin matters: using a local reference genome improves measures in population genomics. *Mol Ecol Resour*. 2023;**23**:1706–1723. https://doi.org/10.1111/1755-0998.13838

Tir JK, Watts HE. Declining food availability alters vocal behavior of a nomadic finch. *Behav Ecol*. 2025;**36**:araf026. https://doi.org/10.1093/beheco/araf026

Van Laar TA, Biehler L, Slade JWG. Draft genome sequence of a multidrug-resistant strain of *Salmonella enterica serovar* Typhimurium isolated from a pine siskin (*Spinus pinus*). *Microbiol Resour Announc*. 2022;**11**:e0098222. https://doi.org/10.1128/mra.00982-22

Vernasco BJ, Emmerson MG, Gilbert ER, Sewall KB, Watts HE. Migratory state and patterns of steroid hormone regulation in the pectoralis muscle of a nomadic migrant, the pine siskin (*Spinus pinus*). *Gen Comp Endocrinol*. 2021;**309**:113787. https://doi.org/10.1016/j.ygcen.2021.113787

Wu Z, Miedzinska K, Krause JS, Pérez JH, Wingfield JC, Meddle SL, Smith J. A chromosome-level genome assembly of a free-living white-crowned sparrow (*Zonotrichia leucophrys gambelii*). *Sci Data*. 2024;**11**:86. https://doi.org/10.1038/s41597-024-02929-6

Journal of Heredity, 2026, 117, 469–477
https://doi.org/10.1093/jhered/esaf076
Genome Resources