1 **Title**

2 *De Novo* Assembly of the Northern Cardinal (*Cardinalis cardinalis*) Genome Reveals

3 Candidate Regulatory Regions for Sexually Dichromatic Red Plumage Coloration

4

5 **Author**

6 Simon Yung Wa Sin †‡*, Lily Lu †, Scott V. Edwards †

7

8 **Author affiliation**

9 † Department of Organismic and Evolutionary Biology, Museum of Comparative Zoology,

10 Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA.

11 ‡ School of Biological Sciences, The University of Hong Kong, Pok Fu Lam Road, Hong

12 Kong.

13

14

15 **Running title**

16 Northern cardinal genome assembly

17

18 **Keywords**

19 AllPaths-LG

20 *Cis*-regulatory elements

21 CYP2J19 gene

22 Ketocarotenoid pigments

23 Transcription factors

24

25 *** Corresponding author**

26 Simon Y. W. Sin

27 Email: yungwa.sin@gmail.com

28 Mailing address: School of Biological Sciences, Kadoorie Biological Sciences Building, The

29 University of Hong Kong, Pok Fu Lam Road, Hong Kong.

30

Abstract

Northern cardinals (*Cardinalis cardinalis*) are common, mid-sized passerines widely distributed in North America. As an iconic species with strong sexual dichromatism, it has been the focus of extensive ecological and evolutionary research, yet genomic studies investigating the evolution of genotype–phenotype association of plumage coloration and dichromatism are lacking. Here we present a new, highly contiguous assembly for *C. cardinalis*. We generated a 1.1 Gb assembly comprised of 4,762 scaffolds, with a scaffold N50 of 3.6 Mb, a contig N50 of 114.4 kb and a longest scaffold of 19.7 Mb. We identified 93.5% complete and single-copy orthologs from an Aves dataset using BUSCO, demonstrating high completeness of the genome assembly. We annotated the genomic region comprising the CYP2J19 gene, which plays a pivotal role in the red coloration in birds. Comparative analyses demonstrated non-exonic regions unique to the CYP2J19 gene in passerines and a long insertion upstream of the gene in *C. cardinalis*. Transcription factor binding motifs discovered in the unique insertion region in *C. cardinalis* suggest potential androgen-regulated mechanisms underlying sexual dichromatism. Pairwise Sequential Markovian Coalescent (PSMC) analysis of the genome reveals fluctuations in historic effective population size between 100,000–250,000 in the last 2 millions years, with declines concordant with the beginning of the Pleistocene epoch and Last Glacial Period. This draft genome of *C. cardinalis* provides an important resource for future studies of ecological, evolutionary, and functional genomics in cardinals and other birds.

Introduction

The northern cardinal (*Cardinalis cardinalis*) is a mid-sized (~42-48 g) passerine broadly distributed in eastern and central North America, with a range encompassing northern Central America to southeastern Canada (Smith et al. 2011). It has high genetic and phenotypic

diversity and is currently divided into 18 subspecies (Paynter 1970; Smith et al. 2011). Cardinals have been studied extensively on research areas such as song and communication (e.g. Anderson and Conner 1985; Yamaguchi 1998; Jawor and MacDougall-Shackleton 2008), sexual selection (e.g. Jawor et al. 2003), physiology (e.g. DeVries and Jawor 2013; Wright and Fokidis 2016), phylogeography (e.g. Smith et al. 2011), and plumage coloration (e.g. Linville and Breitwisch 1997; Linville et al. 1998; McGraw et al. 2003). *C. cardinalis* is an iconic species in Cardinalidae and has strong sexual dichromatism, with adult males possessing bright red plumage and adult females tan (Fig. 1).

The red plumage coloration of many bird species plays important roles in social and sexual signaling. With the advance of genomic technologies in the last several years, the genetic basis and evolution of plumage coloration in birds is under intense investigation (Lopes et al. 2016; Mundy et al. 2016; Twyman et al. 2016; Toomey et al. 2018; Twyman et al. 2018a; Twyman et al. 2018b; Funk and Taylor 2019; Kim et al. 2019; Gazda et al. 2020). The red color of feathers is generated by the deposition of ingested carotenoids, modified endogenously. A ketolase is involved in an oxidative reaction to convert dietary yellow carotenoids into red ketocarotenoids (Friedman et al. 2014). Recently the gene encoding the carotenoid ketolase has been shown to be a cytochrome P450 enzyme, CYP2J19 (Lopes et al. 2016; Mundy et al. 2016). The identification of CYP2J19 as the gene responsible for red coloration in hybrid canary (*Serinus canaria*) plumage (Lopes et al. 2016) and zebra finch (*Taeniopygia guttata*) bill and legs (Mundy et al. 2016) suggests its general role in red pigmentation of multiple tissues across birds. *C. cardinalis* is an excellent candidate to further our understanding of the regulation and development of red plumage coloration in birds. In particular, little attention has been paid to noncoding, regulatory signatures in the region surrounding CYP2J19, and genomic resources for cardinals would greatly facilitate such work.

Despite extensive ecological and evolutionary research on *C. cardinalis*, we lack a highly contiguous genome of this species and other species in the family Cardinalidae. A genome assembly of *C. cardinalis* will facilitate studies of genotype–phenotype association of sexually dichromatic plumage color in this species, and facilitate comparative genomic analysis in birds generally. Here we generate a new genome assembly from a male collected in Texas, USA, with a voucher specimen in the ornithology collection of the Museum of Comparative Zoology. We used the AllPaths-LG (Gnerre et al. 2011) method to perform *de novo* genome assembly. Given the lack of genomic resources currently available for the Cardinalidae, this *C. cardinalis* draft genome will facilitate future studies on genomic studies of cardinals and other passerines.

Methods & Materials

Sample collection and DNA extraction

We collected a male *C. cardinalis* on the DNA Works Ranch, Afton, Dickens County, Texas, United States (33.76286°, -100.8168°) on 19 January 2016 (MCZ Ornithology cat. no.: 364215). We collected muscle, heart, and liver and snap frozen the tissues in liquid nitrogen immediately in the field. Upon returning from the field, tissues were stored in the cryopreservation facility of the Museum of Comparative Zoology, Harvard University until DNA extraction. We isolated genomic DNA using the DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol. We confirmed the sex of the individual using published PCR primers targeting CHD1 genes with different sizes of intron on the W and Z chromosomes (2550F & 2718R; Fridolfsson and Ellegren 1999) and measured DNA concentration with a Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, USA).

106 **Library preparation and sequencing**

107 We performed whole-genome library preparation and sequencing following Grayson et al.
108 (2017). In brief, a DNA fragment library of 220 bp insert size was prepared using the PrepX
109 ILM 32i DNA Library Kit (Takara), and mate-pair libraries of 3 kb insert size were prepared
110 using the Nextera Mate Pair Sample Preparation Kit (cat. No. FC-132-1001, Illumina). We
111 then assessed library quality using the HS DNA Kit (Agilent) and quantified the libraries
112 with qPCR prior to sequencing (KAPA library quantification kit). We sequenced the libraries
113 on an Illumina HiSeq instrument (High Output 250 kit, PE 125 bp reads) at the Bauer Core
114 facility at Harvard University.

115

116 **De novo genome assembly and assessment**

117 We assessed the quality of the sequencing data using FastQC and removed adapters using
118 Trimmomatic (Bolger et al. 2014) and assembled the genome using AllPaths-LG v52488
119 (Gnerre et al. 2011), which allowed us to estimate the genome size from k-mer frequencies
120 and assess the contiguity of the *de novo* genome. We estimated the completeness of the
121 assembled genome with BUSCO v2.0 (Simão et al. 2015) and used the aves_odb9 dataset to
122 search for 4915 universal single-copy orthologs in birds.

123

124 **Analysis of the CYP2J19 genomic region**

125 We annotated the genomic region that comprises the CYP2J19 gene using blastn with the
126 CYP2J19, CYP2J40, HOOK1 and NFIA genes from *S. canaria* and *T. guttata* as queries.
127 Conserved non-exonic elements (CNEEs) were obtained from a published UCSC Genome
128 Browser track hub containing a progressiveCactus alignment of 42 bird and reptile species
129 and CNEE annotations (viewed in the UCSC genome browser at
130 https://ifx.rc.fas.harvard.edu/pub/ratiteHub8/hub.txt; Sackton et al. 2019). We identified

CYP2J19 genes from 30 bird species (Twyman et al. 2018a) in the NCBI by blasting their genome assemblies. The alignment was compiled using MUSCLE (Edgar 2004) and viewed with Geneious (Kearse et al. 2012). We also aligned the genomic regions up- and downstream of CYP2J19 in 10 passerines to identify any potential conserved regulatory regions present in species with red carotenoid coloration.

We used the MEME Suite (Bailey et al. 2009; Bailey et al. 2015) to discover potential regulatory DNA motifs in the region upstream of the CYP2J19 gene that we found to be unique to *C. cardinalis* and to predict transcription factors (TFs) binding to those motifs. We used MEME (Bailey and Elkan 1994) to identify the top five most statistically significant motifs and their corresponding positions in the region. We used Tomtom (Gupta et al. 2007) to compare the identified motifs against databases (e.g. JASPAR) of known motifs and identify potential TFs specific to those matched motifs. To search for potential biological roles of these motifs, we used GOMo (Buske et al. 2010) to scan all promoters in *Gallus gallus* using the identified motifs to determine if any motifs are significantly associated with genes linked to any Genome Ontology (GO) terms.

Inference of demographic history

To investigate historical demographic changes in the northern cardinal we used the Pairwise Sequential Markovian Coalescent (PSMC) model (Li and Durbin 2011) based on the diploid whole-genome sequence to reconstruct the population history. We generated consensus sequences for all autosomes using SAMtools' (v1.5; Li et al. 2009) *mpileup* command and the *vcf2fq* command from vcfutils.pl. We applied filters for base quality and mapping quality below 30. The settings for the PSMC atomic time intervals were "4+30*2+4+6+10". 100 bootstraps were used to compute the variance in estimates of N_e. To convert inferred population sizes and times to numbers of individuals and years, respectively, we used the

156 estimate of mutation rate of 3.44e-09 per site per generation from the medium ground finch
157 (*Geospiza fortis*, Thraupidae, a closely related passerine clade) (Nadachowska-Brzyska et al.
158 2015). We estimated the generation time of the northern cardinal as age of sexual maturity
159 multiplied by a factor of two (Nadachowska-Brzyska et al. 2015), yielding a generation time
160 of 2 years.
161
162 **Data availability (Data will be available later prior to publication)**
163 Data from all sequencing runs and final genome assembly are available from NCBI
164 (BioProject number will be provided later). Short-fragment and mate-pair libraries are also
165 available from the NCBI SRA (number to be provided later).
166
167 **Results and Discussion**
168 **Genome assembly and evaluation**
169 We generated 703,754,970 total reads from two different sequencing libraries, including
170 335,713,090 reads from the fragment library and 368,041,880 reads from the mate-pair
171 library. The genome size estimated by AllPaths-LG from k-mers was 1.1 Gb (Table 1). The
172 sequence coverage estimated was ~59x. The assembly consisted of 32,783 contigs and 4,762
173 scaffolds. The largest scaffold was 19.7 Mb. The contig N50 was 114.4 kb and the scaffold
174 N50 was 3.6 Mb (Table 1). The number of contigs per Mb was 31.4 and the number of
175 scaffolds per Mb was 4.6. The average GC content of the assembly was 42.1%. BUSCO
176 scores (Simão et al. 2015) suggest high completeness of the genome, with 93.5% of single-
177 copy orthologs for birds identified (Table 2). The genome contiguity and completeness is
178 better than most genomes presented in Jarvis et al. (2014) and Grayson et al. (2017).
179
180 **Candidate non-coding regulatory regions of CYP2J19**

The CYP2J19 gene was identified on scaffold 100 of the *C. cardinalis* genome, flanked by CYP2J40 and NFIA genes (Fig. 2), an arrangement consistent with other species (e.g. Lopes et al. 2016; Mundy et al. 2016). The length of the CYP2J19 gene in *C. cardinalis* is 8925 bp, comprising 9 exons. We identified 26 CNEEs in the CYP2J19 gene, 24 CNEEs within ~6 kb upstream and 1 CNEE within ~6 kb downstream of the gene. Of those 51 CNEEs, 10 CNEEs in CYP2J19 and 7 CNEEs upstream of the gene are at least 50 bp in length.

Three intronic insertions, as well as one non-coding region upstream and one downstream of the gene were found to be present only in passerine species (Fig. 3). In addition, we identified unique insertions of large genomic regions in the two passerine species in our alignment with red carotenoid coloration, i.e. *C. cardinalis* and *T. guttata*. A unique 5920 bp insertion is present 339 bp upstream of the CYP2J19 gene in *C. cardinalis* (Fig. 4a). There is a 11322 bp insertion downstream of CYP2J19 in *T. guttata* comprising the CYP2J19B gene (Fig. 4b; Mundy et al. 2016) that is not present in *C. cardinalis*. The most similar sequence to the *C. cardinalis* upstream insertion identified by blasting (~84% identity, 25% query cover) is a sequence annotated as "RNA-directed DNA polymerase from mobile element jockey-like" in the American crow (*Corvus brachyrhynchos*) genome assembly (Zhang et al. 2014), suggesting that the insertion may contain a non-long terminal repeat (non-LTR) retrotransposon (Ivanov et al. 1991). An endogenous retroviral insertion located upstream of the aromatase gene is proposed to be the mechanism of gene activation that lead to the henny feathers phenotype in chickens (Matsumine et al. 1991). The possibility that the upstream retrovirus may promote CYP2J19 gene activation is worth more investigation.

We discovered a total of 25 motifs clustered in a ~1.7 kb region in the unique insertion sequence of *C. cardinalis* (Fig. 5a). Fourteen TFs are predicted for the 5 identified motif types (Fig. S1, Table 3). No significant GO term appears to be associated with motifs

206 1–3 and 5, whereas significant GO terms associated with motif 3 predict molecular functions
207 of GTP binding and hexose transmembrane transporter activity. Three predicted TFs (i.e.
208 Sp1, SREBF, and RREB1; Table 3) are associated with androgen regulation of gene
209 expression.

210 *Cis*-regulatory elements play a crucial role in the development of sexually dimorphic
211 traits (Williams and Carroll 2009). In birds, sexual dimorphism is strongly affected by steroid
212 sex hormones (Kimball 2006). Sex hormones such as androgen may indirectly influence the
213 expression of androgen-regulated genes, through binding to transcription factors that interact
214 with regulatory elements such as enhancer and cause sex-biased gene expression, which leads
215 to sex-specific phenotypes (Coyne et al. 2008; Mank 2009). Whereas male plumage appears
216 to be testosterone-dependent in passerines (Kimball 2006), the molecular mechanisms
217 controlling sex-biased gene expression and development of sexually dimorphic traits in birds
218 is largely unknown (Kraaijeveld 2019; Gazda et al. in press). In this study, we identified
219 potential TFs that may regulate the expression of the red plumage color gene CYP2J19.
220 Many CYP450s are differentially expressed between the sexes (Rinn and Snyder 2005).
221 Some of our TFs predicted to interact with the insertion unique to *C. cardinalis* are androgen-
222 regulated, including Sp1, which is involved in the androgen activation of the vas deferens
223 protein promoter in mice (Darne et al. 1997) and sterol regulatory element-binding factor
224 (SREBF, aka SREBP), which is involved in androgen-regulated activation mechanisms of
225 target genes (Heemers et al. 2006). Another predicted TF, the zinc finger protein Ras-
226 responsive element binding protein (RREB1), is a co-regulator of androgen receptor (AR)
227 and plays a role in androgenic signaling by affecting AR-dependent transcription
228 (Mukhopadhyay et al. 2007). The interaction between androgens, the implicated TFs and
229 their corresponding binding motifs may underlie the development of sexually dimorphic red
230 plumage color in *C. cardinalis*. In addition, the TF nuclear factor erythroid-derived 2-like 2

(Nfe2l2, also known as nuclear factor erythroid 2-related factor 2, Nrf2) is predicted to bind to all but motif 2 (Fig. 5). The activity of Nrf2 is influenced by changes in oxidative stress, and it regulates the expression of numerous antioxidant proteins (Schultz et al. 2014). Because the production of red ketacarotenoids via CYP2J19 is sensitive to oxidative state (Lopes et al. 2016), this oxidation-dependent TF suggests a possible link between production of red pigments and the oxidative stress faced by the organism, which may reflect an association between red carotenoid coloration and individual quality (Hill and Johnson 2012).

The above non-exonic regions in and outside of the CYP2J19 gene are candidates of regulatory elements responsible for modulating red carotenoid-based coloration in passerines and *C. cardinalis*. Noncoding regulatory regions may be subject to less pleiotropic constraint than protein-coding genes (Carroll et al. 2008), and many CNEEs act as enhancers with regulatory functions (Seki et al. 2017; Sackton et al. 2019). In particular, the large non-exonic region rich with TF-binding motifs upstream of the *C. cardinalis* CYP2J19 gene may play a role in the strong sexual dichromatism and striking bright red male color in this species. The mechanism of plumage color development and sexual dichromatism, and the regulatory role of those genomic regions identified in this study are fruitful areas for future research.

Demographic historic of *C. cardinalis*

The PSMC analysis (Fig. 6) suggested a demographic history of *C. cardinalis* characterized by a fluctuation in the historic effective population size between 100,000–250,000 around ~2,000,000 years ago (ya). The population was at ~160,000 at the beginning of Pleistocene epoch at 2,580,000 ya, then decreased to 110,000 in ~800,000 ya, and increased to ~230,000 at the beginning of the Last Glacial Period around 115,000 ya. The population started to declined again after the beginning of the Last Glacial Period (LGP). The decrease in effective population size observed at the beginning of Pleistocene might be due to the divergence of *C.*

cardinalis into different subspecies (Provost et al. 2018). The population decline after the start of the LGP is consistent with the ecological niche model that indicates a dramatic range reduction for *C. cardinalis* during the LGP (Smith et al. 2011).

Conclusion

We used small-fragment and mate-pair libraries Illumina sequencing data to generate a draft genome assembly of the northern cardinal, *C. cardinalis*. Comparative analyses revealed conserved non-exonic regions unique to the CYP2J19 gene in passerines and in *C. cardinalis*, which may play a role in the regulation of red carotenoid-based plumage coloration. The motifs discovered in the lineage-specific upstream region of CYP2J19 in *C. cardinalis* suggest potential *cis*-regulatory mechanisms underlying sexual dichromatism. The assembled *C. cardinalis* genome is therefore useful for studying genotype–phenotype associations and sexual dichromatism in this species. As one of the first genome sequenced in Cardinalidae, and the highest in quality, it will also be an important resource for the comparative study of plumage color evolution in passerines and birds in general.

Acknowledgments

We would like to thank Jonathan Schmitt and Katherine Eldridge for collecting and preparing the specimen and tissues. We thank Clarence Stewart for sharing the northern cardinal photo. This research was supported by research funds through Harvard University. The computing resources for this study were provided by the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Literature Cited

280 Anderson, M. and R. Conner. 1985. Northern cardinal song in three forest habitats in eastern
281 Texas. The Wilson Bulletin:436–449.
282 Bailey, T., M. Boden, F. Buske, M. Frith, C. Grant, L. Clementi, J. Ren, W. Li, and W.
283 Noble. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids
284 Res. 37:W202–W208.
285 Bailey, T. and C. Elkan. 1994. Fitting a mixture model by expectation maximization to
286 discover motifs in biopolymers. Proceedings of the Second International Conference
287 on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park,
288 California:28–36.
289 Bailey, T., J. Johnson, C. Grant, and W. Noble. 2015. The MEME suite. Nucleic Acids Res.
290 43:W39–W49.
291 Bolger, A., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina
292 sequence data. Bioinformatics 30:2114–2120.
293 Buske, F., M. Bodén, D. Bauer, and T. Bailey. 2010. Assigning roles to DNA regulatory
294 motifs using comparative genomics. Bioinformatics 26:860–866.
295 Carroll, S., B. Prud'homme, and N. Gompel. 2008. Regulating evolution. Sci. Am. 298:60–
296 67.
297 Coyne, J., E. Kay, and S. Pruett-Jones. 2008. The genetic basis of sexual dimorphism in
298 birds. Evolution: International Journal of Organic Evolution 62:214–219.
299 Darne, C., L. Morel, F. Claessens, M. Manin, S. Fabre, G. Veyssiere, W. Rombauts, and C.
300 Jean. 1997. Ubiquitous transcription factors NF1 and Sp1 are involved in the
301 androgen activation of the mouse vas deferens protein promoter. Mol. Cell.
302 Endocrinol. 132:13–23.

303 DeVries, M. and J. Jawor. 2013. Natural variation in circulating testosterone does not predict
304 nestling provisioning rates in the northern cardinal, *Cardinalis cardinalis*. Anim.
305 Behav. 85:957–965.
306 Edgar, R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
307 throughput. Nucleic Acids Res. 32:1792–1797.
308 Fridolfsson, A. K. and H. Ellegren. 1999. A simple and universal method for molecular
309 sexing of non-ratite birds. J. Avian Biol. 30:116–121.
310 Friedman, N., K. McGraw, and K. Omland. 2014. Evolution of carotenoid pigmentation in
311 caciques and meadowlarks (Icteridae): repeated gains of red plumage coloration by
312 carotenoid C4-oxygenation. Evolution 68:791–801.
313 Funk, E. and S. Taylor. 2019. High-throughput sequencing is revealing genetic associations
314 with avian plumage color. The Auk 136:ukz048.
315 Gazda, M. A., P. M. Araújo, R. J. Lopes, M. B. Toomey, P. Andrade, S. Afonso, C. Marques,
316 L. Nunes, P. Pereira, S. Trigo, G. E. Hill, J. C. Corbo, and M. Carneiro. in press. A
317 genetic mechanism for sexual dichromatism in birds. Science.
318 Gazda, M. A., M. B. Toomey, P. M. Araújo, R. J. Lopes, S. Afonso, C. A. Myers, K. Serres,
319 P. D. Kiser, G. E. Hill, J. C. Corbo, and M. Carneiro. 2020. Genetic basis of de novo
320 appearance of carotenoid ornamentation in bare parts of canaries. Mol. Biol. Evol.
321 37:1317–1328.
322 Gnerre, S., I. MacCallum, D. Przybylski, F. Ribeiro, J. Burton, B. Walker, T. Sharpe, G. Hall,
323 T. Shea, S. Sykes, and A. Berlin. 2011. High-quality draft assemblies of mammalian
324 genomes from massively parallel sequence data. Proceedings of the National
325 Academy of Sciences 108:1513–1518.
326 Grayson, P., S. Y. W. Sin, T. B. Sackton, and S. V. Edwards. 2017. Comparative genomics as
327 a foundation for evo-devo studies in birds. Humana Press, New York, NY.

328 Gupta, S., J. Stamatoyannopoulos, T. Bailey, and W. Noble. 2007. Quantifying similarity
329 between motifs. Genome Biology 8:R24.
330 Heemers, H., G. Verhoeven, and J. Swinnen. 2006. Androgen activation of the sterol
331 regulatory element-binding protein pathway: Current insights. Mol. Endocrinol.
332 20:2265–2277.
333 Hill, G. and J. Johnson. 2012. The vitamin A–redox hypothesis: a biochemical basis for
334 honest signaling via carotenoid pigmentation. The American Naturalist 180:E127–
335 E150.
336 Ivanov, V., A. Melnikov, A. Siunov, I. Fodor, and Y. Ilyin. 1991. Authentic reverse
337 transcriptase is coded by jockey, a mobile Drosophila element related to mammalian
338 LINEs. The EMBO journal 10:2489–2495.
339 Jarvis, E., S. Mirarab, A. Aberer, B. Li, P. Houde, C. Li, S. Ho, B. Faircloth, B. Nabholz, J.
340 Howard, and A. Suh. 2014. Whole-genome analyses resolve early branches in the tree
341 of life of modern birds. Science 346:1320–1331.
342 Jawor, J., S. Linville, S. Beall, and R. Breitwisch. 2003. Assortative mating by multiple
343 ornaments in northern cardinals (*Cardinalis cardinalis*). Behav. Ecol. 14:515–520.
344 Jawor, J. and S. MacDougall-Shackleton. 2008. Seasonal and sex-related variation in song
345 control nuclei in a species with near-monomorphic song, the northern cardinal.
346 Neurosci. Lett. 443:169–173.
347 Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, S. Buxton, A.
348 Cooper, S. Markowitz, C. Duran, and T. Thierer. 2012. Geneious Basic: an integrated
349 and extendable desktop software platform for the organization and analysis of
350 sequence data. Bioinformatics 28:1647–1649.
351 Kim, K., B. Jackson, H. Zhang, D. Toews, S. Taylor, E. Greig, I. Lovette, M. Liu, A.
352 Davison, S. Griffith, and K. Zeng. 2019. Genetics and evidence for balancing

353 selection of a sex-linked colour polymorphism in a songbird. Nature Communications
354 10:1–11.
355 Kimball, R. 2006. Hormonal control of coloration. Pp. 431–468 *in* G. Hill, and K. McGraw,
356 eds. Bird coloration. I. Mechanisms and measurements. Harvard Univerasity Press,
357 Cambridge, MA.
358 Kraaijeveld, K. 2019. Genetic architecture of novel ornamental traits and the establishment of
359 sexual dimorphism: insights from domestic birds. Journal of Ornithology 160:861–
360 868.
361 Li, H. and R. Durbin. 2011. Inference of human population history from individual whole-
362 genome sequences. Nature 475:493–496.
363 Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis,
364 and R. Durbin. 2009. The sequence alignment/map format and SAMtools.
365 Bioinformatics 25:2078–2079.
366 Linville, S. and R. Breitwisch. 1997. Carotenoid availability and plumage coloration in a wild
367 population of northern cardinals. The Auk 114:796–800.
368 Linville, S., R. Breitwisch, and A. Schilling. 1998. Plumage brightness as an indicator of
369 parental care in northern cardinals. Anim. Behav. 55:119–127.
370 Lopes, R. J., J. D. Johnson, M. B. Toomey, M. S. Ferreira, P. M. Araujo, J. Melo-Ferreira, L.
371 Andersson, G. E. Hill, J. C. Corbo, and M. Carneiro. 2016. Genetic basis for red
372 coloration in birds. Curr. Biol. 26:1427–1434.
373 Mank, J. 2009. Sex chromosomes and the evolution of sexual dimorphism: lessons from the
374 genome. The American Naturalist 173:141–150.
375 Matsumine, H., M. A. Herbst, S. H. Ou, J. D. Wilson, and M. J. McPhaul. 1991. Aromatase
376 mRNA in the extragonadal tissues of chickens with the henny-feathering trait is
377 derived from a distinctive promoter structure that contains a segment of a retroviral

long terminal repeat. Functional organization of the Sebright, Leghorn, and Campine aromatase genes. J. Biol. Chem. 266:19900–19907.

McGraw, K., G. Hill, and R. Parker. 2003. Carotenoid pigments in a mutant cardinal: implications for the genetic and enzymatic control mechanisms of carotenoid metabolism in birds. The Condor 105:587–592.

Mukhopadhyay, N., B. Cinar, L. Mukhopadhyay, M. Lutchman, A. Ferdinand, J. Kim, L. Chung, R. Adam, S. Ray, A. Leiter, and J. Richie. 2007. The zinc finger protein ras-responsive element binding protein-1 is a coregulator of the androgen receptor: implications for the role of the Ras pathway in enhancing androgenic signaling in prostate cancer. Mol. Endocrinol. 21:2056–2070.

Mundy, N. I., J. Stapley, C. Bennison, R. Tucker, H. Twyman, K. W. Kim, T. Burke, T. R. Birkhead, S. Andersson, and J. Slate. 2016. Red carotenoid coloration in the zebra finch is controlled by a cytochrome P450 gene cluster. Curr. Biol. 26:1435–1440.

Nadachowska-Brzyska, K., C. Li, L. Smeds, G. Zhang, and H. Ellegren. 2015. Temporal dynamics of avian populations during Pleistocene revealed by whole-genome sequences. Curr. Biol. 25:1375–1380.

Paynter, R. J. 1970. Subfamily Emberizinae. Museum of Comparitive Zoology, Cambridge, MA, USA.

Provost, K., W. M. III, and B. Smith. 2018. Genomic divergence in allopatric Northern Cardinals of the North American warm deserts is linked to behavioral differentiation. Ecology and Evolution 8:12456–12478.

Rinn, J. and M. Snyder. 2005. Sexual dimorphism in mammalian gene expression. Trends Genet. 21:298–305.

401 Sackton, T., P. Grayson, A. Cloutier, Z. Hu, J. Liu, N. Wheeler, P. Gardner, J. Clarke, A.
402 Baker, M. Clamp, and S. Edwards. 2019. Convergent regulatory evolution and loss of
403 flight in paleognathous birds. Science 364:74–78.
404 Schultz, M., S. Hagan, A. Datta, Y. Zhang, M. Freeman, S. Sikka, A. Abdel-Mageed, and D.
405 Mondal. 2014. Nrf1 and Nrf2 transcription factors regulate androgen receptor
406 transactivation in prostate cancer cells. PloS One 9.
407 Seki, R., C. Li, Q. Fang, S. Hayashi, S. Egawa, J. Hu, L. Xu, H. Pan, M. Kondo, T. Sato, and
408 H. Matsubara. 2017. Functional roles of Aves class-specific cis-regulatory elements
409 on macroevolution of bird-specific features. Nature Communications 8:1–14.
410 Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov. 2015.
411 BUSCO: assessing genome assembly and annotation completeness with single-copy
412 orthologs. Bioinformatics 31:3210–3212.
413 Smith, B., P. Escalante, B. Baños, A. Navarro-Sigüenza, S. Rohwer, and J. Klicka. 2011. The
414 role of historical and contemporary processes on phylogeographic structure and
415 genetic diversity in the Northern Cardinal, *Cardinalis cardinalis*. BMC Evol. Biol.
416 11:136.
417 Toomey, M., C. Marques, P. Andrade, P. Araújo, S. Sabatino, M. Gazda, S. Afonso, R.
418 Lopes, J. Corbo, and M. Carneiro. 2018. A non-coding region near Follistatin controls
419 head colour polymorphism in the Gouldian finch. Proceedings of the Royal Society B
420 285:20181788.
421 Twyman, H., S. Andersson, and N. Mundy. 2018a. Evolution of CYP2J19, a gene involved in
422 colour vision and red coloration in birds: positive selection in the face of conservation
423 and pleiotropy. BMC Evol. Biol. 18:22.

424 Twyman, H., M. Prager, N. I. Mundy, and S. Andersson. 2018b. Expression of a carotenoid-
425 modifying gene and evolution of red coloration in weaverbirds (Ploceidae). Mol.
426 Ecol. 27:449–458.
427 Twyman, H., N. Valenzuela, R. Literman, S. Andersson, and N. I. Mundy. 2016. Seeing red
428 to being red: conserved genetic mechanism for red cone oil droplets and co-option for
429 red coloration in birds and turtles. Proc. R. Soc. Lond. B Biol. Sci. 283:20161208.
430 Williams, T. and S. Carroll. 2009. Genetic and molecular insights into the development and
431 evolution of sexual dimorphism. Nat. Rev. Genet. 10:797–804.
432 Wright, S. and H. Fokidis. 2016. Sources of variation in plasma corticosterone and
433 dehydroepiandrosterone in the male northern cardinal (*Cardinalis cardinalis*): II.
434 Effects of urbanization, food supplementation and social stress. Gen. Comp.
435 Endocrinol. 235:201–209.
436 Yamaguchi, A. 1998. A sexually dimorphic learned birdsong in the Northern Cardinal. The
437 Condor 100:504–511.
438 Zhang, G., C. Li, Q. Li, B. Li, D. Larkin, C. Lee, J. Storz, A. Antunes, M. Greenwold, R.
439 Meredith, and A. Ödeen. 2014. Comparative genomics reveals insights into avian
440 genome evolution and adaptation. Science 346:1311–1320.
441

Tables

Table 1 *De novo* assembly metrics for northern cardinal genome

Metric	Value
Estimated genome size	1.10 Gb
%GC content	42.1
Total depth of coverage	59x
Total contig length (bp)	1,019,501,986
Total scaffold length (bp, gapped)	1,044,184,327
Number of contigs	32,783
Contig N50	114.4 kb
Number of scaffolds	4,762
Scaffold N50 (with gaps)	3.6 Mb
Largest scaffold	19.7 Mb

Table 2 Output from BUSCO analyses to assess genome completeness by searching for single-copy orthologs from aves dataset

	Aves	%
Complete BUSCOs	4642	94.4
Complete and single-copy BUSCOs	4596	93.5
Complete and duplicated BUSCOs	46	0.9
Fragmented BUSCOs	167	3.4
Missing BUSCOs	106	2.2
Total BUSCO groups searched	4915	

Table 3 Transcription factors (TFs) predicted for the 5 motifs.

Motif no.	TF name [a]	*p*-value
1	KLF13	6.32×10^{-04}
	Nfe2l2	2.15×10^{-03}
	SP1	3.62×10^{-03}
	SREBF1	4.81×10^{-03}
	Ascl2 secondary	5.18×10^{-03}
2	SOX8 DBD5	3.99×10^{-03}
3	**Nfe2l2**	3.72×10^{-04}
	MAF NFE2	2.78×10^{-03}
	Bach1 Mafk	3.14×10^{-03}
	Zfp128 primary	3.40×10^{-03}
	E2F7	5.52×10^{-03}
4	**Nfe2l2**	1.05×10^{-03}
	RREB1	2.86×10^{-03}
	SP1	4.73×10^{-03}
	Bach1 Mafk	4.91×10^{-03}
5	**SP1**	2.04×10^{-03}
	Ascl2 secondary	2.06×10^{-03}
	SREBF1	2.18×10^{-03}
	EBF1	3.42×10^{-03}
	Nfe2l2	3.95×10^{-03}
	KLF16	4.83×10^{-03}
	SP3	4.83×10^{-03}
	RREB1	4.96×10^{-03}

[a] Bold names indicates TFs predicted for more than one motif.

Figures

Figure 1 A pair of northern cardinals, a common, sexually dichromatic passerine bird. The adult male (left) has bright red plumage whereas the adult female (right) is primarily tan in color. Photo © Clarence Stewart.

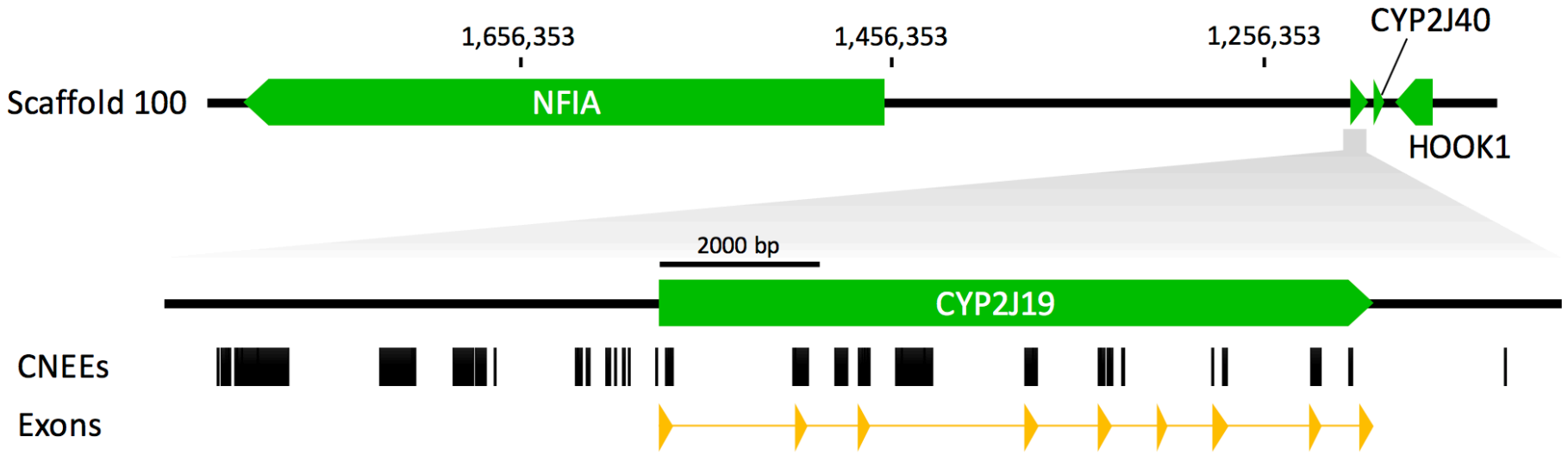

Figure 2 Annotation of the CYP2J19 region in scaffold 100 of the northern cardinal genome. The CYP2J19 gene is flanked by the NFIA and CYP2J40 genes as in other passerines. Conserved non-exonic elements (CNEEs) are shown in black boxes. Exons of the CYP2J19 gene are in yellow triangles.

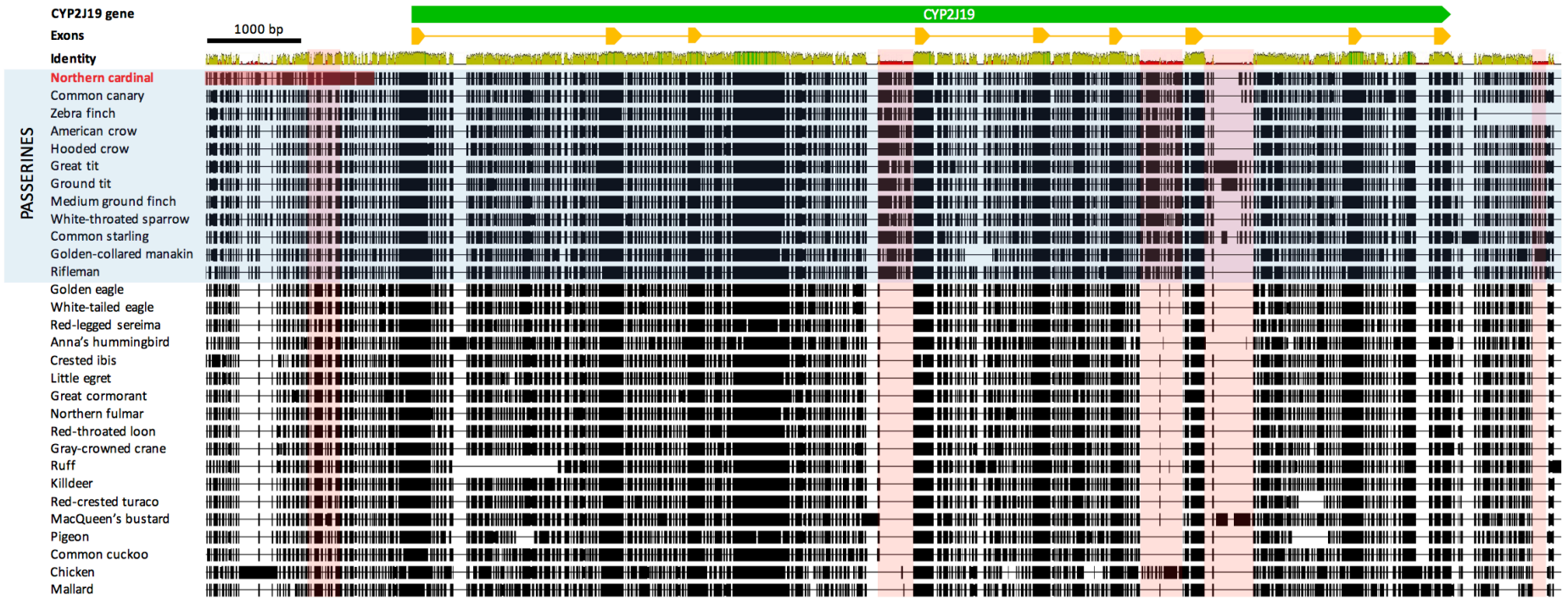

Figure 3 Multiple sequence alignment of the CYP2J19 gene in birds, with red highlighted areas showing regions present in passerines but not other birds. Exons of the CYP2J19 gene are depicted as yellow triangles. The dark red highlighted region upstream of the CYP2J19 gene in the northern cardinal is different from other passerines (see figure 5).

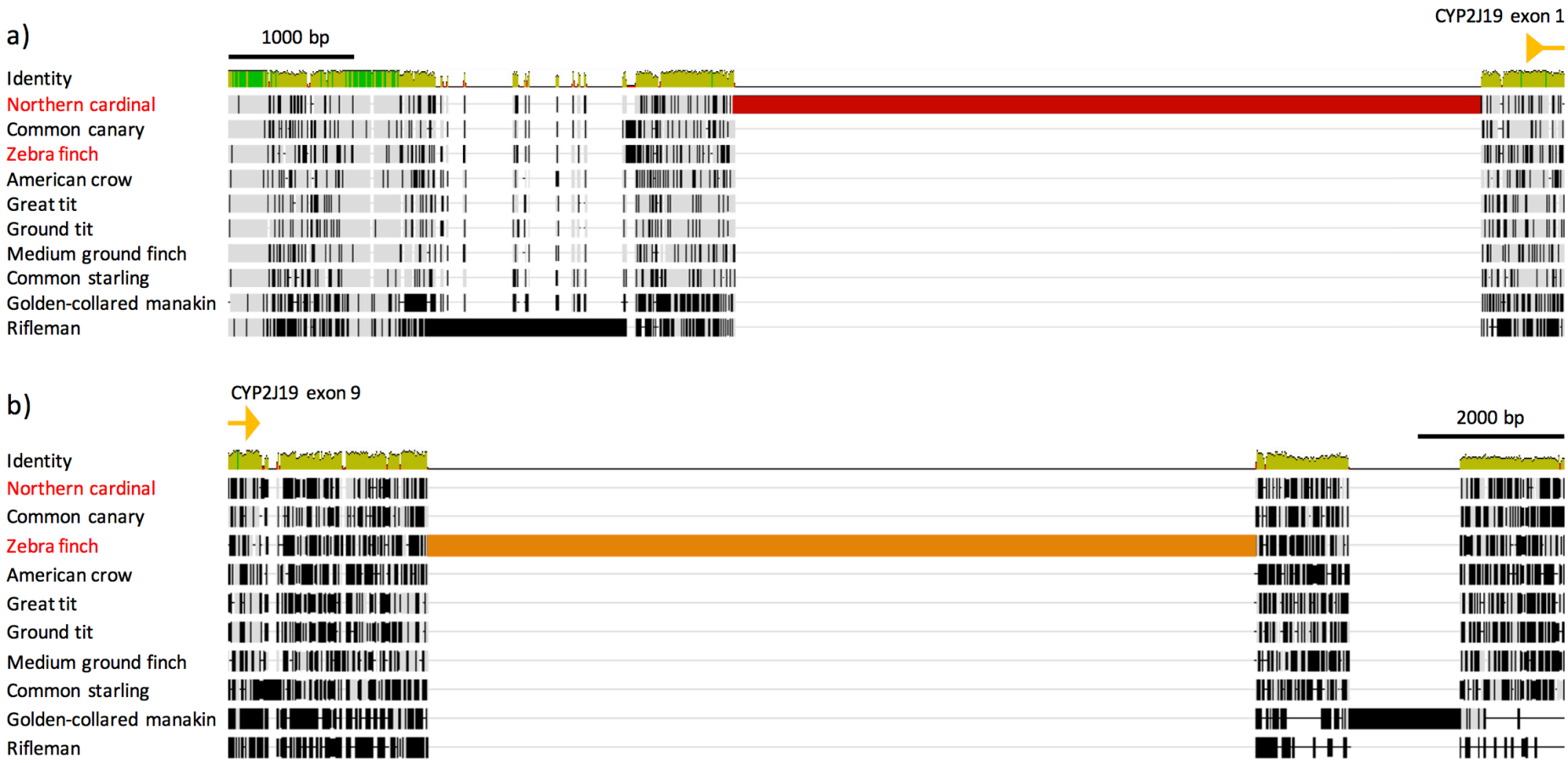

Figure 4 Multiple sequence alignment of (a) upstream and (b) downstream regions of the CYP2J19 gene in passerines. A large insertion upstream of CYP2J19 unique to the northern cardinal is highlighted in red. The insertion of a large region downstream of CYP2J19 highlighted in orange in the zebra finch indicates the CYP2J19B gene in the CYP2J2-like cluster in this species. The names of species possessing red carotenoid coloration (i.e. northern cardinal and zebra finch) are in red. Exons of the CYP2J19 gene are depicted as yellow triangles.

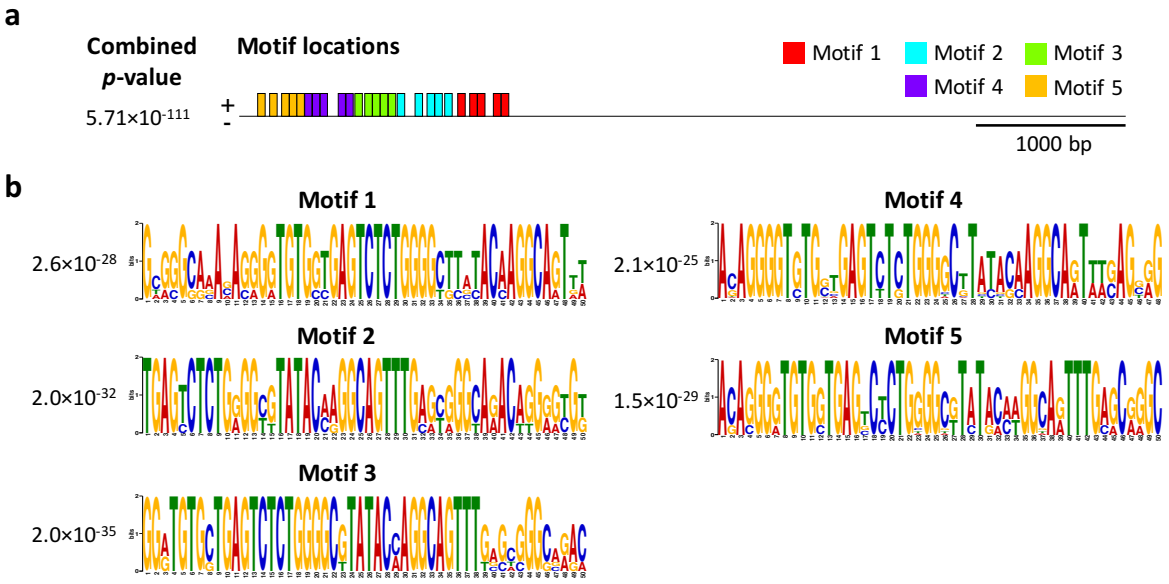

Figure 5 Identification of potential DNA motifs in the unique 5920 bp insertion upstream of the CYP2J19 gene in *C. cardinalis*. (a) Locations of 25 motifs identified and their distribution in the insertion sequence. Sites on the positive (+) strand are shown above the line. Scale is shown below the sequence. (b) Logos of five motifs with statistically significant e-values provided next to the logos.

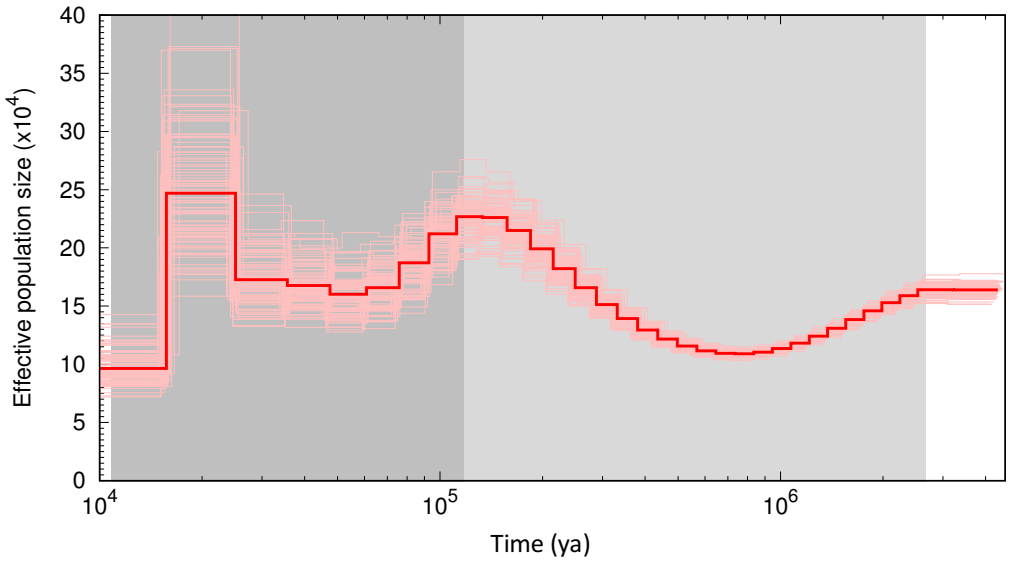

Figure 6 Demographic history of the northern cardinal inferred using PSMC. Bold red line is the median effective population size estimate, whereas thin lines are 100 individual bootstrap replicates. The light and dark grey areas indicate the Pleistocene epoch and Last Glacial Period, respectively.

Supplementary Information

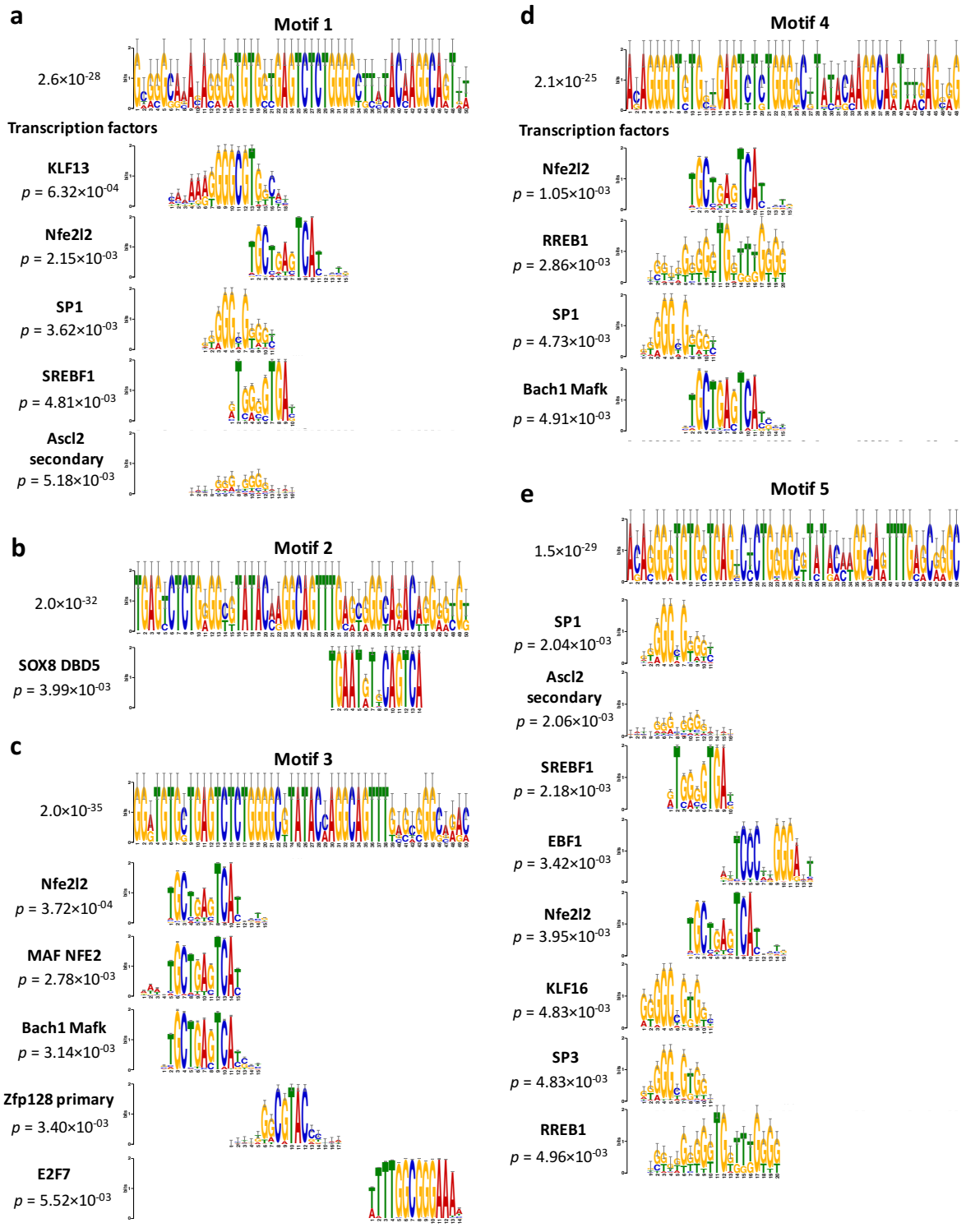

Figure S1 Identification of potential binding transcription factors in the unique 5920 bp insertion upstream of the CYP2J19 gene in *C. cardinalis*. (a–e) Logos of five motifs with statistically significant e-values provided next to the logos. Transcription factors predicted for the 5 motifs are shown below the query motifs.