A high-quality genome assembly and annotation of the dark-eyed junco Junco hyemalis, a recently diversified songbird

Friis, Guillermo; Vizueta, Joel; Ketterson, Ellen D.; Milá, Borja

Published in:
G3: Genes, Genomes, Genetics

DOI:
10.1093/g3journal/jkac083

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

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

Guillermo Friis, Joel Vizueta, Ellen D. Ketterson, Borja Milá

1Department of Biodiversity and Evolutionary Biology, National Museum of Natural Sciences, Spanish National Research Council (CSIC), Madrid 28006, Spain,
2Centre for Social Evolution, University of Copenhagen, Copenhagen 1165, Denmark,
3Department of Biology, Indiana University, Bloomington, IN 47405, USA

*Corresponding author: Center for Genomics and Systems Biology, New York University—Abu Dhabi, PO Box 129188, Abu Dhabi, UAE. Email: gfm3@nyu.edu

Abstract

The dark-eyed junco (*Junco hyemalis*) is one of the most common passerines of North America, and has served as a model organism in studies related to ecophysiology, behavior, and evolutionary biology for over a century. It is composed of at least 6 distinct, geographically structured forms of recent evolutionary origin, presenting remarkable variation in phenotypic traits, migratory behavior, and habitat. Here, we report a high-quality genome assembly and annotation of the dark-eyed junco generated using a combination of shotgun libraries and proximity ligation Chicago and Dovetail Hi-C libraries. The final assembly is ~1.03 Gb in size, with 98.3% of the sequence located in 30 full or nearly full chromosome scaffolds, and with a N50/L50 of 71.3 Mb/5 scaffolds. We identified 19,026 functional genes combining gene prediction and similarity approaches, of which 15,967 were associated to GO terms. The genome assembly and the set of annotated genes yielded 95.4% and 96.2% completeness scores, respectively when compared with the BUSCO avian dataset. This new assembly for *J. hyemalis* provides a valuable resource for genome evolution analysis, and for identifying functional genes involved in adaptive processes and speciation.

Keywords: dark-eyed junco; *Junco hyemalis*; genome assembly; Hi-C

Introduction

The dark-eyed junco (*Junco hyemalis*) is a common and widespread North American passerine that has been the subject of extensive research in multiple scientific disciplines for over 100 years (Miller 1941; Ketterson and Atwell 2016). The species present at least 6 distinct and geographically structured forms, showing marked levels of divergence in plumage coloration, habitat and life-history traits, including the timing of reproduction or migratory behavior (Fig. 1; Supplementary Table 1; Miller 1941; Nolan et al. 2002). These forms represent independent evolutionary lineages that radiated after the Last Glacial Maximum circa 18,000 years ago during a northward recolonization of the North American continent (Milá et al. 2007; Friis et al. 2022). Previous studies based on genome-wide data have documented a unique case of rapid diversification driven by the combined effects of sexual selection, ecological habitat features, and historical demographic factors (Friis et al. 2018; Friis and Milá 2020). The dark-eyed junco has also served as a model organism for the study of endocrine, neurological and behavioral aspects of migration, reproductive phenology (e.g. Ketterson and Nolan 1976; Cristol et al. 2003; Fudíckar et al. 2017; Singh et al. 2021) and other life-history traits (Ketterson and Nolan 1999); ecophysiology (e.g. Stager et al. 2015, 2021); the hormonal basis of phenotypic variation, including traits related to sexual dimorphism, courtship behavior and mate choice (e.g. Ketterson et al. 1996, 2005, 2009); differential gene expression associated with pigmentation (Abolins-Abols et al. 2018); and phenotypic plasticity and adaptation to urban environments (e.g. Yeh and Price 2004; Atwell et al. 2014; Friis et al. 2022). For a review, see Ketterson and Atwell (2016). Despite the extensive research conducted on dark-eyed juncos, the molecular basis of the outstanding diversity found in the group remains poorly understood, in part because high-quality, full genome accounting for such variability have been sparse.

A plethora of avian species has been fully sequenced and analyzed in the last decade (e.g. Warren et al. 2010; Zhang et al. 2012; Jarvis et al. 2014; Poelstra et al. 2014; Liu et al. 2019; Louha et al. 2020; Dussex et al. 2021; Recuerda et al. 2021). Recently, the Ten-Thousand Bird Genomes (B10K) consortium released over 300 avian genomes from 92.4% of all avian families, enabling genome-based comparative studies at phylogenetic scales never achieved before (Zhang et al. 2014; Feng et al. 2020). Yet within birds, the dark-eyed junco complex is of particular interest for the study of evolutionary processes. First, the system (which includes the closely related yellow-eyed junco *Junco phaeonotus* from Mexico and Central America), is composed of recently diversified yet phenotypically differentiated lineages, among which the signals of drift and selection at the molecular level are still recent and detectable. Second, the complex includes forms with broad geographic distributions encompassing heterogeneous...
habitats across ecological clines, but also spatially discontinuous habitats so that selective and neutral processes of divergence can be assessed in different spatial settings. Third, dark-eyed juncos show large variability in the degree of geographical isolation among phenotypically differentiated forms, from extensive gene flow to total isolation, which along with the first 2 points, makes them a suitable system for studying evolutionary processes related to dispersal, directional selection, and neutral evolution (Milaé et al. 2006; Friis et al. 2018, 2022). Fourth, the dark-eyed juncos present highly divergent secondary sexual traits, enabling comparative analyses to test the role of sexual selection as a driver of genetic divergence (Friis and Milaé 2020; Friis et al. 2022).

And finally, as in most bird species, the dark-eyed junco genome is highly syntenic within the class Aves, relatively small, and structurally simple compared with other amniotes (Gregory 2002), which facilitates the identification of polymorphisms and structural variants.

Here, we report a high-quality, chromosome-level assembly obtained using shotgun and proximity ligation libraries as a resource for genome-based studies on J. hyemalis and related bird species. A structural and functional annotation using gene prediction and similarity approaches is also provided. To evaluate the performance of the assembly as a reference for genome-based studies, we conducted a demographic analysis using the pairwise sequentially Markovian coalescent method (PSMC) and the shotgun data generated for the assembly. Shotgun data are also used for the assembly of the mitochondrial genome.

Materials and methods

Genome sequencing and assembly

A high-quality genome was produced combining newly generated shotgun reads and sequence data from proximity ligation libraries. Preparation of proximity ligation libraries Chicago and Hi-C, as well as scaffolding with the software pipeline HiRise (Putnam et al. 2016; https://dovetailgenomics.com, last accessed on April 10, 2022) was conducted at Dovetail Genomics, LLC. The sequenced sample consisted of muscle tissue obtained from a female J. hyemalis carolinensis, collected at Mountain Lake Biological Station in Pembroke, Virginia, USA (37.3751°N, 80.5228°W), currently deposited at the Moore Laboratory of Zoology, Occidental College, Los Angeles, CA, USA (voucher number: MLZ: bird: 69236). Briefly, a de novo draft assembly was first built using shotgun, paired-end libraries (mean insert size ~350 bp) and the Meraculous pipeline (Chapman et al. 2011). For the Chicago and the Dovetail Hi-C library preparation, chromatin was fixed with formaldehyde. Fixed chromatin was then digested with DpnII and free blunt ends were ligated. Crosslinks were reversed and the DNA was purified from protein. Resulting nucleic material was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Sequencing of the libraries was carried out on an Illumina HiSeq X platform. The shotgun reads, Chicago library reads, and Dovetail Hi-C library reads were then used as input data for HiRise, a software pipeline designed specifically for using proximity ligation data to scaffold genome assemblies (Supplementary Fig. 1, Putnam et al. 2016). More details on genome sequencing and assembly methods are available in the Supplementary Information.

A circularized assembly of the mitochondrial genome was generated using NOVOplasty2.7.2 (Dierckxsens et al. 2017) and the shotgun data. The NADH dehydrogenase subunit 2 (ND2) mitochondrial gene sequenced for a previous study (Friis et al. 2022) and available at NCBI (GenBank accession no. KX461682.1) was used for the input seed sequence.

Identification of repetitive regions

We first created a repeat library for the junco genome by modeling ab initio repeats using Repeat Modeler 1.0.11 (Smit and Hubley 2019) in scaffolds longer than 100 Kb with default options. The
under the accession GCA_009859065.2.

We first used BLASTP against the UniProt SwissProt database and conducted the functional annotation of the junco predicted proteins. Gene prediction and functional annotation

Gene prediction was conducted using BRAKER v2.1.5 (Hoff et al. 2019) and GeMoMa v1.7.1 (Keilwagen et al. 2019). We used the repeat soft-masked genome assembly and we first trained Augustus with the conserved orthologous genes from BUSCO Aves_odb10 as proteins from short evolutionary distance (Gremme et al. 2005; Stanke et al. 2006; see Fig. 3b from Hoff et al. 2019). The predicted gene models obtained with Augustus were combined with homology-based annotations using the zebra finch (GCF_008822105.2; Warren et al. 2010) (GCF_008822105.2; Warren et al. 2010) and chicken (GCF_000002315.6; International Chicken Genome Sequencing Consortium 2004) annotated genes with the GeMoMa pipeline, obtaining the final reported annotated genes with the zebra finch Taeniopygia guttata (Fig. 2). The 30 longest scaffolds obtained based on the proximity ligation mate-pair libraries, ranging from 152.01 to 2.67 Mb (median = 18.90 Mb; Supplementary Fig. 4) and accounting for 98.27% of the whole assembled genome, were identified as full or nearly full syntenic blocks with the zebra finch Taeniopygia guttata.

We sequenced and assembled a reference genome of the dark-eyed junco. Shotgun library produced 465 million read pairs (2 × 150 bp). Chicago and DoveTail Hi-C libraries produced 218 million and 121 million read pairs (2 × 151 bp), respectively. Overall, 121 Gb were generated. Genome scaffolding with HiRise yielded an assembly of 4,684 scaffolds and 1.03 Gb, with a sequence coverage of 117X; an L50/N50 equal to 5 scaffolds/71.3 Mb and an L90/N90 of 19 scaffolds/14.1 Mb; and a relatively low number of ambiguous bases (i.e. N) inserted in the genome (3.13%; Table 1).

The dark-eyed junco genome also showed high levels of synteny with the zebra finch Taeniopygia guttata (Fig. 2). The 30 longest scaffolds obtained based on the proximity ligation mate-pair libraries, ranging from 152.01 to 2.67 Mb (median = 18.90 Mb; Supplementary Fig. 4) and accounting for 98.27% of the whole assembled genome, were identified as full or nearly full syntenic blocks with the zebra finch Taeniopygia guttata.

Gene completeness assessment and genome synten

We assessed gene completeness in the genome assembly and the gene annotation using BUSCO (Benchmarking Universal Single-Copy Orthologs) v4.0.5 (–auto-lineage-euk option; Waterhouse et al. 2018). BUSCO evaluations were conducted using the 255 and 8,338 single-copy orthologous genes in Eukaryota_odb10 and Aves_odb10 datasets, respectively. In addition, we used MUMmer (Delcher et al. 2003) to explore synten with the zebra finch (Taeniopygia guttata) genome bTaeGut2.pri.v2 available at NCBI under the accession GCA_009859065.2.

Historic changes in effective population size

The pairwise sequentially Markovian coalescent (PSMC) (Li and Durbin 2011) was used to model historic changes in the effective population size of the dark-eyed junco. The shotgun data produced was mapped with bwa-mem (Li and Durbin 2009) against the fully assembled genome and a whole-genome diploid consensus sequence was generated using SAMtools v0.1.19 (Li et al. 2009) and bcftools (Danecek and McCarthy 2017). The default settings of PSMC were adopted, with a generation time of 1.5 years and a mutation rate of 4.6 × 10⁻⁸ per site per generation (Nam et al. 2010; Smeds et al. 2016). Confidence intervals were estimated based on bootstrapping (Li and Durbin 2011).
chromosomes when compared to the zebra finch (Fig. 2). Both sexual chromosomes were successfully mapped, yet the chromosome W was highly fragmented (293 fragments), possibly due to difficulties assembling highly repetitive genomic regions (Dovetail, pers. comm.). In addition, a circularized assembly of the mitochondrial genome was 16,894 bp long with a 46.6% GC content.

The integrity assessment of the *J. hyemalis* genome retrieved 92.9% and 95.4% of the tested BUSCO groups for the eukaryotes and aves databases, respectively, evidencing the completeness of our assembly (Table 1).

**Genome annotation**

We identified 19,026 protein-coding genes, from which 15,967 were associated to GO terms. The average gene length was 15.4 Kb, with a mean of 9.86 exons per gene. BUSCO integrity analysis reported 89.0% of recovered complete BUSCOs for the eukaryota database, and 96.2% in the case of Aves. Quantitatively, our genome annotation identified a considerably higher number of genes, coding sequences, exons and introns when compared with other passerines (i.e. *Fringilla coelebs*, *Melospiza melodia*, *Taeniopygia guttata*, *Ficedula albicollis*, *Manacus vitellinus*, and *Geospiza fortis*), while averaged lengths for these elements remained similar to other species annotations (Table 2). We also found that a total of 5.77% (59.5 Mb) of the *J. hyemalis* assembly consisted of repetitive elements, a value within the range expected for birds which is at 4%–10% of the genome (Zhang et al. 2014). Of the total of repetitive elements, the greatest proportions corresponded to long interspersed nuclear elements (LINEs, 52.51%) and to long terminal repeats (LTRs, 36.57%; Supplementary Table 2). In a WEGO analysis, the most represented GO terms where those associated to “cell,” “cell parts,” and “membrane part” in the cellular component category; “binding” in molecular functions; and “cellular processes” in the category of biological processes (Supplementary Fig. 4).

**Historic changes in effective population size**

A PSMC analysis for the dark-eyed junco revealed an ancestral demographic expansion that started 500K years ago approximately. The trend was sustained till circa 100K years ago, followed by a decrease in the effective population size. This decrease roughly overlaps with the onset of the last glaciation period (Richmond and Fullerton 1986), which likely caused a demographic decline of *Junco* populations in the North American continent (Fig. 3).

### Table 2. Summary statistics for the genome annotation of *Junco hyemalis* compared with other similarly sized avian species (*Fringilla coelebs*, *Melospiza melodia*, *Taeniopygia guttata*, *Ficedula albicollis*, *Manacus vitellinus*, and *Geospiza fortis*), modified from Recuerda et al. (2021).

<table>
<thead>
<tr>
<th>Genome annotation</th>
<th><em>J. hyemalis</em></th>
<th><em>F. coelebs</em></th>
<th><em>M. melodia</em></th>
<th><em>T. guttata</em></th>
<th><em>F. albicollis</em></th>
<th><em>M. vitellinus</em></th>
<th><em>G. fortis</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes</td>
<td>19,026</td>
<td>17,703</td>
<td>15,086</td>
<td>17,561</td>
<td>16,763</td>
<td>18,976</td>
<td>14,399</td>
</tr>
<tr>
<td>Average gene length (bp)</td>
<td>15,402</td>
<td>15,818</td>
<td>14,457</td>
<td>26,458</td>
<td>31,394</td>
<td>27,847</td>
<td>30,164</td>
</tr>
<tr>
<td>Number of CDS</td>
<td>23,245</td>
<td>17,703</td>
<td>15,086</td>
<td>17,561</td>
<td>16,763</td>
<td>18,976</td>
<td>14,399</td>
</tr>
<tr>
<td>Average CDS length (bp)</td>
<td>1,647</td>
<td>1,679</td>
<td>1,325</td>
<td>1,677</td>
<td>1,942</td>
<td>1,929</td>
<td>1,766</td>
</tr>
<tr>
<td>Number of exons</td>
<td>229,210</td>
<td>221,872</td>
<td>131,940</td>
<td>171,767</td>
<td>189,043</td>
<td>190,590</td>
<td>164,721</td>
</tr>
<tr>
<td>Average exon length (bp)</td>
<td>167</td>
<td>165</td>
<td>153</td>
<td>255</td>
<td>253</td>
<td>264</td>
<td>195</td>
</tr>
<tr>
<td>Average number of exons/gene</td>
<td>9.86</td>
<td>10.16</td>
<td>8.67</td>
<td>10.25</td>
<td>12.22</td>
<td>11.51</td>
<td>11.41</td>
</tr>
<tr>
<td>Number of introns</td>
<td>205,965</td>
<td>200,041</td>
<td>116,724</td>
<td>153,909</td>
<td>171,236</td>
<td>171,089</td>
<td>149,563</td>
</tr>
<tr>
<td>Average intron length (bp)</td>
<td>1,945</td>
<td>1,902</td>
<td>1,695</td>
<td>2,930</td>
<td>3,290</td>
<td>3,284</td>
<td>2,813</td>
</tr>
</tbody>
</table>

**BUSCO Eukaryota database:** C: 89.0% [S: 85.1%, D: 3.9%], F: 5.1%, M: 5.9%, N: 255
**BUSCO Aves database:** C: 96.2% [S: 83.7%, D: 12.5%], F: 1.5%, M: 2.3%, N: 8,338

BUSCO parameters are C: Complete BUSCO; S: Complete and single-copy BUSCOs; D: Complete and duplicated BUSCOs; F: Fragmented BUSCOs; M: Missing BUSCOs; and N: Total BUSCO groups searched. CDS indicates protein-coding sequences.

Fig. 3. Changes in the ancestral effective population size of the dark-eyed junco. The dark red line represents the original effective population size through time, and light red lines represent 100 bootstrap estimates. The indices g and µ denote the generation time and the mutation rate, respectively.
Conclusion

We report here a high-quality assembly for the *J. hyemalis* genome along with a comprehensive functional annotation, a valuable resource to address a range of biological questions using genomic approaches. The combination of shotgun and proximity ligation libraries yielded an assembly 1.03 Gb long, with 98.3% of the sequence located in 30 full or nearly full chromosome scaffolds. The genome is highly contiguous and complete, enabling its use as a reference for variant calling and for the identification of candidate genes potentially involved in phenotypic variation. Improved scaffolding also enables the identification of regions putatively under selection, including structural variants such as chromosome rearrangements, repetitive elements or copy number variation, all relevant for investigating questions related to evolutionary biology and molecular ecology in the dark-eyed junco system and related species.

Data availability

Genome assembly, mitochondrial genome, genome annotation and related supporting resources have been deposited at DRYAD (doi: https://doi.org/10.5061/dryad.c59zw3r87). The genome assembly, including the raw shotgun sequencing data, Chicago and Hi-C libraries have been deposited at NCBI under accession QZWM00000000.2; BioProject PRJNA493001; BioSample: SAMN10120167. 

Supplemental material is available at G3 online.

Acknowledgments

The authors thank Maria Recuerda for her assistance with syneny and PSMC analyses.

Funding

Funding was provided by grant CGL-2011-25866 from the Spanish Ministerio de Ciencia e Innovación to BM and by multiple awards from the National Science Foundation to EK.

Conflicts of interest

None declared.

Literature cited


