A comparison of genomic forecasts based on genotypes versus allele frequencies

Brandon M. Lind*, Katie E. Lotterhos

Department of Marine and Environmental Sciences
Northeastern University Marine Science Center
430 Nahant Road, Nahant, MA 01908, USA.10 July 2025

10 July 2025

Running Title: *Population- and individual-level genomic offsets*

Keywords: genomic forecasting, genomic offset, random forest, climate change, genotypes, allele frequencies

***Corresponding Author**
Email: lind.brandon.m@gmail.com

Brandon M. Lind https://orcid.org/0000-0002-8560-5417
Katie E. Lotterhos https://orcid.org/0000-0001-7529-2771

Abstract

Accelerating land use and climate change threaten to disrupt relationships between adaptive variation and environmental optima of many species. Consequently, management must increasingly identify non-local genetic sources for restoration programs. Genomic offset methods, like gradientForests, have shown promise in identifying these sources using genomic data, potentially bypassing the need for traditional, time-consuming transplant experiments. However, previous studies primarily used population-level allele frequencies (AF) for training and population-mean fitness for evaluation, ignoring individual variation within populations. Here, we used simulation data to compare the accuracy of genotype- and AF-based models, factorially evaluated using both individual and population-mean fitness. With over 810,000 evaluations of such models, we found that the number of loci had little impact on model performance. As expected, population-level evaluation provided an optimistic view of predictive performance for both genomic inputs. While genotype- and AF-based models showed similar qualitative and quantitative aspects, genotype-based models improved predictions in landscapes that differed from strict environmental clines by incorporating additional loci beyond those used by AF-based models. This suggests genotype-based models may enhance offset predictions in environments that are discontinuous and have multiple populations in geographically distant yet similar environments. We close with recommendations for future use and evaluation of these tools.

1 | Introduction

Species regularly inhabit environmentally heterogenous ranges. Natural selection pressures often differ spatially across these habitats, and can lead to genotype-environment interactions that affect fitness (i.e., survival and reproduction). When the strength of local selection pressures overcomes the forces of gene flow and drift (Blanquart et al. 2012), local adaptation can evolve in the metapopulation. Local adaptation is defined by a higher mean fitness of populations in sympatry than in allopatry (Kawecki and Ebert 2004; Blanquart et al. 2013). Indeed, local adaptation is common across many species, such as plants(Leimu and Fischer 2008; Fournier-Level et al. 2011; Leites and Garzón 2023), insects (Arguello et al. 2016), fish (Fraser et al. 2011), corals (Thomas et al. 2022), and marine invertebrates (Burford et al. 2014; Bible and Sanford 2016).

Rapid environmental change, such as that projected for future climate, threatens to decouple locally adaptive genetic variation from local fitness optima for many species (Aitken et al. 2008; Hoffmann and Sgrò 2012). Previous methods used to understand the impact of environmental change on species, such as species distribution models (SDMs), model an organism's ecological niche to make predictions about the distribution and demographic performance of populations across space (Guisan and Thuiller 2005; Thuiller et al. 2008; Pacifici et al. 2015). However, these models ignore intraspecific variation and patterns of local adaptation (Capblancq et al. 2020). While these models have been found to have some predictive ability relative to species occurrence, their performance is often

poor for predicting fitness of populations within their native range from independent, ground-truth data (Lee-Yaw et al. 2022).

Over the past decade, ecological forecasting methods called genomic offsets have become increasingly popular as an alternative, or complement (e.g., Chen et al. 2022), to SDMs. These genomic offset methods model the relationship between intraspecific variation from landscape genomic data and environmental variables across populations, and then use this relationship to predict the extent of maladaptation of populations under environmental change (Capblancq et al. 2020). Genomic offsets have been interpreted as the degree of genetic change necessary for a population to maintain past relationships between genetic variation and optimal environments (Rellstab et al. 2021). Whether these genomic offset measures are well suited for inferring how the fitness of populations will differ between current and future climates on the landscape, or for inferring fitness differences among genotypes for a specific environment, has recently been debated (Lotterhos 2024*a*).

Regardless of debate around the technical interpretation of genomic offsets, investigators generally agree that the performance of their predictions can be evaluated as the relationship between genomic offsets and empirical estimates of fitness measured in an experimental context of different genotypes being moved into different environments (Lotterhos 2024*b*). There is also general agreement that if genomic offsets are predictive of genotype responses, there should be a strong negative relationship between genomic offsets and fitness proxies in such experiments (Capblancq et al. 2020; Rellstab et al. 2021;

Lotterhos 2024*b*). We refer to an experiment designed to test for this negative relationship between genomic offsets and fitness proxies as an "evaluation".

Genomic offset studies have been applied across myriad taxa, including birds (Bay et al. 2018; Ruegg et al. 2018; Chen et al. 2022; DeSaix et al. 2022), fish (Brauer et al. 2023), herbaceous plants (Exposito-Alonso et al. 2019), mammals (Rivkin et al. 2024), and trees (Fitzpatrick and Keller 2015; Capblancq and Forester 2021; Fitzpatrick et al. 2021; Gougherty et al. 2021; Gugger et al. 2021; Lachmuth et al. 2023, 2024; Lind et al. 2024). However, despite the increase in popularity of these methods, there have been only a few empirical evaluations of their predictive performance using experimental measures relating to fitness (but see Rhoné et al. 2020; Capblancq and Forester 2021; Fitzpatrick et al. 2021; Gain et al. 2023; Lachmuth et al. 2023; Lind et al. 2024).

While there are now several methods used to predict genomic offsets (see Table 1 in Capblancq et al. 2020), the gradientForests method (sensu Fitzpatrick and Keller 2015) has undergone the most evaluation. For instance, recent *in situ* evaluations of genomic offset from gradientForests (GF_{offset}) in locally adapted tree species have often found the expected negative relationship between fitness-related phenotypes from different populations transplanted to a common garden (e.g., Fitzpatrick et al. 2021; Lachmuth et al. 2023, 2024; Lind et al. 2024). Further, these studies have shown that predictions from GF_{offset} are often better than geographic or climate distance alone (Fitzpatrick et al. 2021; Lind et al. 2024). *In silico* evaluations based on simulation data have also shown that GF_{offset} generally outperformed other genomic offset methods (e.g., Rellstab et al. 2016;

96 Capblancq and Forester 2021; Gain and François 2021) across a range of scenarios (Lind
97 and Lotterhos 2024), or gives similar predictive performance to simple measures of
98 environmental distance that use only causal environmental variables (Láruson et al. 2022).
99 Overall, several genomic offset methods, including GFoffset, have been shown to perform
100 best when there is a high degree of local adaptation in the metapopulation (Lind and
101 Lotterhos 2024). These evaluations of simulation data have generally shown high levels of
102 predictive performance of GF_{offset} within contemporary environments (Láruson et al. 2022;
103 Lind and Lotterhos 2024), but not when populations are moved into novel climates (Lind
104 et al. 2024).

105 Despite these advances, the domain of applicability, or the circumstances under which
106 model predictions are valid (Lotterhos et al. 2022), are still limited for genomic offset
107 measures, further limiting confidence in their application more broadly. These
108 circumstances ultimately encompass the experimental design (e.g., the collection and
109 format of training and evaluation data, the choice of analyses) and the evolutionary
110 history of the targeted populations (e.g., spatial patterns of selection, genomic
111 architecture). To date, the majority of genomic offset predictions and evaluations have
112 made use of population-level allele frequencies in modeling, and have made predictions at
113 the level of local populations for evaluation. On the other hand, many management
114 applications would benefit from genomic offset methods that make predictions at the
115 individual level, yet to date these methods remain largely undeveloped. Individual-level
116 predictions are particularly important when the species has a relatively small census size,

117 requiring offset strategies tailored to existing individuals or when limited management
118 resources or the carrying capacity of targeted restoration sites restrict the number of
119 individuals that can be moved or transplanted, making it essential to select individuals
120 with the highest fitness. Individual-level predictions are also relevant when variation in
121 fitness is high relative to the population average. In contrast, population-level predictions
122 are most applicable when population sizes are large, when the goal is to assess offsets to
123 future climate change on a regional scale, or when individuals for future translocation or
124 transplantation will be randomly selected from a larger population. Population-level
125 predictions also present a cost-saving benefit, for example when using pool-seq data. There
126 are conceptual and statistical considerations regarding how to fairly compare the
127 performance of population-level vs. individual-level evaluations, which we discuss in more
128 detail below.

129 ## *Considerations for model inputs*

130 Although the format of the genomic data that is input into models used for calculating
131 genomic offsets has typically been allele frequencies, many models could also take
132 individual genotypes as input (e.g., gradientForests; redundancy analysis, sensu
133 Capblancq and Forester 2021; and genetic.gap, sensu Gain et al. 2023). A comparison of
134 allele frequency vs. genotype inputs is shown in Fig. S1. We consider the case where
135 individual genotypes collected from the same population are assigned the same
136 environmental data for that location, which would be common when collecting groups of

137 individuals, each from a small geographic area, and environmental data is not available
138 at the spatial scale of individual collections. This would also be the case when the exact
139 locations of individuals are unknown, such as with bulk seed lots collected for
140 reforestation. In contrast, individual-level environmental data could also be used to train
141 the model, which would be common when collecting individuals from a larger geographic
142 area, and environmental data is available for each individual. The important distinction
143 between allele-frequency vs. genotype inputs is that population-level allele frequencies
144 average over within-population variation and may miss key aspects of diversity that could
145 lead to more accurate predictions. For example, recent modeling efforts have revealed that
146 adaptive trait clines can evolve even when the underlying relationships between the
147 frequency of adaptive alleles and the selective environment are not monotonic (Lotterhos,
148 2023a). Individual-level genotype data may therefore offer an advantage in these scenarios.
149 Thus, it is unclear if training models using genotypes from individuals can improve
150 predictions relative to predictions from models that use allele frequencies as input.
151 Similarly, population-level environmental data can introduce unnecessary noise by
152 inaccurately assigning environmental values to individuals, particularly if individuals from
153 a population are sampled over a large geographic area.

154 On the other hand, there are also computational requirements to consider with regard
155 to the format of model inputs. For instance, because genomic offset models are trained
156 using two sources of data (genetic data and environmental data), the size of the dataset
157 (determined by the number of individuals or populations, as well as the number of loci

158 and environmental variables provided for training) will likely affect the computational
159 memory or runtime required, which may limit potential avenues of inference when
160 computational resources are scarce. Individual-level genotype and environmental data will
161 always be a much larger dataset than population-level allele frequency and environmental
162 data, which will impact computational requirements.

163 ### *Considerations for model outputs*

164 From the input training data, the relationship between changes in genomic composition
165 and the multivariate environment is modeled (see Fig. 2 in Rellstab et al. 2016; and Fig.
166 3 in Capblancq et al. 2020 for conceptual illustrations). This model can then be used to
167 estimate the genomic composition for a hypothetical population in a given environment,
168 for instance the genomic composition of a population in its current environment, or its
169 composition in a potential future environment. A genomic offset is the magnitude of the
170 difference between these estimates, and thus uses the current and future environmental
171 values as input into the equation (Box 1, Eq. B1). Therefore, to generate offset predictions
172 that would be different among individuals in GF, one would need to use unique current
173 environmental values as input.

174 When individuals collected from the same population are assigned the same
175 environmental data for that location as model input, the output of model predictions is
176 at the population level no matter the format of input genomic data (genotypes vs. allele
177 frequencies). In other words, any genomic offset method that relies exclusively on

178 environmental data in its calculation of offset (Box 1) will yield the same genomic offset
179 value for any individuals transitioning from the same current environment to the same
180 new environment, irrespective of the model inputs or their genetic compositions. While
181 Box 1 focuses on GF_{offset}, predictions are also restricted to the population level for other
182 genomic offset methods using population-level inputs as well (e.g., genomic offset from
183 redundancy analysis, Capblancq and Forester 2021; and from genetic.offset and
184 genetic.gap, Gain and François 2021; Gain et al. 2023). We consider this consequence of
185 population-level offset predictions in the next section.

186 ## *Considerations for model evaluation*

187 While the format of input genomic data may affect model accuracy, the level of
188 evaluation may affect the inferred accuracy as well (Box 2, Fig. B2). For instance,
189 evaluating genomic offset predictions using population-mean fitness proxies will inflate
190 performance scores relative to evaluation using individual fitness proxies, particularly
191 when populations exhibit high variation in individual fitness. This happens because the
192 residual error in the evaluation statistic (i.e., error in the relationship between a genomic
193 offset and a fitness proxy) is reduced when the fitness proxies are averaged across
194 individuals within a population. As a consequence, evaluations at the population level
195 may provide an overly optimistic view of predictive performance. Thus, our study is
196 careful to compare genomic offset models that were evaluated at the same level (ie.,
197 comparisons that are made within rows of Box 2).

Study Questions

Here, we explore the implications of allele frequency- and genotype-based genomic offset models. Previously, Lind & Lotterhos (2024) found GF_{offset} to be particularly promising due to its consistently high performance across many of the scenarios that were evaluated. Using a subset of the same simulation scenarios evaluated by Lind & Lotterhos (2024), we train new models to investigate potential trade-offs between population- (e.g. allele frequency) and individual-level (e.g., genotype) inputs by comparing predictive performance and computational requirements for genomic offsets estimated from gradientForests. We pose the following five questions: Q1) How does the number of markers used as input affect performance? (We address this question first so that following questions could be addressed with a constant number of markers). Q2) How does the format of evaluation data affect performance? Q3) How does the format of the genetic training data affect performance? Q4) How does the format of the environmental training data affect performance? Q5) How does the size of the dataset affect computational time and memory requirements?

2 | Methods

2.1 / Explanation of input training data for genomic offset models

We used a subset of the simulation datasets previously described and evaluated by Lotterhos (2023a) and Lind & Lotterhos (2024). The first set of simulations were conducted on a 10 x 10 heterogeneous landscape of spatially discrete demes, such that all

individuals from a deme experienced the same environmental values and selection pressures. Here forward, we will refer to demes as populations for the purposes of analysis, but recognize that they are not demographically or evolutionarily independent units. We use these datasets to answer Q1, Q2, Q3, and Q5 (see Explanation of Questions). Specifically, within each simulated dataset, a Wright-Fisher metapopulation of 100 populations adapted to a heterogeneous landscape, where aspects of the evolutionary history varied across simulation levels. Specifically, the subset of simulation levels used here were only those datasets where metapopulations adapted to two environmental variables on the landscape (nlevels = 180/225, using three replicates per simulation level). The levels included: three landscapes that varied the geographic distribution of environmental variables (Figure 1), five demographic scenarios that varied patterns of gene flow and population sizes, three genic levels that varied the number of loci responding to selection (i.e., spanning oligogenic to polygenic), and four pleiotropy levels, the degree of pleiotropy from causal mutations, and different levels for the strength of selection from the causal environments (see Fig. 1 from Lotterhos 2023*a*). Loci were simulated on 20 independent linkage groups, 10 of which allowed mutations with effects on fitness. The univariate effect size of a QTN evolving without pleiotropy was drawn from a normal distribution, and the bivariate effect size of QTNs under two traits with pleiotropy was drawn from a multivariate normal distribution. Neutral mutations were added to all 20 linkage groups with tree sequencing (Lotterhos 2023*a*). The scaled metapopulation-level recombination rate ($\rho = 0.01$) resulted in a resolution of 0.001cM between proximate bases

where each linkage group had a total length of 5cM. This resolution is common in SNP array or SNP chip designs, where loci are sampled broadly across the genome.

The causal environments imposing selection pressure included a temperature-like variable (*temp*) that created a north-to-south cline on all landscapes (top panels Fig. 1), and a second clinal environmental variable (*Env2*) that represented different analogies depending on the landscape - in the *Stepping Stones - Clines* landscape, *Env2* formed a longitudinal cline; in the *Stepping Stones - Mountains* landscape, *Env2* was analogous to elevation; in the *Estuary - Clines* landscape, *Env2* was analogous to salinity gradients within coastal inlets that were only connected by the outer marine environment (top panels Fig. 1). For more information on simulations see Lotterhos (2023*a*) and Lind & Lotterhos (2024).

Ten individuals from each of the 100 populations were sampled for genetic data (1000 total). For the evaluations here, we used a modified version of the *Adaptive Environment* workflow from Lind & Lotterhos (2024). For an illustration of this workflow, see Fig. S2. All replicates were evaluated using models trained with all adaptive environments. Further, while Lind & Lotterhos (2024) used allele frequency (AF) data as input (i.e., allele frequencies calculated across individual genotypes), here we also used the individual-level genotype data as input to compare to AF-based models.

For each simulation replicate, genotypes of individuals were encoded as counts of the derived allele, and derived allele frequencies from individuals were used for input. Loci with minor allele frequency < 0.01 were removed. After identifying a set of loci for a given

replicate, both population- and individual-level data were filtered for the same markers and used as input for training GF_{offset} models (hereafter GF_{AF} and GF_{geno}, respectively, to distinguish from evaluation workflows in Table 1).

The genotypes within the dataframe passed to the gradientForest function were encoded as integer class to ensure random forest regression instead of random forest classification. Regression is more appropriate given the additive nature of the alleles underlying local adaptation simulated here, and to enable direct comparison to models using allele frequencies. While Lind & Lotterhos (2024) varied marker choice (e.g., *adaptive*, *neutral*, or *all* markers for input), we vary only the number of markers randomly chosen from the full set of loci (see Q1 below).

Models were evaluated using *in silico* common gardens in each of 100 populations on the landscape. For each of these 100 common gardens , model performance was quantified as the rank correlation (Kendall's $\boldsymbol{\tau}$) between (i) the projected genomic offset to the common garden and (ii) the known fitness (of either individuals or populations) in the garden environment. The format of the evaluation data (i.e., population- or individual-level fitness) depended upon the workflow (Table 1). If greater magnitudes of estimated genomic offset indicate higher degrees of maladaptation, than a well-performing model would result in a negative relationship between genomic offsets and fitness values. We found that the relationship between offset and fitness, as well as offset and log(fitness), was non-linear for our data (Figs. S3-S7), so we chose the rank correlation Kendall's $\boldsymbol{\tau}$ to

assess model performance. Kendall's $\boldsymbol{\tau}$ is particularly well suited for rank correlations when there are ties among either univariate or bivariate ranks.

In addition to the spatially discrete simulations described above, we evaluated offset predictions using data from a single simulation of a non-Wright-Fisher metapopulation undergoing range expansion from three refugia. We use this data to answer Q1 and Q4 (see Explanation of Questions). Unlike the previous simulations described above, this model was spatially continuous, with individuals occupying distinct locations across the landscape and evolving variable degrees of admixture. Selection pressures on six moderately polygenic traits with unconstrained pleiotropy were driven by six environmental variables derived from real-world bioclimatic data from western Canada (bottom panel Fig. 1). Markers were distributed across linkage groups in a similar fashion to genomes simulated for spatially discrete populations. From this simulation, we sampled 1,000 individuals across the landscape. Using individual-level genotypes, we formatted the resulting data similarly to the data from the spatially discrete simulations described earlier, but the environmental data used as training input was formatted in two distinct ways: i) individual-level environmental data, 2) population-averaged environmental data. These two formats were then used to examine the impact of the format of environmental data used for training (Q4) when trained using individual genotypes. To allow for comparison to the spatially discrete evaluations, evaluation of these spatially continuous datasets took place in 100 common gardens where the environment was the population-

mean environment of the assigned individuals. For more information on coding workflows implemented here, see Supplemental Note S1.

2.2 / Explanation of questions

Q1 / How does the number of markers used as input affect performance?

Modern SNP datasets often have millions of markers distributed across the genome. However, despite the promise of GF_{offset} performance, the current software implementation is often computationally intensive and requires resources (e.g., allocated memory) beyond those commonly available outside of high-performance computing clusters. This computational burden will often limit implementations of GF_{offset} to those using fewer than around 20,000 loci, even on systems with plentiful resources. Even so, previous evaluations of GF_{offset} have shown that random markers often perform on par with adaptive marker sets (Fitzpatrick et al. 2021; Láruson et al. 2022; Lachmuth et al. 2023; Lind and Lotterhos 2024). Still, in some cases the adaptive (candidate) markers in empirical datasets are limited to several hundred markers, and it is not known how well these small marker sets will capture genome-wide patterns of population history. To understand how the number of loci can impact performance, we chose random sets of loci (allowing potential overlap among sets) in the following sample sizes from each replicate from both the spatially discrete as well as spatially continuous simulations: 500, 5 000, 10 000, 20 000. Comparisons of performance across marker set sizes were made within AF- and genotype-based approaches.

320 To further understand the impact of loci that are incorporated into GF models, we
321 used spatially discrete simulations to evaluate distributions of locus-specific R^2 - a measure
322 of the goodness of fit of predicting genetic data using environmental values - estimated
323 from internal random forest models for each locus. To be incorporated into a GF model,
324 the random forest model must have explanatory value (i.e., $R^2 > 0$). R^2 values of loci are
325 used to configure model weights used by GF (Ellis et al. 2012; Smith et al. 2012). We
326 compare distributions of R^2 between *adaptive* loci (i.e., loci with positive effects on fitness)
327 and between two classes of neutral loci. The first class are neutral loci on linkage groups
328 without adaptive loci (hereafter, *neutral* loci). The second class are neutral loci that are
329 on the same linkage groups as *adaptive* loci (hereafter, *neutral-linked* loci). We also
330 compare how distributions of R^2 are impacted by evolutionary history (i.e., simulation
331 parameter settings).

332 In total, there were 216,000 potential evaluations per spatially discrete workflow (4
333 marker sets * 180 levels * 3 replicates per level * 100 common gardens per replicate; Table
334 1). However, 269 replicates that used 20 000 input markers encoded as genotypes failed
335 to complete in less than one day or using less than 250Gb of memory. Therefore, unless
336 otherwise noted, the remaining questions used the 189100 evaluations common to both
337 AF- and genotype-based models.

338 *Q2* / How does the format of evaluation data affect performance?

339 Models of GF_{offset} are restricted to predictions at the environmental level (Box 1). As a
340 likely consequence, previous implementations of genomic offset models have most often
341 used population-mean measures of fitness-related phenotypes to evaluate model
342 predictions. However, using population-mean phenotypes for evaluation may present an
343 overly optimistic performance of these models because individuals sampled within
344 populations are likely to have variation in fitness and may perform better or worse when
345 transplanted to new environments.

346 To illustrate this point, we compared evaluations using population-level fitness data to
347 individual-level fitness data from each in silico common garden experiment, for each set
348 of genomic offset predictions from a single model (either calculated from allele frequency
349 or genotype input data). If performance is greater when evaluated at the population level
350 compared to when performance is evaluated at the individual level for the exact same
351 model, this indicates that population-level evaluation presents an overly optimistic view
352 of model performance than could likely be expected in management practice. To address
353 this question, we used the spatially discrete simulation scenarios.

354 *Q3* / How does the format of the genetic training data affect model performance?

355 The effect of AF- and genotype-based inputs on the performance of genomic offset
356 models is not well understood. To compare the performance of different model inputs on
357 a common ground, we used datasets from spatially discrete simulations to compare AF-

and genotype-based models that were evaluated using population-mean fitness (e.g., bottom row in Figure B2).

Q4 / How does the format of the environmental training data affect model performance?

To understand the impact of the format of environmental data used for training, we created two environmental datasets from the spatially continuous simulation for model training input. For the first case, we created a dataset using individual-level environmental data corresponding to the specific locations of the sampled individuals on the landscape. For the second, we assigned individuals to one of 100 populations using an evenly spaced 10 x 10 grid system and assigned all individuals from a population the population-averaged environmental data (all individuals within a single grid were given the same environmental value). For both of these types of models, we used genotypes as training input and individual fitness for evaluation (Table 1).

Q5 / How does the size of the dataset affect computational time and memory requirements?

The analytical logistics of training and evaluating genomic offset models is computationally burdensome beyond a single model, because often the input data should be varied to understand the sensitivity of predictive outcomes. This input data ranges from the populations or markers used to the uncertainty inherent in future climate projections, or even the climate data used for training (DeSaix et al. 2022; Lachmuth et al. 2023; Lind et al. 2024). Because of this, it is necessary to train potentially dozens of

379 models for comparison. Benchmarking the time and memory requirements necessary for
380 these models will therefore be important during the planning and training execution of
381 such models in future settings. We compared walltime and memory usage from all training
382 phases from spatially discrete simulations and compared how the number of genetic
383 sources (individuals or populations) and number of loci provided to the model for training
384 affected resource usage. We obtained resource information using a custom parallel
385 implementation of the Slurm seff command in python. Training jobs were generally run
386 on Intel Xeon processors, ranging from 2.0GHz - 2.8GHz. More information on compute
387 node processors used for training GF models is available in Supplemental Note S2.

388 # 3 | Results

389 Nearly all targeted simulation replicates were successfully processed through training
390 and prediction phases. However, 49.81% (269/540) of the spatially discrete replicates using
391 individual genotypes from 20000 loci were unable to be trained using less than 250Gb of
392 requested memory and one day of requested run time. Of the 269 datasets that failed to
393 complete training, 12 were due to exceeding run time requests. Given similarity in
394 performance (see Q1 Results), we considered these 269 instances as having failed and did
395 not attempt to run with increased resource requests. Of the datasets that failed, all were
396 from either the *Stepping Stones* - *Clines* landscape (75.6% failure) or the *Estuary* - *Clines*
397 landscape (73.9% failure; SC 05.06). All other replicates across loci sets and genetic sources
398 were able to complete training and prediction phases.

399 *Q1 / How does the number of markers used as input affect performance?*

400 The number of markers used to train both AF- and genotype-based models of GF_{offset}
401 from spatially discrete datasets did not differentially impact performance, as all
402 comparisons within levels had strong linear correlations (Pearson's $r > 0.9960$; Fig. 2; Fig.
403 S8). Across both GF_{AF} and GF_{geno} implementations, comparisons between models that
404 were trained with more than 5000 loci had the strongest relationship, while comparisons
405 to models trained with 500 loci had strong, albeit relatively weaker, relationships
406 (Pearson's $r < 0.9981$, Fig. 2, Fig. S8). We found similar results when comparing models
407 from the continuous space simulation (Fig. S9). While the number of loci had little effect
408 on model performance, we further explored results from our spatially discrete evaluations
409 to understand why this was the case, if there were any differences in the way in which
410 GF treated loci internally within its modeling framework, and if any differences were due
411 to experimental design or evolutionary history of the targeted populations.

412 Q1a - Why do marker sets perform similarly?

413 To understand why marker sets of 500 loci performed similarly to models trained with
414 10000 markers, we ran principal component analysis (PCA) on each marker set for each
415 replicate from spatially discrete datasets using either genotypes or allele frequencies using
416 all loci from the entire dataset. For each replicate and genetic source, we then calculated
417 absolute correlations - abs(Pearson's r) - between axis loadings from the first three PC
418 axes estimated from either 500 or 10 000 loci. We generally found a strong, linear
419 relationship between the first three PC axes loadings when comparing corresponding axes

between the two marker sets (Fig. S10; e.g., between the first PC from each dataset) demonstrating that both higher- and lower-order axes of genetic variation in the simulations were consistently captured by marker sets as small as several hundred loci (Supplemental Code 05.01).

Q1b - How many loci were used by GF models?

While our marker set categories are indicative of the number of loci that were provided to GF for training, this does not necessarily indicate that the same number of loci were used by the model, as GF only incorporates loci with positive R^2 values calculated from internal random forest models. Nevertheless, for the spatially discrete datasets we found that the number of loci incorporated into GF models increased with the number of loci provided to GF for training (Fig. S11).

We also found that the number and identity of loci used by these GF models depended on the landscape. Both GF_{AF} and GF_{geno} models trained using datasets from the *Stepping Stones - Clines* and *Estuary - Clines* landscapes incorporated very similar sets of loci (Fig. 3). In contrast, however, GF_{geno} models used many more loci from *Stepping Stones - Mountain* landscape datasets than GF_{AF} models (Fig. 3).

Q1c - Is the explanatory value of loci affected by experimental design or evolutionary history?

To further understand the effect of input loci on model performance, we explored locus-specific R^2 values from random forest models internal to GF using the spatially discrete datasets. R2 is a measure of the goodness of fit of predicting genetic data using

environmental values that is estimated from internal random forest models for each locus. Overall, the R^2 values assigned to loci by either GF_{geno} or GF_{AF} were positively correlated (Fig. S12). Even so, for these loci that overlapped between GF_{AF} and GF_{geno} models, R^2 was generally lower for loci incorporated into GF_{geno} models than GF_{AF} models (Fig. S12, Fig. S13). The additional loci used only by GF_{geno} models (but not by GF_{AF} models) generally had lower R^2 than the loci that overlapped between GF_{geno} and GF_{AF} (Fig. S14).

We examined how the evolutionary history of sampled populations (i.e., simulation parameters) interacted with R^2 values. Generally, these parameters had little effect on the distribution of R^2 values but there were exceptions. For instance, distributions of R^2 were higher for *adaptive* (i.e., causal) than for neutral (i.e., unlinked to causal) and neutral-linked (i.e., linked to causal) loci when the genetic architecture underlying adaptation was oligogenic (Fig. S15). However, these differences in R2 among the marker sets were less pronounced when the architecture was moderately or highly polygenic. Contrastingly, other parameters had little effect on differences in R^2 between *adaptive* and neutral loci classes, including the effect of landscape (Fig. S16).

Despite little differences between *adaptive* and both classes of neutral markers within landscapes, distributions of R^2 across all loci differed to greater degrees among landscapes (Figs. S13, S17). For instance, median R^2 was greatest from loci in *Estuary* - *Clines* landscapes, followed by *Stepping Stone* - *Clines* and next by *Stepping Stone* - *Mountain* landscapes. While there were landscape differences between distributions of R^2, these patterns were not indicative of overall performance across landscapes. For instance,

462 performance was highest in *Stepping Stone - Clines* landscapes, followed by *Stepping*
463 *Stone - Mountain* and *Estuary - Clines* landscapes (compare Fig. S17 with Fig. S8). This
464 overall pattern of GF_{AF} performance across landscapes has been shown previously using
465 similar datasets from the same simulations (Lind & Lotterhos, 2024) though R^2 was not
466 previously compared.

467 *Q2* / How does the format of evaluation data affect performance?

468 We hypothesized that models evaluated using population-level fitness would have higher
469 performance scores relative to models evaluated using individual-level fitness. This was
470 indeed the case, as can be seen by comparing performance from genotype-based predictions
471 evaluated using fitness at either the individual level ($GO_{\mathrm{geno,ind}}$) or using average fitness
472 at the population level ($GO_{\mathrm{geno,pop}}$; Fig. 4) from spatially discrete datasets.

473 Similar to genotype-based models, comparison of AF-based models showed elevated
474 performance scores when evaluated using population mean fitness ($GO_{\mathrm{AF,pop}}$) compared
475 to predictions from the same models evaluated at the individual level ($GO_{\mathrm{AF,ind}}$; Fig. 4B).
476 This statistical artifact is taken into account in the next section, where we compare
477 performance from evaluations at the population level.

478 *Q3* / How does the format of the genetic training data affect model performance?

479 To understand if within-population diversity captured by individual genotypes
480 improved model performance, we compared performance from models trained with
481 genotypes to models trained using allele frequencies at the same evaluation level

(population-mean fitness). The majority of models saw similar performance, but there were datasets that saw improved performance from genotype inputs in spatially discrete datasets (Fig. 5). These were mainly from *Stepping Stone - Mountain* landscapes where multiple populations on the landscape inhabit the same (but geographically distinct) multivariate environment (top middle panel Fig. 1).

As an example to show the improvement of population-level prediction from individual-level genetic data, we plotted the relationship between the predicted genomic offset from each of the four workflows with fitness using a single replicate from the *Stepping Stones - Mountain* landscape (bottom four panels, Fig. 5). For these cases, genomic offset is projected to the common garden environment of Population 1 (starred, lower left corner of the *Stepping Stones - Mountain* landscape in Fig. 1). In this dataset, performance is lower for the allele frequency training data (left column in Figure 5C) when compared to genotype training data (right column in Figure 5C) at the same level of evaluation (within a row in Figure 5C). Further, $GO_{\mathrm{geno,pop}}$ accurately predicts the top ten populations from which mean fitness will be greatest in the transplant environment while $GO_{\mathrm{AF,pop}}$ does not (see dashed lines, bottom panels Fig. 5).

Q4 / How does the format of the environmental training data affect model performance?

We hypothesized that averaging environmental data at the population level ($GO_{geno,ind(pop\text{-}env)}$) would decrease accuracy relative to individual-level environmental values ($GO_{geno,ind(ind\text{-}env)}$) by assigning incorrect environmental values to individuals.

503 Comparing genotype-based models from spatially continuous datasets, the performance
504 was greater for models that used individual-level environmental data than that from
505 population-level environmental data (Fig. 6).

506 *Q5 | How does the size of the dataset affect computational time and memory*
507 *requirements?*

508 We observed that the increase of memory and time with larger datasets (e.g., more loci
509 or more samples) was nonlinear (Fig. 7). For example, AF models were 1/10 the size of
510 genotype models (100 populations vs. 1000 individuals), but AF implementations required
511 44% less computing time and 57% less memory compared to genotype implementations
512 (Supplemental Code 04.03). Walltime and memory also increased with the number of loci,
513 with 20000-locus datasets requiring more memory and walltime than expected from a
514 linear extrapolation of 500-locus datasets. Additionally, the spatial arrangement of
515 environmental variables on the landscape affected computational resource requirements
516 (Fig. 7), but the current software implementation of GF makes it difficult to investigate
517 the exact causes of why more memory and time resources are needed for different
518 landscapes.

519 # 4 | Discussion

520 Recent evaluations of genomic offset methods suggest that these models may be useful
521 in some systems to guide management action in ameliorating the maladaptive effects of
522 climate change in natural populations. Even so, much of the domain of applicability of

genomic offsets is yet to be described (Lind and Lotterhos 2024; Lotterhos 2024*b*). Previous evaluations have shown that genomic offsets should be rigorously explored before incorporating model inference into management planning. This includes understanding model sensitivity to the choice of populations used for training (Lind et al. 2024), accounting for uncertainty in climate forecasts and projections to novel climates (DeSaix et al. 2022; Lachmuth et al. 2023; Lind and Lotterhos 2024), sampling locally adapted populations (Rellstab et al. 2021; Lind and Lotterhos 2024), as well as considering other important factors affecting genotype-climate relationships, such as neutral demographic effects driven by differences in effective population sizes of sampled populations (Láruson et al. 2022).

Here we add to these considerations by showing that while the performance of models was relatively insensitive to the number of loci provided, there were differences between some models trained using either population allele frequencies or individual genotypes. Although the format of the genetic data had little effect in landscapes in which geographic distance corresponded to both environmental distance and patterns of local adaptation (*Stepping Stone* - *Clines* and *Estuary* - *Clines* landscapes), models that were provided genotypes instead of allele frequencies improved the most in environments where geographic distance did not correspond to environmental distance or patterns of local adaptation (*Stepping Stones* - *Mountain* landscapes). Models were improved in these landscapes by incorporating additional loci beyond those used by AF-based models. These

additional loci affected the configuration of model weights that led to more accurate predictions.

4.1 / The usefulness of individual genotypes

Many previous implementations of GF_{offset} have used population allele frequencies to train models (e.g., Lachmuth et al. 2023; Lind et al. 2024), and in some cases have done so despite the availability of individual-level genotypic data (e.g., Fitzpatrick and Keller 2015; DeSaix et al. 2022; Láruson et al. 2022; Lind and Lotterhos 2024; Tigano et al. 2024). Several factors may contribute to the common use of allele frequencies instead of genotypes for GF_{offset} model training. First, instructional genomic offset vignettes (including published code) often give examples using allele frequencies. Second, investigators may be using pool-seq datasets where DNA extractions from multiple individuals are pooled at the population level prior to sequencing (see Schlötterer et al. 2014 for a technical review) and therefore only have allele frequency information. This pool-seq approach has been used in several studies (e.g., Gugger et al. 2021; Nielsen et al. 2021; Lind et al. 2024). A third consideration could also be computational resources. Indeed, the first implementation of GF_{offset} was run on a single laptop (Fitzpatrick and Keller 2015) , and subsequent analyses on high performance computing clusters have favored allele frequency data because of reduced run times (e.g., Lind and Lotterhos 2024) . Finally, the use of allele frequencies may have been motivated by the fact that methods such as GF_{offset} cannot differentiate performance between individuals that come

563 from the same multivariate environment (e.g., when assigning environmental values to
564 individuals from bulk seed lots collected regionally for reforestation), and therefore since
565 model output takes place at the environmental level, population allele frequencies were
566 used as input.

567 Despite this trend, we showed that genotypic data, which provides information on
568 within-population variation, offers some advantages over allele frequency inputs. The
569 ultimate reason for this gain in performance seems to be from the incorporation of loci in
570 genotype models that were excluded from allele frequency models. These incorporated loci
571 each generally explained much less variation (R^2) within internal random forest models of
572 GF than from loci common between model types. Although these loci were not necessarily
573 adaptive, these additional loci nonetheless changed the weighting of environmental values
574 in internal cumulative importance curves in a way that led to more accurate genomic
575 offset calculations in many *Stepping Stone - Mountain* landscapes. Ultimately the change
576 in weights between model types were due to the differences in numerical values of
577 genotypes (0, 1, or 2) and allele frequencies (a range from 0.01-0.99) that likely impacted
578 the trees within the random forest models. For each tree in a random forest, algorithmic
579 decisions regarding ultimate tree depth or splits at internal nodes (due to differences in
580 the calculation of impurity scores between genotypes and allele frequency values when
581 splitting data using climate values at internal nodes) resulted in differences in the
582 magnitude of R^2 for a given locus. While our implementation of gradientForests ensured
583 that our individual allele counts (i.e., genotypes) were encoded as a continuous variable

584 (i.e., using random forest regression), future studies could investigate the predictive
585 accuracy from models where genetic data is encoded as categorical genotypes (i.e., random
586 forest classification).

587 While we evaluated the impact of genotype and allele frequency inputs on GF_{offset}
588 performance, such considerations are also relevant to other existing genomic offset
589 methods. For instance, genomic offset predictions from redundancy analysis (sensu
590 Capblancq and Forester 2021) can accept both individual genotypes or allele frequencies
591 as input. Furthermore, analyses such as RONA that have traditionally been carried out
592 using allele frequencies (following Rellstab et al. 2016), could also be modified to model
593 genotypes instead. Further investigation is warranted to understand the impacts of
594 genomic data formats for this and other methods.

595 ### *4.2 | How many markers are enough?*

596 Previous evaluations of GF_{offset} have found that predictive accuracy is relatively
597 insensitive to the choice of markers used, for instance between sets of random loci and
598 sets of candidate loci putatively involved in local adaptation (Fitzpatrick et al. 2021;
599 Láruson et al. 2022; Lachmuth et al. 2023; Lind et al. 2024). However, the effect of the
600 number of markers has received less attention. In our study, less dense marker sets (~500
601 loci) performed similarly to more dense marker sets (~20 000 loci) because they were
602 sufficient to capture similar levels of genetic structure that was estimated with PCA axes.
603 For empirical datasets, a small number of markers (e.g.,~ 500) may not perform as well

604 as they do in this study because small marker sets may not capture all aspects of
605 population structure or of adaptive genetic variation through linkage disequilibrium.
606 Model sensitivity to small marker sets may be particularly relevant when using
607 technologies that may sample unevenly across the genome, such as with restriction site-
608 associated sequencing (RAD-seq; Lowry et al. 2017), though future investigation is
609 warranted. To this end, it will be important for future studies to understand the extent
610 to which the genotypic data is distributed across the genome and how well the loci sample
611 across recombination or haplotype blocks. Reporting the extent of linkage disequilibrium
612 decay will be an important step in this direction, as understanding the extent of decay
613 will inform the approximate spacing of markers necessary to capture evolutionary history
614 across the genome. Using annotations from reference genomes will also be important to
615 understand the extent to which loci represent coding (exonic) and non-coding (intronic,
616 intergenic) regions within the genome.

617 When feasible, future studies should demonstrate the sensitivity of their predictions
618 on different sets of loci. For instance, in the case of a candidate marker set, using sets of
619 random loci of similar numbers can be used to understand how predictions change when
620 the input loci are varied. For datasets with a large number of loci, multiple runs of
621 different subsets of loci could be compared. Studies with large datasets could also consider
622 pruning loci for linkage disequilibrium, as it is not yet known how over- or
623 underrepresentation of groups of linked loci affect model outcomes. This is particularly

relevant to gradientForests, where multiple loci are used to weight environmental values that are ultimately used to make predictions.

Our data suggests that loci with low values of R^2 are important for model accuracy, and that adaptive loci are not always those loci with the greatest R^2, even if the underlying genetic basis for fitness is oligogenic. This is particularly true when the genetic architecture is expected to be polygenic. Because of this, GF should not be used to identify adaptive loci from R^2 values. Additionally, a threshold of R^2 should not be applied to loci (e.g., using a test run of GF to get locus R^2, then rerunning using a subset of loci chosen based on R^2) unless sensitivity to such cutoffs is explored.

4.3 | Considerations for future experimental design and evaluation of genomic offsets

Central to decisions regarding data generation for genomic offset models are the specific aims or hypotheses targeted by investigators and how model predictions of genomic offsets are interpreted biologically (Lotterhos 2024*a*, 2024*b*) . Generally, these aims fall into either making predictions for a restoration project or for population-level responses to climate change (Capblancq et al. 2020; Rellstab et al. 2021). For instance, in the case of a restoration scenario, genomic offset predictions could be used to rank potential donor populations for a restoration site, as evaluated in this study. For such cases, investigators are seeking to best identify the population(s) with the greatest fitness in the restoration environment compared to other populations under consideration (i.e., a prediction of fitness differences among genotypes in a single environment). In other cases, investigators

may wish to understand the extent of *in situ* maladaptation of a focal population in relation to the predicted disruptions in environmental optima resulting from climate change (i.e., a prediction of the change in fitness of a single genotype from a current to a future environment). Recently, Lotterhos (2024*a*) showed that there are many ways to calculate fitness differences (i.e., fitness offsets), that the correct calculation depends on the context, and that various calculations of fitness offsets may not be correlated with each other or with genomic offset predictions. These differences are ultimately related to the pattern of local adaptation in the metapopulation. The analysis from Lotterhos (2024*a*) suggested that common gardens, which have been the primary scenario under which genomic offsets have been evaluated (e.g., Capblancq and Forester 2021; Fitzpatrick et al. 2021; Gougherty et al. 2021; Lachmuth et al. 2023; Lind et al. 2024), including this study, may not adequately serve as a proxy of model accuracy when considering potential *in situ* maladaptation to climate change (see also Lind and Lotterhos 2024). Experimental designs have been proposed that could be used to evaluate model accuracy in both scenarios (Lotterhos 2024*a*, 2024*b*) .

To date, there have been several strategies used to quantify the accuracy of genomic offset predictions, ranging from coefficients of rank correlation, to the level of variance explained from linear models. However, no consensus exists, and there may be strengths or motivations for either evaluation statistic (Lotterhos 2024*b*) . Even so, such evaluation statistics determine the relationship between paired estimates of genomic offset with measures of fitness (i.e., are calculated using all data points). In a restoration scenario,

666 the relationship between prediction and ground-truth measurements for all data points
667 may not capture the level of accuracy desired, which is ultimately determined by the
668 purpose of the experiment - e.g., to identify the population source(s) with highest fitness
669 at a restoration site. In these cases, the predictive accuracy of populations that are least
670 suitable for the environment is not a priority. Measures of model accuracy across all data
671 points also complicates model comparison within and between studies, where comparison
672 of the magnitude of the correlation coefficient or R^2 estimated from multiple models is
673 not necessarily indicative of differences between models in identifying the most suitable
674 population(s) for a given site. Furthermore, models estimating the extent of fitness
675 reduction associated with increasing offset (e.g., linear models that report R^2), while
676 potentially valuable for management, are further complicated by the non-linear
677 relationship between these variables, as demonstrated here. Importantly, there may be
678 cases where the relationship between offset and fitness is linear across populations
679 predicted to have intermediate fitness for a given environment (and thus contribute to
680 the significance of a linear model). However, it may also be the case that this relationship
681 becomes increasingly non-linear for individuals predicted to be most or least suited for
682 that environment - as was the case in both the spatially continuous and spatially discrete
683 space Stepping Stones simulations.

684 In this simulation study, both the offset-fitness relationship and the offset-log(fitness)
685 relationship were monotonic but non-linear. Empirical studies have found different shapes
686 in the relationship between fitness proxies and genetic offset. For instance, a study on

687 balsam poplar (*Populus balsamifera*) found a parabolic relationship between height
688 increment and GF_{offset} (Fitzpatrick et al. 2021) , a study on pearl millet landraces
689 (*Cenchrus americanus*) found a linear relationship between mean seed weight and GF_{offset}
690 (Rhoné et al. 2020), and a follow-up study on pearl millet using the same dataset found a
691 linear relationship between log-transformed mean seed weight and GF_{offset} (Gain et al.
692 2023). These different shapes complicate the application of genomic offsets in practice. In
693 our study, we found that for populations with intermediate fitness in a common garden,
694 the relationship between offset and fitness was approximately linear, but the relationship
695 became increasingly more nonlinear when including populations with extremely high or
696 low fitness. Thus, the observed fitness-offset relationship in an empirical context might
697 depend on which subset of populations from the entire metapopulation are included in a
698 common garden experiment.

699 Because accuracy across all data points is not a primary concern for evaluating genomic
700 offset predictions in a restoration scenario, evaluation statistics that capture the accuracy
701 of the predictions for the most suitable populations will likely be useful for model
702 comparisons going forward. Similar arguments can be made for evaluations of genomic
703 offset in the context of *in situ* climate change where the goal is to identify the populations
704 that are least suited for their future climate. In future work, predictive accuracy of models
705 across all data points (e.g., rank correlations or R^2 from linear models) can be compared
706 to evaluations where the underlying hypothesis tests whether the populations with highest
707 (or lowest) fitnesses are those enriched in the extreme ranks of genomic offset predictions.

In practice, the number of population or individual ranks that are relevant will likely vary by system, and may be influenced by management goals relating to genetic diversity and effective population sizes. For instance, predictions that are based on population mean fitness therefore serve as a measure of accuracy when transplanting large numbers of individuals (perhaps from several populations), and may not best serve interests where relatively few individuals will be moved. Such a scenario is analogous to assisted migration, where populations are moved within a species range to match environmental optima in changing climates (Aitken and Whitlock 2013) . While population-level predictions may provide useful insight for management, the development of methods that predict genomic offsets at the individual level warrant both further investigation and development.

4.4 | Conclusions and future directions

An increasing number of studies have shown that signals of environmental adaptation inherent within genomic data hold potential in assisting management to mitigate the maladaptive effects of climate change in some systems. Future studies that implement rigorous exploration of model inputs will be most useful for these purposes, as model sensitivity and the accuracy of model predictions can be jointly assessed.

The experimental design and environmental databases used to generate data is an important factor beyond identifying suitable systems (i.e., populations that are locally adapted to measurable environmental forces). Future evaluations should explore the sensitivity of model predictions to these experimental decisions. Although the number of

markers required to maintain model accuracy may vary by system, the evaluations we present here and elsewhere (e.g., Láruson et al. 2022; Lind and Lotterhos 2024) , suggest that sampling across the climate space of the target organism is likely an important consideration, as is the ability accurately estimate allele frequencies within populations with adequate sample sizes. Indeed, studies comparing predictions from offset models trained using either 240 candidate loci or >335,000 loci from the complete dataset found similar relationships with common garden fitness proxies (e.g., Lachmuth et al. 2023). In this study, we had a relatively large sample size for training and evaluation with coverage of the landscape (1000 individuals - 10 from each of 100 populations). Similarly, our evaluations of continuous space simulations highlight the importance of accurate environmental values at the individual level, a finding further underscored previously by DeSaix et al. (2022) . Even so, empirical evaluations have still found significant negative relationships between offsets and fitness proxies with smaller numbers of individuals or populations used to train or evaluate the model (e.g., Fitzpatrick et al. 2021), so sampling as extensive as what we simulated may not be needed in practice. Nevertheless, the necessary sampling effort is likely to vary among study systems and depend on the heterogeneity of the selective environment and the spatial scale of adaptation. Further, current genomic offset methods are largely phenomenological and lack incorporation of mechanisms underlying local adaptation, such as potential fitness trade-offs underlying species distributions and population dynamics (Bastias et al. 2024). When such mechanisms are uncovered (Wadgymar et al. 2017) they may be useful to improve

749 ecological forecasting models (Getz et al. 2018; Waldvogel et al. 2020). Even when such
750 mechanisms are not incorporated into genomic offset approaches, such understanding may
751 offer further avenues to defining the domain of applicability of these genomic offset
752 methods (Lotterhos et al. 2022).

753 Acknowledgments

754 This research was funded by NSF-- 2043905 (KEL) and Northeastern University. The
755 funding bodies did not have any role in the design of the study, analysis, interpretation
756 of results or in writing of the manuscript. We would like to acknowledge Jason Selwyn,
757 Madeline Eppley, Annabel Hughes, Sarit Truskey, Stephen Keller, and an anonymous
758 reviewer for helpful and constructive comments on earlier versions of our manuscript.

759 Data Availability

760 We reference the analysis code in the text and figure legends by designating Supplemental
761 Code using a directory numbering system from our servers (as opposed to the order listed
762 in the manuscript). Supplemental Code includes Jupyter Notebooks (*.ipynb). For
763 example, for Notebook 4 in Directory 2, we refer to Supplemental Code 02.04. All code,
764 as well as dataframes containing evaluation results, are archived on Zenodo.org (Lind,
765 2025), and includes a link to the GitHub repository. This archive also includes the
766 evaluation results from all workflows. Notebooks are best viewed within a local jupyter or
767 jupyter lab session (to enable cell output scrolling/collapsing), but can also be viewed at
768 nbviewer.jupyter.org using the web link in the archive's README on GitHub. Analyses

were carried out primarily using python v3.8.5 and R v3.5.1. The yml files to reconstruct the coding environments for the Rv3.5.1 (r35.yml) and python v3.8.5 (mvp_env.yml) environments have been previously archived (Lind, 2024). Exact package and code versions are available at the top of each notebook. More information on coding workflows and coding environments can be found in Supplemental Note S1. Data used for analysis have been archived previously (Lotterhos 2023*b*) .

Author Contributions

Brandon M. Lind: conceptualization, data curation, formal analysis, methodology, project administration, software, visualization, writing - original draft, writing - review and editing. **Katie E. Lotterhos**: conceptualization, funding acquisition, methodology, project administration, resources, supervision, writing - original draft, writing - review and editing.

References

Aitken, S. N., and M. C. Whitlock. 2013. Assisted Gene Flow to Facilitate Local Adaptation to Climate Change. Annual Review of Ecology, Evolution, and Systematics 44:367–388.

Aitken, S. N., S. Yeaman, J. A. Holliday, T. Wang, and S. Curtis-McLane. 2008. Adaptation, migration or extirpation: climate change outcomes for tree populations. Evolutionary Applications 1:95–111.

Arguello, J. R., M. Cardoso-Moreira, J. K. Grenier, S. Gottipati, A. G. Clark, and R. Benton. 2016. Extensive local adaptation within the chemosensory system following Drosophila melanogaster's global expansion. Nature Communications 7:ncomms11855.

Bastias, C. C., A. Estarague, D. Vile, E. Gaignon, C.-R. Lee, M. Exposito-Alonso, C. Violle, et al. 2024. Ecological trade-offs drive phenotypic and genetic differentiation of Arabidopsis thaliana in Europe. Nature Communications 15:5185.

Bay, R. A., R. J. Harrigan, V. L. Underwood, H. L. Gibbs, T. B. Smith, and K. Ruegg. 2018. Genomic signals of selection predict climate-driven population declines in a migratory bird. Science 359:83–86.

Bible, J. M., and E. Sanford. 2016. Local adaptation in an estuarine foundation species: Implications for restoration. Biological Conservation 193:95–102.

Blanquart, F., S. Gandon, and S. L. Nuismer. 2012. The effects of migration and drift on local adaptation to a heterogeneous environment. Journal of Evolutionary Biology 25:1351–1363.

Blanquart, F., O. Kaltz, S. L. Nuismer, and S. Gandon. 2013. A practical guide to measuring local adaptation. (D. Ebert, ed.)Ecology Letters 16:1195–1205.

Brauer, C. J., J. Sandoval-Castillo, K. Gates, M. P. Hammer, P. J. Unmack, L. Bernatchez, and L. B. Beheregaray. 2023. Natural hybridization reduces vulnerability to climate change. Nature Climate Change 1–8.

Burford, M., J. Scarpa, B. Cook, and M. Hare. 2014. Local adaptation of a marine invertebrate with a high dispersal potential: evidence from a reciprocal transplant

experiment of the eastern oyster Crassostrea virginica. Marine Ecology Progress Series 505:161–175.

Capblancq, T., M. C. Fitzpatrick, R. A. Bay, M. Exposito-Alonso, and S. R. Keller. 2020. Genomic Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes. Annual Review of Ecology, Evolution, and Systematics 51:245–269.

Capblancq, T., and B. R. Forester. 2021. Redundancy analysis: A Swiss Army Knife for landscape genomics. Methods in Ecology and Evolution 1–12.

Chen, Y., Z. Jiang, P. Fan, P. G. P. Ericson, G. Song, X. Luo, F. Lei, et al. 2022. The combination of genomic offset and niche modelling provides insights into climate change-driven vulnerability. Nature Communications 13:4821.

DeSaix, M. G., T. L. George, A. E. Seglund, G. M. Spellman, E. S. Zavaleta, and K. C. Ruegg. 2022. Forecasting climate change response in an alpine specialist songbird reveals the importance of considering novel climate. Diversity and Distributions 28:2239–2254.

Ellis, N., S. J. Smith, and C. R. Pitcher. 2012. Gradient forests: calculating importance gradients on physical predictors. Ecology 93:156–168.

Exposito-Alonso, M., M. Exposito-Alonso, R. G. Rodríguez, C. Barragán, G. Capovilla, E. Chae, J. Devos, et al. 2019. Natural selection on the Arabidopsis thaliana genome in present and future climates. Nature 573:126–129.

Fitzpatrick, M. C., V. E. Chhatre, R. Y. Soolanayakanahally, and S. R. Keller. 2021. Experimental support for genomic prediction of climate maladaptation using the machine learning approach Gradient Forests. Molecular Ecology Resources 00:1–17.

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

Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, and A. M. Wilczek. 2011. A Map of Local Adaptation in Arabidopsis thaliana. Science 334:86–89.

Fraser, D. J., L. K. Weir, L. Bernatchez, M. M. Hansen, and E. B. Taylor. 2011. Extent and scale of local adaptation in salmonid fishes: review and meta-analysis. Heredity 106:404–420.

Gain, C., and O. François. 2021. LEA 3: Factor models in population genetics and ecological genomics with R. Molecular Ecology Resources 21:2738–2748.

Gain, C., B. Rhoné, P. Cubry, I. Salazar, F. Forbes, Y. Vigouroux, F. Jay, et al. 2023. A Quantitative Theory for Genomic Offset Statistics. Molecular Biology and Evolution 40:msad140.

Getz, W. M., C. R. Marshall, C. J. Carlson, L. Giuggioli, S. J. Ryan, S. S. Romañach, C. Boettiger, et al. 2018. Making ecological models adequate. Ecology Letters 21:153–166.

Gougherty, A. V., S. R. Keller, and M. C. Fitzpatrick. 2021. Maladaptation, migration and extirpation fuel climate change risk in a forest tree species. Nature Climate Change 1–15.

Gugger, P. F., S. T. Fitz-Gibbon, A. Albarrán-Lara, J. W. Wright, and V. L. Sork. 2021. Landscape genomics of Quercus lobata reveals genes involved in local climate adaptation at multiple spatial scales. Molecular Ecology 30:406–423.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Hoffmann, A. A., and C. M. Sgrò. 2012. Climate change and evolutionary adaptation. Nature 470:479–485.

Kawecki, T. J., and D. Ebert. 2004. Conceptual issues in local adaptation. Ecology Letters 7:1225–1241.

Lachmuth, S., T. Capblancq, S. R. Keller, and M. C. Fitzpatrick. 2023. Assessing uncertainty in genomic offset forecasts from landscape genomic models (and implications for restoration and assisted migration). Frontiers in Ecology and Evolution 11:1155783.

Lachmuth, S., T. Capblancq, A. Prakash, S. R. Keller, and M. C. Fitzpatrick. 2024. Novel genomic offset metrics integrate local adaptation into habitat suitability forecasts and inform assisted migration. Ecological Monographs 94:e1593.

Láruson, Á. J., M. C. Fitzpatrick, S. R. Keller, B. C. Haller, and K. E. Lotterhos. 2022. Seeing the forest for the trees: Assessing genetic offset predictions from gradient forest. Evolutionary Applications 15:403–416.

Lee-Yaw, J. A., J. L. McCune, S. Pironon, and S. N. Sheth. 2022. Species distribution models rarely predict the biology of real populations. Ecography 2022.

Leimu, R., and M. Fischer. 2008. A meta-analysis of local adaptation in plants. (A. Buckling, ed.)PLoS ONE 3:e4010.

873 Leites, L., and M. B. Garzón. 2023. Forest tree species adaptation to climate across
874 biomes: Building on the legacy of ecological genetics to anticipate responses to climate
875 change. Global Change Biology 29:4711–4730.

876 Lind, B. M., R. Candido-Ribeiro, P. Singh, M. Lu, D. O. Vidakovic, T. R. Booker, M.
877 C. Whitlock, et al. 2024. How useful is genomic data for predicting maladaptation to
878 future climate? Global Change Biology 30:e17227.

879 Lind, B. M., and K. E. Lotterhos. 2024. The accuracy of predicting maladaptation to
880 new environments with genomic data. Molecular Ecology Resources e14008.

881 Lotterhos, K. E. 2023*a*. The paradox of adaptive trait clines with nonclinal patterns in
882 the underlying genes. Proceedings of the National Academy of Sciences 120.

883 Lotterhos, K. E. 2023*b*. Output model data from paradox of adaptive trait clines with
884 non-clinal patterns in the underlying genes (Model Validation Program project).
885 Biological and chemical oceanography data management office (BCO-DMO). (version 1)
886 version date 2023-02-13. https://doi.org/10.26008/1912/bco-dmo.889769.1.

887 Lotterhos, K. E. 2024*a*. Interpretation issues with "genomic vulnerability" arise from
888 conceptual issues in local adaptation and maladaptation. Evolution Letters qrae004.

889 Lotterhos, K. E. 2024*b*. Principles in experimental design for evaluating genomic
890 forecasts. Methods in Ecology and Evolution 15:1466–1482.

891 Lotterhos, K. E., M. C. Fitzpatrick, and H. Blackmon. 2022. Simulation Tests of
892 Methods in Evolution, Ecology, and Systematics: Pitfalls, Progress, and Principles.
893 Annual Review of Ecology, Evolution, and Systematics 53:113–136.

894 Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, M. F. Antolin, and
895 A. Storfer. 2017. Responsible RAD: Striving for best practices in population genomic
896 studies of adaptation. Molecular Ecology Resources 17:366–369.

897 Nielsen, E. S., R. Henriques, M. Beger, and S. Heyden. 2021. Distinct interspecific and
898 intraspecific vulnerability of coastal species to global change. Global Change Biology
899 27:3415–3431.

900 Pacifici, M., W. B. Foden, P. Visconti, J. E. M. Watson, S. H. M. Butchart, K. M.
901 Kovacs, B. R. Scheffers, et al. 2015. Assessing species vulnerability to climate change.
902 Nature Climate Change 5:215–224.

903 Rellstab, C., B. Dauphin, and M. Exposito-Alonso. 2021. Prospects and limitations of
904 genomic offset in conservation management. Evolutionary Applications 14:1202–1212.

905 Rellstab, C., S. Zoller, L. Walthert, I. Lesur, A. R. Pluess, R. Graf, C. Bodénès, et al.
906 2016. Signatures of local adaptation in candidate genes of oaks (Quercus spp.) with
907 respect to present and future climatic conditions. Molecular Ecology 25:5907–5924.

908 Rhoné, B., D. Defrance, C. Berthouly-Salazar, C. Mariac, P. Cubry, M. Couderc, A.
909 Dequincey, et al. 2020. Pearl millet genomic vulnerability to climate change in West
910 Africa highlights the need for regional collaboration. Nature Communications 11:5274.

911 Rivkin, L. R., E. S. Richardson, J. M. Miller, T. C. Atwood, S. Baryluk, E. W. Born, C.
912 Davis, et al. 2024. Assessing the risk of climate maladaptation for Canadian polar bears.
913 Ecology Letters 27:e14486.

914 Ruegg, K., R. A. Bay, E. C. Anderson, J. F. Saracco, R. J. Harrigan, M. Whitfield, E.
915 H. Paxton, et al. 2018. Ecological genomics predicts climate vulnerability in an
916 endangered southwestern songbird. Ecology Letters 21:1085–1096.

917 Schlötterer, C., R. Tobler, R. Kofler, and V. Nolte. 2014. Sequencing pools of individuals
918 — mining genome-wide polymorphism data without big funding. Scientific Reports
919 15:749–763.

920 Smith, S., N. Ellis, and C. Pitcher. 2012. gradientForest: Random Forest functions for
921 the Census of Marine Life synthesis project - v0.1-24. Ecology 93:156–168.

922 Thomas, L., J. N. Underwood, N. H. Rose, Z. L. Fuller, Z. T. Richards, L. Dugal, C. M.
923 Grimaldi, et al. 2022. Spatially varying selection between habitats drives physiological
924 shifts and local adaptation in a broadcast spawning coral on a remote atoll in Western
925 Australia. Science Advances 8:eabl9185.

926 Thuiller, W., C. Albert, M. B. Araújo, P. M. Berry, M. Cabeza, A. Guisan, T. Hickler,
927 et al. 2008. Predicting global change impacts on plant species' distributions: Future
928 challenges. Perspectives in Plant Ecology, Evolution and Systematics 9:137–152.

929 Tigano, A., T. Weir, H. G. M. Ward, M. K. Gale, C. M. Wong, E. J. Eliason, K. M.
930 Miller, et al. 2024. Genomic vulnerability of a freshwater salmonid under climate
931 change. Evolutionary Applications 17:e13602.

932 Wadgymar, S. M., D. B. Lowry, B. A. Gould, C. N. Byron, R. M. Mactavish, and J. T.
933 Anderson. 2017. Identifying targets and agents of selection: innovative methods to

934 evaluate the processes that contribute to local adaptation. Methods in Ecology and
935 Evolution 8:738–749.

936 Waldvogel, A.-M., B. Feldmeyer, G. Rolshausen, M. Exposito-Alonso, C. Rellstab, R.
937 Kofler, T. Mock, et al. 2020. Evolutionary genomics can improve prediction of species'
938 responses to climate change. Evolution Letters 4.

939

Boxes, Figures, and Tables for:

A comparison of genomic forecasts based on genotypes versus allele frequencies

Brandon M. Lind*, Katie E. Lotterhos

Department of Marine and Environmental Sciences
Northeastern University Marine Science Center
430 Nahant Road, Nahant, MA 01908, USA.10 July 2025

10 July 2025

Running Title: *Population- and individual-level genomic offsets*

Keywords: genomic forecasting, genomic offset, random forest, climate change, genotypes, allele frequencies

***Corresponding Author**
Email: lind.brandon.m@gmail.com

Brandon M. Lind https://orcid.org/0000-0002-8560-5417
Katie E. Lotterhos https://orcid.org/0000-0001-7529-2771

Box 1 Offset predictions from gradientForest models are invariant across inhabitants of the same contemporary environment

1. Model Training

Genetic and environmental data are needed to train a gradientForest (GF) model. Genetic data (whether formatted as genotypes or allele frequencies) is often assigned to a single geographic location. Generally, the geographic location is used to determine values of the contemporary multivariate environment for each population or individual. Genetic data for each genetic source (either individuals or populations) and contemporary environmental values are then used as input for GF model training (see Fig. newS3 for an illustrative example).

2. Offset prediction

After training the model, to estimate GF_{offset} for each population or location of interest, a Euclidean distance is calculated between a contemporary ("current") and future model-transformed prediction of the multivariate environment:

$$\mathrm{GF}_{\mathrm{offset}} = d(\mathbf{e}_{curr}, \mathbf{e}_{fut}) = \sqrt{\sum_{i=1}^{n} (T(e_{fut_i}) - T(e_{curr_i}))^2} \quad \text{(Eq. B1)}$$

where $\mathbf{e}_{curr}$ and $\mathbf{e}_{fut}$ are respectively the current and future multivariate environments of the population or location, e_i is the value of the *i*'th of *n* environmental variables, and $T(e_i)$ is the GF-model-transformed environmental value predicted by GF for the i'th environmental variable.

One limitation of GF_{offset} predictions, therefore, is that genetic identity within the offset calculation is represented only by the multivariate environment, $\mathbf{e}_{curr}$. Because of this, the model predicts the same level of offset to $\mathbf{e}_{fut}$ for all genetic sources (individuals or populations) that inhabit the same multivariate environment, $\mathbf{e}_{curr}$, whether the genetic sources all come from the same location, or from multiple locations on an environmentally patchy landscape. The consequence of this is that no matter the format of genetic data used to train the GF model, ranking suitable individuals or populations from the same source environment based on GF_{offset} values is not possible.

Box 2 The accuracy of offset predictions can be evaluated at the individual or population level

1. Offset evaluations

The accuracy of GF_{offset} predictions can be evaluated using ground-truth data, such as data from individuals that are measured for fitness proxies in a common garden environment.

Offset predictions from both allele frequency- (AF) and genotype-based models (geno; columns Fig. B2) can be evaluated using either population-mean (pop) or individual level fitnesses (ind; rows Fig. B2). For example, allele frequency data could be used to build an genomic offset model, but individual-level fitness data from a common garden experiment could be used to evaluate the model. The accuracy of the model (e.g., Kendall's τ) is then calculated using the relationship between the fitness proxies and the predicted offset values from Eq. 1.

We hypothesized that population-level evaluation will give misleadingly high inference for model accuracy compared to individual-level predictions, for statistical reasons associated with population averaging. We also hypothesized that AF-based models would perform more poorly than genotype-based models because of the exclusion of within-population variation (see Explanation of Questions).

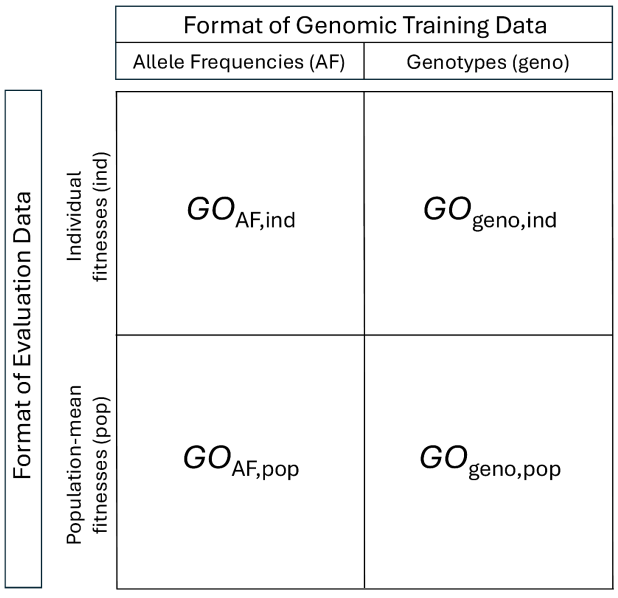


Figure B2 Schematic of the spatially discrete workflows (see Table 1 for more details) used to compare predictive accuracy of genomic offset (GO) from gradientForests (GF_{offset}). We varied the format of input genomic data as either population allele frequencies (AF) or individual genotypes (geno; columns). We also varied the format of fitness data used for evaluation as either population-mean (pop) or individual fitnesses (ind; rows). The subscript abbreviations in the format "$GO_{Training,Evaluation}$" refer to the format of the data used in training (AF or geno) and the format of data used in evaluation (ind or pop). Model accuracy is calculated using the rank correlation (Kendall's $\boldsymbol{\tau}$) between the GF_{offset} calculated from the training data ($GO_{Training,*}$) and fitnesses of individuals or populations in the simulated common garden experiment. In contexts outside of prediction and evaluation, throughout the text we refer to the models by the format of genomic training data that was used: GF_{AF} or GF_{geno}.

Workflow name	(2) Genetic Source for training	(10) Fitness Source for evaluation	(3) Environmental Data	n_{traits}	(1) Simulations Levels (replicates per level)	(9) Total Performance Evaluations completed
$GO_{AF,ind}$	Population-level (allele frequencies)	Individual-level (individual fitness)	Population level (discrete space)	2-trait	180 (3)	216,000
$GO_{AF,pop}$	Population-level (allele frequencies)	Population-level (mean fitness)	Population level (discrete space)	2-trait	180 (3)	216,000
$GO_{geno,ind}$	Individual-level (genotypes)	Individual-level (individual fitness)	Population level (discrete space)	2-trait	180 (3)	189,100*
$GO_{geno,pop}$	Individual-level (genotypes)	Population-level (mean-fitness)	Population level (discrete space)	2-trait	180 (3)	189,100*
$GO_{geno,ind\ (ind\text{-}env)}$	Individual-level (genotypes)	Individual-level (individual fitness)	Individual level (continuous space)	6-trait	1 (1)	400
$GO_{geno,ind\ (pop\text{-}env)}$	Individual-level (genotypes)	Individual-level (individual fitness)	Population average (continuous space)	6-trait	1 (1)	400

* individual-level training did not finish for 269 spatially discrete replicates within either 1 day run time or 250Gb memory

Table 1 Evaluation workflows used to process spatially discrete and spatially continuous simulation data for the evaluation of offset predictions from gradientForests. Numbers given in column names refer to locations in the workflow schematic depicted in Fig newS2. . Subscripts within the names of workflows respectively indicate the levels of training and evaluation, which are indicated in second and third columns (see also Fig. B2). Parenthetical subscripts in workflows indicate the format of environmental data used in training data from the spatially continuous simulation. Total performance evaluations for the spatially discrete workflows (N = 216,000) include evaluations of each replicate (n = 540) from 100 within-landscape common gardens across four implementations of each workflow that further varied the number of loci (500, 5 000, 10 000, 20 000). Total performance evaluations from spatially continuous workflows include 100 common garden evaluations from each of four models varying the number of input loci (500, 5000, 10 000, 20 000). Data from a total of 269 spatially discrete replicates was not able to complete training by gradientForests using genotypes from 20 000 loci. All workflows use only adaptive environmental variables in training (5th column) . Evaluation counts for spatially discrete workflows were tabulated in Supplemental Code 01.02.

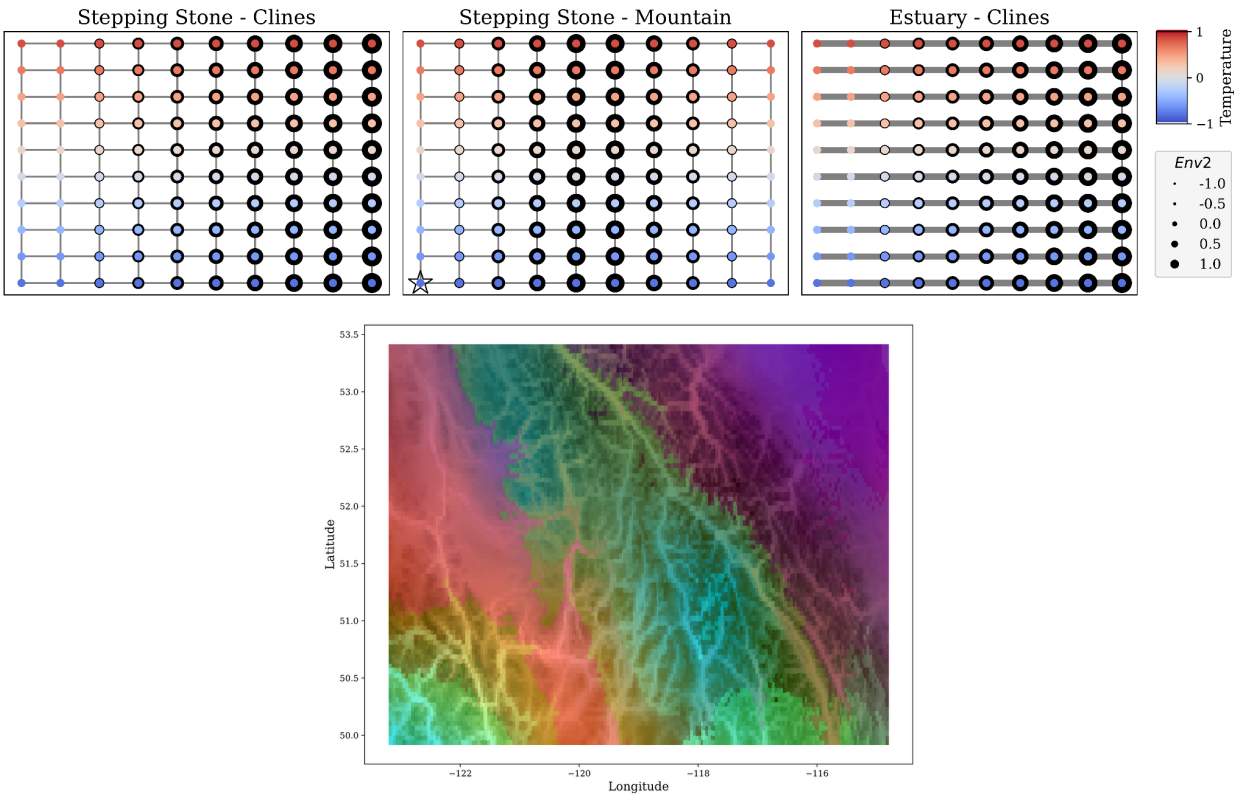


Figure 1 The distribution of environmental variables from spatially discrete (top three panels) and spatially continuous simulations (bottom panel). Points on each spatially discrete landscape represent the sampled populations. The color of each point indicates the value along the latitudinal environmental gradient analogous to temperature (temp), and the size of the black point is indicative of the value along the longitudinal environmental gradient (Env2). Gray lines connecting discrete populations indicate potential for gene flow, the magnitude and direction of which was varied (see Methods; thick gray lines in the Estuary - Clines landscape indicates greater migration rates compared to other landscapes). The star in the Stepping Stones - Mountain landscape designates the evaluation location used in Fig. 5. Multivariate environmental values for the spatially continuous simulation are color-coded with respect to the environments' loading on the first three axes of a principal component analysis of standardized environmental variables - similar environments have similar color shades. Code used to create the spatially discrete landscapes can be found in Supplemental Code 05.08. Code to create the spatially continuous landscape can be found in Supplemental Code 07.03.

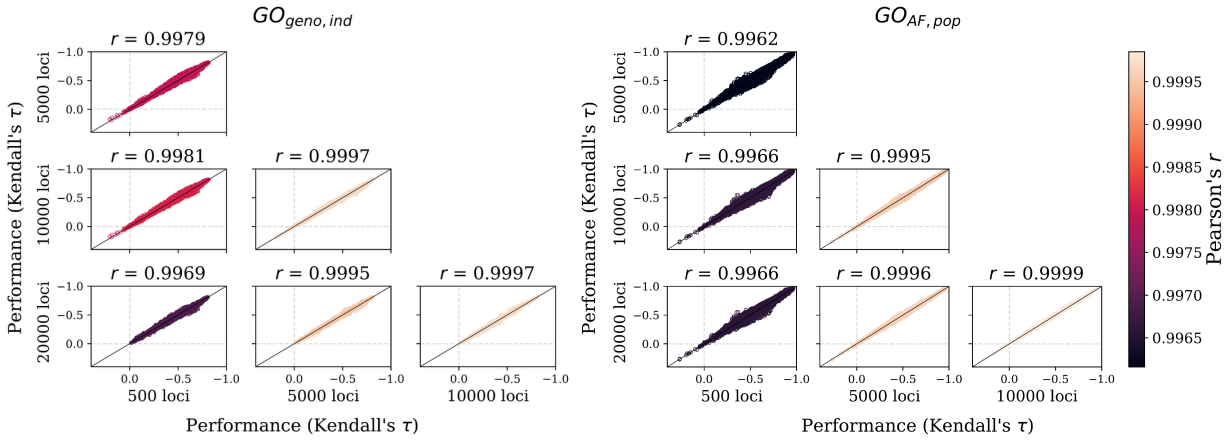


Figure 2 The number of loci used for training has little impact on performance within spatially discrete workflows. Comparison of predictive performance within individual-level ($GO_{\text{geno,ind}}$) and population-level ($GO_{AF,pop}$) offset models when trained with 500, 5 000, 10 000, or 20 000 loci. Data included in this figure is for all comparisons common between evaluations from spatially discrete simulations that completed without failure. Note, for $GO_{geno,ind}$ workflows) that were provided 20,000 loci, only 27,100/54,000 evaluations were able to be compared to other marker set evaluations because of excessive resource requirements. Code to create these figures can be found in Supplemental Code 04.01.

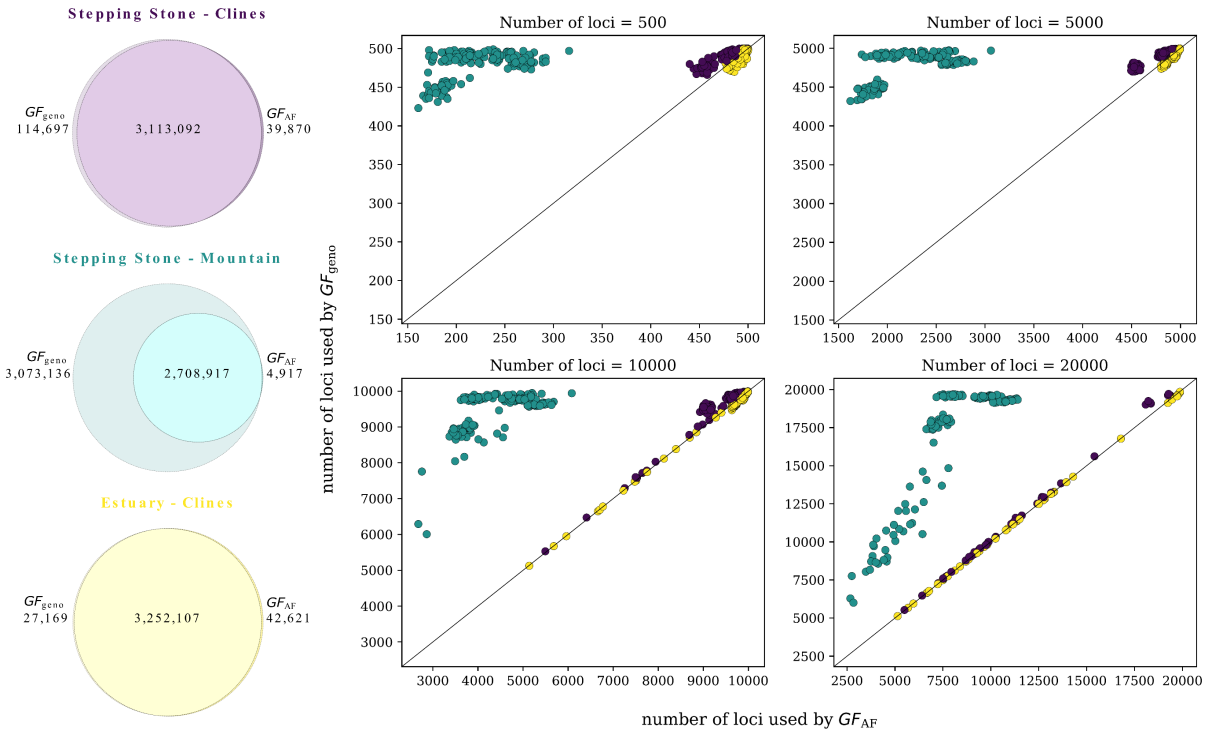
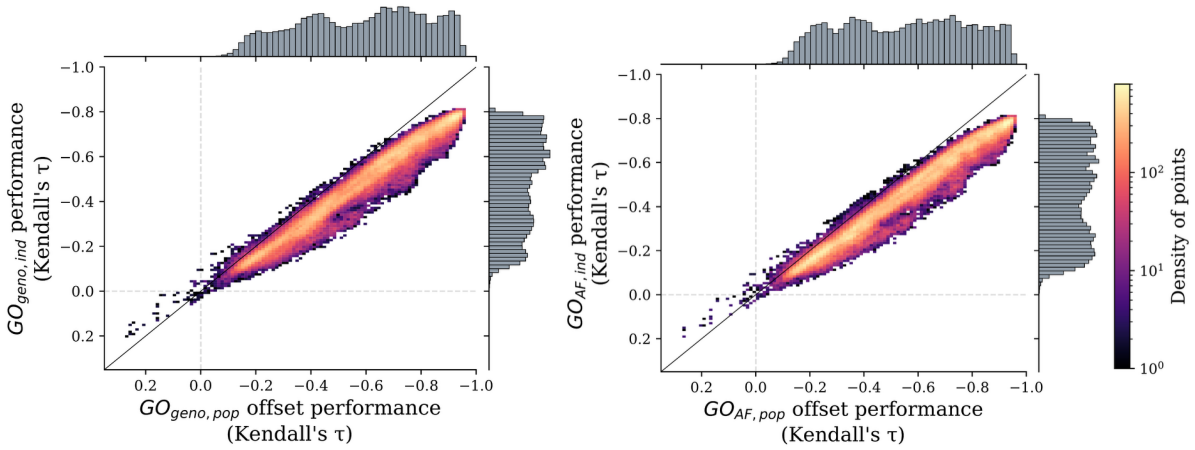


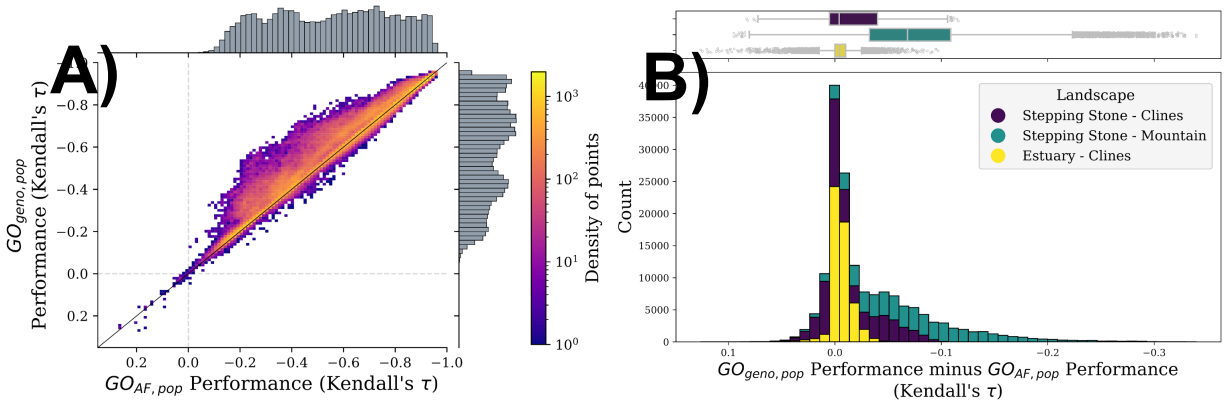
Figure 3 The number and identity of loci used by gradientForest (GF) models differs most substantially between allele frequency- (GFAF) and genotype-based models (GFgeno) from spatially discrete datasets in *Stepping Stone - Mountain* landscapes (green). Left panel - Venn diagrams displaying the extent of overlap of loci used by GF_{AF} and GF_{geno} models; the numbers shown reflect loci across all spatially discrete simulations that were incorporated into GF models. Right panel - relationship between the number of loci used within GF_{AF} (x-axes) and GF_{geno} models (y-axes) for each replicate (points), faceted by the number of loci provided for GF training, and colored with respect to the simulation landscape (see titles in left panel); the 1:1 relationship is represented by a diagonal black line. Data included in these figures are from 189,100 evaluations across replicates and marker sets that finished training for both GF_{AF} and GF_{geno} models. Code used to create these figures can be found in Supplemental Code 05.04.



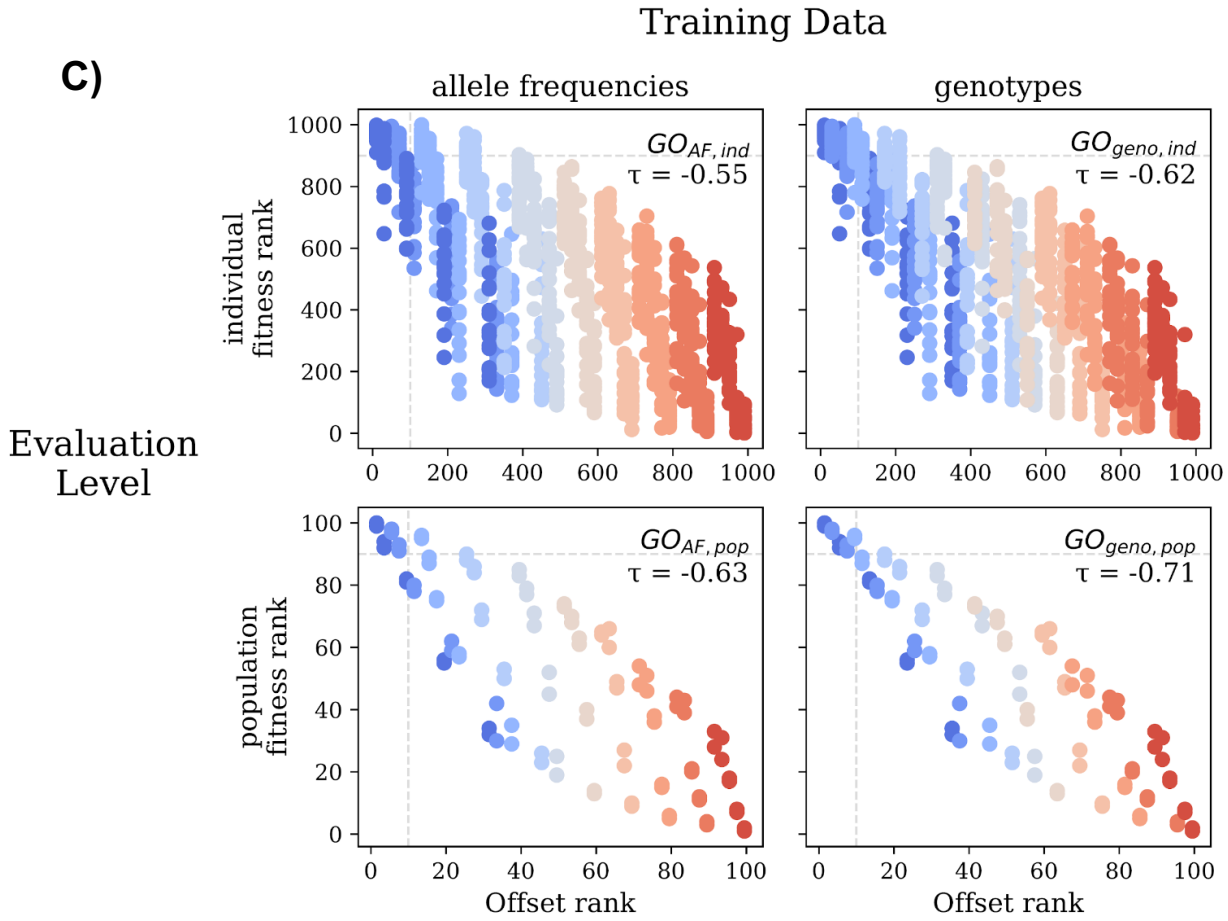
147

148 **Figure 4** Model evaluation using population-mean fitness (x-axes) inflates offset
149 performance scores relative to scores from individual-level evaluation (y-axes). Left
150 panel - genotype-based models are compared using individual- ($GO_{\text{geno,ind}}$, y-axis) and
151 population-level evaluation ($GO_{\text{geno,pop}}$, x-axis). Right panel - performance of allele
152 frequency-based models are compared using individual- ($GO_{\text{AF,ind}}$, y-axis) and
153 population-level evaluation ($GO_{\text{AF,pop}}$, x-axis). The colorbar is standardized across
154 figures; 1:1 relationship is shown as a solid black line. Data included in this figure is
155 for all spatially discrete evaluations that completed without failure. Code to create
156 these figures can be found in Supplemental Code 04.02.

157 (Fig. 5)



158



159

Figure 5 Performance at the population level can sometimes be improved when genetic sources are input as individual genotypes. A) the relationship between performance of population-level predictions from genotype- ($GO_{geno,pop}$, y-axis) and allele frequency-based models ($GO_{AF,pop}$, x-axis; 1:1 relationship is shown as a solid black line). B) the gain in performance of models trained using individual-level genetic data over population-level genetic data (negative values indicate gains in performance). C) for a single replicate, scatter plots of the relationship between fitness and projected offset to Garden 01 (starred on map, middle panel Fig. 1) from each workflow (titles) are shown, color-coded with respect to the home environment on the *Stepping Stones - Mountain* landscape. Vertical and horizontal dashed lines in bottom panels represent the tenth percentile of each axis. Data used in (A) and (B) are from all 162000 spatially discrete evaluation levels from models provided less than 20000 loci. Data used in (C) is from offset models trained using 10000 loci from spatially discrete replicate seed 1231422. Code to create figures can be found in Supplemental Code 05.03 and Supplemental Code 05.07.

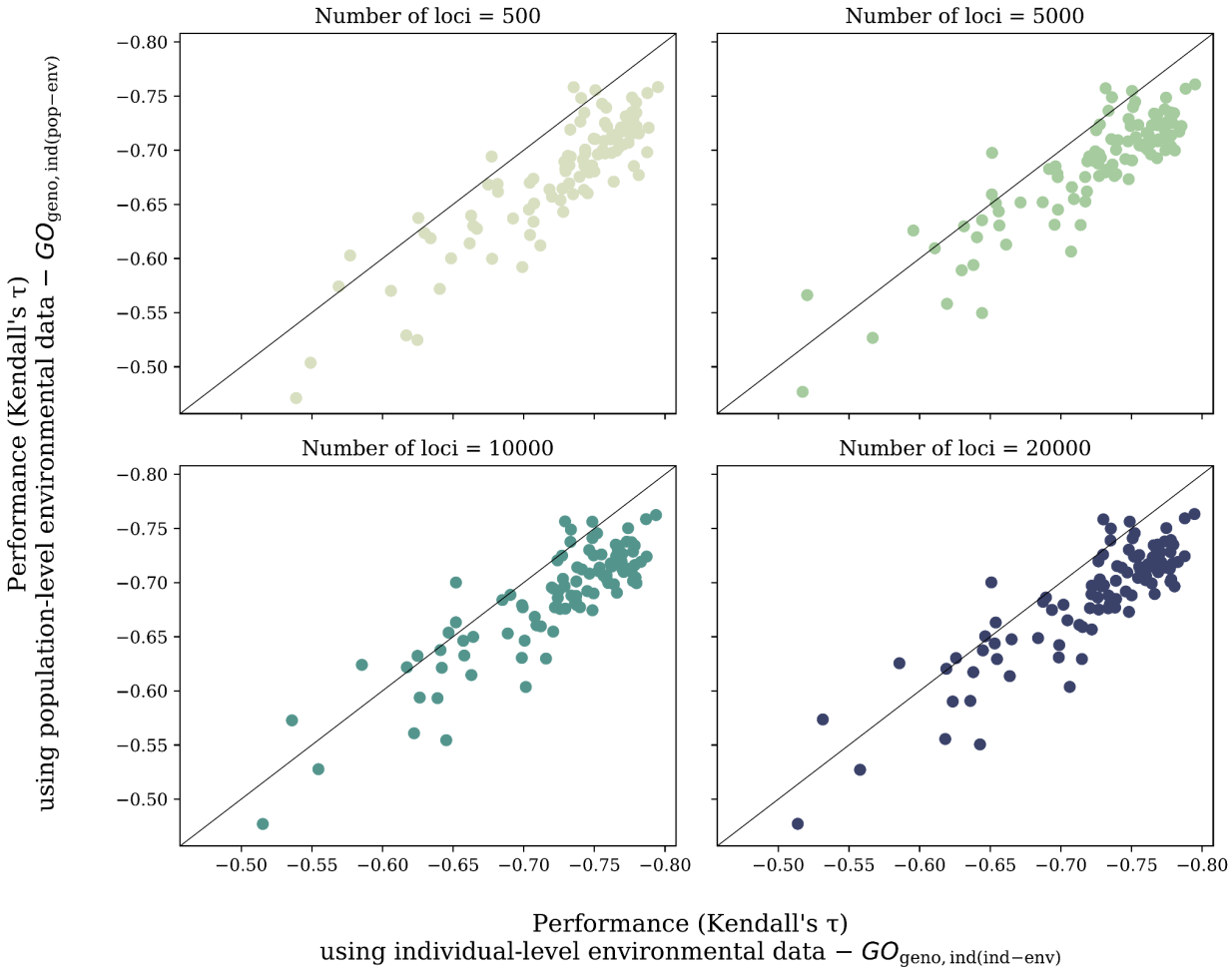


Figure 6 Model performance of GF_{offset} decreases when population-averaged environmental data is used alongside individual genotypes. Shown is the relationship between predictive accuracy of models trained using population-averaged environmental data (y-axes) and individual-level environmental data (x-axes), faceted by the number of markers provided for training. Code to create this figure can be found in Supplemental Code 07.02.

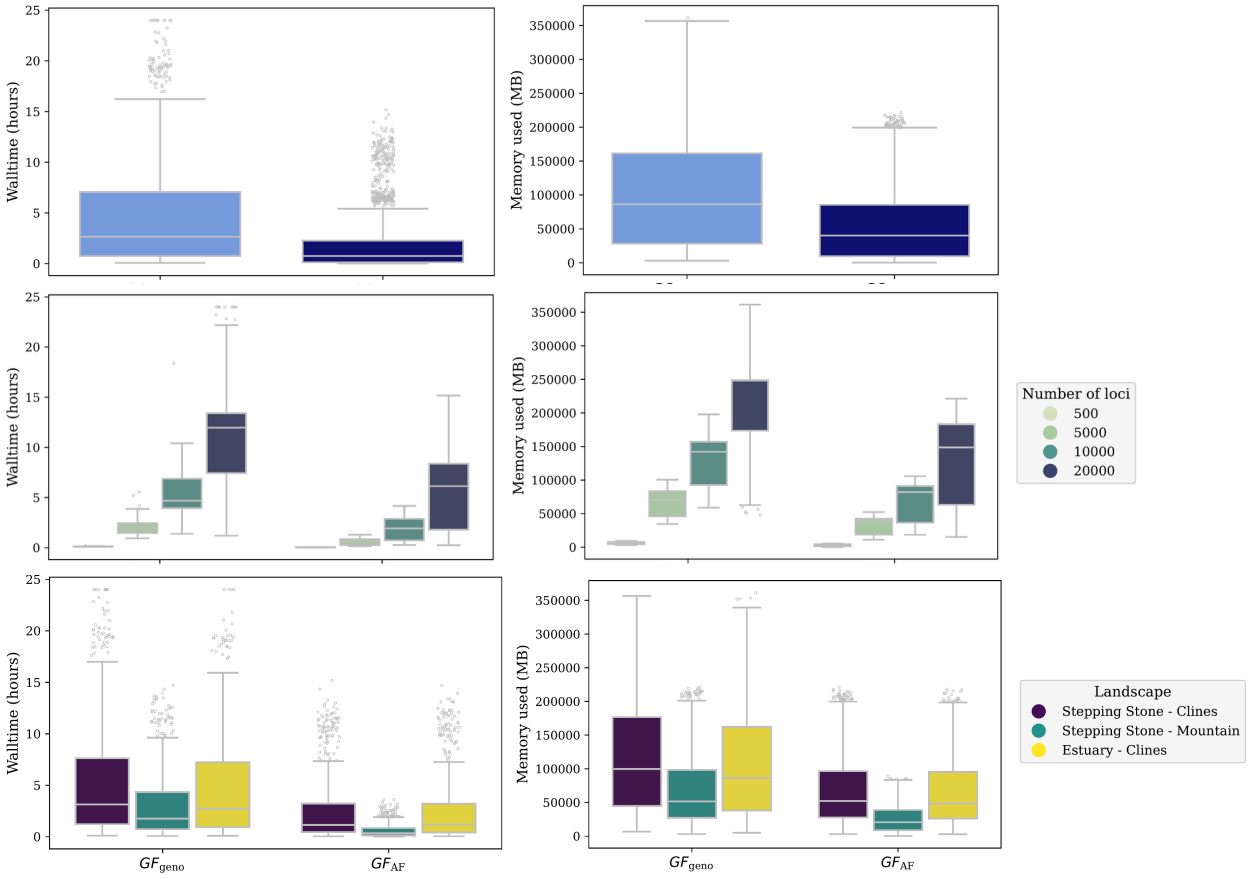


Figure 7 Computation time (y-axes, first column) and memory requirements (y-axes, second column) of gradientForests (GF) model training differs between genotype- (GF_{geno}) and allele frequency-based (GF_{AF}) implementations (x-axes). Differences within implementations are driven primarily by the number of loci (second row) and the pattern of environmental variables (third row). Total runtime across all model training exceeds 417 days for GF_{geno} and 180 days for GF_{AF}. Data included in this figure is for all spatially discrete evaluations that completed without failure. Code used to create this figure can be found in Supplemental Code 04.03.

Supplemental Information for:

A comparison of genomic forecasts based on genotypes versus allele frequencies

Brandon M. Lind*, Katie E. Lotterhos

Department of Marine and Environmental Sciences
Northeastern University Marine Science Center
430 Nahant Road, Nahant, MA 01908, USA.10 July 2025

10 July 2025

Running Title: *Population- and individual-level genomic offsets*

Keywords: genomic forecasting, genomic offset, random forest, climate change, genotypes, allele frequencies

***Corresponding Author**
Email: lind.brandon.m@gmail.com

Brandon M. Lind https://orcid.org/0000-0002-8560-5417
Katie E. Lotterhos https://orcid.org/0000-0001-7529-2771

1 | Supplemental Text

Supplemental Note S1

Implementation of gradientForests

For each set of input loci, we training gradientForest (v0.1-18) using the following parameters: ntree=500, corr.threshold=0.5, and maxLevel=$(0.368 * N/2)$, where N is the number of populations or individuals. We used the default linear extrapolation. The trained models are projected onto the landscape using the `predict` function for each individual or population's home environmental values. This creates the "current" projection used to calculate the offset (Eq. B1 of Box 1 in the main text).

We then used the trained models to create the "future" prediction (Eq. B1 of Box 1 in the main text) for the climates of each of 100 common gardens on the landscape. Specifically, for each garden, the `predict` function is used to take the trained model and the garden's climate to create a projection similar to that using current climate data (previous paragraph). This "future" value is used for all individuals or populations. Then the Euclidean distance is taken between the current and future predictions to calculate genomic offset (Eq. B1 of Box 1 in the main text). These scripts (01_src/MVP_gf_training_script.R and 01_src/MVP_gf_fitting_script.R) were the same used in Lind & Lotterhos (2024) and can be found in the coding archive (Lind 2024a).

Coding workflows

Below we reference the notebooks (*.ipynb) and scripts (*.R, *.py) that were used to analyze data in this manuscript. We reference the analysis code in the text and figure legends by designating Supplemental Code (SC) using a directory numbering system from our servers (as opposed to the order listed in the manuscript). Supplemental Code includes only Jupyter Notebooks (*.ipynb). For example, for Notebook 4 in Directory 2, we refer to SC 02.04.

The remaining notebooks found in the archive not mentioned below are all cited within the main text or within figure legends. Exact versions for python packages are available at the top of each notebook. More information on archived code can be found in the archives' READMEs (Lind 2024a).

Spatially discrete workflows

Datasets of loci (N= 500, 5 000, 10 000, 20 000) were created in SC 00.00. Training of gradientForests (GF, v0.1-18; Ellis et al. 2012; Smith et al. 2012) models using these datasets were executed in SC 01.00 for genotype-based models (GF_{geno}) and in SC 02.01 for allele frequency-based models (GF_{AF}). gradientForests was run in R v3.5.1. Predictions from these models to common garden environments were also carried out in these directories (SC 01.01 - 01.03, SC 02.01 - 02.03).

Evaluation of offset predictions were executed for GF_{geno} models at both the individual ($GO_{geno,ind}$; SC 01.01) and population levels ($GO_{geno,pop}$; SC 03.00). Similarly, evaluations of offset predictions were executed for GF_{AF} models at both the individual ($GO_{AF,ind}$; SC 06.01) and population levels ($GO_{AF,pop}$; SC 02.01).

Scripts used to train, predict, and evaluate GF models were the same as used in previous work (Lind 2024a; Lind and Lotterhos 2024). Specifically, we used the `MVP_gf_training_script.R` script for training, the `MVP_02_fit_gradient_forests.py` script for prediction, and the `MVP_03_validate_gradient_forests.py` script for evaluation. These scripts are available in the 01_src directory of Lind (2024a). Commands used to execute these scripts are provided in the notebooks discussed above.

Spatially continuous workflows

Training datasets were created in SC 07.00. A 10 x 10 grid was used to assign individuals to 100 demes, where the univariate mean of each environmental value among individuals from a deme were used to assign values at the deme level. Scripts to train and predict offset to common gardens were also executed in this notebook.

All workflows

Fitness was estimated for the common gardens in SC 07.01. Performance of predicted offset was evaluated in SC 07.02.

Supplemental Note S2 - Processor Information

We recorded the type and frequency of the processors used to quantify runtime for gradientForest model training. Below, we show the processor information. Code used to retrieve this information is available in Supplemental Code 04.03.

```
Model_name: Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz
Model_name: Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz
Stepping: 2
CPU_MHz: 2999.902
CPU_max_MHz: 3500.0000
CPU_min_MHz: 1200.0000


Model_name: Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz
Model_name: Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz
Stepping: 1
CPU_MHz: 2899.951
CPU_max_MHz: 3300.0000
CPU_min_MHz: 1200.0000


Model_name: Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz
Model_name: Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz
Stepping: 4
CPU_MHz: 3099.926
CPU_max_MHz: 3600.0000
CPU_min_MHz: 1200.0000


Model_name: Intel(R) Xeon(R) Platinum 8276 CPU @ 2.20GHz
Model_name: Intel(R) Xeon(R) Platinum 8276 CPU @ 2.20GHz
Stepping: 7
CPU_MHz: 2200.000
BogoMIPS: 4400.00
Virtualization: VT-x


Model_name: AMD EPYC 7702 64-Core Processor
Model_name: AMD EPYC 7702 64-Core Processor
Stepping: 0
CPU_MHz: 1996.372
BogoMIPS: 3992.74
Virtualization: AMD-V
L1d_cache: 32K L1i


Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz
Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz
Stepping: 4
CPU_MHz: 999.932
CPU_max_MHz: 3500.0000
CPU_min_MHz: 1200.0000
```

2 | Supplemental Figures

This page intentionally left blank.

Allele frequency and population-level environmental input data

	Locus1	Locus2	...	LocusN
Pop1	0.80	0.63	...	0.50
Pop2	0.50	0.75	...	0.17
Pop3	0.75	0.17	...	0.50

	Env1	Env2	...	EnvN
Pop1	225	15.6	...	2.30
Pop2	550	16.4	...	-4.40
Pop3	1525	13.4	...	-10.70

Genotypes and population-level environmental input data

	Locus1	Locus2	...	LocusN
Pop1_1	2	1	...	1
Pop1_2	1	1	...	2
Pop1_3	0	1	...	0
Pop1_4	1	2	...	1
Pop2_1	0	2	...	0
Pop2_2	2	1	...	0
Pop2_3	1	1	...	1
Pop3_1	2	1	...	2
Pop3_2	1	0	...	0
Pop3_3	1	0	...	1

	Env1	Env2	...	EnvN
Pop1_1	225	15.6	...	2.30
Pop1_2	225	15.6	...	2.30
Pop1_3	225	15.6	...	2.30
Pop1_4	225	15.6	...	2.30
Pop2_1	550	16.4	...	-4.40
Pop2_2	550	16.4	...	-4.40
Pop2_3	550	16.4	...	-4.40
Pop3_1	1525	13.4	...	-10.70
Pop3_2	1525	13.4	...	-10.70
Pop3_3	1525	13.4	...	-10.70

Genotypes and individual-level environmental input data

	Locus1	Locus2	...	LocusN
Pop1_1	2	1	...	1
Pop1_2	1	1	...	2
Pop1_3	0	1	...	0
Pop1_4	1	2	...	1
Pop2_1	0	2	...	0
Pop2_2	2	1	...	0
Pop2_3	1	1	...	1
Pop3_1	2	1	...	2
Pop3_2	1	0	...	0
Pop3_3	1	0	...	1

	Env1	En2	...	EnvN
Pop1_1	226	15.60	...	2.32
Pop1_2	223	15.48	...	2.27
Pop1_3	224	15.32	...	2.35
Pop1_4	227	15.87	...	2.29
Pop2_1	553	16.59	...	-4.41
Pop2_2	547	16.40	...	-4.35
Pop2_3	550	16.09	...	-4.45
Pop3_1	1508	13.38	...	-10.61
Pop3_2	1517	13.20	...	-10.73
Pop3_3	1550	13.65	...	-10.76

Figure S1 Examples of the accepted formats of input data to gradientForests model training. Shown are example inputs for three populations that vary the type of genetic and environmental inputs. Environmental data is often input uniformly across individuals from the same population, but can include individual-level data if environmental data can be accessed on the same spatial scales as sampling.

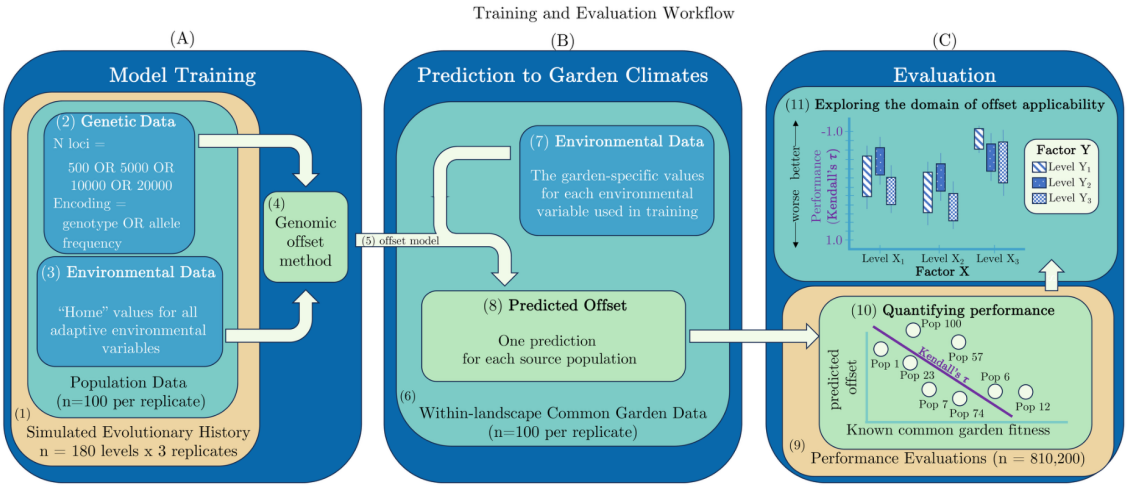


Figure S2 Analysis of spatially discrete simulations included three main phases: (a) model training, (b) model prediction and (c) evaluation of models. Subpanels of this schematic are numbered for referencing in Table 2 and the main text. Evaluation of spatially continuous simulations (not shown) used only individual genotypes and varied the format of the environmental data (individual-level or population-level).

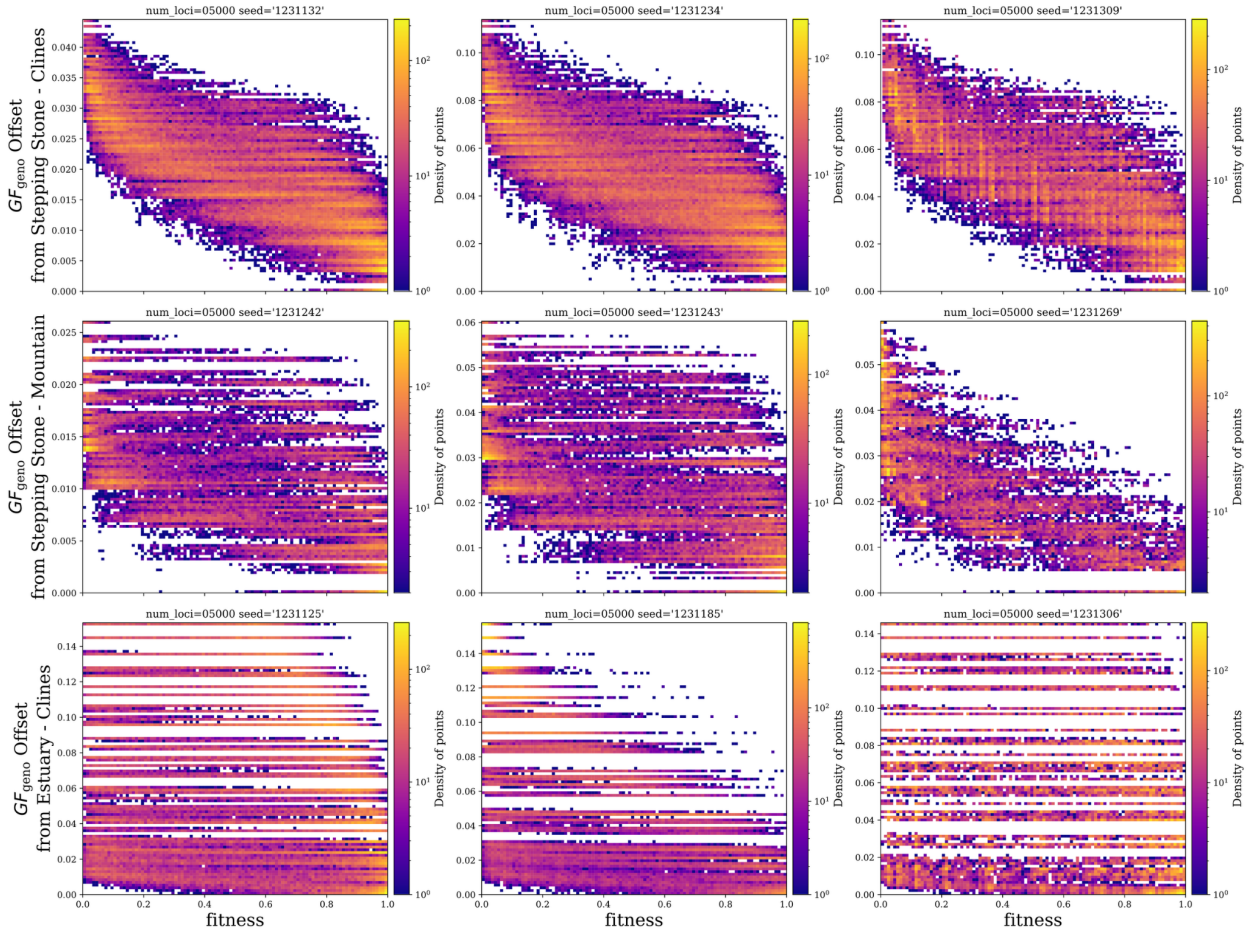


Figure S3 The relationship between fitness and predicted offset from GF_{geno} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored with respect to the density of points. Code to create this figure can be found in SC 05.09. Data included in this figure is from spatially discrete simulations using the fitness and offset predicted for all 100 common gardens on the landscape.

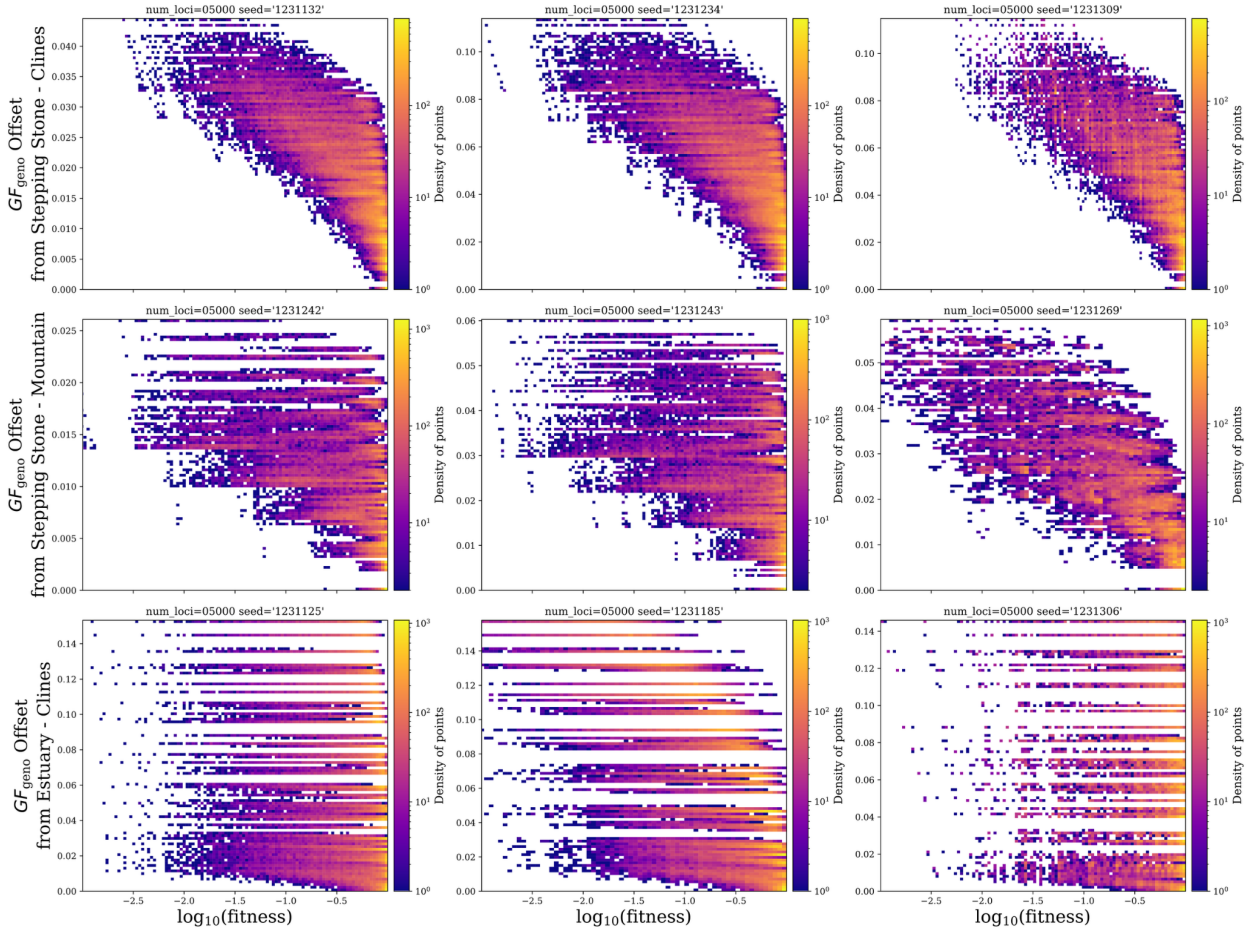


Figure S4 The relationship between $\log_{10}$(fitness) and predicted offset from GF_{geno} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored with respect to the density of points. Code to create this figure can be found in SC 05.09. Data included in this figure is from spatially discrete simulations using the fitness and offset predicted for all 100 common gardens on the landscape.

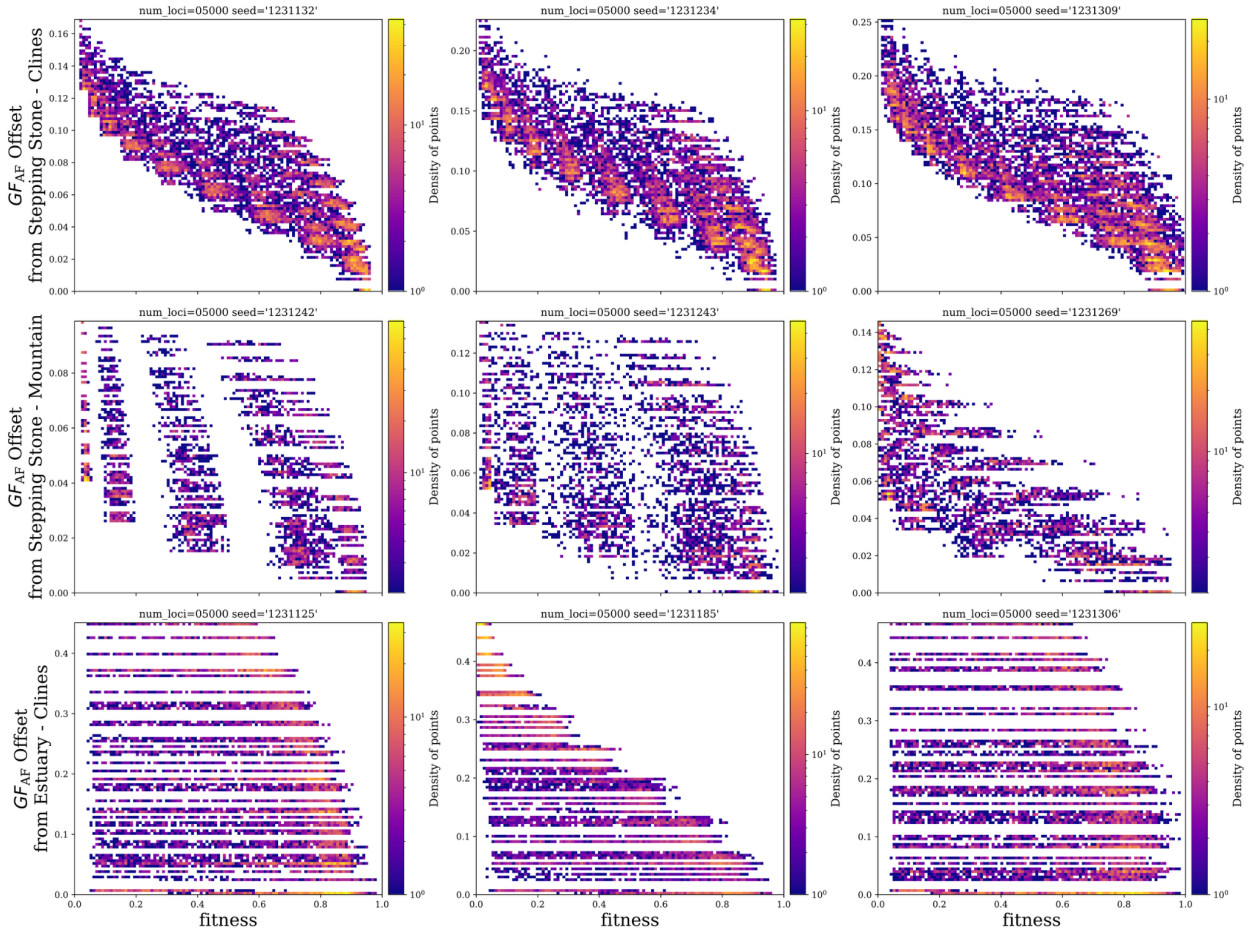


Figure S5 The relationship between fitness and predicted offset from GF_{AF} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored with respect to the density of points. Code to create this figure can be found in SC 05.09. Data included in this figure is from spatially discrete simulations. Data included in this figure is from spatially discrete simulations using the fitness and offset predicted for all 100 common gardens on the landscape.

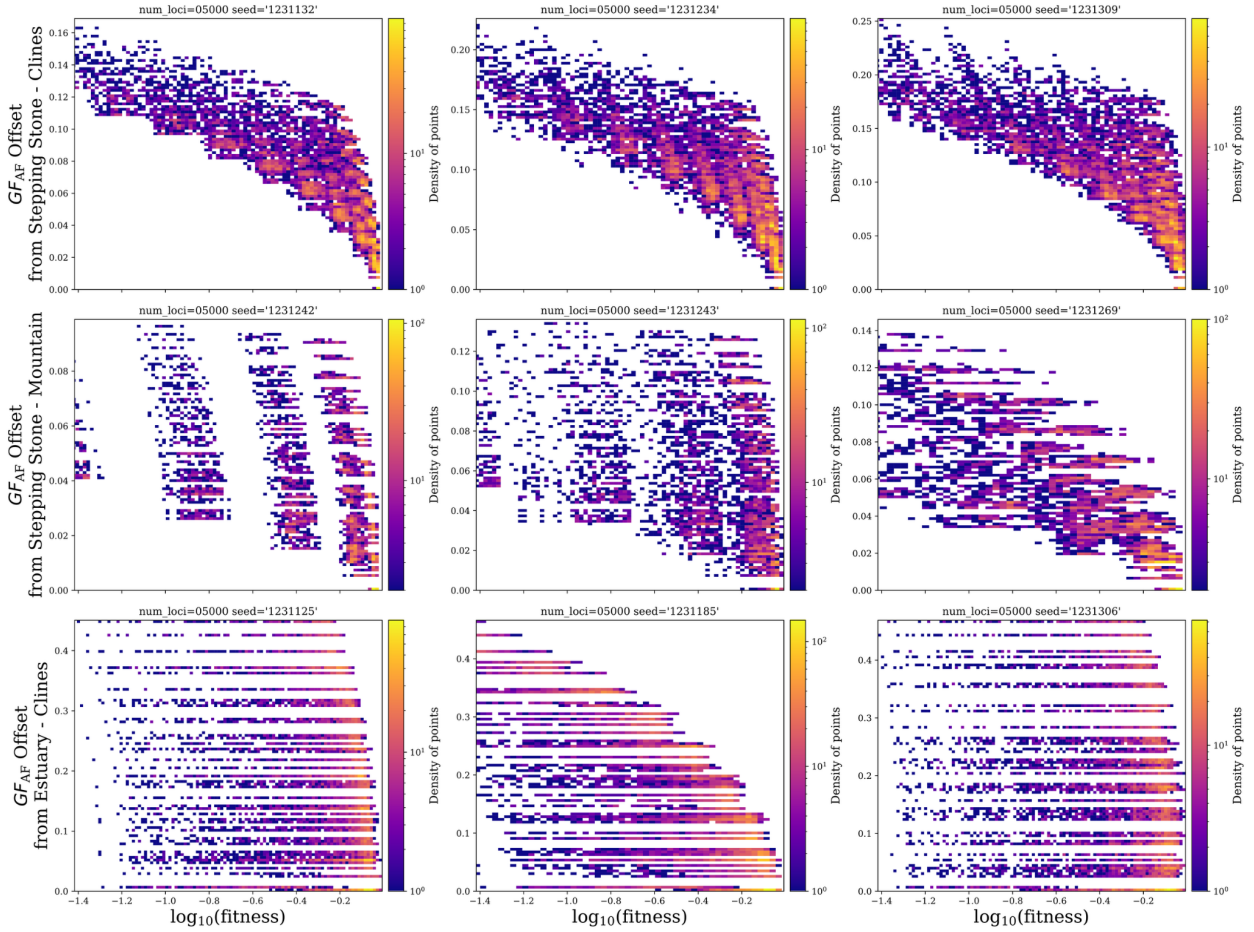


Figure S6 The relationship between fitness and predicted offset from GF_{AF} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored with respect to the density of points. Code to create this figure can be found in SC 05.09. Data included in this figure is from spatially discrete simulations. Data included in this figure is from spatially discrete simulations using the fitness and offset predicted for all 100 common gardens on the landscape.

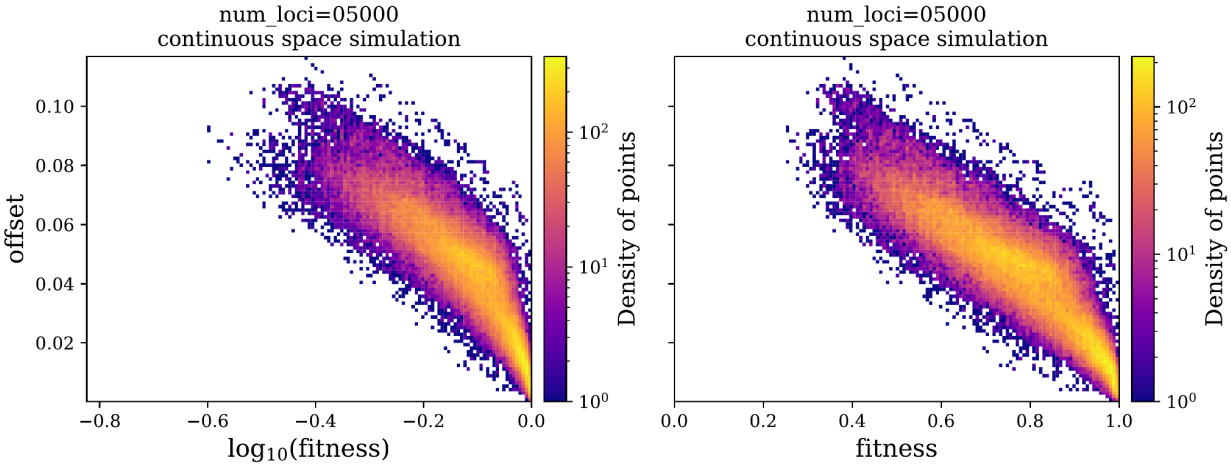


Figure S7 The relationship between fitness or $\log_{10}$(fitness) and predicted offset from the continuous space GF model is non-linear. Figures are colored with respect to the density of points. Code to create this figure can be found in SC 05.09. Data included in this figure is from the spatially discrete simulation using the fitness and offset predicted for all 100 common gardens on the landscape.

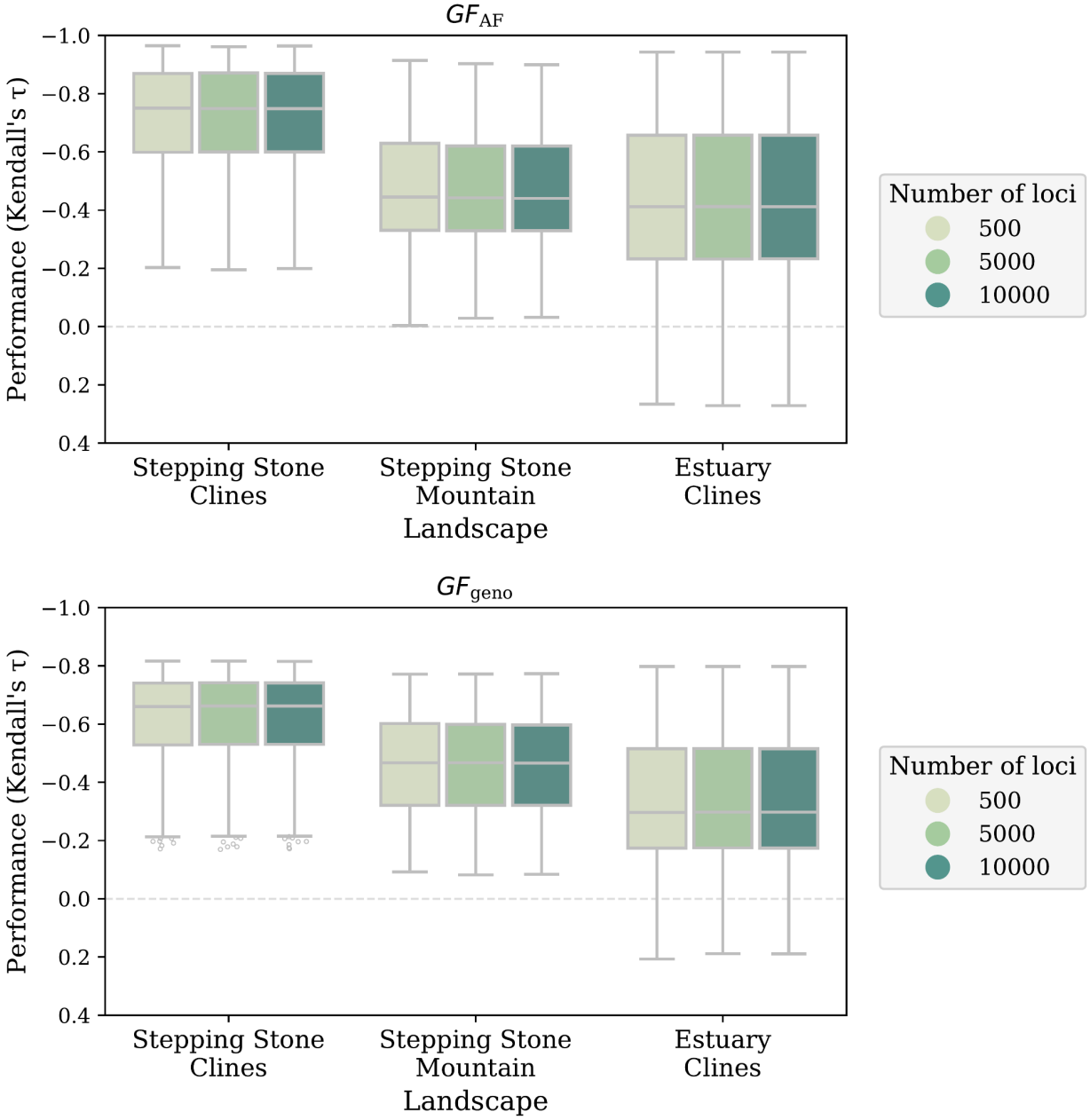


Figure S8 The performance of offset predictions from gradientForests models trained with allele frequencies (GF_{AF}) or genotypes (GF_{geno}) was not differentially affected by the number of loci provided for training. Code to create these figures can be found in Supplemental Code 04.01.

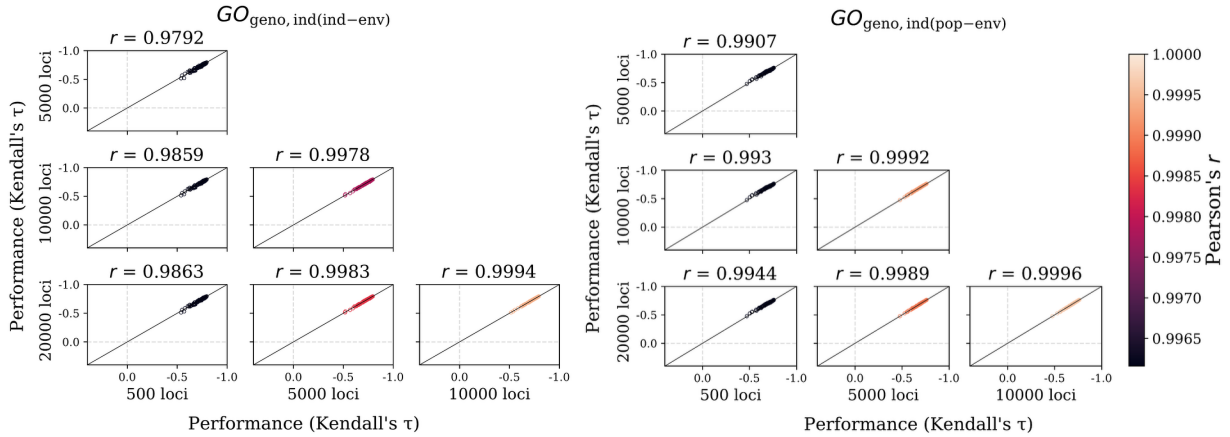


Figure S9 The number of loci used for training has little impact on performance within the spatially continuous workflows. Comparison of predictive performance when environmental data is input at the individual level ($GO_{geno,ind\ (ind\text{-}env)}$) or at the population level ($GO_{geno,ind\ (pop\text{-}env)}$) for offset models trained with 500, 5 000, 10 000, or 20 000 loci encoded as genotypes. Data included in this figure is from all N=100 common garden evaluations from the spatially continuous simulation. Code to create these figures can be found in Supplemental Code 07.02.

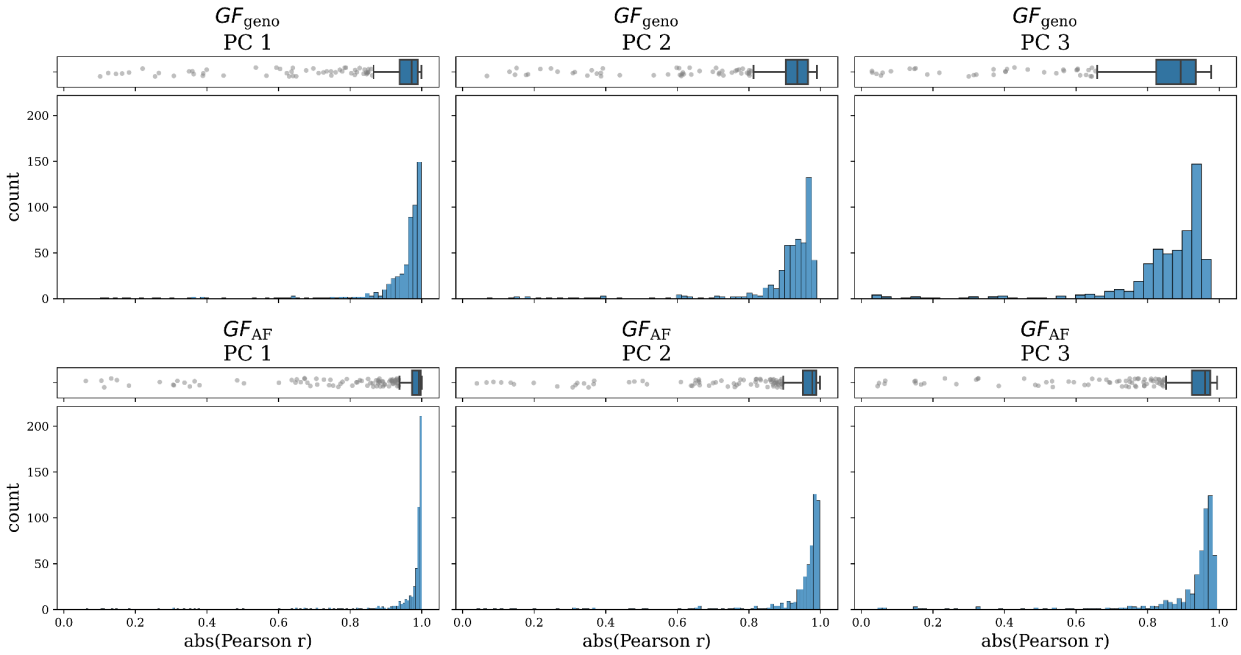


Figure S10 Population structure is captured by small marker sets from both allele frequency- (GF_{AF}) and population-level (GF_{geno}) models of gradientForests (GF). Principal component (PC) analysis was carried out for each marker set from each simulation replicate using either genotype or allele frequencies. Within a replicate, the absolute correlation (abs(Pearson's r), x-axes) between PC axis loadings from PCs estimated using either 500 or 10000 loci was calculated. A-F are histograms of correlations for each PC axis for each workflow (see titles). Code to create these figures can be found in Supplemental Code 05.01.

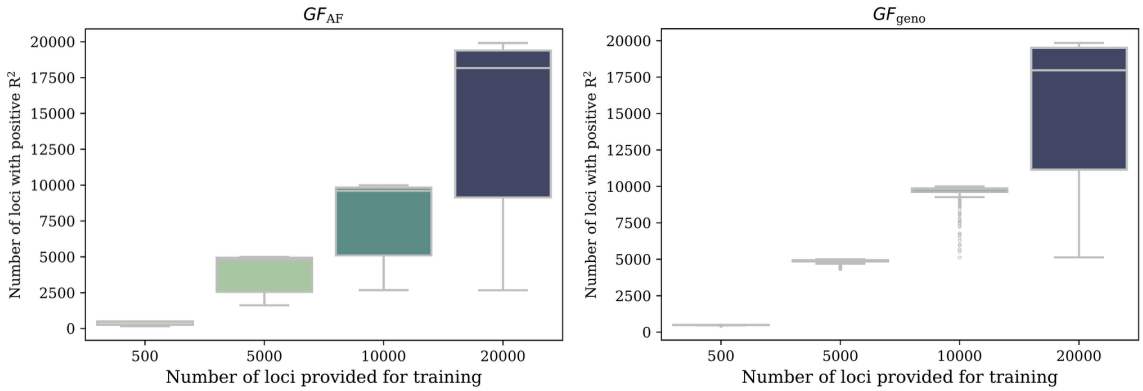


Figure S11 The number of loci that were incorporated into gradientForest (GF) models (y-axes; i.e., loci with $R^2 > 0$ from internal random forest models) were roughly representative of the number of loci provided for model training (x-axes). Data included in this figure is from all models that successfully completed training. Code to create these figures can be found in Supplemental Code 05.04.

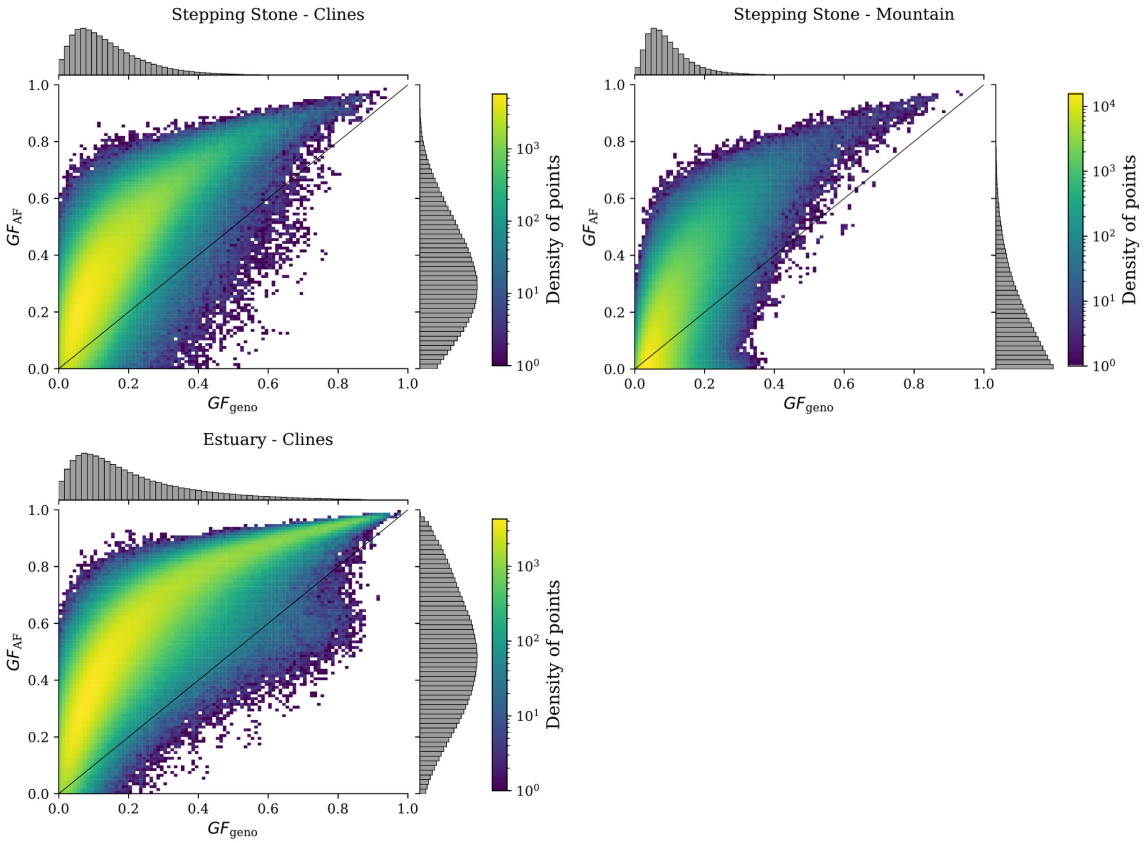


Figure S12 Predictive power of loci (R^2, all axes) from random forest models used by gradientForests (GF) are positively correlated between genotype- (GF_{geno}) and allele frequency-based (GF_{AF}) models. Data included in this figure are all overlapping loci incorporated into GF_{geno} or GF_{AF} models from models that completed training without failure. The 1:1 line is shown as a diagonal black line. Code to create these figures can be found in Supplemental Code 05.05.

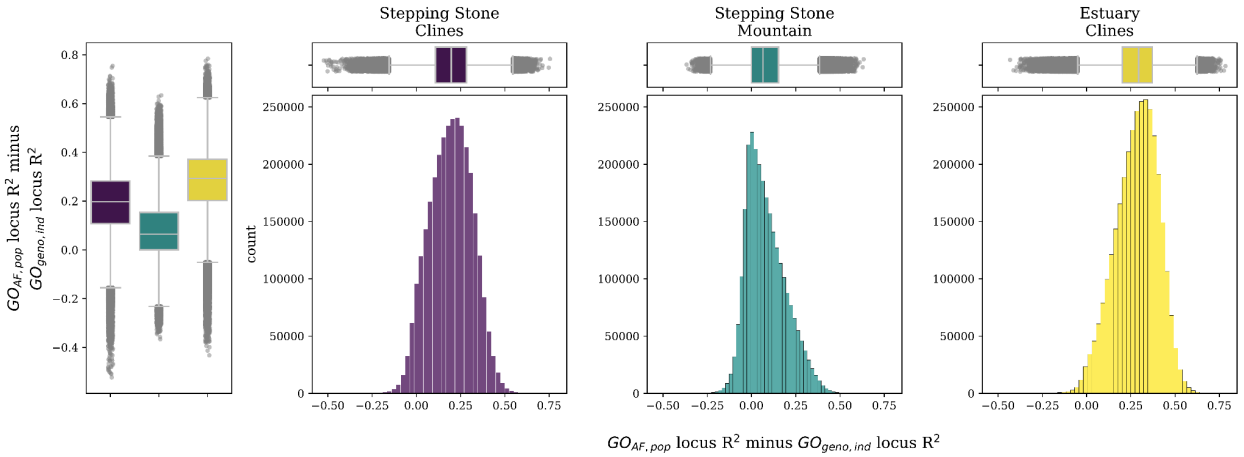


Figure S13 Predictive power of loci (R^2) from random forest models used by gradientForests (GF) are generally greater when encoded as allele frequencies (i.e., GF_{AF} models) than genotypes (i.e., GF_{geno} models). Data used in this figure is R^2 from loci overlapping GF_{AF} and GF_{geno} models. Code to create these figures can be found in Supplemental Code 05.05.

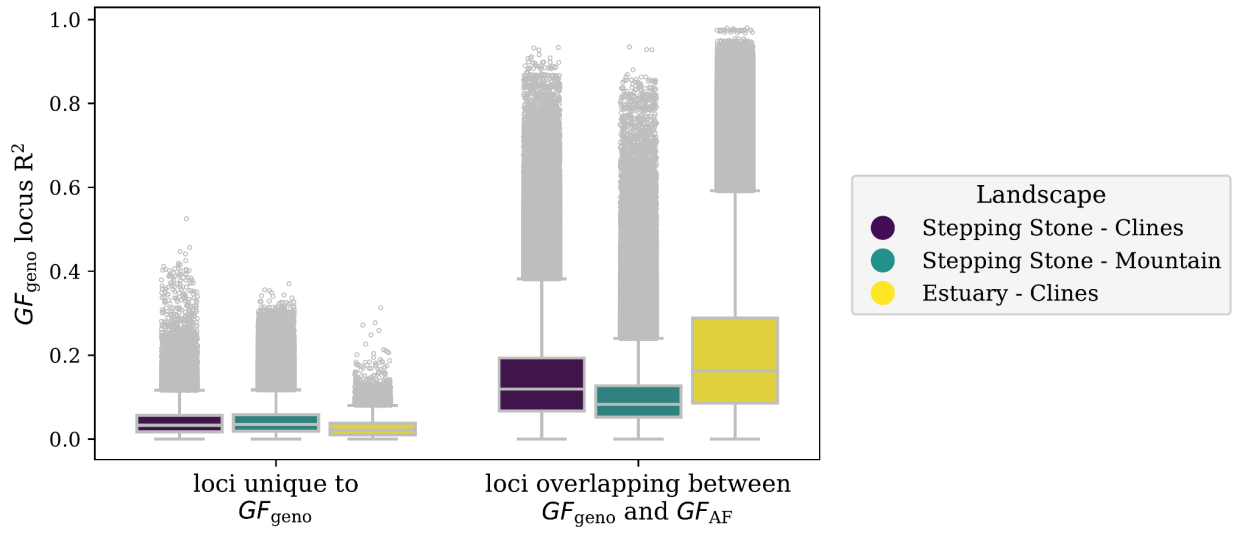


Figure S14 Predictive power (R^2, y-axis) of random forest models (one per locus) from loci incorporated into gradientForest (GF) models. Data included in this figure are R^2 values from GF_{geno} models. Code to create this figure can be found in Supplemental Code 05.05.

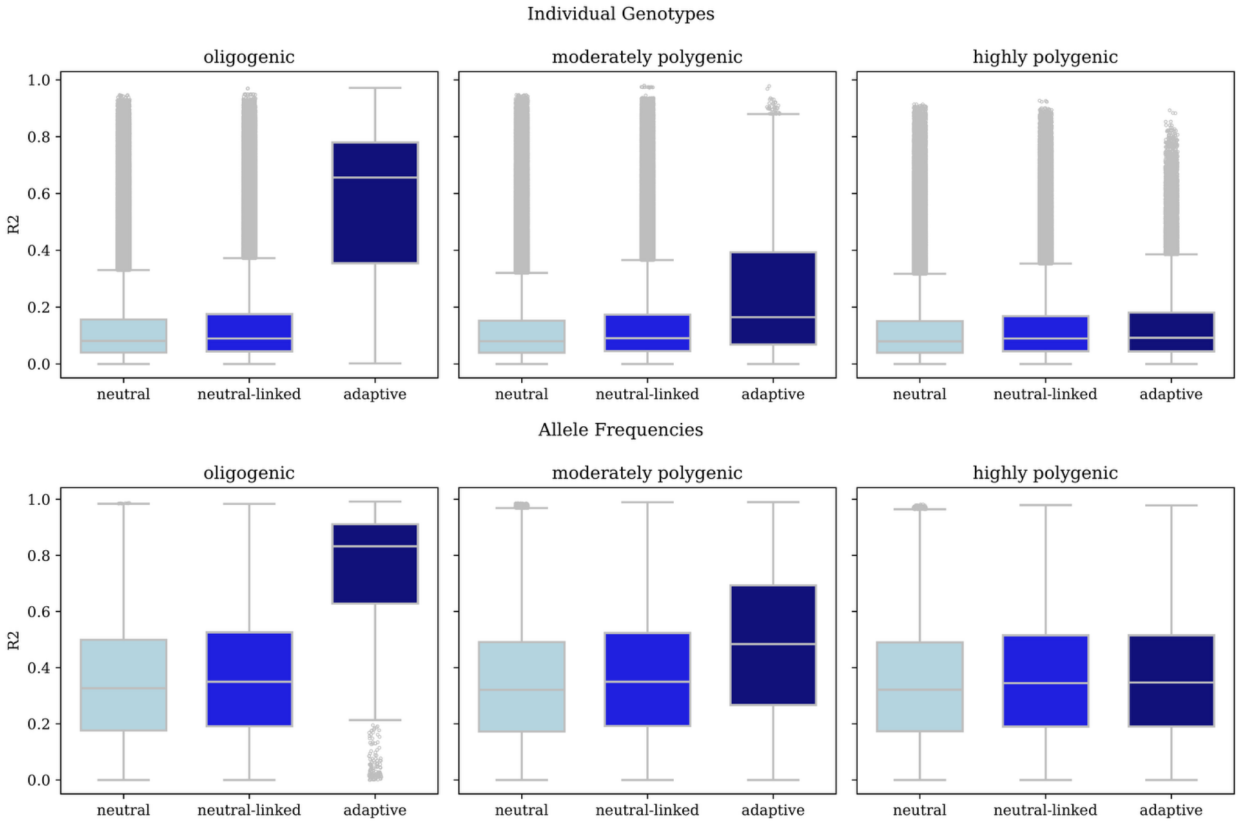


Figure S15 Differences in the predictive power of between adaptive and neutral loci (R^2) from random forest models depend on the genetic architecture underlying local adaptation. Data included in this figure are all overlapping loci incorporated into GF_{geno} and GF_{AF} models that completed training without failure. Code to create these figures can be found in SC 05.05.

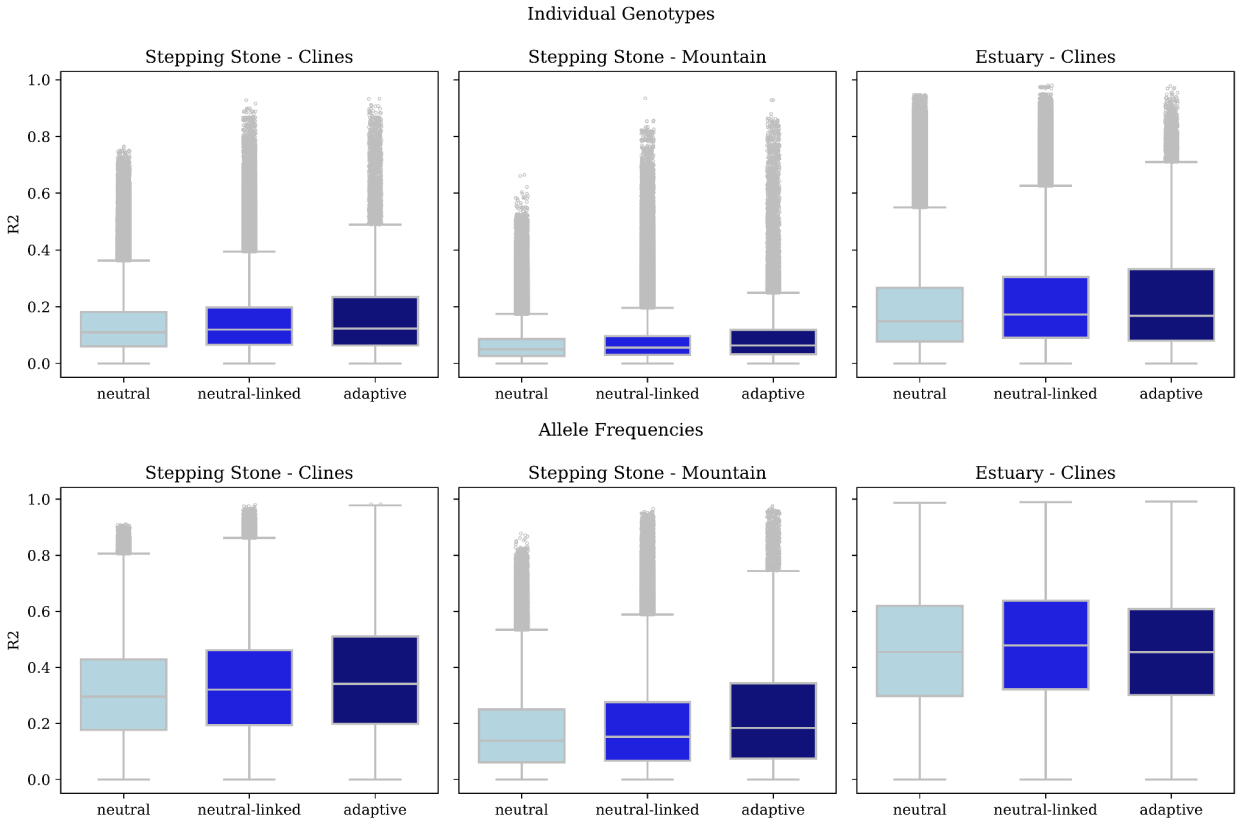


Figure S16 The predictive power of between adaptive and neutral loci (R^2) from random forest models are largely undifferentiated within spatially discrete landscapes (titles). Shown are R^2) values for loci incorporated into genotype-based models (top) and allele frequency-based models (bottom). Data included in this figure are all overlapping loci incorporated into GF_{geno} and GF_{AF} models that completed training without failure. Code to create these figures can be found in SC 05.05.

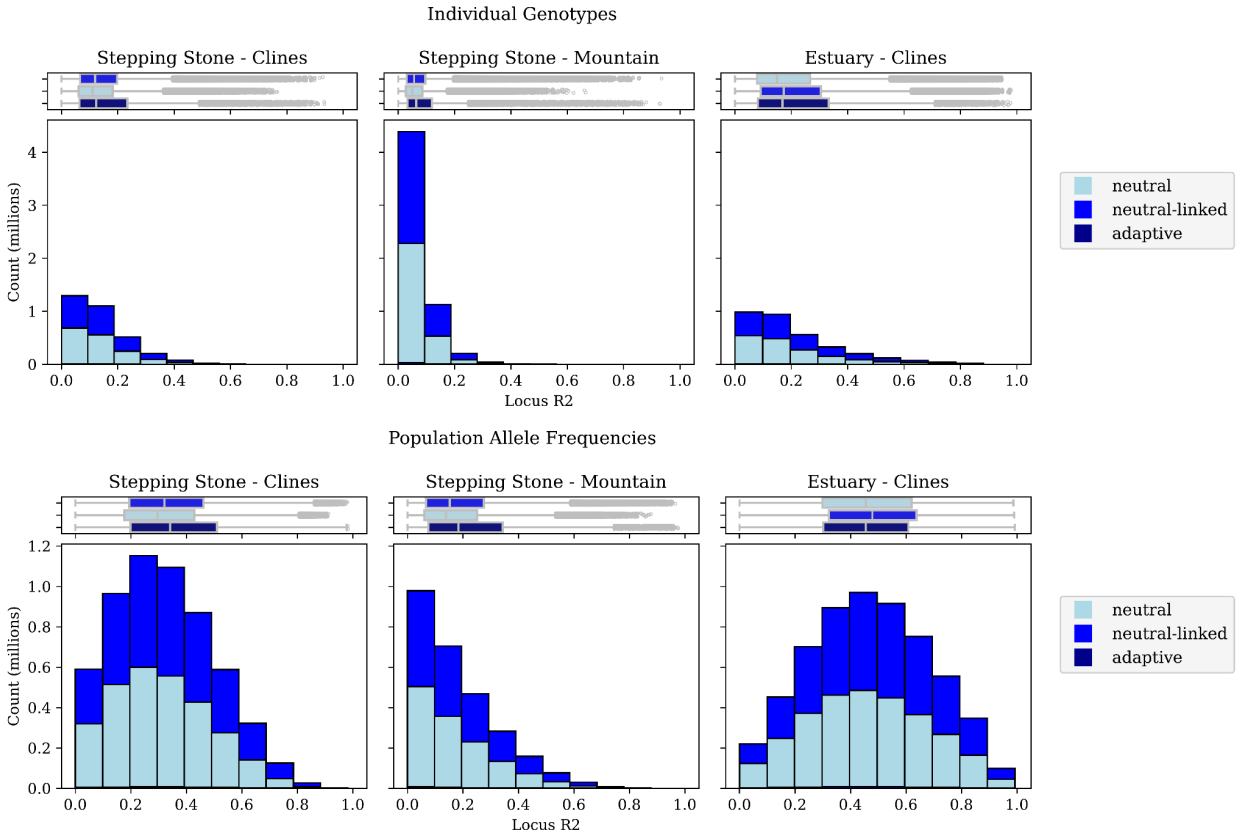


Figure S17 Despite little differences in R^2 between adaptive and neutral loci, distributions of R^2 differed among landscapes. Data included in this figure are all overlapping loci incorporated into GF_{geno} and GF_{AF} models that completed training without failure. Code to create these figures can be found in SC 05.05.

3 | References

Ellis, N., S. J. Smith, *a*nd C. R. Pitcher. 2012. Gradient forests: calculating importance gradients on physical predictors. Ecology 93:156–168.

Lind, B. M. 2024a. GitHub.com/ModelValidationProgram/MVP-offsets. Revision release (v1.0.1).
Zenodo. https://doi.org/10.5281/zenodo.11209812.

Lind, B. M. 2025. GitHub.com/brandonlind/geno_af_gradient_forests. Revision release (v1.1.0). Zenodo. https://doi.org/10.5281/zenodo.14946219.

Lind, B. M., and K. E. Lotterhos. 2024. The accuracy of predicting maladaptation to new environments with genomic data. Molecular Ecology Resources e14008.

Smith, S., N. Ellis, and C. Pitcher. 2012. gradientForest: Random Forest functions for the Census of Marine Life synthesis project - v0.1-24. Ecology 93:156–168.