Title: **Post-glacial expansion dynamics, not refugial isolation, shaped the genetic structure of a migratory bird, the Yellow Warbler (*Setophaga petechia*).**

Authors: Eleanor F. Miller* [a], Michela Leonardi [a], Robert Beyer [a], Mario Krapp [a], Marius Somveille [a, b, c], Gian Luigi Somma [a], Pierpaolo Maisano Delser [a] **, Andrea Manica [a] **

[a] Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK.

[b] Department of Biology, Colorado State University, Fort Collins, CO, 80521 USA

[c] Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK

** Equal contribution

* Corresponding Author: Eleanor F. Miller, Department of Zoology, University of Cambridge, Downing Street, Cambridge, CB2 3EJ, UK. Email: em618@cam.ac.uk

ORCIDs:

E.F.M.	0000-0002-3213-5714
M.L.	0000-0001-8933-9374
M.S.	0000-0002-6868-5080
G.L.S.	0000-0003-0383-5719
M.K.	0000-0002-2599-0683
A.M.	0000-0003-1895-450X
P.M.D.	0000-0002-1844-1715

Abstract

During the glacial periods of the Pleistocene, swathes of the Northern Hemisphere were covered by ice sheets, tundra and permafrost leaving large areas uninhabitable for temperate and boreal species. The glacial refugia paradigm proposes that, during glaciations, species living in the Northern Hemisphere were forced southwards, forming isolated, insular populations that persisted in disjunct regions known as refugia. According to this hypothesis, as ice sheets retreated, species recolonised the continent from these glacial refugia, and the mixing of these lineages is responsible for modern patterns of genetic diversity. However, an alternative hypothesis is that complex genetic patterns could also arise simply from heterogenous post-glacial expansion dynamics, without separate refugia. Both mitochondrial and genomic data from the North American Yellow warbler (*Setophaga petechia)* shows the presence of an eastern and western clade, a pattern often ascribed to the presence of two refugia. Using a climate-informed spatial genetic modelling (CISGeM) framework, we were able to reconstruct past population sizes, range expansions, and likely recolonisation dynamics of this species, generating spatially and temporally explicit demographic reconstructions. The model captures the empirical genetic structure despite including only a single, large glacial refugium. The contemporary population structure observed in the data was generated during the expansion dynamics after the glaciation and is due to unbalanced rates of northward advance to the east and west linked to the melting of the icesheets. Thus, modern population structure in this species is consistent with expansion dynamics, and refugial isolation is not required to explain it, highlighting the importance of explicitly testing drivers of geographic structure.

Introduction

It has frequently been shown that seemingly continuously distributed populations in the Northern Hemisphere harbour geographic structure in their genetic diversity. Indeed, within North America, many widespread and migratory passerines exhibit clear differences in both migration patterns and genomic diversity between eastern and western populations e.g. (1–3). This pattern has been interpreted as the consequence of glaciations, during which species were forced southwards, forming isolated, insular populations that persisted in disjunct regions known as refugia (4,5). According to this narrative, as ice-sheets retreated, species recolonised the continent from these glacial refugia, and the subsequent mixing of these lineages is responsible for modern patterns of genetic diversity.

However, even though the cycles of expansion and contraction could have fragmented ranges, leading to multiple glacial refugia in some species, multiple glacial refugia have not been demonstrated for all species e.g. (6,7). Indeed, it is becoming clear that glaciations in North America might not have driven range fragmentation as ubiquitously as it has previously been assumed, e.g. (8,9).

What other processes might then have shaped the genetics of modern populations? Range expansions have been shown to have the potential to leave profound signatures in the genetic structure of metapopulations through repeated founder events (10). An extreme consequence of this process is gene surfing, when rare variants can become common through stochastic sampling during a founder event, and then be spread widely at high frequency during the subsequent expansion. An important role for the recolonization dynamics in shaping modern-day population structuring has been recently put forward for a trans-continentally distributed species, the painted-turtle, *Chrysemys picta* (11). Reid et al. (11) demonstrated that, for this species, genetic differentiation during range expansion and isolation-by-distance are more likely to have driven modern-day population diversity than isolation in allopatric refugia.

The difficulty in quantifying the role of the range change dynamics on the genetic structure of species is that the results are highly dependent on the detail of the dynamics. Whilst it is straightforward to build simple spatial models that represent a range expansion, capturing the spatial and temporal heterogeneities of the real process is challenging. A possible solution is to use climate informed spatial genetic models (CISGeMs), which use climate reconstructions to condition the local demography of individual populations within a map, and quantify the demographic parameters by Approximate Bayesian Computation comparing predicted and observed genetic quantities (see Fig. 1). This approach has been successfully used to reconstruct the dynamics of the out of Africa expansion of humans (12,13).

In this paper, we use the CISGeM framework to explore the past range dynamics of the Yellow Warbler. This species is an abundant passerine species with a large continuous contemporary range and clear geographic population structuring, for which range-wide genomic data are available (14). Here we test whether today's patterns of genetic structure in the North American Yellow Warbler (*Setophaga petechia*) can be best explained by recolonization from isolated glacial refugia, or if, more simply, heterogenous post-glacial expansion dynamics, without separate refugia, may have been enough to result in observed patterns today. Firstly, we describe genetic patterns that are found in the Yellow Warbler today from an empirical RAD-seq dataset. Then, we fit a spatially explicit model of population growth and expansion that accounts for past climatic variation to the dataset. By simulating the genetics and fitting to the observations with an Approximate Bayesian Computation framework, we investigate to what extent these recolonization dynamics could explain modern genomic patterns.

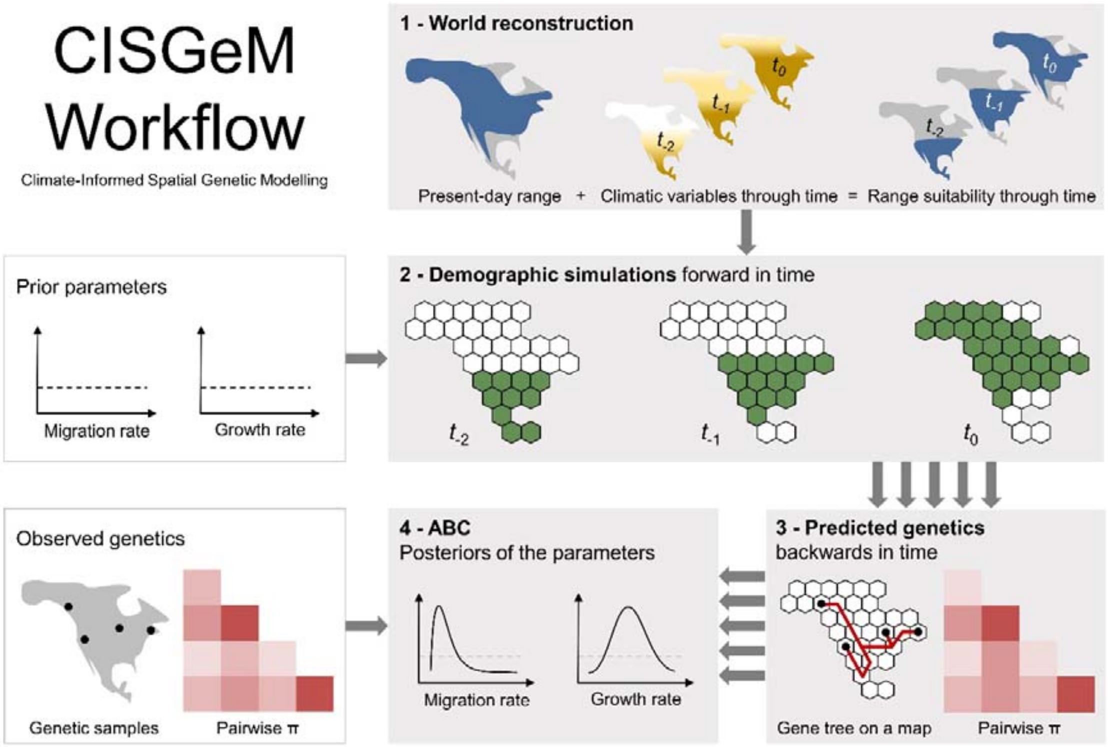

Figure 1. A schematic representation of the Climate-Informed Spatial Genetic Modelling framework implemented in this paper.

Results

Observed genetics:

The 200 samples included in our study came from 21 sites across the modern breeding range of the North American Yellow Warbler. Sample sizes per site ranged from 6 to 20 individuals (see Materials and Methods). Analysis of the genetic structure (15) of the Yellow Warbler population revealed a clear longitudinal divide, with distinct East and West clusters that converge in the centre of the continent. Populations were with a mixture proportion of less than 70% for either of the two clusters (‘East’ and ‘West’) were grouped in the ‘Central’ category. This pattern is congruent with both the distribution of mitochondrial haplotypes (16) and patterns of migratory connectivity (17) in this species.

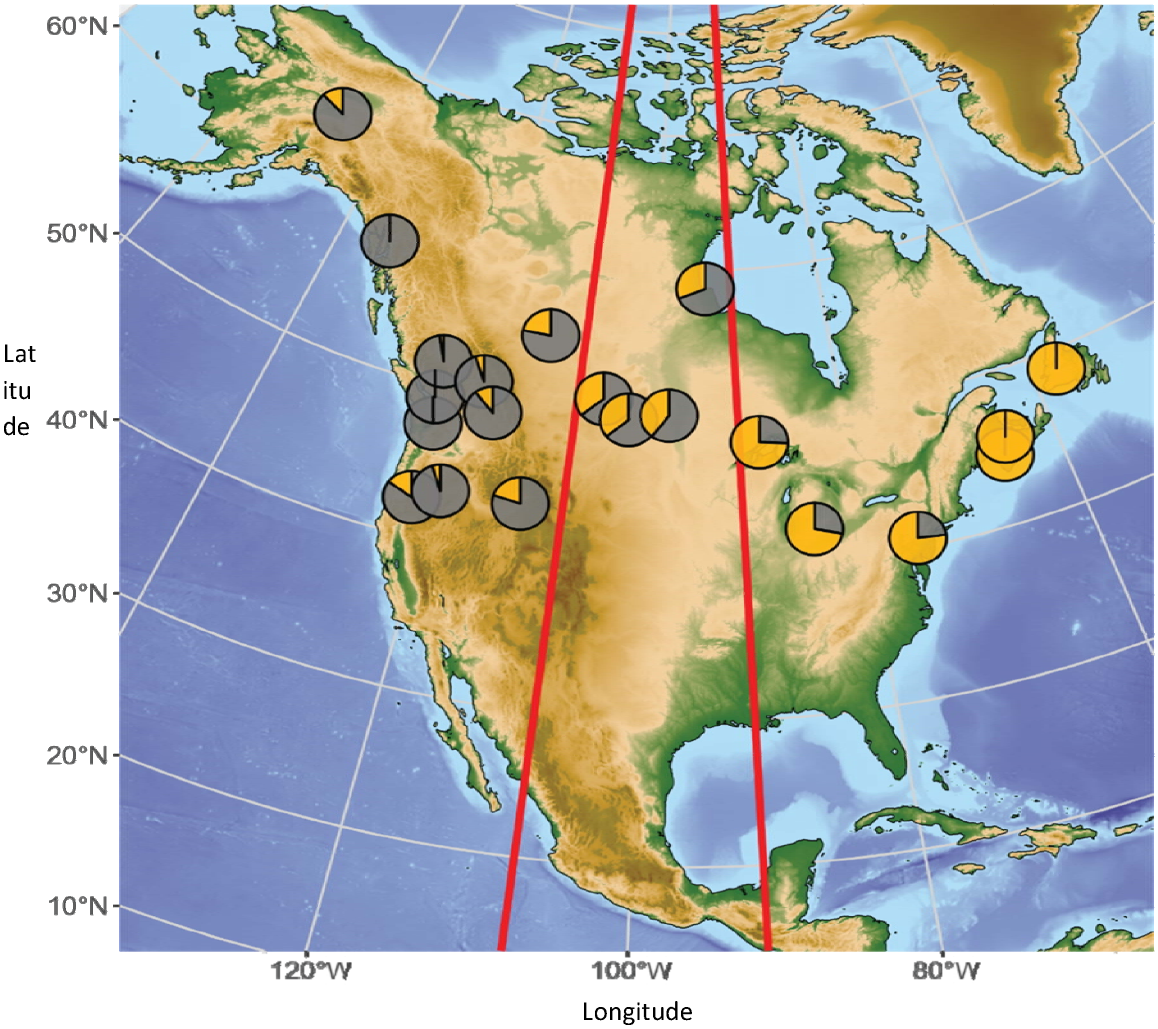
60°N
50°N
40°N
30°N
20°N
10°N
Lat
itu
de
120°W
100°W
80°W
Longitude

Figure 2. Genetic clustering results (*K* = 2) results for all 21 populations in our study. Red lines separate the populations longitudinally into West, Central, and East.

Species Distribution Modelling for world reconstruction:

CISGeM requires the reconstruction of range suitability maps through time (step 1 in Fig. 1). We built a Species Distribution Model (18) for Yellow Warblers based on modern data, and projected back in time using paleoclimate reconstructions. The raw species occurrences data to define the present-day range, downloaded from the Global Biodiversity Information facility (GBIF, our data can be found at https://doi.org/10.15468/dl.jfkwcg), totalled 1,573,147 data points. After filtering for coordinate accuracy, allowing an attributed error of 1km maximum, and filtering to only include points found within the BirdLife breeding and resident geographical ranges (BirdLife International and Handbook of the Birds of the World 2018), we were left with 177,202 data points. As SDM works on presence/absence data and not frequencies, we retained only one presence per 0.5° grid cell, further refining this dataset down to 3,364 observations. With these observations we selected the four most informative, uncorrelated (threshold=0.7), bioclimatic variables to base our model on: Leaf Area Index (LAI), BIO7 (Temperature Annual Range), BIO8 (Mean Temperature of Wettest Quarter), BIO14 (Precipitation of Driest Month). Observations were further thinned based on a maximum distance between points of 70 km, leaving 1,188 presences; this procedure is used to correct for uneven sampling biases (Steen et al. 2020). We fitted SDMs to predict the probability of occurrence in each grid cell. We used an ensemble of four different algorithms: generalised linear models (GLM, (19)), generalized boosting method (GBM,(20)), generalised additive models (GAM, (21)), and random forest (22). Models were run performing spatial cross validation with 80% of the data used to train the algorithm and the remaining 20% to test it.

At present, the predicted potential distribution matches well the best range estimates for the species (Supplementary Fig. 1.). Based on paleoclimate and vegetation reconstruction (see Materials and Methods), range projections from the present day back to 50 thousand years ago suggest that the distribution of habitat suitable for the Yellow Warbler expanded and contracted, to various degrees, multiple times. The potential range underwent a substantial contraction into the south of the continent at the peak of the Last Glacial Maximum (LGM), ~21kya, before beginning to re-expand (Supplementary Fig. 2. A-B), but the range never separated into distinct eastern and western refugia. Following the LGM, the retreat of the Cordillerian Ice Sheet was asymmetric: in the west, the ice started to retreat at about 18ka (23) with the opening of a corridor that progressively expanded to the higher latitudes, whereas the eastern and central part of the Laurentide Ice Sheet began retreating much later (24). By 13kya this deglaciated terrain became habitable for Yellow Warblers according to our SDM (Supplementary Fig. 2. C). From then on, as the ice sheets retreated further, habitat to the east of the continent and in the central area deglaciated, becoming increasingly viable (Supplementary Fig. 2. D-F).

Climate Informed Spatial Genetic Model

The reconstructed range suitability maps over time were used as an input for CISGeM. In this framework, the genetics of multiple populations can be modelled within a spatially explicit reconstruction of the world where the suitability of each deme changes through time according to the SDM back-cast suitability scores (Fig. 1). Using an Approximate Bayesian Computation framework, we fitted basic demographic parameters such as population growth rate and migration, as well as the link between SDM suitability scores and local population sizes. The mean pairwise genetic differentiation (π) between populations in each of the three clades (East, Central and West) were used as summary statistics that had to be matched by the model. We performed a Monte-Carlo sweep of the input parameters (Table 1.), generating a total of 61,504 simulations.

Visual inspection of the values of pairwise differentiation among the clades revealed that the model was able to recapitulate the observations in a realistic fashion (see Supplementary Fig. 3. A for a PCA plot of the values predicted by the model vs the observations, and Supplementary Fig. 4. for all individual summary statistics). We then formally tested model fit with '*gfit*' from the '*abc*' package (25) in R (see Materials and Methods for details). This test verifies that the distance between the observed and the simulated data is not significantly larger than the distance of a random simulation to other simulations (and thus that the model is able to capture the patterns seen in the data): our model recovered a p value of 0.379 which implies a good fit (Supplementary Fig. 3. B).

We used a random forest algorithm (ABC-RF) (26) to generate posterior probabilities of the input demographic parameters given the observed levels of pairwise population differentiation (Supplementary Fig. 5.). The metapopulation dynamics was characterised by an expansion dynamics with moderate to strong bottlenecks (as determined by a relatively low directed expansion coefficient, m_d, which defines the proportion of individuals that move into an unoccupied area, Supplementary Fig. 5. C), followed by limited subsequent migration (low values in the undirected expansion coefficient rates m_r, Supplementary Fig. 5. E, in accordance with observations that this species tends to be philopatric in its breeding range). Such signals suggest an expansion characterised by sequential founder events that would have set up a pattern of isolation by distance along the colonisation routes that was preserved by the limited migration afterwards.

From the top 2.5% best fitting simulations (n=1055 runs), we reconstructed the demography of the species through space and time. The average demographic profile, calculated as a weighted mean of population size across these simulations, shows that the Yellow Warbler was forced to contract its range at the peak of the Last Glacial Maximum (~21kya) as the ice sheets grew across the north of the continent (Fig. 3. A&B). At this point, the population existed in a restricted but broadly continuous range in the south of the continent. As the climate ameliorated, northward range expansion became possible. However, the pattern of recolonization was uneven. By 13kya, our model reconstructs an expansion mostly following the corridor that opened between the Laurentide and Cordilleran icesheets on the west of the continent, whilst expansion on the eastern side was limited (Fig. 3. C). The western spread continued at a pace with the melting of the Cordilleran ice sheet (Fig. 3. D), but the eastern expansion lagged behind due to the slower melting of the

Laurentide icesheet (Fig. 3. E). The central and eastern part of the continent were fully colonised by 5kya, when the ice sheets were fully melted (Fig. 3. F).

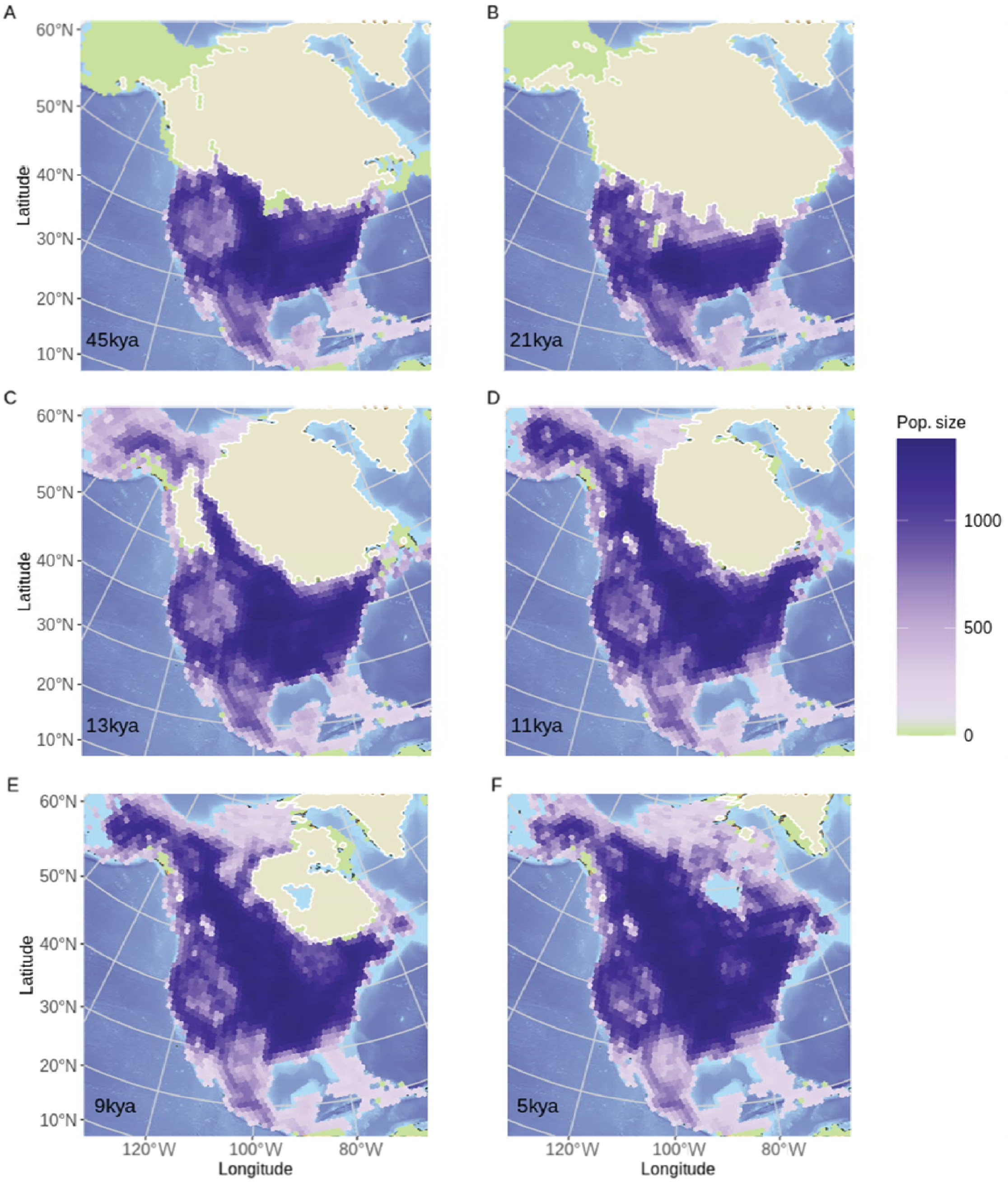

Figure 3. Weighted mean population size (per deme) of Yellow Warbler at A) 45kya, B) 21kya, C) 13kya, D) 11, E) 9kya, F) 5kya, from 1055 simulations retained during the parameter estimation. Dark grey regions are areas uninhabitable for yellow warblers at the given point in time.

The importance of this asymmetric expansion in setting up the patterns of genetic diversity across the range of Yellow warblers can be seen by mapping geolocation of common ancestor (CA) events that occurred between populations. These events allow us to reconstruct gene flow through time, as shaped by colonisations and subsequent connectivity, revealing how the patterns of diversity have emerged. Using two populations from two regions each time, we plotted locations of the CA events between the East and West, Central and East, and Central and West regions (Fig. 4. A). When we considered populations from the East and West (Fig. 4. B), we can see that common ancestor events between these two clusters show a “v” shape that matches closely the shape of the ice sheets at 13kya, when the postglacial expansion occurred. Importantly, the same pattern was also found when we considered only West and Central populations (Fig. 4. C), albeit with a greater intensity of events in the corridor between the two populations. Even though we did not have any population from the East, common ancestors event reveal that central populations are linked to that area, thus representing a mix of the western and eastern arms of the expansion. The same is true when we considered only Central and East populations Fig. 4. D). Together with the reconstructed demography in Fig. 3., this pattern shows the importance of the early expansion up the west coast, followed by subsequent expansion up the east coast, in setting up an initial divergence of the clades, which then mixed in the central region comparatively recently. The signature left by the relatively strong founder events that occurred during the expansion have not yet been eroded by the relatively low levels of migration, explaining the current patterns of genetic diversity and structure in this species.

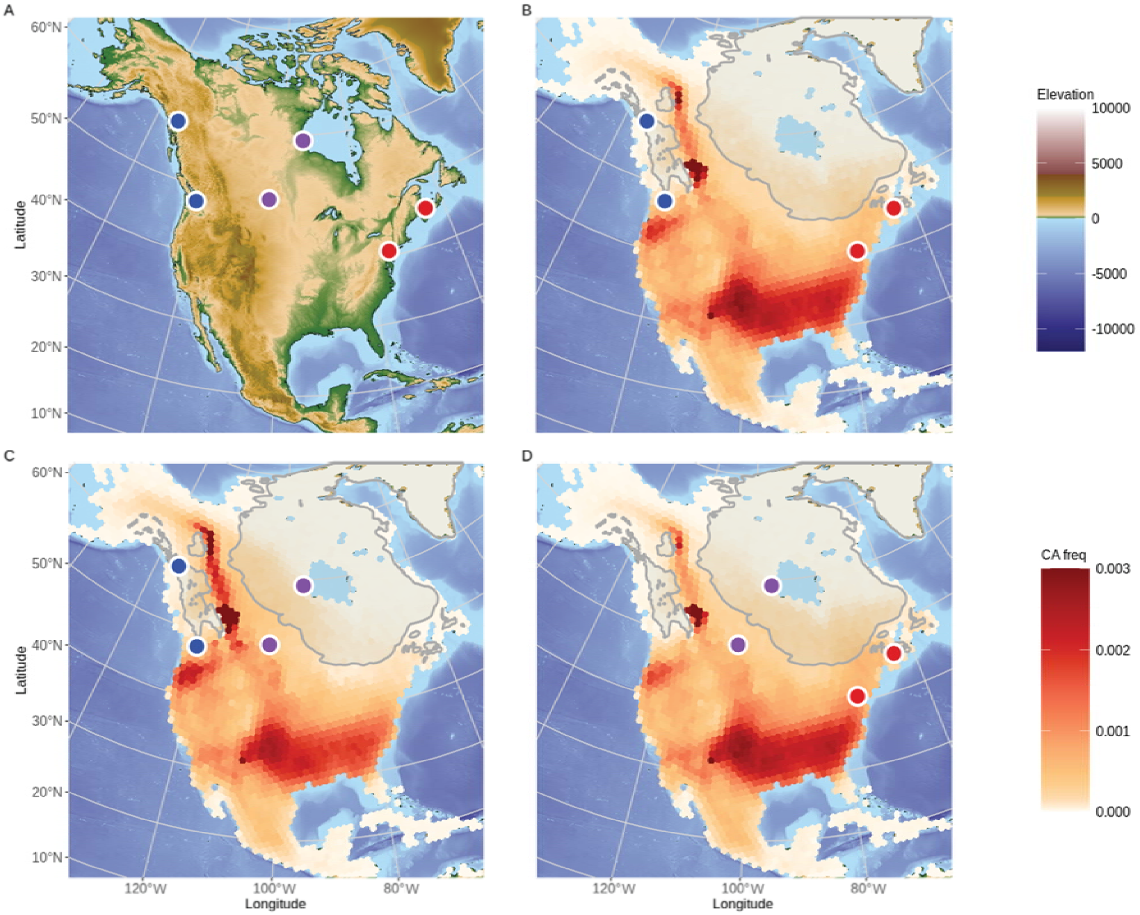

Figure 4. The location of common ancestor (CA) events are plotted across a map of North America. A) is an elevation map of the region with all six sampling locations labelled. In B-D colour density represents proportion of total CA events on the map that occur in each deme. B) is an analysis based on two populations each from West (blue) and Central (purple) regions, C) is based on Central and East (red) regions, and finally D) the West and East regions.

Discussion

In this study, we examined the relative roles of different forces that may have driven modern day genetic structuring in a widespread species. We used a set of complementary datasets to explore structure in the North American Yellow Warbler (*Setophaga petechia*), a common passerine species. By integrating genetic data and climatic and environmental variables through time into a spatially-explicit modelling framework (CISGeM), we were able to build a detailed reconstruction of the population dynamics for this species, stretching back through the last fifty thousand years. Our model was able to reconstruct population size changes, track potential range expansions, and simulate recolonisation dynamics, whilst capturing the genetic structure found in the modern population. With this information, we were able to explore the extent to which expansion dynamics could explain modern genomic patterns of the Yellow Warbler.

East-west population structure, as found in the Yellow Warbler, is not an uncommon pattern in North America. These genetic differences, as well as variation in other traits such as migratory behaviour, are often considered to support the existence of isolated refugia during glaciations (e.g. (16,27)). However, recent work on refugia has shown that the patterns of diversity found in the Northern Hemisphere only fit the expectations from cyclic expansion-contraction fragmenting ranges and driving genetic variation at a coarse level (8,9,28).

By explicitly modelling the recolonization dynamics, we have demonstrated a plausible explanation for the formation of genetic structure over time, without the need of multiple glacial refugia. The dynamics of the modelled Yellow Warbler recolonization show that, to a large extent, this passerine species tracked the uneven (asynchronous) retreat of the Laurentide Ice Sheet, with a longitudinally unequal progression northward (Fig. 3.). Despite the species exhibiting a single large glacial refugium, the asymmetrical pattern of re-expansion generates the genetic structure of east, west, and central population clusters found in the empirical genetic data. This implies a key role for post-glacial re-expansion in shaping modern-day populations.

The important role of re-expansion dynamics has recently been highlighted in a range of different species e.g. (8,11,28), though it would be naïve to assume that the complex patterns of diversity found in real populations could be easily explained by a single mechanistic process (29). Our work highlights that, at the very least, modern population diversity and structure may have originated from a combination of different processes, each of which needs to be carefully considered.

We acknowledge that, within this framework, we were unable to consider the possible influence of biotic interactions which may have impacted the pattern of recolonization (30). Our model also works with demes that are discrete spatial units of a fixed size, allowing for a step change in the likelihood of common ancestor events occurring within the deme and outside it. Moving away from the discretisation of space could help further 'naturalise' our model, and indeed models that incorporate continuous space are rapidly advancing (31). However, there are still major computational challenges to overcome before these tools would be suitable for an area on the scale of this study.

Whilst theories that describe broad patterns have been crucial to increasing our understanding of the likely impacts environmental changes have had on populations, we now realise that North American avifauna is probably a composite of species with different histories (32). Species have responded individually to the rapid climate changes faced in the Pleistocene and therefore we would not wish to claim our findings refute the existence and effect of North American glacial refugia for birds. However, now the resources and techniques exist to study the idiosyncratic responses of different species, and it will be possible to assess the importance of isolated refugia in shaping the genetic structure of species. Furthermore, an increased understanding on the different population dynamics that underlined species responses to the large climatic changes that occurred over the last glacial cycle might provide an important tool to refine our ability to predict the responses of species to anthropogenic change in the future.

Materials and Methods

Study species

The North American Yellow Warbler (*Setophaga petechia*) is a small, riparian, migratory passerine. Today, this common species is widely distributed across the continent. However, despite its large and well-connected contemporary range, the Yellow Warbler exhibits spatial structure across its range, including multiple mitochondrial clades (16) and clear isolation by distance (33). Although not a species of concern, the Yellow Warbler has recorded a declining population trend in the North American Breeding Bird Survey between 1966-2015, triggering several studies looking into the species ability to cope in the face of a rapidly changing climate (33,34). One such study by Bay et al. (33) built RAD-seq data from individuals sampled across the species' range in order to explore potential population trends in response to future climate scenarios. Such data was made available on GenBank and forms the basis of our empirical dataset here.

Raw genetic data

RAD sequence data for North American Yellow Warblers (*Setophaga petechia*) from 21 populations (33) were downloaded from the NCBI Sequence Read Archive (SRA). From the 269 accessions associated with the Bay et al. paper we chose to focus on only the individuals included in the original analysis (n = 223), individuals for which full information about their breeding population was available. A further 22 samples were dropped as the file sizes were under 75MB and, therefore, were likely to have low coverage. One final exclusion was made, GenBank accession number SRR6366039, as the sample was found to be an outlier with a measure of diversity higher than the range of all other samples, despite comparable levels of coverage and number of sites. This left 200 samples for further analysis. These individuals were sampled from across the modern population range, providing a good overview of the population genetics of this species, see Fig 2. for sampling locations.

Clustering analysis

RAD-seq methods are known to create specific biases in estimated allele frequencies, potentially affecting downstream analysis of the data (35). Using allele frequencies derived directly from the sequence data in a genotype-free method has been shown to account for RAD-seq specific issues, improving population genetic inferences (35). Therefore, we used Analyses of Next-Generation Sequencing Data (ANGSD) (36,37) to infer genotype likelihoods directly from aligned BAM files. Filters were set to only include SNPs with a p value of $< 2 \times 10^{-6}$ and only keep sites with at least 100 informative individuals. These ANGSD genotype likelihood values were then used as input for NGSadmix to calculate population admixture, setting a *K* (presumed cluster number) value of 2 and keeping minimum informative individuals at 100.

Observed genetics for CISGeM

In order to calculate pairwise π (the average number of differences between two sequences, normalised by the number of available positions), we first calculated genotype likelihoods in ANGSD. Input files were aligned BAM files, we used the samtools genotype likelihood method and inferred the major and minor allele from these likelihoods, with the command below:

```
angsd -GL 1 -out genolike -doGlf 1 -doMajorMinor 1 -bam bam.filelist
```

We then computed pairwise π from the ANGSD output; since our population genetic simulations (see below) modelled haploid samples (as it is the case of most genetic simulators, e.g. msprime (38)), we used the below formula:

$$\frac{pAB_{AB} \cdot 0.5 + pAA_{BB} + pAA_{AB} \cdot 0.5 + pAB_{AA} \cdot 0.5}{nSites}$$

In order to make the modelling computationally feasible, we then investigated how many samples were needed to get a reliable estimate of π for each population (Supplementary Fig. 6.). This analysis showed that five diploid individuals, or ten chromosomes, provided a reasonable compromise for noise. All estimates of pairwise π were therefore re-computed with only five individuals per population. As estimates were consistent with the values from the full dataset (Supplementary Fig.

7.), for computation efficiency of the model, all future analyses were based on this subset of the data.

Species Distribution Modelling for world reconstruction

The range and population size of a species changes in time and space according to fluctuations in resources and environmental conditions. In order to build a spatially explicit model it is first necessary to use Species Distribution Modelling (SDM) to reconstruct how population ranges and demographics may have changed through time. For this study an SDM analysis was undertaken using an R (39) pipeline.

Climate reconstructions:

Climate data for North America were drawn from a 0.5° resolution dataset for 19 bioclimatic variables; Net Primary productivity (NPP), Leaf Area Index (LAI) and all the BioClim variables (40) with the exclusion of BIO2 and BIO3; covering the last 50,000 years in 1,000 year time steps from the present to 22kya and in 2,000years time steps before that date (41). This dataset was originally constructed from a combination of HadCM3 climate simulations of the last 120,000 years (42), high-resolution HadAM3H simulations of the last 21,000 years (43), and empirical present-day data. The data had been downscaled and bias-corrected using the Delta Method (44). Bioclimatic variables through time were then used as input data to inform the SDM.

SDM data preparation:

Species occurrences data for the present day were initially downloaded from the GBIF database (https://www.gbif.org), the original downloads are available at the following DOI: S. petechia 10.15468/dl.jfkwcg (GBIF.org). These data were then filtered based on the attributed accuracy of the coordinates (maximum error: 1 km) and additionally, only points that were within Birdlife breeding and resident geographical ranges (45) were retained. Remaining occurrences were then matched to the 0.5° resolution grid used for the palaeoclimatic reconstructions and, as the method works on presence/absence data and not frequency, only one presence per grid cell was kept.

This cleaned observation dataset was then used to define a set of informative bioclimatic variables with the most influence on the species distribution for use in the Species Distribution Model (SDM), through visual check of how much the distribution of the variable values differed between the observation points and the whole area. We selected the variables with highest differences between the two curves, which are most likely to be relevant for the species, and then, in order to avoid using highly correlated variables, which may increase noise in the data, we constructed a correlation matrix between the variables associated with each of the retained observations. Where two values were highly correlated, the variable with the lowest overall correlation across the matrix was kept, allowing us to select a set of uncorrelated variables (threshold = 0.7) leaving us with the following ones to be used for SDM modelling: LAI (leaf area index), BIO7 (Temperature Annual Range), BIO8 (Mean Temperature of Wettest Quarter), BIO14 (Precipitation of Driest Month).

Geographic biases in sampling effort are common when observation data are collected opportunistically, such as the data in the GBIF database. In order to reduce this bias, we thinned our dataset using the R package *spThin* (46) enforcing a minimum distance of 70 km between observations. Given the random nature of removing nearest-neighbour data points, we repeated this step 100 times (‘rep’ = 100) retaining for further analysis the result with the maximum number of observations after thinning.

SDM modelling:

The SDM was built with the R package biomod2 (47) following the same procedure used in Miller et al. (48). The thinned observation dataset was used as presences whilst the landmass of North America was considered as background. The same number of pseudo-absences as presences were then drawn five separate times, at random, from outside the BirdLife resident and breeding masks: creating five independent datasets for analysis. For each data set, following Bagchi et al. (49), models were then run independently using four different algorithms: generalised linear models (GLM), generalized boosting method (GBM), generalised additive models (GAM), and random forest.

Spatial cross-validation was used to evaluate the model; 80% of the data were used to train the algorithm and the remaining 20% to test it. Initially, both the presences and the five pseudoabsences datasets were subdivided in 14 latitudinal bands using the R package BlockCV (50). Each band was given a 'band ID number', looping sequentially through numbers 1-5 until all bands were labelled. Then the bands were assembled into five working data splits grouped by their band ID (numbers 1-5). This was performed to maximise the probability of having at least some presences in all five data splits as a data split cannot be used for evaluation if it contains only absences. Each of the four models (GLM, GBM, GAM, and random forest) were then run five times (once for each pseudoabsence run), using in turn four of the five defined data splits to calibrate and one to evaluate based on TSS (threshold = 0.7).

Finally, a full ensemble combining all algorithms and pseudoabsences runs (51) was created, using only models with TSS > 0.7, averaged using four different statistics: mean, median, committee average and weighted mean. The statistic showing the highest TSS, the mean, was then used to predict the probability of occurrence in each grid cell. This was then projected for all available time slices from the present to 50 thousand years ago.

CISGeM Demography:

CISGeM's demographic module consists of a spatial model that simulates long-term and global growth and migration dynamics of Yellow Warblers. These processes depend on a number of parameters (see Table 1.), which we later estimate statistically based on empirical genetic data.

α	Allometric scaling factor for population size
β	Allometric scaling exponent for population size
t_0	Temporal origin of the population
r	Intrinsic growth rate
m_r	Non-directed (random) mobility parameter
m_d	Directed mobility parameter

Table 1. Details of parameters used in CISGeM.

The model operates on a global hexagonal grid of 40962 cells that represent the whole world (the distance between the centre of two hexagonal cells is 241 ±15 km); 2422 grid cells make up North America. Each time step represents 1 year, the generation time of Yellow Warblers. Each time step of a simulation begins with the computation of the carrying capacity of each grid cell, i.e. the maximum number of YWs theoretically able to live in the cell for the environmental resources at the given point in time. Here, we estimate the carrying capacity in a grid cell x at a time t as

$$K(x,t) \overset{\text{def}}{=} \begin{cases} \alpha \cdot p(x,t)^{\beta}, \text{ if } x \text{ is on land at time } t \\ 0, \text{else} \end{cases}$$

where $p(x,t)$ denotes the probability of a species inhabiting cell x at time t (see section ‘Species Distribution Modelling’). The particular function used here was chosen based on analysis of SDM projections and census data of Holarctic birds (R. Green, pers. comm.).

The estimated carrying capacities are used to simulate spatial population dynamics as follows. We begin a simulation by initialising a population of yellow warblers in a grid cell x_0 (represents the spatial origin of yellow warbler in our model) at a point in time t_0 with $K(x_0, t_0)$ individuals.

At each subsequent time step between $t_0$and the present, CISGeM simulates two processes: the local growth of populations within grid cells, and the spatial migration of individuals across cells. We used the logistic function to model local population growth, estimating the net number of individuals by which the population of size $N(x,t)$ in the a x at time t increases within the time step as

$$r \cdot N(x,t) \cdot \left(1 - \frac{N(x,t)}{K(x,t)}\right),$$

where r denotes the intrinsic growth rate. Thus, growth is approximately exponential at low population sizes, before decelerating, and eventually levelling off at the local carrying capacity.

Across-cell migration is modelled as two separate processes, representing a non-directed, spatially uniform movement into all neighbouring grid cells on the one hand, and a directed movement along a resource availability gradient on the other hand. Under the first movement type, the number of individuals migrating from a cell x into any one of the up to six neighbouring cells is estimated as

$$m_r \cdot N(x,t),$$

where m_r is a mobility parameter. This mechanism is equivalent to a spatially uniform diffusion process, which has previously been used to model random movement in other species (52). Under the second movement type, an additional number of individuals moving from a grid cell x_1 to a neighbouring cell x_2 is estimated as

$$m_d \cdot N(x_1,t) \cdot max\left(0, \frac{K(x_2,t) - N(x_2,t)}{K(x_2,t)} - \frac{K(x_1,t) - N(x_1,t)}{K(x_1,t)}\right)$$

The number $\frac{K(x,t)-N(x,t)}{K(x,t)}$ represents the relative availability of unused resources in the cell x at time t, equalling 1 if all natural resources in x are potentially available for yellow warblers ($N(x,t) = 0$), and 0 if all resources are used ($N(x,t) = K(x,t)$). Thus, individuals migrate in the direction of increasing relative resource availability, and the number of migrants is proportional to the steepness of the gradient. The distinction between directed and non-directed movement allows us to examine to which extent migration patterns can be explained by random motion alone or requires us to account for more complex responses to available resources.

For some values of the mobility parameters m_r and m_d, it is possible for the calculated number of migrants from a cell to exceed the number of individuals in that cell. In this scenario, the number of migrants into neighbouring cells are rescaled proportionally such that the total number of migrants from the cell is equal to the number of individuals present.

Similarly, it is in principle possible that the number of individuals present in a cell after all migrations are accounted for (i.e., the sum of local non-migrating individuals, minus outgoing migrants, plus incoming migrants from neighbouring cells) exceeds the local carrying capacity. In this case, incoming migrants are rescaled proportionally so that the final number of individuals in the cell is equal to the local carrying capacity. In other words, some incoming migrants perish before establishing themselves in the destination cell, and these unsuccessful migrants are not included in the model's output of migration fluxes between grid cells. In contrast, non-migrating local residents remain unaffected in this step. They are assumed to benefit from a residential advantage (53), and capable of outcompeting incoming migrants.

CISGeM's demographic module outputs the number of individuals in each grid cell, and the number of migrants between neighbouring grid cells, across all time steps of a simulation. These quantities are the used to reconstruct genetic lineages.

CISGeM predicted genetics

Once a global population demography has been constructed, gene trees are simulated. This process is dependent on the population dynamics recorded in the demography stage and assumes local random mating according to the Wright-Fisher dynamic. From the present, ancestral lines of sampled individuals are tracked back through the generations, recording which cell each line belongs to. Every generation, the lines are randomly assigned to a gamete from the individuals within its present cell. If the assigned individual is a migrant or coloniser, the line moves to the cell of origin for that individual before 'reproduction'. Whenever two lines are assigned to the same parental gamete, this is recorded as a coalescent event, and the two lines merge into a single line representing their common ancestor. This process is repeated until all the lineages have met, reaching the common ancestor of the whole sample. If multiple lineages are still present when the model reaches the generation and deme from which the demography was initialised, the lines enter a single ancestral population (K_0) until sufficient additional coalescent events have occurred for the gene tree to close.

ABC parameter estimation

Parameter space was explored with a Monte Carlo sweep in which demographic parameters were randomly sampled from flat prior ranges: directed expansion coefficient [0.0,0.14], undirected expansion coefficient [0.0,0.04], intrinsic growth rate [0.02,0.15], allometric scaling exponent [0.1,1], and allometric scaling factor [20,5000] on a $\log_{10}$ scale. A fixed mutation rate of $2.3\text{x}10^{-9}$ μ/Site/Year was used (54).

Model fit was initially calculated within an Approximate Bayesian Computation (abc) framework using the results of the Monte Carlo sweep. To compute summary statistics, populations were clustered into three groups representing the West, Central, and East regions of the North American continent, based on the NGSadmix outputs. The mean pairwise π for populations was then computed within each group and between each pair of groups, giving us a total of 6 summary statistics.

We performed parameter estimation with the R package '*abc*' (25) using a local linear abc algorithm, setting the tolerance to 0.025. For each of the simulations retained by the abc analysis, demographic simulations were then recorded and combined to create an average, representative, profile of the population's demographic history.

ABC model fitting: gfit & gfitpca

We also confirmed the quality of the model fit using formal hypothesis testing approaches from the R package '*abc*' (25). Firstly we used the '*gfit'* function (55) to confirm that our model outperformed a series on null models. In this function the goodness of fit test statistic, or D-statistic, is the median Euclidean distance between the observed summary statistics and the nearest (accepted) summary statistics. For comparison, a null distribution of D is then generated from summary statistics of 1000 pseudo-observed datasets. A goodness of fit p-value can then be calculated as the proportion of D based on pseudo-observed data sets that are larger than the empirical value of D. Consequently, a non-significant p-value signifies that the distance between the observed and accepted summary statistics is not larger than the expectation, confirming that the model fits the observed data well.

We then further performed an a priori goodness of fit test using the *'gfitpca'* function which captures and plots the two first components obtained with a principle component analysis. We used a '*cprob*' value of 0.1, 0.15, and 0.2, leaving a different proportion of points from the model outside the displayed envelope. The observed summary statistics is then marked to check that it is contained within these envelopes, indicating a good fit.

ABC model fitting: abcrf

We further evaluated model fit and posterior distributions with an abc random forest (RF) approach implemented via the R package '*abcrf*' (26,56). Forests of 1,000 trees were used.

Author contribution:

E.F.M. and A.M. devised the project. E.F.M. ran the genetic analysis under the supervision of P.M.D. and A.M.. M.L. ran the Species Distribution Modelling. M.K. and R.B. and A.M. developed the modelling software with help from E.F.M, P.M.D. and G.L.S.. M.S. helped with the interpretation of results. E.F.M and A.M. wrote the manuscript with feedback from all the co-authors.

Competing interests:

The authors declare no competing interests.

1. Kelly JF, Hutto RL. An East-West Comparison of Migration in North American Wood Warblers. Condor. 2005;107(2):197–211.

2. Lovette IJ, Clegg SM, Smith TB. Limited Utility of mtDNA Markers for Determining Connectivity among Breeding and Overwintering Locations in Three Neotropical Migrant Birds. Conserv Biol. 2004;18(1):156–66.

3. Peters JL, Gretes W, Omland KE. Late Pleistocene divergence between eastern and western populations of wood ducks (Aix sponsa) inferred by the "isolation with migration" coalescent method. Mol Ecol. 2005;14(11):3407–18.

4. Hewitt GM. Postglacial re-colonisation of European biota. Biol J Linn Soc. 1999;68(May):87–112.

5. Haffer J. Speciation in amazonian forest birds. Vol. 165, Science. 1969. p. 131–7.

6. Davis LA, Roalson EH, Cornell KL, Mcclanahan KD, Webster MS. Genetic divergence and migration patterns in a North American passerine bird: Implications for evolution and conservation. Mol Ecol. 2006;15(8):2141–52.

7. Colbeck GJ, Gibbs HL, Marra PP, Hobson K, Webster MS. Phylogeography of a widespread North American migratory songbird (Setophaga ruticilla). J Hered. 2008;99(5):453–63.

8. Bemmels JB, Dick CW. Genomic evidence of a widespread southern distribution during the Last Glacial Maximum for two eastern North American hickory species. J Biogeogr. 2018;45(8):1739–50.

9. Lumibao CY, Hoban SM, McLachlan J. Ice ages leave genetic diversity 'hotspots' in Europe but not in Eastern North America. Ecol Lett. 2017;20(11):1459–68.

10. Excoffier L, Foll M, Petit RJ. Genetic consequences of range expansions. Annu Rev Ecol Evol Syst. 2009;40:481–501.

11. Reid BN, Kass JM, Wollney S, Jensen EL, Russello MA, Viola EM, et al. Disentangling the genetic effects of refugial isolation and range expansion in a trans-continentally distributed species. Heredity (Edinb) [Internet]. 2019;122(4):441–57. Available from: http://dx.doi.org/10.1038/s41437-018-0135-5

12. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science (80-). 2015;349(6250).

13. Eriksson A, Betti L, Friend AD, Lycett SJ, Singarayer JS, von Cramon-Taubadel N, et al. Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc Natl Acad Sci [Internet]. 2012 Oct 2;109(40):16089–94. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1209494109

14. Bay R, Harrigan R, Underwood V Le, Gibbs HL, Smith TB, Ruegg K. Response to Comment on "Genomic signals of selection predict climate-driven population declines in a migratory bird" Science. 2018;361(August):2–4.

15. Skotte L, Korneliussen TS, Albrechtsen A. Estimating individual admixture proportions from next generation sequencing data. Genetics. 2013;195(3):693–702.

16. Boulèt M, Gibbs HL, Hobson KA. Integrated analysis of genetic, stable isotope, and banding data reveal migratory connectivity and flyways in the northern yellow warbler (Dendroica petechia; aestiva group). Ornithol Monogr. 2006;61(July 2015):29–78.

17. Bay RA, Karp DS, Saracco JF, Anderegg WRL, Frishkoff L, Wiedenfeld D, et al. Genetic variation reveals individual-level climate tracking across the full annual cycle of a migratory bird. bioRxiv [Internet]. 2020;(preprint). Available from: https://www.biorxiv.org/content/10.1101/2020.04.15.043331v1.abstract

18. Elith J, Leathwick JR. Species Distribution Models: Ecological Explanation and Prediction

Across Space and Time. Annu Rev Ecol Evol Syst. 2009;40(1):677–97.

19. McCullagh P, Nelder JA. Generalized Linear Models, 2nd Edn. Vol. 39. Chapman and Hall; 1990. 385 p.

20. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008;77(4):802–13.

21. Hastie TJ, Tibshirani. RJ. Generalized additive models. Chapman and Hall; 1990.

22. Breiman L. Random Forest. Mach Learn [Internet]. 2001 Oct 31;5–32. Available from: https://www.taylorfrancis.com/books/9780429890277/chapters/10.1201/9780429469275-8

23. Darvill CM, Menounos B, Goehring BM, Lian OB, Caffee MW. Retreat of the Western Cordilleran Ice Sheet Margin During the Last Deglaciation. Geophys Res Lett. 2018;45(18):9710–20.

24. Margold M, Stokes CR, Clark CD. Reconciling records of ice streaming and ice margin retreat to produce a palaeogeographic reconstruction of the deglaciation of the Laurentide Ice Sheet. Quat Sci Rev [Internet]. 2018;189:1–30. Available from: https://doi.org/10.1016/j.quascirev.2018.03.013

25. Csilléry K, François O, Blum MGB. Abc: An R package for approximate Bayesian computation (ABC). Methods Ecol Evol. 2012;3(3):475–9.

26. Pudlo P, Marin JM, Estoup A, Cornuet JM, Gautier M, Robert CP. Reliable ABC model choice via random forests. Bioinformatics. 2016;32(6):859–66.

27. Ruegg KC, Smith TB. Not as the crow flies: A historical explanation for circuitous migration in Swainson’s thrush (Catharus ustulatus). Proc R Soc B Biol Sci. 2002;269(1498):1375–81.

28. Markova S, Hornikova M, Lanier HC, Henttonen H, Searle JB, Weider LJ, et al. High genomic diversity in the bank vole at the northern apex of a range expansion: the role of multiple

colonizations and end-glacial refugia. Mol Ecol [Internet]. 2020;0–3. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25688

29. Cheviron ZA, Hackett SJ, Capparella AP. Complex evolutionary history of a Neotropical lowland forest bird (Lepidothrix coronata) and its implications for historical hypotheses of the origin of Neotropical avian diversity. Mol Phylogenet Evol. 2005;36(2):338–57.

30. Pearson RG, Dawson TP. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob Ecol Biogeogr [Internet]. 2003 Sep;12(5):361–71. Available from: http://doi.wiley.com/10.1046/j.1466-822X.2003.00042.x

31. Haller BC, Messer PW. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol Biol Evol. 2019;36(3):632–7.

32. Zink RM. Comparative phylogeography in North American birds. Evolution (N Y). 1996;50(1):308–317.

33. Bay RA, Harrigan RJ, Underwood V Le, Gibbs HL, Smith TB, Ruegg K. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science (80-). 2018 Jan 5;359(6371):83–6.

34. Mazerolle DF, Dufour KW, Hobson KA, Haan HE Den. Effects of large-scale climatic fluctuations on survival and production of young in a Neotropical migrant songbird, the yellow warbler Dendroica petechia. J Avian Biol [Internet]. 2005 Feb;36(2):155–63. Available from: https://doi.org/10.1111/j.0908-8857.2005.03289.x

35. Warmuth VM, Ellegren H. Genotype-free estimation of allele frequencies reduces bias and improves demographic inference from RADSeq data. Mol Ecol Resour. 2019 May 17;19(3):586–96.

36. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One.

2012;7(7).

37. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics. 2014 Dec 25;15(1):356.

38. Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLoS Comput Biol. 2016;12(5):1–22.

39. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. 2019. Available from: http://www.r-project.org/

40. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25(15):1965–78.

41. Beyer RM, Krapp M, Manica A. High-resolution terrestrial climate, bioclimate and vegetation for the last 120,000 years. Sci Data. 2020;7(1):1–9.

42. Singarayer JS, Valdes PJ. High-latitude climate sensitivity to ice-sheet forcing over the last 120 kyr. Quat Sci Rev [Internet]. 2010;29(1–2):43–55. Available from: http://dx.doi.org/10.1016/j.quascirev.2009.10.011

43. Armstrong E, Hopcroft PO, Valdes PJ. Reassessing the Value of Regional Climate Modeling Using Paleoclimate Simulations. Geophys Res Lett. 2019;46(21):12464–75.

44. Beyer R, Krapp M, Manica A. An empirical evaluation of bias correction methods for palaeoclimate simulations. Clim Past. 2020;16(4):1493–508.

45. BirdLife International and Handbook of the Birds of the World. Bird species distribution maps of the world. Version 2018.1. 2018.

46. Aiello-Lammens ME, Boria RA, Radosavljevic A, Vilela B, Anderson RP. spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography (Cop). 2015;38(5):541–5.

47. Thuiller W, Georges D, Engler R, Breiner F. biomod2: Ensemble Platform for Species Distribution Modeling [Internet]. 2019. Available from: https://cran.r-project.org/package=biomod2

48. Miller EF, Green RE, Balmford A, Beyer R, Somveille M, Leonard M, et al. mtDNA-based reconstructions of change in effective population sizes of Holarctic birds do not agree with their reconstructed range sizes based on paleoclimates. bioRxiv. 2019;

49. Bagchi R, Crosby M, Huntley B, Hole DG, Butchart SHM, Collingham Y, et al. Evaluating the effectiveness of conservation site networks under climate change: Accounting for uncertainty. Glob Chang Biol. 2013;19(4):1236–48.

50. Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol. 2019;10(2):225–32.

51. Araújo MB, New M. Ensemble forecasting of species distributions. Trends Ecol Evol. 2007;22(1):42–7.

52. Kot M. Elements of Mathematical Ecology [Internet]. Cambridge University Press; 2001. 453 p. Available from: https://books.google.co.uk/books?id=Zh3GNd9M1oUC&printsec=frontcover&source=gbs_ge_summary_r&cad=0

53. Pérez-Tris J, Tellería JL. Migratory and sedentary blackcaps in sympatric non-breeding grounds: Implications for the evolution of avian migration. J Anim Ecol. 2002;71(2):211–24.

54. Smeds L, Qvarnström A, Ellegren H. Direct estimate of the rate of germline mutation in a bird. Genome Res. 2016;26(9):1211–8.

55. Lemaire L, Jay F, Lee I-H, Csilléry K, Blum MGB. Goodness-of-fit statistics for approximate Bayesian computation. 2016;1–30. Available from: http://arxiv.org/abs/1601.04096

56. Raynal L, Marin JM, Pudlo P, Ribatet M, Robert CP, Estoup A. ABC random forests for Bayesian parameter inference. Bioinformatics. 2019;35(10):1720–8.