Title: Remarkably high repeat content in the genomes of sparrows: the importance of genome assembly completeness for transposable element discovery.

Authors: Phred M. Benham[1,2], Carla Cicero[1], Merly Escalona[3], Eric Beraut[4], Colin Fairbairn[4], Mohan P. A. Marimuthu[5], Oanh Nguyen[5], Ruta Sahasrabudhe[5], Benjamin L. King[6], W. Kelley Thomas[7], Adrienne I. Kovach[8], Michael W. Nachman[1,2], Rauri C. K. Bowie[1,2]

Author affiliations:

(1) Museum of Vertebrate Zoology, University of California Berkeley, Berkeley, CA 94720, USA.

(2) Department of Integrative Biology, University of California Berkeley, Berkeley, CA 94720, USA.

(3) Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA.

(4) Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.

(5) DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California-Davis, CA 95616, USA.

(6) Department of Molecular and Biomedical Sciences, University of Maine, Orono, ME 04469

(7) Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH 03824

(8) Department of Natural Resources and the Environment, University of New Hampshire, Durham, NH 03824

Corresponding author: Phred M. Benham, phbenham@gmail.com

Running title: repeat content and sparrow genomes

ABSTRACT: Transposable elements (TE) play critical roles in shaping genome evolution. However, the highly repetitive sequence content of TEs is a major source of assembly gaps. This makes it difficult to decipher the impact of these elements on the dynamics of genome evolution. The increased capacity of long-read sequencing technologies to span highly repetitive regions of the genome should provide novel insights into patterns of TE diversity. Here we report the generation of highly contiguous reference genomes using PacBio long read and Omni-C technologies for three species of sparrows in the family Passerellidae. To assess the influence of sequencing technology on TE annotation, we compared these assemblies to three chromosome-level sparrow assemblies recently generated by the Vertebrate Genomes Project and nine other sparrow species generated using a variety of short- and long-read technologies. All long-read based assemblies were longer in length (range: 1.12-1.41 Gb) than short-read assemblies (0.91-1.08 Gb). Assembly length was strongly correlated with the amount of repeat content, with longer genomes showing much higher levels of repeat content than typically reported for the avian order Passeriformes. Repeat content for the Bell's sparrow (31.2% of genome) was the highest level reported to date for a songbird genome assembly and was more in line with woodpecker (order Piciformes) genomes. CR1 LINE elements retained from an expansion that occurred 25-30 million years ago were the most abundant TEs in the song sparrow genome. Although the other five sparrow species also exhibit evidence for a spike in CR1 LINE activity at 25-30 million years ago, LTR elements stemming from more recent expansions were the most abundant elements in these species. LTRs were uniquely abundant in the Bell's sparrow genome

deriving from two recent peaks of activity. Higher levels of repeat content (79.2-93.7%) were found on the W chromosome relative to the Z (20.7-26.5) or autosomes (16.1-30.9%). These patterns support a dynamic model of transposable element expansion and contraction underpinning the seemingly constrained and small sized genomes of birds. Our work highlights how the resolution of difficult-to-assemble regions of the genome with new sequencing technologies promises to transform our understanding of avian genome evolution.

KEYWORDS: Passerellidae, transposable elements, genome size, California Conservation Genomics Project, C-value.

INTRODUCTION:

The dynamics of transposable element (TE) activity within host genomes are a major driver of genome evolution (Ågren & Wright 2011). Transposable elements proliferate throughout the genome either through copy-and-paste mechanisms (Class I elements; e.g. long interspersed nuclear elements [LINEs]) or through cut-and-paste mechanisms (class II elements; e.g. DNA transposons). The mobility of these elements in the genome contributes to structural variation (e.g. indels, inversions), alterations to gene expression, and the evolution of gene regulatory networks (Feschotte 2008; Schrader & Schmitz 2018). Given the genomic disruption potentially caused by TEs, the majority of new TE insertions are likely deleterious and species exhibit a wide range of defense mechanisms to both silence and delete TEs from the genome (Goodier 2016). Nevertheless, co-option of TEs by the host genome has led to the evolution of novel phenotypes (Mi et al. 2000; Cornelis et al. 2017) including color polymorphisms (van't Hof et al. 2016; Kratochwil et al. 2022), increased

52 immunity (Brosh et al. 2022), insecticide resistance (Daborn et al. 2002), and speciation (Serrato-Capuchina & Matute 2018). To date,
53 much of TE biology has focused on model organisms with well-characterized genomic resources. The generation of high-quality
54 genomes for a diversity of non-model organisms (Teeling et al. 2018; Feng et al. 2020; Rhie et al. 2021; Lewin et al. 2022) promises to
55 broaden our understanding of how co-evolutionary dynamics between TEs and their host shape genome evolution.

56 Avian genomes provide an illuminating case of how expanding the diversity of available genome assemblies has altered our
57 understanding of TE dynamics. Among amniotes, birds exhibit the smallest and most constrained genomes. Although contraction of
58 avian genomes likely began prior to the evolution of flight (Organ et al. 2007), the high metabolic demand of flight is the leading
59 hypothesis for continued constraint on avian genome size evolution (Hughes & Hughes 1995; Andrews et al. 2009; Wright et al. 2014).
60 Consistent with a hypothesis of constrained genome evolution, the first avian genomes sequenced revealed low repeat content (<10%),
61 little recent TE activity, and high chromosomal stability (Ellegren 2010). Detailed TE annotation of an increasing diversity of avian
62 genome assemblies has since challenged the early narrative of low repeat content and high stability. First, comparative analyses across
63 12 avian genomes showed that the apparent stability in avian genome size was actually the product of a more dynamic history of genomic
64 expansions offset by large-scale deletions (Kapusta et al. 2017). Second, extensive variation in the timing and proliferation of TE
65 elements has been discovered across birds (Kapusta & Suh 2017; Suh et al. 2018; Galbraith et al. 2021). This includes the discovery of
66 relatively high repeat content of 20-30% in the orders Piciformes (woodpeckers and allies) and Bucerotiformes (hornbills and hoopoes;
67 Zhang et al. 2014; Manthey et al. 2018; Feng et al. 2020). Third, novel TEs have been discovered in avian lineages that derive from
68 horizontal gene transfer from filarial nematodes (Suh et al. 2016). Finally, highly contiguous assemblies have confirmed that previous

challenges to assembling the W chromosome were due in part to its role as a refugium for long terminal repeat (LTR) retrotransposons (Peona et al. 2021; Warmuth et al. 2022).

Our understanding of TE dynamics in avian genomes is poised to advance further with the increased use of long-read sequencing technologies (Kapusta & Suh 2017; Rhie et al. 2021). Repetitive regions of the genome, including centromeres, telomeres, and the W chromosome, are a major source of assembly gaps. Consequently, repetitive DNA is thought to make up a large proportion of the 7-42% of the genomic DNA missing from short-read genome assemblies relative to flow cytometry or densitometry estimates of genome size (hereafter the C-value; Peona et al. 2018). Indeed, a recent comparison of assembly methods for the Paradise Crow (*Lycocorax pyrrhopterus*) showed that gaps in short-read assemblies were primarily caused by LTR retrotransposons and simple repeats (Peona et al. 2020). Further, a recent comparison of activity levels of the chicken repeat 1 (CR1) retrotransposon across 117 avian genomes found a relationship between assembly contiguity (scaffold N50) and number of full length CR1s identified in individual genomes (Galbraith et al. 2021). This pattern was found across all genomes analyzed and also explained intra-generic variation in CR1 insertions. Detailed TE annotation of highly contiguous genomes will be essential for overcoming the confounding influence of assembly quality on patterns of TE diversity. In particular, studies leveraging highly contiguous genomes to explore TE dynamics across shorter evolutionary time scales are lacking but will be essential for understanding the contributions of these elements to the generation of avian diversity.

To this end, we performed in-depth TE annotations of highly contiguous genomes generated from six closely related sparrow species in the family Passerellidae. Passerellidae sparrows are a diverse clade of oscine Passeriformes, with 132 recognized species that are found throughout the Americas from northern Canada to southern Chile (Winkler et al. 2020). We generated *de novo* genome

assemblies for Bell's sparrow (*Artemisiospiza belli*), Savannah sparrow (*Passerculus sandwichensis*), and song sparrow (*Melospiza melodia*) for this paper as part of the California Conservation Genomics Project (CCGP; Shaffer et al. 2022). We analyze these assemblies alongside three genomes recently sequenced by the Vertebrate Genomes Project (VGP; Rhie et al. 2021) for saltmarsh (*Ammospiza caudacutus*), Nelson's (*Ammospiza nelsoni*), and swamp sparrow (*Melospiza georgiana*), for a study of the genomic basis of tidal marsh adaptation. These new genome assemblies come from six members of the 'grassland' sparrow clade (Klicka et al. 2014). True to their name, all species can be generally found in a variety of shrub and grassland habitats across North America. Savannah and song sparrow are the two most ecologically and geographically widespread species occupying a broad range of tundra, alpine, meadow, prairie, marsh, and shrub habitats from Alaska and northern Canada south through Mexico to Guatemala (Arcese et al. 2002; Wheelwright & Rising 2008). Bell's sparrow is found primarily in more arid chaparral and coastal sage habitat from northwestern California south into Baja California, Mexico and east into the southern San Joaquin valley and Mojave desert of southeastern California (Cicero & Koo 2012). Nelson's and Swamp sparrow can primarily be found in central to eastern North America, principally in marsh habitats (Shriver et al. 2018; Herbert & Mowbray 2019). Saltmarsh sparrow is exclusively found in tidal marsh habitats of the Atlantic coast, and Nelson's, swamp, song and Savannah sparrow all include tidal marsh specialist subspecies (Greenberg et al. 2006; Walsh et al. 2019a).

Song and Savannah sparrow are two of the most polytypic North America bird species, with 25 and 17 subspecies described in the song (Patten & Pruett 2009) and Savannah sparrow (Wheelwright & Rising 2008), respectively. In general, subspecific divergence across ecological gradients has long made all six species the focus of geographic variation and speciation studies (Marshall 1948; Aldrich

1984; Rising 2001; Cicero & Johnson 2006; Walsh et al. 2017; Mikles et al. 2020; Walsh et al. 2019b, 2021; Clark et al. 2022). The Bell's sparrow also forms a narrow hybrid zone with the sagebrush sparrow (*Artemisiospiza nevadensis*) in the Owen's Valley of eastern California (Cicero & Johnson 2007; Cicero & Koo 2012), while Nelson's and saltmarsh sparrow hybridize along the coast of southern Maine (Rising & Avise 1993; Shriver et al. 2005; Walsh et al. 2015). Additionally, studies of these six sparrow species have provided important insights into avian life history and demography (Nice 1937; Johnston 1954; Keller et al. 1994; Keller & Arcese 1998; Marr et al. 2002; Freeman-Gallant et al. 2005; Ruskin et al. 2017a,b; Field et al. 2018), physiology (Poulson 1965; Greenberg et al. 2012; Benham & Cheviron 2020), vocal learning and behavior (Marler & Peters 1977; Searcy & Marler 1981; Williams et al. 2022), and migratory behavior (Moore et al. 1978; Able & Able 1996). The generation of highly contiguous reference genomes for these sparrow species with in-depth TE annotations will thus provide a critically important resource for future research in this intensively studied clade.

In addition to the six new sparrow genome assemblies, nine other assemblies were analyzed from across the Passerellidae . The previous assemblies were produced using a variety of short- and long-read sequencing approaches. Previously sequenced genomes also include short-read assemblies for both the song and saltmarsh sparrows, which allows for intra-specific comparisons to assess the impact of sequencing technology on repeat annotation. We take advantage of the diverse genomic resources available from within this single avian family to ask: **(1)** what is the impact of sequencing technology and assembly completeness on TE element annotation? And **(2)** how do the evolutionary dynamics of TEs vary among closely related sparrow species. Addressing these questions will be important for determining how different sequencing approaches may introduce bias into comparative genomics analyses. In addition, our comparisons provide insights into how analyses based on short-read assemblies may miss important dynamics of avian genome evolution.

METHODS:

CCGP genome sampling:

We sequenced liver from an adult female Bell's sparrow (*Artemisiospiza belli canescens*) collected on 25 June 2018 at Hunter Cabin, 1.5 Mi. east of Jackass Spring, Death Valley National Park, Inyo Co., California (36.54758ºN, 117.48786ºW; elevation: 6860 ft.). Blood and liver tissue samples were sequenced from an immature female song sparrow (*Melospiza melodia gouldii*) captured on 5 September 2020 in oak woodland habitat at Mitsui Ranch, Sonoma Mountain, Sonoma Co., California (38.33131ºN; 122.57720ºW). Liver tissue for sequencing was collected from an adult male Savannah sparrow (*Passerculus sandwichensis alaudinus*) on 20 May 2015 in tidal marsh habitat at the San Francisco Bay National Wildlife Refuge, Santa Clara Co., California (37º 26.029'N; 122º 0.996'W: elevation: 4 m). Bell's and song sparrow individuals were collected with approval of California Department of Fish and Wildlife (CDFW permit #: SCP-458), the U.S. Fish and Wildlife Service (USFWS permit #: MB153526), Death Valley National Park (Bell's sparrow only; permit#: DEVA-2015-SCI-0040), and following protocols approved by the University of California, Berkeley IACUC (AUP-2016-04-8665-1). The Savannah sparrow sample was also collected with approval from CDFW (permit #: SCP-012913), USFWS (permit #: MB24360B-0), the San Francisco Bay National Wildlife Refuge (special use permit: 2015-015), and using methods approved by the University of Illinois, Urbana-Champaign IACUC (protocol #: 13418). Voucher specimens were deposited at the Museum of Vertebrate Zoology, Berkeley, CA for the Bell's sparrow (https://arctos.database.museum/guid/MVZ:Bird:192114) and song sparrow

(https://arctos.database.museum/guid/MVZ:Bird:193390). A specimen voucher of the Savannah sparrow was deposited at the Field Museum of Natural History in Chicago, IL (FMNH:Birds:499929).

DNA extraction, library preparation, and sequencing for CCGP genomes:

High molecular weight (HMW) genomic DNA (gDNA) for PacBio HiFi library preparation was extracted from 27 mg and 15 mg of liver tissue from the Bell's and Savannah sparrow samples, respectively. Extractions were performed using the Nanobind Tissue Big DNA kit following the manufacturer's instructions (Pacific BioSciences - PacBio, Menlo Park, CA). For song sparrow, HMW gDNA was isolated from whole blood preserved in EDTA. A total of 30µl of whole blood was added to 2 ml of lysis buffer containing 100mM NaCl, 10 mM Tris-HCl pH 8.0, 25 mM EDTA, 0.5% (w/v) SDS and 100µg/ml Proteinase K. Lysis was carried out at room temperature for a few hours until the solution was homogenous. The lysate was treated with 20µg/ml RNase A at 37ºC for 30 minutes and cleaned with equal volumes of phenol/chloroform using phase lock gels (Quantabio Cat # 2302830). DNA was precipitated by adding 0.4X volume of 5M ammonium acetate and 3X volume of ice-cold ethanol. The DNA pellet was washed twice with 70% ethanol and resuspended in an elution buffer (10mM Tris, pH 8.0), and purity was estimated using absorbance ratios (260/280 = 1.81-1.84 and 260/230 = 2.29-2.40) on a NanoDrop ND-1000 spectrophotometer. The final DNA yield (Bell's: 13 µg; Savannah: 16 µg; song: 150 µg total) was quantified using the Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay; Promega, Madison, WI). The size distribution of the HMW DNA was estimated using the Femto Pulse system (Agilent, Santa Clara, CA): 62% of the fragments were

>140 Kb for Bell's sparrow; 60% of the fragments were >140 Kb for Savannah sparrow; and 85% of the DNA was found in fragments >120 Kb for song sparrow.

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 following the manufacturer's protocols (Pacific Biosciences - PacBio, Menlo Park, CA; Cat. #100-938-900). HMW gDNA was sheared to a target DNA size distribution between 15-20 Kb and concentrated using 0.45X of AMPure PB beads (PacBio Cat. #100-265-900) for the removal of single-strand overhangs at 37ºC for 15 minutes. Enzymatic steps of DNA damage repair were performed at 37ºC for 30 minutes, followed by the end repair and A-tailing steps at 20ºC for 10 minutes and 65ºC for 30 minutes. Ligation of overhang adapter v3 was performed at 20ºC for 60 minutes with subsequent heating to 65ºC for 10 minutes to inactivate the ligase. Finally, DNA product was nuclease treated at 37ºC for 1 hour. To collect fragments greater than 9 Kb, the resulting SMRTbell library was purified and concentrated with 0.45X Ampure PB beads (PacBio, Cat. #100-265-900) for size selection using the BluePippin system (Sage Science, Beverly, MA; Cat #BLF7510). The 15-20 Kb average HiFi SMRTbell library was sequenced at the University of California Davis DNA Technologies Core (Davis, CA) using two 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-hour movies each on a PacBio Sequel II sequencer.

The Omni-C library was prepared using the Dovetail™ Omni-C™ Kit (Dovetail Genomics, CA) according to the manufacturer's protocol with slight modifications. First, specimen tissue was ground thoroughly with a mortar and pestle while cooled with liquid nitrogen. Subsequently, chromatin was fixed in place in the nucleus and then passed through 100 µm and 40 µm cell strainers to remove large debris. Fixed chromatin was digested under various conditions of DNase I until a suitable fragment length distribution of DNA

molecules was obtained. Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed and the DNA purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. An NGS library was generated using an NEB Ultra II DNA Library Prep kit (NEB, Ipswich, MA) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The library was sequenced at the Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq platform (Illumina, San Diego, CA) to generate over 100 million 150 bp paired end reads per species. See supplemental Table S1 for details on PacBio and Illumina sequencing.

Assembly of CCGP genomes:

We assembled the genome of the three CCGP sparrows following the CCGP assembly pipeline Version 3.0 (see Lin et al. 2022; Supplemental Table S2). The pipeline takes advantage of long and highly accurate PacBio HiFi reads alongside chromatin capture Omni-C data to produce high quality and highly contiguous genome assemblies while minimizing manual curation.

In brief, we removed remnant adapter sequences from the PacBio HiFi dataset for all three assemblies using HiFiAdapterFilt (Sim et al. 2022) and obtained the initial dual assembly with the filtered PacBio reads using HiFiasm (Cheng et al. 2021). The dual assembly consists of a primary and alternate assembly: the primary assembly is more complete and consists of longer phased blocks, while the alternate consists of haplotigs (contigs with the same haplotype) in heterozygous regions and is more fragmented. Given the

characteristics of the latter, it cannot be considered a complete assembly of its own but rather is a complement of the primary assembly (https://lh3.github.io/2021/04/17/concepts-in-phased-assemblies, https://www.ncbi.nlm.nih.gov/grc/help/definitions/).

Next, we identified sequences corresponding to haplotypic duplications, contig overlaps and repeats on the primary assembly with purge_dups (Guan et al 2020) and transferred these sequences to the corresponding alternate assembly. We aligned the Omni-C data to both assemblies following the Arima Genomics Mapping Pipeline (https://github.com/ArimaGenomics/mapping_pipeline) and used SALSA to produce scaffolds for the primary assembly (Ghurye et al. 2017, Ghurye et al. 2019).

Omni-C contact maps for the primary assembly were produced by aligning the Omni-C data with BWA-MEM (Li 2013), identifying ligation junctions, and generating Omni-C pairs using pairtools (Golobordko et al. 2018). We generated a multi-resolution Omni-C matrix with cooler (Abdennur an Mirny 2020) and balanced it with hicExplorer (Ramírez et al. 2018). To visualize and check contact maps for mis-assemblies, we used HiGlass (Kerpedjiev et al. 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/PretextSnapshot). In detail, if we identified a strong off-diagonal signal in the proximity of a join that was made by the scaffolder, and a lack of signal in the consecutive genomic region, we dissolved it by breaking the scaffolds at the coordinates of the join. After this process, no further manual joins were made. Some of the remaining gaps (joins generated by the scaffolder) were closed using the PacBio HiFi reads and YAGCloser (https://github.com/merlyescalona/yagcloser). We checked for contamination using the BlobToolKit Framework (Challis et al. 2020). Finally, upon submission of the assemblies to NCBI, we trimmed remnants of sequence adaptors and mitochondrial contamination identified during NCBI's own contamination screening.

We assembled the mitochondrial genomes for each of the sparrows from their corresponding PacBio HiFi reads starting from the same mitochondrial sequence of *Zonotrichia albicollis* (NCBI:NC_053110.1; Feng et al. 2020, B10K Project Consortium) and using the reference-guided pipeline MitoHiFi (Uliano-Silva et al 2021; Allio et al. 2020). After completion of the nuclear genomes, we searched for matches of the resulting mitochondrial assembly sequence in their nuclear genome assembly using BLAST+ (Camacho et al. 2009), filtering out contigs and scaffolds from the nuclear genome with a sequence identity >99% and a size smaller than the mitochondrial assembly sequence.

Genome assembly assessment:

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). GenomeScope2.0 (Ranallo-Benavidez et al. 2020) was used to estimate genome features including genome size, heterozygosity, and repeat content from the resulting k-mer spectrum. To obtain general contiguity metrics, we ran QUAST (Gurevich et al. 2013). Genome quality and completeness were quantified with BUSCO (Manni et al. 2021) using the 8,338 genes in the Aves ortholog database (aves_odb10) for the three CCGP genomes. Assessment of base level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and merqury (Rhie et al. 2020). We further estimated genome assembly accuracy via a BUSCO gene set frameshift analysis using the pipeline described in Korlach et al. (2017). We followed the quality metric nomenclature established by Rhie et al. (2021), with the genome quality code x.y.Q.C, where, x = log10[contig NG50]; y = log10[scaffold NG50]; Q = Phred base accuracy QV (quality value); C = % genome represented by the first ‘n’ scaffolds, following a known karyotype of 2n=74 for *P.*

sandwichensis (Bird Chromosome database - V3.0/2022; Degrandi et al 2020) and estimated 2n=80 for both *M. melodia* and *A. belli;* this estimation is the median number of chromosomes from closely related species (Genome on a Tree - GoaT; tax_tree(1729112); Challis et al. 2023). Quality metrics for the notation were calculated on the primary assemblies. Finally, we used the JupiterPlot pipeline (https://github.com/JustinChu/JupiterPlot) to visualize higher level synteny between the scaffolds of each sparrow assembly and chromosomes of the zebra finch (*Taeniopygia guttata*) genome assembly (Warren et al. 2010). Scaffolds representing 80% of each draft assembly (ng=80) were mapped to zebra finch chromosomes exceeding 1Mb in length (m=1000000).

VGP genome sampling:

The three CCGP were compared genomes to three chromosome-level assemblies of closely related sparrows generated by the Vertebrate Genomes Project following protocols outlined in Rhie et al. (2021). These samples were sequenced from blood samples (100 – 200 µl in ethanol) taken from female Nelson's and swamp sparrows and a saltmarsh sparrow that was identified as a female in the field but lacked the W chromosome in the final assembly. The Nelson's sparrow sample was collected by Nicole Guido (Saltmarsh Habitat and Avian Research Program) in South Branch Marsh River, Waldo County, Maine (44.5864ºN; 68.8591ºW) on 31 July 2020. The swamp sparrow sample was collected by Jonathan Clark (University of New Hampshire) in Durham, Rockingham County, New Hampshire (43.14ºN; 71.00ºW) on 23 July 2020. The saltmarsh sparrow sample was collected by Chris Elphick (University of Connecticut) from Barn Island Wildlife Management Area, New London County, Connecticut (41.338ºN; 71.8677ºW) on August 19 2020. Sample collection occurred under permits of the Maine Division of Inland Fisheries and Wildlife (#2020-314), Connecticut

Department of Energy and Environmental Protection (#0221012b), and New Hampshire Fish and Game, and followed protocols approved under the University of New Hampshire IACUC (#190401). All individuals were released at the capture location immediately after sampling. DNA extracted from these three species was sequenced to 31.6-34.6x coverage of PacBio Sequel II HiFi long reads, 254-450x coverage of Bionano Genomics DLS, and 103-112x coverage for Arima Hi-C v2. Genome assemblies were generated from these data with the VGP standard assembly pipeline version 2.0, which included hifiasm v0.15.4, purge_dups v. 1.2.5, solve v. 3.6.1, salsa v. 2.3. Further details on raw data, sequence evaluations and curated assemblies can be found on VGP Genome Ark pages for the Nelson's sparrow (https://genomeark.github.io/genomeark-all/Ammospiza_nelsoni.html), saltmarsh sparrow (https://genomeark.github.io/genomeark-all/Ammodramus_caudacutus.html), and swamp sparrow (https://genomeark.github.io/genomeark-all/Melospiza_georgiana.html).

Genome size variation within sparrows:

We estimated the distribution of genome sizes in Passerellidae sparrows with c-value data from 33 individuals of 21 species archived in the Animal Genome Size Database (Gregory 2022). The majority of these C-value estimates were generated using Feulgen image analysis densitometry with original values reported in Andrews et al. (2009) and Wright et al. (2014). This includes all species for which we compared C-values to assembly lengths. C-value estimates of genome size were converted from picograms to base pairs using a conversion of 1pg = 0.978 Gb (Dolezel et al. 2003). We additionally obtained genome assembly length from nine publicly available assemblies that were sequenced previously for members of the Passerellidae. Six of these were based on Illumina short-read

sequence data and include a second song sparrow genome from Alaska (Louha et al. 2019), a short-read genome of the saltmarsh sparrow (*Ammospiza caudacuta*; Walsh et al. 2019a), plus genomes for the white-throated sparrow (*Zonotrichia albicollis*; Tuttle et al. 2016), dark-eyed junco (*Junco hyemalis*; Friis et al. 2022), chipping sparrow (*Spizella passerina*; Feng et al. 2020), and grasshopper sparrow (*Ammodramus savannarum;* Carneiro 2021). The remaining three assemblies were generated using PacBio long-read sequence data and include a third song sparrow genome from British Columbia (Feng et al. 2020), a white-crowned sparrow genome (*Zonotrichia leucophrys*; unpublished), and a contig-level assembly of the California towhee (*Melozone crissalis;* Black et al. 2023). For complete GenBank accession details, sequencing, and assembly methods for these genomes see supplemental Table S3. We compared C-value estimates of genome size and genome assembly length for all of the assemblies that included both these estimates of genome size (10 of 15 total).

Repeat annotation:

We performed detailed *de novo* repeat annotation and manual curation of repeat libraries for the Bell's, song, and Savannah sparrow genomes sequenced by the CCGP in the program RepeatModeler2 with the ltrstruct option selected to improve identification of LTR elements (Flynn et al. 2020). Consensus transposable element libraries generated from RepeatModeler2 were then curated manually following protocols and methods of Goubert et al. (2022). First, we removed any redundancy in the *de novo* repeat libraries using cd-hit-est (Li & Godzik 2006) to cluster any consensus sequences together that were 80 base pairs in length and shared >80% similarity over more than 80% of their length. This corresponds to the 80-80-80 rule of Wicker et al. (2007) frequently used to classify

TE elements as a single family. We then prioritized elements for manual curation that were at least 1000 base pairs in length and had at least 10 blastn hits in the genome assembly. For each consensus sequence prioritized for manual curation, we used blastn (Camacho et al. 2009) to identify other members of each TE family in the genome, and for each blastn hit we added 2000 bp of flanking sequence to both ends of the sequence. We then aligned the extended sequences using mafft (Katoh & Standley 2013) and removed gaps automatically using T-coffee (Notredame et al. 2000). The multiple sequence alignment produced by mafft was visualized in aliview (Larsson et al. 2014), and the termini of each element were identified based on canonical motifs of different element classes (e.g. 5' TG and 3' CA dinucleotides in LTR elements). A consensus sequence of the trimmed multiple sequence alignment was then generated using the cons tool in EMBOSS (Rice et al. 2000). Finally, the program TE-Aid (https://github.com/clemgoub/TE-Aid) was used to confirm structural properties and the presence of open reading frames for the expected proteins characteristic of each class of TE element. This process of blast, extension, and alignment was repeated iteratively for each element until termini were discovered. Following manual curation of TE sequences, we again used cd-hit-est with the same settings as above to cluster sequences belonging to the same family. The final set of manually curated sequences was then compared against a library of avian TE elements downloaded from repbase as well as other recently published TE datasets (e.g. *Dromaius novaehollandiae*, Peona et al. 2020) using cd-hit-est and the 80-80-80 rule above as a threshold to classify curated elements as belonging to previously identified TE families. We assigned the following species-specific prefixes for newly identified repeat elements in each sparrow species: pasSan (*Passerculus sandwichensis*), melMel (*Melospiza melodia*), and artBel (*Artemisiospiza belli*). For new TE families shared across two or more of the sparrow species we assigned the

prefix Passerellidae. For each element the prefix was followed by the superfamily identity (e.g., LTR/ERV1). For elements where we could not confidently identify the complete consensus sequence we added the suffix .inc.

Curated TE libraries for all three sparrow species were merged into a single Passerellidae repeat library that was then used to annotate transposable element diversity in the genome assemblies of the three CCGP sparrow genomes using RepeatMasker v. 4.1.2 (Smit et al. 2015). We additionally used the *de novo* Passerellidae TE library to annotate the genome assemblies of the 3 VGP sparrow genomes and the 9 previously sequenced sparrow genomes available on Genbank. Secondly, for seven sparrow species with a contig N50 >1Mb and at least a scaffold level assembly, we performed separate RepeatMasker runs on the autosomes, Z chromosome, and W chromosome (if present). For the CCGP genomes, scaffolds were assigned to these different chromosomes based on homology with the Zebra Finch chromosomes using Minimap2 (Li 2018). Finally, we assessed temporal patterns of TE activity for these seven sparrow species. For autosomal, Z, and W chromosomes, we used the calcDivergenceFromAlign.pl script in the RepeatMasker package to estimate the Kimura 2-parameter (K2P) distance of each TE element from the consensus sequence. K2P distances were used to generate barplots for the LTR, SINE, LINE, and DNA classes of TE elements found in the genome. For autosomal loci, we additionally used a mutation rate estimate of 2.3×10^{-9} per generation from another Passerine species (*Ficedula albicollis,* Smeds et al. 2016) to estimate the approximate timing of repeat activity.

Sparrow phylogeny construction:

To further contextualize the history of transposable element proliferation across sparrows, we constructed a time-calibrated phylogeny using ultra-conserved elements (UCE) extracted from all available sparrow genomes (14 taxa including 12 species and 3 song sparrow subspecies) as well as the medium ground finch (*Geospiza fortis*), island canary (*Serinus canaria*), and zebra finch (*Taeniopygia guttata*) as outgroup taxa. We used the UCE-5k-probe-set and the phyluce pipeline (Faircloth 2016) to align probes to each genome, extended sequences by 1000 bp on either flank, aligned sequences using mafft v. 7.49 (Katoh & Standley 2013), and produced a final PHYLIP file of the UCEs. We used RAxML-NG (Stamatakis 2014) on the CIPRES science portal (Miller et al. 2010) to generate a maximum likelihood phylogeny of the concatenated UCE loci using the GTRCAT model, rapid bootstrapping, and the autoMRE bootstrapping criterion. Secondly, we used MCMCtree to estimate a time-calibrated phylogeny based on the topology generated by RAxML-NG. We used the fossil sparrow *Ammodramus hatcheri* (Steadman 1981) to calibrate the node between the grasshopper sparrow (*Ammodramus savannarum*) and all other sparrows. Following Oliveros et al. (2019), we set the calibration time to 12 Mya bounded by a minimum age of 7.5 Mya and a maximum age of 18.6 Mya. We implemented a model of independent rates among branches and drawing from a log-normal distribution. We performed two independent runs of MCMCtree with each run starting from a different random seed. The first 50,000 iterations were removed as burnin before running for another 100 million iterations with a sample taken every 1000 iterations. We assessed convergence of the two runs using tracer v. 1.6.0 (Rambaut et al. 2018).

RESULTS:

Genome assemblies:

The three CCGP assemblies included a low of 337 scaffolds in the Savannah sparrow and a high of 1,339 scaffolds in Bell's sparrow; total assembly lengths ranged from 1.15-1.40 Gb (Table 1; Supplemental figures S1-S3). All assembly metrics indicate that the genomes are highly contiguous with contig N50 ranging from 5.98-8.31 Mb and scaffold N50 from 17.08-25.78 Mb. The largest contig length was over 32.13 Mb and the longest scaffold over 99.81 Mb. Over 96% of the genes in the avian orthologous database were found to be complete and single copy in BUSCO. These metrics indicate that the genomes generated *de novo* by the CCGP pipeline are in line with overall contiguity and completeness metrics for the three genomes generated by the VGP. Genomes for swamp, saltmarsh, and Nelson's sparrow showed similar contig N50 ranging from 8.25-12.04 Mb, but scaffold N50s were approximately 3x as large (74.25-78.44 Mb). The VGP genomes were also assembled into chromosome-level assemblies with 36-40 chromosomes identified based on decreasing order of size. Finally, BUSCO scores were slightly higher for the VGP genomes, exceeding 98% complete and single copy genes in all three species, but this could also reflect the different databases used in BUSCO analyses of the CCGP and VGP genomes. Jupiter plots showed CCGP sparrow scaffolds mapping to most chromosomes of the zebra finch genome, with little evidence for inversions or translocations that may be indicative of misassemblies (Fig. 1). Similarly, although contact maps for the primary assemblies of the three CCGP genomes show some level of fragmentation, they also reveal little evidence for inversions or translocations. (Supplementary Figure S4). Given their greater contiguity, we only describe the primary assemblies here, but the sequences corresponding to both primary and alternate assemblies for each of the CCGP species are available on NCBI (See supplemental Table S4 and Data availability for details).

Genome size variation in sparrows:

Adjusted genome size estimates from C-values of Passerellidae sparrows varied by 0.5 Gb from 1.13 Gb in Savannah sparrow (*Passerculus sandwichensis*) to 1.63 Gb in gray-browed brushfinch (*Arremon assimilis*), with a mean of 1.36 Gb (Fig. 2a). Previous short-read genome assemblies of sparrows varied in length from 0.91 Gb in the grasshopper sparrow to 1.05 Gb in the white-throated sparrow assembly, which were 0.16 to 0.42 Gb smaller than the corresponding C-value estimates of genome size for these species. Assembly lengths for the CCGP and VGP sparrow genomes varied from 1.15 Gb in the Savannah sparrow to 1.40 Gb in the Bell's sparrow (Fig. 2a). Recently released long-read assemblies of the white-crowned sparrow (1.12 Gb) and California towhee (1.41 Gb) span a similar range. The length of the assemblies reported here closely approximated the C-value estimates of genome size for the Savannah sparrow (mean 1.18 Gb vs. 1.15 Gb assembly) and the song sparrow (mean 1.40 Gb vs. 1.35 Gb assembly), but was more divergent in the swamp sparrow (1.46 Gb vs. 1.16 Gb assembly). No C-value estimates exist for the other three sparrow species; however, alternate estimates of genome size are available from the kmer profiles analyzed in GenomeScope (supplemental Fig. S1-S3; https://www.genomeark.org/genomeark-all/). These profiles suggest that the Nelson's (assembly: 1.18 Gb; GenomeScope: 1.19 Gb) and saltmarsh sparrow (1.24 vs. 1.22) assemblies closely match the expected genome length estimated from the kmer profile. In contrast, all three CCGP sparrow genome assemblies exceed the kmer profile genome size estimates (for example Bell's sparrow assembly: 1.40 Gb vs. GenomeScope estimate: 1.13); whereas, the curated swamp sparrow assembly (1.16 Gb) was considerably shorter than the estimated length from GenomeScope (1.33 Gb). Together these data underscore the high level of completeness of the assemblies generated using long-read approaches, with less than 3% of the genome missing from most species. The swamp sparrow assembly appears to be an

354 exception with 12-20% of the genome content potentially missing (depending on GenomeScope or C-value estimate). Purged repeat
355 content may explain some of the missing data from the swamp sparrow assembly. Repeat content in the swamp sparrow was estimated
356 to span 18.26% of the genome using our Passerellidae repeat library in RepeatMasker, while k-mer estimates of repeat content were
357 30.4% from GenomeScope. In contrast, prior sparrow genome assemblies that used primarily short-read data were inferred to be missing
358 as much as 12-30% of the genomic DNA (Fig. 2b).
359
360 *Passerellidae de novo repeat library:*
361 *De novo* identification of transposable elements in RepeatModeler2 followed by manual curation led to the identification of 514,
362 704, and 650 TE families within the Savannah, song, and Bell's sparrows, respectively. Merging of the three sparrow libraries produced
363 a final Passerellidae TE library with 1,272 elements. This includes 361 elements shared by two or more sparrow species and 234, 341,
364 and 336 elements unique to Savannah, song, and Bell's sparrows, respectively. Similar to other avian species, LINE and LTR elements
365 represent the majority of TEs identified in these sparrow species. These include 15 shared LINE families and 68 shared LTR families
366 across all three species. Song sparrows had the most unique LINE elements (n=58), whereas Bell's sparrow had the most unique LTR
367 elements (n=122). Savannah sparrow had the least number of both elements identified (supplemental Fig. S5).
368 The curated Passerellidae repeat library was used to annotate all 15 sparrow assemblies. Annotation results from RepeatMasker
369 showed that repeat content comprises a considerably higher percentage of the genome in the more contiguous, long-read assemblies
370 (Fig. 3a). Bell's sparrow, song sparrow, and California towhee showed the greatest proportion of repeat content with repeats comprising

371 over 29% of the genome (Table 2; Fig. 3a). The Savannah sparrow exhibited the lowest proportion of repeats (16.5%) among the long-
372 read assemblies, but this still exceeded the 6.5-10.3% of the genome covered by repeat content identified in other sparrow species.
373 Indeed, we found contig N50 to be highly predictive of total repeat content discovered in genome assemblies. All eight sparrow
374 assemblies with a contig N50 greater than 1 Mb had significantly more repeat content than assemblies with contig N50 less than 1 Mb
375 (Fig. 3b; t=-6.174; df=7.8; p=0.0003). All assemblies with a contig N50 > 1 Mb were assembled using PacBio long-reads; whereas only
376 1 of 7 of the assemblies with contig N50 less than 1 Mb included PacBio long-read sequencing in the assembly. Variation in assembly
377 length among species also strongly predicted repeat content (adjusted R^2=0.95, p-value <<0.0001; Fig. 4c). This pattern was replicated
378 among different assemblies of the song sparrow and saltmarsh sparrow. Repeat content increased from 7.1% (0.978 Gb assembly) to
379 10.3% (1.06 Gb assembly) to 29.5% (1.36 Gb assembly) as assembly length increased in the three song sparrow assemblies. In the
380 saltmarsh sparrow, repeat content more than doubled from 10.6% in the short-read assembly (1.07 Gb) to 24.2% in the long-read
381 assembly (1.24 Gb). Finally, the amount of repeat content significantly decreased as the percent missing DNA increased for each
382 assembly with missing DNA inferred from the difference between C-value estimate and assembly length (adjusted R^2=0.57, p-value =
383 0.007; Fig. 3d).

384 The comparison among the three song sparrow genome assemblies showed that LINEs and LTRs comprised the greatest number
385 and total base pairs of newly discovered TE sequence (Fig. 4a). We found an additional 8,375 (a ~5% increase) line elements in the
386 California versus British Columbia song sparrow assemblies. Despite only a small increase in the total number of elements, we find that
387 these LINE elements span an additional 155 Mb of DNA in the California song sparrow genome (Fig. 4b). This discrepancy likely stems

388 both from different elements segregating at different frequencies in each population and LINEs in the California genome being of greater
389 length on average. One of the most abundant LINE elements in the California song sparrow genome was found across all three of the
390 CCGP sparrow genomes but was missing from the British Columbia song sparrow genome. This element was >6500 bp in length and
391 nearly 4000 full length copies were found across the California song sparrow genome. Comparisons between a short-read and long-read
392 assembly of the saltmarsh sparrow revealed a similar pattern (Fig. 4c-d). A small increase in the total number of LINE (~3%) and LTR
393 (~16%) elements led to respective increases of 43.1 Mb and 43.8 Mb of TE DNA discovered in these genomes. These results further
394 support the inference that missing DNA from previous assemblies corresponds to longer TE elements and may have been a major
395 contributor to gaps.

396 The prevalence of different TE classes varied across our three genome assemblies. LTRs were the most abundant element within
397 all sparrow genomes except the song sparrow assembly where LINE elements were the most abundant (14.58%; Table 2). Across
398 different chromosomes, the density of repeat content (79.23-93.73%) was highest on the W chromosomes. W chromosome repeat
399 content was particularly high in the genus *Melospiza* with over 90% of the W chromosome spanned by repetitive elements in both the
400 song and swamp sparrow. Z chromosome repeat content tended to be higher than autosomal repeat content for all species except song
401 and Bell's sparrow (Table 2).

402

403 *Timing of repeat proliferation:*

We extracted an average of 4,663 (range: 3,699-4,839) ultra-conserved element (UCE) loci from the 17 reference genomes queried (Supplemental Table S5). From these loci we constructed a concatenated data matrix of 4,196 UCE loci shared across 95% of samples with a total of 4,815,326 base pairs. The concatenated maximum likelihood tree was well-resolved with all nodes receiving bootstrap support of 100. This topology was used as input into MCMCtree to estimate divergence times among the focal species (see supplemental Figure S6 for full phylogeny). This time-calibrated phylogeny indicated that the white-crowned sparrow split from the other sequenced sparrow species at 13.3 Mya (95% HPD: 7.4-17.8; Fig. 5), Bell's sparrow diverged from other species in the grassland sparrow clade 7.9 Mya (95% HPD: 4.5-10.8), *Ammospiza* sparrows (Nelson's and saltmarsh) split from Savannah, swamp, and song sparrow 6.8 Mya (95% HPD: 3.8-9.3), and Savannah sparrow diverged from the *Melospiza* sparrows 5.8 Mya (95% HPD: 3.3-7.9). Within the context of these divergence times, members of the Passerellidae family show sharply divergent histories of transposable element proliferation (Fig. 5). All members of the grassland sparrow clade show evidence for a spike in LINE element activity in the autosomes ~25-30 Ma. In contrast, the white-crowned sparrow does not show evidence for this spike, but rather shows a normal distribution of LINE element divergence centered at ~40-50 Ma. Although the timing of LINE proliferation in grassland sparrows appears to predate divergence estimates among Passerellidae species (Fig. 5), the contrast with white-crowned sparrow suggests it may have occurred more recently following divergence of these different sparrow lineages. Despite shared evidence for this period of LINE activity, song sparrows have retained more LINE elements from this proliferation (~14% of the genome) than the other five sparrow species (only 1-3% of genomes). Bell's sparrow shows a unique pulse of LTR proliferation approximately 12 Mya and a very recent (<5 Mya) proliferation of both LINE and LTR elements in the autosomes. For species with an assembled W chromosome, all show a steady

accumulation of LTR elements on the W chromosome, with endogenous retroviruses being the most prolific and representing up to a maximum of 69.2% in the song sparrow.

DISCUSSION:

Highly contiguous and complete genomes reveal high repeat content:

We generated highly contiguous and complete genomes of three sparrow species in the family Passerellidae that we compared with three chromosome-level genomes generated by the Vertebrate Genome Project. Contig N50 for the three newly generated genomes exceeded 92%, and the scaffold N50 exceeded 85% of all avian genomes recently surveyed by Bravo et al. (2021). Assembly length for the six species analyzed here also exceeds assembly lengths for all short-read based assemblies generated to date (1.16-1.40Gb vs. 0.91-1.05Gb). The longer length of these assemblies more closely approximates independent estimates of genome size from Feulgen image analysis densitometry (C-value), with song and Savannah sparrow missing only 2-3% of genomic sequence relative to C-value size estimates. Longer assemblies were also associated with greater levels of repeat content. The high percentage of total interspersed repeats discovered in the song sparrow (31.2%), Bell's sparrow (29.5%), and California towhee (30.9%; also see Black et al. 2023) genomes are the highest levels ever reported for Passeriformes and more closely resemble levels of repeat content described in the avian orders Piciformes and Bucerotiformes (Manthey et al. 2018; Feng et al. 2020). Although our finding is a novel result for passerine genome assemblies, reassociation kinetic studies found about 36% of the dark-eyed junco genome to be repetitive DNA (Shields & Straus 1975). Recent long-read assemblies for jays in the passerine family Corvidae also show repeat content in-line with the results presented here

(Benham et al. 2023; DeRaad et al. 2023). Further, high levels of repeat content in these sparrows matches predictions that much of the missing genomic data from avian short-read assemblies are likely repetitive DNA (Elliot & Gregory 2015; Kapusta & Suh 2017; Peona et al. 2018). We expect that the generation of additional highly contiguous and complete genomes using third generation sequencing technology will also find higher levels of repeat content in avian genomes than previously appreciated.

Previous sparrow genome assemblies generated using short-read methods were found to be missing ~12-30% of DNA sequence relative to C-values (Fig. 2b). The majority of this missing DNA is likely associated with highly repetitive regions of the genome that caused gaps in prior assemblies. Gaps associated with repeat regions is a well-established phenomenon and recent comparisons among sequencing technologies point to long contiguous reads as essential for spanning these gaps (Rhie et al. 2021). Similarly, we find that sparrow assemblies generated using PacBio long-read sequence data exhibited elevated contig N50 and higher percent repeat content, and that percent repeat content decreased in assemblies inferred to be missing a greater percentage of DNA (Fig. 3). This was also true for intra-specific comparisons of multiple song and saltmarsh sparrow genomes where repeat content increased with assembly length and contiguity. The diversity of transposable elements can vary significantly within and among populations (e.g. *Ficedula* flycatchers; Suh et al. 2018). Although we compared song sparrow genomes from three different subspecies that could differ in repeat content and genome size, the patterns for song sparrow are consistent with work showing increased TE discovery and abundance in long-read versus short-read assemblies of the same individual (Peona et al. 2021). Our results are also consistent with prior work linking metrics of genome contiguity with levels of repeat content detected in genome assemblies (Galbraith et al. 2021). Overall, our analyses among

454 closely related sparrow species confirms that sequencing technology appears to be a major confounding factor in comparative research
455 on TE diversity.
456
457 *Genome size evolution in birds:*
458 Transposable elements are widely recognized as an important driver of genome size evolution across Eukaryotes (Kidwell 2002;
459 Elliott & Gregory 2015). Low repeat content and high synteny contributed to an early view that avian genomes were relatively stable
460 and constrained in size as an adaptation for the metabolic demands of flight (Hughes & Hughes 1995; Wright et al. 2014). This
461 perspective has been challenged with evidence of a more dynamic history of avian genome expansions followed by large-scale deletions
462 (Kapusta et al. 2017). Genome size variation in the Passerellidae, as measured by densitometry and cytometry methods, ranges from
463 1.13-1.63 Gb, with the Savannah sparrow possessing the smallest genome of any sparrow measured to date. Differences in assembly
464 length across all sparrow species were entirely related to repeat content (Fig. 3). Further, TE composition differed significantly even in
465 species with similar genome assembly lengths (e.g. Bell's and song sparrow).
466 Unlike the other sparrow species analyzed, CR1 LINE elements were found to be the most abundant TE class within the song
467 sparrow genome. The TE landscape of the song sparrow genome indicates that the majority of LINE DNA stems from a period of
468 increased activity 25-30 million years ago. All six sparrow species in the grassland clade show a spike in LINE activity during this
469 period, but much of the LINE DNA from this period was eliminated in species other than the song sparrow. In contrast, the white-
470 crowned sparrow shows a more 'bell-curve' shape of LINE element proliferation with a peak of less than 0.5% at ~30-40 Mya (about

471 20% divergence from consensus). The white-crowned sparrow pattern more closely resembles TE landscapes observed in other
472 Passerifomes birds such as *Ficedula* flycatchers (Muscicapidae; Suh et al. 2017) and Estrildidae finches (Boman et al. 2019). These
473 patterns point to a proliferation of LINE elements within the grassland sparrow clade that more likely occurred after divergence from
474 the white-crowned sparrow ~13.3 Mya. This discrepancy in the timing of activity could reflect the use of a genome wide estimate of
475 mutation rate from *Ficedula* flycatchers (Smeds et al. 2016) that may be underestimating the true mutation rate for Passerellidae sparrows
476 and/or transposable elements. Indeed, many of the host genome's defense mechanisms against TE proliferation involve DNA-editing
477 enzymes, such as APOBECs, which mutate TE sequences to silence their activity in the genome (Goodier et al. 2016; Knisbacher &
478 Levanon 2016).

479 LTR elements were the most abundant TE within all sparrow genomes except the song sparrow. Proliferation of these elements
480 has also been more recent, beginning ~12 million years ago and continuing to the present. Recent proliferation of LTR elements was
481 especially pronounced in the Bell's sparrow. Recent proliferation of LTR elements more closely aligns with patterns of TE expansion
482 observed in the zebra finch (Kapusta & Suh 2017) and the blackcap (*Sylvia atricapilla*; Bours et al. 2023). The reasons for recent LTR
483 expansions in songbirds versus other avian lineages (e.g. Chicken; Warren et al. 2017) are not entirely clear. One possibility is that
484 competition for similar genomic insertion sites between LTR and LINE elements could be mediated by host defenses. Recent work in
485 the deer mouse (*Peromyscus maniculatus;* Gozashti et al. 2023) provided evidence for a cycle initiated by greater host repression of
486 ancient endogenous retroviruses (ERV, a type of LTR) that allowed for greater LINE proliferation in the genome. This was followed by
487 the invasion of the deer mouse genome by a novel ERV that was hypothesized to have a greater immunity to host defense mechanisms

and greater potential to outcompete LINE elements for insertion sites. A related possibility could reflect the accumulation of LTR elements on the W chromosome, many of which remain transcriptionally active and could seed invasions of the autosomal chromosomes (see below; Peona et al. 2021). Whether either of these scenarios contributes to the recent expansions of LTR elements in songbird genomes awaits further study; however, the expanding number of avian genomes assembled using long-read sequence data will be essential for understanding the dynamics of TE proliferation and deletion in the evolution of avian genomes.

Differences in genome size and repeat content could be the result of a number of different mechanisms involved in TE silencing, deletion, or expansion in host genomes (Goodier et al. 2016). A wide range of epigenetic mechanisms exist to silence TE activity in plants and animals (reviewed in Slotkin & Martienssen 2007). In birds, methylation of CpG and non-CpG sites in TEs with DNA methyltransferases is the primary mechanism of TE silencing that has been documented (Derks et al. 2016; Kapusta & Suh 2017). Mutating TE sequences is another mechanism hosts deploy to defend against TE proliferation. APOBEC genes induce C-to-U mutations in retrotransposons leading to inactivation and degradation of these elements. The genomes of zebra finch and other bird species exhibit signatures of high APOBEC activity (Knisbacher & Levanon 2015). An important mechanism for the removal of LTR elements from the genome is ectopic recombination. This process deletes most of the element sequence leaving only a single LTR and correlates with recombination rate variation across the genome in birds (Ji & DeWoody 2016). Finally, demographic differences among populations could influence TE dynamics, with TEs predicted to insert and spread more rapidly in populations with a small effective population size (Ne; Lynch & Conery 2003). Demographic analyses of the different sparrow species provide some support for this hypothesis as both Nelson's and saltmarsh sparrows have been inferred to experience historical bottlenecks and lower Ne than other species (Walsh et al.

505 2019a,b; Walsh et al. 2021). In contrast, the Savannah sparrow has the lowest repeat content and has been inferred to maintain high and
506 constant effective population sizes (Benham & Cheviron 2019), which may be important for combating TE proliferation and maintaining
507 a smaller genome. However, the relationship between Ne and TE proliferation is not necessarily straightforward (Whitney & Garland
508 2010) and some authors argue that TE expansions may be even more likely in species with large Ne (Ågren & Wright 2011).

509 Disruption of a host's TE repression mechanisms can lead to TE expansions. Stressful conditions (e.g. thermal stress) can disrupt
510 epigenetic silencing of TEs in the host genome, leading to TE expansions (Capy et al. 2000, Slotkin & Martienssen 2007). Furthermore,
511 co-evolutionary arms races between TEs and the host genomes could lead to divergence in TE repressors among populations or closely
512 related species. Subsequent hybridization among these lineages could allow TEs to escape their repressors and proliferate throughout
513 the genome of hybrids (Bingham et al. 1982; Serrato-Capuchina & Matute 2018). Examples of TE re-activation following hybridization
514 have been documented in both plants (Josefsson et al. 2006) and animals (O'Neill et al. 1998). In hybrid *Helianthus* sunflower species,
515 proliferation of LTR elements was found to contribute significantly to a 50% increase in the genome size of hybrids relative to parental
516 species (Ungerer et al. 2006). Intriguingly, the Bell's sparrow individual used to generate the reference assembly for this project comes
517 from the same subspecies known to hybridize with sagebrush sparrow in a contact zone centered ca. 120-150 km. to the northwest of
518 the collecting locality. Whether recent hybridization between Bell's and sagebrush or Nelson's and saltmarsh sparrow lineages led to a
519 TE expansion remains to be determined. However, the dynamic patterns of genome size evolution within sparrows indicates that the
520 Passerellidae are an exciting model for future research on the dynamics of TE evolution.
521

TE element proliferation on sex chromosomes:

The potential deleterious effects of TE insertions is thought to explain a general trend of TE prevalence in regions of lower recombination rate (Rizzon et al. 2002; Ji & DeWoody 2016; Kent et al. 2017). These patterns are especially pronounced on the Y/W sex chromosomes where a lack of recombination, low gene density, and small effective population sizes are thought to allow for TE accumulation (Charlesworth & Langly 1988; Bachtrog 2003). This high TE abundance is thought to be a major contributing factor to the challenges of sequencing and assembling the W chromosome in birds and Y chromosome in mammals (Tomaskiewicz et al. 2017). Consistent with these expectations for the non-recombining W chromosome, we found that repeat content on the W chromosome was dramatically higher across all four female birds sequenced relative to autosomal or Z chromosomes. Interspersed repeats comprised 79.2% and 82.6% of the Nelson's and Bell's sparrow W chromosome, respectively, while the song and swamp sparrow (both *Melospiza*) possessed W chromosomes with over 90% repeat content. Previous reports of repeat content on the W chromosome range from 22% in the emu (*Dromaius novaehollandiae*; Peona et al. 2021) to over 84% in the hooded crow (e.g *Corvus cornix*; Warmuth et al. 2022) and 89% in the Steller's jay (e.g. *Cyanocitta stelleri;* Benham et al. 2023). Similar to other avian W chromosome assemblies, endogenous retroviral elements are the dominant element representing 42.3% of the song to 69.2% of the Bell's sparrow W chromosome assembly. Peona et al. (2021) also showed that a disproportionately large percentage of LTR elements on the avian W chromosome are full length retroviral elements that continue to be actively transcribed. The capacity of active elements to spread from the W to other regions of the genome makes the W chromosome a likely source for the recent activity and abundance of endogenous retroviral elements in Passeriformes (Warren et al. 2010; Zhang et al. 2014; Warmuth et al. 2022). Kapusta & Suh (2017) posited that the abundance of these

539 elements in Passeriformes may have played critical roles in their high levels of diversification. The highly complete genome assemblies
540 generated via third generation sequencing techniques will provide new opportunities to test this hypothesis.
541
542 *Conclusions:* Here we report on the release of three highly contiguous assemblies of sparrows in the family Passerellidae. The
543 combination of long-read and Omni-C technology enabled the generation of nearly complete and highly contiguous assemblies. Analysis
544 of these genomes revealed a previously underappreciated abundance of repetitive elements in the genomes of songbirds and suggests
545 that much of the missing data from other avian assemblies are likely comprised of repeat content. As third generation sequencing
546 technologies become the standard in avian genome assembly, the dynamics of TE element proliferation and genome size evolution
547 across different evolutionary timescales will become better understood. Our results point to the strong role repetitive element
548 proliferation and deletion plays in the dynamics of avian genome size evolution, even among closely related species.
549
550 **FUNDING:** This work was supported by the California Conservation Genomics Project, with funding provided to the University of
551 California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224]. Additional funding was provided by
552 National Science Foundation Grant #1826777.
553
554 **ACKNOWLEDGEMENTS:**

We thank Joshua Ho for assistance with tissue subsampling, and Nicole Guido, Jonathan Clark and Chris Elphick for sample collection. PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the UC Davis DNA Technologies and Expression Analysis Cores and the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high quality sequence data. We thank Erich Jarvis and personnel at the Vertebrate Genomes Project for conducting the sequencing and assembly of the Nelson's, saltmarsh, and swamp sparrow.

DATA AVAILIBILITY:

Data generated for this study are available under NCBI BioProject PRJNA720569. Raw sequencing data for the Savannah sparrow sample FMNH:Bird:499929 (NCBI BioSample SAMN24839580) are deposited in the NCBI Short Read Archive (SRA) under SRS12336030. Raw sequencing data for the song sparrow sample MVZ:Bird:193390 (NCBI BioSample SAMN24817870, SAMN24817871) are deposited in the NCBI SRA under SRS12452128. Raw sequencing data for the Bell's sparrow sample MVZ:Bird:192114 (NCBI BioSample SAMN24224802) are deposited in the NCBI SRA under SRS11988259. See supplemental Table S3 for the GenBank accession, BioProject, and BioSample numbers associated with the VGP and other genomes analyzed. For the

571 CCGP genomes, assembly scripts and other data for the analyses presented can be found at the following GitHub repository:
572 www.github.com/ccgproject/ccgp_assembly. Transposable element library, UCE sequences, and other supplemental data and code can
573 be found on Dryad: https://doi.org/10.5061/dryad.cjsxksncs

574

575 **REFERENCES:**

576 Abdennur, N., and L. A. Mirny. 2020. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics,
577 36(1), 311–316. doi: 10.1093/bioinformatics/btz540

578 Able, K. P., & Able, M. A. (1996). The flexible migratory orientation system of the Savannah sparrow (Passerculus sanadwichensis).
579 *Journal of Experimental Biology*, *199*, 3–8. https://doi.org/10.1016/S0003-3472(05)80791-2

580 Ågren, J. A., & Wright, S. I. (2011). Co-evolution between transposable elements and their hosts: A major factor in genome size
581 evolution? *Chromosome Research*, *19*(6), 777–786. https://doi.org/10.1007/s10577-011-9229-0

582 Aldrich, J. W. (1984). Ecogeographical Variation in Size and Proportions of Song Sparrows (Melospiza melodia). *Ornithological*
583 *Monographs*, (35), iii–134. https://doi.org/10.2307/40166779

584 Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F.

2020. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources, 20(4), 892–905. doi: 10.1111/1755-0998.13160

Andrews, C. B., Mackenzie, S. A., & Gregory, T. R. (2009). Genome size and wing parameters in passerine birds. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1654), 55–61. https://doi.org/10.1098/rspb.2008.1012

Arcese, P., M. K. Sogge, A. B. Marr, and M. A. Patten (2020). Song Sparrow (*Melospiza melodia*), version 1.0. In Birds of the World (A. F. Poole and F. B. Gill, Editors). Cornell Lab of Ornithology, Ithaca, NY, USA. https://doi.org/10.2173/bow.sonspa.01

Bachtrog, D. (2003). Accumulation of Spock and Worf, two novel non-LTR retrotransposons, on the neo-Y chromosome of Drosophila miranda. *Molecular Biology and Evolution*, *20*(2), 173–181. https://doi.org/10.1093/molbev/msg035

Benham, P. M., & Cheviron, Z. A. (2019). Divergent mitochondrial lineages arose within a large, panmictic population of the Savannah sparrow (Passerculus sandwichensis). *Molecular Ecology*, *28*(7), 1765–1783. https://doi.org/10.1111/mec.15049

Benham, P. M., & Cheviron, Z. A. (2020). Population history and the selective landscape shape patterns of osmoregulatory trait divergence in tidal marsh Savannah sparrows (Passerculus sandwichensis). *Evolution*, *74*(1), 57–72. https://doi.org/10.1111/evo.13886

598 Benham, P.M., Cicero, C., DeRaad, D.A., McCormack, J.E., Wayne, R.K., Escalona, M., Beraut, E., Marimuthu, M.P.A., Nguyen, O.,
599 Nachman, M.W., Bowie, R.C.K. 2023. A highly contiguous reference genome for the Steller's hay (*Cyanocitta stelleri*). *Journal*
600 *of Heredity*, accepted article. early view: https://doi.org/10.1093/jhered/esad042

601 Bingham, P. M., Kidwell, M. G., & Rubin, G. M. (1982). The molecular basis of P-M hybrid dysgenesis: The role of the P element, a
602 P-strain-specific transposon family. *Cell*, *29*(3), 995–1004. https://doi.org/10.1016/0092-8674(82)90463-9

603 Black, A., Yoon, J., McCreedy, C., Janjua, S., Heenkenda, E., Mathur, S., Ferree, E., Fesnock, A., Hernandez, A., DeWoody, A.
604 Conservation genomics of California towhee (*Melozone crissalis*) in relation to the official list of endangered and threatened
605 wildlife. *Authorea.* May 10, 2023. DOI: 10.22541/au.168371288.85881657/v1

606 Bours, A., Pruisscher, P., Bascón-Cardozo, K., Odenthal-Hesse, L., & Liedvogel, M. (2023). The blackcap (Sylvia atricapilla) genome
607 reveals a species-specic accumulation of LTR retrotransposons. *Scientific Reports*, 1–17. https://doi.org/10.1038/s41598-023-
608 43090-1

609 Bravo, G. A., Schmitt, C. J., & Edwards, S. V. (2021). What Have We Learned from the First 500 Avian Genomes? *Annual Review of*
610 *Ecology, Evolution, and Systematics*. https://doi.org/10.1146/annurev-ecolsys-012121-085928

611 Brosh, O., Fabian, D. K., Cogni, R., Tolosana, I., Day, J. P., Olivieri, F., … Jiggins, F. M. (2022). A novel transposable element-
612 mediated mechanism causes antiviral resistance in Drosophila through truncating the Veneno protein. *Proceedings of the*
613 *National Academy of Sciences*, *119*(29), 1–10. https://doi.org/10.1073/pnas.2122026119

614 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and
615 applications. *BMC Bioinformatics*, *10*, 1–9. https://doi.org/10.1186/1471-2105-10-421

616 Capy, P., Gasperi, G., Biémont, C., & Bazin, C. (2000). Stress and transposable elements: Co-evolution or useful parasites? *Heredity*,
617 *85*(2), 101–106. https://doi.org/10.1046/j.1365-2540.2000.00751.x

618 Carneiro, C.M. (2021) Genomic insight into the demographic history and structure of the grasshopper sparrow (*Ammodramus*
619 *savannarum*). M.sc. thesis, University of Florida, Gainesville.

620 Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive quality assessment of genome
621 assemblies. bioRxiv, doi: 10.1101/844852

622 Challis, R., Kumar, S., Sotero-Caio, C., Brown, M., & Blaxter, M. (2023). Genomes on a Tree (GoaT): A versatile, scalable search
623 engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Research*, *8*, 1–27.
624 https://doi.org/10.12688/wellcomeopenres.18658.1

625 Charlesworth, B., & Langley, C. H. (1989). The population genetics of Drosophila transposable elements. *Annual Review of Genetics*,
626 *23*, 251–287. https://doi.org/10.1146/annurev.ge.23.120189.001343

627 Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Robust haplotype-resolved assembly of diploid individuals
628 without parental data. arXiv:2109.04785. 2021. Available: http://arxiv.org/abs/2109.04785

629 Chesser, T.R., Banks, R. C., Keith Barker, F., Cicero, C., Dunn, J. L., Kratter, R. W., … Winker, K. (2013). Fifty-fourth supplement
630 to the American ornithologists' union check-list of north American birds. *Auk*, *130*(3), 558–571.
631 https://doi.org/10.1525/auk.2013.130.3.1

632 Cicero, C., & Johnson, N. K. (2006). Diagnosability of subspecies: Lessons from Sage Sparrows (Amphispiza belli) for analysis of
633 geographic variation in birds. *Auk*. https://doi.org/10.1642/0004-8038(2006)123[0266:DOSLFS]2.0.CO;2

634 Cicero, C., & Johnson, N. K. (2007). Narrow contact of desert sage sparrows (Amphispiza belli nevadensis and A. B. Canescens) in
635 owens valley, Eastern California: Evidence from mitochondrial DNA, morphology, and gis-based niche models. *Ornithological*
636 *Monographs*, *63*(63), 78–95. https://doi.org/10.1642/0078-6594(2007)63[78:NCODSS]2.0.CO;2

637 Clark, J. D., Benham, P. M., Maldonado, J. E., Luther, D. A., & Lim, H. C. (2022). Maintenance of local adaptation despite gene flow
638 in a coastal songbird. *Evolution*, 1–14. https://doi.org/10.1111/evo.14538

639 Cornelis, G., Funk, M., Vernochet, C., Leal, F., Tarazona, O. A., Meurice, G., … Roberts, R. M. (2017). An endogenous retroviral
640 envelope syncytin and its cognate receptor identified in the viviparous placental Mabuya lizard. *Proceedings of the National*
641 *Academy of Sciences of the United States of America*, *114*(51), E10991–E11000. https://doi.org/10.1073/pnas.1714590114

642 Daborn, P. J., Yen, J. L., Bogwitz, M. R., Le Goff, G., Feil, E., Jeffers, S., … Ffrench-Constant, R. H. (2002). A single P450 allele
643 associated with insecticide resistance in Drosophila. *Science*, *297*(5590), 2253–2256. https://doi.org/10.1126/science.1074170

644 Degrandi T, M, Barcellos S, A, Costa A, L, Garnero A, D, V, Hass I, Gunski R, J: Introducing the Bird Chromosome Database: An
645 Overview of Cytogenetic Studies in Birds. Cytogenet Genome Res 2020. doi: 10.1159/000507768

646 Derks, M. F. L., Schachtschneider, K. M., Madsen, O., Schijlen, E., Verhoeven, K. J. F., & van Oers, K. (2016). Gene and
647 transposable element methylation in great tit (Parus major) brain and blood. *BMC Genomics*, *17*(1), 1–13.
648 https://doi.org/10.1186/s12864-016-2653-y

649 Dolezel, J., Bartos, J., Voglmayr, H., & Greilhuber, J. (2003). Letter to the editor: nuclear DNA content and genome size of trout and
650 human. *Cytometry*, *51A*(2), 127–128. https://doi.org/10.1002/cyto.a.10013

651 Eisermann, K., Avendaño, C., & Matías, E. (2017). Nesting evidence, density and vocalisations in a resident population of Savannah
652 Sparrow Passerculus sandwichensis wetmorei in Guatemala. *Bulletin of the British Ornithologists' Club*, *137*(1), 37–45.
653 https://doi.org/10.25226/bboc.v137i1.2017.a4

654 Ellegren, H. (2010). Evolutionary stasis: the stable chromosomes of birds. *Trends in Ecology and Evolution*, *25*(5), 283–291.
655 https://doi.org/10.1016/j.tree.2009.12.004

656 Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content.
657 *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1678). https://doi.org/10.1098/rstb.2014.0331

658 Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786–788.
659 https://doi.org/10.1093/bioinformatics/btv646

660 Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., … Zhang, G. (2020). Dense sampling of bird diversity increases
661 power of comparative genomics. *Nature*, *587*(7833), 252–257. https://doi.org/10.1038/s41586-020-2873-9

662 Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405.
663 https://doi.org/10.1038/nrg2337

664 Field, C.R, K. J. Ruskin, B. Benvenuti, A. Borowske, J.B. Cohen, L. Garey, T.P. Hodgman, R.A. Kern, E. King, A.R. Kocek, A.I.
665 Kovach, K.M. O'Brien, B.J. Olsen, N. Pau, S.G. Roberts, E. Shelly, W.G. Shriver, J. Walsh, and C.S. Elphick. 2018.
666 Quantifying the importance of geographic replication and representativeness when estimating demographic rates, using a coastal
667 species as a case study. *Ecography 41: 971–981*. DOI: *10.1111/ecog.02424*

668 Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated
669 genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of*
670 *America*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

671 Freeman-Gallant, C. R., Wheelwright, N. T., Meiklejohn, K. E., States, S. L., & Sollecito, S. V. (2005). Little effect of extrapair
672 paternity on the opportunity for sexual selection in Savannah sparrows (Passerculus sandwichensis). *Evolution*, *59*(2), 422–430.
673 https://doi.org/10.1111/j.0014-3820.2005.tb01000.x

674 Friis, G., Vizueta, J., Ketterson, E. D., & Milá, B. (2022). A high-quality genome assembly and annotation of the dark-eyed junco
675 Junco hyemalis, a recently diversified songbird. *G3: Genes, Genomes, Genetics*, *12*(6). https://doi.org/10.1093/g3journal/jkac083

676 Galbraith, J. D., Kortschak, R. D., Suh, A., & Adelson, D. L. (2021). Genome Stability Is in the Eye of the Beholder: CR1
677 Retrotransposon Activity Varies Significantly across Avian Diversity. *Genome Biology and Evolution*, *13*(12), 1–14.
678 https://doi.org/10.1093/gbe/evab259

679 Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C.-S. 2017. Scaffolding of long read assemblies using long range contact
680 information. BMC Genomics, 18(1), 527. doi: 10.1186/s12864-017-3879-z
681 Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., … Koren, S. 2018. Integrating Hi-C links with assembly graphs
682 for chromosome-scale assembly. PLoS Computation Biology 15(8): e1007273. doi: 10.1371/journal.pcbi.1007273
683 Goloborodko, A., N. Abdennur, S. Venev, H. B. Brandao, and G. Fudenberg. 2018. mirnylab/pairtools: v0.2.0. doi:
684 10.5281/zenodo.1490831

685 Goodier, J. L. (2016). Restricting retrotransposons: A review. *Mobile DNA*, *7*(1). https://doi.org/10.1186/s13100-016-0070-z

686 Goubert, C., Craig, R. J., Bilat, A. F., Peona, V., Vogan, A. A., & Protasio, A. V. (2022). A beginner's guide to manual curation of
687 transposable elements. *Mobile DNA*, *13*(1), 1–19. https://doi.org/10.1186/s13100-021-00259-7

688 Gozashti, L., Feschotte, C., & Hoekstra, H. E. (2023). Transposable Element Interactions Shape the Ecology of the Deer Mouse
689 Genome. *Molecular Biology and Evolution*, *40*(4), 1–17. https://doi.org/10.1093/molbev/msad069

690 Greenberg, R., Cadena, V., Danner, R. M., & Tattersall, G. (2012). Heat loss may explain bill size differences between birds
691 occupying different habitats. *PLoS ONE*, *7*(7), 1–9. https://doi.org/10.1371/journal.pone.0040933

692 Gregory, T.R. (2022). Animal Genome Size Database. http://www.genomesize.com.

693 Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. 2020. Identifying and
694 removing haplotypic duplication in primary genome assemblies. Bioinformatics, 36(9), 2896–2898. doi:
695 10.1093/bioinformatics/btaa025
696 Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics,
697 29(8), 1072–1075. doi: 10.1093/bioinformatics/btt086

698 Hughes, A. L., & Hughes, M. K. (1995). Small genomes for better flyers. *Nature*, *377*(6548), 391. https://doi.org/10.1038/377391a0

699 Ji, Y., & DeWoody, J. A. (2016). Genomic Landscape of Long Terminal Repeat Retrotransposons (LTR-RTs) and Solo LTRs as
700 Shaped by Ectopic Recombination in Chicken and Zebra Finch. *Journal of Molecular Evolution*, *82*(6), 251–263.
701 https://doi.org/10.1007/s00239-016-9741-0

702 Johnston, R. F. (1954). Variation in Breeding Season and Clutch Size in Song Sparrows of the Pacific Coast. *The Condor*, *56*(5), 268–
703 273. https://doi.org/10.2307/1364850

704 Josefsson, C., Dilkes, B., & Comai, L. (2006). Parent-Dependent Loss of Gene Silencing during Interspecies Hybridization. *Current*
705 *Biology*, *16*(13), 1322–1328. https://doi.org/10.1016/j.cub.2006.05.045

706 Kapusta, A., & Suh, A. (2017). Evolution of bird genomes—a transposon's-eye view. *Annals of the New York Academy of Sciences*,
707 *1389*(1), 164–185. https://doi.org/10.1111/nyas.13295

708 Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National*
709 *Academy of Sciences*, 201616702. https://doi.org/10.1073/pnas.1616702114

710 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and
711 usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

712 Keller, L. F., & Arcese, P. (1998). No evidence for inbreeding avoidance in a natural population of song sparrows (Melospiza
713 melodia). *American Naturalist*, *152*(3), 380–392. https://doi.org/10.1086/286176

714 Keller, L. F., Arcese, P., Smith, J. N. M., Hochachka, W. M., & Stearns, S. C. (1994). Selection against inbred song sparrows during a
715 natural population bottleneck. *Nature*, *372*(6504), 356–357. https://doi.org/10.1038/372356a0

716 Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical*
717 *Transactions of the Royal Society B: Biological Sciences*, *372*(1736). https://doi.org/10.1098/rstb.2016.0458

718 Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., … Gehlenborg, N. 2018. HiGlass: web-based
719 visual exploration and analysis of genome interaction maps. Genome Biology, 19(1), 125. doi: 10.1186/s13059-018-1486-1

720 Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, *115*(1), 49–63.
721 https://doi.org/10.1023/A:1016072014259

722 Klicka, J., Keith Barker, F., Burns, K. J., Lanyon, S. M., Lovette, I. J., Chaves, J. A., & Bryson, R. W. (2014). A comprehensive
723 multilocus assessment of sparrow (Aves: Passerellidae) relationships. *Molecular Phylogenetics and Evolution*, *77*(1), 177–182.
724 https://doi.org/10.1016/j.ympev.2014.04.025

725 Knisbacher, B. A., & Levanon, E. Y. (2016). DNA editing of LTR retrotransposons reveals the impact of APOBECs on vertebrate
726 genomes. *Molecular Biology and Evolution*, *33*(2), 554–567. https://doi.org/10.1093/molbev/msv239

727 Korlach, J., Gedman, G., Kingan, S. B., Chin, C.-S., Howard, J. T., Audet, J.-N., …Jarvis, E. D. 2017. De novo PacBio long-read and
728 phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. GigaScience,
729 6(10), 1–16. doi: 10.1093/gigascience/gix085

730 Kratochwil, C. F., Kautt, A. F., Nater, A., Härer, A., Liang, Y., Henning, F., & Meyer, A. (2022). An intronic transposon insertion
731 associates with a trans-species color polymorphism in Midas cichlid fishes. *Nature Communications*, *13*(1), 296.
732 https://doi.org/10.1038/s41467-021-27685-8

733 Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–
734 3278. https://doi.org/10.1093/bioinformatics/btu531

735 Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. Retrieved from
736 http://arxiv.org/abs/1303.3997

737 Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.
738 https://doi.org/10.1093/bioinformatics/bty191

739 Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences.
740 *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

741 Lin M, Escalona M, Sahasrabudhe R, Nguyen O, Beraut E, Buchalski MR, Wayne RK. 2022. A reference genome assembly of the
742 bobcat, Lynx rufus. J Hered. 113:615-623

743 Louha, S., Ray, D. A., Winker, K., & Glenn, T. C. (2020). A High-Quality Genome Assembly of the North American Song Sparrow,
744 Melospiza melodia. *G3: Genes|Genomes|Genetics*, g3.400929.2019. https://doi.org/10.1534/g3.119.400929

745 Lynch, M., & Conery, J. S. (2003). The Origins of Genome Complexity. *Science*, *302*(5649), 1401–1404.
746 https://doi.org/10.1126/science.1089370

747 Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined
748 Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.
749 *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

750 Manthey, J. D., Moyle, R. G., Boissinot, S., Dhabi Abu Dhabi, A., Manthey Stéphane Boissinot, J. D., & Dhabi, A. (2018). Multiple
751 and independent phases of transposable element amplification in the genomes of Piciformes (woodpeckers and allies), *10*(June),
752 1–35. https://doi.org/10.1093/gbe/evy105/5020728

753 Marr, A. B., Keller, L. F., & Arcese, P. (2002). Heterosis and Outbreeding Depression in Descendants of Natural Immigrants To an
754 Inbred Population of Song Sparrows (Melospiza Melodia). *Evolution*, *56*(1), 131–142. https://doi.org/10.1554/0014-
755 3820(2002)056[0131:haodid]2.0.co;2

756 Marshall, J. T. (1948). Ecologic races of song sparrows in the San Francisco Bay Region: Part II. Geographic Variation. *The Condor*,
757 *50*(6), 233–256. Retrieved from https://sora.unm.edu/sites/default/files/journals/condor/v050n06/p0233-p0256.pdf

758 Mi, S. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, *403*(February),
759 785–789. Retrieved from www.nature.com

760 Mikles, C. S., Aguillon, S. M., Chan, Y. L., Arcese, P., Benham, P. M., Lovette, I. J., & Walsh, J. (2020). Genomic differentiation and
761 local adaptation on a microgeographic scale in a resident songbird. *Molecular Ecology*, *29*(22), 4295–4307.
762 https://doi.org/10.1111/mec.15647

763 Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees.
764 In *2010 gateway computing environments workshop (GCE)* (pp. 1–8).

765 Moore, F. R. (1978). Sunset and the orientation of a nocturnal migrant bird. *Nature*, *274*(5667), 154–156.
766 https://doi.org/10.1038/274154a0

767 Nice, M. M. (1937). Studies in the life history of the song sparrow. I. A population study of the song sparrow. *Transactions of the*
768 *Linnaean Society of New York,* 4:1-247.

769 Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment.
770 *Journal of Molecular Biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

771 O'Neill, R., O'Neill, M. & Graves, J. (1998). Undermethylation associated with retroelement activation and chromosome remodelling
772 in an interspecific mammalian hybrid. *Nature* 393, 68–72. https://doi.org/10.1038/29985

773 Oliveros, C. H., Field, D. J., Ksepka, D. T., Keith Barker, F., Aleixo, A., Andersen, M. J., … Faircloth, B. C. (2019). Earth history and
774 the passerine superradiation. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(16), 7916–
775 7925. https://doi.org/10.1073/pnas.1813206116

776 Organ, C. L., Shedlock, A. M., Meade, A., Pagel, M., & Edwards, S. V. (2007). Origin of avian genome size and structure in non-
777 avian dinosaurs. *Nature*, *446*(7132), 180–184. https://doi.org/10.1038/nature05621

778 Patten, M. A., & Pruett, C. L. (2009). The Song Sparrow, Melospiza melodia, as a ring species: Patterns of geographic variation, a
779 revision of subspecies, and implications for speciation. *Systematics and Biodiversity*, *7*(1), 33–62.
780 https://doi.org/10.1017/S1477200008002867

781 Peona, V., Weissensteiner, M. H., & Suh, A. (2018). How complete are "complete" genome assemblies?—an avian perspective.
782 *Molecular Ecology Resources*, *18*(6), 1188–1195. https://doi.org/10.1111/1755-0998.12933

783 Peona, V., Palacios-Gimenez, O. M., Blommaert, J., Liu, J., Haryoko, T., Jønsson, K. A., … Suh, A. (2021). The avian W
784 chromosome is a refugium for endogenous retroviruses with likely effects on female-biased mutational load and genetic
785 incompatibilities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1833).
786 https://doi.org/10.1098/rstb.2020.0186

787 Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., … Suh, A. (2021). Identifying the causes and consequences of
788 assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*, *21*(1), 263–286.
789 https://doi.org/10.1111/1755-0998.13252

790 Poulson, T. L. (1965). Countercurrent multipliers in avian kidneys. *Science*, *148*(3668), 389–391.
791 https://doi.org/10.1126/science.148.3668.389

792 Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics
793 using Tracer 1.7. *Systematic Biology*, *67*(5), 901–904. https://doi.org/10.1093/sysbio/syy032

794 Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., … Manke, T. 2018. High-resolution TADs reveal
795 DNA sequences underlying genome organization in flies. Nature Communications, 9. doi: 10.1038/s41467-017-02525-w
796 Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of
797 polyploid genomes. Nature Communications, 11(1), 1432. doi: 10.1038/s41467-020-14998-3
798 Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. 2020. Merqury: reference-free quality, completeness, and phasing assessment
799 for genome assemblies. Genome Biology, 21(1), 245. doi:10.1186/s13059-020-02134-9

800 Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., … Jarvis, E. D. (2021). Towards complete and error-free
801 genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

802 Rice, P., Longden, L., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*,
803 *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

804 Rising, James D. (2001). Geographic variation in size and shape of Savannah Sparrows (Passerculus sandwichensis). *Studies in Avian*
805 *Biology* (23):1-65.

806 Rizzon, C., Marais, G., Gouy, M., & Biémont, C. (2002). Recombination Rate and the Distribution of Transposable Elements in the
807 Drosophila melanogaster Genome . *Genome Research*, *12*(3), 400–407. https://doi.org/10.1101/gr.210802

808 Ruskin, K.J., MA. Ettersona, T.P. Hodgman, A. Borowske, J.B. Cohen, C.S. Elphick, C.R. Field, R.A. Kern, E. King, A.R. Kocek,
809 A.I. Kovach, K.M. O'Brien, N. Paul, W.G. Shriver, J. Walsh, and B.J. Olsen. 2017a. Demographic analysis demonstrates
810 systematic but independent abiotic and biotic stressors across 59% of a global species range. *Auk. 134: 903–916*. DOI:
811 10.1642/AUK-16-230.1

812 Ruskin, K.J., MA. Etterson, T.P. Hodgman, A. Borowske, J.B. Cohen, C.S. Elphick, C.R. Field, R.A. Kern, E. King, A.R. Kocek, A.I.
813 Kovach, K.M. O'Brien, N. Paul, W.G. Shriver, J. Walsh, and B.J. Olsen. 2017b. Seasonal fecundity is not related to range
814 position across a species' global range despite a central peak in abundance. *Oecologia 183:291–301*. doi:10.1007/s00442-016-
815 3745-8

816 Schrader, L., & Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Molecular Ecology*, *28*(6), 1537–1549.
817 https://doi.org/10.1111/mec.14794

818 Searcy, W. A., & Marler, P. (1981). A Test for Responsiveness to Song Structure and Programming in Female Sparrows. *Science*,
819 *213*(4510), 926–928.

820 Serrato-Capuchina, A., & Matute, D. R. (2018). The role of transposable elements in speciation. *Genes*, *9*(5).
821 https://doi.org/10.3390/genes9050254

822 Shaffer, H. B., Toffelmier, E., Corbett-Detig, R. B., Escalona, M., Erickson, B., Fiedler, P., … Wang, I. J. (2022). Landscape
823 genomics to enable conservation actions: The California Conservation Genomics Project. *Journal of Heredity*, (April), 1–12.
824 https://doi.org/10.1093/jhered/esac020

825 Shields, G. F., & Straus, N. A. (1975). DNA-DNA Hybridization Studies of Birds. *Evolution*, *29*(1), 159.
826 https://doi.org/10.2307/2407149

827 Sim SB, Corpuz RL, Simmonds TJ, Geib SM. 2022. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents
828 occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics, 23: 157.
829 doi:10.1186/s12864-022-08375-1

Slotkin, R. K., & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, *8*(4), 272–285. https://doi.org/10.1038/nrg2072

Smeds, L., Qvarnstrom, A., & Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome Research* , 1211–1218. https://doi.org/10.1101/gr.204669.116

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Steadman, D. W. (1981). A re-examination of Palaeostruthus Hatcheri (Shufeldt), a late Miocene sparrow from Kansas. *Journal of Vertebrate Paleontology*, *1*(2), 171–173. https://doi.org/10.1080/02724634.1981.10011889

Suh, A., Witt, C., Menger, J. *et al.* (2016). Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. *Nature Communications* 7, 11396 (2016). https://doi.org/10.1038/ncomms11396

Suh, A., Smeds, L., & Ellegren, H. (2018). Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. *Molecular Ecology*, *27*(1), 99–111. https://doi.org/10.1111/mec.14439

843 Tomaszkiewicz, M., Medvedev, P., & Makova, K. D. (2017). Y and W Chromosome Assemblies: Approaches and Discoveries.
844 *Trends in Genetics*, *33*(4), 266–282. https://doi.org/10.1016/j.tig.2017.01.008

845 Tuttle, E. M., Bergland, A. O., Korody, M. L., Brewer, M. S., Newhouse, D. J., Minx, P., … Balakrishnan, C. N. (2016). Divergence
846 and functional degradation of a sex chromosome-like supergene. *Current Biology*, *26*(3), 344–350.
847 https://doi.org/10.1016/j.cub.2015.11.069

848 Uliano-Silva M, Ferreira Nunes JG, Krasheninnikova K, McCarthy SA. marcelauliano/MitoHiFi: mitohifi_v2.0 (v2.0). *Zenodo*. 2021.
849 doi:10.5281/zenodo.5205678

850 Ungerer, M. C., Strakosh, S. C., & Zhen, Y. (2006). Genome expansion in three hybrid sunflower species is associated with
851 retrotransposon proliferation. *Current Biology*, *16*(20), 872–873. https://doi.org/10.1016/j.cub.2006.09.020

852 Van't Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., … Saccheri, I. J. (2016). The industrial
853 melanism mutation in British peppered moths is a transposable element. *Nature*, *534*(7605), 102–105.
854 https://doi.org/10.1038/nature17951

855 Walsh, J., P.M. Benham, P.E. Deane-Coe, P. Arcese, B.G. Butcher, Y.L. Chan, Z.A. Cheviron, C.S. Elphick, A.I. Kovach, B.J. Olsen, W.G.
856 Shriver, V.L. Winder, and I.J. Lovette. 2019a. Genomics of rapid ecological divergence and parallel adaptation in songbirds. *Evolution*
857 *Letters 3-4: 324–338* doi:10.1002/evl3.126

Walsh, J., G. Clucas, M. MacManes, W. K. Thomas, Kovach, A.I. 2019b. Divergent selection and drift shape the genomes of two avian sister species spanning a saline-freshwater ecotone. *Ecology & Evolution* DOI: 10.1002/ece3.5804

Walsh, J., A.I. Kovach, P.M. Benham, G.V. Clucas, G.I. Winder, I. Lovette. 2021. Genomic data reveal the biogeographic and demographic history of *Ammospiza* sparrows in Northeast Tidal Marshes. *Journal of Biogeography 48:2360-2374*. https://doi.org/10.1111/jbi.14208

Walsh, J., I. Lovette, V. Winder, C. Elphick, B. Olsen, G. Shriver, and A.I. Kovach. 2017. Subspecies delineation amid phenotypic, geographic, and genetic discordance. *Molecular Ecology 26: 1242-1255. doi: 10.1111/mec.14010*

Warmuth, V. M., Weissensteiner, M. H., & Wolf, J. B. W. (2022). Accumulation and ineffective silencing of transposable elements on an avian W Chromosome. *Genome Research*, *32*(4), 671–681. https://doi.org/10.1101/gr.275465.121

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., … Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*(7289), 757–762. https://doi.org/10.1038/nature08819

Warren, W. C., Hillier, L. D. W., Tomlinson, C., Minx, P., Kremitzki, M., Graves, T., … Cheng, H. H. (2017). A new chicken genome assembly provides insight into avian genome structure. *G3: Genes, Genomes, Genetics*, *7*(1), 109–117. https://doi.org/10.1534/g3.116.035923

873 Wheelwright, N. T. and J. D. Rising (2020). Savannah Sparrow (*Passerculus sandwichensis*), version 1.0. In Birds of the World (A. F.
874 Poole, Editor). Cornell Lab of Ornithology, Ithaca, NY, USA. https://doi.org/10.2173/bow.savspa.01

875 Whitney, K. D., & Garland, T. (2010). Did genetic drift drive increases in genome complexity? *PLoS Genetics*, *6*(8), 1–6.
876 https://doi.org/10.1371/journal.pgen.1001080

877 Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., … Schulman, A. H. (2007). A unified classification
878 system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982. Retrieved from
879 http://www.nature.com/articles/nrg2165

880 Williams, H., Scharf, A., Ryba, A. R., Ryan Norris, D., Mennill, D. J., Newman, A. E. M., … Blackwood, J. C. (2022). Cumulative
881 cultural evolution and mechanisms for cultural selection in wild bird songs. *Nature Communications*, *13*(1), 4001.
882 https://doi.org/10.1038/s41467-022-31621-9

883 Winkler, D. W., S. M. Billerman, and I. J. Lovette (2020). New World Sparrows (*Passerellidae*), version 1.0. In Birds of the World
884 (S. M. Billerman, B. K. Keeney, P. G. Rodewald, and T. S. Schulenberg, Editors). Cornell Lab of Ornithology, Ithaca, NY,
885 USA. https://doi.org/10.2173/bow.passer3.01

886 Wright, N. A., Gregory, T. R., & Witt, C. C. (2014). Metabolic "engines" of flight drive genome size reduction in birds. *Proceedings*
887 *of the Royal Society B: Biological Sciences*, *281*(1779). https://doi.org/10.1098/rspb.2013.2780

888 Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., … Al., E. (2014). Comparative genomics reveals insights into avian genome
889 evolution and adaptation. *Science*, *346*(6215), 1311–1321. https://doi.org/10.1126/science.12513

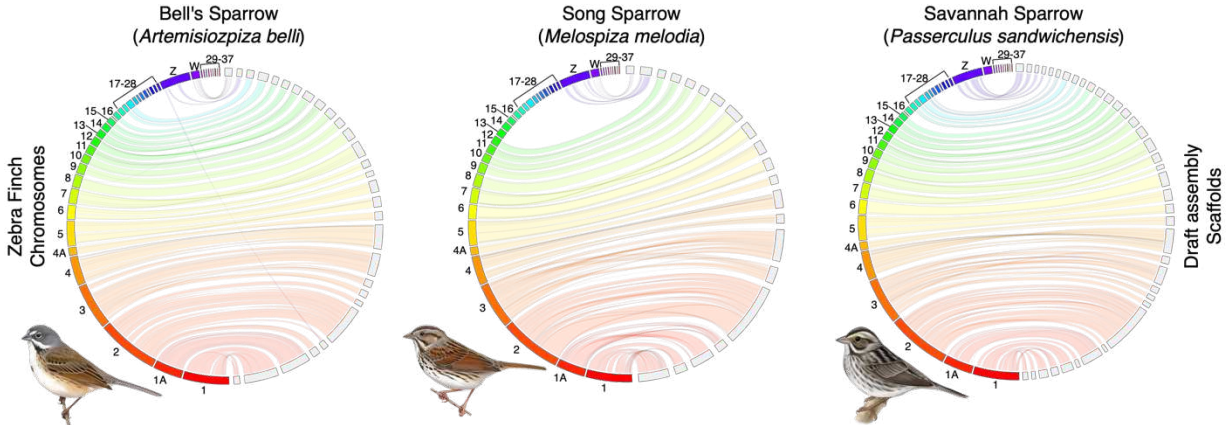


Figure 1: Jupiter plot comparing higher level synteny and completeness between the zebra finch (*Taeniopygia guttata*) genome (bTaeGut.4) and each of the three CCGP draft assemblies of Passerellidae sparrow species. Zebra finch chromosomes are on the left in each plot (colored) and sparrow scaffolds are on the right (light gray). Twists represent reversed orientation of scaffolds between assemblies. Song and Bell's sparrow reference genome samples were both from females, whereas the Savannah sparrow reference was from a male. Song and Bell's sparrow illustrations reproduced with the permission of https://birdsoftheworld.org with permission from Lynx Edicions. Savannah sparrow illustration contributed by Jillian Nichol Ditner.

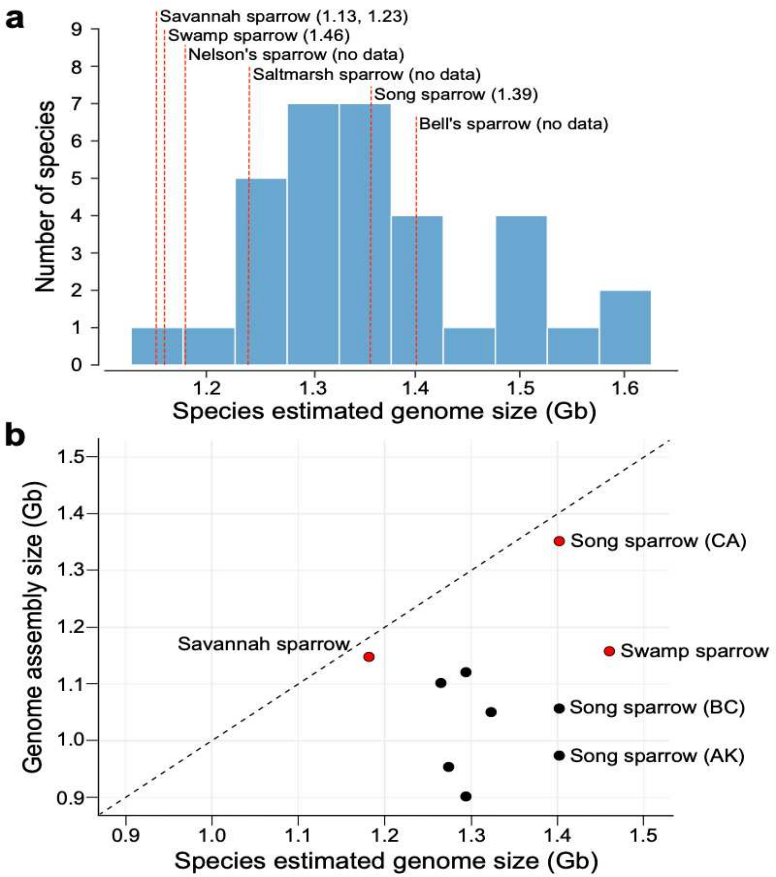


Figure 2: (a) Histogram showing variation in the genome sizes for 21 species (33 individuals) of Passerellidae sparrows estimated from Feulgen image analysis densitometry methods. C-values were adjusted based on the assumption that 1pg = 0.978 Gb (Dolezel et al. 2003). Genome size of the draft assemblies generated through the CCGP and VGP shown as red dashed lines with corrected c-value estimate (Gb) in parentheses. For *Passerculus sandwichensis* two estimates of genome size were found in the Animal Genome Size Database. (**b**) Comparison of estimated genome size from C-values versus genome assembly size (Gb). Black dashed line indicates the 1-to-1 line indicating equal estimates of genome size from the two metrics. Red dots are newly generated assemblies reported here. Nelson's, saltmarsh, and Bell's sparrow lack independent estimates of C-values and are not shown on this plot. Black dots denote previously published genome assemblies, which all show a shorter assembly length relative to the C-value estimate of genome size.

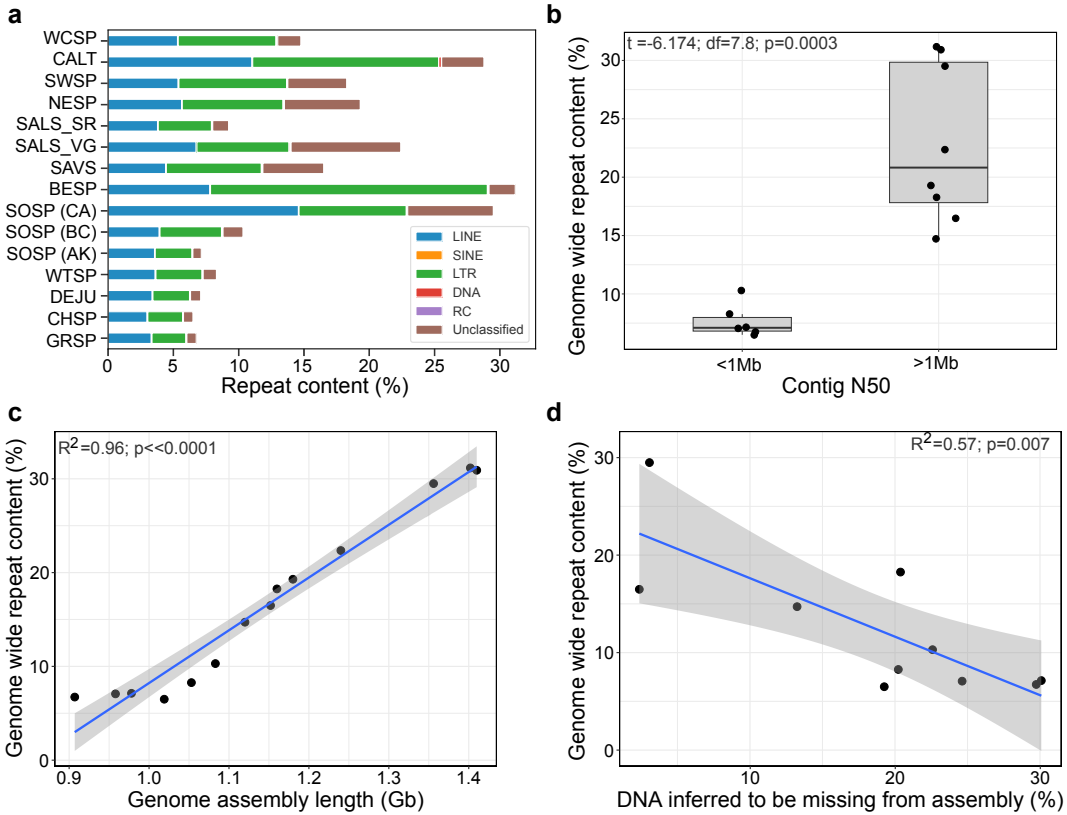


Figure 3: (a) Percentage of the genome comprising interspersed repeats, including: retroelements (LINE, SINE, LTR), DNA transposons (DNA), rolling-circles (RC), and unclassified elements. (white-crowned sparrow: WCSP; California towhee: CALT; swamp sparrow: SWSP; Nelson's sparrow: NESP; saltmarsh sparrow short-read: SALS_SR; saltmarsh sparrow long-read: SALS_VG; Savannah sparrow: SAVS; Bell's sparrow: BESP; song sparrow: SOSP; white-throated sparrow: WTSP; dark-eyed junco: DEJU; chipping sparrow: CHSP; grasshopper sparrow: GRSP). (**b**) The relationship between contig N50 and genome wide repeat content. Significantly higher levels of repeat content were discovered in genomes with a contig N50 greater than 1 Mb. All of which were generated with PacBio long-read technology. **(c)** Correlation between percent repeat content identified in each genome and the length of the assembled genome in Gb. (**d**) Correlation between percent repeat content and the amount of DNA inferred to be missing from each of the sparrow assemblies. C-value is assumed to be the more accurate estimate of total genome length. Percent missing DNA from each sparrow assembly is estimated as the difference between the c-value and assembly length.

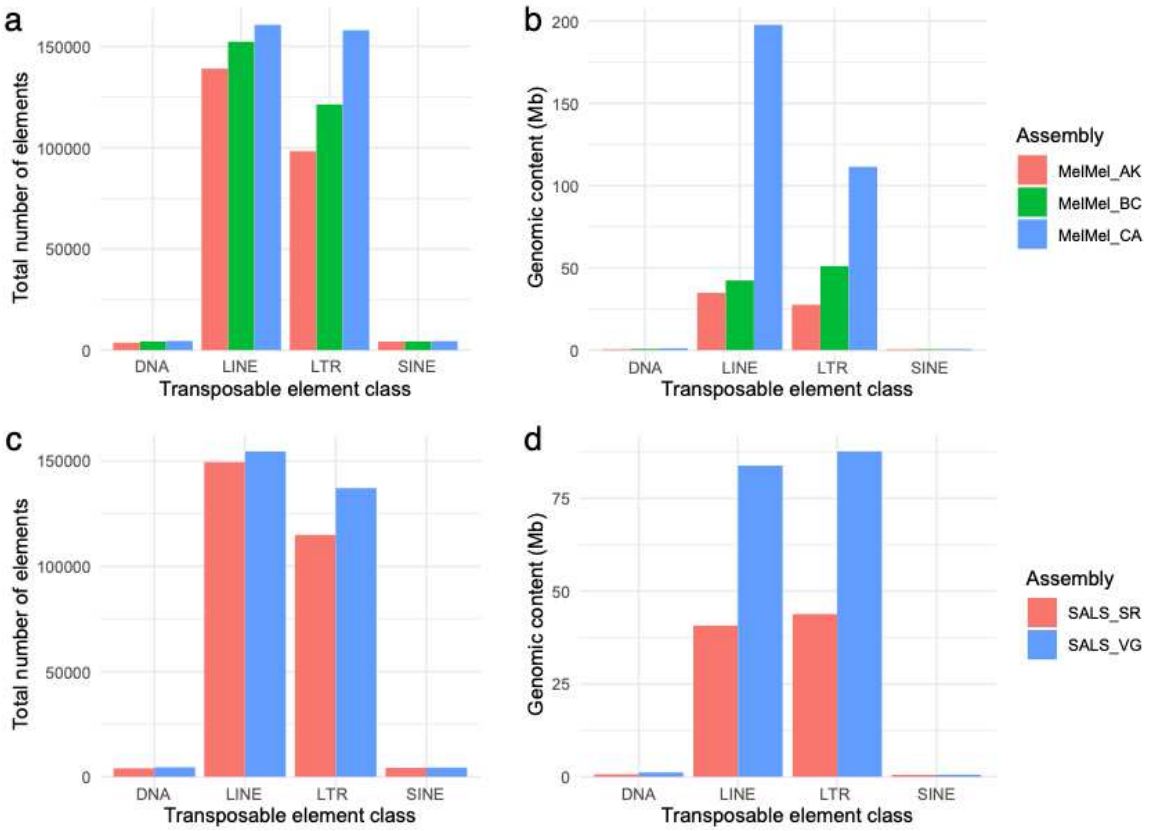
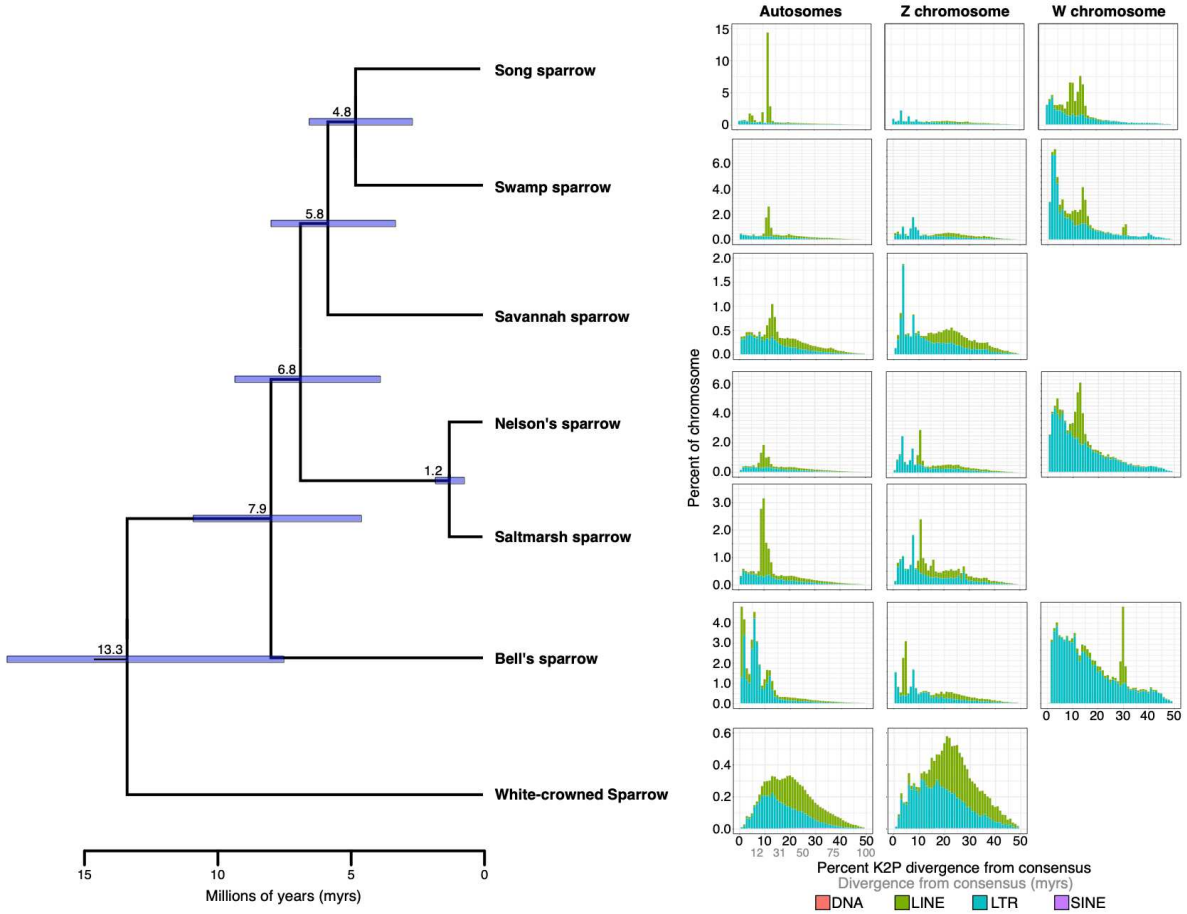


Figure 4: (**a-b**) Comparison of transposable element annotation across three song sparrow (*Melospiza melodia*) assemblies. (**a**) Total number of transposable elements found in each assembly. (**b**) Total genomic content (Mb) of transposable element content identified in each of the three song sparrow assemblies. MelMel_AK (red) is an assembly from an Alaskan bird sequenced using short-read and Chicago library technology. MelMel_BC (green) is a bird from British Columbia sequenced using Illumina short-read and PacBio SMRT long-read approaches MelMel_CA was generated using Hi-C and PacBio long read approaches for this paper. (**c-d)** Comparison of TE annotations in two saltmarsh sparrow assemblies (*Ammospiza caudacuta*). (**c**) Total number of transposable elements found in each assembly. (**b**) Total genomic content (Mb) of transposable element content identified in each of the three song sparrow assemblies. SALS_SR (Red) was assembled from Illumina short reads and the SALS_VG (blue) assembly was assembled using PacBio long read and Omni-C approaches with the Vertebrate Genomics Project pipeline.

933



934

935 **Figure 5:** Transposable element (TE) landscapes for the autosomal, Z, and W chromosomes. Left panel shows time-calibrated UCE
936 phylogeny of seven sparrow species. Branch labels indicate mean estimate of divergence time for each node with purple bars indicating
937 95% HPD error around that estimate. All nodes in the topology received bootstrap support of 100%. Right panel shows TE landscapes
938 for each species. Percent divergence on the x-axis was calculated as the percent Kimura 2-parameter (K2P) distance with CpG sites
939 excluded. The abundance of TEs in each percent divergence bin was normalized as a percentage of the chromosome length on the y-
940 axis.

941 **Table 1:** Comparison of assembly quality statistics and BUSCO search results among the three CCGP (left three) and three VGP (right
942 three) genomes. BUSCO results for the CCGP genomes were obtained using the 8,338 universal single copy genes in birds found in the
943 aves_odb10 database. BUSCO results from the VGP genomes were obtained using the 10,844 genes from the passeriformes_odb10
944 database.
945

Genome metrics	**Savannah Sparrow (*Passerculus sandwichensis*)**	**Bell's Sparrow (*Artemisiospiza belli*)**	**Song Sparrow (*Melospiza melodia*)**	**Swamp sparrow (*Melospiza georgiana*)**	**Nelson's sparrow (*Ammospiza nelsoni*)**	**Saltmarsh sparrow (*Ammospiza caudacuta*)**
Chromosomes	NA	NA	NA	36	37	40
# contigs	676	1539	823	276	292	645
Largest contig (bp)	32,137,824	35,931,659	59,497,540	29,976,604	50,792,433	43,812,934
Total length (bp)	1,152,258,190	1,401,798,777	1,356,272,071	1,160,782,308	1,180,370,373	1,239,216,328
GC (%)	43.1	43.46	44.45	43.24	43.01	43.5
N50	5,981,027	8,253,817	8,311,625	10,446,106	12,036,358	8,252,193
N75	2,762,208	1,578,947	3,378,466	3,855,203	4,591,448	3,265,797
L50	50	45	39	36	27	39
# scaffolds	337	1339	501	40	77	282
Largest scaffold (bp)	124,432,526	99,814,828	153,992,920	155,044,423	155,447,619	157,152,855
Total length (bp)	1,152,292,115	1,401,818,823	1,356,304,709	1,162,015,399	1,185,463,352	1,241,209,685
GC (%)	43.1	43.46	44.45	43.24	43.01	43.5
N50	18,220,233	17,082,054	25,784,215	74,254,230	74,723,840	78,443,464
N75	6,722,078	2,980,250	6,297,809	23,937,375	21,551,278	22,481,186
L50	17	20	14	6	6	6
# N's per 100 kbp	2.94	1.43	2.38	106.12	429.62	160.60
BUSCO_results (%)						
complete	97.3	97.0	96.8	98.5	99.0	99.2

complete & single copy	96.7	96.2	96	98.2	98.7	98.9
complete & duplicate	0.6	0.8	0.8	0.4	0.3	0.4
fragmented	0.9	0.8	0.8	0.2	1.2	0.2
missing	1.8	2.2	2.4	1.3	0.8	0.6

948 **Table 2:** Percentage of each genome spanned by different classes of repeats. Estimates of each class of repeat region identified within
949 RepeatMasker using the sparrow TE libraries generated *de novo* with RepeatModeler2.
950

Element class:	**Savannah sparrow (*Passerculus sandwichensis*)**	**Bell's sparrow (*Artemisiospiza belli*)**	**Song sparrow (*Melospiza melodia*)**	**Swamp sparrow (*Melospiza georgiana*)**	**Nelson's sparrow (*Ammospiza nelsoni*)**	**Saltmarsh sparrow (*Ammospiza caudacuta*)**
LINE	4.41	7.8	14.58	5.37	5.63	6.75
SINE	0.05	0.04	0.02	0.05	0.05	0.05
LTR	7.28	21.21	8.22	8.25	7.72	7.06
DNA transposons	0.09	0.10	0.09	0.08	0.08	0.1
Rolling-circles	0.02	0.02	0.01	0.01	0.02	0.05
Unclassified	4.66	2.01	6.56	4.51	5.81	8.40
Total interspersed repeats:	16.49	31.16	29.49	18.26	19.29	22.35
Autosomal chromosomes:	16.17	30.91	28.84	15.85	16.11	19.48
Z chromosome:	20.72	23.75	25.28	22.40	26.42	26.46
W chromosome:	NA	82.59	91.03	93.73	79.23	NA
Other repeat regions:						
Small RNA	0.04	0.04	0.03	0.06	0.05	0.04
Satellites	0.33	0.21	0.24	0.25	0.22	0.25
Simple repeats	1.35	1.11	1.05	1.25	1.19	1.21
Low complexity	0.25	0.20	0.20	0.23	0.28	0.32

951
952