The WZA: A window-based method for characterizing genotype-environment association

Tom R. Booker[1,2,3,*], Sam Yeaman[1], James R. Whiting[1], Michael C. Whitlock[2,3]

1. Department of Biological Sciences, University of Calgary, Calgary, Canada
2. Department of Zoology, University of British Columbia, Vancouver, Canada
3. Biodiversity Research Centre, University of British Columbia, Vancouver, Canada

*Corresponding author: booker@zoology.ubc.ca

Abstract

Genotype environment association (GEA) studies have the potential to identify the genetic basis of local adaptation in natural populations. Specifically, GEA approaches look for a correlation between allele frequencies and putatively selective features of the environment. Genetic markers with extreme evidence of correlation with the environment are presumed to be tagging the location of alleles that contribute to local adaptation. In this study, we propose a new method for GEA studies called the weighted-Z analysis (WZA) that combines information from closely linked sites into analysis windows in a way that was inspired by methods for calculating F_{ST}. We analyze simulations modelling local adaptation to heterogeneous environments to compare the WZA with existing methods. In the majority of cases we tested, the WZA either outperformed single-SNP based approaches or performed similarly. In particular, the WZA outperformed individual SNP approaches when a small number of individuals or demes was sampled. We apply the WZA to previously published data from lodgepole pine and identified candidate loci that were not found in the original study.

KEYWORDS: Local adaptation, population genetics, landscape genomics, GEA

Introduction

Studying local adaptation can provide a window into the process of evolution, yielding insights about the nature of evolvability, constraints to diversification, and how the interplay between a species and its environment shapes intraspecific genetic variation (e.g. Savolainen 2013). Understanding local adaptation can also benefit practical applications such as in forestry where many species of economic interest exhibit pronounced trade-offs in productivity across environments. Characterizing such trade-offs may help identify alleles involved in local adaptation, revealing candidate genes important for breeding or informing conservation management programs for buffering against the consequences of anthropogenic climate change (Aitken and Whitlock 2013). Whatever the aim or application, a first step in studying the basis of local adaptation is to identify the genes that are driving it.

A potentially powerful method for identifying the genomic regions involved in local adaptation is genotype-environment association (GEA) analysis, which has been widely adopted in recent years. Alleles may vary in frequency across a species' range in response to local environmental conditions that give rise to spatially varying selection pressures (Haldane 1948). For that reason, genetic variants that exhibit strong correlations with putatively selective features of the environment are often interpreted as a signature of local adaptation (Coop et al. 2010). Genotype-environment association (GEA) studies examine such correlations. Allele frequencies for many genetic markers, typically single nucleotide polymorphisms (hereafter SNPs), are estimated in numerous locations across a species' range. Correlations between allele frequency and environmental variables are calculated then contrasted for sites across the genome. It is assumed in GEA studies that current heterogeneity in the environment (whether biotic or abiotic) reflects the history of selection and that the local populations contain genetic variation that maximise fitness in those environments.

The most straightforward way to perform a GEA analysis is to simply examine the correlation between allele frequencies and environmental variables measured in multiple populations, for example using rank correlations such as Spearman's ρ or Kendall's τ. This simple approach may commonly lead to false positives, however, if there is environmental variation across the focal species' range that is correlated with patterns of gene flow or historical selection (Meirmans 2012; Novembre and Di Rienzo 2009). For example, consider a hypothetical species inhabiting a large latitudinal range. If this species had restricted migration and exhibited isolation-by-distance, neutral alleles may be correlated with any environmental variable that happened to correlate with latitude, as population structure would also correlate with latitude.

Several approaches have been proposed to identify genotype-environment correlations above and beyond what is expected given an underlying pattern of population structure and environmental variation. For example, the commonly used *BayPass* package (Gautier 2015), an extension of *BayEnv* by Coop et al. (2010), estimates correlations between alleles and environmental variables in a two-step process. First, a population covariance matrix ($\mathbf{\Omega}$) is estimated from SNP data. Second, correlations between the

frequencies of individual SNPs and environmental variables are estimated treating $\boldsymbol{\Omega}$ in a manner similar to a random effect in a generalized mixed model. In a recent study, Lotterhos (2019) compared several of the most commonly used packages for performing GEA; including *BayPass* (Gautier 2015), latent-factor mixed models as implemented in the LEA package (*LFMM-LEA*; Frichot et al. 2013; Frichot and François 2015), redundancy analysis (RDA; see Forester et al. 2016, 2018) and a comparatively simple analysis calculating Spearman's ρ between allele frequency and environment. Of the methods they tested, Lotterhos (2019) found that the GEA approaches that did not correct for population structure (i.e., Spearman's ρ) had higher power to detect local adaptation compared to *BayPass* or *LFMM-LEA*

Individual SNPs may provide very noisy estimates of summary statistics, but closely linked SNPs are not independently inherited and may have highly correlated evolutionary histories. As a way to reduce noise, genome scan studies often aggregate data across adjacent markers into analysis windows based on a fixed physical or genetic distance or number of SNPs (Hoban et al. 2016). In the case of F_{ST}, the standard measure of population differentiation, there are numerous methods for combining estimates across sites (see Bhatia et al. (2013)). In Weir and Cockerham's (1984) method, for example, estimates of F_{ST} for individual loci are combined into a single value with each marker's contribution weighted by its expected heterozygosity.

In the context of GEA studies, each marker or SNP provides a test of whether a particular genealogy is correlated with the pattern of environmental variation. In the extreme case of a non-recombining region, all SNPs would share the same genealogy and thus provide multiple tests of the same hypothesis. For recombining portions of the genome, however, linked sites will not have the same genealogy, but genealogies may be highly correlated. Similar to combining estimates of F_{ST} to decrease statistical noise, combining GEA tests performed on individual markers may decrease noise and increase the power of GEA studies to identify genomic regions that contribute to local adaptation. In addition, there are several practical benefits of a window-based approach over a SNP-based approach. The number of analysis windows will be substantially less than the number of SNPs in a genome-wide analysis, so there will be fewer multiple comparisons to correct for — corrections which can severely reduce power (Benjamini and Hochberg 1995). Additionally, wide variation in SNP number across the genome may lead to varying false positive rates across genes. Finally, window-based metrics are more readily compared across species.

In this study, we propose a general method for combining the results of single SNP GEA scores into analysis windows that we call the weighted-Z analysis (WZA), and we test its efficacy using simulations. The WZA is capable of using many different GEA summary statistics as input. We generate datasets modelling a sequencing project where estimates of allele frequency are obtained for numerous populations across a species' range. Using our simulated data, we compare the performance of the WZA to Kendall's τ as well as other widely used GEA methods. Additionally, we compare the WZA to another window-based GEA approach proposed by Yeaman et al. (2016). We found that the WZA is particularly useful when GEA analysis is performed on small samples and when results for individual SNPs are statistically noisy. We re-analyze

previously published lodgepole pine (*Pinus contorta*) data using the WZA and find several candidate loci that were not identified using the methods of the original study.

The Weighted-*Z* Analysis

In this study, we propose the Weighted-*Z* Analysis (hereafter, the WZA) for combining information across linked sites in the context of GEA studies. Specifically, we aim to combine information from multiple SNPs within the same small genomic region to ask whether that region shows associations between local allele frequencies and local environment.

The WZA uses the weighted-*Z* test from the meta-analysis literature that combines *p*-values from multiple independent hypothesis tests into a single score (Mosteller and Bush 1954; Liptak 1958; Stouffer et al. 1949). In the weighted-*Z* test, each of the n independent tests is given a weight that is proportional to the inverse of its error variance (Whitlock 2005). We use the expected heterozygosity of each SNP in a gene or window for the weights in the WZA, following Weir and Cockerham (1984), as their classic method performs well in a similar evolutionary context, where the aim is to quantify divergence in allele frequencies among populations. At a given polymorphic site, we denote the average frequency of the minor allele across populations as $\overline{p}$ ($\overline{q}$ corresponds to the frequency of the major allele). Sites with higher values of $\overline{p}\overline{q}$ will carry more information about the underlying genealogy.

We combine information about genetic correlations with the environment from biallelic markers (typically SNPs) present in a focal genomic region into a single weighted-*Z* score (Z_W). The genomic region in question could be a gene or genomic analysis window. For each SNP with a minor allele frequency greater than 0.05 in the genomic window, we measure the association between the SNP's local allele frequency and the local environment in some way (for example rank correlation between allele frequency and environmental variation) and use the *p*-value of a test of no association for each SNP. (The exact measure of evidence for association used here may vary; in this paper we test the use of several such measures, described below.)

These *p*-values from each SNP in a window are combined using the weighted version of Stouffer's weighted *Z* approach (Whitlock 2005). We calculate $Z_{W,k}$ for genomic region *k*, which contains *n* SNPs, as

$$Z_{W,k} = \frac{\sum_{i=1}^{n} \bar{p}_i \, \bar{q}_i z_i}{\sqrt{\sum_{i=1}^{n} (\bar{p}_i \, \bar{q}_i)^2}}, \quad (1)$$

where $\bar{p}_i$ is the mean allele frequency across populations and z_i is the standard normal deviate calculated from the one-sided *p*-value for SNP *i*. A given *p*-value can be converted into a z_i score by finding the corresponding quantile of the standard normal distribution, for example using the *qnorm* function in R.

Under the null hypothesis that there is no correlation between allele frequency and environment and no spatial population structure, the expected distribution of correlation coefficients in a GEA would be normal with mean 0, with a uniform distribution of *p*-values. However, as will often be the case in nature, there may be an underlying correlation between population structure and environmental variation that will cause these genome-wide distributions to deviate from this null expectation. The average effect of population structure on individual SNP scores can be incorporated into an analysis by converting an individual SNP's squared correlation coefficient or parametric *p*-value into empirical *p*-values based on the genome-wide distribution (following the approach of Hancock et al. [2011]). Empirical *p*-values are simply the rank-transformed data, so to calculate them, we rank all values (from smallest to largest in the case of *p*-values) and divide the ranks by the total number of tests performed (i.e. the number of SNPs or markers in the analysis window). Note that in practice, we calculated empirical *p*-values after removing SNPs with minor allele frequency less than 0.05 and would recommend that others perform similar filtering. In empirical studies with varying levels of missing data across the genome, it may be preferable to rank the parametric *p*-values rather than the correlation coefficients themselves as there may be varying power to calculate correlations across the genome. With the empirical *p*-value procedure, aggregating information using the WZA will identify genomic regions with a pattern of GEA statistics that deviate from the average genome-wide. A feature of the WZA is that many tests can potentially be used as input as long as individual *p*-values provide a measure for the strength of evidence against a null hypothesis.

Wide variation in the density of SNPs across the genome may inf4uence the performance of the WZA (see Results). We account for variation in SNP number in the WZA as follows: We order all WZA scores by the number of SNPs in each window. Then, for a sliding bin of 50 analysis windows (with a step of 1 window) we calculate the mean and standard deviation of WZA scores. We then fit separate 1-dimensional polynomials to both the means and standard deviations of these sliding bin data to obtain a predictive model of the mean and standard deviation of WZA scores for an arbitrary number of SNPs. We use the "*poly1d*" function from *Numpy* to fit these models. Then, for each analysis window we calculate its *p*-value based on its predicted mean and standard deviation under the assumption of normality. We use the $-\log_{10}$(*p*-values) of WZA scores as our summary statistic.

Materials and Methods

In the previous section we described the mechanics of our new method, the WZA. The rest of this paper is devoted to a test of the relative efficacy of the WZA compared to widely used GEA approaches. Note that Lotterhos (2019) identified a simple rank correlation on individual SNPs as having among the highest power of the GEA analyses that they tested, making such a method a good standard of comparison. In addition, we also compare the WZA to commonly used GEA methods.

To do these tests, we simulate populations evolving on a variety of different environmental landscapes, with the selective optima varying over space. We simulate cases of relatively weak selection and strong selection.

Simulating local adaptation

We performed forward-in-time population genetic simulations of local adaptation to determine how well the WZA was able to identify the genetic basis of local adaptation. GEA studies are often performed on large spatially extended populations that may be comprised of hundreds of thousands of individuals. However, it is computationally infeasible to model selection and linkage in long chromosomal segments (>1Mbp) for such large populations. For that reason, we simulated relatively small populations containing 19,600 diploid individuals in total and scaled population genetic parameters to model a large population. We based our choice of population genetic parameters on estimates for conifer species. Note, while our simulations were motivated by conifers, we were not aiming to model a particular species. A representative set of parameters is given in Table S1 and in the Appendix we give a breakdown and justification of the parameters we chose. All simulations were performed in *SLiM* v3.7 (Haller and Messer 2019).

We simulated meta-populations inhabiting and adapting to heterogeneous environments and modelled the population structure on an idealized conifer species. In conifers, strong isolation-by-distance has been reported and overall mean $F_{ST} < 0.10$ has been estimated in several species (Mimura and Aitken 2007; Mosca et al. 2014). We thus simulated individuals inhabiting a 2-dimensional stepping-stone population made up of 196 demes (i.e. a 14×14 grid). Each deme consisted of N_d = 100 diploid individuals. We assumed a Wright-Fisher model so demes did not fluctuate in size over time. Migration was limited to neighboring demes in the cardinal directions and the reciprocal migration rate between demes (m) was set to 0.0375 in each possible direction to achieve an overall F_{ST} for the metapopulation of around 0.04 (Figure S1). As expected under restricted migration, our simulations exhibited a strong pattern of isolation-by-distance (Figure S1). Additionally, we simulated metapopulations with no spatial structure (i.e., finite island models). In these simulations, we used the formula

$$m = \frac{\frac{1}{F_{ST}} - 1}{4N_d 196}$$

(Charlesworth and Charlesworth 2010; pp319) to determine that a migration rate between each pair of demes of m = 4.12 x 10^{-4} would give a target F_{ST} of 0.03.

The simulated organism had a genome containing 1,000 genes evenly distributed on 5 chromosomes. We simulated a chromosome structure in *SLiM* by including nucleotides that recombined at r = 0.5 at the hypothetical chromosome boundaries. Each chromosome contained 200 segments of 10,000bp each. We refer to these segments as genes for brevity, although we did not model an explicit exon/intron or codon structure. It has been reported that linkage disequilibrium (LD) decays rapidly in

conifers, with LD between pairs of SNPs decaying to background levels within 1,000bp or so in several species (Pavy et al. 2012). In our simulations, recombination within genes was uniform and occurred at a rate of $r = 10^{-7}$ per base-pair, giving a population-scaled recombination rate ($4N_d r$) of 0.0004. The recombination rate between the genes was set to 0.005, effectively modelling a stretch of 50,000bp of intergenic sequence. Given these recombination rates, LD decayed rapidly in our simulations with SNPs that were approximately 600bp apart having, on average, half the LD of immediately adjacent SNPs in neutral simulations (Figure S1). Thus, patterns of LD decay in our simulations were broadly similar to the patterns reported for conifers.

We incorporated spatial variation in the environment into our simulations using a discretized map of degree days below 0 (DD0) across British Columbia (BC). We generated the discretized DD0 map by first downloading the map of DD0 for BC from ClimateBC (*http://climatebc.ca/*; Wang et al. 2016; Figure 1A). Using Dog Mountain, BC as the reference point in the South-West corner (Latitude = 49.37, Longitude = -122.97), we extracted data in a rectangular grid with edges 3.6 degrees long in terms of both latitude and longitude, an area of approximately $266 \times 400 km^2$ (Figure 1A). We divided this map into a 14×14 grid, calculated the mean DD0 scores in each grid cell, converted them into standard normal deviates (i.e. *Z*-scores) and rounded up to the nearest third. We used the number of thirds of a *Z*-score as phenotypic optima in our simulations. We refer to this map of phenotypic optima as the *BC* map (Figure 1B).

We used data from the *BC* map to generate two additional maps of environmental variation. First, we ordered the data from the *BC* map along one axis of the 14×14 grid and randomized optima along the non-ordered axis. We refer to this re-ordered map as the *Gradient* map (Figure 1C). Second, we generated a map where selection differed over only a small portion of the environmental range. For some species, fitness optima may differ only beyond certain environmental thresholds (*e.g.* temperature above vs. below 0ºC), leading to a non-normal distribution of phenotypic optima. To model such a situation, we set the phenotypic optimum of 20 demes in the top-right corner of the meta-population to +3 and set the optimum for all other populations to –1. We chose 20 demes as it represented approximately 10% of the total population. We refer to this map as the *Truncated* map (Figure 1D).

We simulated local adaptation using a model of directional selection. There were 12 causal genes distributed evenly across four simulated chromosomes that potentially contributed to local adaptation. Mutations affecting fitness could only occur at a single nucleotide position in the center of the 12 potentially selected genes. Selected mutations had a spatially antagonistic effect on fitness. In deme *d* with phenotypic optimum θ_d, the fitness of an individual homozygous for the selected allele was $1 + s_a\theta_d$ (selected alleles were semi-dominant). The fitness affecting alleles had a mutation rate of 3×10^{-7} and a fixed s_a = 0.003 (hereafter weak selection) or s_a = 0.0136 (hereafter strong selection; see *Appendix* for a defense of these parameter choices).

We ran simulations for a total of 200,102 generations. The 19,600 individuals initially inhabited a panmictic population that evolved neutrally. After 100 generations, the

panmictic population divided into a 14×14 stepping-stone population and evolved strictly neutrally. After 180,000 generations, we imposed the various maps of phenotypic optima and simulated for a further 20,000 generations. For selected mutations, we used the "*f*" option for *SLiM*'s mutation stack policy, so only the first mutational change was retained. Using the tree-sequence option in *SLiM* (Haller et al. 2019), we tracked the coalescent history of each individual in the population. For each combination of map and mode of selection, we performed 30 replicate simulations.

We used the same simulated coalescent histories to model constant and varying mutation rates. At the end of each simulation, neutral mutations were added using *PySLiM* (*https://pyslim.readthedocs.io/en/latest/*). To model a constant mutation rate, mutations were added at a constant rate of 10^{-8}. To model variation in SNP density, we sampled mutation rates for individual genes uniformly between 1×10^{-9} and 7.3×10^{-8}. Simulations with a uniform mutation rates and varying mutation rates had similar mean numbers of SNPs per gene.

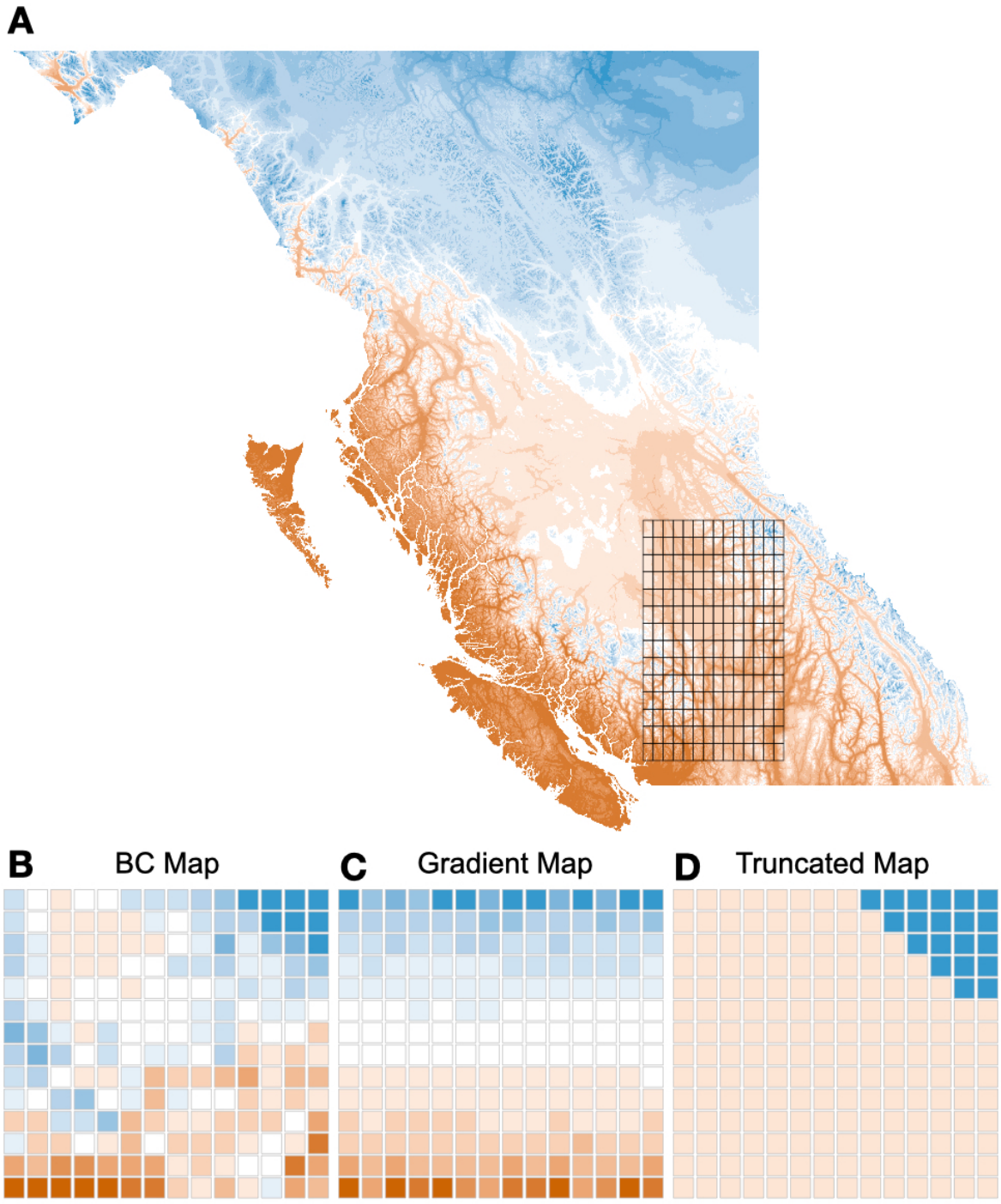


Figure 1 A) Degree days below zero across British Columbia, the overlain grid in A shows the locations we used to construct phenotypes for our simulated populations. B) A discretized map of DD0 in Southern British Columbia, we refer to the map in B as the *BC map*. C) A 1-dimensional gradient of phenotypic optima, we refer to this as the Gradient map. D) A model of selection acting on a small proportion of the population, we refer to this map as the Truncated map.

Classifying simulated genes as locally adapted

To evaluate the performance of different GEA methods, we needed to identify which of the 12 causal genes contributed to local adaptation and which did not in each simulation replicate. As described above, our simulations incorporated a stochastic mutation model so from replicate to replicate the genes that contributed to local adaptation varied.

We identified locally adapted genes from our simulations based on the mean fitness of their alleles at the single variable site in each gene with a polymorphism. Our measure of local adaptation was the covariance between the mean fitness contributed by the selected allele in each population and the environment.

We defined locally adapted genes as those with a covariance between environment and fitness greater than 0.005. When modelling weak directional selection, an average of 6.35, 6.50 and 5.80 genes (out of 12) contained genetic variants that established and contributed to local adaptation for the *BC map*, the *Gradient map* and the *Truncated map*, respectively. When modelling strong directional selection, an average of 12, 11 and 10 genes (out of 12) contained genetic variants that established and contributed to local adaptation for the *BC map*, the *Gradient map* and the *Truncated map*, respectively. Strong directional selection led to a tight distribution of effect sizes, while weak selection led to a wider spread of effect sizes (Figure S2).

Analysis of simulation data

We compared the performance of the WZA on our simulated data to several other GEA methods. We used Kendall's τ-b (hereafter Kendall's τ), a rank correlation that does not model population structure, *BayPass* (Gautier 2015), latent factor mixed models as implemented in the LEA package, redundancy analysis (RDA) (Forester et al. 2016) and the top-candidate method as described by Yeaman et al (2016). For all analyses, except where specified, we analyzed data for a set of 40 randomly selected demes and sampled 20 individuals from each to estimate allele frequencies. The demes from which individuals were sampled for each of the maps are shown in Figure S3. Each simulation replicate included 1,000 genes, and after excluding alleles with a minor allele frequency less than 0.05 there was an average of 23.3 SNPs per gene. We ran *BayPass* following the "worked example" in section 5.1.2 of the manual provided with the software. For RDA, we based our analysis on the tutorial given at (*https://popgen.nescent.org/2018-03-27_RDA_GEA.html*). For LFMM, we used the worked example in the manual distributed with the software assuming three latent factors (*http://membres-timc.imag.fr/Olivier.Francois/LEA/files/LEA_1.html*).

We compared performance of window-based GEA methods (the WZA and the top-candidate method) to single SNP-based methods as follows. For the window-based methods we simply used the scores obtained for individual genes. For single SNP-based methods, the SNP with the most extreme test statistic (e.g. the smallest *p*-value or largest Bayes factor) for each gene was recorded and other SNPs in the gene were subsequently ignored. This was done to prevent multiple outliers that are closely linked from being counted as separate hits. The single-SNP based method is perhaps most

similar to how GEA analyses are typically interpreted, as it relies upon the evidence from the most strongly associated SNP to assess significance for a closely linked gene.

We implemented a simplified version of the top-candidate method proposed by Yeaman et al. (2016), which aggregates GEA results in analysis windows. The top-candidate method attempts to identify regions of the genome involved in local adaptation under the assumption that such regions may contain multiple sites that exhibit strong correlation with environmental variables. The top-candidate is essentially a binomial test looking at whether a particular region has an excess of "outlier" SNPs based on the genome wide average. We defined outliers as those with the 99th percentile of scores genome wide. The *p*-value from the binomial test is used as a continuous index due to non-independence of SNPs within windows.

We performed the WZA using four different statistics as input: the genome-wide distribution of parametric *p*-values from Kendall's τ (referred to as WZAτ), the genome-wide distribution of Bayes factors as obtained using *BayPass* (referred to as WZA_{BP}), *p*-values from LFMM-LEA (referred to as WZA_{LFMM}) and individual-SNP loadings from RDA (referred to as WZA_{RDA}).

To assess the performance of the different methods, we calculated the area under precision-recall curves (AUC-PR) for each GEA method. AUC-PR is a widely used metric for comparing tests and is particularly useful when datasets have an unbalanced combination of true and false positives (Davis and Goadrich 2006), as in our simulated data. To construct precision-recall curves, confusion matrices were constructed

We examined the effect of variation in recombination rates on the properties of the WZA by manipulating the tree-sequences that we recorded in *SLiM.* In our simulations, genes were 10,000 bp long, so to model genomic regions of low recombination rate, we extracted the coalescent trees that corresponded to the central 1,000bp or 100bp of each gene. For the 1,000bp and 100bp intervals, we added mutations at 10× and 100× the standard mutation rate, respectively.

By default, all SNPs present in each 10,000bp gene in our simulations were analyzed together. However, to explore the effect of window size on the performance of the WZA, we calculated WZA scores for variable numbers of SNPs. In these cases, we calculated WZA scores for all non-overlapping sets of a particular number of SNPs.

Tree sequences were manipulated using the *tskit* package. Mutations were added to trees using the *msprime* (Kelleher et al. 2016; *https://tskit*.dev/msprime/docs/stable/intro.html), *tskit* and *PySLiM* workflow (*https://pyslim*.readthedocs.io/en/latest/). F_{ST} and r^2 (an estimator of linkage disequilibrium) were calculated using custom Python scripts that invoked the *scikit-allel* package (*https://scikit*-allel.readthedocs.io/en/stable/).

Analysis of data from lodgepole pine

We re-analyzed a previously published population genomic dataset for lodgepole pine, *Pinus contorta*, a conifer that is widely distributed across the Northwest of North America. Briefly, Yeaman et al. (2016) collected samples from 254 populations across British Columbia and Alberta, Canada and Northern Washington, USA. The lodgepole pine genome is very large (approximately 20Gbp), so Yeaman et al. (2016) used a sequence capture technique based on the *P. contorta* transcriptome. Allele frequencies were estimated for many markers across the captured portion of the genome by sequencing 1-4 individuals per population. Yeaman et al. (2016) performed GEA on each SNP using Spearman's ρ and used their top-candidate method (see above) to aggregate data across sites within genes. We downloaded the data for individual SNPs from the Dryad repository associated with Yeaman et al. (2016) (*https://doi.org/10.5061/dryad.0t407*). We converted Spearman's ρ *p*-values into empirical *p*-values and performed WZA on the same genes analyzed by Yeaman et al. (2016). We also repeated the top-candidate method, classifying SNPs with empirical *p-values* < 0.01 as outliers. However, as above, we use the *p*-value from the top-candidate method as a continuous index.

Data and Code Availability

The simulation configuration files and code to perform the analysis of simulated data and generate the associated plots are available at *https://github.com/TBooker/WZA*. Analyses were performed using a combination of R and Python. All plots were made using *ggplot2* (Wickham 2016). An implementation of the WZA written in Python can be downloaded from *https://github.com/TBooker/WZA.*

Results

The statistical properties of the WZA

To assess the statistical properties of the WZA, we first analyzed populations evolving neutrally with a constant mutation rate genome-wide. Under neutrality, our simulated metapopulations exhibited a clear pattern of population structure and isolation-by-distance (Figure S1). Figure 2A shows the distribution of WZAτ scores for such populations. The null expectation for WZA scores in this case is the standard normal distribution (mean of 0 and standard deviation of 1), but we found that the distribution of WZAτ scores deviated slightly from this even under neutrality, where the mean and standard deviation of WZAτ scores from individual simulation replicates were approximately 0.089 and 1.38, respectively. Additionally, the inset histogram in Figure 2A shows that distribution of WZAτ scores had a somewhat thicker right-hand tail than expected under the normal distribution. A similar deviation from normality was observed when data were simulated under an island model, or when WZA was performed using Bayes factors from *BayPass* (Figure S4).

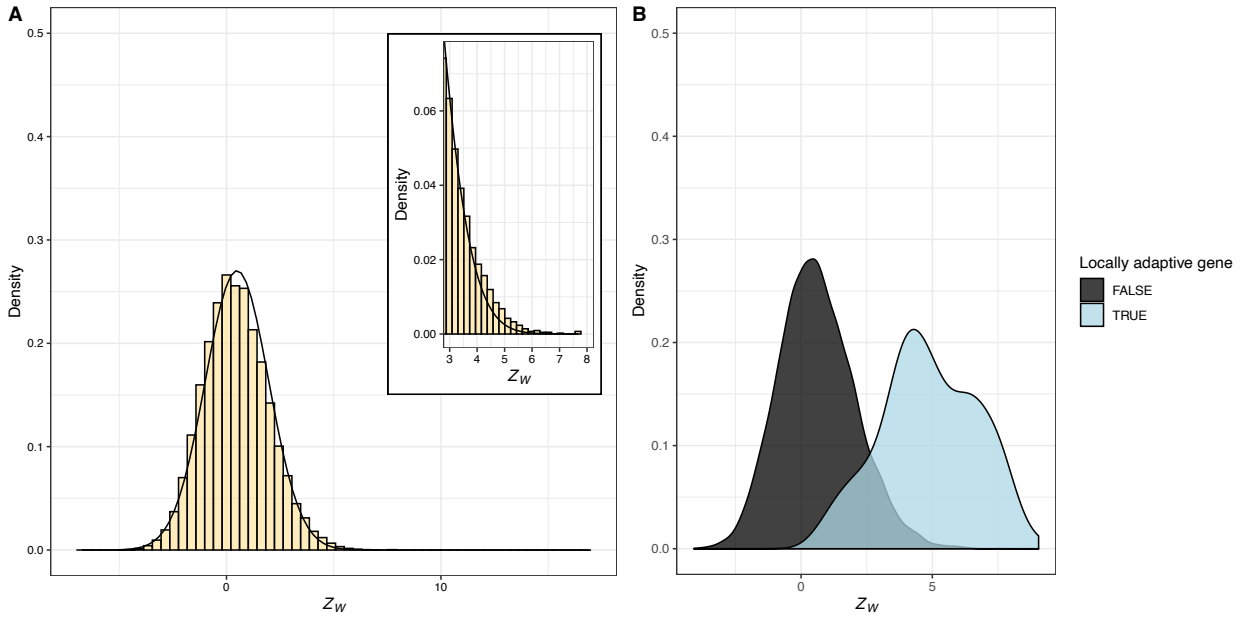


Figure 2. The distribution of WZA scores under neutrality and a model of local adaptation. A) A histogram of WZAτ scores under strict neutrality across a set of 20 replicate simulations, inset is a close-up view of the upper tail of the distribution of Z_W scores. The black line indicates the standard normal distribution. B) A density plot showing the separation of WZAτ scores for genes that are locally adaptive versus evolving neutrally across the genome of 20 simulation replicates. GEA was performed on 40 demes sampled from the *BC Map*.

The deviation from the standard normal distribution is driven by non-independence of SNPs within the analysis windows we used to calculate WZAτ scores. To demonstrate this, we re-calculated WZAτ scores for data simulated under an island model, but permuted the locations of SNPs across the genome, effectively erasing the signal of linkage within genes. The distribution of WZAτ scores in this permuted dataset closely matched the null expectation and did not have a thick right-hand tail (Figure S4; shuffled); each of 30 simulation replicates had a mean WZAτ indistinguishable from 0 with a standard deviation very close to 1. It is worth noting that we modelled populations that did not change in size over time. Non-equilibrium population dynamics such as population expansion may influence the distribution of WZA scores.

The distribution of WZA scores for regions of the genome subject to selection is clearly distinct from that of neutrally evolving genes. Figure 2B shows separation of WZAτ scores for genes that contribute to local adaptation from those that are evolving neutrally (similar results were found for both the *Gradient* and *Truncated* maps; Figure S5). The separation of the distributions of WZAτ scores for locally adaptive genes versus neutrally evolving genes indicates that it may be a powerful method for identifying the genetic basis of local adaptation.

Despite the deviation from strict normality, parametric *p*-values calculated from WZA scores are fairly well behaved, yielding a distribution that is close to uniform for genes not involved in local adaptation (Figure S6). In empirical analyses, the number of SNPs within genes may vary across the genome for many reasons (e.g. variation in sequencing coverage or mutation rate variation). Our implementation of the WZA corrects for variation in SNP number across the genome, and we observe similar distributions of *p*-values when there is wide variation in the number of SNPs in genes (Figure S6). Because the WZA leads to a quasi-uniform distribution of *p*-values under the null hypothesis (Figure S6), parametric *p*-values obtained from the WZA may be used as the basis of explicit hypothesis testing.

The effect of recombination and mutation rate variation on the WZA

Random drift may cause genealogies in some regions of the genome to correlate with environmental variables more than others. Many of the SNPs present in an analysis window that consisted of genealogies that were highly correlated with the environment may be highly significant in a GEA analysis, leading to a large WZA score. This effect would lead to a larger variance in WZA scores for analysis windows that were present in regions of low recombination. To demonstrate this, we down-sampled the tree-sequences we recorded for our simulated populations to model analysis windows present in low recombination regions and performed the WZA on the resulting data. As expected, we found that the variance of the distribution of WZA scores was greater when there was a lower recombination rate (Figure 3A). This is a similar effect to that we described in a previous paper focusing on F_{ST} (Booker et al. 2020).

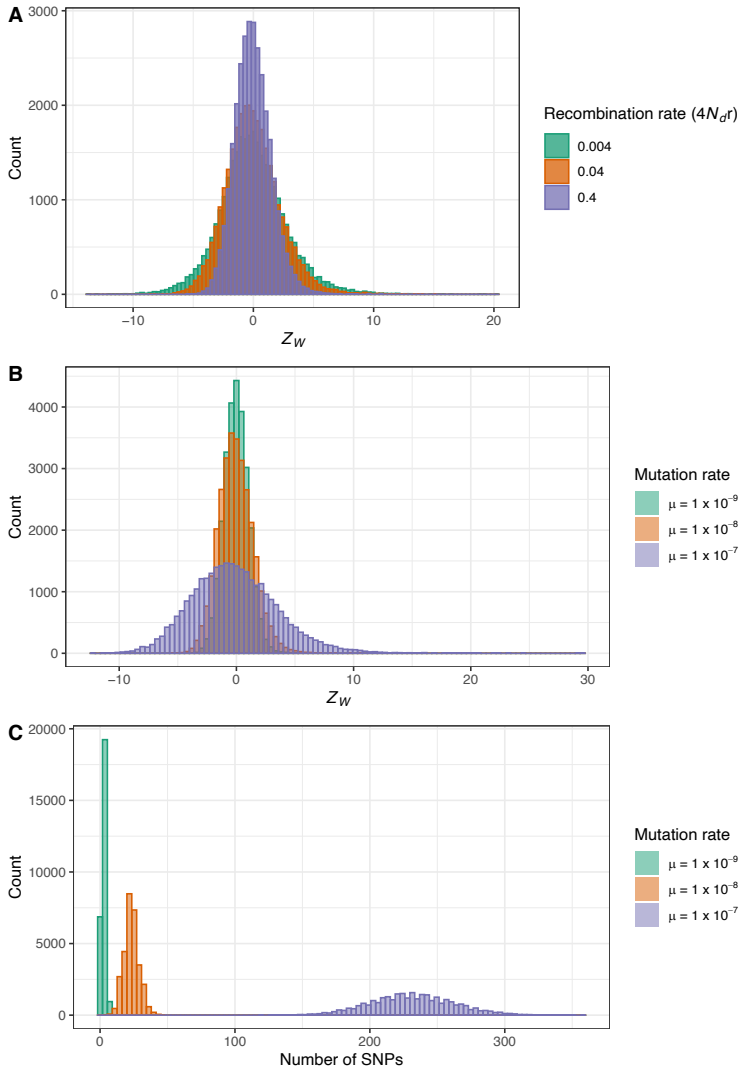


Figure 3 The distribution of Z_W scores under different recombination rates (A), mutation rates (B) and the distribution of the numbers of SNPs associated with different mutation rates (C). Results are shown for neutral simulations using the *BC Map.* WZA scores were calculated from a sample of 40 demes where 50 individuals were sampled in each.

Essentially, the WZA is a method that summarizes evidence for excess correlation with the environment. So, any source of variation in the quantity of evidence will influence the properties of WZA scores. Of particular importance in empirical analysis will be variation in the number of SNPs present in analysis windows across the genome. Numerous factors may contribute to variation in SNP density such as mutation rate variation or targeted sequence capture. Figures 3B-C show how variation in SNP number may lead to heteroscedasticity in Z_W scores, though our method for computing parametric *p*-values from Z_W scores accounts for this (Figure S6). All subsequent analyses focus on cases with wide variation in SNP number across genes.

Comparing the performance of the WZA to other GEA approaches

We compared the performance of the WZA to several widely used GEA analysis methods as well as to a simple rank correlation analysis performed using Kendall's τ and the "top-candidate method" employed by Yeaman et al. (2016). Figure 4 compares the area under precision-recall curves (AUC-PR) for the various GEA methods across the three maps of environmental heterogeneity we simulated. While the various GEA methods varied in their relative performance depending on the map of environmental heterogeneity modelled, the WZA always exhibited the highest or close to the highest AUC-PR. Figure 4 shows results modelling strong selection on locally adaptive alleles, but in the case of more weakly selected alleles all GEA methods had fairly low AUC-PR, though the WZA tended to outperform all other methods (Figure S7).

As expected, the number of sampled demes had a large effect on the performance of GEA methods — sampling fewer demes obviously led to less powerful analyses. However, the WZA still exhibited large AUC-PR even in analyses of only 10 demes (Figure 4, S7). The analyses summarized in Figure 4 modelled a study where 20 individuals were sampled in each deme. Decreasing the number of individuals, and thus increasing the sampling variance of allele frequencies, reduced performance of GEA methods overall, but did not substantially influence the rank order of the performance of the GEA methods (Figure S8). Furthermore, weighting the contribution of individual SNPs to the WZA by *pq* slightly increased performance of the WZA when locally adaptive alleles were weakly selected (Figure S9).

In each of the maps of environmental variation that we simulated, there was a strong correlation between environmental variables and gene flow. There was also a strong pattern of isolation-by-distance in our simulated populations (Figure S1). The combination of these two factors makes it difficult to control the false positive rate in GEA studies (Meirmans 2012). Thus, it is notable the WZA often outperformed *BayPass* and LFMM-LEA, two methods which explicitly control for population structure (Figure 4). When applying the WZA, one could use results from a single-SNP-based method that controls for population structure as input to the WZA. However, we found that empirical *p*-values calculated from the results of Kendall's τ generally provided the highest performance (Figure S10).

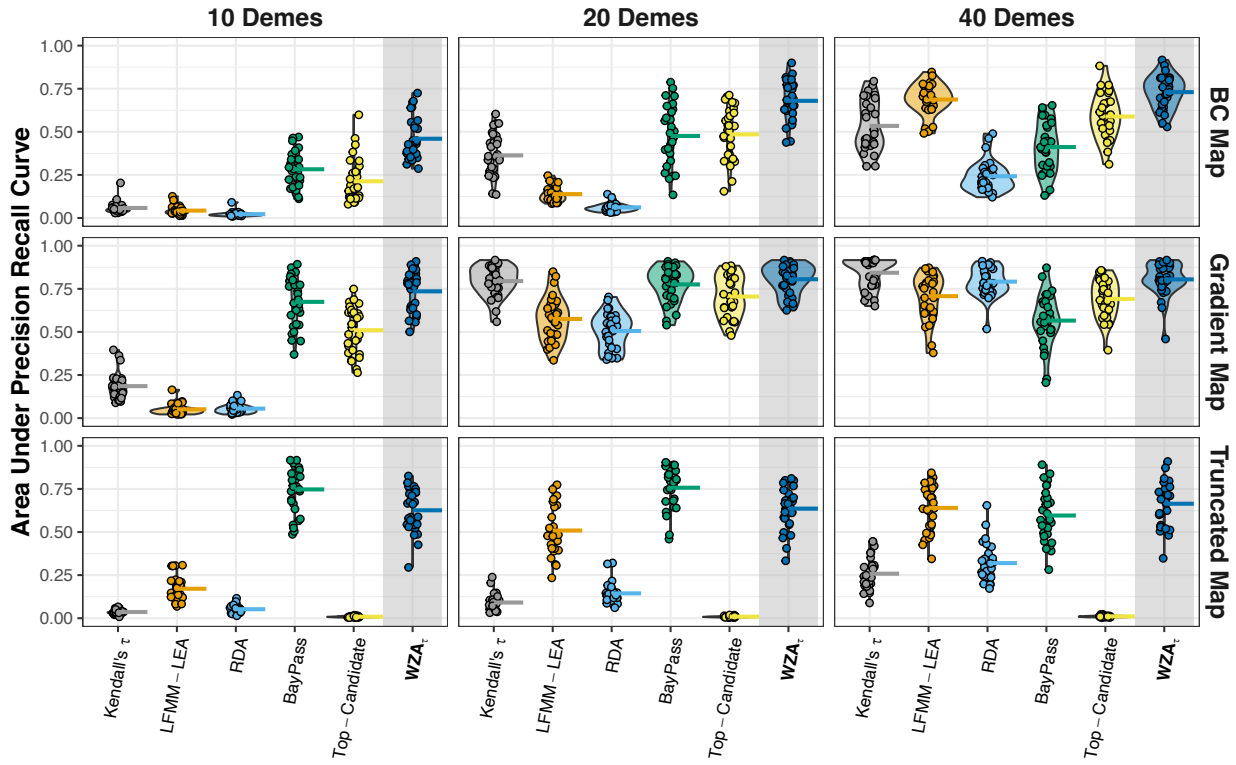


Figure 4 The performance of different GEA methods to identify locally adaptive genes as measured using area under precision-recall curves. Violins indicate the relative density of points and the colored horizontal bars indicate the mean of 30 simulation replicates.

Power and False Discovery Rates of GEA methods

In empirical analyses, GEA summary statistics are often treated as an index of evidence that a particular marker is tagging the location of locally adaptive genetic variation. It is common to see analyses focus on the top x^{th} percentile of GEA scores rather than to treat GEA results as explicit hypothesis tests. Up to this point, we have compared the performance of the WZA to other GEA methods using AUC-PR, a method that characterizes the separation of true positives from true negatives. A large AUC-PR value indicates that a particular statistical test may be a useful index for identifying true positives, but it does not convey the performance of an explicit hypothesis testing framework.

In Figure 5, we compare the performance of GEA methods when applying a genome-wide significance threshold to our simulated datasets. In all cases except *BayPass*, we converted parametric *p*-values into FDR corrected *q*-values using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) and applied a genome-wide significance threshold of $q < 0.05$. For *BayPass* we applied a significance threshold of Bayes Factors > 20dB (i.e. Jeffrey's rule for "decisive evidence"). Using these thresholds we computed the power and false discovery rates (FDR) of the various GEA methods.

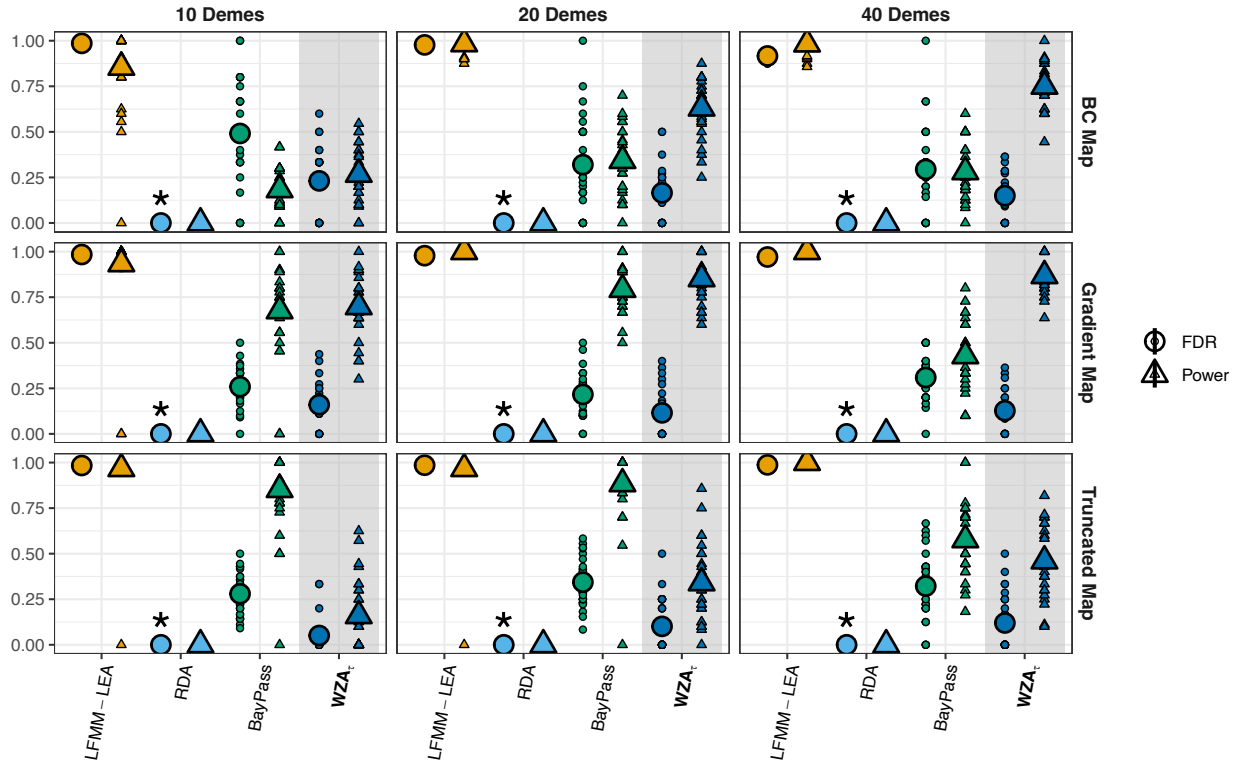


Figure 5 The power and false discovery rate of various GEA methods after applying a stringent genome-wide significance threshold. Small points indicate values for 30 individual simulation replicates, while large shapes indicate the means of the respective statistics. Simulation results shown were obtained by assuming strong selection on locally adaptive alleles and variation in the mutation rate. The asterisk indicates that FDR for RDA was undefined as there were no genome-wide significant hits in any replicates.

The WZA exhibited the best balance of power and FDR across the different sample sizes and maps of environmental heterogeneity (Figure 5). In our analyses, LFMM-LEA had extremely high power, but a large excess of false positives, with FDR values close to 1. The WZA exhibited a higher FDR than expected (i.e. $q < 0.05$), but these were lower than those observed when applying Jeffrey's rule to *BayPass* results (Figure 5). However, *BayPass* exhibited higher power than the WZA when analyzing data simulated under the *Truncated* map. Application of RDA did not lead to parametric *p*-values that were significant genome-wide. Qualitatively similar results were obtained when modelling local adaptation via weakly selected alleles (Figure S11).

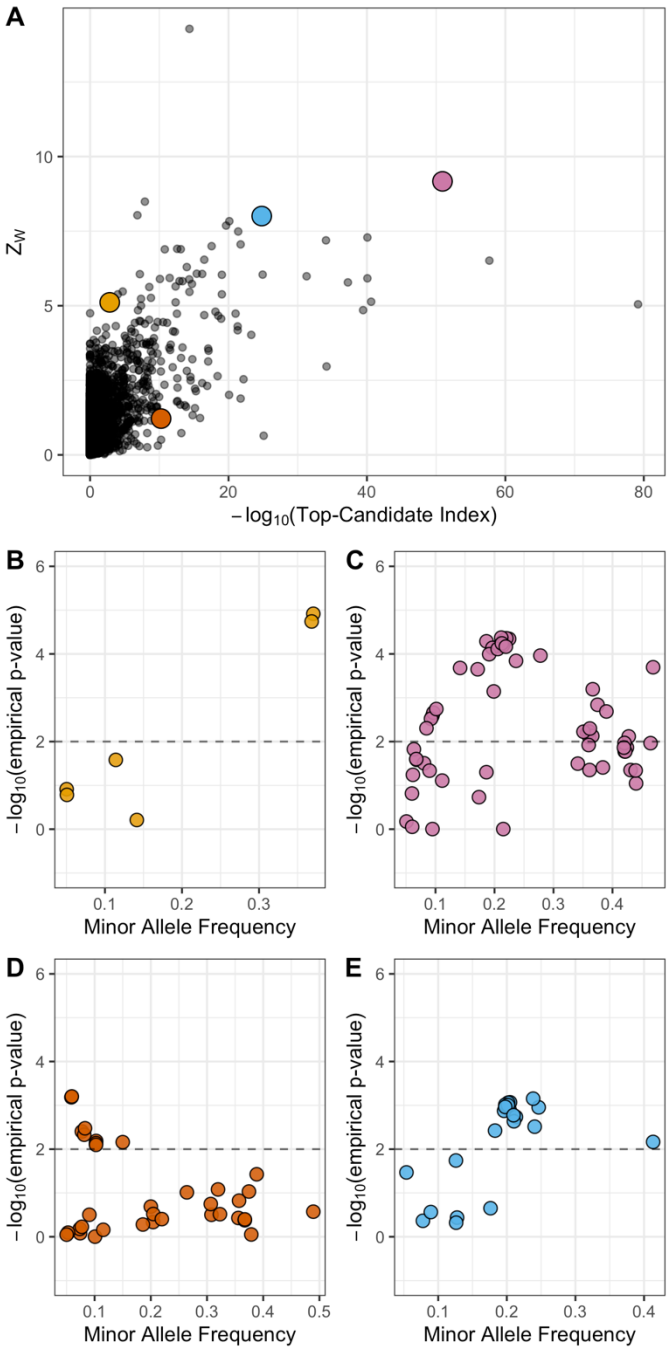
Application of the WZA to lodgepole pine data

We re-analyzed a previously published (Yeaman et al. 2016) lodgepole pine (*Pinus contorta*) dataset and compared the WZA to the top-candidate method, which had been developed for the original study. Following their approach, we analyzed windows spanning the start and end-points of genes when we re-analyzed their data. We applied

562 the WZA to all genes, but for genes that contained more than 21 SNPs (the 75th
563 percentile), we resampled sets of 21 SNPs and calculated WZA scores 100 times taking
564 the average of the resampled WZA scores as our point estimate.

565 Overall, the WZA and top candidate statistic were broadly correlated and identified
566 many of the same genes as the most strongly associated loci, but also differed in
567 important ways. Figure 6A shows the relationship between WZA scores and the
568 $-log_{10}(p\text{-value})$ from the top-candidate method, which were positively correlated
569 (Kendall's τ = 0.213, *p*-value < 10^{-16}). There were several genes that had strong
570 evidence for environmental association from WZA, but only very modest top-candidate
571 scores (Figure 6A). Figure 6B shows that for one such region, there were several SNPs
572 with high minor allele frequency that have small *p*-values. Conversely, Figure 6E shows
573 a region that only had a very modest WZA score, but an extreme score from the top-
574 candidate method. In this case, there were numerous SNPs that passed the top-
575 candidate outlier threshold, but they were mostly at low allele frequency. Figures 6C&E
576 show the relationship between allele frequency and the empirical *p*-value for SNPs
577 present in two genes that had extreme scores from both the top-candidate method and
578 the WZA.

579



580

581 **Figure 6** The WZA applied to GEA results on lodgepole pine for degree days below 0
582 (DD0). A) Z_w scores compared to scores from the top-candidate method for each of the
583 genes analyzed by Yeaman et al. (2016). Panels B-E show the results for $-log_{10}$(p-
584 values) for Spearman's ρ applied to individual SNPs against minor allele frequency
585 (MAF) for the colored points in A. The dashed horizontal lines in B-E indicates the
586 significance threshold used for the top-candidate method (i.e. 99^{th} percentile of GEA
587 $-log_{10}$(*p*-values) genome-wide).

588

Discussion

In this study, we have shown that combining information across linked sites in GEA analyses is a potentially powerful way to identify genomic loci involved in local adaptation. The method we propose, the WZA, was typically more powerful than standard methods that look at individual sites in isolation, particularly when working with small samples or local adaptation generated by weakly selected alleles (Figures 4 and 5).

In a hypothetical world where one had perfect knowledge of allele frequency variation across a species' range for all sites across the genome, a single marker approach would likely be the best way to perform a GEA analysis, as one would be able to determine the true correlation between genetic and environmental variation for each site in the genome. However, such a situation is unrealistic, and empirical GEA studies will likely always be limited to samples from only some of the populations of interest. Thus, leveraging the correlated information present among closely linked sites in GEA studies may provide a powerful method for identifying the genetic basis of local adaptation.

The effects of population structure on GEA analyses

A striking result from our comparison of the various GEA methods we tested in this study was how Kendall's τ often outperformed other single-SNP analyses (Figure 4). As mentioned in the Introduction, Lotterhos (2019) obtained a similar result in a previous study. This presumably occurs because genome-wide population genetic structure is oriented along a similar spatial axis as adaptation, and the methods that *BayPass* and LFMM-LEA use to incorporate population structure cause a reduction in the signal of association at genes involved in adaptation. In such cases, the use of simple rank correlations such as Spearman's ρ or Kendall's τ, which assume that all demes are independent, may often yield a skewed distribution of *p*-values. Such a distribution would lead to a large number of false positives if a standard significance threshold were used (Meirmans 2012). Here, we avoid standard significance testing, and instead make use of an attractive quality of the distribution of *p*-values: SNPs in regions of the genome that contribute to adaptation tend to have extreme p-values, relative to the genome-wide distribution. By converting them to empirical *p*-values, we retain the information contained in the rank-order of *p*-values, but reduce the inflation of their magnitude, which increases the power of the test (Figure S12). While the empirical *p*-value approach may partially and indirectly correct for false positives due to population structure genome-wide, it loses information contained in the raw *p*-value that represents the deviation of the data from the null model for our summary statistic of interest. It is possible that a GEA approach that produced parametric *p*-values that was adequately controlled for population structure may provide a more powerful input statistic to the WZA, although that was not the case when we tested WZA based on results from *BayPass* and *LFMM-LEA* (Figure S10).

Perhaps more striking is that the false discovery rate of GEA methods were often much higher than expected (Figure 5, S11). This implies that many of the empirical studies

that have employed those methods may have higher false positive rates than stated or assumed. Furthermore, we also found that RDA did not yield *p*-values that were significant genome-wide (Figure 5, S11), though it is worth pointing out that we were performing univariate GEA analyses and one of the strengths of RDA an an approach is that it is capable of modelling multi-variate environments (Capblancq and Forester 2021). Using the results of a multi-variate RDA as input to the WZA may prove to be a powerful GEA method.

Why use analysis windows?

Theoretical studies of local adaptation suggest that we should expect regions of the genome subject to spatially varying selection pressures to exhibit elevated linkage disequilibrium (LD) relative to the genomic background for a number of reasons. Under local adaptation, alleles are subject to spatial fluctuation in the direction of selection. As a locally adaptive allele spreads in the locations where it is beneficial, it may cause some linked neutral variants to hitchhike along with it (Sakamoto and Innan 2019). LD can be increased further as non-beneficial genetic variants introduced to local populations via gene flow are removed by selection. This process can be thought of as a local barrier to gene flow acting in proportion to the linkage with a selected site (Barton and Bengtsson 1986). Beyond this hitchhiking signature, there is a selective advantage for alleles that are involved in local adaptation to cluster together, particularly in regions of low recombination (Rieseberg 2001; Noor et al. 2001; Kirkpatrick and Barton 2006; Yeaman 2013). For example, in sunflowers and *Littorina* marine snails, there is evidence that regions of suppressed recombination cause alleles involved in local adaptation to be inherited together (Morales et al. 2019; Todesco et al. 2020). The processes we have outlined are not mutually exclusive, but overall, genomic regions containing strongly selected alleles that contribute to local adaptation may have elevated LD and potentially exhibit GEA signals at multiple linked sites. Window-based GEA scans can potentially take advantage of the LD that is induced by local adaptation, aiding in the discovery of locally adaptive genetic variation.

The two window-based GEA methods we compared in this study, the WZA and the top-candidate method of Yeaman et al. (2016), were fairly similar in power in some cases, but the WZA was most often better (Figure 5). Moreover, there are philosophical reasons as to why WZA should be preferred over the top-candidate method. Firstly, the top-candidate method requires the use of more arbitrary significance thresholds. Secondly, the top-candidate method gives equal weight to all SNPs that have exceeded the significance threshold. For example, with a threshold of α = 0.01, genomic regions with only a single outlier are treated in the same way whether that outlier has a *p*-value of 0.009 or 10^{-5}. It is desirable to retain information about particularly strong outliers. It should be kept in mind, however, that the WZA (and the top-candidate method for that matter) does not explicitly test for local adaptation and only provides an indication of whether a particular genomic region has a pattern that deviates from the genome-wide average. Indeed, numerous processes other than local adaptation may cause excessive correlation between environmental variables and allele frequencies in particular genomic regions. For example, population expansions can cause allelic surfing, where regions of the genome “surf" to high frequency at leading edges of expanding

674 populations. Allelic surfing can leave heterogeneous patterns of variation across a
675 species range leaving signals across the genome that may resemble local adaptation
676 (Novembre and Di Rienzo 2009; Klopfstein, Currat, and Excoffier 2006).

677 Combining information from multiple association tests in genomic wide analyses is not
678 unique to the present study. There are numerous methods that have been proposed for
679 combining *p*-values from genome-wide association studies within genes or specific
680 genomic regions; e.g. MAGMA (de Leeuw et al. 2015) and comb-*p* (Pedersen et al.
681 2012). In comb-*p*, for example, *p*-values within genes are combined in such a way as to
682 diminish the influence of LD from linked sites, which is conceptually similar to LD
683 pruning or clumping. Such approaches reduce the burden of multiple comparisons and
684 the effects of pseudoreplication in genome-wide association studies, where the goal
685 may be to identify loci that are not expected to be present in regions of high LD. In
686 contrast, with the WZA we are searching for genomic regions with an evolutionary
687 history that correlates with environmental heterogeneity. With that goal in mind, we use
688 all the information available (i.e. all SNPs) to try and characterize whether there is truly
689 an association between evolutionary history and environmental heterogeneity in a part
690 of region of the genome. Our approach has the benefit of potentially capitalizing on the
691 LD that is expected to be generated by local adaptation.

692 ## *Choosing the width of analysis windows for the WZA*

693 When performing a genome-scan using a windowed approach a question that inevitably
694 arises is, how to choose the width of analysis windows? In window-based genome
695 scans, summary statistics sensitive to particular evolutionary processes (such as
696 nucleotide diversity or Tajima's D) are calculated for analysis windows sized such that
697 the coalescent history across the window is more or less homogeneous. If analysis
698 windows were too narrow, there may be little benefit in using a windowed approach over
699 a single-SNP approach, while if analysis windows are too wide the evolutionary signal of
700 interest may be diluted by unlinked sites. Regions of the genome in tight linkage will
701 recombine less frequently than more loosely linked sites. Sites that are separated by an
702 effective recombination fraction much less than the reciprocal of the time to the most
703 recent common ancestor are not expected to recombine in the coalescent history of a
704 sample (Wakeley 2005). If there has been little to no recombination across a window in
705 the coalescent history of a sample, SNPs present in that window will all reflect the
706 underlying genealogy and potentially the evolutionary processes that have shaped it.
707 This idea forms the logic behind the choice of analysis window width in genome scan
708 studies.

709 The WZA is aimed at identifying regions of the genome that contribute to local
710 adaptation by combining information across closely linked sites that have similar
711 evolutionary histories. In the absence of information about recombination rates, one can
712 get a sense for the average distance over which recombination breaks down
713 associations among sites by examining the decay of linkage disequilibrium (LD) among
714 pairs of SNPs. Regions of the genome that contribute to local adaptation are expected
715 to exhibit elevated LD compared to neutrally evolving sites (see above), which is what

716 we see in our simulated data (Figure S1). High LD across an analysis window indicates
717 a homogeneous coalescent history.

718 When setting the width of analysis windows for the WZA, we recommend that users aim
719 for a window size that is wider than the expected pattern of LD decay for neutral sites,
720 to capitalize on the LD-inducing effects of local adaptation. For example, in our
721 simulations LD at neutral sites decayed rapidly, on the order of 1Kbp or so (Figure S1).
722 When performing the WZA on our simulated data, we used windows of 10Kbp as we
723 found narrower windows were intermediate in performance between the single-SNP and
724 10,000bp approaches (Figure S13). If the width of analysis windows is close to the
725 width over which LD typically decays, neutrally evolving regions that happen to have a
726 coalescent history that correlates with the environment may exhibit extreme WZA
727 scores and there may be little to distinguish them from regions that are affected by
728 adaptation. The inclusion of loosely linked SNPs for neutral regions will dilute the
729 information about segments of the genome that have coalescent histories that closely
730 align with environmental variation.

731 Of course, if recombination rate varies widely across the genome, that will influence the
732 ability to interpret the results (Figure 3; Booker et al. 2020). If possible, one should
733 incorporate information on recombination rate variability into their analyses; for example
734 by altering the size of windows as a function of the recombination rate.

735 ## *Future directions*

736 Ultimately, performing GEA analyses using analysis windows is an attempt to leverage
737 information from closely linked sites to identify loci involved in local adaptation. The
738 WZA could potentially be used with other statistics where LD is expected to result in
739 correlated signals across physically linked nucleotides, for example *p*-values from
740 genome-wide association studies on the basis of phenotypic standing variation, but
741 power in this context would need to be assessed by further testing. With the advent of
742 methods for reconstructing ancestral recombination graphs from population genomic
743 data (Hejase et al. 2020), perhaps a GEA method could be developed that explicitly
744 analyzes inferred genealogies rather than individual markers in a manner similar to
745 regression of phenotypes on genealogies proposed by Ralph et al. (2020). Such a
746 method would require large numbers of individuals with phased genome sequences,
747 which may now be feasible given recent technological advances (Meier et al. 2021).

748 However, there are scenarios where incorporating information from linked sites in GEA
749 analyses may obscure the signal of local adaptation. For example, the power of the
750 WZA could be reduced if causal alleles contributed to local adaptation along multiple
751 gradients (e.g. to altitudinal gradients in several distinct mountain ranges). If such
752 gradients were semi-independent (i.e. medium/high F_{ST} among gradients), and then
753 there may be a different combination of neutral variants in high LD with the causal allele
754 in each case. In such a scenario, the species-wide LD in regions flanking the causal
755 locus may be reduced, which would likely also reduce the power of the WZA.
756 Furthermore, if local adaptation is typically caused by rare alleles, GEA may simply be
757 an underpowered analysis to detect the genetic basis of adaptation.

Conclusions

Theoretical models of local adaptation suggest that we should expect elevated LD in genomic regions subject to spatially varying selection pressures. For that reason, GEA analyses may gain power by making use of information encoded in patterns of tightly linked genetic variation. The method we propose in this study, the WZA, aims to do that. The WZA outperforms single-SNP approaches in a range of settings and so provides researchers with a powerful tool to characterize the genetic basis of local adaptation in population and landscape genomic studies.

Acknowledgements

Thanks to Finlay Booker for moral support throughout the course of this project. Thanks to Pooja Singh for many helpful discussions, to Tongli Wang for help with BC climate data and to Simon Kapitza for help with wrangling raster files. Thanks to Jared Grummer, Tyler Kent and Isabela Jerônimo Bezerra do Ó for comments on the manuscript and to three anonymous reviewers for valuable suggestions and comments. Funding for this work was provided by Genome Canada, Genome Alberta and NSERC Discovery Grants awarded to MCW and SY. SY is supported by an AIHS research chair. Computational Support was provided by Compute Canada. This study is part of the CoAdapTree project which is funded by Genome Canada (241REF), Genome BC and 16 other sponsors (http://coadaptree.forestry.ubc.ca/sponsors/).

Bibliography

Aitken, SN, and Whitlock, MC. 2013. Assisted gene flow to facilitate local adaptation to climate change. *Annu. Rev. Ecol. Evol. Syst.* 44 (1): 367–88.

Barton N., and Bengtsson B. O. 1986. The barrier to genetic exchange between hybridising populations. *Heredity* 57 (3): 357–76.

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), pp.289-300.

Bhatia G., Patterson N., Sankararaman, S. and Price AL. 2013. Estimating and interpreting FST: The impact of rare variants. *Genome Research* 23 (9): 1514–21.

Bontrager, M., Muir CD, Mahony C, Gamble DE, Germain RM, Hargreaves AL, Kleynhans EJ, Thompson KA, and Angert AL. 2020. Climate warming weakens local adaptation. bioRxiv. *https://doi.org/10.1101/2020.11.01.364349*.

Booker, TR., Yeaman S, and Whitlock MC. 2020. Variation in recombination rate affects detection of outliers in genome scans under neutrality. *Molecular Ecology* 29 (22): 4274–9.

Capblancq, T. and Forester, B.R., 2021. Redundancy analysis: A Swiss Army Knife for landscape genomics. *Methods in Ecology and Evolution*, 12(12), pp.2298-2309.

Charlesworth B and Charlesworth D. 2010. *Elements of Evolutionary Genetics*. Greenwood Village, Colorado: Roberts & Company.

Coop, G, Witonsky D, Rienzo A, and Pritchard J. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185 (4): 1411–23.

Davis, J. and Goadrich, M., 2006, June. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240).

de Leeuw, C.A., Mooij, J.M., Heskes, T. and Posthuma, D., 2015. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Computational Biology*, 11(4), p.e1004219.

Forester BR, Jones MR, Joost S, Landguth EL, and Lasky JR. 2016. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Molecular Ecology* 25 (1): 104–20.

Forester, BR., Lasky JR, Wagner HH, and Urban DL. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. *Molecular Ecology* 27 (9): 2215–33.

812 Frichot E, and François O. 2015. LEA: An R package for landscape and ecological
813 association studies. Edited by Brian O'Meara. *Methods in Ecology and Evolution* 6 (8):
814 925–29.

815 Frichot E, Schoville SD, Bouchard G, and François O. 2013. Testing for associations
816 between loci and environmental gradients using latent factor mixed models. *Molecular*
817 *Biology and Evolution* 30 (7): 1687–99.

818 Gautier M. 2015. Genome-wide scan for adaptive divergence and association with
819 population-specific covariates. *Genetics* 201 (4): 1555–79.

820 Haldane JBS. 1948. The theory of a cline. *Journal of Genetics* 48 (3): 277–84.

821 Haller, B.C. and Messer, P.W., 2019. SLiM 3: forward genetic simulations beyond the
822 Wright–Fisher model. *Molecular Biology and Evolution*, 36(3), pp.632-637.

823 Haller BC, Galloway J, Kelleher J, Messer PW, and Ralph PL. 2019. Tree-sequence
824 recording in SLiM opens new horizons for forward-time simulation of whole genomes.
825 *Molecular Ecology Resources* 19 (2): 552–66.

826 Hancock AM, Brachi B, Faure N, Horton MW, Jarymowycz LB, Sperone FG, Toomajian
827 C, Roux F, and Bergelson J. 2011. Adaptation to climate across the Arabidopsis
828 thaliana genome. *Science* 334 (6052): 83–86.

829 Hejase HA, Dukler N, and Siepel A. 2020. From summary statistics to gene trees:
830 Methods for inferring positive selection. *Trends in Genetics* 36(4): 243-258.

831 Hereford, Joe. 2009. A quantitative survey of local adaptation and fitness trade-offs.
832 *American Naturalist* 173 (5): 579-8

833 Hoban S, Kelley JL, Lotterhos KE, Antolin MF, Bradburd G, Lowry DB, Poss ML, Reed
834 LK, Storfer A, and Whitlock MC. 2016. Finding the genomic basis of local adaptation:
835 Pitfalls, practical solutions, and future directions. *American Naturalist* 188 (4): 379–97.

836 Kelleher J, Etheridge AM, and McVean G. 2016. Efficient coalescent simulation and
837 genealogical analysis for large sample sizes. *PLOS Computational Biology*, *12*(5),
838 e1004842.

839 Kirkpatrick M, and Barton N. 2006. Chromosome inversions, local adaptation and
840 speciation. *Genetics* 173 (1): 419–34.

841 Klopfstein S, Currat M, and Excoffier L. 2006. The fate of mutations surfing on the wave
842 of a range expansion. *Molecular Biology and Evolution*. 23(3): 482-490.

843 Liptak, T. 1958. On the combination of independent tests*. Magyar Tud. Akad. Mat.*
844 *Kutato Int. Kozl*. 3: 171– 197.

Lotterhos KE. 2019. The effect of neutral recombination variation on genome scans for selection. *G3* 9 (6): 1851–67.

Meier, J.I., Salazar, P.A., Kučka, M., Davies, R.W., Dréau, A., Aldás, I., Box Power, O., Nadeau, N.J., Bridle, J.R., Rolian, C. and Barton, N.H., 2021. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species*. Proceedings of the National Academy of Sciences*, 118(25), p.e2015005118.

Meirmans PG. 2012. The trouble with isolation by distance. *Molecular Ecology* 21 (12): 2839–46.

Mimura M, and Aitken SN. 2007. Adaptive gradients and isolation-by-distance with postglacial migration in *Picea sitchensis*. *Heredity* 99 (2): 224–32.

Morales HE, Faria R, Johannesson K, Larsson T, Panova M, Westram AM, and Butlin RK. 2019. Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast. *Science Advances* 5 (12): eaav9963.

Mosca E, González-Martínez SC, and Neale DB. 2014. Environmental versus geographical determinants of genetic structure in two subalpine conifers. *New Phytologist* 201 (1): 180–92.

Mosteller, F. & Bush, R.R. 1954. Selected quantitative techniques. In*: Handbook of Social Psychology,* Vol. 1 (G. Lindzey, ed.), pp. 289– 334. Addison-Wesley, Cambridge, Mass.

Noor MAF, Gratos KL, Bertucci LA and Reiland J. 2001. Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences of the United States of America* 98 (21): 12084–8.

Novembre J, and Rienzo A. 2009. Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet.* 10: 645-755

Pavy N, Namroud MC, Gagnon F, Isabel N, and Bousquet J. 2012. The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers. *Heredity* 108 (3): 273–84.

Pedersen, B.S., Schwartz, D.A., Yang, I.V. and Kechris, K.J., 2012. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics*, 28(22), pp.2986-2988.

Ralph P, Thornton K, and Kelleher K. 2020. Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes. *Genetics* 215 (3): 779–97.

Rieseberg LH. 2001. Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution* 16(7): 351-358.

Sakamoto T, and Innan H. 2019. The evolutionary dynamics of a genetic barrier to gene flow: From the establishment to the emergence of a peak of divergence. *Genetics* 212 (4): 1383–98.

Savolainen, O., Lascoux, M. and Merilä, J., 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14(11), pp.807-820.

Stapley J, Feulner PGD, Johnston SE, Santure AW, and Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: Patterns and processes. *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1736).

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams, R.M. Jr. 1949. *The American Soldier, Vol. 1: Adjustment during Army Life*. Princeton University Press, Princeton.

Todesco M, Owens GL, Bercovich N, Légaré JS, Soudi S, Burge DO, Huang K, et al. 2020. Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* 584 (7822): 602–7.

Wang T, Hamann A, Spittlehouse D and Carroll C. 2016. Locally downscaled and spatially customizable climate data for historical and future periods for North America. *PLOS ONE* 11 (6): e0156720.

Wakeley, J., 2009. Coalescent theory. *Roberts & Company.*

Weir BS and Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* 38 (6): 1358–70.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, *https://ggplot2.tidyverse.org*.

Whitlock MC. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* 18 (5): 1368–73.

Yeaman S. 2013. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America* 110 (19): E1743–E1751.

Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, Nurkowski KA, et al. 2016. Convergent local adaptation to climate in distantly related conifers. *Science* 353 (6306): 1431–3.

Appendix

Parametrizing simulations of local adaptation

Consider a hypothetical species of conifer inhabiting British Columbia, Canada. There may be many hundreds of millions of individuals in this hypothetical species distributed across the landscape. It would be computationally intractable to simulate all individuals forward-in-time incorporating adaptation to environmental variation across the landscape with recombining chromosomes, even with modern population genetic simulators. In our simulations we scaled several population genetic parameters to model a large population when simulating a much smaller one. In the following sections, we outline and justify the approach we used to scale pertinent population genetic parameters.

Mutation rate

We set the neutral mutation rate such that there would be an average of around 20 SNPs in each gene after applying a minor allele frequency threshold of >0.05. This number was motivated by the average number of SNPs per gene in the lodgepole pine dataset described by Yeaman et al. (2016). We found that a neutral mutation rate (μ_{neu}) of 10^{-8} in our simulations achieved an average of 23.3. Note that this μ_{neu} gave a very low population-mutation rate within demes, $4N_d\mu_{neu} = 4.0 \times 10^{-6}$.

There are no estimates available of the mutation rate to locally adaptive alleles. We opted to use mutation rates that resulted in multiple locally beneficial alleles establishing in our simulations. For directional selection, we found that a mutation rate of $\mu_{alpha} = 3 \times 10^{-7}$ resulted in around 6 locally adaptive genes establishing. For stabilizing selection, a mutation rate of $\mu_{alpha} = 1 \times 10^{-10}$, resulted in similar numbers of genes establishing. Note that in our model of directional selection, only a single nucleotide in each of 12 genes could mutate to a locally beneficial allele. In the case of stabilizing selection, all 10,000bp in the simulated gene could give rise to mutations that affected phenotype.

Recombination rates

We based our choice of recombination rate on patterns of LD decay reported for conifers. The pattern of LD decay in a panmictic population can be predicted by the population-scaled recombination parameter ($\rho = 4N_e r$; Charlesworth and Charlesworth 2010), but the pattern of LD decay in structured populations is less well described. In conifers, LD decays very rapidly in conifers and $\rho \approx 0.005$ has been estimated (Pavy et al. 2012). However, per basepair recombination rates (r) in conifers are extremely low, estimated to be on the order of 0.05 cM/Mbp - more than 10× lower than the average for humans (Stapley et al. 2017). This implies a very large effective population size of roughly $\frac{0.005}{4 \times 0.5 \times 10^{-8}} = 2.5 \times 10^6$, much larger than is feasible to simulate. To acheive a

similar number of recombination events through time in our simulated populations, we needed to increase r above what has been empirically estimated. We chose a recombination rate that gave us a pattern of LD decay that was similar to what has been observed in conifers. We found that a per base pair recombination $r = 1 \times 10^{-7}$ (i.e. roughly $200\times$ greater than in natural populations) gave a pattern of LD in our simulated populations that was similar to what has been reported for conifers.

Selection coefficients

It is difficult to choose a realistic set of selection parameters for modelling local adaptation because there are, at present, no estimates of the distribution of fitness effects for mutations that have spatially divergent effects. However, common garden studies of a variety of taxa have estimated fitness differences of up to 35-45% between populations grown in home-like conditions versus away-like conditions (Hereford 2009; Bontrager et al. 2020). Motivated by such studies, we chose to parametrize selection using the fitness difference between home versus away environments.

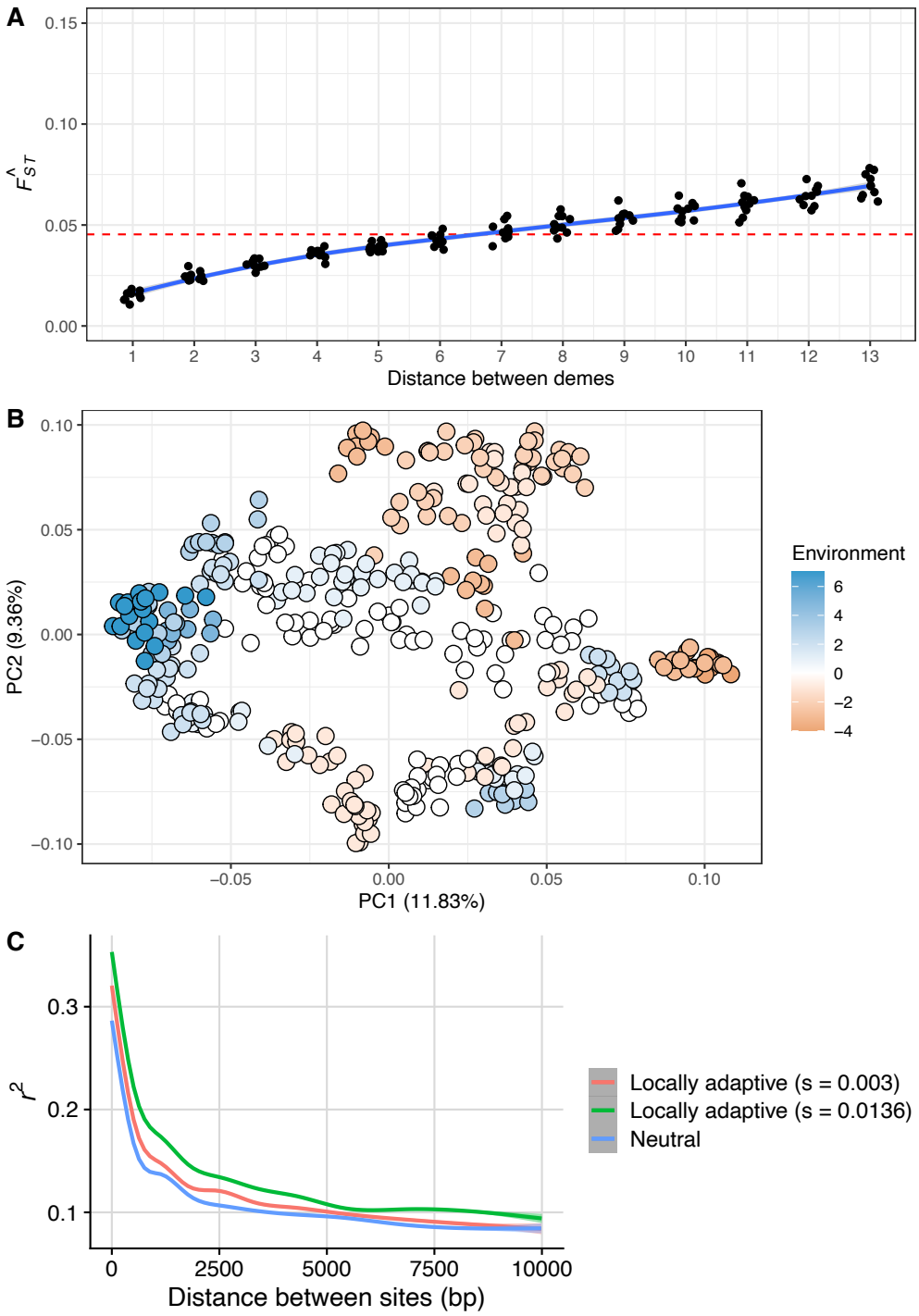
Our simulations contained 12 loci that could mutate to generate a locally beneficial allele. The phenotypic optima that we simulated ranged from -7 to 7 and we modelled selection on a locus as $1 + s_a\theta$ for a homozygote and $1 + hs_a\theta$ for a heterozygote, where s_a is the selection coefficient, θ is the phenotypic optimum and h is the dominance coefficient. With a selection coefficient of $s_a = 0.003$, the maximum relative fitness was $(1 + 7 \times s_a)^{12} = 1.28$ for an individual homozygous for all locally beneficial alleles. An individual homozygous for those alleles, but in the oppositely selected environment (i.e. present in the wrong deme) had a fitness of $(1 - 7 \times s_a)^{12} = 0.775$. Thus, there would be approximately 40% difference in fitness between well locally adapted individuals at home versus away in the most extreme case. Note, however, that approximately 6 genes established in each simulation replicate, so the realized fitness difference was closer to a 20% difference. We also simulated stronger selection with a selection coefficient of $s_a = 0.0136$, which corresponds to approximately 90% difference in fitness between well locally adapted individuals at home versus away in the most extreme case. In these simulations 12 genes established in most cases.

Migration rate

We wanted to model populations with F_{ST} across the metapopulation of approximately 0.05, as has been reported for widely distributed conifer species such as lodgepole pine and interior spruce (Yeaman et al. 2016). For the stepping-stone simulations, we chose a migration rate of $\frac{7.5}{2N_d}$ as we found that this gave a mean F_{ST} of 0.04. For an island model, we used the analytical formulae given in the main text to set m to achieve a mean F_{ST} of 0.03.

Table S1 Population genetic parameters of a hypothetical organism, and how they are scaled in the simulations. The meta-population inhabits a 14×14 2-dimensional stepping stone. Parameters are shown for a population with 12 loci subject to directional selection.

Parameter	Hypothetical Biological Value	Scaled Parameter	Unscaled (Simulation)
Global population size (N_e)	10^6	-	19,600
Number of demes (d)	196	-	196
Local population size (N_d)	5,100	-	100
Recombination rate (r)	2.00×10^{-9}	$4N_d r = 0.00004$	1×10^{-7}
Selection coefficient (s_a)	0.0001	$2N_d s_a = 0.6$	0.003
Migration rate (m)	7.35×10^{-4}	$2N_d m = 7.5$	0.0375
Neutral mutation rate (μ_{neu})	2×10^{-10}	$4N_e \mu_{neu} = 0.000004$	10^{-8}
Functional mutation rate (μ_α)	2×10^{-9}	$4N_e \mu_\alpha = 0.00004$	3×10^{-7}



992

993 **Figure S1** Summary statistics from simulations. A) F_{ST} between pairs of demes in
994 stepping-stone populations from neutral simulations. The average F_{ST} across replicates
995 is 0.042. B) Principal components plot of data simulated under the BC map showing that
996 the first two axes of variation. C) LOESS smoothed LD, as measured by r^2, between
997 pairs of SNPs in genes that are either evolving neutrally are locally adaptation as
998 indicated by the color. Smoothing was performed using the ggplot2 package in R.

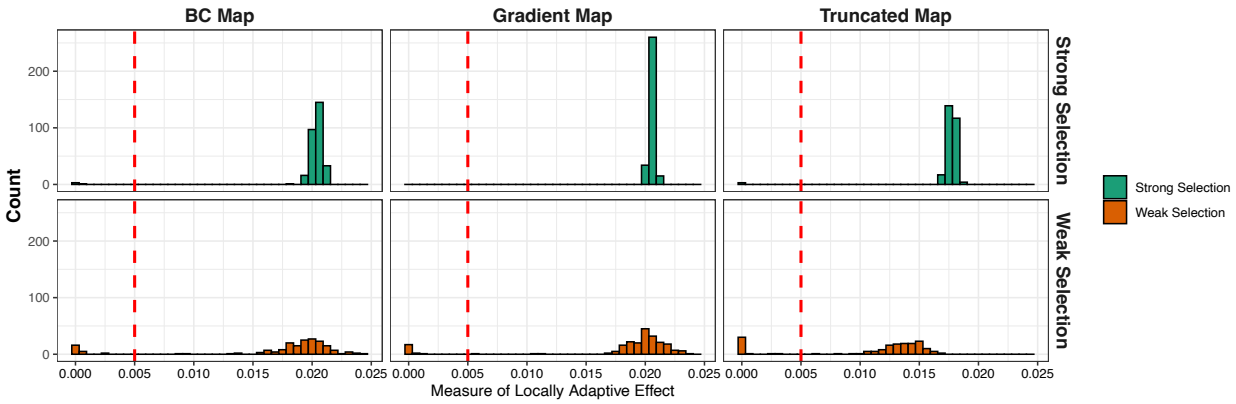


Figure S2 The distribution of effect sizes per gene from simulations modelling local adaptation. The effect size was we used the covariance between the fitness of a gene and the environment. The vertical line indicates the threshold we applied to the simulated data to classify genes as locally adaptive or not.

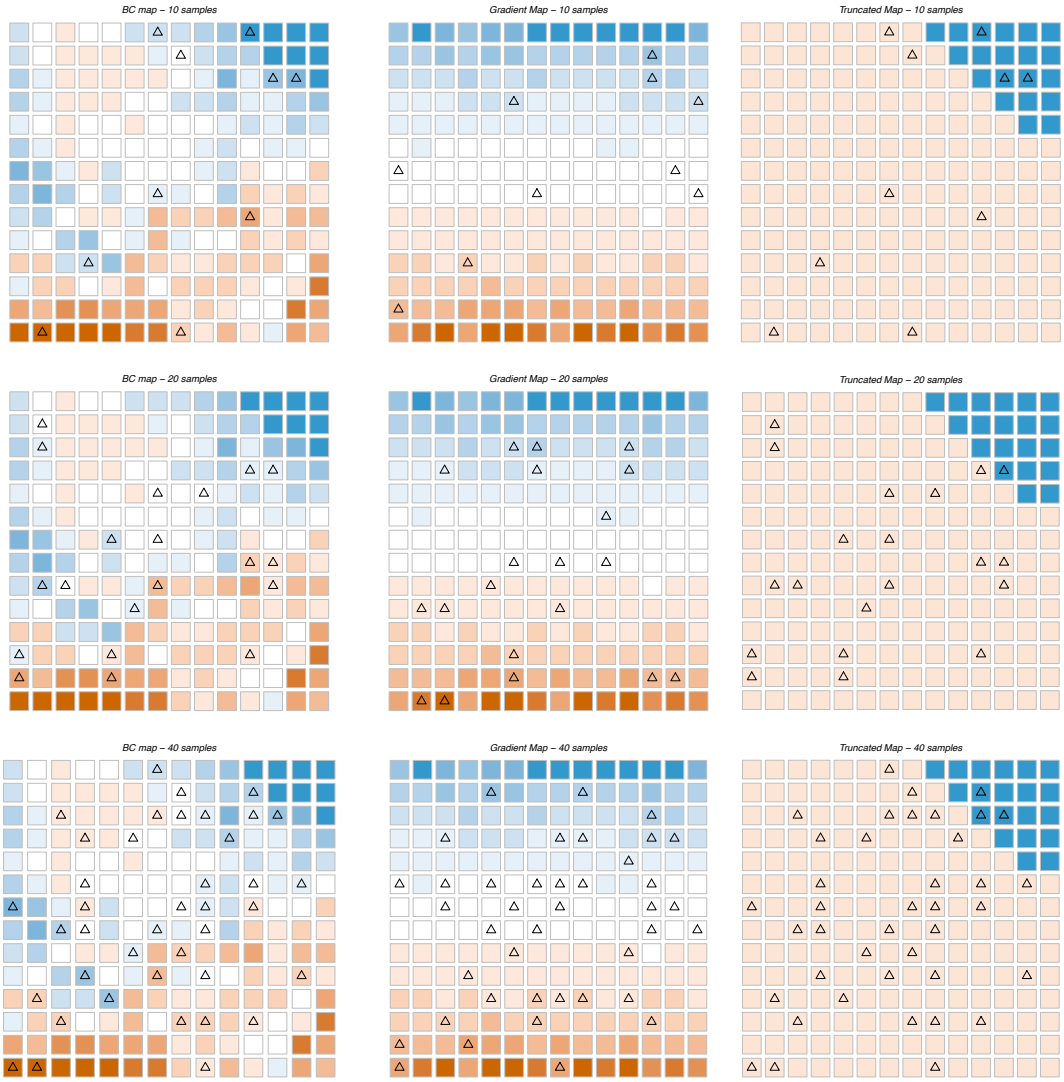


Figure S3 Locations of sampled demes on the maps of environmental variation we assumed in the simulations. Triangles indicate the locations where individuals were sampled in each case. Colors represent the optimal phenotype in each population, using the same color scheme as Figure 1 in the main text.

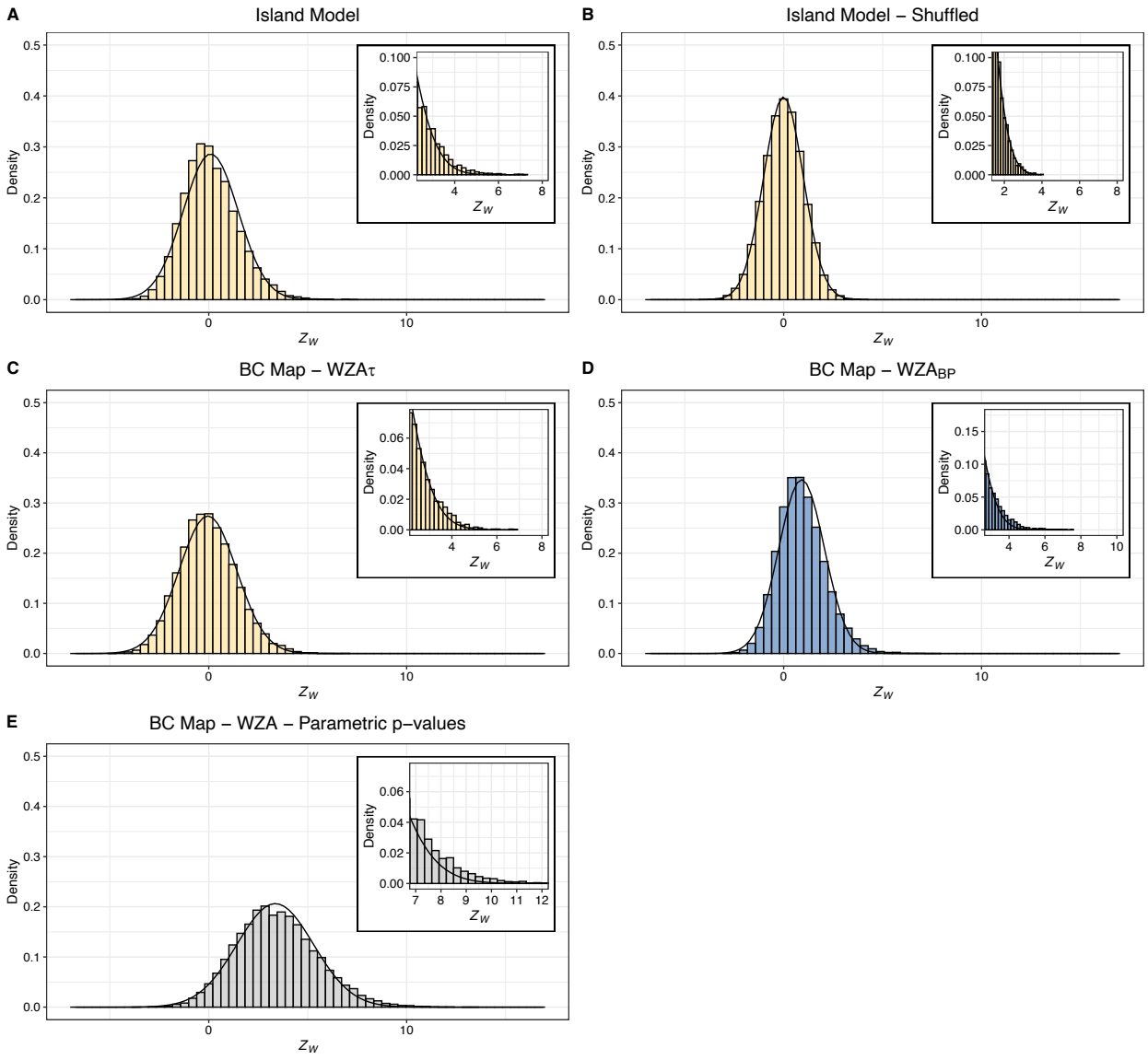


Figure S4 The distribution of WZA scores from neutral simulations with details of the right tail in the insets. Overlaid on each panel is the normal distribution fitted to each dataset. In all cases, results from 20 simulation replicates are plotted together. Results shown were obtained from simulations assuming a constant mutation rate.

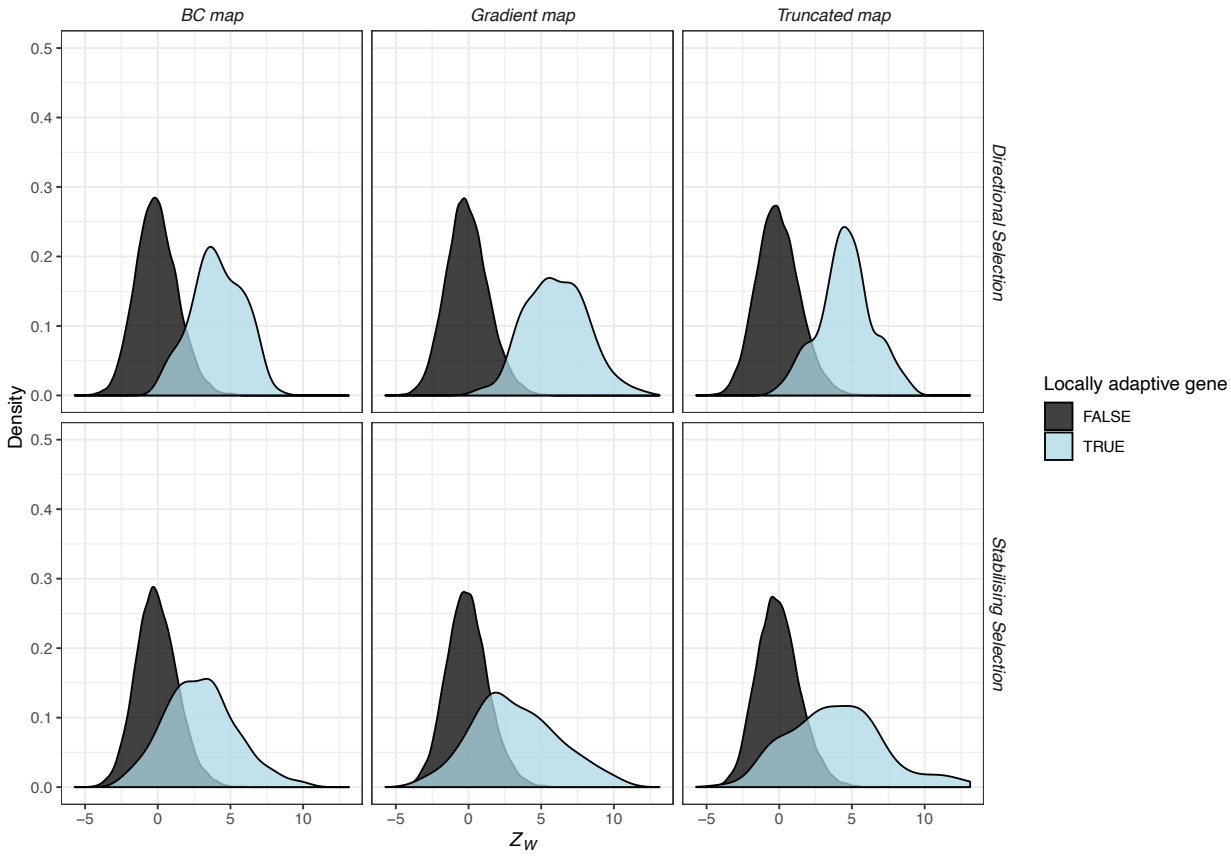


Figure S5 The distribution of WZA scores from simulations of local adaptation. Note, the plot does not indicate the relative frequency of genes that are or are not locally adaptive. Results shown are for samples of 40 demes with 20 individuals sampled in each. In all cases, results from 30 simulation replicates are plotted together. Results shown were obtained from simulations assuming a constant mutation rate.

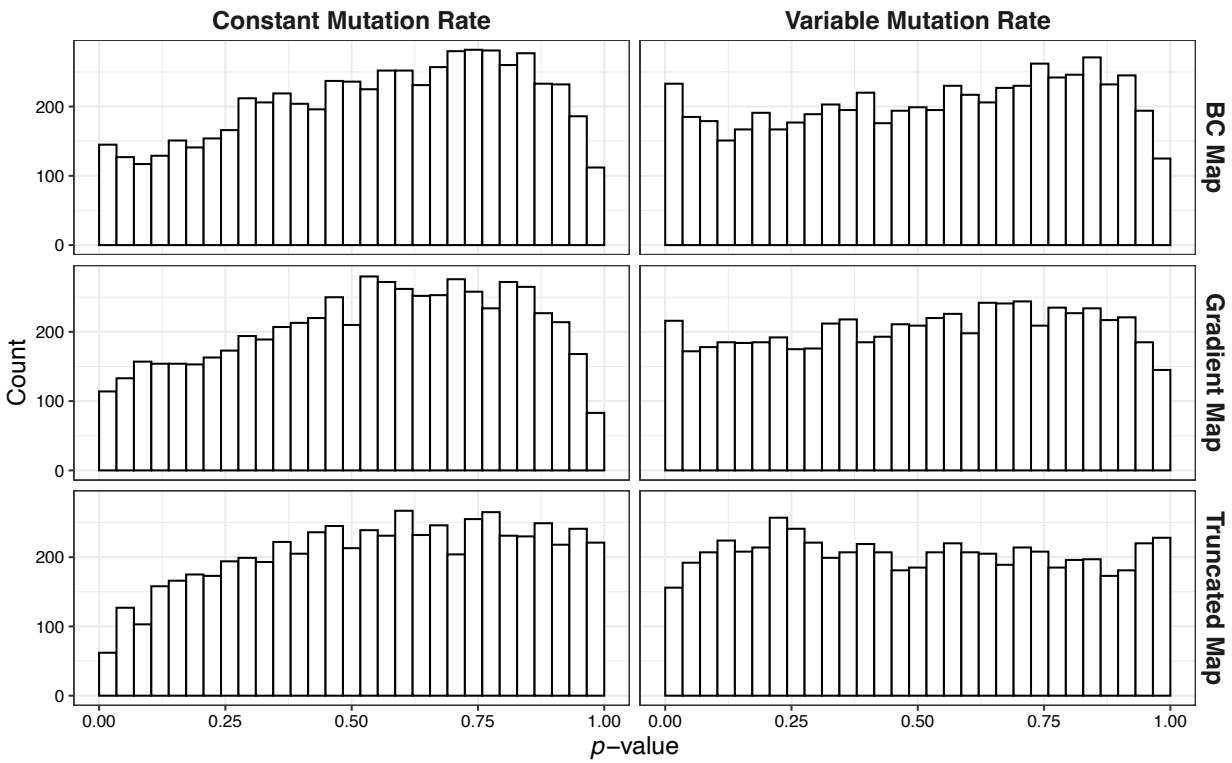


Figure S6 The distribution of parametric *p*-values obtained from the WZA for simulations modelling local adaptation. In each case, results shown are for genes that are evolving under strict neutrality, but may be affected by the effects of selection genome-wide.

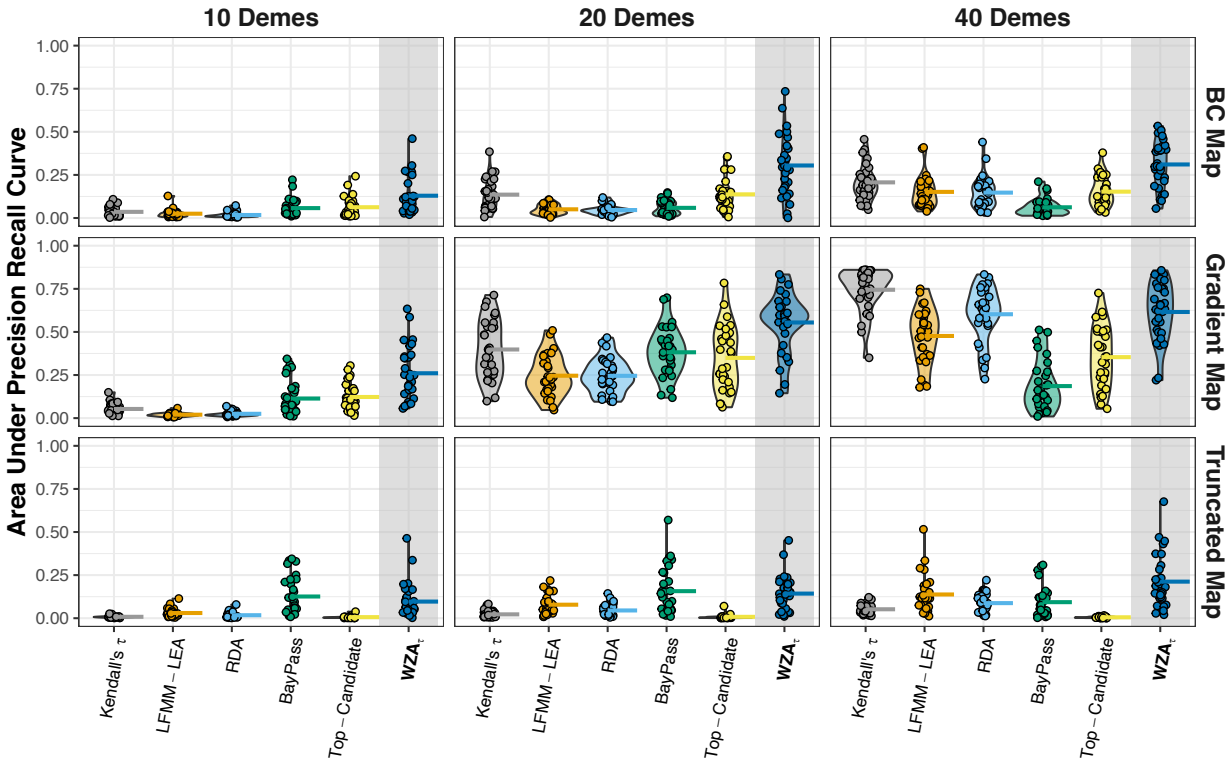


Figure S7 The relative performance of various GEA methods as evaluated using AUC-PR when locally adaptive genes have a weak effect on fitness. The violins show the relative density of points and the horizontal lines indicate the mean of 30 simulation replicates. The grey box simply highlights the performance of the WZA as compared to the other methods.

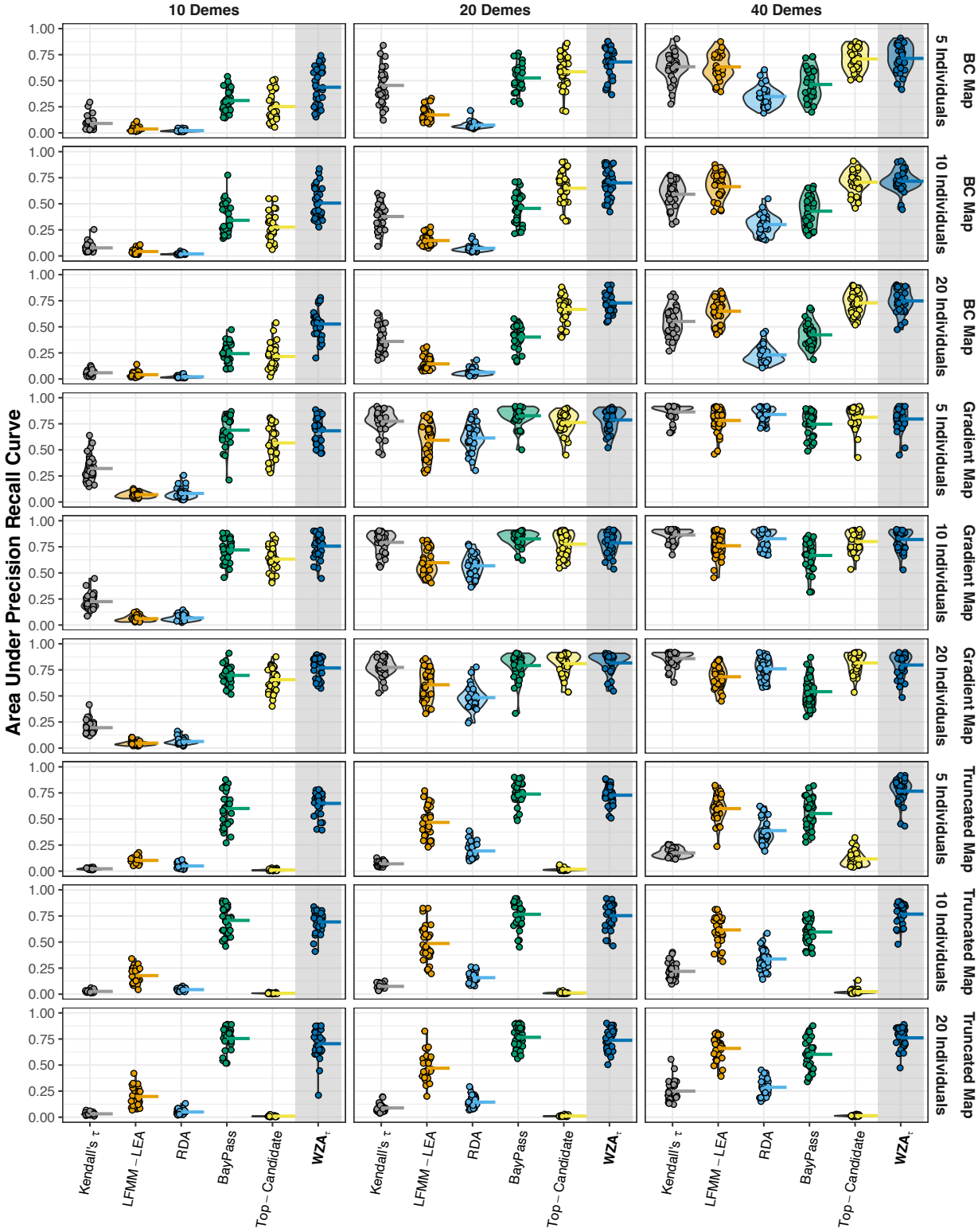


Figure S8 The relative performance of various GEA methods as evaluated using AUC-PR with varying numbers of individuals sampled per deme. The violins show the relative density of points, and the horizontal lines indicate the mean of 30 simulation replicates.

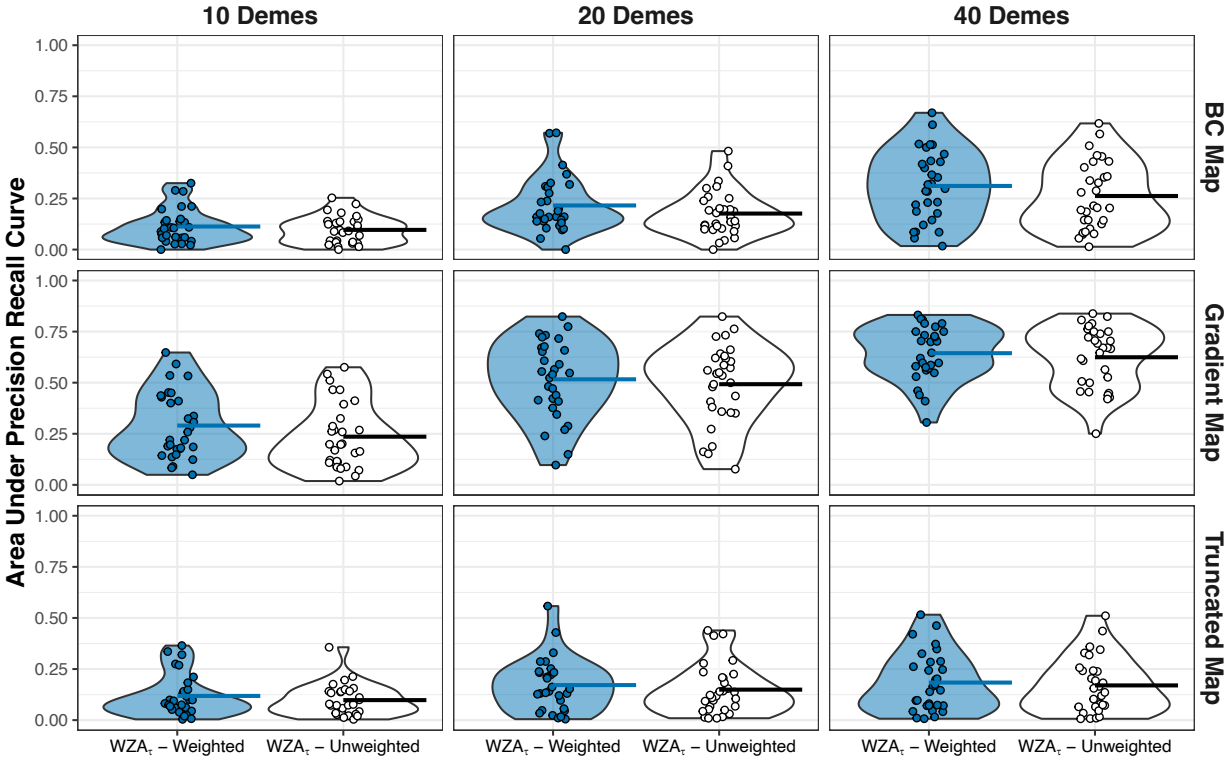


Figure S9 Comparison of the WZA using $\overline{p}\overline{q}$ as weights in the Equation 1 (WZAτ) and an unweighted version of the WZA (WZAτ - Unweighted). In each case, the results were obtained using a sample of 50 individuals sampled from each of 40 demes. Lines represent the means of 20 replicates. See the caption of Figure 3 for a description of the x-axis.

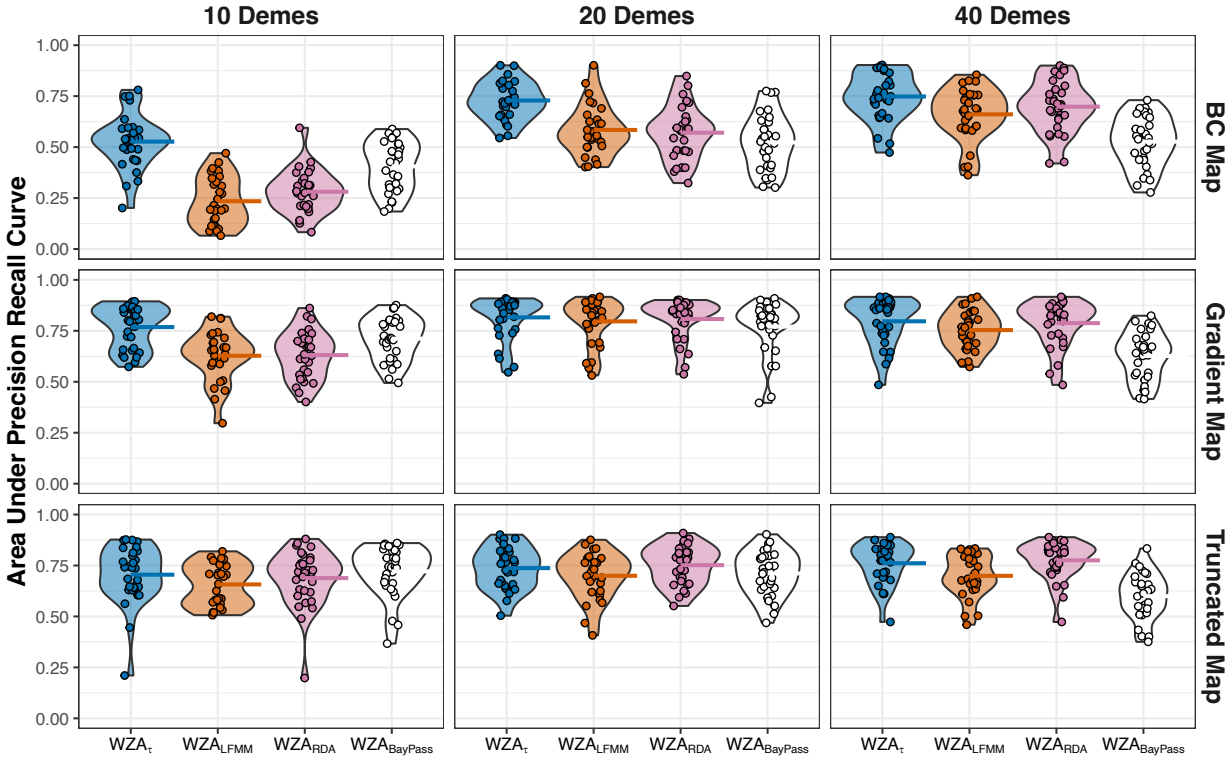


Figure S10 Comparison of performance of the WZA when using various GEA summary statistics as input.

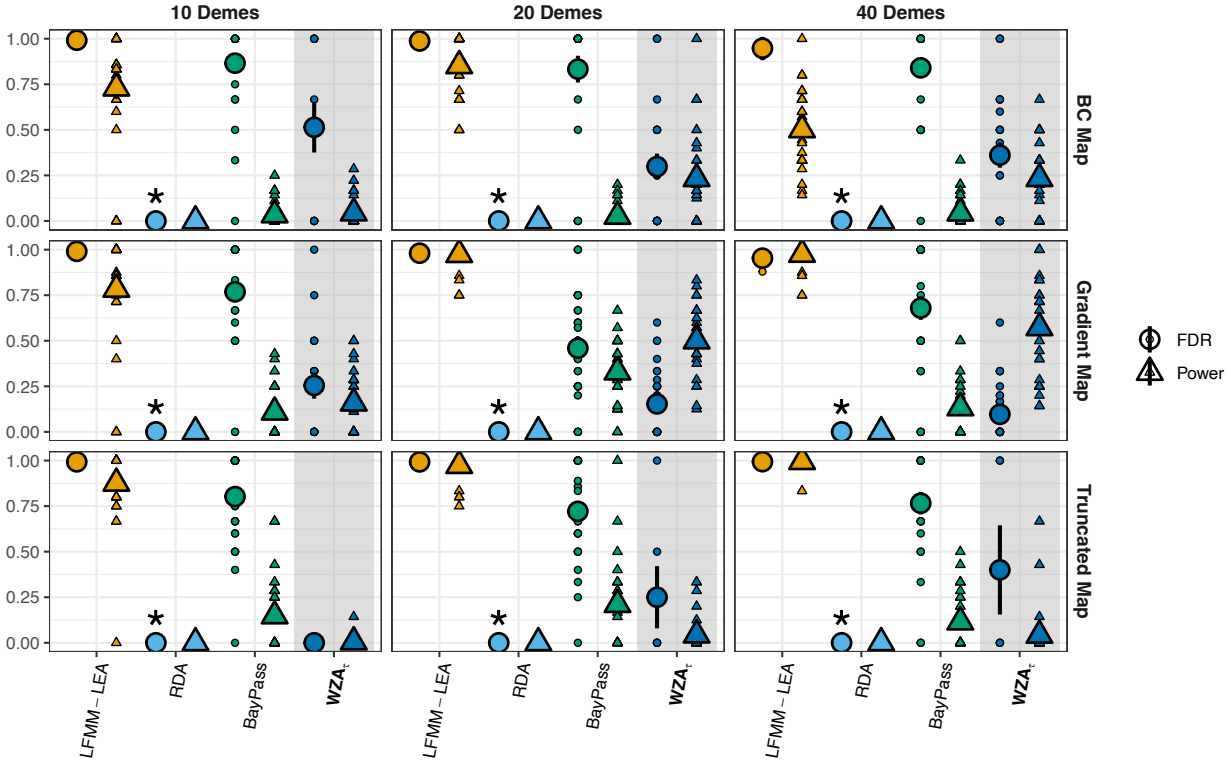


Figure S11 The power and false discovery rate of various GEA methods after applying a stringent genome-wide significance threshold. Small points indicate values for 30 individual simulation replicates, while large shapes indicate the means of the respective statistics. Simulation results shown were obtained by assuming weak selection on locally adaptive alleles and variation in the mutation rate.

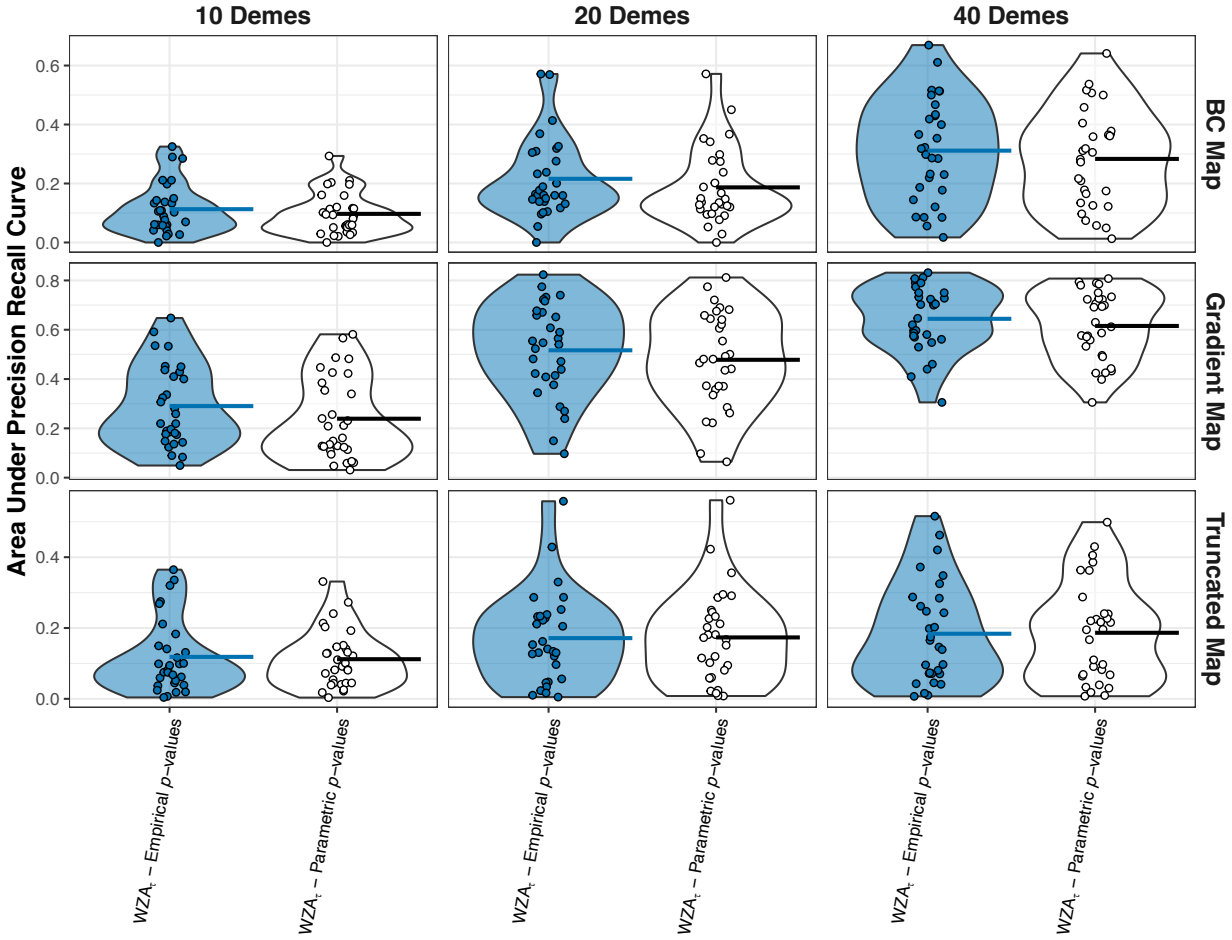
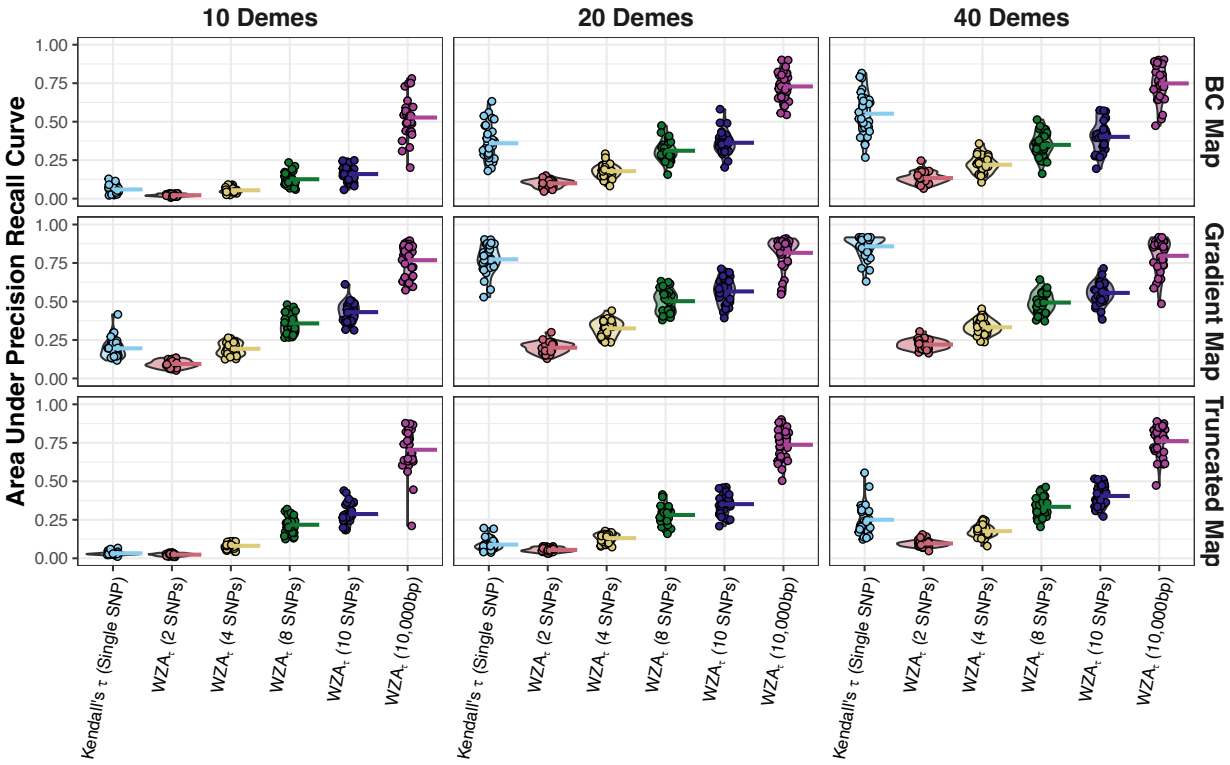


Figure S12 Comparison of the WZA performed using empirical *p*-values (WZAτ) or using parametric *p*-values from Kendall's τ (WZAτ – Parametric *p*-values). Results were obtained assuming weak selection on the alleles that contribute to local adaptation.



1066

1067 **Figure S13** Comparing the performance of the WZA genes identified using the WZA,
1068 using analysis windows analyzing a fixed number of SNPs. Lines represent the means
1069 of 20 replicates. Analysis was performed on results for a sample of 40 demes with 50
1070 individuals taken in each location. For a description of the axes in this plot see the
1071 legend to Figure 3 in the main text.

1072

1073