

1 **A comparison of genomic forecasts based on**
2 **genotypes versus allele frequencies**

3 Brandon M. Lind*, Katie E. Lotterhos

4 Department of Marine and Environmental Sciences
5 Northeastern University Marine Science Center
6 430 Nahant Road, Nahant, MA 01908, USA.10 July 2025

7 10 July 2025

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

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

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

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

15 Abstract

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

33

34 **1 | Introduction**

35 Species regularly inhabit environmentally heterogenous ranges. Natural selection pressures
36 often differ spatially across these habitats, and can lead to genotype-environment interactions
37 that affect fitness (i.e., survival and reproduction). When the strength of local selection pressures
38 overcomes the forces of gene flow and drift (Blanquart et al. 2012), local adaptation can evolve
39 in the metapopulation. Local adaptation is defined by a higher mean fitness of populations in
40 sympatry than in allopatry (Kawecki and Ebert 2004; Blanquart et al. 2013). Indeed, local
41 adaptation is common across many species, such as plants(Leimu and Fischer 2008; Fournier-
42 Level et al. 2011; Leites and Garzón 2023), insects (Arguello et al. 2016), fish (Fraser et al. 2011),
43 corals (Thomas et al. 2022), and marine invertebrates (Burford et al. 2014; Bible and Sanford
44 2016).

45 Rapid environmental change, such as that projected for future climate, threatens to decouple
46 locally adaptive genetic variation from local fitness optima for many species (Aitken et al. 2008;
47 Hoffmann and Sgrò 2012). Previous methods used to understand the impact of environmental
48 change on species, such as species distribution models (SDMs), model an organism's ecological
49 niche to make predictions about the distribution and demographic performance of populations
50 across space (Guisan and Thuiller 2005; Thuiller et al. 2008; Pacifici et al. 2015). However, these
51 models ignore intraspecific variation and patterns of local adaptation (Capblancq et al. 2020).
52 While these models have been found to have some predictive ability relative to species
53 occurrence, their performance is often poor for predicting fitness of populations within their
54 native range from independent, ground-truth data (Lee-Yaw et al. 2022).

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

66 Regardless of debate around the technical interpretation of genomic offsets, investigators
67 generally agree that the performance of their predictions can be evaluated as the relationship
68 between genomic offsets and empirical estimates of fitness measured in an experimental context
69 of different genotypes being moved into different environments (Lotterhos 2024b). There is also
70 general agreement that if genomic offsets are predictive of genotype responses, there should be
71 a strong negative relationship between genomic offsets and fitness proxies in such experiments
72 (Capblancq et al. 2020; Rellstab et al. 2021; Lotterhos 2024b). We refer to an experiment
73 designed to test for this negative relationship between genomic offsets and fitness proxies as an
74 “evaluation”.

75 Genomic offset studies have been applied across myriad taxa, including birds (Bay et al. 2018;
76 Ruegg et al. 2018; Chen et al. 2022; DeSaix et al. 2022), fish (Brauer et al. 2023), herbaceous

77 plants (Exposito-Alonso et al. 2019), mammals (Rivkin et al. 2024), and trees (Fitzpatrick and
78 Keller 2015; Capblancq and Forester 2021; Fitzpatrick et al. 2021; Gougherty et al. 2021; Gugger
79 et al. 2021; Lachmuth et al. 2023, 2024; Lind et al. 2024). However, despite the increase in
80 popularity of these methods, there have been only a few empirical evaluations of their predictive
81 performance using experimental measures relating to fitness (but see Rhoné et al. 2020;
82 Capblancq and Forester 2021; Fitzpatrick et al. 2021; Gain et al. 2023; Lachmuth et al. 2023; Lind
83 et al. 2024).

84 While there are now several methods used to predict genomic offsets (see Table 1 in Capblancq
85 et al. 2020), the gradientForests method (*sensu* Fitzpatrick and Keller 2015) has undergone the
86 most evaluation. For instance, recent *in situ* evaluations of genomic offset from gradientForests
87 (GF_{offset}) in locally adapted tree species have often found the expected negative relationship
88 between fitness-related phenotypes from different populations transplanted to a common
89 garden (e.g., Fitzpatrick et al. 2021; Lachmuth et al. 2023, 2024; Lind et al. 2024). Further, these
90 studies have shown that predictions from GF_{offset} are often better than geographic or climate
91 distance alone (Fitzpatrick et al. 2021; Lind et al. 2024). *In silico* evaluations based on simulation
92 data have also shown that GF_{offset} generally outperformed other genomic offset methods (e.g.,
93 Rellstab et al. 2016; Capblancq and Forester 2021; Gain and François 2021) across a range of
94 scenarios (Lind and Lotterhos 2024), or gives similar predictive performance to simple measures
95 of environmental distance that use only causal environmental variables (Láruson et al. 2022).
96 Overall, several genomic offset methods, including GF_{offset} , have been shown to perform best
97 when there is a high degree of local adaptation in the metapopulation (Lind and Lotterhos 2024).
98 These evaluations of simulation data have generally shown high levels of predictive performance

99 of GF_{offset} within contemporary environments (Láruson et al. 2022; Lind and Lotterhos 2024), but
100 not when populations are moved into novel climates (Lind et al. 2024).

101 Despite these advances, the domain of applicability, or the circumstances under which model
102 predictions are valid (Lotterhos et al. 2022), are still limited for genomic offset measures, further
103 limiting confidence in their application more broadly. These circumstances ultimately encompass
104 the experimental design (e.g., the collection and format of training and evaluation data, the
105 choice of analyses) and the evolutionary history of the targeted populations (e.g., spatial patterns
106 of selection, genomic architecture). To date, the majority of genomic offset predictions and
107 evaluations have made use of population-level allele frequencies in modeling, and have made
108 predictions at the level of local populations for evaluation. On the other hand, many
109 management applications would benefit from genomic offset methods that make predictions at
110 the individual level, yet to date these methods remain largely undeveloped. Individual-level
111 predictions are particularly important when the species has a relatively small census size,
112 requiring offset strategies tailored to existing individuals or when limited management resources
113 or the carrying capacity of targeted restoration sites restrict the number of individuals that can
114 be moved or transplanted, making it essential to select individuals with the highest fitness.
115 Individual-level predictions are also relevant when variation in fitness is high relative to the
116 population average. In contrast, population-level predictions are most applicable when
117 population sizes are large, when the goal is to assess offsets to future climate change on a
118 regional scale, or when individuals for future translocation or transplantation will be randomly
119 selected from a larger population. Population-level predictions also present a cost-saving benefit,
120 for example when using pool-seq data. There are conceptual and statistical considerations

121 regarding how to fairly compare the performance of population-level vs. individual-level
122 evaluations, which we discuss in more detail below.

123 ***Considerations for model inputs***

124 Although the format of the genomic data that is input into models used for calculating genomic
125 offsets has typically been allele frequencies, many models could also take individual genotypes
126 as input (e.g., gradientForests; redundancy analysis, sensu Capblancq and Forester 2021; and
127 genetic.gap, sensu Gain et al. 2023). A comparison of allele frequency vs. genotype inputs is
128 shown in Fig. S1. We consider the case where individual genotypes collected from the same
129 population are assigned the same environmental data for that location, which would be common
130 when collecting groups of individuals, each from a small geographic area, and environmental data
131 is not available at the spatial scale of individual collections. This would also be the case when the
132 exact locations of individuals are unknown, such as with bulk seed lots collected for reforestation.

133 In contrast, individual-level environmental data could also be used to train the model, which
134 would be common when collecting individuals from a larger geographic area, and environmental
135 data is available for each individual. The important distinction between allele-frequency vs.
136 genotype inputs is that population-level allele frequencies average over within-population
137 variation and may miss key aspects of diversity that could lead to more accurate predictions. For
138 example, recent modeling efforts have revealed that adaptive trait clines can evolve even when
139 the underlying relationships between the frequency of adaptive alleles and the selective
140 environment are not monotonic (Lotterhos, 2023a). Individual-level genotype data may
141 therefore offer an advantage in these scenarios. Thus, it is unclear if training models using
142 genotypes from individuals can improve predictions relative to predictions from models that use

143 allele frequencies as input. Similarly, population-level environmental data can introduce
144 unnecessary noise by inaccurately assigning environmental values to individuals, particularly if
145 individuals from a population are sampled over a large geographic area.

146 On the other hand, there are also computational requirements to consider with regard to the
147 format of model inputs. For instance, because genomic offset models are trained using two
148 sources of data (genetic data and environmental data), the size of the dataset (determined by
149 the number of individuals or populations, as well as the number of loci and environmental
150 variables provided for training) will likely affect the computational memory or runtime required,
151 which may limit potential avenues of inference when computational resources are scarce.
152 Individual-level genotype and environmental data will always be a much larger dataset than
153 population-level allele frequency and environmental data, which will impact computational
154 requirements.

155 ***Considerations for model outputs***

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

165 When individuals collected from the same population are assigned the same environmental
166 data for that location as model input, the output of model predictions is at the population level
167 no matter the format of input genomic data (genotypes vs. allele frequencies). In other words,
168 any genomic offset method that relies exclusively on environmental data in its calculation of
169 offset (Box 1) will yield the same genomic offset value for any individuals transitioning from the
170 same current environment to the same new environment, irrespective of the model inputs or
171 their genetic compositions. While Box 1 focuses on GF_{offset} , predictions are also restricted to the
172 population level for other genomic offset methods using population-level inputs as well (e.g.,
173 genomic offset from redundancy analysis, Capblancq and Forester 2021; and from `genetic.offset`
174 and `genetic.gap`, Gain and François 2021; Gain et al. 2023). We consider this consequence of
175 population-level offset predictions in the next section.

176 ***Considerations for model evaluation***

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

187 ***Study Questions***

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

201 **2 | Methods**

202 ***2.1 | Explanation of input training data for genomic offset models***

203 We used a subset of the simulation datasets previously described and evaluated by Lotterhos
204 (2023a) and Lind & Lotterhos (2024). The first set of simulations were conducted on a 10 x 10
205 heterogeneous landscape of spatially discrete demes, such that all individuals from a deme
206 experienced the same environmental values and selection pressures. Here forward, we will refer
207 to demes as populations for the purposes of analysis, but recognize that they are not
208 demographically or evolutionarily independent units. We use these datasets to answer Q1, Q2,

209 Q3, and Q5 (see Explanation of Questions). Specifically, within each simulated dataset, a Wright-
210 Fisher metapopulation of 100 populations adapted to a heterogeneous landscape, where aspects
211 of the evolutionary history varied across simulation levels. Specifically, the subset of simulation
212 levels used here were only those datasets where metapopulations adapted to two environmental
213 variables on the landscape (`nlevels = 180/225`, using three replicates per simulation level). The
214 levels included: three landscapes that varied the geographic distribution of environmental
215 variables (Figure 1), five demographic scenarios that varied patterns of gene flow and population
216 sizes, three genic levels that varied the number of loci responding to selection (i.e., spanning
217 oligogenic to polygenic), and four pleiotropy levels, the degree of pleiotropy from causal
218 mutations, and different levels for the strength of selection from the causal environments (see
219 Fig. 1 from Lotterhos 2023a). Loci were simulated on 20 independent linkage groups, 10 of which
220 allowed mutations with effects on fitness. The univariate effect size of a QTN evolving without
221 pleiotropy was drawn from a normal distribution, and the bivariate effect size of QTNs under two
222 traits with pleiotropy was drawn from a multivariate normal distribution. Neutral mutations were
223 added to all 20 linkage groups with tree sequencing (Lotterhos 2023a). The scaled
224 metapopulation-level recombination rate ($\rho = 0.01$) resulted in a resolution of 0.001cM between
225 proximate bases where each linkage group had a total length of 5cM. This resolution is common
226 in SNP array or SNP chip designs, where loci are sampled broadly across the genome.

227 The causal environments imposing selection pressure included a temperature-like variable
228 (`temp`) that created a north-to-south cline on all landscapes (top panels Fig. 1), and a second
229 clinal environmental variable (`Env2`) that represented different analogies depending on the
230 landscape - in the *Stepping Stones - Clines* landscape, `Env2` formed a longitudinal cline; in the

231 *Stepping Stones - Mountains* landscape, *Env2* was analogous to elevation; in the *Estuary - Clines*
232 landscape, *Env2* was analogous to salinity gradients within coastal inlets that were only
233 connected by the outer marine environment (top panels Fig. 1). For more information on
234 simulations see Lotterhos (2023a) and Lind & Lotterhos (2024).

235 Ten individuals from each of the 100 populations were sampled for genetic data (1000 total).
236 For the evaluations here, we used a modified version of the *Adaptive Environment* workflow from
237 Lind & Lotterhos (2024). For an illustration of this workflow, see Fig. S2. All replicates were
238 evaluated using models trained with all adaptive environments. Further, while Lind & Lotterhos
239 (2024) used allele frequency (AF) data as input (i.e., allele frequencies calculated across individual
240 genotypes), here we also used the individual-level genotype data as input to compare to AF-
241 based models.

242 For each simulation replicate, genotypes of individuals were encoded as counts of the derived
243 allele, and derived allele frequencies from individuals were used for input. Loci with minor allele
244 frequency < 0.01 were removed. After identifying a set of loci for a given replicate, both
245 population- and individual-level data were filtered for the same markers and used as input for
246 training GF_{offset} models (hereafter GF_{AF} and GF_{geno}, respectively, to distinguish from evaluation
247 workflows in Table 1).

248 The genotypes within the dataframe passed to the gradientForest function were encoded
249 as integer class to ensure random forest regression instead of random forest classification.
250 Regression is more appropriate given the additive nature of the alleles underlying local
251 adaptation simulated here, and to enable direct comparison to models using allele frequencies.
252 While Lind & Lotterhos (2024) varied marker choice (e.g., *adaptive*, *neutral*, or *all* markers for

253 input), we vary only the number of markers randomly chosen from the full set of loci (see Q1
254 below).

255 Models were evaluated using *in silico* common gardens in each of 100 populations on the
256 landscape. For each of these 100 common gardens , model performance was quantified as the
257 rank correlation (Kendall's τ) between (i) the projected genomic offset to the common garden
258 and (ii) the known fitness (of either individuals or populations) in the garden environment. The
259 format of the evaluation data (i.e., population- or individual-level fitness) depended upon the
260 workflow (Table 1). If greater magnitudes of estimated genomic offset indicate higher degrees of
261 maladaptation, than a well-performing model would result in a negative relationship between
262 genomic offsets and fitness values. We found that the relationship between offset and fitness, as
263 well as offset and log(fitness), was non-linear for our data (Figs. S3-S7), so we chose the rank
264 correlation Kendall's τ to assess model performance. Kendall's τ is particularly well suited for
265 rank correlations when there are ties among either univariate or bivariate ranks.

266 In addition to the spatially discrete simulations described above, we evaluated offset
267 predictions using data from a single simulation of a non-Wright-Fisher metapopulation
268 undergoing range expansion from three refugia. We use this data to answer Q1 and Q4 (see
269 Explanation of Questions). Unlike the previous simulations described above, this model was
270 spatially continuous, with individuals occupying distinct locations across the landscape and
271 evolving variable degrees of admixture. Selection pressures on six moderately polygenic traits
272 with unconstrained pleiotropy were driven by six environmental variables derived from real-
273 world bioclimatic data from western Canada (bottom panel Fig. 1). Markers were distributed
274 across linkage groups in a similar fashion to genomes simulated for spatially discrete populations.

275 From this simulation, we sampled 1,000 individuals across the landscape. Using individual-level
276 genotypes, we formatted the resulting data similarly to the data from the spatially discrete
277 simulations described earlier, but the environmental data used as training input was formatted
278 in two distinct ways: i) individual-level environmental data, 2) population-averaged
279 environmental data. These two formats were then used to examine the impact of the format of
280 environmental data used for training (Q4) when trained using individual genotypes. To allow for
281 comparison to the spatially discrete evaluations, evaluation of these spatially continuous
282 datasets took place in 100 common gardens where the environment was the population-mean
283 environment of the assigned individuals. For more information on coding workflows
284 implemented here, see Supplemental Note S1.

285 **2.2 | Explanation of questions**

286 *Q1 | How does the number of markers used as input affect performance?*

287 Modern SNP datasets often have millions of markers distributed across the genome. However,
288 despite the promise of GF_{offset} performance, the current software implementation is often
289 computationally intensive and requires resources (e.g., allocated memory) beyond those
290 commonly available outside of high-performance computing clusters. This computational burden
291 will often limit implementations of GF_{offset} to those using fewer than around 20,000 loci, even on
292 systems with plentiful resources. Even so, previous evaluations of GF_{offset} have shown that
293 random markers often perform on par with adaptive marker sets (Fitzpatrick et al. 2021; Láruson
294 et al. 2022; Lachmuth et al. 2023; Lind and Lotterhos 2024). Still, in some cases the adaptive
295 (candidate) markers in empirical datasets are limited to several hundred markers, and it is not
296 known how well these small marker sets will capture genome-wide patterns of population

297 history. To understand how the number of loci can impact performance, we chose random sets
298 of loci (allowing potential overlap among sets) in the following sample sizes from each replicate
299 from both the spatially discrete as well as spatially continuous simulations: 500, 5 000, 10 000,
300 20 000. Comparisons of performance across marker set sizes were made within AF- and
301 genotype-based approaches.

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

313 In total, there were 216,000 potential evaluations per spatially discrete workflow (4 marker
314 sets * 180 levels * 3 replicates per level * 100 common gardens per replicate; Table 1). However,
315 269 replicates that used 20 000 input markers encoded as genotypes failed to complete in less
316 than one day or using less than 250Gb of memory. Therefore, unless otherwise noted, the
317 remaining questions used the 189100 evaluations common to both AF- and genotype-based
318 models.

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

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

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

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

335 The effect of AF- and genotype-based inputs on the performance of genomic offset models is
336 not well understood. To compare the performance of different model inputs on a common
337 ground, we used datasets from spatially discrete simulations to compare AF- and genotype-based
338 models that were evaluated using population-mean fitness (e.g., bottom row in Figure B2).

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

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

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

350 The analytical logistics of training and evaluating genomic offset models is computationally
351 burdensome beyond a single model, because often the input data should be varied to understand
352 the sensitivity of predictive outcomes. This input data ranges from the populations or markers
353 used to the uncertainty inherent in future climate projections, or even the climate data used for
354 training (DeSaix et al. 2022; Lachmuth et al. 2023; Lind et al. 2024). Because of this, it is necessary
355 to train potentially dozens of models for comparison. Benchmarking the time and memory
356 requirements necessary for these models will therefore be important during the planning and
357 training execution of such models in future settings. We compared walltime and memory usage
358 from all training phases from spatially discrete simulations and compared how the number of
359 genetic sources (individuals or populations) and number of loci provided to the model for training
360 affected resource usage. We obtained resource information using a custom parallel

361 implementation of the Slurm `sbatch` command in python. Training jobs were generally run on Intel
362 Xeon processors, ranging from 2.0GHz - 2.8GHz. More information on compute node processors
363 used for training GF models is available in Supplemental Note S2.

364 **3 | Results**

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

375 *Q1 | How does the number of markers used as input affect performance?*

376 The number of markers used to train both AF- and genotype-based models of GF_{offset} from
377 spatially discrete datasets did not differentially impact performance, as all comparisons within
378 levels had strong linear correlations (Pearson's $r > 0.9960$; Fig. 2; Fig. S8). Across both GF_{AF} and
379 GF_{geno} implementations, comparisons between models that were trained with more than 5000
380 loci had the strongest relationship, while comparisons to models trained with 500 loci had strong,
381 albeit relatively weaker, relationships (Pearson's $r < 0.9981$, Fig. 2, Fig. S8). We found similar

382 results when comparing models from the continuous space simulation (Fig. S9). While the
383 number of loci had little effect on model performance, we further explored results from our
384 spatially discrete evaluations to understand why this was the case, if there were any differences
385 in the way in which GF treated loci internally within its modeling framework, and if any
386 differences were due to experimental design or evolutionary history of the targeted populations.

387 Q1a - Why do marker sets perform similarly?

388 To understand why marker sets of 500 loci performed similarly to models trained with 10000
389 markers, we ran principal component analysis (PCA) on each marker set for each replicate from
390 spatially discrete datasets using either genotypes or allele frequencies using all loci from the
391 entire dataset. For each replicate and genetic source, we then calculated absolute correlations -
392 $\text{abs}(\text{Pearson's } r)$ - between axis loadings from the first three PC axes estimated from either 500
393 or 10 000 loci. We generally found a strong, linear relationship between the first three PC axes
394 loadings when comparing corresponding axes between the two marker sets (Fig. S10; e.g.,
395 between the first PC from each dataset) demonstrating that both higher- and lower-order axes
396 of genetic variation in the simulations were consistently captured by marker sets as small as
397 several hundred loci (Supplemental Code 05.01).

398 Q1b - How many loci were used by GF models?

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

403 incorporated into GF models increased with the number of loci provided to GF for training (Fig.
404 S11).

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

410 Q1c - Is the explanatory value of loci affected by experimental design or evolutionary history?
411 To further understand the effect of input loci on model performance, we explored locus-
412 specific R² values from random forest models internal to GF using the spatially discrete datasets.
413 R² is a measure of the goodness of fit of predicting genetic data using environmental values that
414 is estimated from internal random forest models for each locus. Overall, the R² values assigned
415 to loci by either GF_{geno} or GF_{AF} were positively correlated (Fig. S12). Even so, for these loci that
416 overlapped between GF_{AF} and GF_{geno} models, R² was generally lower for loci incorporated into
417 GF_{geno} models than GF_{AF} models (Fig. S12, Fig. S13). The additional loci used only by GF_{geno} models
418 (but not by GF_{AF} models) generally had lower R² than the loci that overlapped between GF_{geno}
419 and GF_{AF} (Fig. S14).

420 We examined how the evolutionary history of sampled populations (i.e., simulation
421 parameters) interacted with R² values. Generally, these parameters had little effect on the
422 distribution of R² values but there were exceptions. For instance, distributions of R² were higher
423 for *adaptive* (i.e., causal) than for neutral (i.e., unlinked to causal) and neutral-linked (i.e., linked
424 to causal) loci when the genetic architecture underlying adaptation was oligogenic (Fig. S15).

425 However, these differences in R^2 among the marker sets were less pronounced when the
426 architecture was moderately or highly polygenic. Contrastingly, other parameters had little effect
427 on differences in R^2 between *adaptive* and neutral loci classes, including the effect of landscape
428 (Fig. S16).

429 Despite little differences between *adaptive* and both classes of neutral markers within
430 landscapes, distributions of R^2 across all loci differed to greater degrees among landscapes (Figs.
431 S13, S17). For instance, median R^2 was greatest from loci in *Estuary - Clines* landscapes, followed
432 by *Stepping Stone - Clines* and next by *Stepping Stone - Mountain* landscapes. While there were
433 landscape differences between distributions of R^2 , these patterns were not indicative of overall
434 performance across landscapes. For instance, performance was highest in *Stepping Stone - Clines*
435 landscapes, followed by *Stepping Stone - Mountain* and *Estuary - Clines* landscapes (compare Fig.
436 S17 with Fig. S8). This overall pattern of GF_{AF} performance across landscapes has been shown
437 previously using similar datasets from the same simulations (Lind & Lotterhos, 2024) though R^2
438 was not previously compared.

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

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

445 Similar to genotype-based models, comparison of AF-based models showed elevated
446 performance scores when evaluated using population mean fitness ($GO_{AF,pop}$) compared to

447 predictions from the same models evaluated at the individual level ($GO_{AF,ind}$; Fig. 4B). This
448 statistical artifact is taken into account in the next section, where we compare performance from
449 evaluations at the population level.

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

451 To understand if within-population diversity captured by individual genotypes improved
452 model performance, we compared performance from models trained with genotypes to models
453 trained using allele frequencies at the same evaluation level (population-mean fitness). The
454 majority of models saw similar performance, but there were datasets that saw improved
455 performance from genotype inputs in spatially discrete datasets (Fig. 5). These were mainly from
456 *Stepping Stone - Mountain* landscapes where multiple populations on the landscape inhabit the
457 same (but geographically distinct) multivariate environment (top middle panel Fig. 1).

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

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

469 We hypothesized that averaging environmental data at the population level ($GO_{geno,ind(pop-env)}$)
470 would decrease accuracy relative to individual-level environmental values ($GO_{geno,ind(ind-env)}$) by
471 assigning incorrect environmental values to individuals. Comparing genotype-based models from
472 spatially continuous datasets, the performance was greater for models that used individual-level
473 environmental data than that from population-level environmental data (Fig. 6).

474 Q5 | How does the size of the dataset affect computational time and memory
475 requirements?

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

486 4 | Discussion

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

490 yet to be described (Lind and Lotterhos 2024; Lotterhos 2024b). Previous evaluations have shown
491 that genomic offsets should be rigorously explored before incorporating model inference into
492 management planning. This includes understanding model sensitivity to the choice of
493 populations used for training (Lind et al. 2024), accounting for uncertainty in climate forecasts
494 and projections to novel climates (DeSaix et al. 2022; Lachmuth et al. 2023; Lind and Lotterhos
495 2024), sampling locally adapted populations (Rellstab et al. 2021; Lind and Lotterhos 2024), as
496 well as considering other important factors affecting genotype-climate relationships, such as
497 neutral demographic effects driven by differences in effective population sizes of sampled
498 populations (Láruson et al. 2022).

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

510 **4.1 | The usefulness of individual genotypes**

511 Many previous implementations of GF_{offset} have used population allele frequencies to train
512 models (e.g., Lachmuth et al. 2023; Lind et al. 2024), and in some cases have done so despite the
513 availability of individual-level genotypic data (e.g., Fitzpatrick and Keller 2015; DeSaix et al. 2022;
514 Láruson et al. 2022; Lind and Lotterhos 2024; Tigano et al. 2024). Several factors may contribute
515 to the common use of allele frequencies instead of genotypes for GF_{offset} model training. First,
516 instructional genomic offset vignettes (including published code) often give examples using allele
517 frequencies. Second, investigators may be using pool-seq datasets where DNA extractions from
518 multiple individuals are pooled at the population level prior to sequencing (see Schlötterer et al.
519 2014 for a technical review) and therefore only have allele frequency information. This pool-seq
520 approach has been used in several studies (e.g., Gugger et al. 2021; Nielsen et al. 2021; Lind et
521 al. 2024). A third consideration could also be computational resources. Indeed, the first
522 implementation of GF_{offset} was run on a single laptop (Fitzpatrick and Keller 2015), and
523 subsequent analyses on high performance computing clusters have favored allele frequency data
524 because of reduced run times (e.g., Lind and Lotterhos 2024). Finally, the use of allele
525 frequencies may have been motivated by the fact that methods such as GF_{offset} cannot
526 differentiate performance between individuals that come from the same multivariate
527 environment (e.g., when assigning environmental values to individuals from bulk seed lots
528 collected regionally for reforestation), and therefore since model output takes place at the
529 environmental level, population allele frequencies were used as input.

530 Despite this trend, we showed that genotypic data, which provides information on within-
531 population variation, offers some advantages over allele frequency inputs. The ultimate reason

532 for this gain in performance seems to be from the incorporation of loci in genotype models that
533 were excluded from allele frequency models. These incorporated loci each generally explained
534 much less variation (R^2) within internal random forest models of GF than from loci common
535 between model types. Although these loci were not necessarily adaptive, these additional loci
536 nonetheless changed the weighting of environmental values in internal cumulative importance
537 curves in a way that led to more accurate genomic offset calculations in many *Stepping Stone -*
538 *Mountain* landscapes. Ultimately the change in weights between model types were due to the
539 differences in numerical values of genotypes (0, 1, or 2) and allele frequencies (a range from 0.01-
540 0.99) that likely impacted the trees within the random forest models. For each tree in a random
541 forest, algorithmic decisions regarding ultimate tree depth or splits at internal nodes (due to
542 differences in the calculation of impurity scores between genotypes and allele frequency values
543 when splitting data using climate values at internal nodes) resulted in differences in the
544 magnitude of R^2 for a given locus. While our implementation of gradientForests ensured that our
545 individual allele counts (i.e., genotypes) were encoded as a continuous variable (i.e., using
546 random forest regression), future studies could investigate the predictive accuracy from models
547 where genetic data is encoded as categorical genotypes (i.e., random forest classification).

548 While we evaluated the impact of genotype and allele frequency inputs on GF_{offset} performance,
549 such considerations are also relevant to other existing genomic offset methods. For instance,
550 genomic offset predictions from redundancy analysis (sensu Capblancq and Forester 2021) can
551 accept both individual genotypes or allele frequencies as input. Furthermore, analyses such as
552 RONA that have traditionally been carried out using allele frequencies (following Rellstab et al.

553 2016), could also be modified to model genotypes instead. Further investigation is warranted to
554 understand the impacts of genomic data formats for this and other methods.

555 **4.2 | How many markers are enough?**

556 Previous evaluations of GF_{offset} have found that predictive accuracy is relatively insensitive to
557 the choice of markers used, for instance between sets of random loci and sets of candidate loci
558 putatively involved in local adaptation (Fitzpatrick et al. 2021; Láruson et al. 2022; Lachmuth et
559 al. 2023; Lind et al. 2024). However, the effect of the number of markers has received less
560 attention. In our study, less dense marker sets (~500 loci) performed similarly to more dense
561 marker sets (~20 000 loci) because they were sufficient to capture similar levels of genetic
562 structure that was estimated with PCA axes. For empirical datasets, a small number of markers
563 (e.g., ~500) may not perform as well as they do in this study because small marker sets may not
564 capture all aspects of population structure or of adaptive genetic variation through linkage
565 disequilibrium. Model sensitivity to small marker sets may be particularly relevant when using
566 technologies that may sample unevenly across the genome, such as with restriction site-
567 associated sequencing (RAD-seq; Lowry et al. 2017), though future investigation is warranted. To
568 this end, it will be important for future studies to understand the extent to which the genotypic
569 data is distributed across the genome and how well the loci sample across recombination or
570 haplotype blocks. Reporting the extent of linkage disequilibrium decay will be an important step
571 in this direction, as understanding the extent of decay will inform the approximate spacing of
572 markers necessary to capture evolutionary history across the genome. Using annotations from
573 reference genomes will also be important to understand the extent to which loci represent
574 coding (exonic) and non-coding (intronic, intergenic) regions within the genome.

575 When feasible, future studies should demonstrate the sensitivity of their predictions on
576 different sets of loci. For instance, in the case of a candidate marker set, using sets of random
577 loci of similar numbers can be used to understand how predictions change when the input loci
578 are varied. For datasets with a large number of loci, multiple runs of different subsets of loci could
579 be compared. Studies with large datasets could also consider pruning loci for linkage
580 disequilibrium, as it is not yet known how over- or underrepresentation of groups of linked loci
581 affect model outcomes. This is particularly relevant to gradientForests, where multiple loci are
582 used to weight environmental values that are ultimately used to make predictions.

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

590 **4.3 | Considerations for future experimental design and evaluation of genomic offsets**

591 Central to decisions regarding data generation for genomic offset models are the specific aims
592 or hypotheses targeted by investigators and how model predictions of genomic offsets are
593 interpreted biologically (Lotterhos 2024a, 2024b). Generally, these aims fall into either making
594 predictions for a restoration project or for population-level responses to climate change
595 (Capblancq et al. 2020; Rellstab et al. 2021). For instance, in the case of a restoration scenario,
596 genomic offset predictions could be used to rank potential donor populations for a restoration

597 site, as evaluated in this study. For such cases, investigators are seeking to best identify the
598 population(s) with the greatest fitness in the restoration environment compared to other
599 populations under consideration (i.e., a prediction of fitness differences among genotypes in a
600 single environment). In other cases, investigators may wish to understand the extent of *in situ*
601 maladaptation of a focal population in relation to the predicted disruptions in environmental
602 optima resulting from climate change (i.e., a prediction of the change in fitness of a single
603 genotype from a current to a future environment). Recently, Lotterhos (2024a) showed that
604 there are many ways to calculate fitness differences (i.e., fitness offsets), that the correct
605 calculation depends on the context, and that various calculations of fitness offsets may not be
606 correlated with each other or with genomic offset predictions. These differences are ultimately
607 related to the pattern of local adaptation in the metapopulation. The analysis from Lotterhos
608 (2024a) suggested that common gardens, which have been the primary scenario under which
609 genomic offsets have been evaluated (e.g., Capblancq and Forester 2021; Fitzpatrick et al. 2021;
610 Gougherty et al. 2021; Lachmuth et al. 2023; Lind et al. 2024), including this study, may not
611 adequately serve as a proxy of model accuracy when considering potential *in situ* maladaptation
612 to climate change (see also Lind and Lotterhos 2024). Experimental designs have been proposed
613 that could be used to evaluate model accuracy in both scenarios (Lotterhos 2024a, 2024b).

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

619 calculated using all data points). In a restoration scenario, the relationship between prediction
620 and ground-truth measurements for all data points may not capture the level of accuracy desired,
621 which is ultimately determined by the purpose of the experiment - e.g., to identify the population
622 source(s) with highest fitness at a restoration site. In these cases, the predictive accuracy of
623 populations that are least suitable for the environment is not a priority. Measures of model
624 accuracy across all data points also complicates model comparison within and between studies,
625 where comparison of the magnitude of the correlation coefficient or R^2 estimated from multiple
626 models is not necessarily indicative of differences between models in identifying the most
627 suitable population(s) for a given site. Furthermore, models estimating the extent of fitness
628 reduction associated with increasing offset (e.g., linear models that report R^2), while potentially
629 valuable for management, are further complicated by the non-linear relationship between these
630 variables, as demonstrated here. Importantly, there may be cases where the relationship
631 between offset and fitness is linear across populations predicted to have intermediate fitness for
632 a given environment (and thus contribute to the significance of a linear model). However, it may
633 also be the case that this relationship becomes increasingly non-linear for individuals predicted
634 to be most or least suited for that environment - as was the case in both the spatially continuous
635 and spatially discrete space Stepping Stones simulations.

636 In this simulation study, both the offset-fitness relationship and the offset-log(fitness)
637 relationship were monotonic but non-linear. Empirical studies have found different shapes in the
638 relationship between fitness proxies and genetic offset. For instance, a study on balsam poplar
639 (*Populus balsamifera*) found a parabolic relationship between height increment and GF_{offset}
640 (Fitzpatrick et al. 2021), a study on pearl millet landraces (*Cenchrus americanus*) found a linear

641 relationship between mean seed weight and GF_{offset} (Rhoné et al. 2020), and a follow-up study
642 on pearl millet using the same dataset found a linear relationship between log-transformed mean
643 seed weight and GF_{offset} (Gain et al. 2023). These different shapes complicate the application of
644 genomic offsets in practice. In our study, we found that for populations with intermediate fitness
645 in a common garden, the relationship between offset and fitness was approximately linear, but
646 the relationship became increasingly more nonlinear when including populations with extremely
647 high or low fitness. Thus, the observed fitness-offset relationship in an empirical context might
648 depend on which subset of populations from the entire metapopulation are included in a
649 common garden experiment.

650 Because accuracy across all data points is not a primary concern for evaluating genomic offset
651 predictions in a restoration scenario, evaluation statistics that capture the accuracy of the
652 predictions for the most suitable populations will likely be useful for model comparisons going
653 forward. Similar arguments can be made for evaluations of genomic offset in the context of *in*
654 *situ* climate change where the goal is to identify the populations that are least suited for their
655 future climate. In future work, predictive accuracy of models across all data points (e.g., rank
656 correlations or R^2 from linear models) can be compared to evaluations where the underlying
657 hypothesis tests whether the populations with highest (or lowest) fitnesses are those enriched
658 in the extreme ranks of genomic offset predictions. In practice, the number of population or
659 individual ranks that are relevant will likely vary by system, and may be influenced by
660 management goals relating to genetic diversity and effective population sizes. For instance,
661 predictions that are based on population mean fitness therefore serve as a measure of accuracy
662 when transplanting large numbers of individuals (perhaps from several populations), and may

663 not best serve interests where relatively few individuals will be moved. Such a scenario is
664 analogous to assisted migration, where populations are moved within a species range to match
665 environmental optima in changing climates (Aitken and Whitlock 2013) . While population-level
666 predictions may provide useful insight for management, the development of methods that
667 predict genomic offsets at the individual level warrant both further investigation and
668 development.

669 ***4.4 | Conclusions and future directions***

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

675 The experimental design and environmental databases used to generate data is an important
676 factor beyond identifying suitable systems (i.e., populations that are locally adapted to
677 measurable environmental forces). Future evaluations should explore the sensitivity of model
678 predictions to these experimental decisions. Although the number of markers required to
679 maintain model accuracy may vary by system, the evaluations we present here and elsewhere
680 (e.g., Láruson et al. 2022; Lind and Lotterhos 2024) , suggest that sampling across the climate
681 space of the target organism is likely an important consideration, as is the ability accurately
682 estimate allele frequencies within populations with adequate sample sizes. Indeed, studies
683 comparing predictions from offset models trained using either 240 candidate loci or >335,000
684 loci from the complete dataset found similar relationships with common garden fitness proxies

685 (e.g., Lachmuth et al. 2023). In this study, we had a relatively large sample size for training and
686 evaluation with coverage of the landscape (1000 individuals - 10 from each of 100 populations).
687 Similarly, our evaluations of continuous space simulations highlight the importance of accurate
688 environmental values at the individual level, a finding further underscored previously by DeSaix
689 et al. (2022) . Even so, empirical evaluations have still found significant negative relationships
690 between offsets and fitness proxies with smaller numbers of individuals or populations used to
691 train or evaluate the model (e.g., Fitzpatrick et al. 2021), so sampling as extensive as what we
692 simulated may not be needed in practice. Nevertheless, the necessary sampling effort is likely to
693 vary among study systems and depend on the heterogeneity of the selective environment and
694 the spatial scale of adaptation. Further, current genomic offset methods are largely
695 phenomenological and lack incorporation of mechanisms underlying local adaptation, such as
696 potential fitness trade-offs underlying species distributions and population dynamics (Bastias et
697 al. 2024). When such mechanisms are uncovered (Wadgymar et al. 2017) they may be useful to
698 improve ecological forecasting models (Getz et al. 2018; Waldvogel et al. 2020). Even when such
699 mechanisms are not incorporated into genomic offset approaches, such understanding may offer
700 further avenues to defining the domain of applicability of these genomic offset methods
701 (Lotterhos et al. 2022).

702 Acknowledgments

703 This research was funded by NSF-- 2043905 (KEL) and Northeastern University. The funding
704 bodies did not have any role in the design of the study, analysis, interpretation of results or in
705 writing of the manuscript. We would like to acknowledge Jason Selwyn, Madeline Eppley,

706 Annabel Hughes, Sarit Truskey, Stephen Keller, and an anonymous reviewer for helpful and
707 constructive comments on earlier versions of our manuscript.

708 **Data Availability**

709 We reference the analysis code in the text and figure legends by designating Supplemental Code
710 using a directory numbering system from our servers (as opposed to the order listed in the
711 manuscript). Supplemental Code includes Jupyter Notebooks (*.ipynb). For example, for
712 Notebook 4 in Directory 2, we refer to Supplemental Code 02.04. All code, as well as dataframes
713 containing evaluation results, are archived on Zenodo.org (Lind, 2025), and includes a link to the
714 GitHub repository. This archive also includes the evaluation results from all workflows.
715 Notebooks are best viewed within a local jupyter or jupyter lab session (to enable cell output
716 scrolling/collapsing), but can also be viewed at nbviewer.jupyter.org using the web link in the
717 archive's README on GitHub. Analyses were carried out primarily using python v3.8.5 and R
718 v3.5.1. The yml files to reconstruct the coding environments for the Rv3.5.1 (r35.yml) and python
719 v3.8.5 (mvp_env.yml) environments have been previously archived (Lind, 2024). Exact package
720 and code versions are available at the top of each notebook. More information on coding
721 workflows and coding environments can be found in Supplemental Note S1. Data used for
722 analysis have been archived previously (Lotterhos 2023b) .

723 **Author Contributions**

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

728 References

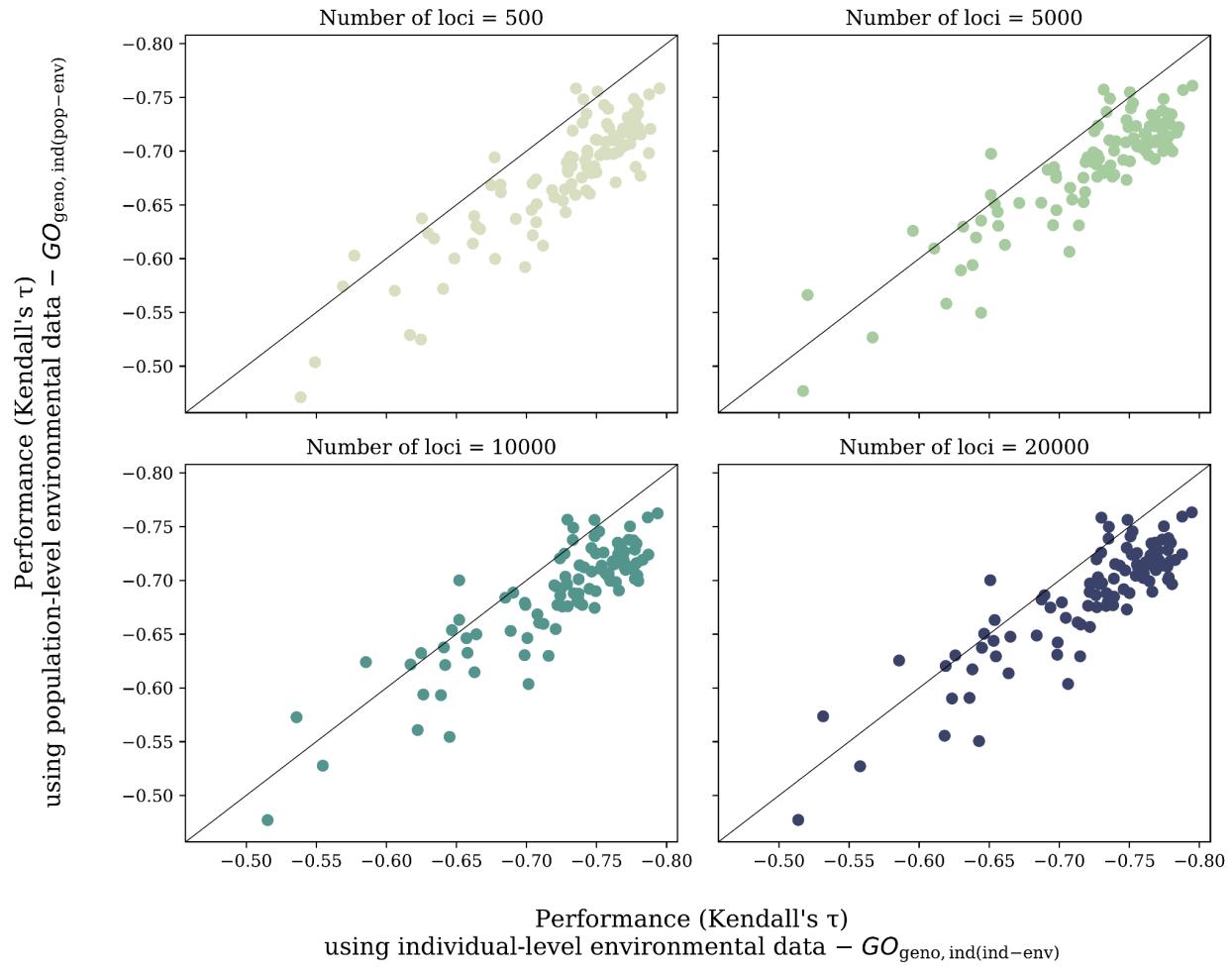
- 729
- 730
- 731 Aitken, S. N., and M. C. Whitlock. 2013. Assisted Gene Flow to Facilitate Local Adaptation to
732 Climate Change. *Annual Review of Ecology, Evolution, and Systematics* 44:367–388.
- 733 Aitken, S. N., S. Yeaman, J. A. Holliday, T. Wang, and S. Curtis-McLane. 2008. Adaptation,
734 migration or extirpation: climate change outcomes for tree populations. *Evolutionary
735 Applications* 1:95–111.
- 736 Arguello, J. R., M. Cardoso-Moreira, J. K. Grenier, S. Gottipati, A. G. Clark, and R. Benton. 2016.
737 Extensive local adaptation within the chemosensory system following *Drosophila
738 melanogaster*'s global expansion. *Nature Communications* 7:ncomms11855.
- 739 Bastias, C. C., A. Estarague, D. Vile, E. Gaignon, C.-R. Lee, M. Exposito-Alonso, C. Violle, et al.
740 2024. Ecological trade-offs drive phenotypic and genetic differentiation of *Arabidopsis thaliana*
741 in Europe. *Nature Communications* 15:5185.
- 742 Bay, R. A., R. J. Harrigan, V. L. Underwood, H. L. Gibbs, T. B. Smith, and K. Ruegg. 2018. Genomic
743 signals of selection predict climate-driven population declines in a migratory bird. *Science*
744 359:83–86.
- 745 Bible, J. M., and E. Sanford. 2016. Local adaptation in an estuarine foundation species:
746 Implications for restoration. *Biological Conservation* 193:95–102.
- 747 Blanquart, F., S. Gandon, and S. L. Nuismer. 2012. The effects of migration and drift on local
748 adaptation to a heterogeneous environment. *Journal of Evolutionary Biology* 25:1351–1363.
- 749 Blanquart, F., O. Kaltz, S. L. Nuismer, and S. Gandon. 2013. A practical guide to measuring local
750 adaptation. (D. Ebert, ed.) *Ecology Letters* 16:1195–1205.
- 751 Brauer, C. J., J. Sandoval-Castillo, K. Gates, M. P. Hammer, P. J. Unmack, L. Bernatchez, and L. B.
752 Beheregaray. 2023. Natural hybridization reduces vulnerability to climate change. *Nature
753 Climate Change* 1–8.
- 754 Burford, M., J. Scarpa, B. Cook, and M. Hare. 2014. Local adaptation of a marine invertebrate
755 with a high dispersal potential: evidence from a reciprocal transplant experiment of the eastern
756 oyster *Crassostrea virginica*. *Marine Ecology Progress Series* 505:161–175.
- 757 Capblancq, T., M. C. Fitzpatrick, R. A. Bay, M. Exposito-Alonso, and S. R. Keller. 2020. Genomic
758 Prediction of (Mal)Adaptation Across Current and Future Climatic Landscapes. *Annual Review of
759 Ecology, Evolution, and Systematics* 51:245–269.

- 760 Capblancq, T., and B. R. Forester. 2021. Redundancy analysis: A Swiss Army Knife for landscape
761 genomics. *Methods in Ecology and Evolution* 1–12.
- 762 Chen, Y., Z. Jiang, P. Fan, P. G. P. Ericson, G. Song, X. Luo, F. Lei, et al. 2022. The combination of
763 genomic offset and niche modelling provides insights into climate change-driven vulnerability.
764 *Nature Communications* 13:4821.
- 765 DeSaix, M. G., T. L. George, A. E. Seglund, G. M. Spellman, E. S. Zavaleta, and K. C. Ruegg. 2022.
766 Forecasting climate change response in an alpine specialist songbird reveals the importance of
767 considering novel climate. *Diversity and Distributions* 28:2239–2254.
- 768 Ellis, N., S. J. Smith, and C. R. Pitcher. 2012. Gradient forests: calculating importance gradients
769 on physical predictors. *Ecology* 93:156–168.
- 770 Exposito-Alonso, M., M. Exposito-Alonso, R. G. Rodríguez, C. Barragán, G. Capovilla, E. Chae, J.
771 Devos, et al. 2019. Natural selection on the *Arabidopsis thaliana* genome in present and future
772 climates. *Nature* 573:126–129.
- 773 Fitzpatrick, M. C., V. E. Chhatre, R. Y. Soolanayakanahally, and S. R. Keller. 2021. Experimental
774 support for genomic prediction of climate maladaptation using the machine learning approach
775 Gradient Forests. *Molecular Ecology Resources* 00:1–17.
- 776 Fitzpatrick, M. C., and S. R. Keller. 2015. Ecological genomics meets community-level modelling
777 of biodiversity: mapping the genomic landscape of current and future environmental
778 adaptation. (M. Vellend, ed.) *Ecology Letters* 18:1–16.
- 779 Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, and A. M. Wilczek. 2011. A
780 Map of Local Adaptation in *Arabidopsis thaliana*. *Science* 334:86–89.
- 781 Fraser, D. J., L. K. Weir, L. Bernatchez, M. M. Hansen, and E. B. Taylor. 2011. Extent and scale of
782 local adaptation in salmonid fishes: review and meta-analysis. *Heredity* 106:404–420.
- 783 Gain, C., and O. François. 2021. LEA 3: Factor models in population genetics and ecological
784 genomics with R. *Molecular Ecology Resources* 21:2738–2748.
- 785 Gain, C., B. Rhoné, P. Cubry, I. Salazar, F. Forbes, Y. Vigouroux, F. Jay, et al. 2023. A Quantitative
786 Theory for Genomic Offset Statistics. *Molecular Biology and Evolution* 40:msad140.
- 787 Getz, W. M., C. R. Marshall, C. J. Carlson, L. Giuggioli, S. J. Ryan, S. S. Romañach, C. Boettiger, et
788 al. 2018. Making ecological models adequate. *Ecology Letters* 21:153–166.
- 789 Gougherty, A. V., S. R. Keller, and M. C. Fitzpatrick. 2021. Maladaptation, migration and
790 extirpation fuel climate change risk in a forest tree species. *Nature Climate Change* 1–15.

- 791 Gugger, P. F., S. T. Fitz-Gibbon, A. Albarrán-Lara, J. W. Wright, and V. L. Sork. 2021. Landscape
792 genomics of *Quercus lobata* reveals genes involved in local climate adaptation at multiple
793 spatial scales. *Molecular Ecology* 30:406–423.
- 794 Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple
795 habitat models. *Ecology Letters* 8:993–1009.
- 796 Hoffmann, A. A., and C. M. Sgrò. 2012. Climate change and evolutionary adaptation. *Nature*
797 470:479–485.
- 798 Kawecki, T. J., and D. Ebert. 2004. Conceptual issues in local adaptation. *Ecology Letters*
799 7:1225–1241.
- 800 Lachmuth, S., T. Capblancq, S. R. Keller, and M. C. Fitzpatrick. 2023. Assessing uncertainty in
801 genomic offset forecasts from landscape genomic models (and implications for restoration and
802 assisted migration). *Frontiers in Ecology and Evolution* 11:1155783.
- 803 Lachmuth, S., T. Capblancq, A. Prakash, S. R. Keller, and M. C. Fitzpatrick. 2024. Novel genomic
804 offset metrics integrate local adaptation into habitat suitability forecasts and inform assisted
805 migration. *Ecological Monographs* 94:e1593.
- 806 Láruson, Á. J., M. C. Fitzpatrick, S. R. Keller, B. C. Haller, and K. E. Lotterhos. 2022. Seeing the
807 forest for the trees: Assessing genetic offset predictions from gradient forest. *Evolutionary
808 Applications* 15:403–416.
- 809 Lee-Yaw, J. A., J. L. McCune, S. Pironon, and S. N. Sheth. 2022. Species distribution models
810 rarely predict the biology of real populations. *Ecography* 2022.
- 811 Leimu, R., and M. Fischer. 2008. A meta-analysis of local adaptation in plants. (A. Buckling,
812 ed.)*PLoS ONE* 3:e4010.
- 813 Leites, L., and M. B. Garzón. 2023. Forest tree species adaptation to climate across biomes:
814 Building on the legacy of ecological genetics to anticipate responses to climate change. *Global
815 Change Biology* 29:4711–4730.
- 816 Lind, B. M., R. Candido-Ribeiro, P. Singh, M. Lu, D. O. Vidakovic, T. R. Booker, M. C. Whitlock, et
817 al. 2024. How useful is genomic data for predicting maladaptation to future climate? *Global
818 Change Biology* 30:e17227.
- 819 Lind, B. M., and K. E. Lotterhos. 2024. The accuracy of predicting maladaptation to new
820 environments with genomic data. *Molecular Ecology Resources* e14008.
- 821 Lotterhos, K. E. 2023a. The paradox of adaptive trait clines with nonclinal patterns in the
822 underlying genes. *Proceedings of the National Academy of Sciences* 120.

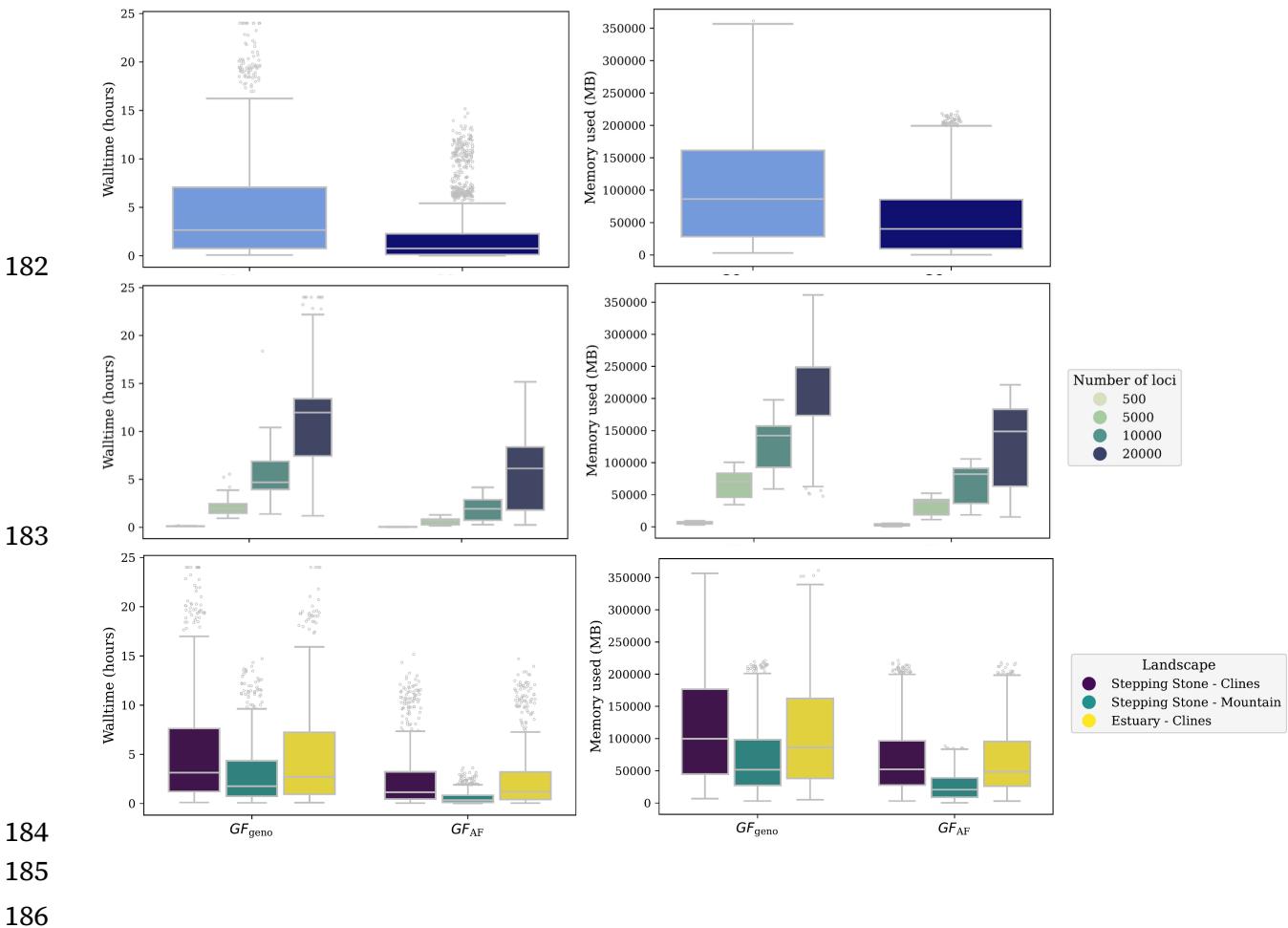
- 823 Lotterhos, K. E. 2023b. Output model data from paradox of adaptive trait clines with non-clinal
824 patterns in the underlying genes (Model Validation Program project). Biological and chemical
825 oceanography data management office (BCO-DMO). (version 1) version date 2023-02-13.
826 <https://doi.org/10.26008/1912/bco-dmo.889769.1>.
- 827 Lotterhos, K. E. 2024a. Interpretation issues with “genomic vulnerability” arise from conceptual
828 issues in local adaptation and maladaptation. Evolution Letters qrae004.
- 829 Lotterhos, K. E. 2024b. Principles in experimental design for evaluating genomic forecasts.
830 Methods in Ecology and Evolution 15:1466–1482.
- 831 Lotterhos, K. E., M. C. Fitzpatrick, and H. Blackmon. 2022. Simulation Tests of Methods in
832 Evolution, Ecology, and Systematics: Pitfalls, Progress, and Principles. Annual Review of Ecology,
833 Evolution, and Systematics 53:113–136.
- 834 Lowry, D. B., S. Hoban, J. L. Kelley, K. E. Lotterhos, L. K. Reed, M. F. Antolin, and A. Storfer. 2017.
835 Responsible RAD: Striving for best practices in population genomic studies of adaptation.
836 Molecular Ecology Resources 17:366–369.
- 837 Nielsen, E. S., R. Henriques, M. Beger, and S. Heyden. 2021. Distinct interspecific and
838 intraspecific vulnerability of coastal species to global change. Global Change Biology 27:3415–
839 3431.
- 840 Pacifici, M., W. B. Foden, P. Visconti, J. E. M. Watson, S. H. M. Butchart, K. M. Kovacs, B. R.
841 Scheffers, et al. 2015. Assessing species vulnerability to climate change. Nature Climate Change
842 5:215–224.
- 843 Rellstab, C., B. Dauphin, and M. Exposito-Alonso. 2021. Prospects and limitations of genomic
844 offset in conservation management. Evolutionary Applications 14:1202–1212.
- 845 Rellstab, C., S. Zoller, L. Walthert, I. Lesur, A. R. Pluess, R. Graf, C. Bodénès, et al. 2016.
846 Signatures of local adaptation in candidate genes of oaks (*Quercus* spp.) with respect to present
847 and future climatic conditions. Molecular Ecology 25:5907–5924.
- 848 Rhoné, B., D. Defrance, C. Berthouly-Salazar, C. Mariac, P. Cubry, M. Couderc, A. Dequincey, et
849 al. 2020. Pearl millet genomic vulnerability to climate change in West Africa highlights the need
850 for regional collaboration. Nature Communications 11:5274.
- 851 Rivkin, L. R., E. S. Richardson, J. M. Miller, T. C. Atwood, S. Baryluk, E. W. Born, C. Davis, et al.
852 2024. Assessing the risk of climate maladaptation for Canadian polar bears. Ecology Letters
853 27:e14486.

- 854 Ruegg, K., R. A. Bay, E. C. Anderson, J. F. Saracco, R. J. Harrigan, M. Whitfield, E. H. Paxton, et al.
855 2018. Ecological genomics predicts climate vulnerability in an endangered southwestern
856 songbird. *Ecology Letters* 21:1085–1096.
- 857 Schlötterer, C., R. Tobler, R. Kofler, and V. Nolte. 2014. Sequencing pools of individuals —
858 mining genome-wide polymorphism data without big funding. *Scientific Reports* 15:749–763.
- 859 Smith, S., N. Ellis, and C. Pitcher. 2012. gradientForest: Random Forest functions for the Census
860 of Marine Life synthesis project - v0.1-24. *Ecology* 93:156–168.
- 861 Thomas, L., J. N. Underwood, N. H. Rose, Z. L. Fuller, Z. T. Richards, L. Dugal, C. M. Grimaldi, et
862 al. 2022. Spatially varying selection between habitats drives physiological shifts and local
863 adaptation in a broadcast spawning coral on a remote atoll in Western Australia. *Science*
864 *Advances* 8:eabl9185.
- 865 Thuiller, W., C. Albert, M. B. Araújo, P. M. Berry, M. Cabeza, A. Guisan, T. Hickler, et al. 2008.
866 Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives*
867 in Plant Ecology, Evolution and Systematics 9:137–152.
- 868 Tigano, A., T. Weir, H. G. M. Ward, M. K. Gale, C. M. Wong, E. J. Eliason, K. M. Miller, et al. 2024.
869 Genomic vulnerability of a freshwater salmonid under climate change. *Evolutionary*
870 *Applications* 17:e13602.
- 871 Wadgymar, S. M., D. B. Lowry, B. A. Gould, C. N. Byron, R. M. Mactavish, and J. T. Anderson.
872 2017. Identifying targets and agents of selection: innovative methods to evaluate the processes
873 that contribute to local adaptation. *Methods in Ecology and Evolution* 8:738–749.
- 874 Waldvogel, A.-M., B. Feldmeyer, G. Rolshausen, M. Exposito-Alonso, C. Rellstab, R. Kofler, T.
875 Mock, et al. 2020. Evolutionary genomics can improve prediction of species' responses to
876 climate change. *Evolution Letters* 4.
- 877



175

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



187 **Figure 7** Computation time (y-axes, first column) and memory requirements (y-
 188 axes, second column) of gradientForests (GF) model training differs between
 189 genotype- (GF_{geno}) and allele frequency-based (GF_{AF}) implementations (x-axes).
 190 Differences within implementations are driven primarily by the number of loci
 191 (second row) and the pattern of environmental variables (third row). Total runtime
 192 across all model training exceeds 417 days for GF_{geno} and 180 days for GF_{AF}. Data
 193 included in this figure is for all spatially discrete evaluations that completed without
 194 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.

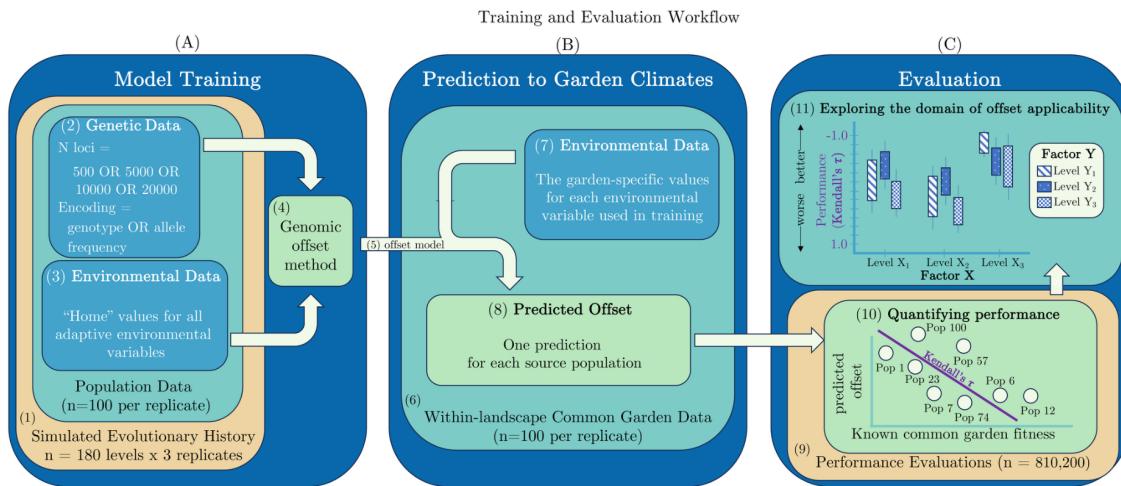


Figure S2 Analysis of spatially discrete simulations included three main phases: (a) model training, (b) model prediction and (c) evaluation of models. Subpanels of this schematic are numbered for referencing in Table 2 and the main text. Evaluation of spatially continuous simulations (not shown) used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

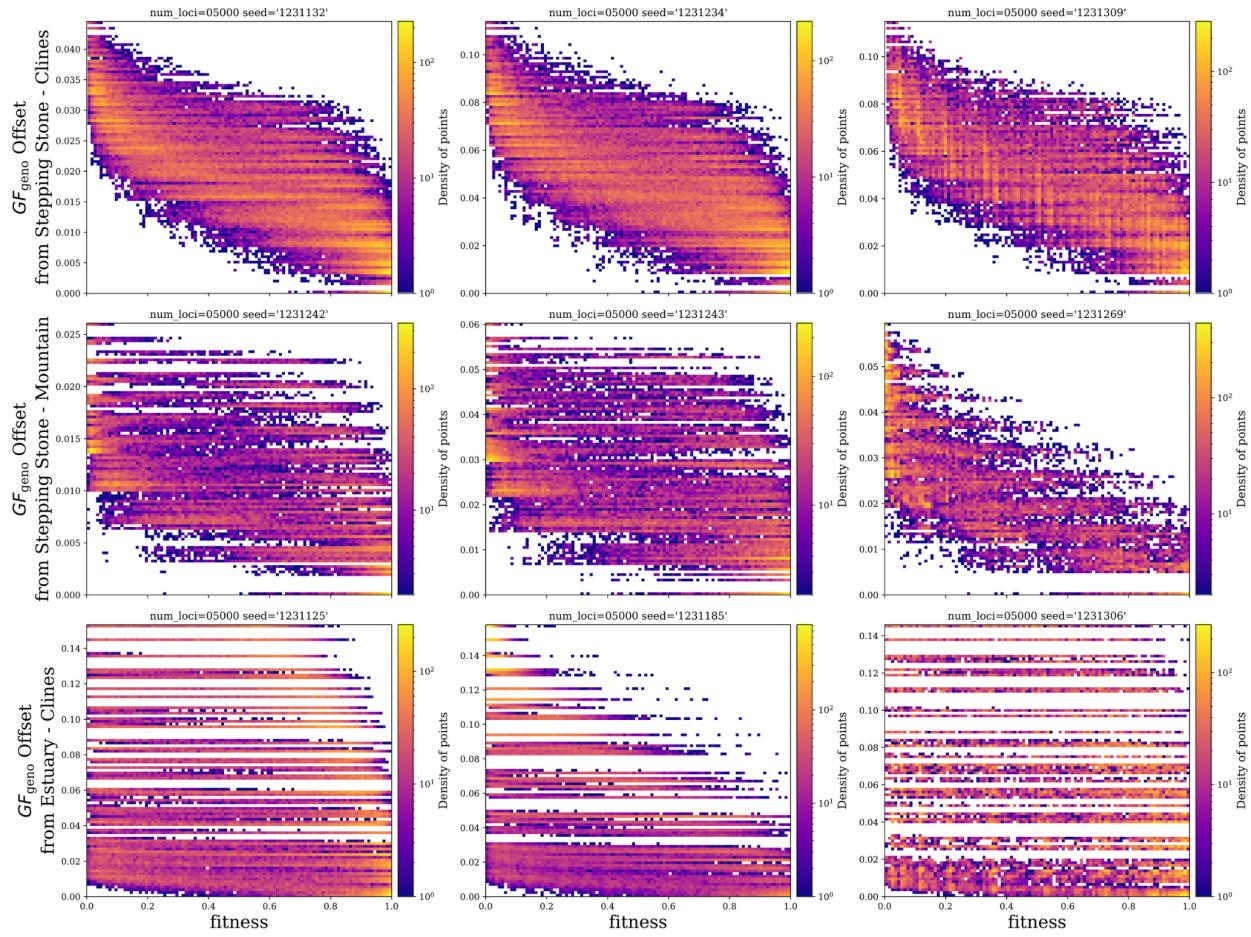


Figure S3 The relationship between fitness and predicted offset from GF_{geno} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored with respect to the 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.

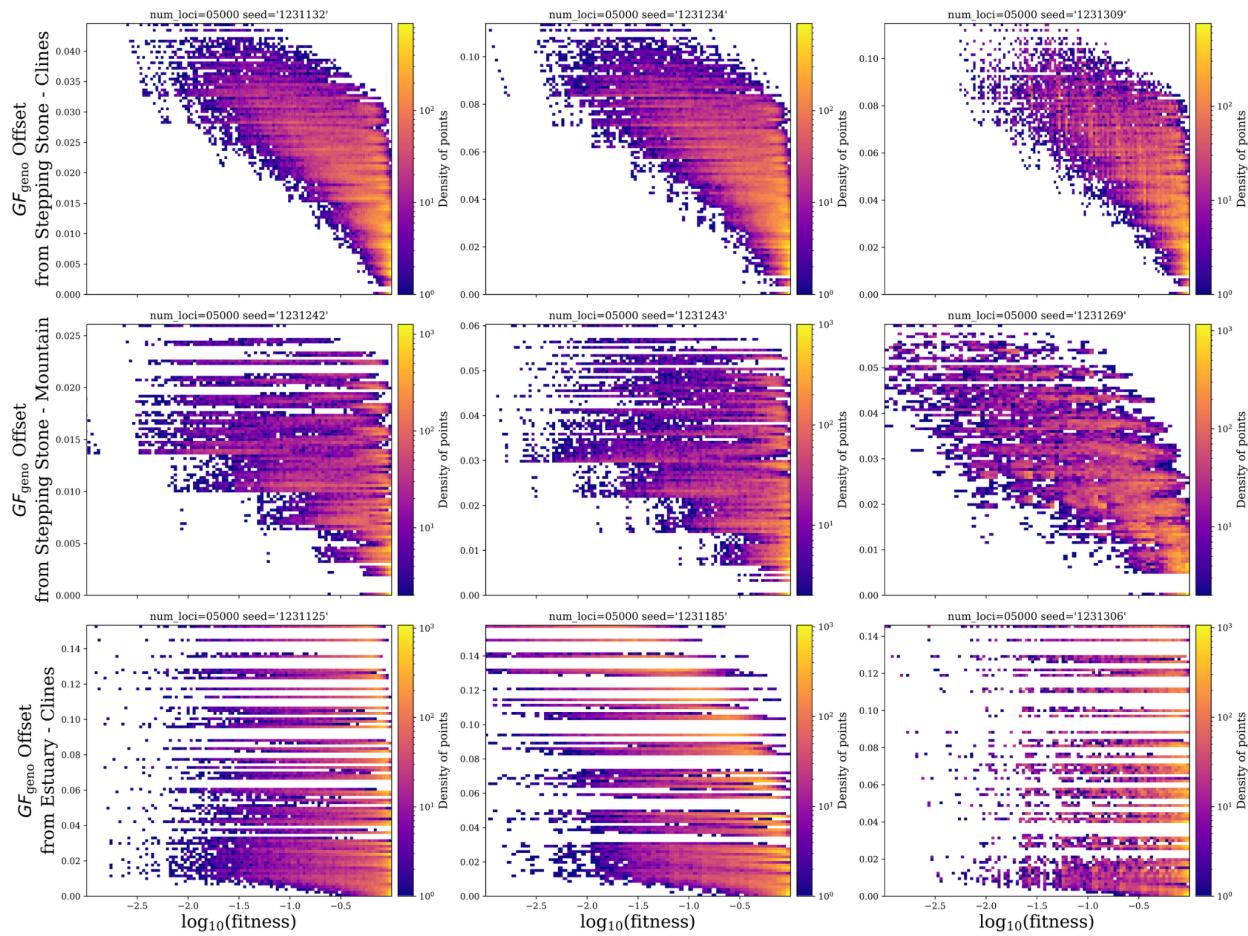


Figure S4 The relationship between $\log_{10}(\text{fitness})$ and predicted offset from GF_{geno} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored 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.

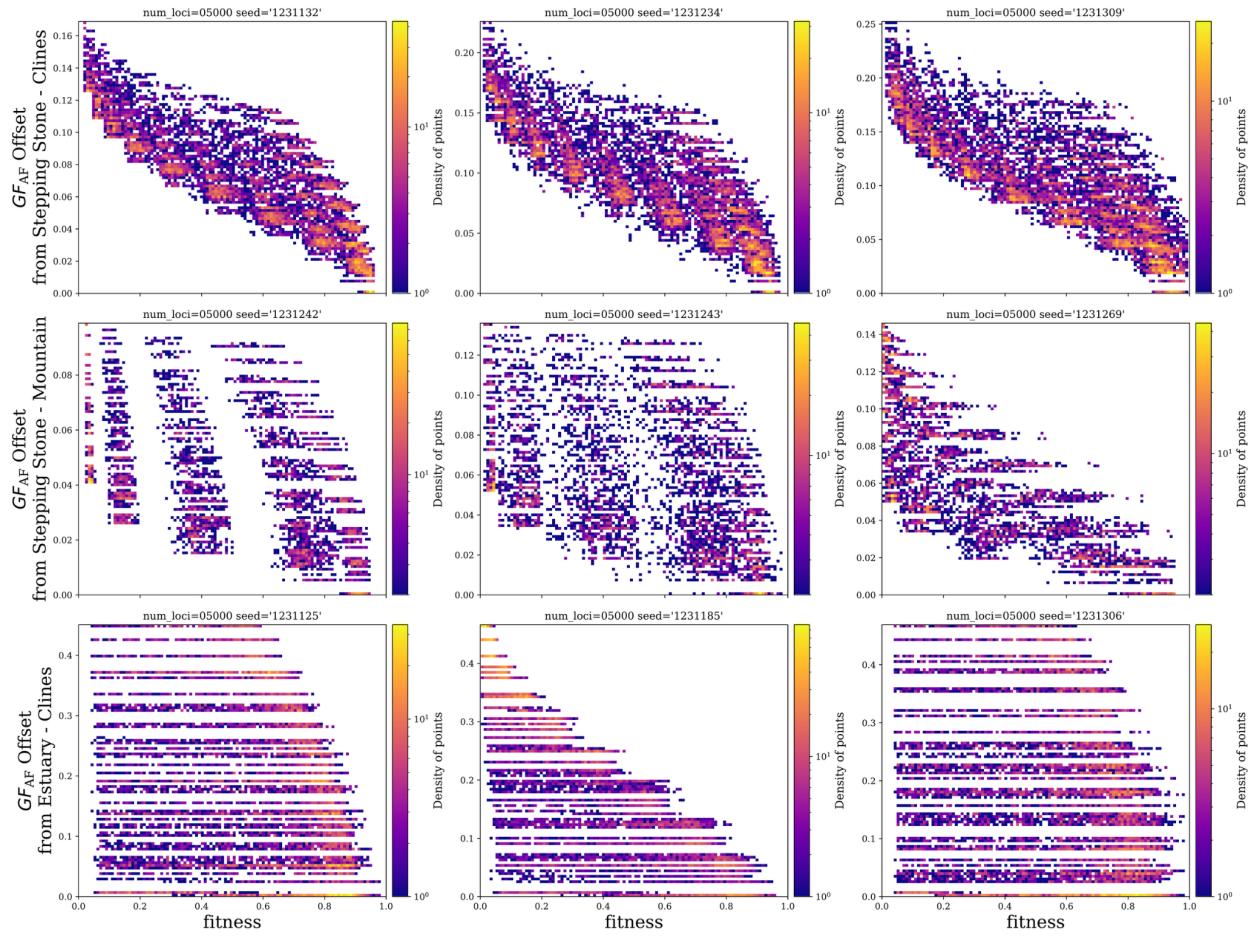


Figure S5 The relationship between fitness and predicted offset from GF_{AF} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored with respect to the density 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.

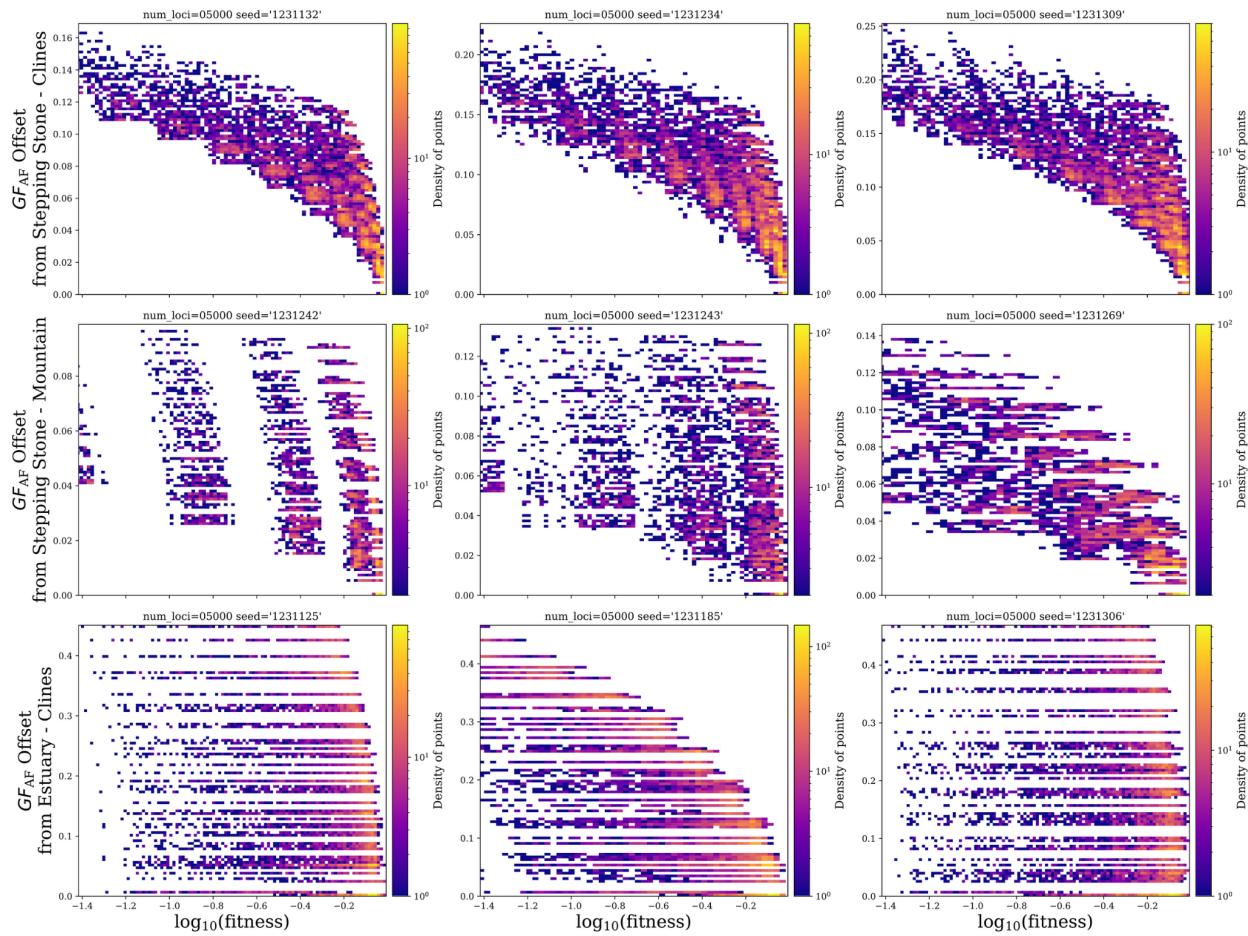


Figure S6 The relationship between fitness and predicted offset from GF_{AF} models is non-linear. From each landscape, the relationship between offset and fitness is plotted for three random levels (seeds - these are the same seeds used across Figs S12-S16). Figures are colored with respect to the density 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.

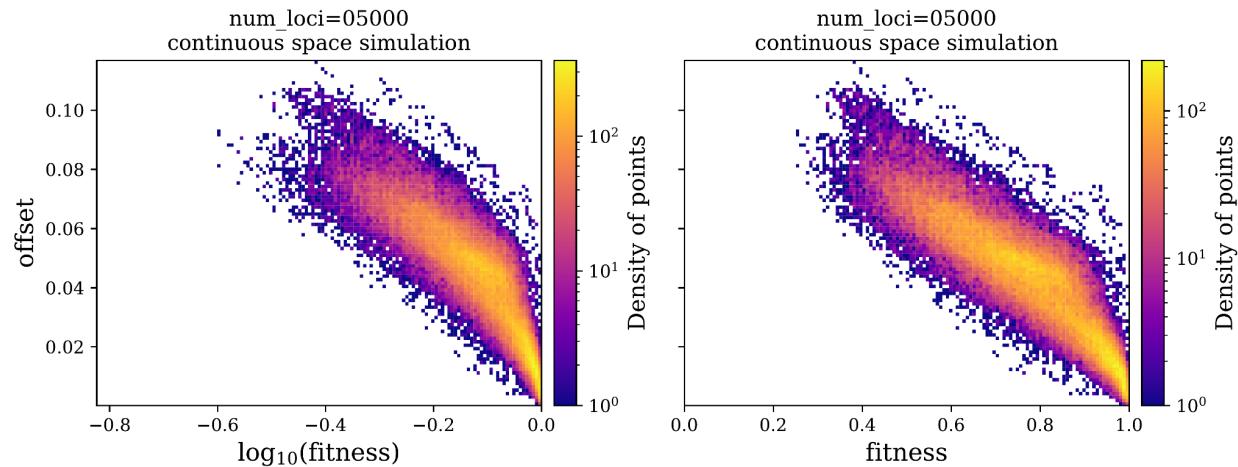


Figure S7 The relationship between fitness or $\log_{10}(\text{fitness})$ and predicted offset from the continuous space GF model is non-linear. Figures are colored with respect to the density of points. Code to create this figure can be found in SC 05.09. Data included in this figure is from the spatially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

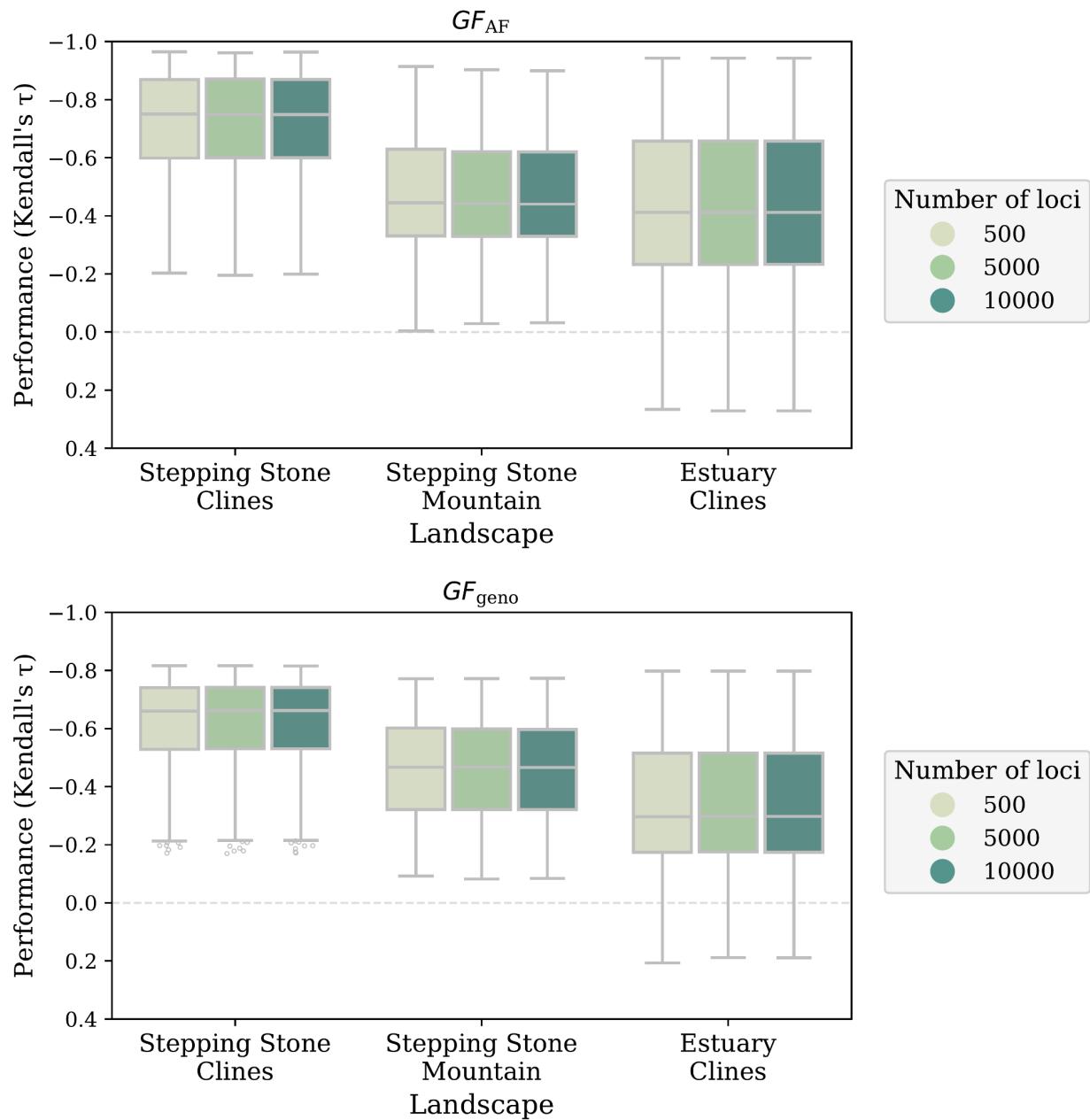


Figure S8 The performance of offset predictions from gradientForests models trained with allele frequencies (GF_{AF}) or genotypes (GF_{geno}) was not differentially affected by the number of loci provided for training. Code to create these figures can be found in Supplemental Code 04.01.

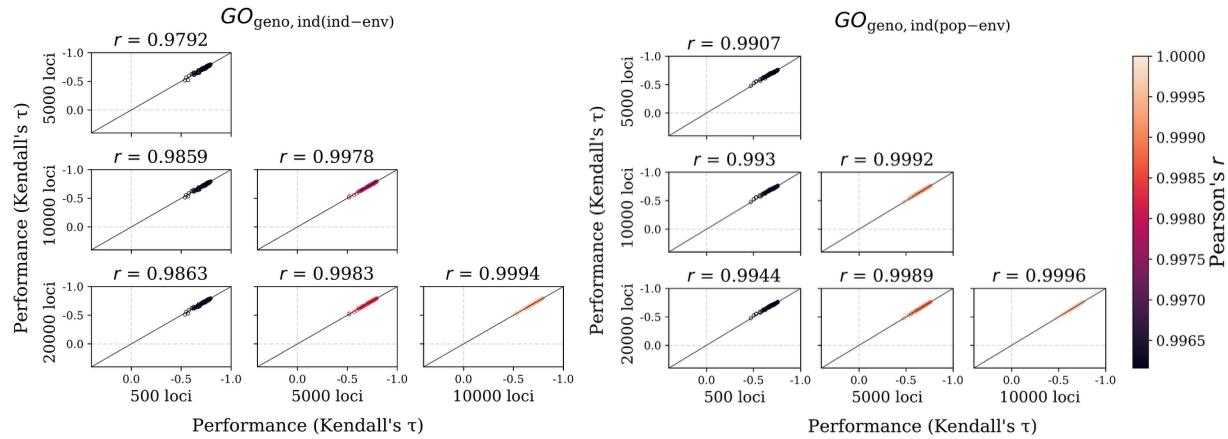


Figure S9 The number of loci used for training has little impact on performance within the spatially continuous workflows. Comparison of predictive performance when environmental data is input at the individual level ($GO_{\text{geno,ind(ind-env)}}$) or at the population level ($GO_{\text{geno,ind(pop-env)}}$) 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.

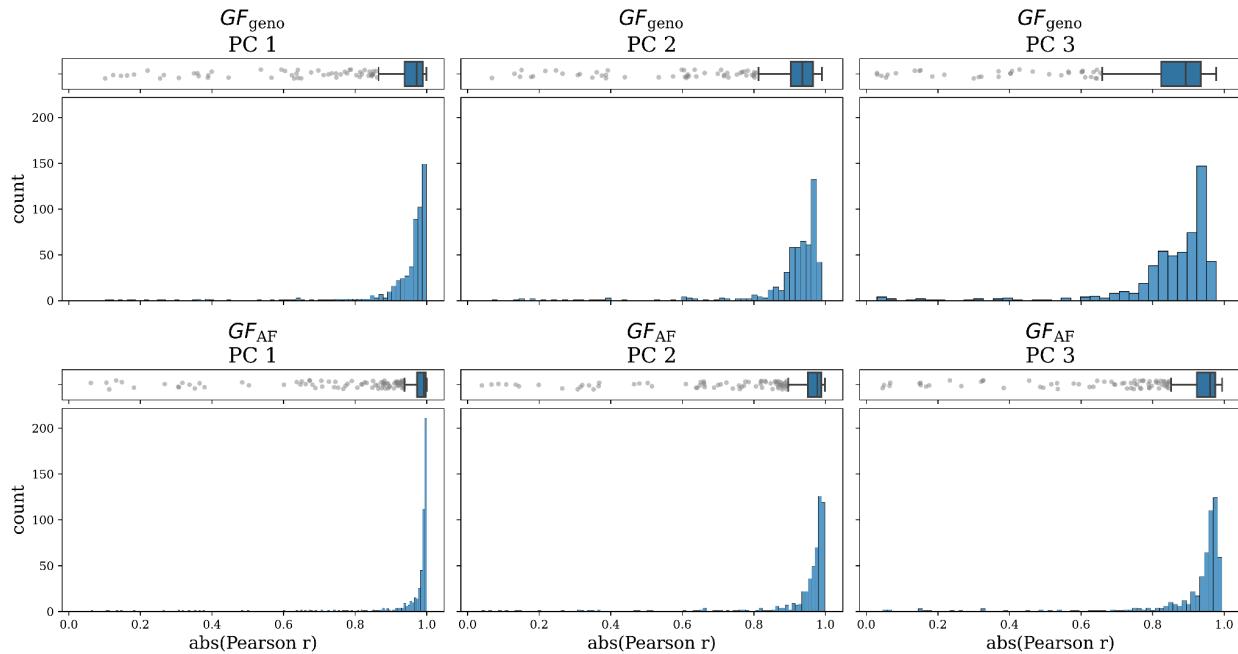


Figure S10 Population structure is captured by small marker sets from both allele frequency- (GF_{AF}) and population-level (GF_{geno}) models of gradientForests (GF). Principal component (PC) analysis was carried out for each marker set from each simulation replicate using either genotype or allele 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.

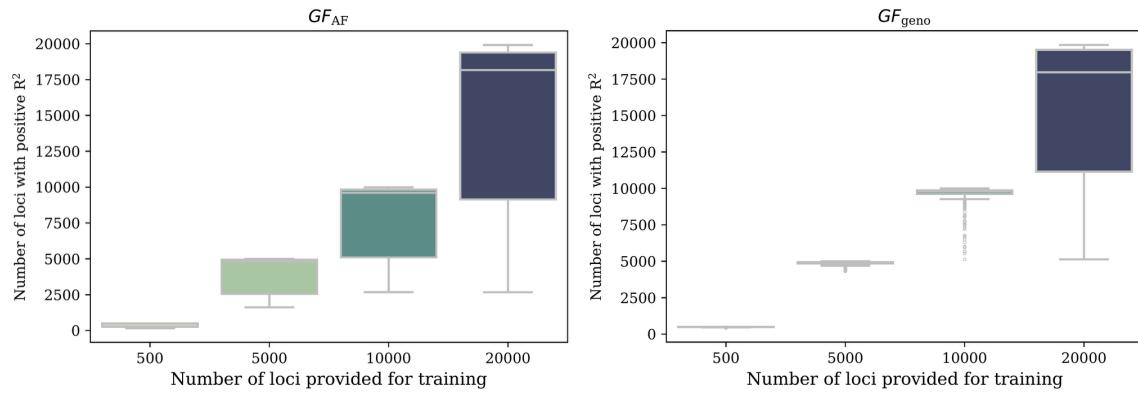


Figure S11 The number of loci that were incorporated into gradientForest (GF) models (y-axes; i.e., loci with $R^2 > 0$ from internal random forest models) were roughly representative of the number of loci provided for model training (x-axes). Data included in this figure is from all models that successfully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

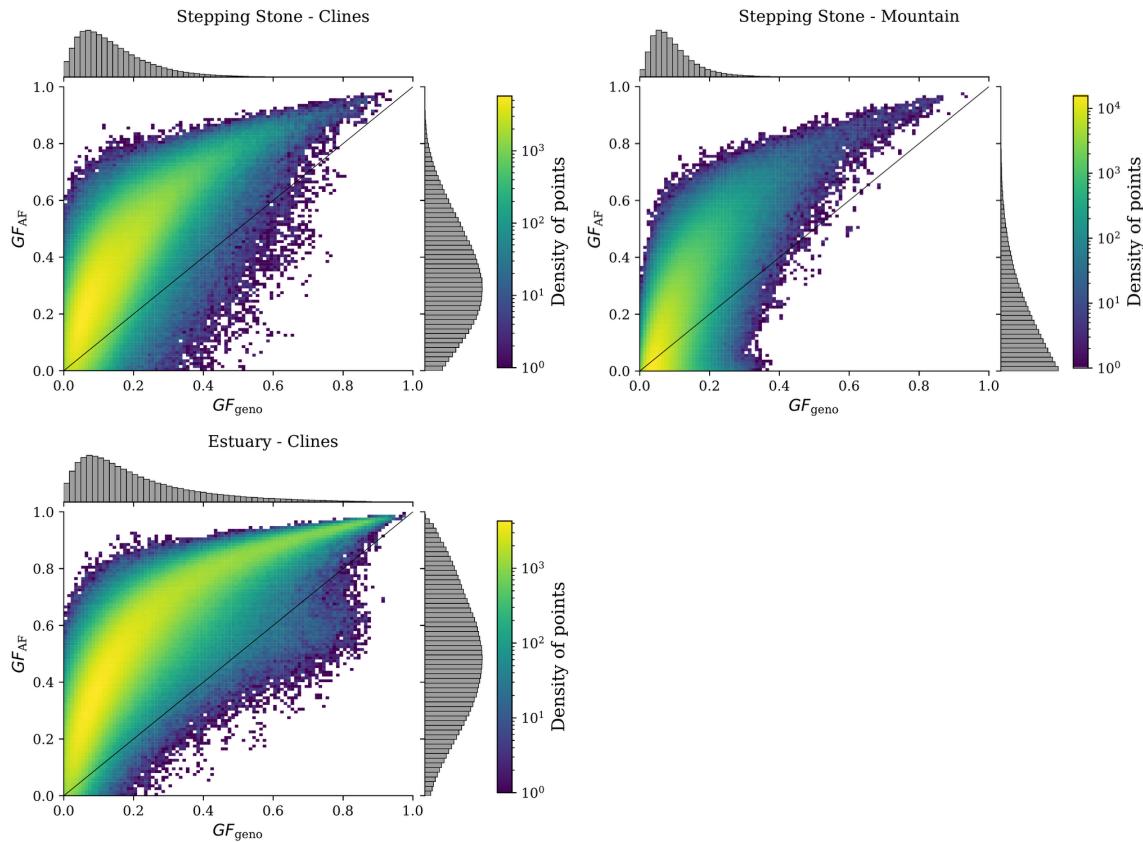


Figure S12 Predictive power of loci (R^2 , all axes) from random forest models used by gradientForests (GF) are positively correlated between genotype- (GF_{geno}) and allele frequency-based (GF_{AF}) models. Data included in this figure are all overlapping loci incorporated into GF_{geno} or GF_{AF} 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.

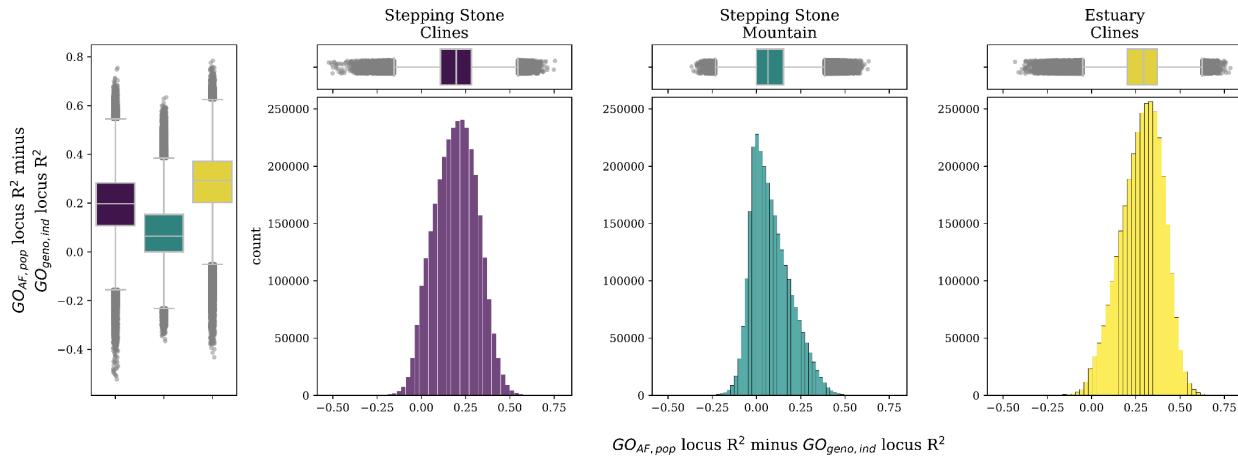


Figure S13 Predictive power of loci (R^2) from random forest models used by gradientForests (GF) are generally greater when encoded as allele frequencies (i.e., GF_{AF} models) than genotypes (i.e., GF_{geno} models). Data used in this figure is R^2 from loci overlapping GF_{AF} and GF_{geno} models.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

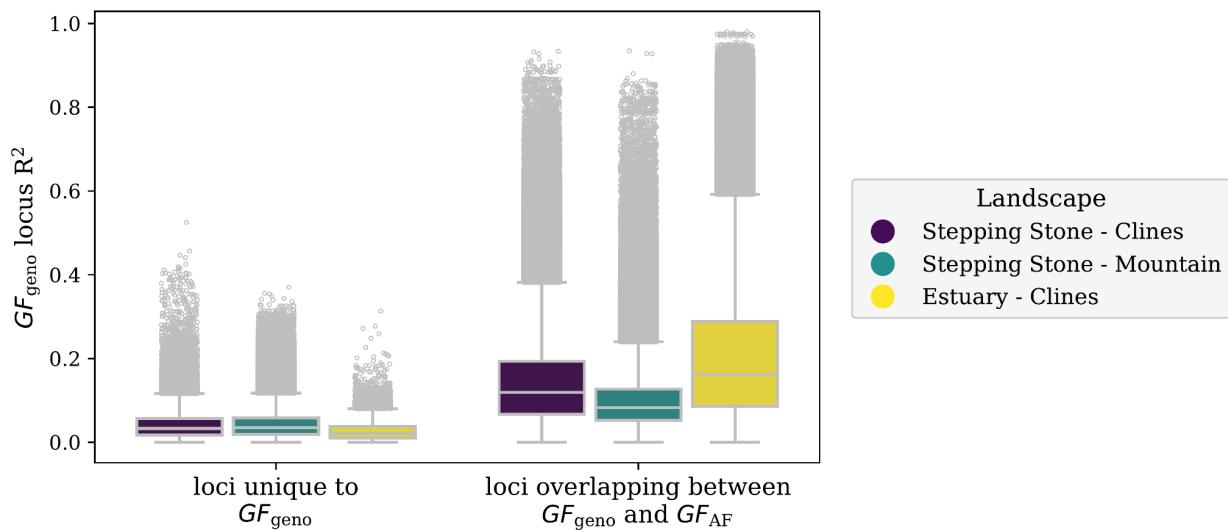


Figure S14 Predictive power (R^2 , y-axis) of random forest models (one per locus) from loci incorporated into gradientForest (GF) models. Data included in this figure are R^2 values from GF_{geno} models. Code to create this figure can be found in Supplemental Code 05.05.

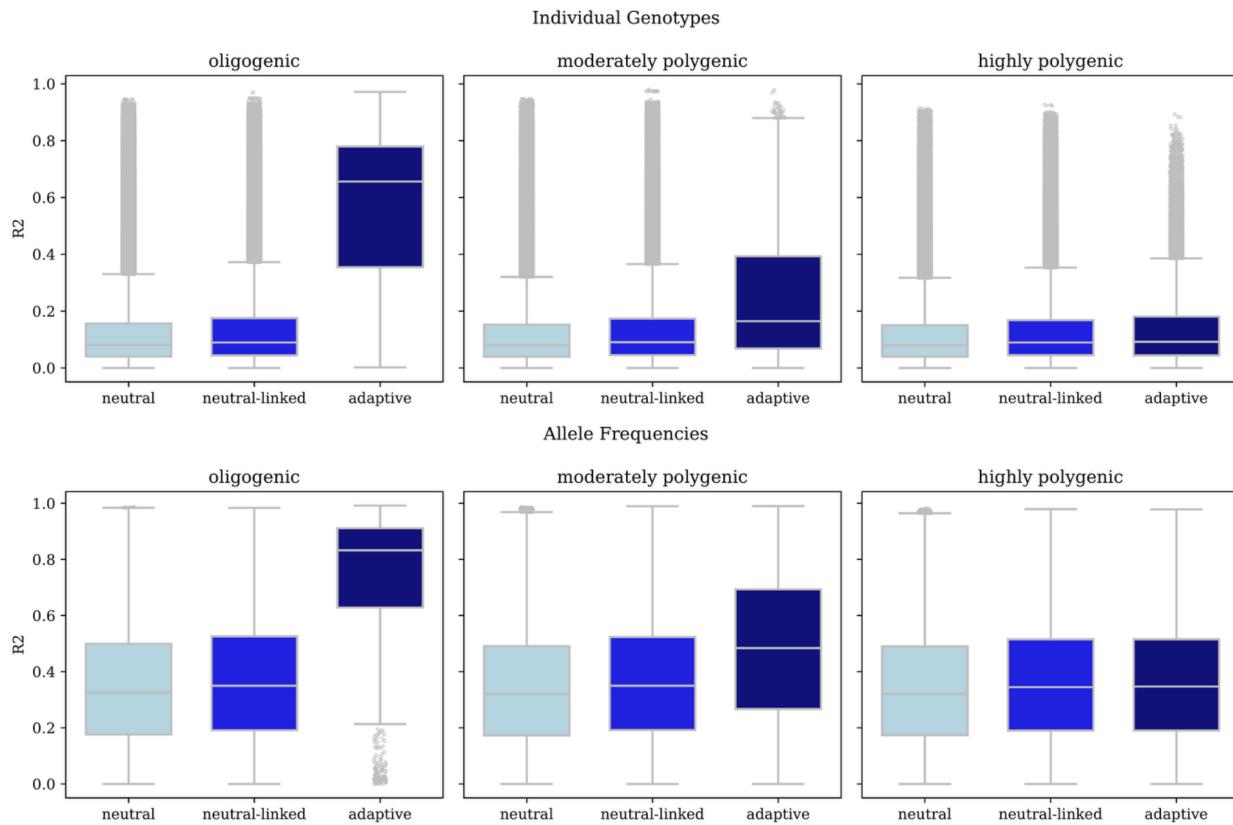


Figure S15 Differences in the predictive power of between adaptive and neutral loci (R^2) from random forest models depend on the genetic architecture underlying local adaptation. Data included in this figure are all overlapping loci incorporated into GF_{geno} and GF_{AF} models that completed training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

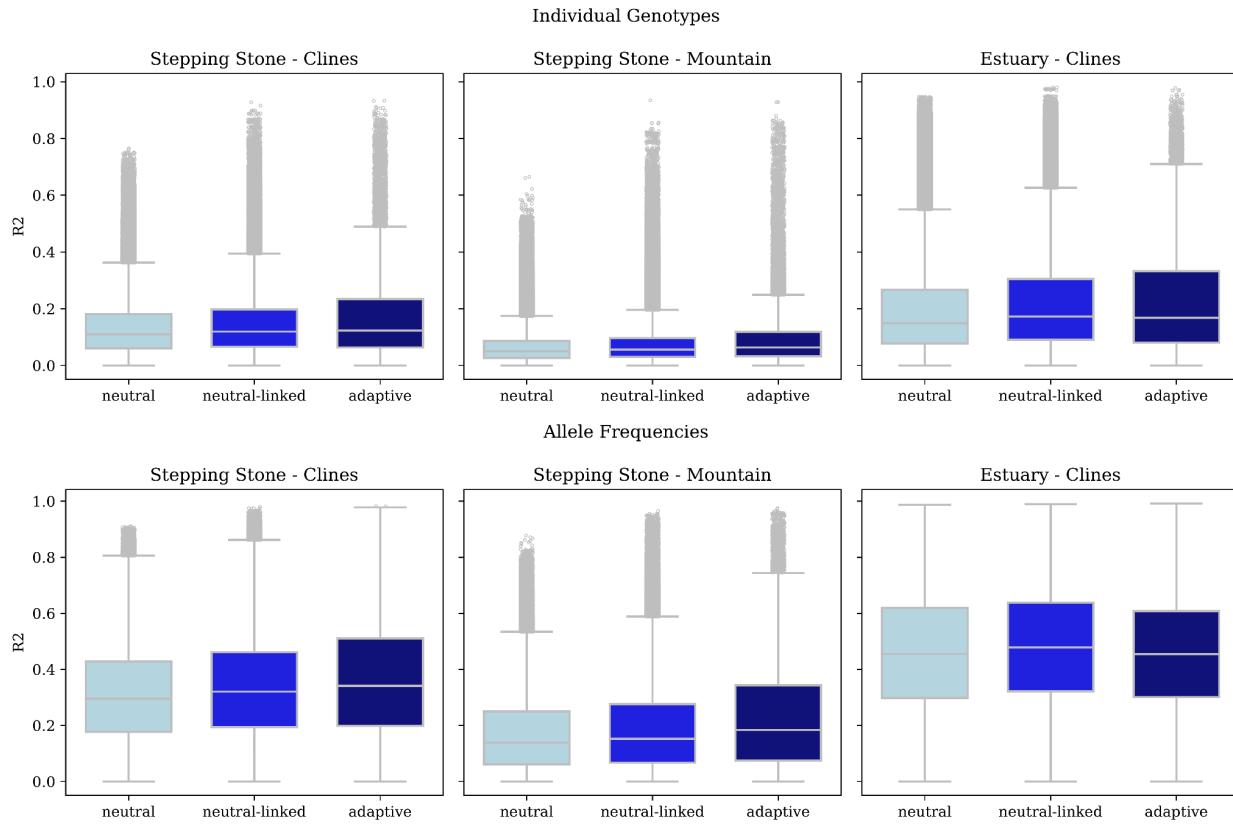


Figure S16 The predictive power of between adaptive and neutral loci (R^2) from random forest models are largely undifferentiated within spatially discrete landscapes (titles). Shown are R^2 values for loci incorporated into genotype-based models (top) and allele frequency-based models (bottom). Data 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.

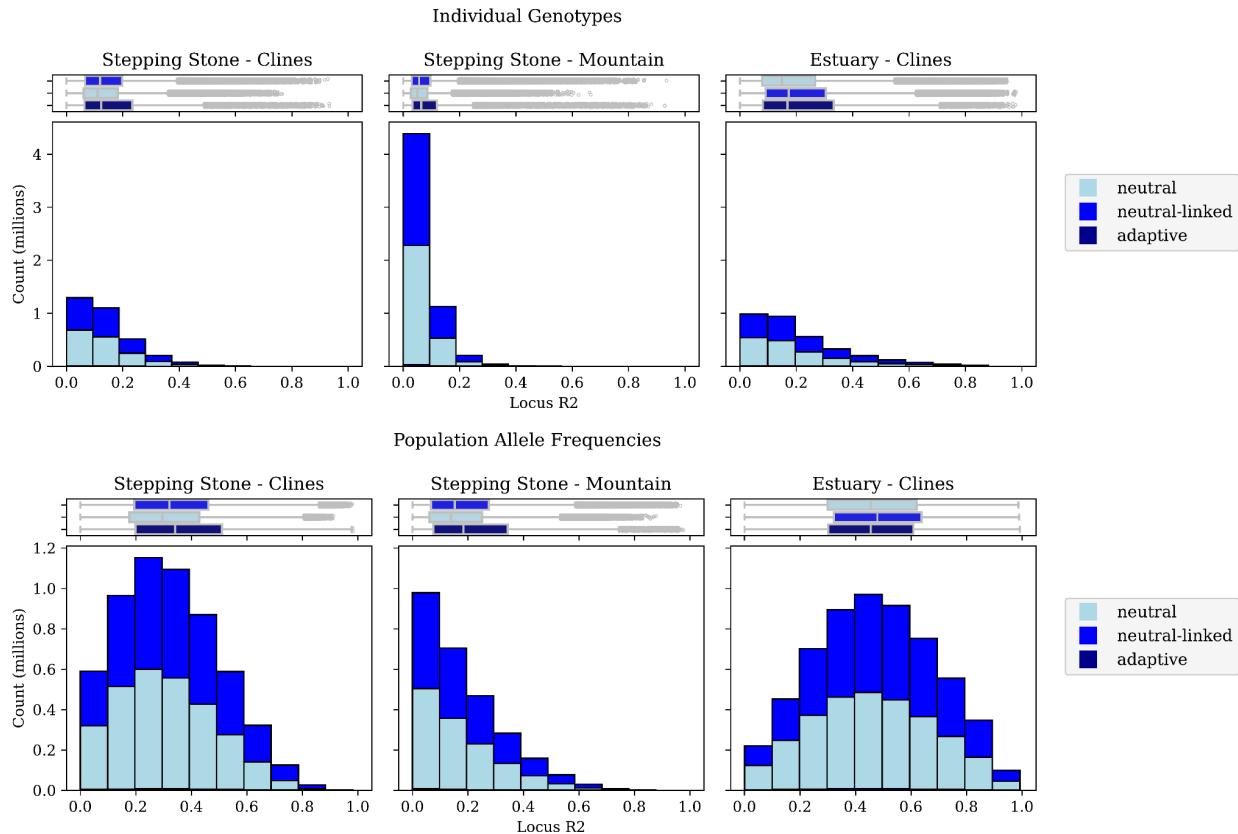


Figure S17 Despite little differences in R^2 between adaptive and neutral loci, distributions of R^2 differed among landscapes. Data included in this figure are all overlapping loci incorporated into GF_{geno} and GF_{AF} models that completed training without failure. Code to create these figures can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

3 | References

- Ellis, N., S. J. Smith, and 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.