A high-quality genome assembly of the North American Song Sparrow, *Melospiza melodia*.

Swarnali Louha[1,*], David A. Ray[2], Kevin Winker[3], Travis Glenn[1,4]

[1]Institute of Bioinformatics, University of Georgia, Athens, GA, USA

[2]Department of Biological Science, Texas Tech University, Lubbock, TX, USA

[3]University of Alaska Museum, Fairbanks, AK, USA

[4]Department of Environmental Health Science, College of Public Health, University of Georgia, Athens, GA, USA

*Correspondence: Swarnali Louha, Institute of Bioinformatics, University of Georgia, Athens, GA 30602; E-mail: sl50708@uga.edu

Keywords: *Melospiza melodia*, Dovetail genomics, Passeriforms, whole genome sequencing, *de novo* assembly

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 coverage of 24.9×, N50 scaffold size of 5.6 Mb and N50 contig size of 31.7 Kb. Genes within our genome assembly are largely 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.

INTRODUCTION

The oscine passerines (Order Passeriformes) are songbirds having specialized vocal learning capabilities. Many species of songbirds have been widely used by neuroscientists to study the processes underlying memory and learning and social interactions (Doupe et al. 1999,

24 White 2010). The song sparrow (*Melospiza melodia*) is one of the most morphologically diverse
25 songbirds found in North America, with 26 recognized subspecies (Pruett et al. 2008). It has
26 been recognized as a model vertebrate species for field studies of birds and has been the subject
27 of extensive research integrating behavioral and ecological studies over the last 70 years (Arcese
28 et al. 2002). Though several species of songbirds have been sequenced and studied (Warren et al.
29 2010, Jarvis et al. 2014), few offer the plethora of biomedical research potential presented by the
30 song sparrow. For example, differences in neural growth and song-center brain development
31 among different populations of song sparrows could play an instrumental role in basic research
32 involving human brain neurogenesis (NIH 2001). The molecular processes underlying the
33 regeneration of "hair" cells in the auditory system of song sparrows might also provide therapies
34 useful in human hearing loss (Hawkins et al. 2003, Hawkins and Lovett 2004). Several other
35 areas of biomedical research in which the song sparrow might serve as a model system are
36 hepatic lipogenesis (Gosler 1996, Schubert et al. 2007), craniofacial development (Brugmann et
37 al. 2010, Powder et al. 2012), and variations in body size (Sutter et al. 2007, Allen et al. 2010).
38 The last of these is a polygenic trait, and elucidation of the underlying gene network affecting
39 different metabolic pathways can help clarify several biological phenomena, including human
40 diseases. Given its significant biomedical potential and experimental tractability in the field and
41 aviary, the song sparrow will continue to be used for answering research questions related to
42 mechanisms causing variation in behavior, morphology, and demographics across populations
43 (Arcese et al. 2002, Nietlisbach et al. 2015).
44 Prior work on population genetics of song sparrows in Alaska has shown how sequential
45 bottlenecking of populations in the Aleutian Archipelago has given rise to a naturally inbred
46 population with large body size (Pruett and Winker 2005). Previous work has also been done on

47 the song sparrow transcriptome, developing genomic markers to screen at population levels
48 (Srivastava et al. 2012). A high-quality genome assembly of *M. melodia* furthers the
49 development of genomic markers to screen loci associated with phenotypic traits of interest. An
50 ever-growing number of songbirds have sequenced genomes, but relatively few have been
51 published so far, including the American crow (*Corvus brachyrhynchos*), golden-collared
52 manakin (*Manacus vitellinus*; Jarvis et al. 2014), Zebra finch (*Taeniopygia guttata*; Warren et al.
53 2010), medium ground finch (*Geospiza fortis*; Parker et al. 2012) and the dark-eyed junco (*Junco*
54 *hyemalis*; Friis et al. 2018). In this study, we provide the genome assembly of *Melospiza*
55 *melodia*, a member of the family Passerellidae; this will facilitate genomic and phylogenetic
56 comparisons with other songbirds, and will also serve as a reference genome for comparative
57 genomic studies.
58 Our high-quality draft genome assembly of *M. melodia* was created by combining both
59 traditional Illumina paired-end libraries and a *de novo* proximity-ligation Chicago library. The
60 Chicago library method together with Dovetail Genomics' HiRise software pipeline is designed
61 to significantly reduce gaps in alignment arising from repetitive elements in the genome (Putnam
62 et al. 2016) and increases assembly contiguity. The draft genome was annotated using
63 transcribed RNA and protein sequences from *M. melodia* and related songbird species, *Junco*
64 *hyemalis* and *Taeniopygia guttata*. Genomic features of interest other than coding sequences,
65 such as microsatellites, repeat elements, transposable elements, and non-coding RNA, were also
66 annotated and the genome assembly was evaluated for quality by comparing it to related avian
67 species.
68
69 **METHODS**

70 **Library preparation and *de novo* shotgun assembly**

71 The *de novo* assembly of the song sparrow genome was constructed using Illumina paired
72 end libraries. Genomic DNA was isolated from a blood sample of a single male song sparrow,
73 obtained from the wild in the Aleutian Islands of Alaska on 16 Sep 2003 and archived as a
74 voucher specimen at the University of Alaska Museum
75 (http://arctos.database.museum/guid/UAM:Bird:31500). We sheared the genomic DNA using the
76 Covaris S2 (Covaris, Woburn, MA, USA) targeting a 600bp average fragment size. The sheared
77 DNA was end-repaired, adenylated, and ligated to TruSeq LT adapters using a TruSeq DNA
78 PCR-Free Library Preparation Kit (Illumina, San Diego, CA, USA). We purified the ligation
79 reaction using a Qiaquick Gel Extraction Kit (Qiagen, Venlo, The Netherlands) from a 2%
80 agarose gel. We sequenced the library on an Illumina HiSeq 2500 at the HudsonAlpha Institute
81 for Biotechnology (Huntsville, AL, USA) to obtain paired-end (PE) ∼100 bp reads. The
82 sequence data consisted of 276 million read pairs sequenced from a total of 41.3 Gbp of paired-
83 end libraries (~50× coverage). Reads were trimmed for quality, sequencing adapters, and mate
84 pair adapters using Trimmomatic (Bolger et al. 2015). The reads were assembled at Dovetail
85 Genomics (Santa Cruz, CA, USA) using Meraculous 2.0.4 (Chapman et al. 2011) with a *k-mer*
86 size of 29. This yielded a 972.4 Mbp assembly with a contig N50 of 22.5 Kbp and a scaffold N50
87 of 33 Kbp.

88 **Chicago library preparation and scaffolding the draft genome**

89 To improve the *de novo* assembly, a Chicago library was prepared at Dovetail Genomics
90 using previously described methods (Putnam et al. 2016). Around 500 ng of high-molecular-
91 weight genomic DNA (mean fragment length = 50 kbp) was used for chromatin reconstitution *in*
92 *vitro* and fixed with formaldehyde. Fixed chromatin was digested with *Dpn*II, the 5' overhangs

93 filled in with biotinylated nucleotides, and free blunt ends were ligated together. After ligation,
94 crosslinks were reversed and DNA was purified from protein. Purified DNA was treated to
95 remove biotin that was not internal to ligated fragments. Next, DNA was sheared to ~350 bp
96 mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes
97 (New England Biolabs, Ipswich, MA, USA) and Illumina-compatible adapters. Biotin-containing
98 fragments were isolated using streptavidin beads before PCR enrichment of the library. The
99 Chicago library was sequenced on an Illumina HiSeq 2500 to produce 47 million 150 bp paired
100 end reads (1-50 kb pairs).

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

108 **Identification of microsatellites and transposable elements**

109 Transposable elements (TEs) in the song sparrow genome were identified using a
110 combination of *de novo* and homology-based TE identification methods, in addition to a manual
111 curation step (Platt et al. 2016). First, we used RepeatModeler v1.0.11 (Smit and Hubley 2008-
112 2015) with default parameters (File S1) to generate a custom repeat library consisting of 672
113 consensus repeat sequences. RepeatModeler uses two *de novo* repeat identification programs,
114 RECON v1.08 (Bao and Eddy 2002) and RepeatScout v1.0.6 (Pevzner et al. 2005), for
115 identifying repetitive elements from sequence data. To ensure accurate and complete

116 representation of putative TEs, the RepeatModeler derived consensus sequences were filtered for
117 size (>100 bp), and then subjected to iterative homology-based searches against the genome,
118 followed by manual curation (Platt et al. 2016). The final set of manually curated TEs was
119 queried against CENSOR (Kohany et al. 2006) and TEclass (Abrusan et al. 2009) for
120 classification. TEs not identifiable in CENSOR were also searched against the NCBI nucleotide
121 and protein databases using BLASTN and BLASTX respectively. Finally, a custom *de novo*
122 repeat library comprising song sparrow-specific TEs and existing simple and low complexity
123 repeats in birds was used to screen for repeats in the song sparrow genome assembly with
124 RepeatMasker v4.0.9.

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

128 ***De novo* gene annotation and function prediction**

129 Genes were predicted in the song sparrow genome with the MAKER v2.31.9
130 genome annotation pipeline (Campbell et al. 2014). A custom repeat library of 900 repeat
131 sequences consisting of TEs identified in the song sparrow genome and other existing avian
132 repeat elements was used to soft mask the genome. Transcriptome evidence sets for MAKER
133 included the assembled song sparrow transcriptome (Srivastava et al. 2012) and Trinity (v2.4.0)
134 mRNA-seq assemblies from multiple tissues of *Junco hyemalis* (Peterson et al. 2012, NCBI
135 BioProject Accession: PRJNA256328). Protein evidence sets used by MAKER included
136 annotated proteins for song sparrow, *Junco hyemalis*, and *Taeniopygia guttata* from the NCBI
137 Protein database. The MAKER pipeline consisted of the following steps: 1) Transcriptomic and
138 protein evidence sets were used to make initial evidence-based annotations with MAKER; 2) the

139 initial annotations were used to train two *ab initio* gene predicters: Augustus (Stanke et al. 2006),
140 which was trained once, and SNAP (Korf 2004), which was iteratively trained twice; and 3) the
141 trained gene prediction tools SNAP and Augustus were used to generate the final set of gene
142 annotations (File S3-S8).

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

148 To quantitatively assess the completeness of the song sparrow genome assembly and
149 annotated gene set, we ran BUSCO (Benchmarking Universal Single-Copy Orthologs) v3.0.2
150 (Waterhouse et al. 2017) with 4,915 single-copy orthologous genes in the Aves lineage group
151 (Aves_odb9; https://busco.ezlab.org/), using "chicken" as the Augustus reference species (File
152 S10). The 4,915 orthologous genes are present in at least 90% of the 40 species included within
153 the Aves lineage group, and thus are likely to be found in the genome of related species. We also
154 used the JupiterPlot pipeline (https://github.com/JustinChu/JupiterPlot) to visually compare the
155 zebra finch (*T. guttata*) genome assembly (Warren et al. 2010) to our assembly in a Circos plot,
156 using the largest scaffolds making up 95% of our genome assembly, and all scaffolds greater
157 than 100 kbp in the Zebra finch genome (File S11). The zebra finch was selected because it was
158 the first fully sequenced songbird genome, it is one of the most complete bird genomes, and it is
159 often used for comparative genomic studies in birds.

160 **Non-coding RNA prediction**

161 Transfer RNAs (tRNAs) were predicted in the song sparrow genome with tRNAscan-SE
162 v2.0 (Lowe and Chan 2016, File S12). A training set comprising eukaryotic tRNAs was used to
163 train the covariance models employed by tRNAscan-SE, and tRNAs were searched against the
164 genome with Infernal v1.1.2 (Nawrocki et al. 2014). tRNAscan-SE also provides functional
165 classification of tRNAs based on a comparative analysis using a suite of isotype-specific tRNA
166 covariance models. A random sample of 10 predicted tRNAs were selected and searched against
167 the tRNA databases GtRNAdb (Chan and Lowe 2016) and tRNAdb (Juhling et al.2009).
168 Identification of miRNAs (microRNAs), snoRNAs (small nucleolar RNAs), snRNAs
169 (small nuclear RNAs), rRNAs (ribosomal RNAs), and lncRNAs (long non-coding RNAs) was
170 achieved by using a homology-based prediction method. Structural homologs to eukaryotic
171 ncRNA covariance models from the Rfam database v14.1 (Gardner et al. 2009) were searched
172 against the song sparrow genome using Infernal's "cmscan" program (File S13). All low-scoring
173 overlapping hits and hits with an E-value greater than 10^{-5} were discarded, and the remaining
174 ncRNAs were grouped into different classes.
175 Lastly, we compared the predicted classes of different ncRNAs in the song sparrow
176 genome to those reported in the genomes of related birds, *Taeniopygia guttata* and *Ficedula*
177 *albicollis* (collared flycatcher).
178

179 **DATA AVAILABILITY**

180 Raw reads have been deposited in the NCBI Sequence Read Archive (SRR10491484 and
181 SRR10451714 for the Meraculous assembly, and SRR10424475 for the Chicago HiRise
182 assembly). Supplemental File S1 contains submission script for RepeatModeler. Supplemental
183 File S2 contains primary configuration file used to run GMATA (default_cfg.txt). Supplemental

184 File S3 contains submission script for MAKER. Supplemental File S4 contains MAKER
185 executable file (maker_exe.ctl). Supplemental File S5 contains specifications for downstream
186 filtering of BLAST and Exonerate alignments (maker_bopts.ctl). Supplemental File S6 contains
187 primary configuration of MAKER specific options (maker_opts.ctl). Supplemental File S7
188 contains scripts for training SNAP. Supplemental File S8 contains scripts for training Augustus.
189 Supplemental File S9 contains scripts for running BLASTP and InterProScan for functional
190 annotation of predicted genes; and scripts for adding the functional annotations to gene
191 annotation files. Supplemental File S10 contains submission script for BUSCO. Supplemental
192 File S11 contains submission scripts for JupiterPlot pipeline. Supplemental File S12 contains
193 submission script for tRNAscan-SE. Supplemental File S13 contains submission script for
194 Infernal. Supplemental File S14 contains classification of predicted transposable elements.
195 Supplemental File S15 contains annotation of microsatellites with their genomic locations.
196 Supplemental File S16 contains percentage of different microsatellites present in the genome.
197 Supplemental File S17 contains frequency of occurrence of microsatellites in each scaffold of the
198 genome. Supplemental File S18 contains the distribution of the length of microsatellites.
199 Supplemental File S19 contains predicted function of annotated genes by BLASTP.
200 Supplemental File S20 contains prediction of protein domains, GO annotations and pathway
201 annotations of predicted genes by InterProScan. Supplemental File S21 contains sequence and
202 structure of tRNAs identified in the Song Sparrow genome. Supplemental File S22 contains
203 classification of predicted tRNAs. Supplemental File S23 contains classification of different
204 ncRNAs predicted in the genome with Infernal. Supplemental File S24 contains the M. melodia
205 Chicago HiRise genome sequence (Mmel_1.0). Supplemental File S25 contains annotations for
206 the M. melodia genome assembly. Supplemental Table S1 contains genome sizes of birds related

207 to M. melodia. Supplemental Figure S1 contains the distribution of the percentage of annotated
208 genes with their corresponding AED scores.
209

210 **RESULTS AND DISCUSSION**

211 **Assembly**

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

219 **Microsatellites and Transposable Elements**

220 In total, 89 as yet unnamed TEs were identified in the song sparrow genome. Fifty-five of
221 these did not have any significant matches in CENSOR (Kohany et al. 2006) and are considered
222 novel (File S14). A TE was considered to have a significant match to a known element in
223 CENSOR only when it had a length of at least 80 bp and 80% identity to the known element over
224 80% of its length, the 80-80-80 rule (Schulman et al. 2007). The predicted TEs were classified
225 into DNA transposons and retrotransposons (i.e. LINEs, LTRs, and SINEs) using CENSOR and
226 TEclass (File S14). Approximately 7.4% of the genome comprises repeats with the majority of
227 that consisting of TEs (~ 48%). Among the different TEs, LTRs (~ 40%) and LINEs (~ 49%)
228 were found to be most abundant (Table 2). The song sparrow genome assembly was found to be

229 less repetitive when compared to sequenced genomes of related songbirds, primarily due to the
230 lower content of LTRs and LINEs than other songbirds (Fig 2).
231 Overall, 112,419 microsatellites with motifs ranging in size from 2-20 bp were found in
232 the song sparrow genome (File S15 contains all microsatellites with their genomic locations).
233 The majority of the microsatellites were made up of 2-, 3-, 4-, and 5-mers, with 2-mers making
234 up about 71% of all microsatellites identified (Fig 3, File S16). The distribution of the top base-
235 pair composition of microsatellite motifs present in the genome is shown in Fig 4. The frequency
236 of occurrence of microsatellites in every scaffold and a distribution of their lengths are provided
237 in Files S17 and S18, respectively.
238 **Gene annotation and function prediction**
239 The MAKER genome annotation pipeline predicted 15,086 genes and 139 pseudogenes
240 in the song sparrow genome, fewer than *T. guttata, F. albicollis*, and *M. vitellinus*, but higher
241 than *G. fortis* (Table 3). The average gene length, exon length, intron length, and the total
242 number of exons and introns predicted are also less compared to closely related species (Table
243 3). Out of the 15,086 predicted genes, 12,541 genes were assigned putative functions with
244 BLASTP (File S19). InterProScan assigned functional domains to 11,298 (74.9%) predicted
245 genes (File S20). A total of 7,010 genes obtained GO annotations. Pathway annotations were
246 assigned to 2,716 genes.
247 Annotated genes were assigned annotation edit distance (AED) scores with values
248 ranging from 0 to 1. AED is a distance metric score that signifies how closely gene models
249 match transcript and protein evidence. Gene models with AED scores closer to 0 have better
250 alignment with the evidence provided in the MAKER pipeline. A distribution of the percentage

251 of genes with their corresponding AED scores shows close similarity of the annotated genes with
252 the transcript and protein evidence provided in the MAKER pipeline (Fig S1).
253 The song sparrow genome assembly contained 4,318 complete universal single-copy
254 orthologs (BUSCOs; 87.9%) from a total of 4,915 BUSCO groups searched. Among all complete
255 BUSCOs, 99.4% were present as single-copy genes and 0.6% were duplicated. About 7.4%
256 (356) of the orthologous gene models were partially recovered, and 4.9% (241) had no
257 significant matches. The incomplete and missing gene models could either be partially present or
258 missing, or could indicate genes that are too divergent or have very complex structures, making
259 their prediction difficult. Incomplete and missing gene models could also suggest problems
260 associated with the genome assembly and gene annotation. The results from the BUSCO analysis
261 are in agreement with the Circos plot (Fig 5), in which few scaffolds in the zebra finch genome
262 assembly are not represented in our assembly and very few small translocations exist between
263 the two genome assemblies.
264 **Non-coding RNA prediction and identification**
265 A total of 267 tRNAs were detected in the song sparrow genome by tRNAscan-SE (see
266 File S21 for sequence and structure of tRNAs), out of which 129 were found coding for the
267 standard twenty amino acids. The predicted output from tRNAscan-SE (File S22) contained 114
268 tRNAs with low Infernal as well as Isotype scores; these were characterized as pseudogenes
269 lacking tRNA-like secondary structures (Lowe and Chan 2016). Two tRNAs had undetermined
270 isotypes and 22 were chimeric, with mismatch isotypes. Chimeric tRNAs contain point
271 mutations in their anticodon sequence, rendering different predicted isotypes than those predicted
272 by structure-specific tRNAscan-SE covariance models. Among all predicted tRNAs, 11
273 contained introns within their sequences. No suppressor tRNAs and tRNAs coding for

274 selenocysteine were predicted. The subset of 10 randomly selected tRNAs was also predicted in
275 many other species in both GtRNAdb and tRNAdb databases.
276 Infernal searches predicted a total of 364 ncRNAs in the song sparrow genome,
277 comprising 166 miRNAs, 8 rRNAs, 154 snoRNAs, 16 snRNAs, and 20 lncRNAs (File S23).
278 Compared to the genomes of related avian species (*T. guttata* and *F. albicollis*), the song sparrow
279 genome has the highest number of predicted tRNAs, but fewer other ncRNAs (Table 4).
280

281 **CONCLUSION**

282 The Chicago and shotgun sequencing libraries along with the HiRise assembly software
283 enabled accurate and highly contiguous *de novo* assembly of the song sparrow genome. The
284 genome assembly is 978.3 Mb, with 48 scaffolds (L50) making up half the genome size. A
285 previous estimate of genome size of *M. melodia* from densitometry analysis provided a C-value
286 of 1.43 pg (1,398.54 Mb) (Gregory et al. 2009). Our own *k-mer* based estimate of genome size
287 from paired reads in the shotgun and Chicago libraries using Kmergenie v1.7044 (Chikhi and
288 Medvedev 2013) yielded an estimated size of 1,127.25 Mb. Both these genome size estimates
289 and the genome sizes of related birds (Table S1) are slightly higher than our genome assembly
290 (978.3 Mb); this could be due to the low TE content observed in the song sparrow. Vertebrate
291 genome sizes are typically correlated with TE content and birds are no exception (Kidwell 2002,
292 Gregory and Elliott 2015). Indeed, birds, along with bats, are proposed to have been subject to
293 extensive DNA loss via an 'accordion model' of genome evolution (Kapusta et al. 2017). Our
294 data suggest that, among birds, the song sparrow may have been particularly susceptible to the
295 segmental deletions that are critical to this model. The small genome size of the song sparrow
296 can also be attributed to the compression of repetitive regions, which is generally observed in

297 assemblies generated from short-read sequencing data. This is also consistent with the fact that
298 our genome contains fewer repeats when compared to related songbirds (Fig 2). Although short
299 reads limit our ability to characterize all genomic repeats, we have been able to resolve the vast
300 majority of repeats into LINEs, SINEs, LTRs, and DNA retrotransposons (Fig 2, Table 2).

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

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

ACKNOWLEDGEMENTS

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 also thank David Sonneborn and Jack Withrow for their roles in making a vouchered specimen available for our work. We thank 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.

337 **References**

338 Abrusan G., Grundmann N., DeMeester L., Makalowski W., 2009 TEclass: a tool for automated
339 classification of unknown eukaryotic transposable elements. Bioinformatics. 25:1329-1330
340

341 Andrews C.B., Mackenzie S.A., Gregory T.R., 2009 Genome size and wing parameters in
342 passerine birds. Proc Biol Sci. 276(1654):55-61. doi:10.1098/rspb.2008.1012
343

344 Arcese P., Sogge M.K., Marr A.B., Patten M.A., 2002 Song Sparrow (*Melospiza melodia*) In:
345 Poole A., Gill F., editors. The Birds of North America. Philadelphia, PA: The Birds of North
346 America Inc. No. 704
347

348 Bolger A.M., Lohse M., Usadel B., 2014 Trimmomatic: a flexible trimmer for Illumina sequence
349 data. Bioinformatics. 30: 2114–2120. doi:10.1093/bioinformatics/btu170
350

351 Brugmann S.A., Powder K.E., Young N.M., Goodnough L.H., Hahn S.M., James A.W., Helms
352 J.A., Lovett M., 2010 Comparative gene expression analysis of avian embryonic facial structures
353 reveals new candidates for human craniofacial disorders. Human Molecular genetics. 19: 920-
354 930
355

356 Campbell M. S., Holt C., Moore B., Yandell M., 2014 Genome annotation and curation using
357 MAKER and MAKER-P. Current Protocols in Bioinformatics. 48:4.11.1–39.
358 doi:10.1002/0471250953.bi0411s48
359

360 Chan P. P., Lowe T. M., 2016 GtRNAdb 2.0: An expanded database of transfer RNA genes
361 identified in complete and draft genomes. Nucleic Acids Res. 44: D184–D189.
362 doi:10.1093/nar/gkv1309
363
364 Chapman J.A., Ho I., Sunkara S., Luo S., Schroth G.P., Rokhsar D.S., 2011 Meraculous: de novo
365 genome assembly with short paired-end reads. PLoS One. 6(8):e23501.
366 doi:10.1371/journal.pone.0023501
367
368 Chikhi R, Medvedev P., 2014 Informed and automated k-mer size selection for genome
369 assembly. Bioinformatics. 30(1):31-37. doi: 10.1093/bioinformatics/btt310
370
371 Doupe A.J., Kuhl P.K., 1999 Birdsong and Human Speech: Common Themes and Mechanisms.
372 Annual Review of Neuroscience. 22(1) 567-631. DOI: 10.1146/annurev.neuro.22.1.567
373 Elliott T.A., Gregory T.R., 2015 What's in a genome? The C-value enigma and the evolution of
374 eukaryotic genome content. Philos Trans R Soc Lond B Biol Sci. 370(1678):20140331. doi:
375 10.1098/rstb.2014.0331
376
377 Friis G., Fandos G., Zellmer A.J., McCormack J.E.,4, Faircloth B.C., Milá B., 2018 Genome-
378 wide signals of drift and local adaptation during rapid lineage divergence in a songbird. Mol
379 Ecol. 27(24):5137-5153. doi:10.1111/mec.14946
380

381 Gardner P.P., Daub J., Tate J., Moore B.L., Osuch I.H., Griffiths-Jones S., Finn R.D., Nawrocki
382 E.P., Kolbe D.L., Eddy S.R., Bateman A., 2011 Rfam: Wikipedia, clans and the 'decimal'
383 release. Nucleic Acids Res. 39(Database issue):D141-5. doi:10.1093/nar/gkq1129
384
385 Gosler A.G., 1996 Environmental and social determinants of winter fat storage in the Great Tit
386 Parus major. J. Animal Ecology. 65:1-17
387
388 Hawkins R.D., Bashiardes S., Helms C.A., Hu L., Saccnone N.L., Warchol M.E., Lovett M.,
389 2003 Gene expression differences in quiescent versus regenerating hair cells of avian sensory
390 epithelia: implications for human hearing and balance disorders. Human Molecular Genetics. 12:
391 1261-1272
392
393 Hawkins R.D., Lovett M., 2004 The developmental genetics of auditory hair cells. Human
394 Molecular Genetics. 13: R289-R296
395
396 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., Faircloth B.C., et al., 2014
397 Whole-genome analyses resolve early branches in the tree of life of modern birds. Science.
398 12;346(6215):1320-31. doi: 10.1126/science.1253451
399
400 Juhling F., Morl M., Hartmann R.K., Sprinzl M., Stadler P.F., Pütz J., 2009 tRNAdb 2009:
401 Compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 37: D159–D162.
402 doi:10.1093/nar/gkn772
403

404 Kapusta A., Suh A., Feschotte C., 2017 Dynamics of genome size evolution in birds and
405 mammals. Proc Natl Acad Sci USA. 114(8):E1460-E1469. doi:10.1073/pnas.1616702114
406
407 Kidwell M.G., 2002 Transposable elements and the evolution of genome size in eukaryotes.
408 Genetica. 115(1):49-63
409
410 Kohany O., Gentles A.J., Hankus L., Jurka J., 2006 Annotation, submission and screening of
411 repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 25;7:474
412
413 Korf I., 2004 Gene finding in novel genomes. BMC Bioinformatics 5:59. doi:10.1186/1471-
414 2105-5-59
415
416 Lango A.H., Estrada K., Lettre G., Berndt S.I., Weedon M.N., et al., 2010 Hundreds of variants
417 clustered in genomic loci and biological pathways affect human height. Nature. 467:832-838
418
419 Lowe T.M., Chan P.P., 2016 tRNAscan-SE On-line: Search and Contextual Analysis of Transfer
420 RNA Genes. Nucleic Acids Res. 44:W54-57
421
422 Nawrocki E.P., 2014 Annotating functional RNAs in genomes using Infernal. Methods Mol Biol.
423 1097:163–97. doi: 10.1007/978-1-62703-709-9_9
424

425 Nietlisbach P., Camenisch G., Bucher T., Slate J., Keller L.F., Postma E., 2015 A microsatellite-
426 based linkage map for song sparrows (*Melospiza melodia*). Molecular Ecology Resources
427 15:1486-1496
428
429 NIH (2001) What we learned from songbirds: The adult brain can generate new nerve cells. NIH
430 Publication No. 01-4602
431
432 Parker P., Li B., Li H., Wang J., 2012 The genome of Darwin's Finch (Geospiza fortis).
433 GigaScience. http://dx.doi.org/10.5524/100040
434
435 Peterson M.P., Whittaker D.J., Ambreth S., Sureshchandra S., Buechlein A., Podicheti R., Choi
436 J.H., Lai Z., Mockatis K., Colbourne J., Tang H., Ketterson E.D., 2012 De novo transcriptome
437 sequencing in a songbird, the dark-eyed junco (*Junco hyemalis*): genomic tools for an ecological
438 model system. BMC Genomics. 13:305. doi: 10.1186/1471-2164-13-305
439
440 Platt R.N. 2nd, Blanco-Berdugo L., Ray D.A., 2016 Accurate Transposable Element Annotation
441 Is Vital When Analyzing New Genome Assemblies. Genome Biology and Evolution. 8(2):403-
442 10. doi: 10.1093/gbe/evw009
443
444 Powder K.E., Ku Y.C., Brugmann S.A., Veile R.A., Renaud N.A., Helms J.A., Lovett M., 2012
445 A cross-species analysis of microRNAs in the developing avian face. PLoS One. 7(4):e35111.
446 doi: 10.1371/journal.pone.0035111
447

448 Price A.L., Jones N.C., Pevzner P.A., 2005 De novo identification of repeat families in large
449 genomes. Bioinformatics. Suppl 1:i351-8
450
451 Pruett C.L., Arcese P., Chan Y.L., Wilson A.G., Patten M.A., Keller L.F., Winker K., 2008
452 Concordant and discordant signals between genetic data and described subspecies of Pacific
453 Coast Song Sparrows. The Condor. 110(2) 359-364. DOI: 10.1525/cond.2008.8475
454
455 Pruett C.L., Winker K., 2005 Northwestern Song Sparrow populations show genetic effects of
456 sequential colonization. Molecular Ecology. 14:1421-1434
457
458 Pundir S., Martin M. J., O'Donovan C., 2016 UniProt Tools. Curr. Protoc. Bioinformatics.
459 53:1.29.1-15. doi: 10.1002/0471250953.bi0129s53
460
461 Putnam N.H., O'Connell B.L., Stites J.C., Rice B.J., Blanchette M., Calef R., Troll C.J., Fields
462 A., Hartley P.D., Sugnet C.W., Haussler D., Rokhsar D.S., Green R.E., 2016 Chromosome-scale
463 shotgun assembly using an in vitro method for long-range linkage. Genome Research. 26:342-
464 350
465
466 Schubert K.A., Mennill D.J., Ramsay S.M., Otter K.A., Boag P.T., Ratcliffe L.M., 2007
467 Variation in social rank acquisition influences lifetime reproductive success in black-capped
468 chickadees. Biol. J. Linn. Soc. 90:85-95
469

470 Smit A.F.A., Hubley R., RepeatModeler. Open-1.0.11, 2008–2015
471 (http://www.repeatmasker.org)
472
473 Srivastava A., Winker K., Shaw T.I., Jones K.L., Glenn T.C., 2012 Transcriptome analysis of a
474 North American songbird, Melospiza melodia. DNA Research. 19:325-333
475
476 Stanke M., Tzvetkova A., Morgenstern B., 2006 AUGUSTUS at EGASP: using EST, protein
477 and genomic alignments for improved gene prediction in the human genome. BMC Genome
478 Biology. 7(Suppl 1):S11. 1-8
479
480 Sutter N.B., Bustamante C.D., Chase K., Gray M.M., Zhao K., Zhu L., Padhukasahasram B.,
481 Karlins E., Davis S., Jones P.G., Quignon P., Johnson G.S., Parker H.G., Fretwell N., Mosher
482 D.S., Lawler D.F., Satyaraj E., Nordborg M., Lark K.G., Wayne R.K., Ostrander E.A., 2007 A
483 single IGF1 allele is a major determinant of small size in dogs. Science. 316(5821):112-115
484
485 Wang X., Wang L., 2016 GMATA: An Integrated Software Package for Genome-Scale SSR
486 Mining, Marker Development and Viewing. Front Plant Sci. 13;7:1350. doi:
487 10.3389/fpls.2016.01350. eCollection 2016
488
489 Warren W.C., Clayton D.F., Ellegren H., Arnold A.P., Hillier L.W., et al., 2010 The genome of a
490 songbird. Nature. 464(7289):757-62. doi: 10.1038/nature08819
491

492 Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G.,
493 Kriventseva E.V., Zdobnov E.M., 2017 BUSCO applications from quality assessments to gene
494 prediction and phylogenomics. Mol Biol Evol. doi: 10.1093/molbev/msx319
495
496 White S.A., 2010 Genes and vocal learning, Brain and Language. 115(1) 21-28.
497 DOI:10.1016/j.bandl.2009.10.002
498
499 Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., Leroy P.,
500 Morgante M., Panaud O., Paux E., SanMiguel P., Schulman A.H., 2007 A unified classification
501 system for eukaryotic transposable elements. Nat Rev Genet. 8(12):973-82
502
503 Z. Bao, S. R. Eddy., 2002 Automated de Novo Identification of Repeat Sequence Families in
504 Sequenced Genomes. Genome Research. 12:1269-1276
505
506 Zdobnov E.M., Apweiler R., 2001 InterProScan-an integration platform for the signature-
507 recognition methods in InterPro. Bioinformatics. 17:847-848
508
509 Zhang G., Li C., Li Q., Li B., Larkin D.M., Lee C., Storz J.F., Antunes A., et al., 2014
510 Comparative genomics reveals insights into avian genome evolution and adaptation. Science.
511 346(6215):1311-20. doi:10.1126/science.1251385
512
513
514

515 **Tables and Figures**

516 Table 1: A comparison of assembly quality statistics from the initial shotgun sequencing

517 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%

518

519

520 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**

521

522

523 Table 3: Numbers of genes predicted in the *M. melodia* genome compared to *Taeniopygia*
524 *guttata* (zebra finch) and *Ficedulla albicollis* (collared flycatcher), *Manacus vitellinus* (golden-
525 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

526

527 [1]https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Taeniopygia_guttata/103/

528 [2]https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ficedula_albicollis/101/

529 [3]https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Manacus_vitellinus/102/

530 [4]https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Geospiza_fortis/101/

531

532 Table 4: Numbers of ncRNAs predicted in the *Melospiza melodia* genome compared to

533 *Taeniopygia guttata* (zebra finch) and *Ficedulla albicollis* (collared flycatcher).

534

	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

535

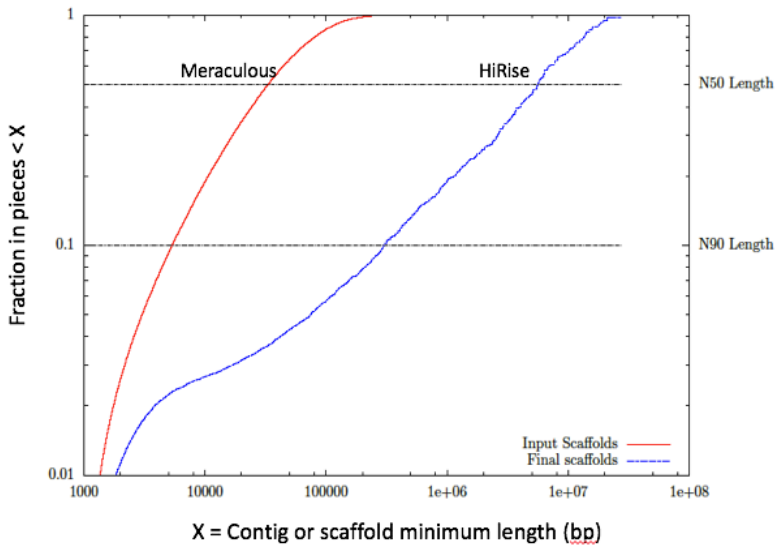
536 [1]http://useast.ensembl.org/info/data/ftp/index.html

537 [2]https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Taeniopygia_guttata/103/

538 [3]https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ficedula_albicollis/101/

539

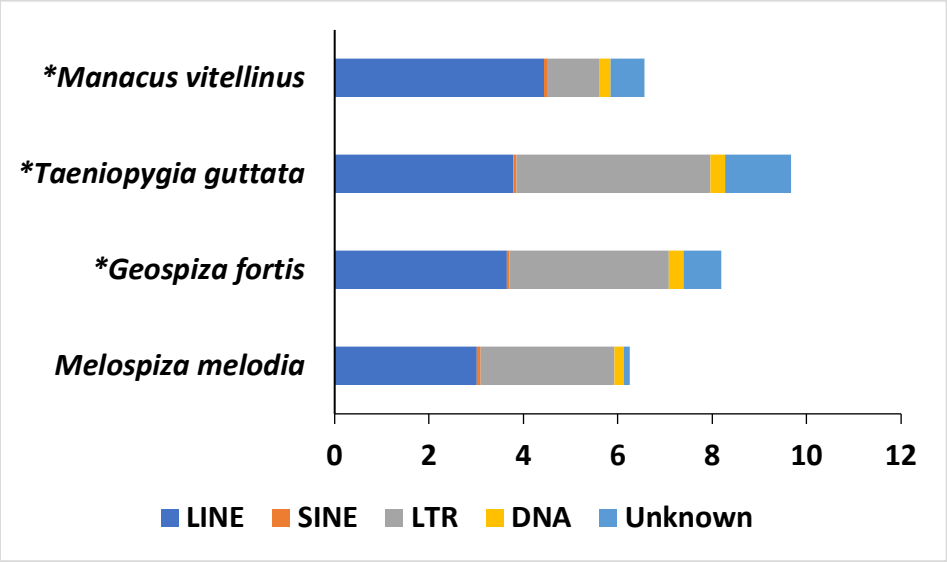
540 Figure 1: Comparison of assembly contiguity.



541

542

543 Figure 2: Comparison of percentages of transposable elements (TEs) among related songbirds.



544

545 *Zhang et al. Science. 2014 Dec 12;346(6215):1311-20. DOI: 10.1126/science.1251385.

546

547 Figure 3: Distribution of microsatellite repeat motif size classes in the *M. melodia* genome

548 assembly (details are given in Supplemental File S15).

549

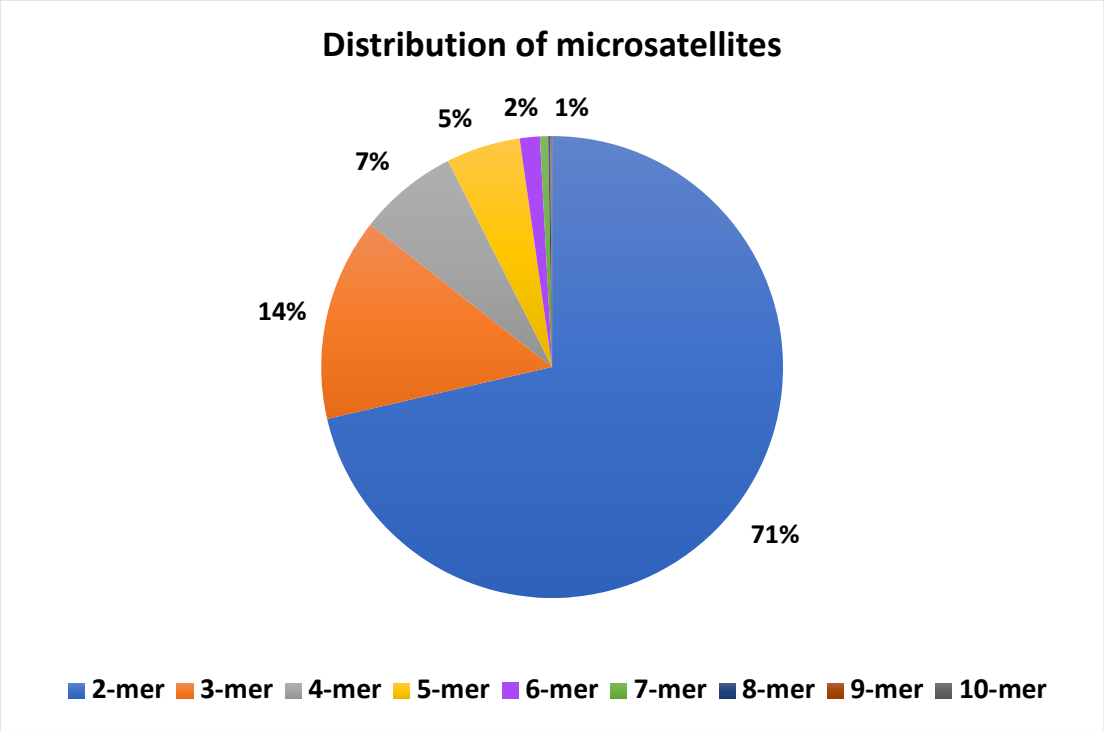


Figure 4: Distribution of the top base-pair composition of microsatellite motifs in the *M. melodia* genome

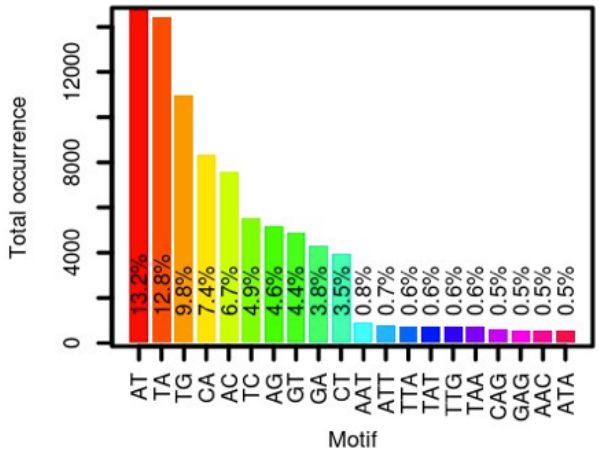
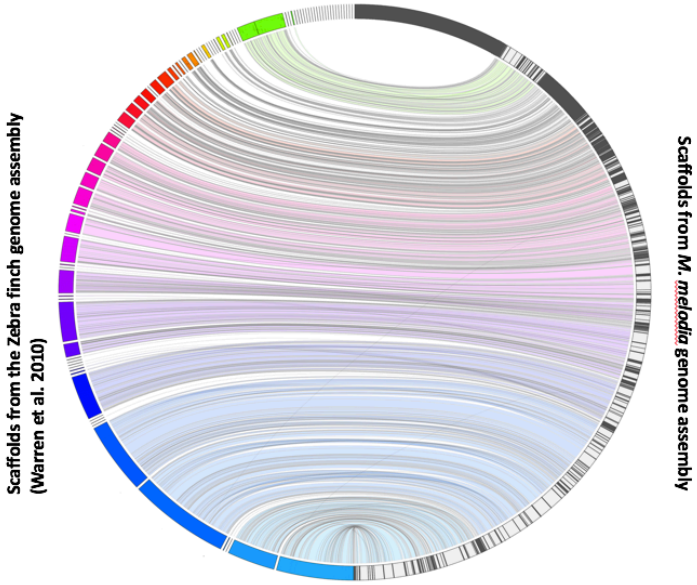



Figure 5: Jupiter plot correlating zebra finch and song sparrow genome assemblies, considering scaffolds greater than 100 kbp in the reference zebra finch genome and the largest scaffolds representing 95% of the query song sparrow genome.