G3 Genes | Genomes | Genetics

A High-Quality Genome Assembly of the North American Song Sparrow, *Melospiza melodia*

Swarnali Louha,*,[1] David A. Ray,† Kevin Winker,‡ and Travis C. Glenn*,§
*Institute of Bioinformatics, §Department of Environmental Health Science, University of Georgia, Athens, GA
†Department of Biological Science, Texas Tech University, Lubbock, TX and ‡University of Alaska Museum, Fairbanks, AK
ORCID IDs: 0000-0002-0777-8507 (S.L.); 0000-0002-3340-3987 (D.A.R.); 0000-0002-8985-8104 (K.W.); 0000-0001-7725-3637 (T.C.G.)

ABSTRACT The song sparrow, *Melospiza melodia*, is one of the most widely distributed species of songbirds found in North America. It has been used in a wide range of behavioral and ecological studies. This species' pronounced morphological and behavioral diversity across populations makes it a favorable candidate in several areas of biomedical research. We have generated a high-quality *de novo* genome assembly of *M. melodia* using Illumina short read sequences from genomic and *in vitro* proximity-ligation libraries. The assembled genome is 978.3 Mb, with a physical coverage of 24.9×, N50 scaffold size of 5.6 Mb and N50 contig size of 31.7 Kb. Our genome assembly is highly complete, with 87.5% full-length genes present out of a set of 4,915 universal single-copy orthologs present in most avian genomes. We annotated our genome assembly and constructed 15,086 gene models, a majority of which have high homology to related birds, *Taeniopygia guttata* and *Junco hyemalis*. In total, 83% of the annotated genes are assigned with putative functions. Furthermore, only ~7% of the genome is found to be repetitive; these regions and other non-coding functional regions are also identified. The high-quality *M. melodia* genome assembly and annotations we report will serve as a valuable resource for facilitating studies on genome structure and evolution that can contribute to biomedical research and serve as a reference in population genomic and comparative genomic studies of closely related species.

KEYWORDS
Melospiza melodia
Dovetail genomics
Passeriformes
whole genome sequencing
de novo assembly

The oscine passerines (Order Passeriformes) are songbirds having specialized vocal learning capabilities (Liu *et al.* 2013). Many species of songbirds have been widely used by neuroscientists to study the processes underlying memory and learning and social interactions (Doupe and Kuhl 1999, White 2010). The song sparrow (*Melospiza melodia*) is one of the most morphologically diverse songbirds found in North America, with 26 recognized subspecies (Pruett *et al.* 2008). It has been recognized as a model vertebrate species for field studies of birds and has been the subject of extensive research integrating behavioral and ecological studies over the last 70 years (Arcese *et al.* 2002). The species is widespread across North America, occupying diverse ecosystems and exhibiting pronounced phenotypic variation in plumage color, seasonal migration and sedentariness, body size, and bill size (Arcese *et al.* 2002, Pruett & Winker 2010, Greenberg *et al.* 2012).

Though several species of songbirds have been sequenced and studied (Warren *et al.* 2010, Jarvis *et al.* 2014), few offer the plethora of biomedical research potential presented by the song sparrow. This species might serve as a model system in areas such as hepatic lipogenesis (through phenotypic variation in seasonal fat deposition for migration; Gosler 1996, Schubert *et al.* 2007), craniofacial development (through variation in bill size and shape; Brugmann *et al.* 2010, Powder *et al.* 2012), and variations in body size (Sutter *et al.* 2007, Lango Allen *et al.* 2010). The latter is a polygenic trait, and elucidation of the underlying gene network affecting different metabolic pathways can help clarify several biological phenomena, including human diseases. Other areas of interest are differences in neural growth and song-center brain development among different song sparrow populations and potential applications in brain neurogenesis (NIH 2001), and also the regeneration of "hair" cells in the song sparrow auditory system and potential therapies useful in hearing loss (Hawkins *et al.* 2003, Hawkins & Lovett 2004). Given its significant biomedical potential and experimental tractability in the field and aviary, the song sparrow will continue to

doi: https://doi.org/10.1534/g3.119.400929
Manuscript received November 19, 2019; accepted for publication February 13, 2020; published Early Online February 19, 2020.

Supplemental material available at figshare: https://doi.org/10.25387/g3.11676441.
[1] Corresponding author: 206 Environmental Health Sci Bld, Athens, GA 30602. E-mail: sl50708@uga.edu

be used for answering research questions related to mechanisms causing variation in behavior, morphology, and demographics across populations (Arcese *et al.* 2002, Nietlisbach *et al.* 2015).

Prior work on song sparrows in Alaska has shown how the song sparrow population in the Aleutian Archipelago is thought to have colonized from the mainland since the last glacial maximum and undergone a series of population bottlenecks to give rise to a naturally inbred population with large body size (Pruett and Winker 2005). The lower genetic variability in this naturally inbred population makes song sparrows from the Aleutian islands a favorable resource for generating a reference genome assembly, because lower levels of polymorphism between both copies of a diploid genome can improve assembly quality. Previous work has also been done on the song sparrow transcriptome, developing genomic markers to screen at population levels (Srivastava *et al.* 2012). A high-quality genome assembly of *M. melodia* furthers the development of genomic markers to screen loci associated with phenotypic traits of interest. An ever-growing number of songbirds have sequenced genomes, but relatively few have been published so far, including the American crow (*Corvus brachyrhynchos*), golden-collared manakin (*Manacus vitellinus*; Jarvis *et al.* 2014), Zebra finch (*Taeniopygia guttata*; Warren *et al.* 2010), medium ground finch (*Geospiza fortis*; Parker *et al.* 2012) and the dark-eyed junco (*Junco hyemalis*; Friis *et al.* 2018). In this study, we provide the genome assembly of *Melospiza melodia*, a member of the family Passerellidae. This genome assembly will serve as a reference genome for this species as well as facilitating genomic and phylogenetic comparisons among songbirds and other taxa.

Our high-quality draft genome assembly of *M. melodia* was created by combining both traditional Illumina paired-end libraries and a *de novo* proximity-ligation Chicago library. The Chicago library method together with Dovetail Genomics' HiRise software pipeline is designed to significantly reduce gaps in alignment arising from repetitive elements in the genome (Putnam *et al.* 2016) and increases assembly contiguity. The draft genome was annotated using transcribed RNA and protein sequences from *M. melodia* and related songbird species, *Junco hyemalis* and *Taeniopygia guttata*. Genomic features of interest other than coding sequences, such as microsatellites, repeat elements, transposable elements, and non-coding RNA, were also annotated and the genome assembly was evaluated for quality by comparing it to related avian species.

METHODS

Library preparation and de novo shotgun assembly

The *de novo* assembly of the song sparrow genome was constructed using Illumina paired end libraries. A blood sample from a single male song sparrow was obtained from the wild in the Aleutian Islands of Alaska (Coordinates: 52.8275 / 173.206) on 16 Sep 2003 and archived as a voucher specimen at the University of Alaska Museum (http://arctos.database.museum/guid/UAM:Bird:31500). We chose a male because females are the heterogametic sex in birds and sex chromosomes are known to have highly repetitive DNA content. This together with the selection of an individual from a population known to have lower genetic variation can improve the quality of our assembled genome, without changing the genome structurally. Whole blood was preserved during specimen preparation and shipped overnight in lysis buffer to UGA, where PCI extraction of DNA was performed. We sheared the genomic DNA using a Covaris S2 (Covaris, Woburn, MA, USA) targeting a 600bp average fragment size. The sheared DNA was end-repaired, adenylated, and ligated to TruSeq LT adapters using a TruSeq DNA PCR-Free Library Preparation Kit (Illumina, San Diego, CA, USA). We purified the ligation reaction using a Qiaquick Gel Extraction Kit (Qiagen, Venlo, The Netherlands) from a 2% agarose gel. We sequenced the library on an Illumina HiSeq 2500 at the HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA) to obtain paired-end (PE) ~100 bp reads. The sequence data consisted of 276 million read pairs sequenced from a total of 41.3 Gbp of paired-end libraries (~49× sequencing coverage). Reads were trimmed for quality, sequencing adapters, and mate pair adapters using Trimmomatic (Bolger *et al.* 2014). The reads were assembled at Dovetail Genomics (Santa Cruz, CA, USA) using Meraculous 2.0.4 (Chapman *et al.* 2011) with a *k-mer* size of 29. This yielded a 972.4 Mbp assembly with a contig N50 of 22.5 Kbp and a scaffold N50 of 33 Kbp.

Chicago library preparation and scaffolding the draft genome

To improve the *de novo* assembly, a Chicago library was prepared at Dovetail Genomics using previously described methods (Putnam *et al.* 2016). In brief, about 500 ng of high-molecular-weight genomic DNA (mean fragment length = 50 kbp) was used for chromatin reconstitution *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with *Dpn*II, the 5′ overhangs filled in with biotinylated nucleotides, and free blunt ends were ligated together. After ligation, crosslinks were reversed and DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. Next, DNA was sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes (New England Biolabs, Ipswich, MA, USA) and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of the library. The Chicago library was sequenced on an Illumina HiSeq 2500 to produce 47 million 150 bp paired end reads (1-50 kb pairs).

Dovetail Genomics' HiRise scaffolding software pipeline (Putnam *et al.* 2016) was used to map the shotgun and Chicago library sequences to the draft *de novo* assembly using a modified SNAP read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After scaffolding, shotgun sequences were used to close gaps between contigs.

Identification of microsatellites and transposable elements

Transposable elements (TEs) in the song sparrow genome were identified using a combination of *de novo* and homology-based TE identification methods, in addition to a manual curation step (Platt *et al.* 2016). First, we used RepeatModeler v1.0.11 (Smit and Hubley 2008-2015) with default parameters (File S1) to generate a custom repeat library consisting of 672 consensus repeat sequences. RepeatModeler uses two *de novo* repeat identification programs, RECON v1.08 (Bao and Eddy 2002) and RepeatScout v1.0.6 (Price *et al.* 2005), for identifying repetitive elements from sequence data. To ensure accurate and complete representation of putative TEs, the RepeatModeler derived consensus sequences were filtered for size (>100 bp), and then subjected to iterative homology-based searches against the genome, followed by manual curation (Platt *et al.* 2016). The final set of manually curated TEs was queried against CENSOR (Kohany *et al.* 2006) and TEclass (Abrusan *et al.* 2009) for classification. TEs not identifiable in CENSOR were also searched against the NCBI nucleotide and protein databases using BLASTN and BLASTX respectively. Finally, a

Table 1 A comparison of assembly quality statistics from the initial shotgun sequencing assembled by Meraculous and the final HiRise assembly

	Meraculous Assembly	Chicago HiRise Assembly
Total length	972.4 Mb	978.3 Mb
Scaffold N50	33 kb	5.58 Mb
Scaffold N90	5 kb	303 kb
Scaffold L50	7,552 scaffolds	48 scaffolds
Scaffold L90	35,731 scaffolds	324 scaffolds
Longest scaffold	366,149	26,942,064
Number of scaffolds	74,832	13,785
Number of scaffolds > 1kb	74,806	13,768
Contig N50	22.5 kb	31.7 kb
Number of gaps	53,577	95,490
Percent of genome in gaps	1.427%	1.847%
Number of N's per 100 kbp	1427.15	1847.03
GC content	41.07%	41.08%

custom repeat library consisting of 900 repeat elements (File S24) comprising song sparrow-specific TEs and existing repeats in other related avian species was used to screen for repeats in the song sparrow genome assembly with RepeatMasker v4.0.9.

Microsatellites in the song sparrow genome were identified and described with GMATA v2.01 (Wang and Wang 2016) with sequence motifs ranging in length from 2-20 bp, and each motif repeated at least 5 times (File S2).

De novo gene annotation and function prediction

Genes were predicted in the song sparrow genome with the MAKER v2.31.9 genome annotation pipeline (Campbell *et al.* 2014). A custom repeat library of 900 repeat sequences (File S24) consisting of TEs identified in the song sparrow genome and other existing avian repeat elements was used to soft mask the genome. Transcriptome evidence sets for MAKER included the assembled song sparrow transcriptome (Srivastava *et al.* 2012) and Trinity (v2.4.0) mRNA-seq assemblies from multiple tissues of *Junco hyemalis* (Peterson *et al.* 2012, NCBI BioProject Accession: PRJNA256328). Protein evidence sets used by MAKER included annotated proteins for song sparrow, *Junco hyemalis*, and *Taeniopygia guttata* from the NCBI Protein database. The MAKER pipeline consisted of the following steps: 1) Transcriptomic and protein evidence sets were used to make initial evidence-based annotations with MAKER; 2) the initial annotations were used to train two *ab initio* gene predicters: Augustus (Stanke *et al.* 2006), which was trained once, and SNAP (Korf 2004), which was iteratively trained twice; and 3) the trained gene prediction tools SNAP and Augustus were used to generate the final set of gene annotations (File S3-S8).

Functional annotations of the predicted genes were obtained by making homology-based searches with BLASTP against the Uniprot/Swiss-Prot protein database (Pundir *et al.* 2016, File S9). InterProScan v5.29 (Zdobnov and Apweiler 2001) was used to find protein domains associated with the genes. The putative functions and protein domains were added to the gene annotations using scripts provided with MAKER (File S9).

To quantitatively assess the completeness of the song sparrow genome assembly and annotated gene set, we ran BUSCO (Benchmarking Universal Single-Copy Orthologs) v3.0.2 (Waterhouse *et al.* 2017) with 4,915 single-copy orthologous genes in the Aves lineage group (Aves_odb9; https://busco.ezlab.org/), using "chicken" as the Augustus reference species (File S10). The 4,915 orthologous genes are present in at least 90% of the 40 species included within the Aves lineage group, and thus are likely to be found in the genome of related species. Additionally, we used the JupiterPlot pipeline (https://github.com/JustinChu/JupiterPlot) to visually compare the zebra finch (*T. guttata*) genome assembly (Warren *et al.* 2010) to our assembly in a Circos plot, using the largest scaffolds making up 85% of our genome assembly, and all scaffolds greater than 100 kbp in the Zebra finch genome (File S11). We also used the JupiterPlot pipeline to compare our assembly to the genome assemblies of the collared flycatcher (*Ficedulla albicollis*), great tit (*Parus major*) and house sparrow (*Passer domesticus*). These birds were selected for comparison because they have highly complete genomes, and are often used for comparative genomic studies in birds.

Non-coding RNA prediction

Transfer RNAs (tRNAs) were predicted in the song sparrow genome with tRNAscan-SE v2.0 (Lowe and Chan 2016, File S12). A training set comprising eukaryotic tRNAs was used to train the covariance models employed by tRNAscan-SE, and tRNAs were searched against the

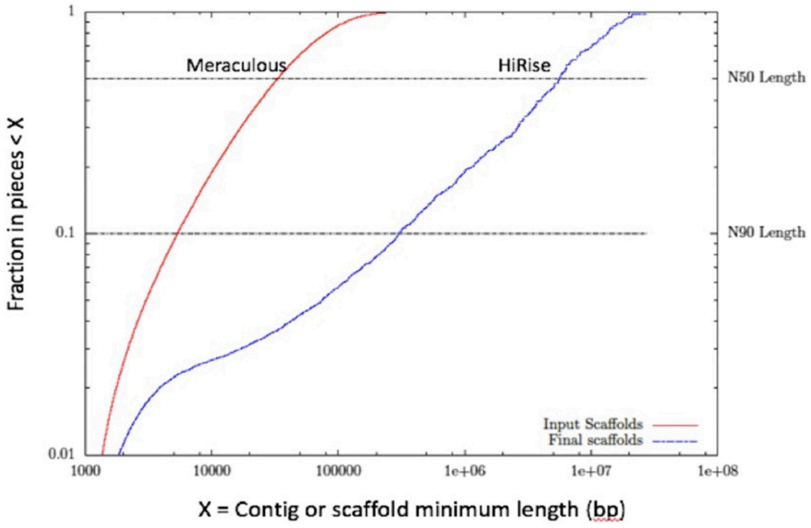

Figure 1 Comparison of assembly contiguity.

Table 2 Number and percentage of repeats in the *M. melodia* genome assembly

Classification	Number of copies	Percentage of assembly
LINEs	104,032	3.01
LTRs	85,276	2.83
SINEs	6,695	0.08
DNA Transposons	13,521	0.21
Unclassified	4,884	0.12
Total transposable elements	**214,408**	**6.25**
Satellites	569	0.00
Low complexity repeats	38,561	0.20
Microsatellites	192,996	0.90
Total	**446,534**	**7.35**

genome with Infernal v1.1.2 (Nawrocki 2014). tRNAscan-SE also provides functional classification of tRNAs based on a comparative analysis using a suite of isotype-specific tRNA covariance models. A random sample of 10 predicted tRNAs were selected and searched against the tRNA databases GtRNAdb (Chan and Lowe 2016) and tRNAdb (Jühling *et al.* 2009).

Identification of miRNAs (microRNAs), snoRNAs (small nucleolar RNAs), snRNAs (small nuclear RNAs), rRNAs (ribosomal RNAs), and lncRNAs (long non-coding RNAs) was achieved by using a homology-based prediction method. Structural homologs to eukaryotic ncRNA covariance models from the Rfam database v14.1 (Gardner *et al.* 2009) were searched against the song sparrow genome using Infernal's (v1.1.2) "cmscan" program (File S13). All low-scoring overlapping hits and hits with an E-value greater than 10^{-5} were discarded, and the remaining ncRNAs were grouped into different classes.

Lastly, we compared the predicted classes of different ncRNAs in the song sparrow genome to those reported in the genomes of related birds, *Taeniopygia guttata* and *Ficedula albicollis* (collared flycatcher).

Data availability

Raw reads have been deposited in the NCBI Sequence Read Archive (SRR10491484 and SRR10451714 for the Meraculous assembly, and SRR10424475 for the Chicago HiRise assembly). The *M. melodia* Chicago HiRise genome sequence (Mmel_1.0), and annotations are available in GenBank under the accession RZID00000000 (NCBI BioProject accession: PRJNA511035). Supplemental File S1 contains submission script for RepeatModeler. Supplemental File S2 contains primary configuration file used to run GMATA (default_cfg.txt). Supplemental File S3 contains submission script for MAKER. Supplemental File S4 contains MAKER executable file (maker_exe.ctl). Supplemental File S5 contains specifications for downstream filtering of BLAST and Exonerate alignments (maker_bopts.ctl). Supplemental File S6 contains primary configuration of MAKER specific options (maker_opts.ctl). Supplemental File S7 contains scripts for training SNAP. Supplemental File S8 contains scripts for training Augustus. Supplemental File S9 contains scripts for running BLASTP and InterProScan for functional annotation of predicted genes; and scripts for adding the functional annotations to gene annotation files. Supplemental File S10 contains submission script for BUSCO. Supplemental File S11 contains submission scripts for JupiterPlot pipeline. Supplemental File S12 contains submission script for tRNAscan-SE. Supplemental File S13 contains submission script for Infernal. Supplemental File S14 contains classification of predicted transposable elements. Supplemental File S15 contains annotation of microsatellites with their genomic locations. Supplemental File S16 contains percentage of different microsatellites present in the genome. Supplemental File S17 contains frequency of occurrence of microsatellites in each scaffold of the genome. Supplemental File S18 contains the distribution of the length of microsatellites. Supplemental File S19 contains predicted function of annotated genes by BLASTP. Supplemental File S20 contains prediction of protein domains, GO annotations and pathway annotations of predicted genes by InterProScan. Supplemental File S21 contains sequence and structure of tRNAs identified in the song sparrow genome. Supplemental File S22

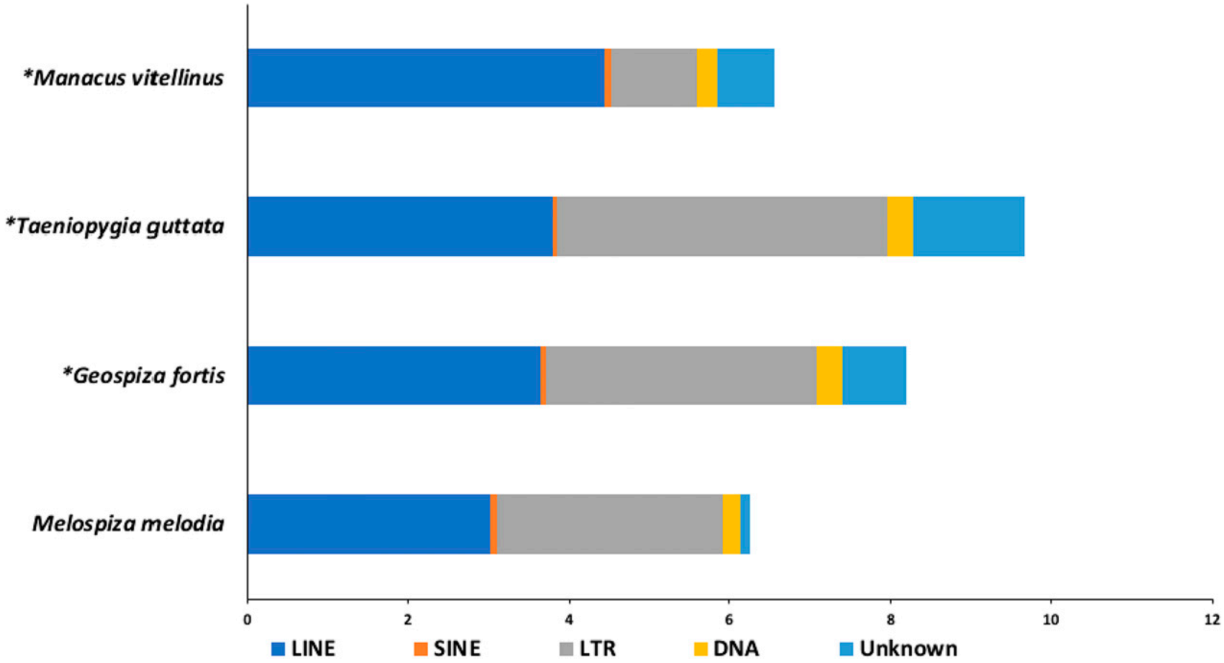

Figure 2 Comparison of percentages of transposable elements (TEs) among related songbird genome assemblies. * Data from: Zhang *et al.* (2014) Science. 346: 1311-1320.

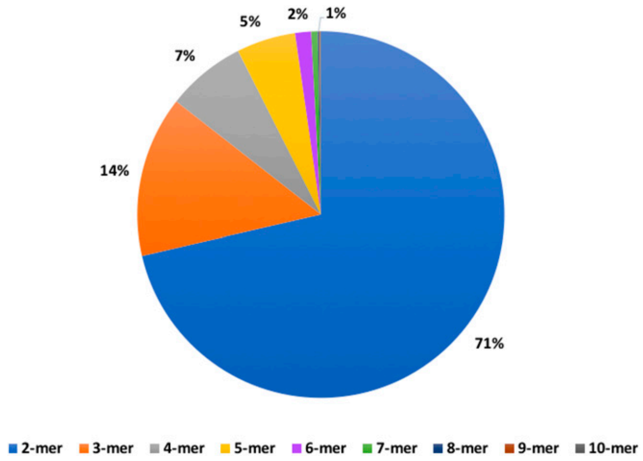

Figure 3 Abundance of microsatellite repeat motif size classes in the *M. melodia* genome assembly (details are given in Supplemental File S16).

contains classification of predicted tRNAs. Supplemental File S23 contains classification of different ncRNAs predicted in the genome with Infernal. Supplemental File S24 contains custom repeat library used to screen for repeats in the song sparrow genome. Supplemental Table S1 contains genome sizes of birds related to *M. melodia*. Supplemental Figure S1 contains the distribution of the percentage of annotated genes with their corresponding AED scores. Supplemental Figure S2 contains the distribution of the top base-pair composition of microsatellite motifs in the *M. melodia* genome. Supplemental Figure S3 contains comparison of the *M. melodia* genome assembly with genome assemblies of related birds. Supplemental material available at figshare: https://doi.org/10.25387/g3.11676441.

RESULTS AND DISCUSSION

Assembly

We produced the *de novo* genome assembly of song sparrow, with a total length of 978.3 Mb, using a Chicago library and the HiRise assembly pipeline. The N50 scaffold size was 5.6 Mb and contig size was 31.7 Kb. This assembly showed significant improvement over the initial shotgun assembly, with a 169-fold increase in scaffold N50 and a 60-fold increase in scaffold N90 (Table 1). These increases in scaffold size were also accompanied by an increase in assembly contiguity, with the total number of scaffolds decreasing from 74,832 to 13,785 (Figure 1, Table 1).

Microsatellites and transposable elements

In total, 88 as yet unnamed TEs were identified in the song sparrow genome. Fifty-five of these did not have any significant matches in CENSOR (Kohany *et al.* 2006) and are considered novel (File S14). A TE was considered to have a significant match to a known element in CENSOR only when it had a length of at least 80 bp and 80% identity to the known element over 80% of its length, the 80-80-80 rule (Wicker *et al.* 2007). The predicted TEs were classified into DNA transposons and retrotransposons (*i.e.*, LINEs, LTRs, and SINEs) using CENSOR and TEclass (File S14). Approximately 7.4% of the genome comprises repeats with the majority of that consisting of TEs (~48%). Among the different TEs, LTRs (~40%) and LINEs (~49%) were found to be most abundant (Table 2). The song sparrow genome assembly was found to be less repetitive when compared to sequenced genomes of related songbirds, primarily due to the lower content of LTRs and LINEs than other songbirds (Figure 2).

Overall, 112,419 microsatellites with motifs ranging in size from 2-20 bp were found in the song sparrow genome (File S15 contains all microsatellites with their genomic locations). The majority of the microsatellites were made up of 2-, 3-, 4-, and 5-mers, with 2-mers making up about 71% of all microsatellites identified (Figure 3, File S16). The distribution of the top base-pair composition of microsatellite motifs present in the genome is shown in Fig S2. The frequency of occurrence of microsatellites in every scaffold and a distribution of their lengths are provided in Files S17 and S18, respectively.

Gene annotation and function prediction

The MAKER genome annotation pipeline predicted 15,086 genes and 139 pseudogenes in the song sparrow genome, fewer than *T. guttata*, *F. albicollis*, and *M. vitellinus*, but higher than *G. fortis* (Table 3). The average gene length, exon length, intron length, and the total number of exons and introns predicted are also less compared to closely related species (Table 3). Of the 15,086 predicted genes, 12,541 genes were assigned putative functions with BLASTP (File S19). InterProScan assigned functional domains to 11,298 (74.9%) predicted genes (File S20). A total of 7,010 genes obtained GO annotations. Pathway annotations were assigned to 2,716 genes.

Annotated genes were assigned annotation edit distance (AED) scores with values ranging from 0 to 1. AED is a distance metric score that signifies how closely gene models match transcript and protein evidence. Gene models with AED scores closer to 0 have better alignment

■ **Table 3 Characteristics of genes predicted in the *M. melodia* genome compared to *Taeniopygia guttata* (zebra finch), *Ficedulla albicollis* (collared flycatcher), *Manacus vitellinus* (golden-collared manakin) and *Geospiza fortis* (medium ground finch)**

	M. melodia	*T. guttata*[1]	*F. albicollis*[2]	*M. vitellinus*[3]	*G. fortis*[4]
Number of genes	15,086	17,561	16,763	18,976	14,388
Mean gene length (bp)	14,457	26,458	31,394	27,847	30,164
Mean CDS length (bp)	1,325	1,677	1,942	1,929	1,766
Number of exons	131,940	171,767	189,043	190,390	164,721
Mean exon length (bp)	153	225	253	264	195
Mean number of exons/gene	8.67	10.25	12.22	11.51	11.41
Number of introns	116,724	153,909	171,236	171,089	149,563
Mean intron length (bp)	1,695	2,930	3,257	3,294	2,813

[1] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Taeniopygia_guttata/103/
[2] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ficedula_albicollis/101/
[3] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Manacus_vitellinus/102/
[4] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Geospiza_fortis/101/

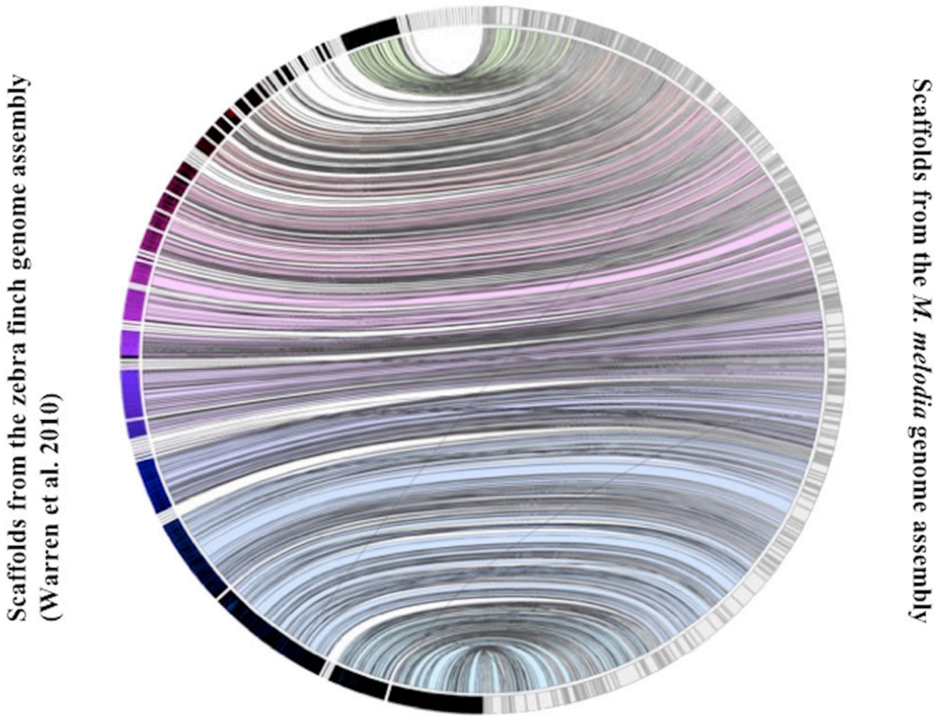

Figure 4 Jupiter plot correlating zebra finch and song sparrow genome assemblies, considering scaffolds greater than 100 kbp in the reference zebra finch genome and the largest scaffolds representing 85% of the song sparrow genome.

with the evidence provided in the MAKER pipeline. A distribution of the percentage of genes with their corresponding AED scores shows close similarity of the annotated genes with the transcript and protein evidence provided in the MAKER pipeline (Fig S1).

The song sparrow genome assembly contained 4,318 complete universal single-copy orthologs (BUSCOs; 87.9%) from a total of 4,915 BUSCO groups searched. Among all complete BUSCOs, 99.4% were present as single-copy genes and 0.6% were duplicated. About 7.4% (356) of the orthologous gene models were partially recovered, and 4.9% (241) had no significant matches. The incomplete and missing gene models could either be partially present or missing, or could indicate genes that are too divergent or have very complex structures, making their prediction difficult. Incomplete and missing gene models could also suggest problems associated with the genome assembly and gene annotation. The results from the BUSCO analysis are in agreement with the Circos plot (Figure 4), in which few scaffolds in the *T. guttata* genome assembly are not represented in our assembly and very few inconsistent arrangements of scaffolds exist between the two genome assemblies. Comparison of our assembly to *F. albicollis*, *P. major*, and *P. domesticus* genome assemblies showed many more inconsistencies in the arrangements of scaffolds between the genomes of these birds and *M. melodia* (Fig S3) than between *T. guttata* and *M. melodia*.

Non-coding RNA prediction and identification

A total of 267 tRNAs were detected in the song sparrow genome by tRNAscan-SE (see File S21 for sequence and structure of tRNAs), out of which 129 were found coding for the standard twenty amino acids. The predicted output from tRNAscan-SE (File S22) contained 114 tRNAs with low Infernal as well as Isotype scores; these were characterized as pseudogenes lacking tRNA-like secondary structures (Lowe and Chan 2016). Two tRNAs had undetermined isotypes and 22 were chimeric, with mismatch isotypes. Chimeric tRNAs contain point mutations in their anticodon sequence, rendering different predicted isotypes than those predicted by structure-specific tRNAscan-SE covariance models. Among all predicted tRNAs, 11 contained introns within their sequences. No suppressor tRNAs and tRNAs coding for selenocysteine were predicted. The subset of 10 randomly selected tRNAs was also predicted in many other species in both GtRNAdb and tRNAdb databases.

Infernal searches predicted a total of 364 ncRNAs in the song sparrow genome, comprising 166 miRNAs, 8 rRNAs, 154 snoRNAs, 16 snRNAs, and 20 lncRNAs (File S23). Compared to the genomes of related avian species (*T. guttata* and *F. albicollis*), the song sparrow genome has the highest number of predicted tRNAs, but fewer other ncRNAs (Table 4).

CONCLUSION

The Chicago and shotgun sequencing libraries along with the HiRise assembly software enabled accurate and highly contiguous *de novo* assembly of the song sparrow genome. The genome assembly is 978.3 Mb, with 48 scaffolds (L50) making up half the genome size. A previous estimate of genome size of *M. melodia* from densitometry analysis provided a C-value of 1.43 pg (1,398.54 Mb) (Andrews *et al.* 2009). Our own *k-mer* based estimate of genome size from paired reads

■ **Table 4 Number of ncRNAs predicted in the *Melospiza melodia* genome compared to *Taeniopygia guttata* (zebra finch) and *Ficedulla albicollis* (collared flycatcher)**

	M. melodia	*T. guttata*[1,2]	*F. albicollis*[1,3]
tRNA	267	184	179
miRNA	166	302	510
snRNA	16	44	32
snoRNA	154	241	199
rRNA	8	100	22
lncRNA	20	908	1473

[1] http://useast.ensembl.org/info/data/ftp/index.html
[2] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Taeniopygia_guttata/103/
[3] https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ficedula_albicollis/101/

in the shotgun and Chicago libraries using Kmergenie v1.7044 (Chikhi and Medvedev 2014) yielded an estimated size of 1,127.25 Mb. Both these genome size estimates and the genome sizes of related birds (Table S1) are slightly higher than our genome assembly (978.3 Mb). Our small assembly size may be attributed to the compression of repetitive regions, which is generally observed in assemblies generated from short-read sequencing data. This is also consistent with the fact that our genome contains fewer repeats when compared to related songbirds (Figure 2). Although short reads limit our ability to characterize the total number of repeats within long tandem arrays, we have been able to characterize vast majority of repeats, resolving them into LINEs, SINEs, LTRs, and DNA retrotransposons (Figure 2, Table 2).

Our genome is highly complete, with 87.5% full-length genes present out of 4,915 universal orthologous genes in avian species. A large set of genes (15,086) with known homology to related birds was annotated in our study. A majority of these genes (83%) were assigned with putative functions. The improved scaffold lengths and gene model annotations will facilitate studies to identify genes responsible for multiple phenotypic traits of interest. Additionally, longer scaffolds in the Chicago HiRise assembly will help detect regions under selection, including SNPs and structural variants such as insertions/deletions or copy number variations which are potentially responsible for the phenotypic diversity observed in this species.

Though we report fewer miRNAs, snRNAs, snoRNAs, rRNAs, and lncRNAs in this genome than in related songbirds, we have high confidence in the predicted ncRNAs we report because we used conservative cutoffs to reduce false positives. Pending the availability of long-read data, this genome assembly provides an excellent reference for a range of genetic, ecological, functional, and comparative genomic studies in song sparrows and other songbirds.

ACKNOWLEDGMENTS

This research was supported in part by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the NIH (P20GM103395). The content is solely the responsibility of the authors and does not necessarily reflect the official views of the NIH. We thank David Sonneborn and Jack Withrow for their roles in making a vouchered specimen available for our work. We also thank Troy Kieran for assistance with laboratory work and sequencing logistics, Roger Nilsen and the staff of Georgia Genomics and Bioinformatics Core for constructing the genomic library, and the staff of Dovetail Genomics for help in preparing and processing the Chicago library and HiRise assemblies. The high performance computing cluster at Georgia Advanced Computing Resource Center (GACRC) at the University of Georgia provided computational infrastructure and technical support throughout the work.

LITERATURE CITED

Abrusan, G., N. Grundmann, L. DeMester, and W. Makalowski, 2009 TEclass: a tool for automated classification of unknown eukaryotic transposable elements. Bioinformatics 25: 1329–1330. https://doi.org/10.1093/bioinformatics/btp084

Andrews, C. B., S. A. Mackenzie, and T. R. Gregory, 2009 Genome size and wing parameters in passerine birds. Proc. Biol. Sci. 276: 55–61. https://doi.org/10.1098/rspb.2008.1012

Arcese, P., M. K. Sogge, A. B. Marr, and M. A. Patten, 2002 Song Sparrow (*Melospiza melodia*), version 2.0, *The Birds of North America*, edited by Poole, A. F., and F. B. Gill. Cornell Lab of Ornithology, Ithaca, NY.

Bao, Z., and S. R. Eddy, 2002 Automated de Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res. 12: 1269–1276. https://doi.org/10.1101/gr.88502

Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brugmann, S. A., K. E. Powder, N. M. Young, L. H. Goodnough, S. M. Hahn *et al.*, 2010 Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. Hum. Mol. Genet. 19: 920–930. https://doi.org/10.1093/hmg/ddp559

Campbell, M. S., C. Holt, B. Moore, M. Yandell, 2014 Genome annotation and curation using MAKER and MAKER-P. Curr. Protoc. Bioinformatics. 48: 4.11.1–39. https://doi.org/10.1002/0471250953.bi0411s48

Chan, P. P., and T. M. Lowe, 2016 GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Res. 44: D184–D189. https://doi.org/10.1093/nar/gkv1309

Chapman, J. A., I. Ho, S. Sunkara, S. Luo, G. P. Schroth *et al.*, 2011 Meraculous: de novo genome assembly with short paired-end reads. PLoS One 6: e23501. https://doi.org/10.1371/journal.pone.0023501

Chikhi, R., and P. Medvedev, 2014 Informed and automated k-mer size selection for genome assembly. Bioinformatics 30: 31–37. https://doi.org/10.1093/bioinformatics/btt310

Doupe, A. J., and P. K. Kuhl, 1999 Birdsong and Human Speech: Common Themes and Mechanisms. Annu. Rev. Neurosci. 22: 567–631. https://doi.org/10.1146/annurev.neuro.22.1.567

Friis, G., G. Fandos, A. J. Zellmer, J. E. McCormack, B. C. Faircloth *et al.*, 2018 Genome-wide signals of drift and local adaptation during rapid lineage divergence in a songbird. Mol. Ecol. 27: 5137–5153. https://doi.org/10.1111/mec.14946

Gardner, P. P., J. Daub, J. Tate, B. L. Moore, I. H. Osuch *et al.*, 2011 Rfam: Wikipedia, clans and the 'decimal' release. Nucleic Acids Res. 39: D141–D145. https://doi.org/10.1093/nar/gkq1129

Gosler, A. G., 1996 Environmental and social determinants of winter fat storage in the Great Tit *Parus major*. J. Anim. Ecol. 65: 1–17. https://doi.org/10.2307/5695

Greenberg, R., V. Cadena, R. M. Danner, and G. J. Tattersall, 2012 Heat Loss May Explain Bill Size Differences between Birds Occupying Different Habitats. PLoS One 7: e40933. https://doi.org/10.1371/journal.pone.0040933

Hawkins, R. D., S. Bashiardes, C. A. Helms, L. Hu, N. L. Saccone *et al.*, 2003 Gene expression differences in quiescent *vs.* regenerating hair cells of avian sensory epithelia: implications for human hearing and balance disorders. Hum. Mol. Genet. 12: 1261–1272. https://doi.org/10.1093/hmg/ddg150

Hawkins, R. D., and M. Lovett, 2004 The developmental genetics of auditory hair cells. Hum. Mol. Genet. 13: R289–R296. https://doi.org/10.1093/hmg/ddh249

Jarvis, E. D., S. Mirarab, A. J. Aberer, B. Li, P. Houde *et al.*, 2014 Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346: 1320–1331. https://doi.org/10.1126/science.1253451

Jühling, F., M. Mörl, R. K. Hartmann, M. Sprinzl, P. F. Stadler *et al.*, 2009 tRNAdb 2009: Compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 37: D159–D162. https://doi.org/10.1093/nar/gkn772

Kohany, O., A. J. Gentles, L. Hankus, and J. Jurka, 2006 Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474. https://doi.org/10.1186/1471-2105-7-474

Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5: 59. https://doi.org/10.1186/1471-2105-5-59

Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon *et al.*, 2010 Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832–838. https://doi.org/10.1038/nature09410

Liu, W. C., K. Wada, E. D. Jarvis, and F. Nottebohm, 2013 Rudimentary substrates for vocal learning in a suboscine. Nat. Commun. 4: 2082. https://doi.org/10.1038/ncomms3082

Lowe, T. M., and P. P. Chan, 2016 tRNAscan-SE On-line: integrating Search and Context for Analysis of Transfer RNA Genes. Nucleic Acids Res. 44: W54–W57. https://doi.org/10.1093/nar/gkw413

Nawrocki, E. P., 2014 Annotating functional RNAs in genomes using Infernal. Methods Mol. Biol. 1097: 163–197. https://doi.org/10.1007/978-1-62703-709-9_9

Nietlisbach, P., G. Camenisch, T. Bucher, J. Slate, L. F. Keller *et al.*, 2015 A microsatellite-based linkage map for song sparrows (*Melospiza melodia*). Mol. Ecol. Resour. 15: 1486–1496. https://doi.org/10.1111/1755-0998.12414

NIH, 2001 What we learned from songbirds: The adult brain can generate new nerve cells. NIH Publication No. 01–4602.

Parker, P., B. Li, H. Li, and J. Wang, 2012 The genome of Darwin's Finch (*Geospiza fortis*). Gigascience. https://doi.org/10.5524/100040

Peterson, M. P., D. J. Whittaker, S. Ambreth, S. Sureshchandra, A. Buechlein *et al.*, 2012 De novo transcriptome sequencing in a songbird, the dark-eyed junco (*Junco hyemalis*): genomic tools for an ecological model system. BMC Genomics 13: 305. https://doi.org/10.1186/1471-2164-13-305

Platt, 2nd, R. N., L. Blanco-Berdugo, and D. A. Ray, 2016 Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. Genome Biol. Evol. 8: 403–410. https://doi.org/10.1093/gbe/evw009

Powder, K. E., Y. C. Ku, S. A. Brugmann, R. A. Veile, N. A. Renaud *et al.*, 2012 A cross-species analysis of microRNAs in the developing avian face. PLoS One 7: e35111. https://doi.org/10.1371/journal.pone.0035111

Price, A. L., N. C. Jones, and P. A. Pevzner, 2005 De novo identification of repeat families in large genomes. Bioinformatics 21: i351–i358. https://doi.org/10.1093/bioinformatics/bti1018

Pruett, C. L., P. Arcese, Y. L. Chan, A. G. Wilson, M. A. Patten *et al.*, 2008 Concordant and discordant signals between genetic data and described subspecies of Pacific Coast Song Sparrows. Condor 110: 359–364. https://doi.org/10.1525/cond.2008.8475

Pruett, C. L., and K. Winker, 2005 Northwestern Song Sparrow populations show genetic effects of sequential colonization. Mol. Ecol. 14: 1421–1434. https://doi.org/10.1111/j.1365-294X.2005.02493.x

Pruett, C. L., and K. Winker, 2010 Alaska Song Sparrows (*Melospiza melodia*) demonstrate that genetic marker and method of analysis matter in subspecies assessments. Ornithol. Monogr. 67: 162–171. https://doi.org/10.1525/om.2010.67.1.162

Pundir, S., M. J. Martin, C. O'Donovan, The UniProt Consortium, 2016 UniProt Tools. Curr. Protoc. Bioinformatics. 53: 1.29.1–1.29.15.

Putnam, N. H., B. L. O'Connell, J. C. Stites, B. J. Rice, M. Blanchette *et al.*, 2016 Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26: 342–350. https://doi.org/10.1101/gr.193474.115

Schubert, K. A., D. J. Mennill, S. M. Ramsay, K. A. Otter, P. T. Boag *et al.*, 2007 Variation in social rank acquisition influences lifetime reproductive success in black-capped chickadees. Biol. J. Linn. Soc. Lond. 90: 85–95. https://doi.org/10.1111/j.1095-8312.2007.00713.x

Smit, A. F. A., and R. Hubley, 2008–2015 *RepeatModeler Open-1.0.11*, http://www.repeatmasker.org.

Srivastava, A., K. Winker, T. I. Shaw, K. L. Jones, and T. C. Glenn, 2012 Transcriptome analysis of a North American songbird, *Melospiza melodia*. DNA Res. 19: 325–333. https://doi.org/10.1093/dnares/dss015

Stanke, M., A. Tzvetkova, B. Morgenstern, 2006 AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol. 7 Suppl 1: S11.1–8.

Sutter, N. B., C. D. Bustamante, K. Chase, M. M. Gray, K. Zhao *et al.*, 2007 A single IGF1 allele is a major determinant of small size in dogs. Science 316: 112–115. https://doi.org/10.1126/science.1137045

Wang, X., and L. Wang, 2016 GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. Front. Plant Sci. 7: 1350.

Warren, W. C., D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier *et al.*, 2010 The genome of a songbird. Nature 464: 757–762. https://doi.org/10.1038/nature08819

Waterhouse, R. M., M. Seppey, F. A. Simão, M. Manni, P. Ioannidis *et al.*, 2017 BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol. Biol. Evol. 35: 543–548. https://doi.org/10.1093/molbev/msx319

White, S. A., 2010 Genes and vocal learning. Brain Lang. 115: 21–28. https://doi.org/10.1016/j.bandl.2009.10.002

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8: 973–982. https://doi.org/10.1038/nrg2165

Zdobnov, E. M., and R. Apweiler, 2001 InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848. https://doi.org/10.1093/bioinformatics/17.9.847

Zhang, G., C. Li, Q. Li, B. Li, D. M. Larkin *et al.*, 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. Science 346: 1311–1320. https://doi.org/10.1126/science.1251385

Communicating editor: M. Hufford