

Identifying adaptive variation in spatially structured populations using low-coverage whole-genome sequencing data

Nikunj Goel ^{1,2*}, Christen M. Bossu ³, Seorim Yi ¹, Timothy M. Brown ⁴, Erica C.N. Robertson ³, Peri E. Bolton ^{5,6}, Ben J. Vernasco ⁷, Reza Goljani Amirkhiz ⁴, Erika S. Zavaleta ⁴, Kristen C. Ruegg ³, and Mevin B. Hooten ¹

¹Department of Statistics and Data Science, The University of Texas at Austin, Austin, TX 78712, USA

²Department of Biology, Emory University, Atlanta, GA 30322, USA

³Department of Biology, Colorado State University, Fort Collins, CO 80523, USA

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

⁵Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH 43210, USA

⁶Department of Biology, East Carolina University, Greenville, NC 27858, USA

⁷Department of Biology, Whitman College, Walla Walla, WA 99362, USA

*Corresponding author: E-mail: nikunj410@gmail.com.

Associate editor: Diogo Meyer

Abstract

Successful implementation of evolutionary programs to rescue climatically threatened species requires identification of adaptive variation. Although many genotype–environment association methods have been successful in identifying adaptive variation, current approaches can be improved in two important aspects. First, most existing methods do not account for genotype uncertainty in widely available low-coverage whole-genome sequencing data. Researchers often restrict analysis to loci for which genotypes can be inferred reliably or call the most probable genotype, allowing the use of genotype-based methods. However, discarding data and false genotype calls increase the uncertainty in estimates of genetic variation and can introduce systematic biases. Second, most methods use phenomenological approaches, such as logistic regression, to partition estimated variation into adaptive and nonadaptive components. Consequently, current approaches may fail to account for evolutionary processes, such as migration–selection balance. Structured migration between climatically disparate locations can produce deviations from a smooth S-shaped response curve, which can be difficult to accommodate using generalized linear models. To overcome these challenges, we developed a method that accounts for genotype uncertainty in sequencing data and propagates this uncertainty to inform the parameters of an evolutionary model. A key feature of this model is that it describes mechanistically how genetic variation arises from joint interactions between local adaptation, structured migration, mutation, and drift. Our synthetic simulation tests reveal that accounting for genotype uncertainty and structured migration substantially reduces false negatives. We also applied our approach to analyze data on North American rosy-finches (3.7 million SNPs), a high-alpine, climatically threatened clade of bird species.

Keywords genotype–environment association, migration–selection balance, low-coverage whole-genome sequencing data, rosy-finches, effective sample size, hierarchical Bayesian model

Introduction

Contemporary climate change is expected to drive many species to extinction, especially those with limited dispersal capacity to track their environmental niches (Schloss et al. 2012) and those that occupy shrinking suitable habitats, such as sky islands (Freeman et al. 2018). Conservationists argue that evolutionary

management policies, such as assisted gene flow and genome editing (Phelps et al. 2020; Kosch et al. 2022), may provide a viable path to rescue these species that are otherwise likely to go extinct (Harrisson et al. 2014; Iverson 2024). A key prerequisite for the successful implementation of these management programs is identifying adaptive loci, which may enable managers to target specific regions of the genome for amelioration.

Received: January 20, 2026. **Revised:** August 5, 2026. **Accepted:** September 24, 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Over the past two decades, several statistical methods have been developed to identify adaptive genomic regions in climatically threatened species (see [Rellstab et al. 2015](#) for a review of studies). Broadly, these statistical methods, referred to as genotype–environment association methods, work in three steps ([Goel et al. 2025](#)). First, individuals of a species are sampled across its geographical range and are subsequently sequenced in the lab to estimate standing genetic variation. Second, the estimated genetic variation is partitioned into adaptive and non-adaptive (neutral) components. The adaptive variation corresponds to clinal patterns in allele frequencies or genotypes that are congruent with environmental variation. The nonadaptive variation corresponds to genetic patterns that arise from neutral evolutionary processes, such as mutation, migration, and genetic drift. Finally, a statistical technique is used to evaluate the relative contribution of adaptive and non-adaptive evolutionary forces in determining variation at a given locus. Despite substantial methodological progress in ecological genomics, the capabilities of current genotype–environment association methods can be further expanded to enable more robust statistical inferences.

When estimating genetic variation using low-coverage whole-genome sequencing data, many genotype–environment association (GEA) methods do not faithfully account for, and propagate, genotype uncertainty. Most current methods that use individual-level sequencing data require genotypes as inputs to estimate genetic variation ([Frichot et al. 2013](#); [Günther and Coop 2013](#); [De Villemereuil and Gaggiotti 2015](#)). However, in low-coverage sequencing data, an individual's genotype can only be inferred probabilistically ([Nielsen et al. 2011, 2012](#)). To address this issue, current genotype–environment association methods take the following routes. Researchers either use the most probable genotype, impute genotypes based on linkage maps ([Li et al. 2011](#); [Lou et al. 2021](#)), or analyze only a subset of loci for which genotypes can be called with high certainty ([Andrews et al. 2023](#)). However, this approach is suboptimal because it can introduce systematic biases from false genotype calls and result in loss of information due to reduced sample size and limited genomic coverage ([Han et al. 2014](#)). Alternatively, some researchers first estimate allele frequencies that account for sampling and genotype uncertainty using statistical software, such as ANGSD ([Korneliussen et al. 2014](#)), and then use point estimates (e.g. the mean) as inputs in the downstream analysis ([Capblancq et al. 2018](#); [Forester et al. 2018](#)). However, this approach is also suboptimal because statistical models underestimate uncertainty in relevant downstream parameters ([Fumagalli et al. 2013](#)). An improved approach to accommodate low-coverage whole-genome sequencing data is to simultaneously account for genotype and sampling uncertainty. Such an approach facilitates obtaining probabilistic estimates of genetic variation that can be used to inform downstream parameters ([Buerkle and Gompert 2013](#); [Lou et al. 2021](#)).

Another major concern with current genotype–environment association models is that they use phenomenological approaches, such as generalized linear regression models, to describe how evolutionary forces (e.g. selection, mutation, migration, and drift) interact to shape spatial genetic patterns ([Goel et al. 2025](#)). Although these phenomenological models are qualitatively consistent with evolutionary principles, they model genetic variation using statistical relationships that are often weakly motivated by

evolutionary theory. Phenomenological statistical models may not account for important evolutionary processes, making it challenging to critique and improve them. For example, current genotype–environment association models account for migration mainly in the context of how it interacts with drift ([Coop et al. 2010](#); [Frichot et al. 2013](#)): migration between demes produces spatial correlation in neutral allele frequencies that decays with distance, a pattern commonly referred to as isolation-by-distance ([Slatkin 1993](#); [McRae 2006](#)). However, interactions between migration and selection at non-neutral loci are seldom accounted for in current methods, even though these interactions can be substantial when migration is structured and a species experiences steep environmental gradients ([Kirkpatrick and Barton 1997](#); [Hoekstra et al. 2004](#)). One important consequence of the migration–selection balance is that a species may not exhibit a typical S-shaped response curve to environmental variation. Two locations experiencing high gene flow and divergent selection in response to the local environment can result in deviations from the S-shaped response curve. These deviations or irregularities stem from maladaptive gene flow, which cannot be easily accommodated by linear models.

A more effective approach is to use evolutionary theory to mechanistically model patterns of genetic variation directly (e.g. [Goel et al. 2025](#)). Because mechanistic models of evolution are constructed from first principles, we can evaluate statistical models based on their underlying biological assumptions ([Rice 2004](#)). More importantly, if necessary, we can refine the statistical model based on the species' demographic history. This mechanistic approach to identifying adaptive loci has several advantages. The precise mathematical relationship between evolutionary parameters constrains the flow of information using scientific knowledge, rather than relying on heuristic functional forms ([Hilborn and Mangel 2013](#)). Because the parameters in an evolutionary model have biological meaning, the estimated parameters may provide valuable biological insights that are otherwise challenging to obtain using phenomenological models ([Wikle 2003](#); [Hooten and Hefley 2019](#)).

In this paper, we developed a hierarchical Bayesian model (henceforth, referred to as the structured genotype-likelihood method) to address the aforementioned limitations. Our structured genotype-likelihood method consists of three levels: the data, process, and parameter models ([Berliner 1996](#)). The data model describes a probabilistic link between genetic data and genetic variation. The process model partitions this genetic variation into adaptive and non-adaptive components using a mechanistic evolutionary model. The parameter models represent our prior knowledge about the parameters. For example, neutral evolutionary theory suggests that most loci in the genome do not contribute to local adaptation ([Kimura 1983](#)) and, as such, should have a flat response curve.

The structured genotype-likelihood method has two key features. First, the method enables us to use genotype-likelihoods obtained from low-coverage whole-genome sequencing data to improve our statistical learning of genetic variation, which is then used to inform the parameters of a mechanistic evolutionary model in the downstream analysis. Second, our statistical approach relies on a new probabilistic metapopulation model of evolution to account for how structured migration (described using coalescent theory) interacts with local adaptation to modify the S-shaped response curve. To evaluate the robustness of our

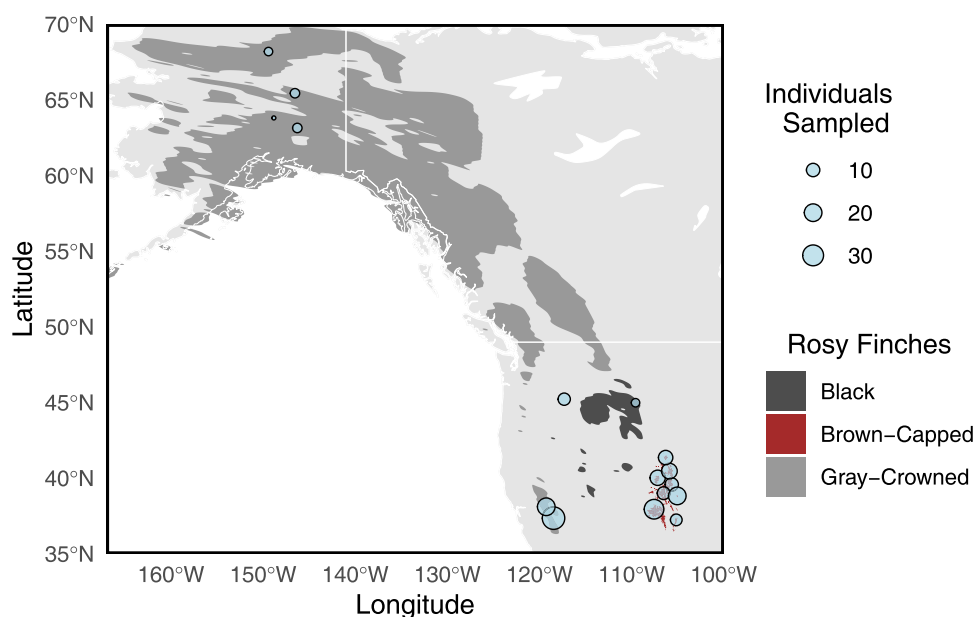


Figure 1 The breeding range of the North American rosy-finches spans western North America (DeSaix et al. 2022; Fink et al. 2023). The blue points denote the sampling sites, and the size of the points represents the number of birds sampled at the site. A total of 192 individuals were sampled. The sample sizes were highly heterogeneous, ranging from 1 to 36 individuals per sampling site.

statistical method, we test and compare our statistical model against redundancy analysis (RDA) (Capblancq and Forester 2021), and latent factor mixed models (LFMM) (Caye et al. 2019), two commonly used genotype–environment association methods, across multiple synthetic datasets with known parameter values. We then analyzed genotype-likelihood data corresponding to 3.7 million SNPs to identify adaptive variation in rosy finches (genus *Leucosticte*), a group of songbirds that predominantly breed in alpine environments.

Rosy finches

The radiation of North American rosy-finches comprises three main species: the Black Rosy-Finch, the Brown-capped Rosy-Finch, and the Gray-crowned Rosy-Finch. Together, the species complex spans western North America, ranging from Alaska, where they are found at sea level, to northern New Mexico, where they breed at elevations up to 4,000 m (Funk et al. 2023). The majority of rosy-finch populations breed in tundra-like vegetation near and above the timberline (Fig. 1).

Data

A total of 192 individuals of rosy-finches were sampled at 24 sites across the United States. Because of the proximity of some sites, we merged those within 80 km of each other, yielding a total of 16 sites (Fig. 1, Table S1). Sample sizes were highly heterogeneous across sites, ranging from 1 to 36 individuals. Using a gradient forest method (Ellis et al. 2012), we identified the five least correlated environmental variables for our genotype–environment association analysis (see Table S2 for correlations between environmental variables): (i) mean July snowpack, (ii) soil water content, (iii) mean summer precipitation, (iv) relative humidity, and (v) continentality. These environmental variables represent

the three primary axes—snow persistence, seasonal moisture availability, and broad macroclimatic gradients—that structure the rosy-finch breeding habitat. July snow patches are key foraging sites during the breeding season when chick-provisioning is at its peak (Antor 1995; Brown et al. 2026). Soil water content is an integrative proxy for habitat quality in snowmelt microhabitat zones that support invertebrate prey base and seed-producing vegetation (Epanchin et al. 2010). Summer precipitation and relative humidity, together, capture seasonal moisture availability during the breeding period, which governs the reliability of snow-adjacent and wet-meadow foraging resources (Amirkhiz et al. 2025). Continentality captures the relative influence of oceanic and continental climate on temperature seasonality and precipitation regime. This climatic variable represents a broad-scale axis governing selective environments across the mainland range of North American rosy-finches (Funk et al. 2023). Please refer to the Supplementary Information for further details on the sampling and sequencing protocols.

Model formulation

In what follows, we describe our hierarchical Bayesian model for identifying loci that may contribute to local adaptation. Henceforth, we refer to this model as the structured genotype-likelihood method. To facilitate the interpretation of the statistical model, we provide a directed acyclic graph (Fig. 2) that describes how the model parameters are linked to each other and the genetic and environmental datasets.

Data model

We assume that individuals are sampled at $\mathcal{D}$ sites across a geographical region representative of the variation in the environment experienced by the species. A total of N_k individuals are

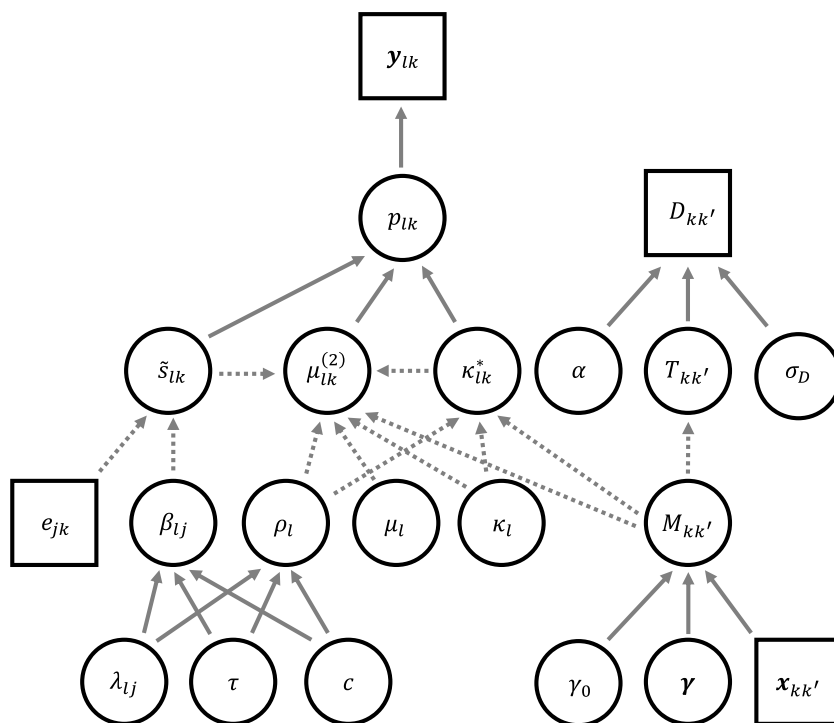


Figure 2 Directed acyclic graph (DAG) of the structured genotype-likelihood model. The squares represent data (reads and environmental variables), the circles represent parameters, the dashed arrows represent deterministic relationships, and the solid arrows represent stochastic relationships (see Eqs. 2, 9, 12, 14, and 17–26).

sampled at site k . Each diploid individual is barcoded and sequenced at L polymorphic loci. However, due to finite sequencing depth, only N_{lk} individuals have read data corresponding to locus l ($N_k \geq N_{lk}$). We assume that the loci are biallelic and the genotype of an individual is an unknown parameter. To characterize the uncertainty in the genotypes of N_{lk} individuals, the read data (generically referred to as y_{lk}) for the i th individual are aligned to a reference genome. The alignment process yields genotype-likelihoods, $\theta_{ilk\tilde{g}}(y_{ilk})$, which corresponds to the probability that the genotype of the i th individual (g_{ilk}) is equal to $\tilde{g}$ (i.e. $\theta_{ilk\tilde{g}} = P(y_{ilk} | g_{ilk} = \tilde{g})$). In this case, $\tilde{g}$ takes discrete values—zero, one, and two—and is equal to the number of copies of the reference allele. Intuitively, genotype-likelihoods, $\theta_{ilk} = [\theta_{ilk0}, \theta_{ilk1}, \theta_{ilk2}]^T$, form a three-dimensional simplex vector, such that $\sum_{\tilde{g}=0}^2 \theta_{ilk\tilde{g}} = 1$. These genotype-likelihoods reflect

uncertainty in the individual's genotype arising from low-coverage, misalignment, and sequencing errors (Kim et al. 2011). To model genetic variation and coalescent time, we split the genetic dataset into two parts— L_1 loci to model genetic variation and L_2 loci to model coalescent time ($L_1 \gg L_2$ and $L = L_1 + L_2$).

Assuming individuals are randomly sampled, we model the genetic variation at L_1 loci as

$$y_{lk} \sim \prod_{i=1}^{N_{lk}} \left(\sum_{\tilde{g}=0}^2 \underbrace{P(y_{ilk} | g_{ilk} = \tilde{g})}_{\theta_{ilk\tilde{g}}} \text{Binomial}(g_{ilk} = \tilde{g} | 2, p_{lk}) \right), \quad (1)$$

where $y_{lk} = [y_{1lk}, y_{2lk}, \dots, y_{N_{lk}lk}]^T$ and p_{lk} are the read data and reference allele frequency in site k at locus l , respectively

(Buerkle and Gompert 2013). Implicitly, the data model assumes conditional independence. We can re-express equation (1) as

$$y_{lk} \sim \sum_{n=0}^{2N_{lk}} a_{nlk} \text{Binomial}(n | 2N_{lk}, p_{lk}), \quad (2)$$

where the discrete parameter n denotes the total number of reference alleles in N_{lk} individuals. The Binomial distribution in Equation (2) accounts for sampling uncertainty, and the scalar coefficients, a_{nlk} , account for genotype uncertainty. These scalar coefficients are a function of $\theta_{ilk\tilde{g}}$ and N_{lk} and can be pre-computed using discrete convolution, such that

$$a_{nlk} = \frac{1}{\binom{2N_{lk}}{n}} \sum_{\sum_{j=1}^{N_{lk}} g_{jlk} = n} \prod_{i=1}^{N_{lk}} \theta_{ilk g_{ilk}} \binom{2}{g_{ilk}}. \quad (3)$$

Note that Equation (2) allows us to simultaneously account for uncertainty in the estimates of genetic variation arising from finite sampling and finite sequencing depth.

To quantify information content in the whole-genome sequencing data, we calculate the effective sample size (DeSaix et al. 2024)—the number of individuals with known genotypes that contain the same information about allele frequency as the genotype-likelihood data. More specifically, we calculate the number of sampled individuals with known genotypes and corresponding total reference allele counts that would yield the same posterior distribution of allele frequency as if one were to fit a standalone data model in Equation (2) to genotype-likelihoods with $p_{lk} \sim \text{Beta}(1, 1)$ prior. Using Bayes'

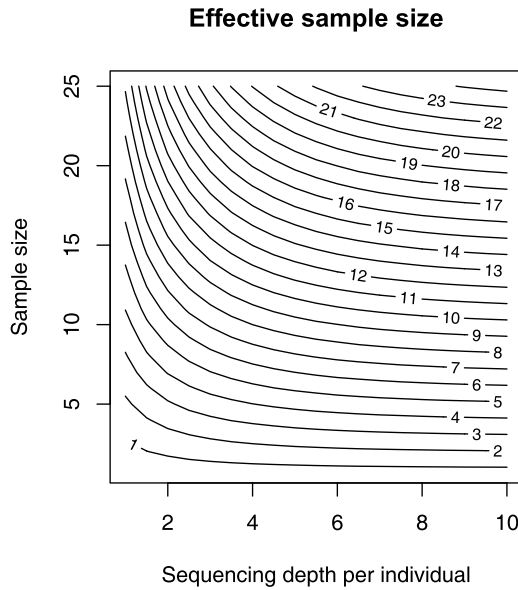


Figure 3 Effective sample size as a function of sequencing depth per individual and number of individuals sampled at a location. The contours represent combinations of sequencing depth and sample size for which the effective sample size is constant. Based on the desired level of precision in estimates of allele frequency (determined by the effective sample size), a biologist can make an informed decision about the relative allocation of resources between collecting individuals and sequencing effort.

theorem, we can show that the posterior distribution of the allele frequency at locus l and site k is given by

$$P(p_{lk} | \{y_{lk}\}) = \sum_{n=0}^{2N_{lk}} \tilde{a}_{nlk} \text{Beta}(p_{lk} | n+1, 2N_{lk} - n+1), \quad (4)$$

where $\tilde{a}_{nlk} = a_{nlk} / \sum_{n=0}^{2N_{lk}} a_{nlk}$. By moment matching, we can show that the effective sample size is

$$N_{lk}^{\text{eff}} = \frac{\omega_{lk}(1 - \omega_{lk})}{2\zeta_{lk}^2} - \frac{3}{2} \quad (5)$$

and the effective total counts of reference alleles is

$$G_{lk}^{\text{eff}} = 2(N_{lk}^{\text{eff}} + 1)\omega_{lk} - 1, \quad (6)$$

where

$$\omega_{lk} = \frac{1 + \sum_{n=0}^{2N_{lk}} \tilde{a}_{nlk}n}{2N_{lk} + 2} \quad (7)$$

and

$$\zeta_{lk}^2 = \frac{\sum_{n=0}^{2N_{lk}} \tilde{a}_{nlk} (n+1)(n+2)}{(2N_{lk} + 2)(2N_{lk} + 3)} - \omega_{lk}^2. \quad (8)$$

By defining effective sample size, N_{lk}^{eff} , and the effective total counts of reference alleles, G_{lk}^{eff} , this way (Eqs. 5–8), we can faithfully capture the uncertainty in genetic variation that is embedded in the raw read data. For reference, Fig. 3 shows how effective sample size varies with sequencing depth and the number of sequenced individuals.

To model coalescent times, we used L_2 putatively neutral loci. Because the “true” genotypes are unknown, we simulated $N_{\text{sim}} = 50$ genotype datasets (g_{ilk}) using a categorical

distribution with parameter θ_{ilk} and categories zero, one, and two. For each dataset, we calculated per-deme reference allele frequencies $\hat{p}_{lk} = (2N_{lk})^{-1} \sum_i g_{ilk}$. We then calculated within-

and between-deme pairwise diversities $\pi_{kk} = 2\hat{p}_k(1 - \hat{p}_k)$ and $\pi_{kk'} = \hat{p}_k(1 - \hat{p}_{k'}) + \hat{p}_{k'}(1 - \hat{p}_k)$ (averaged over L_2 loci and N_{sim} samples). Following Petkova et al. (2016), the genetic dissimilarity between two diploid individuals is $D_{kk'} = 4\pi_{kk'} - \pi_{kk} - \pi_{k'k'}$ and its expectation under neutral coalescent is proportional to $4T_{kk'} - T_{k'k'} - T_{kk}$ (McVean 2009), where $T_{kk'}$ is the expected coalescent time of two randomly chosen sequences from sampling site k and k' . We statistically model the observed genetic dissimilarity as

$$D_{kk'} \sim \text{Normal}(\alpha(4T_{kk'} - T_{k'k'} - T_{kk}), \sigma_D). \quad (9)$$

Although we formulate the data model for low-coverage whole-genome sequencing data, we can also specify data models for other types of commonly used genetic datasets, such as read data from sequencing whole genomes of pools of individuals (pool-seq for short, Gautier et al. 2013) and genotype data obtained from sequencing restriction site-associated DNA (RAD-seq for short, Baird et al. 2008) (see Supplementary Information for corresponding data models; Figures S1–S3).

Process model

We developed a structured metapopulation model to describe evolutionary dynamics, in which demes correspond to the sites where individuals were sampled. In each deme, the genetic variation is regulated by directional and nondirectional evolutionary forces (Rice 2004). Directional forces, such as mutation, migration, and adaptation to local climate, yield a nonzero expected evolutionary change. In contrast, nondirectional forces, such as genetic drift, yield stochastic evolutionary changes stemming from finite population size. We model these directional and nondirectional evolutionary forces jointly to describe genetic variation at locus l using the following coupled stochastic differential equations (Korolev et al. 2010):

$$dp_{lk} = \left(\underbrace{s_{lk}p_{lk}q_{lk}}_{\text{Selection}} + \underbrace{u_{2l}q_{lk} - u_{1l}p_{lk}}_{\text{Mutation}} + \rho_l \sum_{k'=1}^D \underbrace{m_{kk'}(p_{lk'} - p_{lk})}_{\text{Migration}} \right) dt + \underbrace{\sqrt{\frac{p_{lk}q_{lk}}{2N_e}}}_{\text{Drift}} dB, \quad (10)$$

where $q_{lk} = 1 - p_{lk}$ is the alternate allele frequency, u_{2l} and u_{1l} are the forward and backward mutation rates, $\rho_l m_{kk'}$ is the locus-specific symmetric migration rate between demes k and k' , N_e is the number of diploid individuals in each deme, and s_{lk} is the environmentally determined selection coefficient. To simplify our notation, we rescale time as $t = 4N_e t'$, which yields the following equations of evolution:

$$dp_{lk} = \left(\tilde{s}_{lk}p_{lk}q_{lk} + \mu_l \kappa_l - \kappa_l p_{lk} + \rho_l \sum_{k'=1}^D \tilde{m}_{kk'}(p_{lk'} - p_{lk}) \right) dt' + \sqrt{2p_{lk}q_{lk}} dB', \quad (11)$$

where $\tilde{s}_{lk} = 4N_e s_{lk}$, $\mu_l = u_{2l}/(u_{1l} + u_{2l})$, $\kappa_l = 4N_e(u_{1l} + u_{2l})$, and $\tilde{m}_{kk'} = 4N_e m_{kk'}$. For simplicity, we use a linear relationship between (rescaled) selection coefficients and E environmental variables:

$$\tilde{s}_{lk} = \alpha_l + \sum_{j=1}^E \beta_{lj} e_{jk}. \quad (12)$$

where e_{jk} is the j th environmental variable (standardized) in deme k , β_{lj} is the selection coefficient sensitivity to variation in the j th environmental variable, and α_l can be interpreted as the selection coefficient at mean environmental conditions (Goel et al. 2025).

The long-term evolutionary dynamics in Equation (11) can be thought of as a balance between directional and non-directional evolutionary forces. The directional evolutionary forces drive the mean allele frequencies in the metapopulation to an equilibrium determined by selection, mutation, and migration. However, stochastic fluctuations in allele frequencies due to genetic drift result in deviations from the equilibrium allele frequencies. As the magnitude of the deviation increases, the directional evolutionary forces become stronger, pushing allele frequencies in the opposite direction and creating a range of probable values around the equilibrium. This range can be characterized by the stationary distribution, which is the solution of the multivariate Fokker–Planck equation corresponding to coupled stochastic differential equations that describe the evolutionary dynamics in equation (11) (Wright 1931).

We use this stationary distribution—the probability distribution of allele frequencies after a long time—as the process model to partition genetic variation into adaptive and neutral components (Goel et al. 2025). Although the exact solution of the multivariate Fokker–Planck equation is analytically intractable, we derive an approximation of the stationary distribution using an iterative substitution approach (Ortega and Rheinboldt 2000). First, we consider the baseline case in which migration is absent. In this case, the stationary distribution of the reference allele frequency is given by

$$\psi_s(p_{lk} | \tilde{s}_{lk}, \mu_l, \kappa_l) = \frac{e^{\tilde{s}_{lk} p_{lk}}}{{}_1F_1(\mu_l \kappa_l, \kappa_l, \tilde{s}_{lk})} \text{Beta}(p_{lk} | \mu_l, \kappa_l), \quad (13)$$

where ${}_1F_1$ is the confluent hypergeometric function (Wright 1931), and μ_l and κ_l are mean and precision parameters of the Beta distribution. The mean of this stationary distribution is given by

$$\bar{p}_{lk} = \mu_l \frac{{}_1F_1(\mu_l \kappa_l + 1, \kappa_l + 1, \tilde{s}_{lk})}{{}_1F_1(\mu_l \kappa_l, \kappa_l, \tilde{s}_{lk})}. \quad (14)$$

To include migration, we back-substitute $\bar{p}_{lk}$ in equation (11), to get

$$dp_{lk} = \left(\tilde{s}_{lk} p_{lk} q_{lk} + \underbrace{\left(\mu_l \kappa_l + \rho_l \sum_{k' \neq k}^D \tilde{m}_{kk'} \bar{p}_{lk'} \right)}_{\mu_{lk}^{(1)} \kappa_{lk}^{(1)}} - \underbrace{\left(\kappa_l + \rho_l \sum_{k' \neq k}^D \tilde{m}_{kk'} \right)}_{\kappa_{lk}^*} p_{lk} \right) dt' + \sqrt{2p_{lk} q_{lk}} dB' \quad (15)$$

The substitution decouples the stochastic evolutionary dynamics of allele frequencies in demes, which allows us to analytically

approximate the stationary distribution of allele frequencies under the joint influence of structured migration, local adaptation, mutation, and drift as

$$\psi_s^{(1)}(p_{lk} | \tilde{s}_{lk}, \mu_{lk}^{(1)}, \kappa_{lk}^*) = \frac{e^{\tilde{s}_{lk} p_{lk}}}{{}_1F_1(\mu_{lk}^{(1)} \kappa_{lk}^*, \kappa_{lk}^*, \tilde{s}_{lk})} \text{Beta}(p_{lk} | \mu_{lk}^{(1)}, \kappa_{lk}^*), \quad (16)$$

where

$$\mu_{lk}^{(1)} = \kappa_{lk}^{*-1} \left(\mu_l \kappa_l + \rho_l \sum_{k' \neq k}^D M_{kk'} \bar{p}_{lk'} \right) \quad (17)$$

and

$$\kappa_{lk}^* = \kappa_l - \rho_l M_{kk}, \quad (18)$$

are the migration-corrected mean and precision parameters of the Beta distribution, respectively. We let $\mathbf{M}$ denote the symmetric migration-rate (generator) matrix, with $M_{kk'} = \tilde{m}_{kk'}$ and $M_{kk} = -\sum_{k' \neq k}^D \tilde{m}_{kk'}$. The migration-corrected mean of the stationary distribution, or response curve, is given by

$$\bar{p}_{lk}^{(1)} = \mu_{lk}^{(1)} \frac{{}_1F_1(\mu_{lk}^{(1)} \kappa_{lk}^* + 1, \kappa_{lk}^* + 1, \tilde{s}_{lk})}{{}_1F_1(\mu_{lk}^{(1)} \kappa_{lk}^*, \kappa_{lk}^*, \tilde{s}_{lk})}. \quad (19)$$

To further improve the approximation, we again substitute $\bar{p}_{lk}^{(1)}$ in equation (11) and re-calculate second-order migration-corrected $\mu_{lk}^{(2)}$, the stationary distribution $\psi_s^{(2)}$, and the response curve $\bar{p}_{lk}^{(2)}$. Our numerical simulations of the coupled stochastic differential equations (Eq. 11) confirm that $\psi_s^{(2)}$ reasonably approximates the stationary distribution (see Figure S4). More importantly, the migration-corrected response curve, $\bar{p}_{lk}^{(2)}$, faithfully accounts for irregularities in the response curve resulting from maladaptive gene flow. Therefore, we use

$$p_{lk} \sim \frac{e^{\tilde{s}_{lk} p_{lk}}}{{}_1F_1(\mu_{lk}^{(2)} \kappa_{lk}^*, \kappa_{lk}^*, \tilde{s}_{lk})} \text{Beta}(p_{lk} | \mu_{lk}^{(2)}, \kappa_{lk}^*) \quad (20)$$

as our process model. Although higher-order iterative corrections to the stationary distribution ($\psi_s^{(k)}$ with $k \geq 3$) may further improve the process model, our simulations show that these corrections offer only marginal improvements relative to the additional computational complexity.

Finally, we link the rescaled migration rates to coalescent times. Using the first-step analysis (Wakeley 2009), we can show that the coalescent time matrix ($\mathbf{T}$) and migration matrix ($\mathbf{M}$) are related as follows:

$$\mathbf{MT} + \mathbf{TM}^T = 2(\text{diag}([T_{11}, T_{22}, \dots, T_{DD}]^T) - \mathbf{11}^T), \quad (21)$$

To constrain migration rates, we specify a dyadic regression to link migration rates between demes to the intervening features of the landscape (Schwob et al. 2024):

$$M_{kk'} \sim \text{Exponential}(e^{\gamma_0 + \mathbf{x}_{kk'}^T \boldsymbol{\gamma}}). \quad (22)$$

We let $\boldsymbol{\gamma}$, γ_0 , and $\mathbf{x}_{kk'}$ denote the regression coefficients, intercept, and dyadic covariates integrated over the least-cost path from deme k to deme k' , respectively (see Supplementary Information for details on the covariates).

Parameter model

We specify relatively diffuse priors for parameters that respect their natural bounds: $\mu_l \sim \text{Uniform}(0, 1)$, $\kappa_l \sim \text{Normal}^+(0, 4)$, $\gamma \sim \text{Normal}(\mathbf{0}, \mathbf{I})$, $\gamma_0 \sim \text{Normal}(0, 1)$, $\alpha \sim \text{Exponential}(1)$, and $\sigma_D \sim \text{Exponential}(1)$. Due to the degeneracy of the likelihood (Eq. 20), the mean of the Beta distribution in Equation (20) ($\mu_{lk}^{(2)}$) and selection coefficient ($\tilde{s}_{lk}$) are nonidentifiable parameters (Goel et al. 2025). To address this issue, we first fix $\alpha_l = 0$. Next, based on our prior knowledge that only a small fraction of loci likely confer an adaptive advantage to the species, we use regularized horseshoe priors for sensitivity coefficients (β_{ij}) (Piironen and Vehtari 2017; Goel et al. 2025). The regularization is controlled by a global shrinkage parameter, τ , which shrinks all sensitivity coefficients to zero, and a heavy-tailed local shrinkage parameter, $\tilde{\lambda}_{ij}$, that allows a small fraction of sensitivity coefficients to take large values. The cumulative effect of these shrinkage parameters can be expressed as

$$\beta_{ij} \sim \text{Normal}(0, \tau \tilde{\lambda}_{ij}), \quad (23)$$

where

$$\tilde{\lambda}_{ij} = \frac{c \lambda_{ij}}{\sqrt{c^2 + \tau^2 \lambda_{ij}^2}}. \quad (24)$$

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

$$\tau \sim \text{Cauchy}^+\left(0, \frac{f}{LE\sqrt{D}}\right), \quad (25)$$

where f is a free variable. Based on the numerical experiments by Goel et al. (2025), we set f to 20. For the local shrinkage parameter, we use a heavy-tailed prior $\lambda_{ij} \sim \text{Cauchy}^+(0, 1)$ to allow

some sensitivity coefficients to escape shrinkage. However, these large-valued coefficients are also weakly shrunk with an implied regularization of $\text{Normal}(0, c)$ with $c^2 \sim \text{InvGamma}(2, 4)$. We adopt an inverse-gamma prior because the prior mean for c^2 (which controls adaptive dynamics via sensitivity coefficients) is similar in magnitude to prior means of κ_l (which govern neutral dynamics). However, the heavy right tail of the inverse-gamma distribution allows adaptive loci to attain large sensitivity coefficients. These features of the local and global shrinkage parameters effectively induce a spike-and-slab prior (Mitchell and Beauchamp 1988), which prevents numerical pathologies when large-effect loci are weakly identifiable. Finally, we model ρ_l as

$$\rho_l \sim \text{Normal}^+(0, \tau c^{-1} \max_j(\tilde{\lambda}_{lj})). \quad (26)$$

Intuitively, Equation (26) creates a switch mechanism: for neutral loci $\rho_l \rightarrow 0$ and for adaptive loci ρ_l has an implied distribution of $\text{Normal}^+(0, 1)$. Our numerical experiments show that this switch mechanism considerably reduces overfitting and false positives.

Improving sampling efficiency

Our initial model-fitting experiments revealed that the above specification of the structured genotype-likelihood model requires a considerable amount of memory and time to fit, even for reasonably sized datasets. To address these challenges, we first compute relative migration rates ($\mathbf{M}$) using the model of genetic dissimilarity (Eq. 9), coalescent equations (Eq. 21), and dyadic regression (Eq. 22):

$$P(\alpha, \sigma_D, \mathbf{T}, \gamma, \gamma_0 | \{D_{kk'}\}) \propto \left[\prod_{k=1}^D \prod_{k'=k+1}^D \text{Normal}(D_{kk'} | \alpha(4T_{kk'} - T_{k'k'} - T_{kk}), \sigma_D) \text{Exponential}(M_{kk'} | e^{\gamma_0 + \mathbf{x}_{kk'}^T \gamma}) \right] \times \text{Normal}(\gamma_0 | 0, 1) \text{Normal}(\gamma | \mathbf{0}, \mathbf{I}) \text{Exponential}(\alpha | 1) \text{Exponential}(\sigma_D | 1). \quad (27)$$

We then treat mean estimates of migration rates as known inputs in the Bayesian model, while acknowledging that conditioning on them may lead our statistical model to underestimate parameter uncertainty.

Next, we consider the implied data model for the effective total counts of reference alleles

$$G_{lk}^{\text{eff}} \sim \text{Binomial}(2N_{lk}^{\text{eff}}, p_{lk}), \quad (28)$$

which approximately contains the same information about the allele frequencies as the original data model in Equation (2). However, the new data model allows us to marginalize over p_{lk} and the total number of reference alleles (n), which substantially reduces the computational complexity. The marginalization yields the following joint moment-matched likelihood:

$$G_{lk}^{\text{eff}} \sim \frac{{}_1F_1(\mu_{lk}^{(2)} \kappa_{lk}^* + G_{lk}^{\text{eff}}, 2N_{lk}^{\text{eff}} + \kappa_{lk}^*, \tilde{s}_{lk})}{{}_1F_1(\mu_{lk}^{(2)} \kappa_{lk}^*, \kappa_{lk}^*, \tilde{s}_{lk})} \text{BetaBinomial}(G_{lk}^{\text{eff}} | 2N_{lk}^{\text{eff}}, \mu_{lk}^{(2)}, \kappa_{lk}^*) \quad (29)$$

where the $\mu_{lk}^{(2)}$ and κ_{lk}^* are the migration-corrected mean and precision parameters of the Beta-binomial distribution. This simplification reduces memory usage, improves numerical stability, and increases implementation speed while accommodating the

uncertainty inherent in the stochastic nature of sampling, sequencing, and evolutionary dynamics. Cumulatively, our streamlined data, process, and parameter models yield the following posterior distribution:

$$\begin{aligned}
& P(\{\mu_l, \kappa_l, \rho_l, \beta_{lj}, \lambda_{lj}\}, \tau, c | \{G_{lk}^{eff}\}) \\
& \propto \left[\prod_{l=1}^L \prod_{k=1}^D \frac{{}_1F_1(\mu_{lk}^{(2)} \kappa_{lk}^* + G_{lk}^{eff}, 2N_{lk}^{eff} + \kappa_{lk}^*, \tilde{s}_{lk})}{{}_1F_1(\mu_{lk}^{(2)} \kappa_{lk}^*, \kappa_{lk}^*, \tilde{s}_{lk})} \text{BetaBinomial}(G_{lk}^{eff} | 2N_{lk}^{eff}, \mu_{lk}^{(2)}, \kappa_{lk}^*) \right] \\
& \times \left[\prod_{l=1}^L \text{Uniform}(\mu_l | 0, 1) \text{Normal}^+(\kappa_l | 0, 4) \text{Normal}^+(\rho_l | 0, \tau c^{-1} \max_j(\tilde{\lambda}_{lj})) \right] \\
& \times \left[\prod_{l=1}^L \prod_{j=1}^E \text{Normal}(\beta_{lj} | 0, \tau \tilde{\lambda}_{lj}) \text{Cauchy}^+(\lambda_{lj} | 0, 1) \right] \\
& \times \text{Cauchy}^+\left(\tau \middle| 0, f(LE\sqrt{D})^{-1}\right) \text{InvGamma}(c^2 | 2, 4).
\end{aligned} \tag{30}$$

Our numerical experiments showed that pre-computing migration rates and marginalization reduce the fitting time by an order of magnitude, without qualitatively changing the inferences (see Model Testing section below).

Model testing

To test the robustness of our statistical model, we analyzed multiple synthetic datasets. We considered 5 migration regimes (low to high), 5 sequencing depths (low to high), and 5 iterations, yielding a total of 125 synthetic datasets. For each combination of iteration, sequencing depth, and migration regime, we generated the synthetic dataset in three stages. First, we randomly chose 20 coordinates in a 2D landscape, which we treated as demes. Migration between demes decreases exponentially with distance. A deme is characterized by five environmental covariates, each drawn from a standard normal distribution. At each deme, we sampled eight diploid individuals, each with a genome of 2,000 independent loci. We used 500 loci (L_1) to fit the six GEA methods and 1,500 loci (L_2) to estimate migration rates for the structured genotype-likelihood and structured called-genotype methods. We assigned adaptive loci a small value of α_l drawn from $\text{Normal}(0, 0.5)$. We chose three loci per environmental variable and assigned them large sensitivity coefficients ($|\beta_{lj}| = 15$), with their signs determined randomly. For all L_2 loci, we set the selection coefficient to zero.

In the second stage, we simulated allele frequencies, p_{lk} , using the stochastic equations of evolution (Eq. 10) with selection coefficients from the first stage as input. In the final stage, we simulated genotype-likelihoods (θ_{ilk}) for individuals at a given sequencing depth. For the i th individual at locus l in deme k , we sampled its genotype (g_{ilk}) from a $\text{Binomial}(2, p_{lk})$ distribution, where p_{lk} is obtained from stage two of the simulations. Then, we drew a random variable from the Poisson distribution with mean equal to sequencing depth to represent the total number of sequencing reads at locus l (n_{ilk}). To generate read counts corresponding to the reference allele (n'_{ilk}), we sampled from a $\text{Binomial}(n_{ilk}, 0.5g_{ilk}(1 - \epsilon) + (1 - 0.5g_{ilk})\epsilon)$ distribution, where $\epsilon (=0.005)$ is the probability of sequencing error. Finally, we generated genotype-likelihoods using ϵ , n'_{ilk} , and n_{ilk} as inputs in Equation (3) in Buerkle and Gompert (2013). To generate called genotypes, we assigned each individual the genotype corresponding to the highest genotype-likelihood.

For each dataset, we tested six genotype–environment association methods: (i) structured model presented in this paper with genotype-likelihoods as input data, (ii) unstructured model with genotype-likelihoods as input data (adapted from Goel et al. 2025), (iii) unstructured model with called genotypes as input data (adapted from Goel et al. 2025), (iv) structured model presented in this paper with called genotypes as input data, (v) RDA with called genotypes as input data (Capblancq and Forester 2021), and (vi) LFMM with called genotypes as input data (Caye et al. 2019). The comparison between methods (i), (ii), (iii) and (iv) allowed us to evaluate whether including migration and genotype uncertainty improves inferences. The comparison between methods (iv), (v), and (vi) helped us evaluate the performance of our model against commonly used GEA methods. Note that for the second comparison, we did not use the structured genotype-likelihood method. Instead, we used the structured called-genotype method because LFMM and RDA do not accept genotype-likelihoods as inputs.

For the structured called-genotype method, we used the RAD-seq data model presented in the Supplementary Information. For the unstructured genotype-likelihood method, we used effective sample sizes and effective total counts of reference alleles as data in Equation (13) in Goel et al. (2025). For the unstructured called-genotype method, we used called genotypes as data in Equation (13) in Goel et al. (2025). The performance of each method was evaluated by counting false positives and false negatives. The first four models were implemented in Stan (Carpenter et al. 2017), and LFMM and RDA were implemented using the *lmm* and *vegan* packages in the R programming language, respectively (Capblancq et al. 2018; Caye et al. 2019).

Across all synthetic data sets, the structured genotype-likelihood method presented in this paper yielded the fewest false negatives compared to the unstructured genotype-likelihood and structured called-genotype methods, albeit with higher false positives (see Table 1). As expected, the structured genotype-likelihood method performed best in high migration and low sequencing-depth regimes, underscoring the importance of accounting for genotype uncertainty and structured migration for identifying adaptive variation (Figure S5). Furthermore, the structured called-genotype method also yielded lower false negatives (but slightly higher false positives) when compared with LFMM and RDA with called genotypes as inputs (Table 1 and Figure S6).

Table 1 False negative (column 4) and false positive (column 5) rates for 6 genotype–environment association methods (column 1) across all synthetic datasets.

Method	Input data	Citation	False negative %	False positive %
Structured	Genotype likelihoods	This paper	6.99	0.475
Unstructured	Genotype likelihoods	Adapted from Goel et al. (2025)	67.41	0.001
Unstructured	Called genotypes	Adapted from Goel et al. (2025)	67.09	0.002
Structured	Called genotypes	This paper	18.08	0.128
RDA	Called genotypes	Capblancq and Forester (2021)	33.12	0.034
LFMM	Called genotypes	Caye et al. (2019)	31.36	0.073

These methods include (i) structured model presented in this paper with genotype likelihoods as input data, (ii) unstructured model with genotype likelihoods as input data (adapted from [Goel et al. 2025](#)), (iii) unstructured model with called genotypes as input data (adapted from [Goel et al. 2025](#)), (iv) structured model presented in this paper with called genotypes as input data, (v) RDA with called genotypes as input data ([Capblancq and Forester 2021](#)), and (vi) LFMM with called genotypes as input data ([Caye et al. 2019](#)). Also see [Figures S5](#) and [S6](#) for a comparison of synthetic datasets across different combinations of migration rates and sequencing depth.

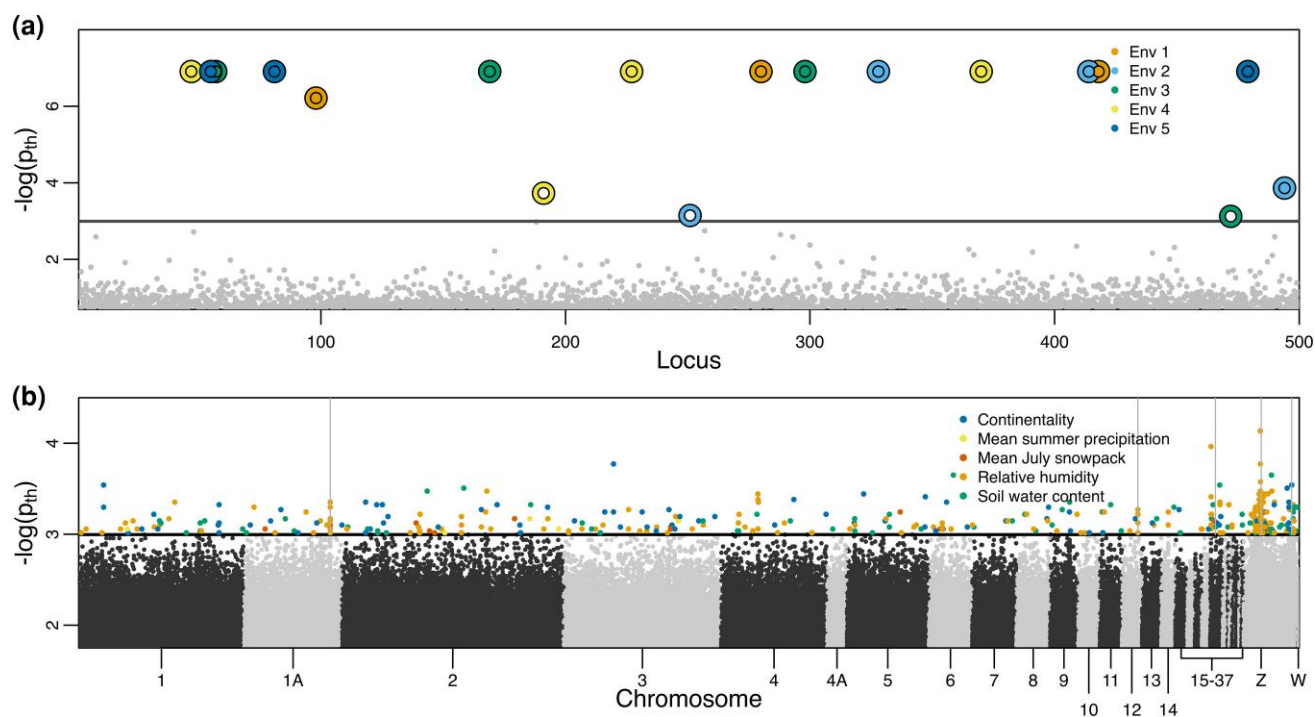


Figure 4 Manhattan plot showing negative log probability ($-\log(p_{th})$) that the posterior distribution of sensitivity coefficients (β_{ij}) includes zero for synthetic (a) and rosy-finch data (b). In both plots, points above the black horizontal line have p_{th} less than 0.05. (a) For synthetic 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. The vertical gray lines represent physically linked loci with statistically significant sensitivity coefficients.

For illustration, the Manhattan plot in [Fig. 4a](#) shows the results from the structured genotype-likelihood method presented in this paper for medium migration and a sequencing depth of 4. To plot the results, we obtained the posterior samples corresponding to sensitivity coefficients (β_{ij}) for L_1 loci. 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 was less than the threshold value 0.05 (p_{th}), we inferred that the locus could have contributed to

local adaptation to the corresponding environmental variable. In [Fig. 4a](#), the y-axis represents the negative log probability that the posterior distribution of sensitivity coefficients included zero ([Wang et al. 2022b](#)). Thus, the points above $y = 2.99$ (the black line) have $p_{th} < 0.05$. These points are characterized by a ring and a solid center. The color of the center represents the true environmental variable associated with the loci, and the color of the ring represents the statistically inferred adaptive environmental variable. We also reanalyzed the data in [Fig. 4a](#) by

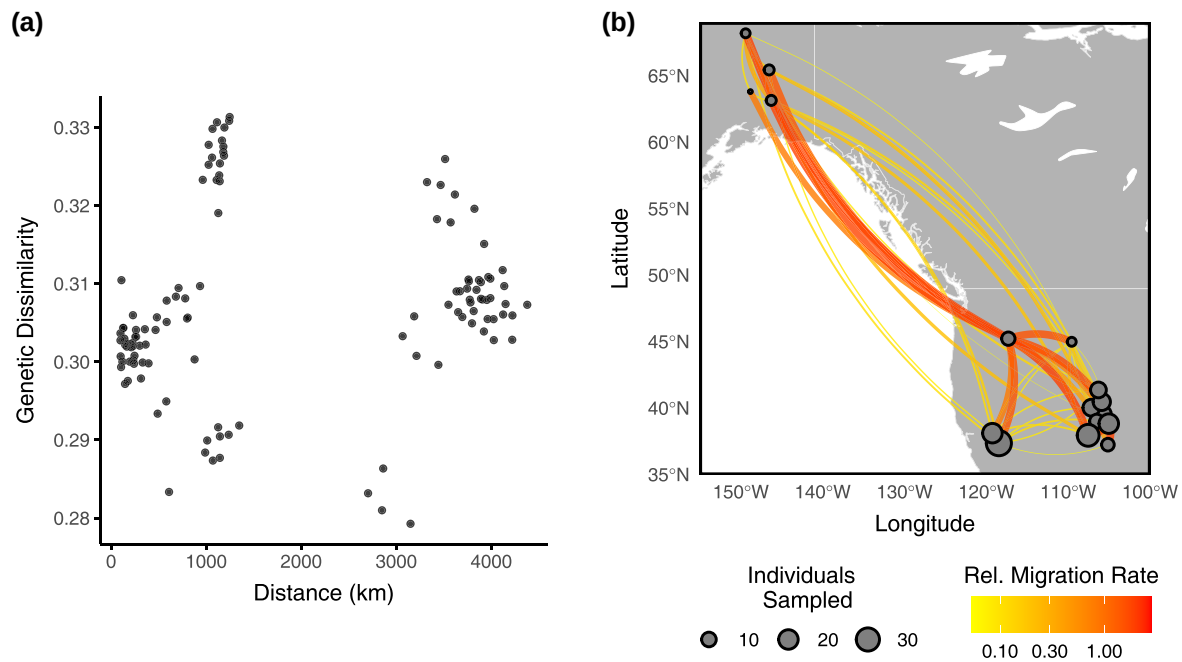


Figure 5 Genetic dissimilarity (a) and migration map (b) of North American rosy-finches. The genetic dissimilarity of rosy-finches does not increase with geographical distance between demes, suggesting heterogeneous landscape connectivity.

fitting the full Bayesian model with migration rates (M) and total number of reference alleles (n) as parameters (see [Supplementary Information](#) for the full Bayesian likelihood). Our results were similar to the results in [Fig. 4a](#) (see [Figure S7](#)).

Data analysis

We identified a total of 3.87 million high-quality variants after filtering for quality, missingness, and batch effects. After aligning the reads with the Zebra Finch reference genome, we recovered approximately 3.75 million SNPs. We used 50 thousand SNPs to estimate mean relative migration rates ([Eq. 27](#)), which were subsequently used as known inputs in the Bayesian model ([Eq. 30](#)). Computing migration rates required 2 min. To analyze the remaining 3.7 million SNPs, we partitioned the genotype-likelihood data into 37 nonoverlapping 100k SNP batches and analyzed each batch concurrently on ICX compute nodes on Stampede3, Texas Advanced Computing Center ([Boerner et al. 2023](#)). Each ICX node features 80 cores with a nominal clock rate of 2.3 GHz (max frequency: 3.4 GHz) and 256 GB of RAM. Our simulations took 16 h (or cumulatively 25 d of compute time) to fit the Bayesian model ([Eq. 30](#)). We also reimplemented our model in JAX ([Bradbury et al. 2018](#)), with sampling in NumPyro's NUTS ([Phan et al. 2019](#)), and it took ~20 min to fit 100k SNPs (or cumulatively ~12 h of compute time) on an H100 compute node on Stampede3. Using one of the SNP batches, we also computed the effective sample size.

Our data simulations revealed that the genotype-likelihood data resulted in an effective sample size of 69% of the sampled birds. Our statistical method identified 434 candidate loci with statistically significant sensitivity coefficients. We found that 57% of these outlier loci were associated with relative humidity, 18% with soil water content, 21% with continentality, 3% with

the average July snowpack, and 1% with mean summer precipitation ([Fig. 4b](#)). About 48% of the statistically significant loci were clustered together on the genome due to genetic hitchhiking (gray vertical lines in [Fig. 4b](#)).

Of the 434 loci, 347 were located within 25 kb upstream or downstream of 231 named genes, which is thought to be the distance within which linkage disequilibrium is expected to break down ([Backström et al. 2006](#)). In addition, we compiled a list of candidate genes from the literature with gene ontology (GO) terms that could be linked to an alpine specialist, for example, associated with aspects of thermal tolerance, as well as genes linked to avian thermal stress and/or bill morphology. We also included genes previously identified in the rosy-finch species complex ([Funk et al. 2021, 2023](#)), as well as migratory-linked candidate genes previously identified in other species. We then searched for genes within 25 kb of these candidate loci in the published candidate list and for genes with relevant GO terms to examine the potential functions of our identified candidate SNPs and further narrowed our focus to 28 candidate genes in close proximity to statistically significant loci ([Table S3](#)). These genes broadly fall into 5 functional categories: pigmentation (7 genes), migration (5 genes), physiology linked to high-altitude adaptation (8 genes), physiology linked to stress response (3 genes), and beak and feather development (5 genes) (see [Table S3](#)).

The relationship between genetic dissimilarity and distance between demes suggests that North American rosy-finches do not show an isolation-by-distance pattern ([Fig. 5a](#)). In particular, we found that the Gray-crowned Rosy-Finches in the Sierra Nevada, California, exhibit low connectivity with the rest of the species complex ([Fig. 5b](#)). Our results show that migration-selection balance can play an important role in shaping genetic variation ([Fig. 6](#)). When migration is absent, allele frequencies (represented by gray squares in [Fig. 6](#)) follow a typical S-shaped response curve. However, when migration is present, we observed that allele

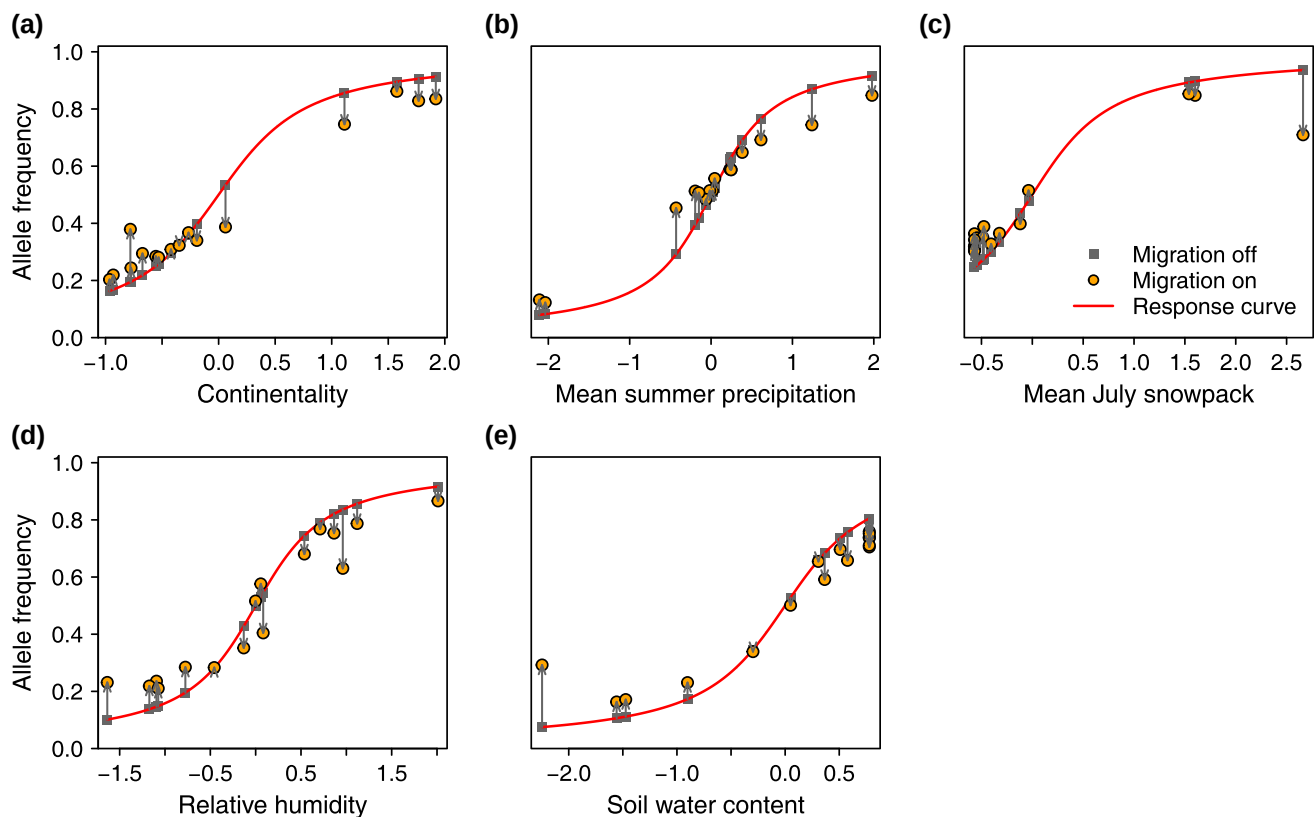


Figure 6 Relationship between allele frequency and environment when migration is present (orange points) and absent (gray squares). In the absence of migration, the response curve, $\bar{p}_{ik}$, takes a typical S-shaped curve (red line). However, when we account for migration, we observe irregularities in the response curve ($\bar{p}_{ik}^{(2)}$), suggesting that migration-selection balance can lead to nonclinical patterns in allele frequencies. To generate these plots, we use migration rates inferred from rosy-finch genetic data (Fig. 5). For all plots, we use parameter values $\mu_i = 0.5$, $\kappa_i = 3.5$, $\rho_i = 1$, and $\beta_{ij} = 10$.

frequencies (represented by orange points in Fig. 6) regress to the mean, and the response curve exhibits irregularities.

Discussion

We presented a highly parallelizable hierarchical Bayesian model to identify genomic adaptation in spatially structured populations (Fig. 2 and Eq. 30). Our statistical model extends the previous approach of Goel et al. (2025) and other currently available genotype–environment association methods (Rellstab et al. 2015) in two respects. We present (a) a probabilistic data model (Eqs. 1–8 and 28) to estimate genetic variation using low-coverage whole-genome sequencing data, and (b) a novel process model of evolution (Eq. 20) to account for migration–selection balance when fitting a clinal relationship between allele frequency and environmental covariates ($\bar{p}_{ik}^{(2)}$). Our approach is one of the first methods for identifying adaptive loci under selection from the environment while accounting for migration.

Data model for low-coverage whole-genome sequencing data

To model low-coverage whole-genome sequencing data, we propose a binomial data model (Eq. 28). This binomial model uses effective sample size and genotype counts (Eqs. 4–8) as inputs,

allowing us to simultaneously capture sampling and genotype uncertainty arising from finite sampling and sequencing effort, respectively. This methodological approach has two advantages (Lou et al. 2021). First, accounting for sampling uncertainty enables us to include sites with extremely low sample sizes (1–3 individuals). Even though these sites individually may not contain much information about underlying genetic variation in the landscape, collectively they may contain nontrivial information to constrain the shape of the response curve, resulting in more robust statistical inferences. Second, by explicitly accounting for and propagating genotype uncertainty, our statistical model avoids systematic biases from false genotype calls and provides more precise estimates of genetic variation across a broader genomic region (see the comparison between structured genotype-likelihood and structured called-genotype methods in Table 1 and Figure S5).

Accounting for genotype uncertainty also offers opportunities to design more affordable sampling and sequencing protocols by spreading the sequencing effort over a larger number of individuals (Buerkle and Gompert 2013; Lou et al. 2021). As an example, our simulations showed that sequencing 20 individuals at 2× (genotypes cannot be called reliably and the total sequencing effort is 40×) yields the same information about allele frequency as sequencing 12 individuals at 30× (genotypes can be called reliably and the total sequencing effort is 360×), even though the two strategies differ by an order of magnitude in total sequencing effort.

To aid practitioners in their study design, we show in Fig. 3 how effective sample size—which determines the precision of

estimates of allele frequency—varies with sequencing and sampling effort. The figure can be used to determine how to distribute resources between sequencing and sampling to achieve a desired effective sample size. When human resources are limited or the species occurs in remote areas, one can devote more resources toward sequencing. Alternatively, for species that are easy to collect, such as many plants, one can benefit by collecting a larger sample size. Surprisingly, using low-coverage whole-genome sequencing data does not incur additional computational cost compared with using traditional called-genotype data, because effective sample size and genotype counts can be evaluated cheaply. For the rosy-finches dataset, it took under 30 min to calculate effective sample sizes and effective total counts of reference alleles, which is substantially less than 25 d of compute time to fit the structured genotype-likelihood method.

A mechanistic structured metapopulation model of evolution

Another main feature of our statistical method is a new mechanistic process model of evolution (Eqs. 16–20). This process model corresponds to the stationary distribution of allele frequencies shaped by the joint interactions between local adaptation, mutation, structured migration, and drift. We derive an approximate stationary distribution (Eq. 20) corresponding to the stochastic evolutionary dynamics in Equation (11). A key feature of our process model is that it allows us to mechanistically account for the interactions between selection (measured by sensitivity coefficients; see Eq. 12) and migration by learning these parameters from genetic data using regularization (Hooten and Hobbs 2015; Eqs. 23–25) and structured coalescent theory (Eq. 21), respectively. These interactions can be substantive when species movement is structured, and the landscape has steep environmental gradients.

Our analysis of rosy-finch genomes and synthetic genetic datasets reveals that accounting for migration can provide new biological insights. Estimates of migration rates suggest that rosy-finches do not exhibit isolation by distance, a commonly observed pattern when migration is distance-dependent (Slatkin 1993) (Fig. 5a). Instead, we find strong connectivity between distant locations (Alaska and populations in the Western United States) and weak connectivity between nearby locations (Sierra Nevada and other populations in the Western United States), indicating heterogeneous migration rates (Fig. 5b). The observed patterns of population connectivity could be related to the diverse nonbreeding season movement behavior of North American rosy-finches. Specifically, populations that breed in Alaska and Canada are known to make long-distance movements to the Western United States during the nonbreeding season and form mixed-species flocks with other species of rosy-finches and other subspecies of Gray-crowned rosy-finches (MacDougall-Shackleton et al. 2000; Campbell et al. 2025). Because pair formation is thought to occur late in the nonbreeding season (MacDougall-Shackleton et al. 2000), such population mixing during the nonbreeding season could facilitate gene flow between spatially distant breeding populations. The Sierra Nevada population, on the other hand, is thought to exhibit short-distance facultative movements to lower elevation areas

east of the Sierra Nevada (MacDougall-Shackleton et al. 2000) which likely provide fewer opportunities for mixing between distant breeding populations during the nonbreeding season.

Accounting for migration can also increase the statistical power for identifying adaptive variation in structured populations (see the comparison between structured genotype-likelihood and unstructured genotype-likelihood methods in Table 1 and Figure S5) because of the better approximate stationary distribution (see Figure S4). In simple cases, such as when migration between patches is either absent or global, theory predicts a smooth S-shaped response curve (gray points in Fig. 6). However, the distributions of most species are spatially structured, with complex migration patterns (e.g. Figure 5). In this realistic scenario, the migration-selection balance can introduce irregularities in the response curve (orange points in Fig. 6) due to maladaptive gene flow from environmentally distinct locations. Consequently, statistical methods that do not account for structured migration will lose substantial power when fitting a smooth S-shaped response curve to allele frequencies that do not vary smoothly with environmental covariates, potentially leading to higher false-negative rates. To address this challenge, our method allows us to separate the irregularities due to structured migration from the observed response curve before fitting a smooth S-shaped function ($\bar{p}_{lk}^{(2)}$).

Analysis of synthetic and real data

Synthetic simulations show that the structured called-genotype method yields lower false negatives than LFMM and RDA in high-migration and low sequencing-depth regimes (Table 1 and Figure S6). Furthermore, analysis of 3.7 million SNPs from rosy-finch genomes revealed 434 loci putatively under selection, of which 48% were clustered together due to genetic hitchhiking (gray vertical lines in Fig. 4b). Gene ontology revealed 28 candidate genes linked to pigmentation, migration, physiology linked to high-altitude adaptation, physiology linked to stress response, and beak and feather development (Table S3). Our method identified eight genes previously documented in rosy-finch genomic analyses, even after adding new differentiated populations and individuals (Funk and Taylor 2019; Funk et al. 2023). Four of the genes linked to high-altitude adaptation (*AGGF1*, *ALDH1A1*, *EGLN1*, and *JMY*) were previously found by Funk et al. (2023). Here, we identified two additional genes with known links to adaptability in hypoxic high-altitude environments, *EPAS1* and *TNFSF11*, and two associated with thermal tolerance in extreme environments, *PPARGC1A* and *RTEL1*. Notably, *EPAS1* forms a part of the HIF (hypoxia-inducible factor) complex that enhances oxygen delivery in hypoxic environments (Beall et al. 2010; Yi et al. 2010; Simonson et al. 2012; Graham and McCracken 2019). *TNFSF11*, also known as RANKL, is enhanced by HIF1 α in simulated hypoxic environments and promotes osteoclast differentiation and bone resorption (Wang et al. 2022a). The latter two genes are associated with extreme environments: *PPARGC1A* with cold tolerance and water scarcity in chickens (Gheys et al. 2021) and *RTEL1* with body temperature changes when subjected to acute heat (Zhuang et al. 2019). Moreover, *RTEL1* variants were associated with the risk of high-altitude pulmonary edema in the Chinese Han population (Rong et al. 2017). We also identified two genes that may be important for adaptation

to the alpine environment—one that regulates the stress response (*CRHBP*; Wan et al. 2022) and another that protects cells from oxidative stress (*GPX8*; Pei et al. 2023; Aryal et al. 2025). Five of seven pigmentation genes are associated with the melanin pathway, which may relate to known differences in the extent of dark coloration on the head and crown across the species complex. Further, melanin pigmentation differences may also shield against intense solar radiation at different elevations in line with Gloger's rule (Rensch 1938; Mayr 1942).

Future improvements

The capabilities of our statistical method can be further expanded by integrating alternate data models to analyze called-genotype data obtained by RAD sequencing and read data from pool sequencing (see [Supplementary Information](#) for a brief description of how to integrate data model for RAD- and pool-seq data with the mechanistic process model in [Eq. 20](#)). To analyze RAD-seq data (Baird et al. 2008), one can calculate the total number of reference alleles in the sampled individuals and model the allele frequency as a binomial distribution to account for sampling uncertainty (Coop et al. 2010; De Villemereuil and Gaggiotti 2015; Goel et al. 2025) ([Figure S1](#)).

To analyze pool-seq data (Gautier et al. 2013), we need to account for both sampling and read uncertainty arising from finite sample size and read counts, respectively, resulting in a mixture data model of two binomial distributions (Coop et al. 2010; Günther and Coop 2013) ([Figures S2 and S3](#)). An advantage of the pool-seq method is that library preparation costs are much lower than those for individual-based sequencing methods. However, in pool-seq data, all individual-level information is lost, making it challenging to reuse the data for other analyses such as pedigree reconstruction.

Our process model can also be improved by accounting for additional evolutionary mechanisms that are currently missing. Our process model assumes that the allele frequencies in demes are independent. However, allele frequencies in demes may not be independent due to coupling via migration. This can yield false positives due to spurious correlations between allele frequencies and environmental gradients (Coop et al. 2010; Frichot et al. 2013). Another potential avenue for improvement is to build a process model that incorporates polygenic adaptation in which multiple loci regulate a phenotypic trait (such as body size). Lotterhos (2023) showed that because of genetic redundancy, a clinal pattern in a phenotypic trait may not always correspond to a clinal pattern in loci that regulate the trait. Consequently, our statistical model may fail to detect adaptive variation.

Acknowledgments

We thank Hooten, Ruegg, and Zavaleta lab members for their feedback on the manuscript. We also thank Timothy H. Keitt for contributing computational resources for the project.

Author contributions

Nikunj Goel and Mevin B. Hooten conceived and designed the study with substantial feedback from Kristen C. Ruegg and

Christen M. Bossu. Kristen C. Ruegg and Christen M. Bossu provided the genetic and climate data. Timothy M. Brown, Erika S. Zavaleta, Erica C. N. Robertson, and Reza Goljani Amirkhiz collected Sierra Nevada Gray-crowned Rosy-Finch samples. Ben J. Vernasco and Peri E. Bolton collected Wallowa samples in Oregon. Nikunj Goel conducted the mathematical and data analysis, Seorim Yi constructed the dyadic regression model of migration, and Nikunj Goel wrote the paper with feedback from Christen M. Bossu, Ben J. Vernasco, Erika Zavaleta, Kristen C. Ruegg, and Mevin B. Hooten. All authors reviewed and approved the manuscript.

Supplementary material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

Funding for this research was provided by grants from the U.S. National Science Foundation: NSF 2222525, NSF 1927177, NSF 2222524, and NSF 2222526. This work used the Texas Advanced Computing Center (TACC) at The University of Texas at Austin through allocations TG-BIO250092 and TG-BIO260362 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) program, which is supported by NSF 2138259, NSF 2138286, NSF 2138307, NSF2137603, and NSF2138296.

Conflicts of interest

Authors declare no conflict of interest.

Data and resource availability

The code and processed data files are available at <https://doi.org/10.5281/zenodo.18307932>. The link contains an “analyze_your_data” folder that contains R and Stan scripts to help users implement our method for low-coverage whole-genome sequencing data, pool sequencing data, and RAD sequencing data. The raw genetic data have been deposited at NCBI under BioProject accessions PRJNA1405131, PRJNA1297057, PRJNA659436, and PRJNA1380501 (see [Supplementary Information](#) for more details).

References

- Amirkhiz RG et al. Hierarchical habitat selection for reconstructing past and present niches and distributions of data-limited species under climate change. *Divers Distrib.* 2025;31:e70128. <https://doi.org/10.1111/ddi.70128>.
- Andrews KR et al. Whole genome resequencing identifies local adaptation associated with environmental variation for redband trout. *Mol Ecol.* 2023;32:800–818. <https://doi.org/10.1111/mec.16810>.
- Antor RJ. The importance of arthropod fallout on snow patches for the foraging of high-alpine birds. *J Avian Biol.* 1995;26:81–85. <https://doi.org/10.2307/3677216>.

- Aryal B *et al.* Major oxidative and antioxidant mechanisms during heat stress-induced oxidative stress in chickens. *Antioxidants*. 2025;14:471. <https://doi.org/10.3390/antiox14040471>.
- Backström N, Qvarnström A, Gustafsson L, Ellegren H. Levels of linkage disequilibrium in a wild bird population. *Biol Lett*. 2006;2:435–438. <https://doi.org/10.1098/rsbl.2006.0507>.
- Baird NA *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*. 2008;3:e3376. <https://doi.org/10.1371/journal.pone.0003376>.
- Beall CM *et al.* Natural selection on EPAS1 (HIF2 α) associated with low hemoglobin concentration in Tibetan highlanders. *Proc Natl Acad Sci U S A*. 2010;107:11459–11464. <https://doi.org/10.1073/pnas.1002443107>.
- Berliner LM. Hierarchical Bayesian time series models. In: Hanson KM, Silver RN, editors. *Maximum entropy and Bayesian methods*. Kluwer Academic Publishers; 1996. p. 15–22.
- Boerner TJ, Deems S, Furlani TR, Knuth SL, Towns J. ACCESS: advancing innovation: NSF's Advanced cyberinfrastructure coordination ecosystem: services & support. In: Practice and Experience in Advanced Research Computing (PEARC '23), July 23–27, 2023, Portland, OR, USA. ACM, New York, NY, USA. pp. 173–176. 2023.
- Bradbury J *et al.* 2018. JAX: composable transformations of Python+NumPy programs [computer program]. Version 0.11.0. <http://github.com/jax-ml/jax>
- Brown TM, *et al.* 2026. Co-limitation by stable, dynamic and directional habitat features shapes climate vulnerability in an alpine specialist. *bioRxiv* 2026.2004.2009.717579. <https://doi.org/10.64898/2026.04.09.717579>
- Buerkle C, Gompert Z. Population genomics based on low coverage sequencing: how low should we go? *Mol Ecol*. 2013;22:3028–3035. <https://doi.org/10.1111/mec.12105>.
- Campbell CJ *et al.* Quantifying rosy-finch migration with stable hydrogen isotope feather markers highlights the need for inter-state collaboration to reach conservation goals. *Avian Conserv Ecol*. 2025;20:6. <https://doi.org/10.5751/ACE-02798-200106>.
- Capblancq T, Forester BR. Redundancy analysis: a Swiss army knife for landscape genomics. *Methods Ecol Evol*. 2021;12:2298–2309. <https://doi.org/10.1111/2041-210X.13722>.
- Capblancq T, Luu K, Blum MG, Bazin E. Evaluation of redundancy analysis to identify signatures of local adaptation. *Mol Ecol Resour*. 2018;18:1223–1233. <https://doi.org/10.1111/1755-0998.12906>.
- Carpenter B *et al.* Stan: a probabilistic programming language. *J Stat Softw*. 2017;76:1–32. <https://doi.org/10.18637/jss.v076.i01>.
- Caye K, Jumentier B, Lepeule J, François O. LFMM 2: Fast and accurate inference of gene–environment associations in genome-wide studies. *Mol Biol Evol*. 2019;36:852–860. <https://doi.org/10.1093/molbev/msz008>.
- Coop G, Witonsky D, Di Rienzo A, Pritchard JK. Using environmental correlations to identify loci underlying local adaptation. *Genetics*. 2010;185:1411–1423. <https://doi.org/10.1534/genetics.110.114819>.
- DeSaix MG *et al.* Forecasting climate change response in an alpine specialist songbird reveals the importance of considering novel climate. *Divers Distrib*. 2022;28:2239–2254. <https://doi.org/10.1111/ddi.13628>.
- DeSaix MG, Rodriguez MD, Ruegg KC, Anderson EC. Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data. *Methods Ecol Evol*. 2024;15:493–510. <https://doi.org/10.1111/2041-210X.14286>.
- De Villemereuil P, Gaggiotti OE. A new FST-based method to uncover local adaptation using environmental variables. *Methods Ecol Evol*. 2015;6:1248–1258. <https://doi.org/10.1111/2041-210X.12418>.
- Ellis N, Smith SJ, Pitcher CR. Gradient forests: calculating importance gradients on physical predictors. *Ecology*. 2012;93:156–168. <https://doi.org/10.1890/11-0252.1>.
- Epanchin PN, Knapp RA, Lawler SP. Nonnative trout impact an alpine-nesting bird by altering aquatic-insect subsidies. *Ecology*. 2010;91:2406–2415. <https://doi.org/10.1890/09-1974.1>.
- Fink D *et al.* *eBird Status and Trends, Data Version: 2022; Released: 2023*. Cornell Lab of Ornithology; 2023. <https://doi.org/10.2173/ebirdst.2022>.
- Forester BR, Lasky JR, Wagner HH, Urban DL. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Mol Ecol*. 2018;27:2215–2233. <https://doi.org/10.1111/mec.14584>.
- Freeman BG, Scholer MN, Ruiz-Gutierrez V, Fitzpatrick JW. Climate change causes upslope shifts and mountaintop extirpations in a tropical bird community. *Proc Natl Acad Sci U S A*. 2018;115:11982–11987. <https://doi.org/10.1073/pnas.1804224115>.
- Frichot E, Schoville SD, Bouchard G, François O. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol*. 2013;30:1687–1699. <https://doi.org/10.1093/molbev/mst063>.
- Fumagalli M *et al.* Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*. 2013;195:979–992. <https://doi.org/10.1534/genetics.113.154740>.
- Funk ER *et al.* A supergene underlies linked variation in color and morphology in a Holarctic songbird. *Nat Commun*. 2021;12:6833. <https://doi.org/10.1038/s41467-021-27173-z>.
- Funk ER *et al.* The genetic basis of plumage coloration and elevation adaptation in a clade of recently diverged alpine and Arctic songbirds. *Evolution*. 2023;77:705–717. <https://doi.org/10.1093/evolut/qpac064>.
- Funk ER, Taylor SA. High-throughput sequencing is revealing genetic associations with avian plumage color. *Auk*. 2019;136:ukz048. <https://doi.org/10.1093/auk/ukz048>.
- Gautier M *et al.* Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Mol Ecol*. 2013;22:3766–3779. <https://doi.org/10.1111/mec.12360>.
- Gheyra AA *et al.* Integrated environmental and genomic analysis reveals the drivers of local adaptation in African indigenous chickens. *Mol Biol Evol*. 2021;38:4268–4285. <https://doi.org/10.1093/molbev/msab156>.
- Goel N *et al.* Identifying genomic adaptation to local climate using a mechanistic evolutionary model. *Methods Ecol Evol*. 2025;16:2448–2460. <https://doi.org/10.1111/2041-210X.70117>.
- Graham AM, McCracken KG. Convergent evolution on the hypoxia-inducible factor (HIF) pathway genes EGLN1 and EPAS1 in high-altitude ducks. *Heredity (Edinb)*. 2019;122:819–832. <https://doi.org/10.1038/s41437-018-0173-z>.
- Günther T, Coop G. Robust identification of local adaptation from allele frequencies. *Genetics*. 2013;195:205–220. <https://doi.org/10.1534/genetics.113.152462>.

- Han E, Sinsheimer JS, Novembre J. Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol Biol Evol*. 2014;31:723–735. <https://doi.org/10.1093/molbev/mst229>.
- Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P. Using genomics to characterize evolutionary potential for conservation of wild populations. *Evol Appl*. 2014;7:1008–1025. <https://doi.org/10.1111/eva.12149>.
- Hilborn R, Mangel M. *The ecological detective: confronting models with data*. Princeton University Press; 2013.
- Hoekstra HE, Drumm KE, Nachman MW. Ecological genetics of adaptive color polymorphism in pocket mice: geographic variation in selected and neutral genes. *Evolution*. 2004;58:1329–1341. <https://doi.org/10.1111/j.0014-3820.2004.tb01711.x>.
- Hooten MB, Hefley TJ. *Bringing Bayesian models to life*. CRC Press; 2019.
- Hooten MB, Hobbs NT. A guide to Bayesian model selection for ecologists. *Ecol Monogr*. 2015;85:3–28. <https://doi.org/10.1890/14-0661.1>.
- Iverson EN. Conservation mitonuclear replacement: facilitated mitochondrial adaptation for a changing world. *Evol Appl*. 2024;17:e13642. <https://doi.org/10.1111/eva.13642>.
- Kim SY et al. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*. 2011;12:231. <https://doi.org/10.1186/1471-2105-12-231>.
- Kimura M. *The neutral theory of molecular evolution*. Cambridge University Press; 1983.
- Kirkpatrick M, Barton NH. Evolution of a species' range. *Am Nat*. 1997;150:1–23. <https://doi.org/10.1086/286054>.
- Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*. 2014;15:356. <https://doi.org/10.1186/s12859-014-0356-4>.
- Korolev KS, Avlund M, Hallatschek O, Nelson DR. Genetic demixing and evolution in linear stepping stone models. *Rev Mod Phys*. 2010;82:1691–1718. <https://doi.org/10.1103/RevModPhys.82.1691>.
- Kosch TA et al. Genetic approaches for increasing fitness in endangered species. *Trends Ecol Evol*. 2022;37:332–345. <https://doi.org/10.1016/j.tree.2021.12.003>.
- Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res*. 2011;21:940–951. <https://doi.org/10.1101/gr.117259.110>.
- Lotterhos KE. The paradox of adaptive trait clines with nonclinal patterns in the underlying genes. *Proc Natl Acad Sci U S A*. 2023;120:e2220313120. <https://doi.org/10.1073/pnas.2220313120>.
- Lou RN, Jacobs A, Wilder AP, Therikildsen NO. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol Ecol*. 2021;30:5966–5993. <https://doi.org/10.1111/mec.16077>.
- MacDougall-Shackleton SA, Johnson RE, Hahn TP. Gray-crowned rosy-finch (*Leucosticte tephrocotis*), version 1.0. In: Gill FB, Poole AF, editors. *Birds of the world*. Cornell Lab of Ornithology; 2000. <https://doi.org/10.2173/bow.gcrfin.01>.
- Mayr E. *Systematics and the origin of species*. Columbia University Press; 1942.
- McRae BH. Isolation by resistance. *Evolution*. 2006;60:1551–1561. <https://doi.org/10.1111/j.0014-3820.2006.tb00500.x>.
- McVean G. A genealogical interpretation of principal components analysis. *PLoS Genet*. 2009;5:e1000686. <https://doi.org/10.1371/journal.pgen.1000686>.
- Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. *J Am Stat Assoc*. 1988;83:1023–1032. <https://doi.org/10.1080/01621459.1988.10478694>.
- 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:e37558. <https://doi.org/10.1371/journal.pone.0037558>.
- Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*. 2011;12:443–451. <https://doi.org/10.1038/nrg2986>.
- Ortega JM, Rheinboldt WC. *Iterative solution of nonlinear equations in several variables*. Academic Press; 2000.
- Pei J, Pan X, Wei G, Hua Y. Research progress of glutathione peroxidase family (GPX) in redox regulation. *Front Pharmacol*. 2023;14:1147414. <https://doi.org/10.3389/fphar.2023.1147414>.
- Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat Genet*. 2016;48:94–100. <https://doi.org/10.1038/ng.3464>.
- Phan D, Pradhan N, Jankowiak M. 2019. Composable effects for flexible and accelerated probabilistic programming in NumPyro. arXiv:1912.11554. <http://arxiv.org/abs/1912.11554v1>.
- Phelps MP, Seeb LW, Seeb JE. Transforming ecology and conservation biology through genome editing. *Conserv Biol*. 2020;34:54–65. <https://doi.org/10.1111/cobi.13292>.
- Piironen J, Vehtari A. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electron J Stat*. 2017;11:5018–5051. <https://doi.org/10.1214/17-EJS1337SI>.
- Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R. A practical guide to environmental association analysis in landscape genomics. *Mol Ecol*. 2015;24:4348–4370. <https://doi.org/10.1111/mec.13322>.
- Rensch B. Some problems of geographical variation and species-formation. *Proc Linn Soc Lond*. 1938;150:275–285. <https://doi.org/10.1111/j.1095-8312.1938.tb00182k.x>.
- Rice SH. *Evolutionary theory: mathematical and conceptual foundations*. Sinauer Associates; 2004.
- Rong H et al. Association between regulator of telomere elongation helicase1 (RTEL1) gene and HAPE risk: a case-control study. *Medicine (Baltimore)*. 2017;96:e8222. <https://doi.org/10.1097/MD.00000000000008222>.
- Schloss CA, Nuñez TA, Lawler JJ. Dispersal will limit ability of mammals to track climate change in the Western Hemisphere. *Proc Natl Acad Sci U S A*. 2012;109:8606–8611. <https://doi.org/10.1073/pnas.1116791109>.
- Schwob MR, Hooten MB, Narasimhan V. Composite dyadic models for spatio-temporal data. *Biometrics*. 2024;80:ujae107. <https://doi.org/10.1093/biomtc/ujae107>.
- Simonson TS, McClain DA, Jorde LB, Prchal JT. Genetic determinants of Tibetan high-altitude adaptation. *Hum Genet*. 2012;131:527–533. <https://doi.org/10.1007/s00439-011-1109-3>.
- Slatkin M. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*. 1993;47:264–279. <https://doi.org/10.1111/j.1558-5646.1993.tb01215.x>.
- Wakeley JH. *Coalescent theory: an introduction*. Roberts and Company Publishers; 2009.

- Wan Y *et al.* Characterization of CRH-binding protein (CRHBP) in chickens: molecular cloning, tissue distribution and investigation of its role as a negative feedback regulator within the hypothalamus–pituitary–adrenal axis. *Genes (Basel)*. 2022;13:1680. <https://doi.org/10.3390/genes13101680>.
- Wang D *et al.* Hypoxia-inducible factor 1alpha enhances RANKL-induced osteoclast differentiation by upregulating the MAPK pathway. *Ann Transl Med*. 2022a;10:1227. <https://doi.org/10.21037/atm-22-4603>.
- Wang J, Yu J, Lipka AE, Zhang Z. Interpretation of Manhattan plots and other outputs of genome-wide association studies. In: Torkamaneh D, Belzile F, editors. *Genome-wide association studies*. Springer; 2022b. p. 63–80.
- Wikle CK. Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*. 2003;84:1382–1394. [https://doi.org/10.1890/0012-9658\(2003\)084\[1382:HBMFPT\]2.0.CO;2](https://doi.org/10.1890/0012-9658(2003)084[1382:HBMFPT]2.0.CO;2).
- Wright S. Evolution in Mendelian populations. *Genetics*. 1931;16:97–159. <https://doi.org/10.1093/genetics/16.2.97>.
- Yi X *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*. 2010;329:75–78. <https://doi.org/10.1126/science.1190371>.
- Zhuang Z-X, Chen S-E, Chen C-F, Lin E-C, Huang S-Y. Genome-wide association study on the body temperature changes of a broiler-type strain Taiwan country chickens under acute heat stress. *J Therm Biol*. 2019;82:33–42. <https://doi.org/10.1016/j.jtherbio.2019.03.007>.