Genomic architecture of migration timing in a long-distance migratory songbird

Evelien de Greef[1*], Alexander Suh[2,3], Matt J. Thorstensen[1], Kira E. Delmore,[4†] Kevin C. Fraser[1†]

[1]Department of Biological Sciences, University of Manitoba, Winnipeg, Canada R3T 2N2.

[2]Department of Organismal Biology, Uppsala University, Uppsala, Sweden SE-752 36.

[3]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom NR4 7TU.

[4]Department of Biology, Texas A&M University, College Station, Texas 77843.

*corresponding author: Evelien de Greef

†contributed equally

Email: eveliendegreef@gmail.com

Abstract

The impact of climate change on spring phenology poses risks to migratory birds, as migration timing is controlled predominantly by endogenous mechanisms. Despite numerous studies on internal cues controlling migration, the underlying genetic basis of migration timing remains largely unknown. We investigated the genetic architecture of migration timing in a long-distance migratory songbird (purple martin, *Progne subis subis*) by integrating genomic data with an extensive dataset of direct migratory tracks. Our findings show migration has a predictable genetic basis in martins and maps to a region on chromosome 1. This region contains genes that could facilitate nocturnal flights and act as epigenetic modifiers. Additionally, we found that genomic variance explained a higher proportion of historic than recent environmental spring phenology data, which may suggest a reduction in the adaptive potential of migratory behavior in contemporary populations. Overall, these results advance our understanding of the genomic underpinnings of migration timing and could provide context for conservation action.

Introduction

Climate change affects spring phenology in temperate zones and could have significant, negative impacts on migratory animals [1]. For example, migrants must synchronize arrival at the breeding grounds to coincide with seasonal resources [2]. These resources are becoming available earlier and it is unclear if migrants will be able to match these advances, potentially leading to substantial population declines [3]. Migratory timing is largely endogenously controlled [4], and thus knowledge of its genetic architecture (e.g., the identity, number, and location of genetic loci involved) is essential for predicting if and how migrants will respond to phenological changes that accompany climate change.

Previous genetic studies provide important insights regarding migration timing, such as in genes associated with circadian and circannual rhythms [5,6]; however, results vary across species [7] and are limited to small portions of the genome. Another limitation associated with earlier studies was an inability to quantify migratory behavior in the wild—prior to 2007, it was not possible to track animals <100 g on migration [8] and, for example, most migratory avian species fall into this size class. We overcame these limitations here, combining high resolution genomic data with an extensive migration tracking dataset for purple martins (*P.s. subis*). The purple martin is a Nearctic-neotropical migrant that travels over 7,000 km between North America and South America [9] and exhibits extensive latitudinal variation in migration timing. It is thus a powerful system to study migration genomics. For example, individuals breeding in the southern edge of the range in Florida may arrive as early as mid-January, while their northern counterparts in Alberta may arrive as late as June [10].

Our first objective was to examine the genomic architecture of migration timing by assembling a reference genome for the purple martin and integrating sequencing data with light-level geolocator tracks. We examined results from genome-wide association studies (GWAS), polygenic scores (PGS), and genomic differentiation analyses. Our second objective was to examine the adaptive potential of migration timing by comparing genomic variation associated with historic and contemporary spring phenology. We reran our GWAS using environmental proxies for spring timing and tested if the proportion of phenotypic variation explained by our genomic data (hereafter "PVE") was lower in contemporary datasets, which could suggest a change in adaptive potential. This study expands our understanding of the whole-genome contribution to migration and yields insight into the adaptability of migration timing behavior.

Results

Reference and resequencing data

The final *P. subis* reference genome assembly based on long reads and linked reads was 1.17 Gb in length, consisted of 2,896 scaffolds, had an N50 scaffold length of 6.13 Mb and an N50 contig length of 3.08 Mb. The annotation included 12,686 genes (SI Appendix, Table S1). The assembly length was similar to other avian genomes, which are typically between 1.0–1.2 Gb [11]. BUSCO analysis revealed that the *P. subis* genome was relatively complete with 91% of avian orthologs detected as complete sequences (89.1% being single-copy and 1.9% being duplicated), which was in range of other non-model avian genomes [12]. We aligned resequencing data for 87 individuals to this reference resulting in 4.6 million SNPs after filtering. All these individuals were tracked on migration with light-level geolocators yielding precise estimates for migratory timing.

Genomic architecture of migration timing

Birds in this study exhibited considerable latitudinal variation in migratory timing (sampling locations in Table S2), ranging over 120 to 131 days for spring departure and arrival dates (Figure S1). Spring departure and arrival locations are displayed in Figure 1, showing weak migratory connectivity between breeding and wintering sites (i.e. mixing of breeding populations at shared wintering areas) such as observed in Fraser et al. 2012 and Fraser et al. 2017 [9,13]. Estimates of PVE from Bayesian sparse linear mixed models (BSLMMs) [14] were high, with a median value of 0.70 (89% ETI: 0.09–1.00). PGS estimated from linear mixed models [15] were strongly correlated with spring timing (R^2 of 0.25, $p = 1.4e^{-148}$). Through jackknife cross-validation partitions, we assessed predictive power of the PGS model and found that birds with

lower PGS deciles exhibited earlier migration timing compared with individuals in higher PGS deciles with later migration timing (Figure 2a). BSLMMs did not identify any specific genomic regions linked to migratory timing, except for one SNP with PIP = 0.18 (Figure S2), 79 kb upstream of gene *tsc-22.* However, a survey of net genomic differentiation (ΔF_{ST}) between the earliest and latest spring migrants in our dataset did reveal a region of elevated differentiation on chromosome 1 (Figure 3). ΔF_{ST} controls for processes unrelated to migration that could elevate F_{ST} (including population structure, see Methods). This elevation was additionally present in comparisons of early and late migrants within populations in Florida (southernmost colony) and Alberta (northernmost colony) (Figure 3b) suggesting population structure did not generate this pattern. Reductions in nucleotide diversity and Tajima's D indicative of a selective sweep are also present in this region, which covers 2 Mb region and consist of 13 genes (Table S3) including *ppfia2* and *nts*, which may be related to sleep [16,17], and *mettl25* and *acss3* that may serve as important epigenetic modifiers [18,19].

Genomic association with ecological spring indices

We extracted green-up [20] and first bloom [21] data for sites where purple martins were tracked. Green-up data were available for all sites (US & Canada; 2001–2015); first bloom data were available for US sites only but spanned a longer period (1981–2015). There was a strong correlation between these environmental variables and spring migration timing (R = 0.76–0.90), thus we used these variables as proxies for current and historic migration timing in our analyses. Estimates of PVE were consistently higher in the historic datasets (green-up PVE = 0.98, SD = 0.06, 89% ETI = 0.91–1.00; first-bloom PVE = 0.99, SD = 0.07, 89% ETI = 0.91–1.00),

105 compared to PVE in recent years (green-up PVE = 0.79, SD = 0.3, 89% ETI = 0.07–1.00; first
106 bloom PVE = 0.89, SD = 0.16, 89% ETI = 0.52–1.00) (Figure S3).
107
108 ## Discussion
109 We used one of the largest tracking datasets available for a long-distance migratory
110 songbird and a genome-wide SNP dataset to reveal the genetic underpinnings of migration
111 timing. Our results demonstrate a strong genetic basis to migration timing in the purple martin.
112 We discovered a previously unidentified 2 Mb genetically differentiated region on chromosome
113 1, illuminating components underlying migration timing. Additionally, lower PVE with recent
114 environmental data suggest a reduced adaptive potential with advancing spring phenology.
115 Phenotypes that vary across individuals are a result of both environmental and genetic
116 factors, and PVE represents the proportion of variance attributed to genetic factors. The large
117 PVE estimate (0.7) is evidence that variation in migration timing is largely determined by
118 genetics. The predictive utility of the polygenic model across multiple deciles showed it is
119 possible to predict early and late migrants using genetic variants, which could potentially aid in
120 estimating a birds' phenotype in the wild. With overlapping wintering grounds among purple
121 martin colonies [9], the model may also help predict an individuals' timing tendency in addition
122 to breeding region when captured during the winter. While the high genomic variation explaining
123 migratory timing does not preclude phenotypic plasticity, it suggests that changes in timing may
124 occur through microevolutionary processes. It is important to understand the source of
125 considerable genetic variation in migratory traits [24], and future work will inform how
126 influences such as standing genetic variation (presence of more than one allele at a locus in a
127 population) may play a role in facilitating rapid microevolutionary changes [25]. While our

sample size limits our estimates of PVE (the ETIs were quite wide), our polygenic scores (PGS) supported prediction of individual migration timing. Given the present sample size, it is remarkable that a substantial proportion of variation was explained through genomics. Whether this proportion could be even greater with larger sample sizes [22,23] could be determined in future studies. However, collecting enough samples to capture strong power in tests of genomic associations with phenotypes in field-based wildlife research will be challenging.

The small number of genomic loci in the GWAS significantly associated with spring migration suggest this trait is controlled by many alleles of small effect, which may have been undetected, or could have been located in assembly gaps [11]. However, elevated genomic differentiation on chromosome 1 illuminates a potential connection between migration timing in purple martins with some genes related to rest and epigenetic modifiers. *Ppfia2* and *nts* are 2 of the 13 genes located in this region. *Ppfia2* has been linked to sleep and wakefulness in white-crowned sparrows [16], and *nts* has been linked to sleep regulation in European mice [17]. While purple martins are primarily diurnal migrants, they can incorporate both day and night flights on spring migration [26], and the former genes could play a role in these nocturnal flights. *Mettl25* and *acss3* could mediate epigenetic changes in response to environmental cues important for migratory timing; *mettl25* encodes a methyltransferase that represses gene expression [18,19] and *acss3* produces acetyl-CoA which promotes gene expression by acetylating histones. Acetyl-CoA is also important for generating, using, and storing energy [27]. While this study suggests associations with these genes, further work could elucidate these mechanisms and their roles in migration timing. While estimates of F_{ST} are often considered bottom-up comparisons, we compared extreme phenotypes while controlling for population differentiation to identify genomic regions associated with migration timing. Results from this approach were further

supported by a comparison of early and late migrants within populations that recovered a similar elevated pattern of genomic differentiation.

Our comparison of contemporary and historical data on spring phenology suggests that the adaptive potential of migration timing in purple martins may have declined in recent years, with lower values of PVE in contemporary datasets. These reductions could derive from selection for earlier arrival on the breeding grounds and will ultimately affect the amount of genetic variation available for future change. Interpretation of these results requires caution (e.g., our analyses assume phenotypic plasticity has not changed, we are using environmental variables as a proxy for migration timing and are using data from a small subset of populations). In addition, environmental and genetic differences between the present day and when historical spring phenology data were recorded (1981–1984 and 2001–2004) may have affected the interactions between genotype, phenotype, and environment in ways not captured by the samples collected more recently (2008 – 2015). Therefore, generalizability of GWAS and PVE over time cannot be rigorously assessed with the present data. Nevertheless, the presence of this pattern in our data set could be a signal of a broader underlying change in adaptive potential which should be further investigated.

This study presents novel findings on migration timing, opening the door to understanding components of the genomic architecture of migration timing in other long-distance migrants. The strong genomic variation and significant regions associated with purple martin migration timing could have important implications for adaptability in long-distance migrants. If the genomic potential for adaptability has decreased in recent years, this could hinder the ability of migrants to keep up with the pace of changing climates. Many portions of the genome are conserved across other bird species and vertebrates [28] and climate change

174 continues to affect migratory animals all over the world. Therefore, these findings bring us closer
175 to understanding a common basis for migration, which may have broad implications for a variety
176 of organisms.
177

178 # Methods

179 ## Reference and resequencing data

180 The reference genome was assembled using PacBio long and 10X linked reads generated
181 for a female martin from Manitoba, Canada. We used FALCON [29,30] to create the initial
182 assembly, then polished and scaffolded the genome with ArrowGrid, Pilon, and ARKS
183 [31,32,33]. Then we annotated the genome through MAKER [34]. We used skimSeq (low-
184 coverage whole-genome sequencing) to generate resequencing data [35] for an additional 87
185 birds (average coverage 2.7x per sample). Missing genotypes were then imputed with Beagle
186 [36], using information from the reference, surrounding genotypes, linkage disequilibrium
187 structure, and haplotype blocks [37]. These included 45 male and 42 female blood samples
188 collected from 13 different breeding colonies across North America between 2008–2015 (Table
189 S2). We filtered SNPs for quality (QUAL>20, MQ>20), max-missing (20%), minor allele
190 frequency (MAF>0.05), Hard-Weinberg equilibrium, and biallelic sites. Details on assembly,
191 annotation, sequencing, and filtering are in supplementary information.
192

193 ## Light-level geolocator analysis

194 Light-level geolocators were mounted during the breeding season using leg-loop
195 backpack harnesses and retrieved through recapture in the following year. Purple martin
196 behavior of aerial foraging and use of open habitats makes light-level geolocators ideal for

capturing sunrises and sunsets with minimal shading. The timing of these twilights is used to estimate the daily locations of birds over the entire year, using the midpoint of rise-set events to determine longitudes and day length for estimating latitudes [38]. We analyzed twilight times with BAStag and GeoLight [39,40], producing estimated daily locations to obtain migratory departure and arrival dates. Due to the correlation of departure with arrival timing for migratory journeys, we ranked individuals in order of timing for both dates and combined these values to determine overall timing phenotypes for spring migration.

Genomic architecture of migration timing

BSLMMs and LMMs were run using GEMMA [14], where we included the covariates of sex, year, age, and the first principal component (PC1) from a PCA summarizing genetic variation in our dataset. We summarized results from these runs for PVE and used posterior inclusion probabilities (PIP) to identify specific SNPs with strong associations to the timing phenotypes. PIP is the probability that the SNP is associated with the phenotypic variation [41] and following [42] we considered SNPs with PIPs > 0.1 important. Polygenic models were created using the PLINK v1.9 [43] and following Choi et al. (2020)'s PGS pipeline [44]. We used VCFtools [45] to estimate F_{ST} between the 10 earliest (originating from two Florida colonies) and 10 latest (originating from two Alberta colonies and one Virginia colony) spring migrants. Since this F_{ST} could be elevated by processes unrelated to migration, including linked background selection and population structure [46], we controlled for these potential effects by subtracting F_{ST} between Alberta and Florida (representing the northernmost and southernmost breeding regions) from values estimated between extreme timing phenotypes. This approached has been used in crows [47] and blackcaps [48] to isolate differentiation associated with specific

phenotypes. Additionally, we estimated F_{ST} between early and late migrants within Alberta and Florida populations separately to examine if we could recover the same signature of elevated F_{ST} in the same genomic region.

Genomic association with ecological spring indices

Green-up data were extracted from MODIS [20] and first bloom dates from the USA National Phenology Network [21]. We extracted data for each purple martin colony location over all available years. We ran BSLMMs for "historic" (2001-2004 for green-up and 1981-1984 for first bloom) and recent phenology data (up till 2015) and ran BSLMMs for each association test, using year, PC1, and colony as covariates. PC1 controls for population structure and colony accounts for the fact that birds from the same colony are assigned the same values for each environmental variable each year. The green-up (MODIS) association used all 87 purple martin individuals, and the first bloom (NPN) dates spanned 63 purple martin individuals (USA only).

References

1. Walther, G. *et al*. Ecological response to recent climate change. *Nature* **416,** 389–395 (2002).

2. Cohen, J. M., Lajeunesse, M. J. & Rohr, J. R. A global synthesis of animal phenological responses to climate change. *Nature Climate Change* **8**, 224–228 (2018).

3. Both, C., Bouwhuis, S., Lessells, C. M. & Visser, M. E. Climate change and population declines in a long-distance migratory bird. *Nature* **441**, 81–83 (2006).

4. Liedvogel, M., Åkesson, S. & Bensch, S. The genetics of migration on the move. *Trends in Ecology and Evolution* **26**, 561–569 (2011).

5. Caprioli, M. *et al*. Clock gene variation is associated with breeding phenology and maybe under directional selection in the migratory barn swallow. *PLoS ONE* **7**, e35140 (2012).

6. Mettler, R., Segelbacher, G. & Schaefer, M. H. Interactions between a candidate gene for migration (ADCYAP1), morphology and sex predict spring arrival in blackcap populations. *PLoS ONE* **10**, e0144587 (2015).

7. Peterson, M. P. *et al.* Variation in candidate genes CLOCK and ADCYAP1 does not consistently predict differences in migratory behavior in the songbird genus Junco. *F1000Research* **2** (2013).

8. McKinnon, E. A. & Love, O. P. Ten years tracking the migrations of small landbirds: Lessons learned in the golden age of bio-logging. *The Auk* **135**, 834–856 (2018).

9. Fraser, K. C. *et al.,* Continent-wide tracking to determine migratory connectivity and tropical habitat associations of a declining aerial insectivore. *Proc. R. Soc. B*. **279**, 4901–4906 (2012).

10. Neufeld, L. R., Muthukumarana, S., Fischer, J. D., Ray, J. D., Siegrist, J. & Fraser, K. C. Breeding latitude is associated with the timing of nesting and migration around the annual calendar among purple martin *Progne subis* populations. *Journal of Ornithology* 162:1009-1024 (2021).

11. Peona, V. *et al*. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology* **21**(1), 263-286 (2020).

12. Coelho, L. A., Musher, L. J., & Cracraft, J. A Multireference-Based Whole Genome Assembly for the Obligate Ant-Following Antbird, *Rhegmatorhina melanosticta* (*Thamnophilidae*). *Diversity*, **11**(19), 144 (2019).

13. Fraser, K. C. *et al*., Determining fine-scale migratory connectivity and habitat selection for a migratory songbird by using new GPS technology. *Journal of Avian Biology* **48**, 001-007 (2017).

14. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* **9**, e1003264 (2013).

15. Fuller, Z. L. *et al*. Population genetics of the coral Acropora millepora: Towards a genomic predictor of bleaching. *Science* **369**(6501). 2019

16. Jones, S., Pfister-Genskow, M. Benca, R. M. & Cirelli, C. Molecular correlates of sleep and wakefulness in the brain of the white-crowned sparrow. *Journal of Neurochemistry* **105**, 46-62 (2008).

17. Ma, C. *et al.* Sleep regulation by neurotensinergic neurons in a thalamo-amygdala circuit. *Neron* **103** (2019).

18. Wong, J. M. & Eirin-Lopez, J. M. Evolution of methyltransferase-like (METTL) proteins in metazoan: a complex gene family involved in epitranscriptomic regulation and other epigenetic processes. *Molecular Biology and Evolution* **38**, 5309-5327 (2021).

19. Saino, N. *et al.* Migration phenology and breeding success are predicted by methylation of a photoperiodic gene in the barn swallow. *Scientific Reports* **7**, 45412 (2017).

20. Friedl, M., Gray, J. & Sulla-Menashe, D. MCD12Q2 MODIS/Terra+Aqua Land Cover Dynamics Yearly L3 Global 500m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. (2019).

21. USA National Phenology Network. Spring Indices, Current Year - First Bloom - Spring Index. https://www.usanpn.org/data/visualizations. Accessed on 17 April 2020. (2017).

22. Colodro-Conde, L. *et al*. A direct test of the diathesis-stress model for depression. *Molecular psychiatry* **23,** 1590-1596. (2017).

23. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genetics* **9**(4) (2013).

24. Liedvogel, M., Åkesson, S. & Bensch, S. The genetics of migration on the move. *Trends in Ecology and Evolution* **26**(11), 561–569 (2011).

25. Barret, R. D. H. & Schluter, D. Adaptation from standing genetic variation. . *Trends in Ecology and Evolution* **23**(1), 38-44. (2008)

26. Lavallée, C. D. *et al*. The use of nocturnal flights for barrier crossing in a diurnally migrating songbird. *Movement Ecology* **9**, 21 (2021).

27. Henry, R. A. *et al.* Changing the selectivity of p300 by Acetyl-CoA modulation of histone acetylation. *ACS Chem. Biol* **10**, 146-156 (2015).

28. Sun, H., Skorgerbø, G., Wang, Z., Liu, W. & Li, Y. Structural relationships between highly conserved elements and genes in vertebrate genomes. *PLoS ONE* **3**, e3727 (2008).

29. Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods* **13**, 1050–1054 (2016).

30. Chin, C. S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563–569 (2013).

31. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **27**, 722–736 (2017).

32. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).

33. Coombe, L. *et al.,* ARKS: Chromosome-scale scaffolding of human genome drafts with linked read kmers. *BMC Bioinformatics* **19**, 1–10 (2018).

34. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics* **48**, 4.11.1-4.11.39 (2014).

35. Malmberg, M. M. *et al.* Evaluation and recommendations for routine genotyping using skim whole genome re-sequencing in canola. *Front. Plant. Sci*. **9** (2018).

36. Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics* **98**, 116–126 (2016).

37. Golicz, A. A., Bayer, P. E. & Edwards, D. Skim-Based Genotyping by Sequencing. *Methods in Molecular Biology* **1245**, 257–270 (2015).

38. Hill, R. D. Theory of geolocation by light levels. In B.J.L. Boeuf, & R.M. Laws (Ed.), Elephant seals: Population ecology, behaviour and physiology p. 227-236. Berkeley, CA: University of California Press (1994).

39. Wotherspoon, S., Summer, M. & Lisovski, S. BAStag: Basic Data Processing for Light Based Geolocation Archival Tags. Version 0.1.3. (2016).

40. Lisovski, S. & Hahn, S. GeoLight-processing and anslysing light-based geolocator data in R. *Methods in Ecology and Evolution* **3**, 1055-1059 (2012).

41. Gompert, Z., Lucas, L. K., Nice, C. C. & Buerkle, C. A. Genome divergence and the genetic architecture of barriers to gene flow between *Lycaeides idas* and *L. melissa*. *Evolution* **67**, 2498–2514 (2013).

42. Pfeifer, S. P. *et al*. The evolutionary history of Nebraska deer mice: local adaptation in the face of strong gene flow. *Molecular Biology and Evolution* **35**, 792-806 (2018).

43. Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559-575 (2007).

44. Choi, S. W., Mak, T. S. & O'Reilly, P. F.. Tutorial: a guide to performing polygenic risk score analysis. *Nat Protoc* **15**, 2759-2772 (2020).

45. Danecek, P. *et al*. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

46. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology* **23**, 3133-3157 (2014).

47. Vijay, N. *et al.* Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nature Communications* **7**, 13195 (2016).

48. Delmore, K. *et al.* The evolutionary history and genomics of European blackcap migration. *eLife* **9**, e54462 (2020).

Acknowledgements

We thank K. Applegate, J. Barrow, L. Burgess, L. Chambers, T. Cheskey, P. Clifton, B. Dietrich, J. Fischer, G. Hvenegaard, P. Kramer, P. Mammenga, N. Mickle, M. North, M. Pearman, J. Ray, A. Ritchie, A. Savage, T. Shaheen, J. Siegrist, C. Silverio, B. Stutchbury, and J. Tautin for sample collection; M. Przeworski, Z. Fuller, and C.J. Garroway for conceptual input; F. Taborsak-Lines, M. Ormestad, P. Ewels, C. Wang, O.V. Pettersson, M.B. Mosbech, and C. Tellgren-Roth for generating data for the reference assembly; K. Jeffries, National Genomics Infrastructure (NGI)/Uppsala Genome Center, and Texas A&M AgriLife Genomics and Bioinformatics Service for laboratory support; Texas A&M High Performance Research Computing and Compute Canada for computing services. This work was supported by funding from the University of Manitoba, Texas A&M University, Natural Sciences and Engineering Research Council, Research Manitoba, the Society of Canadian Ornithologists, Sigma Xi, RFI/VR & Science for Life Laboratory, Sweden, and we further acknowledge the Knut and Alice Wallenberg Foundation and the Swedish Research Council, and SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing and access to the UPPMAX computational infrastructure.

Author contributions

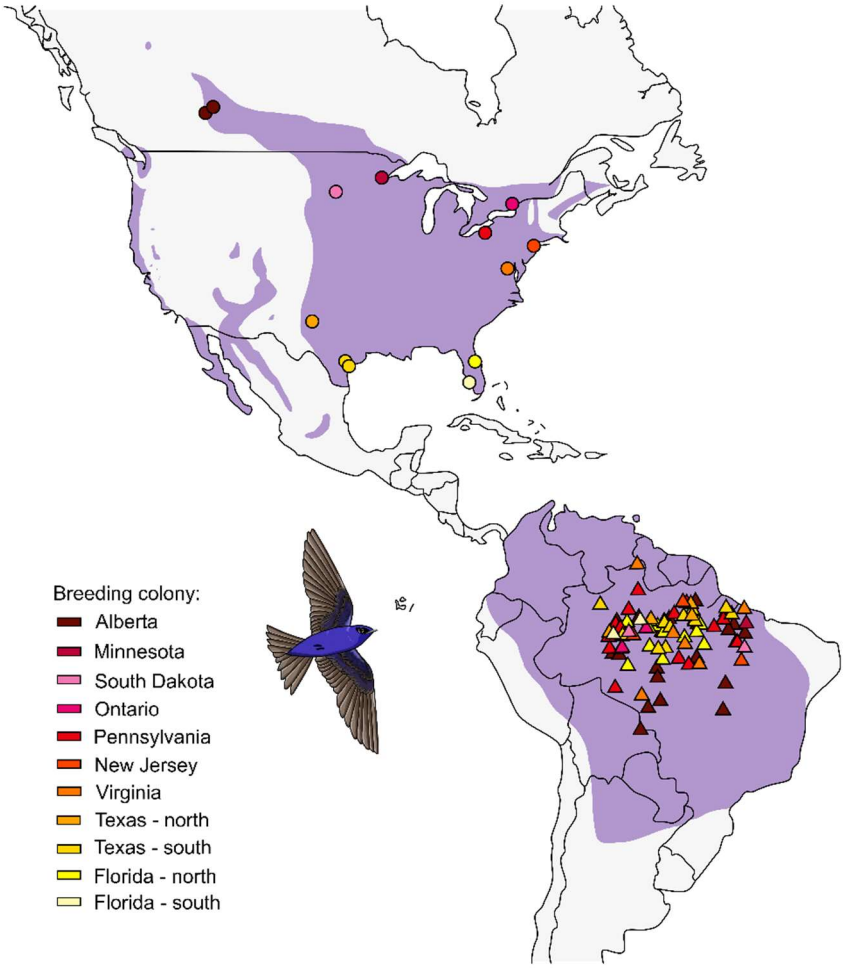
K.F. and K.D. designed and supervised the study. K.F. conducted/coordinated the collection of migration data and blood sampling. E.D. conducted laboratory work and performed genomic analyses with K.D.'s guidance. M.T. assisted with association modelling and conducted polygenic score analyses. A.S. coordinated sequencing for the reference genome and provided guidance on the assembly. E.D. analyzed light-level geolocator data K.F.'s guidance. E.D., K.F., and K.D. wrote the manuscript with feedback from M.T. and A.S.

Data availability statement

Raw sequencing data is available on SRA NCBI BioProject PRJNA772931. Source code for reference genome and resequencing work is available at https://github.com/edegreef/PUMA-reference-genome and https://github.com/edegreef/PUMA-resequencing-data.

Competing interests statement. The authors declare no competing interests.

426 **Figures**

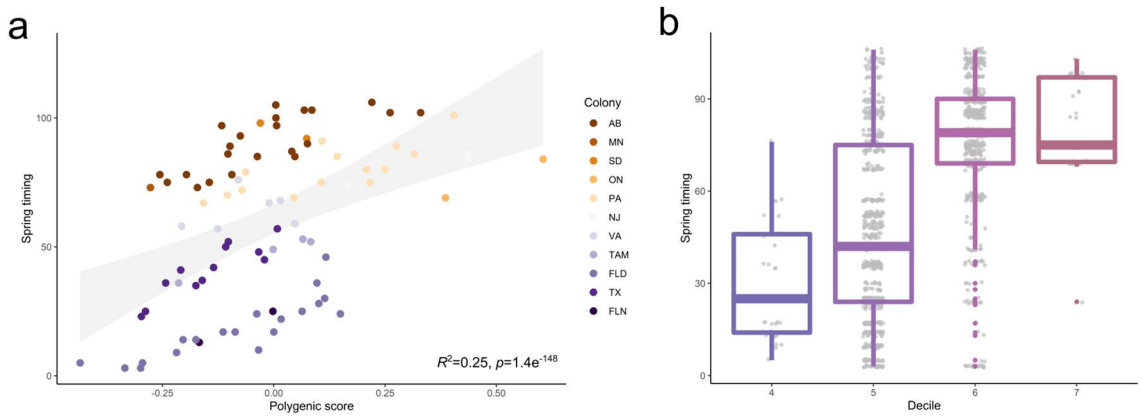


427

428 **Figure 1.** Purple martin breeding and wintering distribution (purple), including sampling sites
429 for 87 individuals in their North American breeding range (circles) and their respective South
430 American wintering destination before spring departure (triangles). Individuals from distinct
431 breeding colonies overlapping at the wintering grounds demonstrate weak connectivity between
432 breeding and wintering sites (such as observed in Fraser et al. 2012, 2017).

433

434



435

436 **Figure 2.** (a) Polygenic scores of spring migration timing for purple martins (n = 87) colored in

437 order by latitude, and linear regression standard error is colored in gray. (b) Individuals in lowest

438 decile of predicated polygenic scores (PGS) had earlier migration timing compared with

439 individuals in higher deciles with later timing.

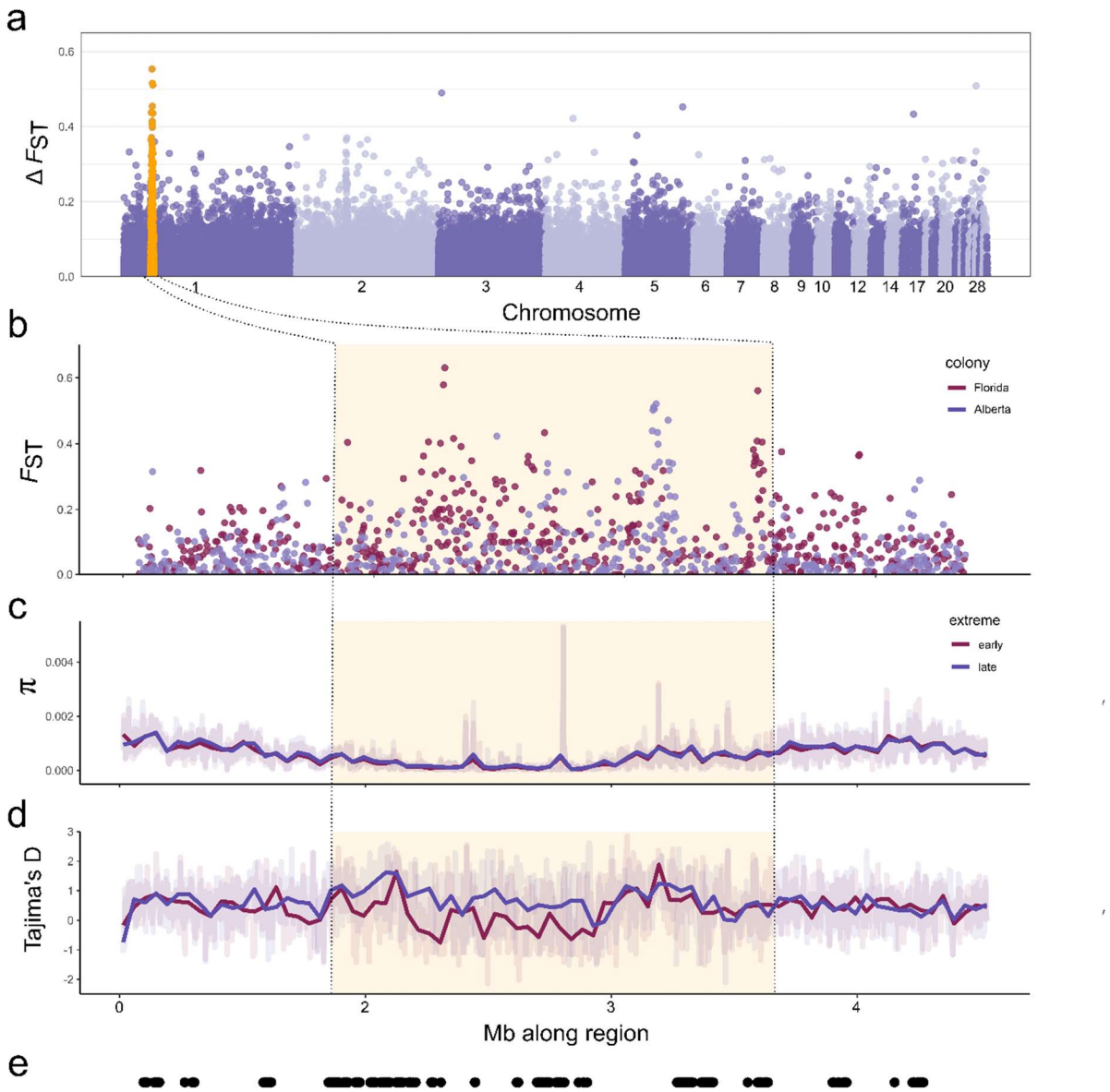


Figure 3. (a) Net genetic differentiation (ΔF_{ST}) across autosomes in 5 kb non-overlapping windows between earliest and latest spring migrants. The elevated region on chromosome 1 is highlighted in orange, with plots examining this region to show (b) F_{ST} within Alberta and within Florida, (c) nucleotide diversity (π), (d) Tajima's D, and (e) location of genes in this region as black dots.