Identifying genomic adaptation to local climate using a mechanistic evolutionary model

Nikunj Goel[1*], Christen M. Bossu[2], Justin J. Van Ee[3], Erika Zavaleta[4], Kristen C. Ruegg[2], and Mevin B. Hooten[1]

[1]Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, Texas, USA, 78705

[2]Department of Biology, Colorado State University, Fort Collins, Colorado, USA, 80523

[3]Department of Statistics, Colorado State University, Fort Collins, Colorado, USA, 80523

[4]Ecology and Evolutionary Biology Department, University of California, Santa Cruz, California, USA, 95060

[*]**Corresponding author; email:** nikunj.goel@utexas.edu and nikunj410@gmail.com

Abstract

1. Identifying genomic adaptation is key to understanding species' evolutionary responses to climate change. However, current methods to identify adaptive variation have two major limitations. First, when estimating genetic variation, most methods do not account for observational uncertainty in genetic data because of finite sampling and missing genotypes. Second, current methods use phenomenological models to partition genetic variation into adaptive and non-adaptive components. These phenomenological models are not mechanistic models of evolution and, therefore, do not faithfully capture the demographic history of the species.

2. We address these limitations by developing a hierarchical Bayesian model that explicitly accounts for both the observational uncertainty and underlying evolutionary processes. The first layer of the hierarchy is the data model that captures observational uncertainty by probabilistically linking RAD-sequence data to genetic variation. The second layer is a process model that provides a mechanistic explanation of how evolutionary forces, such as local adaptation, mutation, migration, and drift, maintain genetic variation. The third layer is the parameter model, which incorporates our knowledge about biological processes. For example, because most loci in the genome are expected to be neutral, the environmental sensitivity coefficients are assigned a regularized prior centered at zero. Together, the three models provide a rigorous probabilistic framework to identify local adaptation in wild organisms.

3. Analysis of simulated RAD-seq data shows that our statistical model can reliably infer adaptive genetic variation. To show the real-world applicability of our method, we re-analyzed RAD-seq data from Willow Flycatchers (*Empidonax traillii*) in the USA. We found 30 genes close to loci that showed a statistically significant association with temperature seasonality. Gene ontology suggests that several of these genes play a crucial role in egg mineralization, feather development, and the ability to withstand extreme temperatures.

4. Moreover, biogeographers can easily modify the data and process models to accommodate a wide range of genetic datasets (*e.g.*, pool and low coverage genome sequencing) and demographic histories (*e.g.,* range shifts), allowing them to construct statistical models specific to their study system.

Introduction

Many species face a looming threat of extinction as global temperature rises (Thomas *et al.* 2004; Urban 2015). This risk is particularly concerning for species with limited dispersal capacity that are slow to track their climatic niches (Carroll *et al.* 2015). Alternatively, some species may adapt to the local climate, potentially alleviating the risk of extinction. Natural selection favors individuals with heritable phenotypes best suited to survive and reproduce in local climates, creating a geographical mosaic in frequencies of adaptive alleles congruent with climatic gradients (Hedrick, Ginevan & Ewing 1976; Hedrick 1986; Hedrick 2006). Therefore, to understand the evolutionary responses of a species to changing climate and incorporate this knowledge to inform evolutionary management policies (Smith *et al.* 2014), we need robust methodologies to identify genes (and their function) that confer local adaptation in natural populations.

With recent advances in high-throughput sequencing and global environmental sensing technologies (Chuvieco 2020; Satam *et al.* 2023), new computational approaches have allowed researchers to identify adaptive genetic variation in wild populations. These approaches, collectively called environmental association analysis (Rellstab *et al.* 2015; Hoban *et al.* 2016), have gained widespread popularity that can be partly attributed to declining sequencing costs, improvements in genomic tools, detailed global climatic maps, and availability of computational resources. These advances have allowed researchers to study climatic adaptation in non-model organisms at a fine genomic resolution and large spatial extent for a reasonable cost (Frichot *et al.* 2013; De Villemereuil & Gaggiotti 2015; Fitzpatrick & Keller 2015; Wagner, Chávez-Pesqueira & Forester 2017).

Broadly, environmental association analysis involves three steps. First, sequencing data from spatially referenced individuals are used to estimate spatial variation in allele frequencies. Next, a statistical model is used to partition this genetic variation into adaptive and non-adaptive components. The adaptive variation corresponds to the response of an allele to changes in prevailing climatic conditions. The non-adaptive variation describes the allelic variation (or null variation) that stems from evolutionary forces other than local adaptation, which includes background selection, gene flow among populations, mutation, and drift. Finally, a statistical criterion is used to select loci that exhibit strong adaptive genetic variation relative to non-adaptive variation.

Despite the success of environmental association studies, current methods face several conceptual and statistical challenges that limit the reliability of the inferences. Most environmental association methods do not account for uncertainty in estimates of allele frequencies (see Coop *et al.* (2010) and Foll and Gaggiotti (2008) for exceptions), which can introduce biases in downstream analyses. Due to practical constraints, biogeographers collect and sequence only a finite number of individuals in a population. The genotype of these individuals, inferred from RAD (restriction site-associated DNA) sequencing (Baird *et* *al.* 2008), for instance, contains imprecise information regarding the underlying allele frequencies that result from sampling a small subset of individuals within a population. Additionally, because of low coverage and alignment errors, the true genotypes of some individuals may be missing (Huang & Knowles 2016), which further reduces sample size and increases uncertainty. Failure to account for these sources of uncertainties results in overly precise estimates of genetic variation that can increase the number of false positives (inferring non-adaptive loci as important). Alternatively, researchers discarding data due to noisy estimates of genetic variation may lose valuable (although imprecise) information, thereby increasing rates of false negatives (inability to identify an adaptive locus).

Another major difficulty is that environmental association methods characterize genetic variation using phenomenological models, such as generalized linear models (Rellstab *et al.* 2015). Although reasonable, phenomenological models are not mechanistic models of evolution and, therefore, they are not usually concerned with how the relationship between data and parameters (e.g., environment and allele frequency) arises because of underlying evolutionary processes (Hilborn & Mangel 2013; Hobbs & Hooten 2015). For example, some phenomenological models assume an S-shape response curve to relate environmental variables and allele frequency (Joost *et al.* 2007; Stucki *et al.* 2017). This response curve has several attractive characteristics: it is bounded between zero and one, and has an incline in the middle that emulates the response of an allele to local climatic adaptation. However, the S-shape response curve is not unique in these characteristics. A biogeographer can construct alternate equally plausible response curves (*e.g.*, linear or probit functions). Similarly, phenomenological models make many other assumptions, including constructing an appropriate null distribution, that are often difficult to rationalize. This makes it hard to evaluate when a phenomenological model will fail and, when it inevitably does, how the model can be improved.

In contrast, mechanistic models of evolution are based on the first principles of birth-death processes, and, therefore, a biogeographer can evaluate model assumptions based on the constraints imposed by the population demography and natural history of the species (Rice 2004). Replacing phenomenological models with mechanistic models of evolution in statistical inferences offers a unique advantage—they constrain the flow of information from data to model parameters using theoretical rather than heuristic arguments. This may allow biogeographers to create bespoke statistical models tailored to match the demographic history of the species (Wikle 2003; Hooten & Hefley 2019), providing robust and reliable inferences.

In this paper, we address the aforementioned limitation of environmental association analysis using hierarchical Bayesian models. Hierarchical models allow us to learn parameters using data, process, and parameter models (Berliner 1996). In the context of environmental association analysis, the data model describes how the unobserved allele frequencies could have led to genetic data. The process model describes the biological processes determining spatial variation in allele frequencies. The parameter model describes the knowledge the biogeographer has about the parameters before the data are collected based on past research.

The structure of the hierarchical model offers several advantages in identifying genomic adaptation to climate. First, the data model allows a biogeographer to quantify observational uncertainty in estimates of genetic variation due to the small sample size and missing genotypes. Second, these imprecise estimates of genetic variation can then be linked to a process model of evolution. This evolution model is based on the demographic history of the species and provides a mechanistic explanation of how to partition genetic variation into adaptive (response curve) and non-adaptive (null) components. Lastly, the parameter model can be used to incorporate knowledge from past research. For example, the theory of molecular evolution suggests that the overwhelming majority of variation in genomes is non-adaptive and stems from the interplay between mutation, drift, and migration, while only a small fraction of the variation stems from local adaptation (Kimura 1983). To incorporate this prior knowledge, a biogeographer can assign a low prior probability that the response curve has a finite slope. This prior effectively shrinks (regularizes) the adaptive genetic variation to zero for most loci, providing a systematic way to evaluate the relative contribution of adaptive and non-adaptive evolutionary forces in determining genetic variation.

Because of these features of the data, process, and parameter models, hierarchical models provide a rigorous probabilistic framework to identify genomic adaptation in wild organisms informed by theoretical principles of evolutionary biology using noisy genetic data. To show the utility of this approach, we develop a demographic Bayesian model and test its robustness by analyzing synthetic data with known parameter values. Next, we use the model to analyze RAD-seq data from Willow Flycatchers (*Empidonax traillii*) and identify candidate genes that may play a functional role in climatic adaptation. Although we apply our model to Willow Flycatchers, the statistical insights are general and broadly applicable to other species.

Willow Flycatchers

The Willow Flycatcher is a migratory songbird that breeds mainly in the United States and Southern Canada. The species is categorized into four distinct subspecies—the Pacific Northwestern form (*E. t. brewsteri*), Western Central form (*E. t. adastus*), Eastern form (*E. t. traillii*), and Southwestern form (*E. t. extimus*). Among them, the Southwestern form has experienced a precipitous decline in abundance, likely due to the loss of riparian habitat along streams and waterways (Sedgwick 2000), which provide respite during extreme temperatures. In addition, in 1995, the Southwestern form was federally declared endangered (Unitt 1987) due to its genetic, ecological, and song distinctiveness (Theimer *et al.* 2016; Ruegg *et al.* 2018; Mahoney *et al.* 2020).

Previous landscape genomic work identified highly significant correlations between allele frequencies in genes linked to thermal tolerance and the intensity of summer heat waves in the southwest (Ruegg *et al.* 2018). Therefore, re-analyzing the Willow Flycatcher genome-wide genetic dataset, initially examined using traditional phenomenological models, offers an ideal opportunity to apply a new environmental association analysis that mechanistically accounts for genetic variation. The resulting analysis provides a strong basis for comparison with previous approaches, and more broadly, the results have important implications for implementing management practices designed to improve the genetic health of the endangered subspecies.

Data

A total of 175 Willow flycatchers were sampled at 23 one-degree squares across the continental United States (Fig. 1). The sampling effort varied from 2-21 individuals per location. DNA was extracted from blood and tissue samples using the QiagenTM DNeasy Blood and Tissue extraction kit and quantified using the Qubit® dsDNA HS Assay kit (Thermo Fisher Scientific). Sequencing was conducted across three lanes of 100 bp paired-end reads on an Illumina HiSeq 2500 at the UC Davis Genome Center. We filtered SNPs using the tradeoff between discarding SNPs with low coverage and discarding individuals with missing genotypes using the R package genoscapeRtools (Anderson 2019), resulting in approximately 105,000 SNPs, of which 3 percent of the loci had a missing genotype (Ruegg *et al.* 2018). At each sampling location, climate data were obtained from WorldClim, which averaged the climate between 1960 and 1990 (Hijmans *et al.* 2005). Due to the high correlation between some of the top-ranked climatic variables identified in Ruegg *et al.* (2018), we limited our analysis to the four least correlated climate variables (standardized) to reduce collinearity: BIO 4 (temperature seasonality), BIO 5 (maximum temperature of the warmest month), BIO 11 (mean temperature of the coldest quarter), and BIO 17 (Precipitation in the driest quarter). For more details about the sampling design, sequencing, and bioinformatic analysis, refer to Ruegg *et al.* (2018).

Model Formulation

To identify putative loci under selection due to local adaptation, we specify a Bayesian model with three hierarchical levels corresponding to the data, process, and parameter models. These three levels are organized such that the output of the parameter model is the input for the process model, whose output is the input for the observer model (Pagel & Schurr 2012). To facilitate a broad understating of how the statistical model works, in Figure 2, we illustrate the three levels of the hierarchical model and show how these levels are connected. In the following section, we provide details about the assumptions associated with our model.

Data model

We assume that Willow Flycatchers were sampled at K sites across the United States, and each bird was genotyped at L genetic (RAD) markers (see sampling and sequencing above). We characterized the genotype of the ith sampled bird at site k using g_{ilk}, which corresponds to the number of copies of the reference allele at locus l. Assuming individuals have biallelic loci, g_{ilk} is a discrete variable that takes values zero, one, or two. Using these individual genotype data, we define a population-level genotype variable

$$y_{lk} = \sum_{i=1}^{N_{lk}} g_{ilk}, \tag{1}$$

that corresponds to the number of reference alleles for N_{lk} birds that were genotyped at locus l. In the data, N_{lk} is always less than or equal to the number of birds sampled at a site because, for some individuals, the genotype at a locus may be missing due to sequencing errors or low coverage. Assuming the birds were randomly sampled, we model the variation in reference allele frequency (p_{lk}) using a binomial distribution,

$$y_{lk} \sim \text{Binomial}(2N_{lk}, p_{lk}). \tag{2}$$

Alternatively, we can also use

$$g_{ilk} \sim \text{Binomial}(2, p_{lk}), \tag{3}$$

as a data model to describe the statistical relationship between allele frequency and an individual's genotype. However, because of computational efficiency, we use equation (2). Note that the observer model accounts for the uncertainty in genetic variation that stems from finite sampling and missing genotypes.

Process model

We consider a metapopulation model proposed by Wright (1931) to model evolutionary dynamics. We rely on this model because it provides a parsimonious explanation of how evolutionary processes maintain genetic variation. We assume that the demes in the metapopulation correspond to the sites where the birds were sampled. Each deme has a population size of N_e, and the migration rate between any pair of demes is equal. We assume that the variation in allele frequencies across demes is maintained by directional and non-directional evolutionary forces. Directional forces—such as local adaptation to climate, mutation, and migration—result in changes in the mean value of allele frequency. Non-directional forces—such as genetic drift—do not change the mean value of allele frequency but create sampling variance due to finite population size. Because of this variance, the frequency of an allele cannot be determined exactly. Instead, allele frequency is characterized probabilistically using a probability distribution, $\psi(p_{lk}, t)$ (Rice 2004; Blanquart, Gandon & Nuismer 2012). Using the Fokker-Planck equation, one can show that the probability distribution of the frequency of the reference allele changes as follows:

$$\frac{\partial \psi(p_{lk}, t)}{\partial t} = -\frac{\partial}{\partial p_{lk}}[\psi(p_{lk}, t)M(p_{lk})] + \frac{1}{2}\frac{\partial^2}{\partial^2 p_{lk}}[\psi(p_{lk}, t)V(p_{lk})], \tag{4}$$

where

$$M(p_{lk}) = \underbrace{s_{lk}p_{lk}q_{lk}}_{\text{Selection}} + \underbrace{u_{2l}q_{lk} - u_{1l}p_{lk}}_{\text{Mutation}} + \underbrace{m(\overline{p_l} - p_{lk})}_{\text{Migration}} \tag{5}$$

is the rate of directional change in the allele frequency and

$$V(p_{lk}) = \frac{p_{lk}q_{lk}}{2N_e} \tag{6}$$

is the variance in the allele frequency due to non-directional (*i.e.*, drift) effects. In equations (4)-(6), $q_{lk}(= 1 - p_{lk})$ is the frequency of the alternate allele, s_{lk} is the environmentally regulated selection coefficient, u_{2l} and u_{1l} are the locus-specific forward and backward mutation rates, m is the migration rate, and $\overline{p_l}$ is the average allele frequency of the immigrants. We consider multiplicative selection dynamics: The relative fitness of individuals with genotype (g_{ilk}) zero, one, and two is $1 + 2s_{lk}$, $1 + s_{lk}$, and 1, respectively.

At the time of sampling, the allele frequency distribution is assumed to be at or close to equilibrium. To obtain this equilibrium distribution, also known as the stationary distribution (ψ_s), we set $\partial\psi(p_{lk}, t)/\partial t = 0$. The stationary distribution of the allele frequency can be expressed in terms of evolutionary parameters as follows (Blanquart, Gandon & Nuismer 2012):

$$\psi_s(p_{lk}|\tilde{s}_{lk}, \mu_l, \kappa_l) = \frac{e^{\tilde{s}_{lk}p_{lk}}}{{}_1F_1(\mu_l\kappa_l, \kappa_l, \tilde{s}_{lk})}\text{Beta}(p_{lk}|\, \mu_l, \kappa_l), \tag{7}$$

where $\tilde{s}_{lk} = 4N_e s_{lk}$, ${}_1F_1$ is the hypergeometric confluent function, and $\mu_l = (u_{2l} + m\overline{p_l})/(u_{1l} + u_{2l} + m)$ and $\kappa_l = 4N_e(u_{1l} + u_{2l} + m)$ are the mean and precision parameters of the beta distribution. To incorporate climate adaptation, we assume a linear relationship between the selection coefficient around E standardized environmental variables:

$$\tilde{s}_{lk} = \alpha_l + \sum_{j=1}^{E} \beta_{lj}e_{jk}, \tag{8}$$

where α_l is the background selection coefficient, e_{jk} is the jth environmental variable (standardized) in deme k, and β_{lj} is the selection coefficient's sensitivity to variation in the environment.

The stationary distribution in equation (7) has several notable features. The response curve of an adaptive allele (relationship between climate and mean allele frequency),

$$\mathbb{E}[p_{lk}] = \mu_l \frac{{}_1F_1(1 + \mu_l\kappa_l, 1 + \kappa_l, \tilde{s}_{lk})}{{}_1F_1(\mu_l\kappa_l, \kappa_l, \tilde{s}_{lk})}, \tag{9}$$

is bounded between zero and one, and its shape is determined by evolutionary parameters with biological interpretation. When selection is absent, the beta distribution captures the non-adaptive genetic variation,

$$\psi_s(p_{lk}|0, \mu_l, \kappa_l) = \text{Beta}(p_{lk}|\, \mu_l, \kappa_l). \tag{10}$$

Therefore, to account for the joint contribution of adaptive and non-adaptive evolutionary forces in determining genetic variation, we use the stationary distribution to statistically model evolutionary dynamics,

$$p_{lk} \sim \frac{e^{\tilde{s}_{lk} p_{lk}}}{{}_1F_1(\mu_l \kappa_l, \kappa_l, \tilde{s}_{lk})} \mathrm{Beta}(\mu_l, \kappa_l). \tag{11}$$

For fast and memory-efficient implementation of the statistical model, we combine the data (Eq. 2) and process (Eq. 11) models by marginalizing over p_{lk},

$$y_{lk} \sim \int_0^1 \mathrm{Binomial}(y_{lk} | 2N_{lk}, p_{lk}) \psi_s(p_{lk} | \tilde{s}_{lk}, \mu_l, \kappa_l)\, dp_{lk}, \tag{12}$$

which results in an integrated likelihood that relates genotype counts to evolutionary dynamics as follows:

$$y_{lk} \sim \frac{{}_1F_1(\mu_l \kappa_l + y_{lk}, 2N_{lk} + \kappa_l, \tilde{s}_{lk})}{{}_1F_1(\mu_l \kappa_l, \kappa_l, \tilde{s}_{lk})} \mathrm{BetaBinomial}(2N_{lk}, \mu_l, \kappa_l). \tag{13}$$

Note that the integrated likelihood in equation (13) accounts for the uncertainty that stems from noisy genetic data and the stochastic nature of evolutionary dynamics.

Parameter model

Finally, we assign priors to parameters in equation (13) based on our prior scientific knowledge. We assign the mean (μ_l) and precision (κ_l) parameters of the beta distribution $\mathrm{Uniform}(0{,}1)$ and $\mathrm{Normal}^{+}(0, 5)$ priors, respectively, which reflect the natural bounds on these parameters. Due to degeneracy in the geometry of the likelihood (Eq. 13, see Fig. S2), the mean of the beta distribution (μ_l) and selection coefficient ($\tilde{s}_{lk}$) are weakly identifiable. We alleviate this by making two biologically reasonable assumptions. First, we fix $\alpha_l = 0$ and re-interpret μ_l as the mean of baseline allele frequency distribution (or null distribution) resulting from non-adaptive evolutionary forces, including background selection. Second, because we expect that most of the genetic variation arises due to neutral processes (Kimura 1983), we use a regularized horseshoe shrinkage parameter model (Piironen & Vehtari 2017) for sensitivity coefficients, β_{lj}.

The horseshoe prior shrinks most of the β_{lj} to zero, allowing a fraction of coefficients to take non-zero values. The horseshoe model achieves this using global (τ) and local ($\tilde{\lambda}_{lj}$) shrinkage parameters. The cumulative effect of these shrinkage parameters is captured by

$$\beta_{lj} \sim \mathrm{Normal}\left(0\,, \tau^2 \tilde{\lambda}_{lj}^2\right), \tag{14}$$

where

$$\tilde{\lambda}_{lj}^2 = \frac{c^2 \lambda_{lj}^2}{c^2 + \tau^2 \lambda_{lj}^2}. \tag{15}$$

To control global shrinkage, we specify a half-Cauchy prior,

$$\tau \sim \text{Cauchy}^{+}\left(0, \frac{f^2}{KE^2}\right), \tag{16}$$

where f is our prior knowledge of the fraction of the RAD sites contributing to local adaptation. Because f is assumed to be small (we use $f = 20L^{-1}$), τ effectively shrinks all sensitivity coefficients to zero. But, the local shrinkage parameter, $\tilde{\lambda}_{lj}$, allows some coefficients to escape shrinkage because of a heavy-tail prior, $\lambda_{lj} \sim \text{Cauchy}^{+}(0,1)$. However, these large-valued coefficients, too, are weakly shrunk and bounded between $\pm 3c$, where $c^2 \sim \text{InvGamma}(2,4)$. This feature of the local shrinkage prevents numerical pathologies when large effect loci are weakly identifiable.

Model Testing

To assess the robustness of the statistical method, we fit the model to synthetic data generated by simulations that share some characteristics of real data. Our simulation study provides evidence that we can obtain reliable inferences for (a) genomic data with sequencing and sampling characteristic that aligns with what we found when analyzing Willow Flycatcher genomic data and (b) when the assumed generative process in the statistical model only partially resembles the true generative process of the real data.

We generated the synthetic genotype data in three stages (see Supplementary code for details). In the first stage, we simulated selection dynamics. We considered diploid individuals with a genome size of one thousand loci. Each locus in the genome was assigned a random non-zero background selection coefficient. We randomly selected fifteen loci that contribute to local adaptation. Each of the fifteen loci was randomly paired with one of the four environmental variables, and the corresponding selection coefficient (s_{lk}) was calculated using equation (8). We used real environmental values (standardized) to calculate s_{lk} to preserve the correlation structure between environmental variables.

In the second stage, we simulated metapopulation dynamics with demes equal to the number of sampling sites in the real data. We sampled random variables from the stationary distribution of allele frequencies by simulating the following stochastic differential equation (Korolev *et al.* 2010):

$$dp_{lk} = \left[\underbrace{s_{lk}p_{lk}q_{lk}}_{\text{Selection}} + \underbrace{u_{2l}q_{lk} - u_{1l}p_{lk}}_{\text{Mutation}} + \sum_{n=1}^{K} \underbrace{e^{-\rho d_{nk}}(p_{ln} - p_{lk})}_{\text{Migration}}\right] dt + \underbrace{\sqrt{p_{lk}q_{lk}/2N_e}}_{\text{Drift}}\, dB, \tag{17}$$

where s_{lk} is the selection coefficient obtained from stage one, B is the standard Brownian motion, d_{nk} is the distance between demes n and k, and ρ^{-1} is the dispersal length scale. The above equation is an Itô representation of the Fokker-Planck equation presented earlier (Eq. 4), with a small modification: migration rate depends on the distance between demes (Fig. 3A).

In the third stage, we simulated genotypes. In each deme, we simulated genotypes for the same number of individuals as in the real data. For each individual, we generated a genotype at a locus by sampling from the distribution Binomial$(2, p_{lk})$ (Eq. 3), where p_{lk} is the reference allele frequency obtained from the second stage. We randomly selected three percent of the loci and treated them as missing genotypes.

To test if the statistical model can identify loci that contribute to local adaptation in synthetic data, we fit the model using Stan programming language (Carpenter *et al.* 2017) to obtain posterior samples of sensitivity coefficients, β_{lj}, using Hamiltonian Monte Carlo (Neal 2011). For each pair of loci and environmental variables, we computed the probability that the posterior distribution of the sensitivity coefficient included zero. If this probability is less than a threshold value of 0.05 (p_{th}), we inferred that the locus may have contributed to local adaptation for the corresponding environmental variable.

The Manhattan plot in Fig. 3A shows the negative log probability that the posterior distribution of sensitivity coefficients included zero (Wang *et al.* 2022). Our model correctly identified nine of the fifteen loci that were assigned large finite sensitivity coefficients in the synthetic data (points above the black line with a ring and solid center). These points have $p_{th} < 0.05$, which on a negative log scale corresponds to points above $y = 2.99$ line. However, the model incorrectly identified the corresponding environmental variable for two of the nine loci (points above the black line with a green ring and purple center). This feature can be explained by a negative correlation (-0.72) between environmental variables, BIO 4 (green) and BIO 11 (purple). As a result, the statistical model used one of the environmental variables as a substitute for another (Rellstab *et al.* 2015). We also found six false negatives (colored points below the black line), but no false positives. These results suggest that our method provides reasonable inferences, even when the assumptions we make to construct the statistical model differ from the generative assumptions of the synthetic data.

Next, we re-analyzed the synthetic data using latent factor mixed model (LFMM; Frichot *et al.* 2013) which was used to analyze Willow Flycatcher data in Ruegg *et al.* (2018). We conducted simulations using two versions of the synthetic dataset. In the first version, we used individual genotypes, and, in the second version, we used raw allele frequencies (*i.e.*, $p_{lk} = 0.5\, y_{lk}/N_{lk}$) as input. In the first version, the LFMM approach identified three of the fifteen loci adaptive loci and had one false positive (Fig. S4A). In the second version, the LFMM approach identified only one correct locus and had no false positives (Fig. S4B).

Data Analysis

Analyzing genomic data from Willow Flycatchers revealed 47 significant loci, all except one were associated with BIO 4 (Fig. 2 and Fig. S3). A closer inspection of these loci suggests that many of them cluster together on the genome (indicated by vertical gray lines in Fig. 2). These clustering patterns are unlikely to happen by random chance. Alternatively, the observed clustering patterns can be explained by gene hitchhiking (Smith & Haigh 1974). The allele frequencies at a locus under selection and its neighboring neutral loci rise or fall in unison due to physical linkage on the chromosome. Consequently, when a gene is under selection, polymorphic sites close to the gene experience pseudo-selection (Barton 1998).

Using the annotated genome of Willow Flycatchers, 36 of the candidate loci were found within or nearby (within 25kb) 30 named genes that may play a functional role in climatic adaptation (Table S1). Twenty-three of these genes are characterized, and 20 have functional roles in chicken (*Gallus gallus*) that span 8 gene ontology categories. The majority of genes cluster in 4 categories: 5 genes are involved in catalytic activity, 5 genes have binding functionality, 3 have transcription regulatory activity, and 3 have transporter activity. A closer investigation into several genes shows that EDIL3 plays a role in the egg mineralization process in Aves (Le Roy *et al.* 2021), PCDH1 is involved in feather development (Lin, Wang & Redies 2013) and GRIK2 acts as a thermoreceptor conferring sensitivity to cold temperatures in mice (Cai *et al.* 2024).

Discussion

Adaptation to climate is pervasive and will continue to play a major role in maintaining biodiversity in the Anthropocene (Thompson 2013). However, current methods aimed at identifying genomic adaptation are limited because they are often unequipped to handle noisy genetic data and do not formally accommodate the demographic history of the species. Our mechanistic Bayesian model addresses these limitations. The important aspects of our approach comprise the following features.

First, we proposed a data model that probabilistically links RAD-seq data to genetic variation (Eq. 2). This probabilistic link allows us to quantify uncertainty in genetic variation that arises due to finite sampling and missing genotypes. Consequently, the data model properly accounts for uncertainty in the estimation of genetic variation and faithfully propagates available information in raw genetic data to the process model for downstream analysis (Hobbs & Hooten 2015).

Second, we used a metapopulation process model to partition estimated genetic variation into adaptive and non-adaptive components (Blanquart, Gandon & Nuismer 2012). The metapopulation model can be implemented using a beta distribution to characterize non-adaptive variation (Eq. 10) and hypergeometric confluent functions to model adaptive variation (Eq. 9). Because our process model is constructed based on theoretical principles from evolutionary biology, we can evaluate it based on the underlying assumptions.

Finally, previous work shows that most of the genetic variation in wild populations is maintained by non-adaptive evolutionary forces (Kimura 1983). We incorporated this prior knowledge in our statistical model using a regularized horseshoe parameter model that shrinks most of the selection coefficients to zero by assigning a global shrinkage prior to sensitivity coefficients (Piironen & Vehtari 2017) (Eq. 14). The level of shrinkage is controlled by the number of sites where the birds were sampled, the number of environmental variables, and our prior understanding of what fraction of the genome contributes to local adaptation (Eq. 16). Thus, the parameter model provides a systematic way to evaluate the relative magnitude of adaptive and non-adaptive evolutionary forces in shaping genetic variation.

To test our statistical model, we conducted simulations to assess the model performance on synthetic RAD-seq data generated by simulating bird genomes of length 1000 loci (Eq. 17) that emulate the characteristics of genetic data from Willow Flycatchers. Our statistical model accurately identified nine out of fifteen adaptive loci with no false positives (Fig. 3A). Out of the nine correctly identified loci, two loci were paired incorrectly with their corresponding environmental variable due to a strong correlation between BIO 4 and BIO 11. This highlights that biogeographers may need to exercise caution when interpreting statistical results or use uncorrelated predictors. The statistically inferred environmental variable might be correlated with the true environmental variable responsible for local adaptation (Rellstab *et al.* 2015). Nevertheless, the synthetic simulations suggest that our statistical model is applicable to use with genetic datasets from wild populations.

Indeed, we identified 30 genes that were within the 25kb region flanking 47 significant loci in RAD-seq data from Willow Flycatchers, most of which were associated with temperature seasonality (Fig. 3B and S3). Some of these genes include EDIL3, PCDH1, and GRIK2, which play a role in the egg mineralization process (Le Roy *et al.* 2021), feather development (Lin, Wang & Redies 2013), and acts as a thermoreceptor conferring sensitivity to cold temperatures (Cai *et al.* 2024), respectively. This suggests that temperature fluctuations could be a key driver of local adaptation in the species, providing evidence that standing genetic variation in Willow Flycatchers could alleviate or buffer extinction risk due to increasing temperature variability predicted by climate projections (Olonscheck *et al.* 2021).

Although these genes differ from those identified in Ruegg *et al.* (2018), these differences are not surprising because our hierarchical statistical model is fundamentally different from traditional environmental association analysis in terms of quantifying uncertainty while estimating genetic variation and using an evolutionary process model to explain sources of genetic variation. Our synthetic data simulations confirm that these differences may play an important role in inferences; re-analyzing synthetic data using LFMM resulted in much higher rates of false positives and false negatives (Fig. S4).

In addition to the conceptual advantages offered by our mechanistic statistical model, synthetic simulations highlight several practical scenarios where our hierarchical model might be better suited to analyze genomic data. Some of these scenarios include low sample size (less than four individuals), wide dispersion in the number of individuals sampled at various locations, and a large fraction of missing genotypes. In these scenarios, rather than requiring a biogeographer to discard or collect new data, our statistical model quantifies genetic variation probabilistically and propagates the corresponding uncertainty for downstream analysis. However, there are some computational tradeoffs: our statistical analysis took 30 hours to analyze 105k SNPs on Mac Studio (M1) and, as a result, our approach may require substantial computational resources to parallelly analyze genomic datasets with millions of SNPs.

Despite these computational requirements, the hierarchical structure of the model provides new opportunities to improve statistical inferences. For example, the hierarchical model is flexible and can integrate a wide range of process and data models, allowing biogeographers to better understand the evolutionary responses of a species to changing climate while dramatically reducing the cost of analysis. The cost of sequencing varies along three axes—the number of sequenced sites in the genome, the depth of sequencing effort, and the number of sampled individuals. Typically, in RAD sequencing, a small proportion of sites in the genome are sequenced deeply to identify true genotypes, which are subsequently used to estimate allele frequencies (Baird *et al.* 2008). Although RAD sequencing is a cost-effective protocol to obtain a reduced representation of the genome, it fails to characterize a large proportion of genetic variation in the genome that could be potentially adaptive (Lowry *et al.* 2017). New sequencing protocols, such as low-depth whole genome sequencing (Alex Buerkle & Gompert 2013) and pool sequencing (Gautier *et al.* 2013), are emerging as attractive alternatives because they provide a wider genomic coverage without increasing cost or sacrificing statistical power. The key idea behind these approaches is to redistribute the same resources to sequence a larger sample of individuals and a greater proportion of the genome (Lou *et al.* 2021). This reduction in sequencing effort (per individual per locus) increases uncertainty in the estimates of individual genotypes. But, instead of discarding SNPs due to low coverage, one can model the true genotype as a parameter and propagate the corresponding uncertainty to inform population-level allele frequencies. In effect, these approaches sacrifice certainty in individual genotypes to gain genomic coverage while keeping the cost and genetic information constant. To leverage information from these sequencing protocols, the data model in the Bayesian hierarchy can be modified to account for genotype uncertainty in estimating allele frequencies.

Another potential avenue to improve inferences is to construct alternate process models incorporating a wider range of demographic histories. For example, most species, including Willow Flycatchers, are geographically structured, and, as such, the migration rates depend on distance between demes. Closer demes exchange more migrants than demes that are far apart. This may create genetic patterns, such as isolation by distance and genetic swamping, that cannot be adequately captured by an unstructured metapopulation model. One possible resolution to this problem is to use stationary distribution corresponding to the structured evolutionary dynamics (Constable & McKane 2015). However, in non-equilibrium settings, such as range shifts due to changing climate or the introduction of invasive species, stationary distributions may not be an appropriate process model because of transient evolutionary dynamics and eco-evolutionary feedback. In such cases, one may consider a probabilistic process model that jointly describes spatiotemporal changes in abundance and genetic variation (Schurr *et al.* 2012; Polechová & Barton 2015). To inform these joint process models, researchers will require temporal and spatial sampling of abundances and genomes, which can be obtained from population surveys (Pardieck *et al.* 2020) and museum collections (Payne & Sorenson 2002), respectively. These improvements in data and process models may allow biogeographers to forecast the response of a species to changing climate using mechanistic models of ecology and evolution.

Author Contributions

NG and MBH conceived and designed the study with substantial feedback from KR and CB. KR and CB provided the genetic and climate data. NG and JJVE conducted the data analysis, and NG wrote the paper with feedback from all authors.

Acknowledgments

We thank Hooten, Ruegg, and Zavaleta lab members for their feedback on the manuscript and related presentations. Upasana Sarraju for making the illustration in Figure 2. Funding for this research was provided by NSF 2222525, NSF 1927177, NSF 2222524, and NSF 2222526.

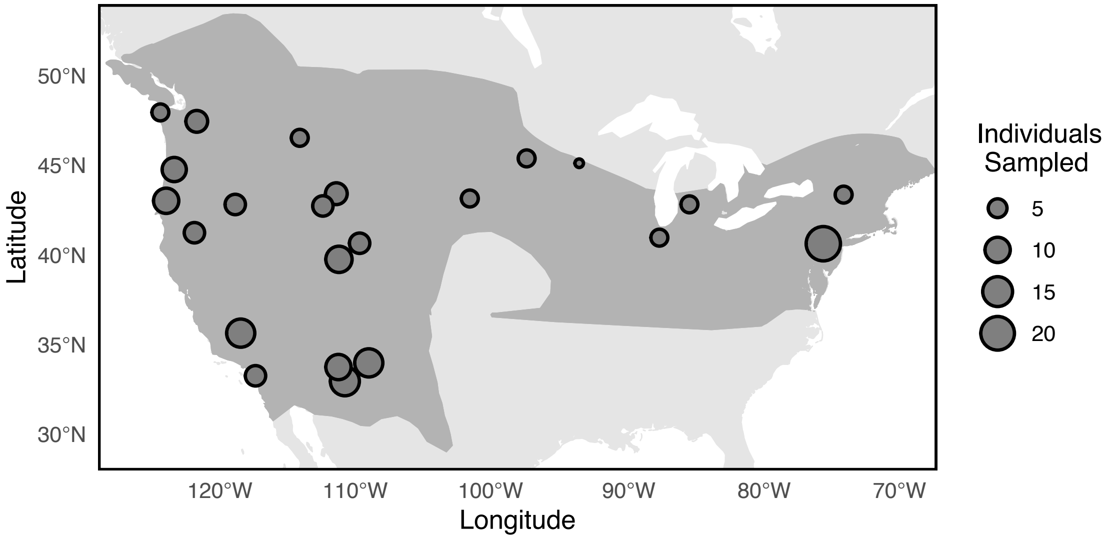

Figure 1: Breeding range of Willow Flycatchers (dark grey region) and 23 sampling locations. The size of the points is proportional to the number of sampled individuals in our study, which varied between 2 and 21 individuals.

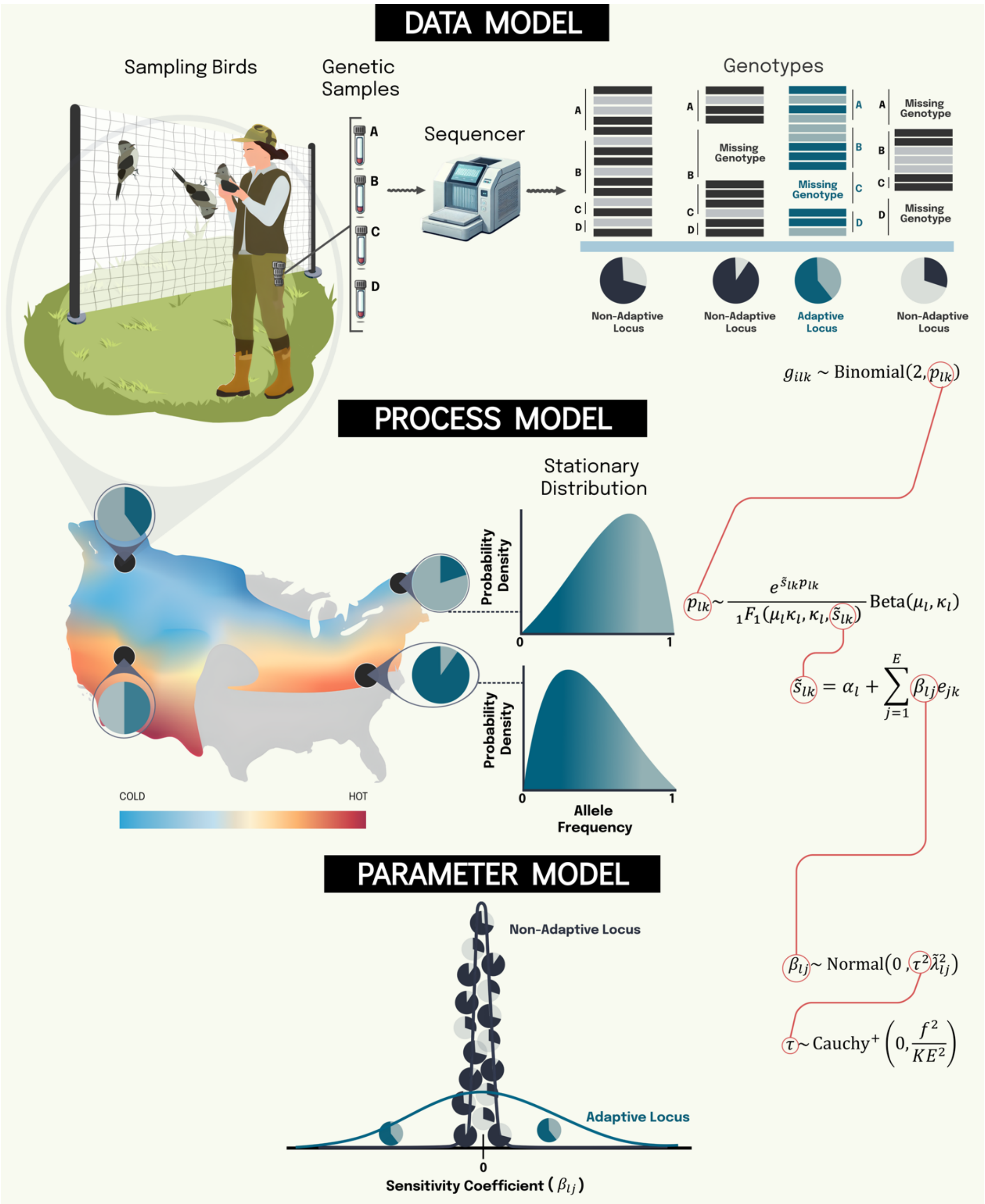
DATA MODEL
Sampling Birds
Genetic Samples
A
B
C
D
Sequencer
Genotypes
Missing Genotype
Missing Genotype
Missing Genotype
Missing Genotype
Non-Adaptive Locus
Non-Adaptive Locus
Adaptive Locus
Non-Adaptive Locus
$g_{ilk} \sim \text{Binomial}(2, p_{lk})$
PROCESS MODEL
Stationary Distribution
Probability Density
Probability Density
0
1
0
1
Allele Frequency
COLD
HOT
$p_{lk} \sim \frac{e^{\tilde{s}_{lk} p_{lk}}}{{}_1F_1(\mu_l \kappa_l, \kappa_l, \tilde{s}_{lk})} \text{Beta}(\mu_l, \kappa_l)$
$\tilde{s}_{lk} = \alpha_l + \sum_{j=1}^{E} \beta_{lj} e_{jk}$
PARAMETER MODEL
Non-Adaptive Locus
Adaptive Locus
0
Sensitivity Coefficient (β_{lj})
$\beta_{lj} \sim \text{Normal}(0, \tau^2 \tilde{\lambda}_{lj}^2)$
$\tau \sim \text{Cauchy}^+\left(0, \frac{f^2}{KE^2}\right)$

Figure 2: A conceptual diagram describing the hierarchical structure of the Bayesian model. The data model (*top*) links the individuals' genotype (g_{ilk}) to allele frequency (p_{lk}) at a locus l in patch k (Eq. 3). The process model (*middle*) describes the stationary distribution of allele frequency resulting from evolutionary forces, such as mutation, migration, drift, and local adaptation. (Eq. 8 and 11). Finally, the parameter model (*bottom*) allows us to regularize sensitivity coefficients (β_{lj}) using prior knowledge from molecular evolutionary theory (Eqs. [14], [15], and [16]).

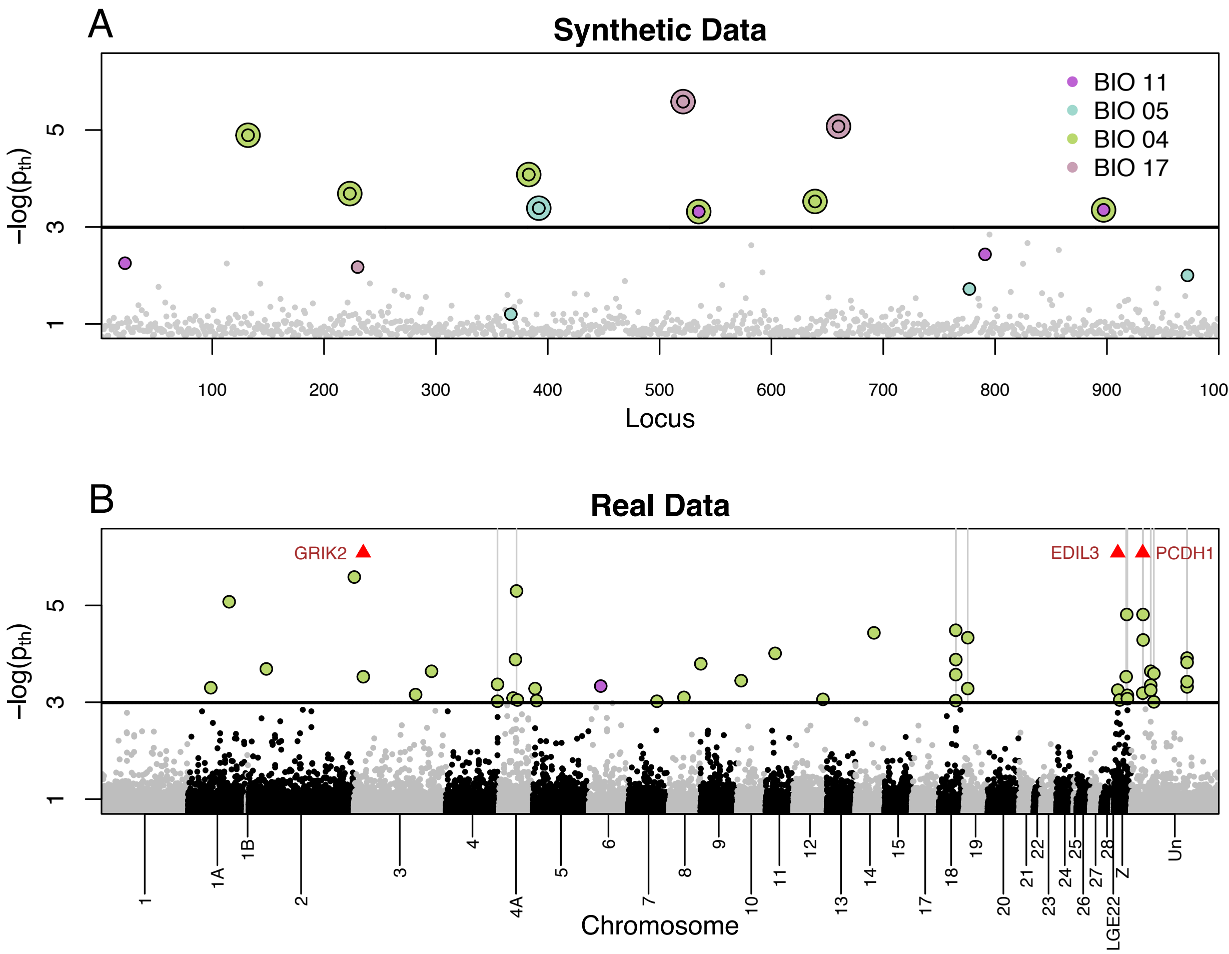

Figure 3: Manhattan plot showing negative log probability ($-\log(p_{th})$) that the posterior distribution of sensitivity coefficients (β_{li}) includes zero for synthetic (A) and real data (B). In both plots, points above the black horizontal line have p_{th} less than 0.05. (A) For synthetic data, we denote points above the black line using an outer ring and a center. The color of the center (ring) corresponds to the true (statistically inferred) environmental variable responsible for local adaptation. (B) For real data, we denote points above the black line with a solid center, with its color corresponding to the statistically inferred environmental variable responsible for local adaptation (also see Fig. S3). The vertical gray lines represent physically linked loci with statistically significant sensitivity coefficients.

References

Alex Buerkle, C. & Gompert, Z. (2013) Population genomics based on low coverage sequencing: How low should we go? *Molecular Ecology,* **22,** 3028-3035.

Anderson, E. (2019) genoscapeRtools: Tools for building migratory bird genoscapes. *R package version 0.1. 0. https://rdrr. io/github/eriqande/genoscapeRtools.*

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. & Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE,* **3,** e3376.

Barton, N.H. (1998) The effect of hitch-hiking on neutral genealogies. *Genetics Research,* **72,** 123-133.

Berliner, L.M. (1996) Hierarchical Bayesian Time Series Models. pp. 15-22. Springer Netherlands, Dordrecht.

Blanquart, F., Gandon, S. & Nuismer, S. (2012) The effects of migration and drift on local adaptation to a heterogeneous environment. *Journal of Evolutionary Biology,* **25,** 1351-1363.

Cai, W., Zhang, W., Zheng, Q., Hor, C.C., Pan, T., Fatima, M., Dong, X., Duan, B. & Xu, X.S. (2024) The kainate receptor GluK2 mediates cold sensing in mice. *Nature Neuroscience,* **27,** 679-688.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software,* **76,** 1–32.

Carroll, C., Lawler, J.J., Roberts, D.R. & Hamann, A. (2015) Biotic and climatic velocity identify contrasting areas of vulnerability to climate change. *PLoS ONE,* **10,** e0140486.

Chuvieco, E. (2020) *Fundamentals of Satellite Remote Sensing: An Environmental Approach*. CRC press, Boca Raton.

Constable, G.W. & McKane, A.J. (2015) Stationary solutions for metapopulation Moran models with mutation and selection. *Physical Review E,* **91,** 032711.

Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J.K. (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics,* **185,** 1411-1423.

De Villemereuil, P. & Gaggiotti, O.E. (2015) A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution,* **6,** 1248-1258.

Fitzpatrick, M.C. & Keller, S.R. (2015) Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters,* **18,** 1-16.

Foll, M. & Gaggiotti, O. (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics,* **180,** 977-993.

Frichot, E., Schoville, S.D., Bouchard, G. & François, O. (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution,* **30,** 1687-1699.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C. & Estoup, A. (2013) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology,* **22,** 3766-3779.

Hedrick, P.W. (1986) Genetic polymorphism in heterogeneous environments: A decade later. *Annual Review of Ecology and Systematics,* **17,** 535-566.

Hedrick, P.W. (2006) Genetic polymorphism in heterogeneous environments: The age of genomics. *Annual Review of Ecology and Systematics,* **37,** 67-93.

Hedrick, P.W., Ginevan, M.E. & Ewing, E.P. (1976) Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics,* **7,** 1-32.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society,* **25,** 1965-1978.

Hilborn, R. & Mangel, M. (2013) *The Ecological Detective: Confronting Models with Data*. Princeton University Press, Princeton, New Jersey.

Hoban, S., Kelley, J.L., Lotterhos, K.E., Antolin, M.F., Bradburd, G., Lowry, D.B., Poss, M.L., Reed, L.K., Storfer, A. & Whitlock, M.C. (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist,* **188,** 379-397.

Hobbs, N.T. & Hooten, M.B. (2015) *Bayesian Models: A Statistical Primer for Ecologists*. Princeton University Press, Princeton, New Jersey.

Hooten, M.B. & Hefley, T.J. (2019) *Bringing Bayesian Models to Life*. CRC Press, Boca Raton, Florida, USA.

Huang, H. & Knowles, L.L. (2016) Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Systematic Biology,* **65,** 357-365.

Joost, S., Bonin, A., Bruford, M.W., Després, L., Conord, C., Erhardt, G. & Taberlet, P. (2007) A spatial analysis method (SAM) to detect candidate loci for selection: Towards a landscape genomics approach to adaptation. *Molecular Ecology,* **16,** 3955-3969.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge UK.

Korolev, K.S., Avlund, M., Hallatschek, O. & Nelson, D.R. (2010) Genetic demixing and evolution in linear stepping stone models. *Reviews of Modern Physics,* **82,** 1691-1718.

Le Roy, N., Stapane, L., Gautron, J. & Hincke, M.T. (2021) Evolution of the avian eggshell biomineralization protein toolkit–new insights from multi-omics. *Frontiers in Genetics,* **12,** 672433.

Lin, J., Wang, C. & Redies, C. (2013) Expression of multiple delta-protocadherins during feather bud formation. *Gene Expression Patterns,* **13,** 57-65.

Lou, R.N., Jacobs, A., Wilder, A.P. & Therkildsen, N.O. (2021) A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology,* **30,** 5966-5993.

Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F. & Storfer, A. (2017) Breaking RAD: An evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular Ecology Resources* **17,** 142–152.

Mahoney, S.M., Reudink, M.W., Pasch, B. & Theimer, T.C. (2020) Song but not plumage varies geographically among willow flycatcher Empidonax traillii subspecies. *Journal of Avian Biology,* **51**.

Neal, R.M. (2011) MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* (eds S. Brooks, A. Gelman, G. Jones & X.-L. Meng), pp. 2. CRC Press, New York.

Olonscheck, D., Schurer, A.P., Lücke, L. & Hegerl, G.C. (2021) Large-scale emergence of regional changes in year-to-year temperature variability by the end of the 21st century. *Nature Communications,* **12,** 7237.

Pagel, J. & Schurr, F.M. (2012) Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography,* **21,** 293—304.

Pardieck, K.L., Jr., Z., D.J., L., M., A., V.I. & and Hudson, M.-A.R. (2020) North American Breeding Bird Survey Dataset 1966—2019 (ed. U.S.G.S.).

Payne, R.B. & Sorenson, M.D. (2002) Museum collections as sources of genetic data. *Bonner Zoologische Beiträge,* **51,** 97-104.

Piironen, J. & Vehtari, A. (2017) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics,* **11,** 5018–5051.

Polechová, J. & Barton, N.H. (2015) Limits to adaptation along environmental gradients. *Proceedings of the National Academy of Sciences,* **112,** 6401-6406.

Rellstab, C., Gugerli, F., Eckert, A.J., Hancock, A.M. & Holderegger, R. (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology,* **24,** 4348-4370.

Rice, S.H. (2004) *Evolutionary Theory: Mathematical and Conceptual Foundations*. Sinauer Associates, MA.

Ruegg, K., Bay, R.A., Anderson, E.C., Saracco, J.F., Harrigan, R.J., Whitfield, M., Paxton, E.H. & Smith, T.B. (2018) Ecological genomics predicts climate vulnerability in an endangered southwestern songbird. *Ecology Letters,* **21,** 1085-1096.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R.P., Banday, S., Mishra, A.K. & Das, G. (2023) Next-generation sequencing technology: current trends and advancements. *Biology,* **12,** 997.

Schurr, F.M., Pagel, J., Cabral, J.S., Groeneveld, J., Bykova, O., O'Hara, R.B., Hartig, F., Kissling, W.D., Linder, H.P., Midgley, G.F., Schröder, B., Singer, A. & Zimmermann, N.E. (2012) How to understand species' niches and range dynamics: a demographic research agenda for biogeography. *Journal of Biogeography,* **39,** 2146--2162.

Sedgwick, J.A. (2000) Willow Flycatcher (Empidonax traillii). *The Birds of North America,* **533**.

Smith, J.M. & Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genetics Research,* **23,** 23-35.

Smith, T.B., Kinnison, M.T., Strauss, S.Y., Fuller, T.L. & Carroll, S.P. (2014) Prescriptive evolution to conserve and manage biodiversity. *Annual Review of Ecology, Evolution, and Systematics,* **45,** 1-22.

Stucki, S., Orozco-terWengel, P., Forester, B.R., Duruz, S., Colli, L., Masembe, C., Negrini, R., Landguth, E., Jones, M.R. & Consortium, N. (2017) High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources,* **17,** 1072-1089.

Theimer, T.C., Smith, A.D., Mahoney, S.M. & Ironside, K.E. (2016) Available data support protection of the Southwestern Willow Flycatcher under the Endangered Species Act. *The Condor: Ornithological Applications,* **118,** 289-299.

Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F., De Siqueira, M.F., Grainger, A. & Hannah, L. (2004) Extinction risk from climate change. *Nature,* **427,** 145-148.

Thompson, J.N. (2013) *Relentless Evolution*. University of Chicago Press, Chicago.

Unitt, P. (1987) Empidonax traillii extimus: an endangered subspecies. *Western Birds,* **18,** 137-162.

Urban, M.C. (2015) Climate change. Accelerating extinction risk from climate change. *Science,* **348,** 571-573.

Wagner, H.H., Chávez-Pesqueira, M. & Forester, B.R. (2017) Spatial detection of outlier loci with Moran eigenvector maps. *Molecular Ecology Resources,* **17,** 1122-1135.

Wang, J., Yu, J., Lipka, A.E. & Zhang, Z. (2022) Interpretation of Manhattan plots and other outputs of genome-wide association studies. *Genome-Wide Association Studies* (eds D. Torkamaneh & F. Belzile), pp. 63-80. Springer, New York.

Wikle, C.K. (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology,* **84,** 1382—1394.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics,* **16,** 97.