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

Tom R. Booker[1,2,3,*], Sam Yeaman[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

Commented [TB1]: This paper is bound to be an instant classic. For that reason, I took out the "new" from the title.

Abstract

Genotype environment association (GEA) studies have the potential to elucidate 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 either using a GEA method that controls for population structure or an uncorrected approach. In the majority of cases we tested, the WZA either outperformed single-SNP based approaches or performed similarly. The WZA outperformed individual SNP approaches when the measured environment is not perfectly correlated with the true selection pressure or when a small number of individuals or demes was sampled. We apply the WZA to previously published data from lodgepole pine identified candidate loci that were not found in the original study.

KEYWORDS: Local adaptation, population genetics, landscape genomics

Commented [SY2]: The tough part here is we don't know the truth, and it's not surprising that we find a different set – I would expect that any method would reveal different results to a previous method.

I don't want to get into mentioning the top candidate method in the abstract here, but perhaps just mention that we illustrate the method on empirical data?

Commented [TB3R2]: You're right of course. I think it's probably a good call to leave the TC out.

Introduction

Studying local adaptation can provide a window into the process of evolution, yielding insights about the nature of evolvability, constraints to diversification, and the how the interplay between a species and its environment shapes its genome (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 fitness 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.

Numerous approaches for performing GEA analyses have been proposed. If individuals are sequenced, GEA can be performed by regressing environments on genotypes as a form of genome-wide association study, for example using the *GEMMA* package (Zhou, Carbonetto, and Stephens 2013). However, to estimate SNP effects with reasonable statistical power, many individuals may need to be sequenced. A cost-effective alternative is pooled sequencing (hereafter pooled-seq), where allele frequencies for populations of individuals are estimated rather than individual genotypes (Schlötterer et al. 2014). In this study, we focus on analyses that can be performed on pooled-seq datasets given the wide adoption of that protocol in the GEA literature.

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 it had restricted migration and exhibited isolation-by-distance, neutral alleles may be

Commented [MOU4]: To make this flow from the previous paragraph, I'd switch the order of this – one way to detect local adaptation is to search for regions of the genome with strong signature of association with environment.

Alternatively, a connecting sentence at the end of the first paragraph might help

Commented [TB5R4]: I opted to add a topic sentence at the beginning to achieve the flow

Commented [MCW6R4]: Good.

Commented [MOU7]: Given how commonly FST outliers are used, I feel like they could be mentioned somewhere around here to mention that other ways are also used, but here we focus on correlation

Commented [TB8R7]: I can't find a way of adding that to this sentence without disturbing the flow.

Commented [MCW9R7]: I don't think we need this.

Commented [SY10]: No need to worry about this here, but just as a writing tip when you eventually write a paper for a glossy journal where space is a real constraint – this could probably be condensed into a single sentence.

Commented [TB11R10]: Fingers crossed that the convergence analysis leads to something that might make me put this advice into practice!

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 $\mathbf{\Omega}$ in a manner similar to a random effect in a generalized mixed model. In a recent study, Lotterhos (2019) compared several the most commonly used packages for performing GEA on pooled-sequencing datasets; including *BayPass* (Gautier 2015), latent-factor mixed models (LFMMs) as implemented in the LEA package (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 and RDA) had higher power to detect local adaptation compared to *BayPass* or LFMMs. In their standard application to genome-wide datasets, all of the GEA analysis methods provide a summary statistic for each marker or SNP.

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 environmental variation. In the extreme case of a non-recombining region, all SNPs present would share the same genealogy and thus provide multiple tests of the same hypothesis. The SNPs that are the most informative in this context are those with the highest heterozygosities as they contain the most information about the shape of the underlying genealogy. For recombining portions of the genome, however, linked sites will not have exactly 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 increase the power of GEA studies to identify genomic regions that contribute to local adaptation.

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). We test the efficacy of WZA using simulations. We generate datasets modelling a pooled-sequencing experiment where estimates of allele frequency are obtained for numerous

Commented [MOU12]: I try to use heterogeneity in this context as a descriptor of how much variation there is in the environment, in which case a correlation with heterogeneity would mean one allele tended to be found in homogeneous regions vs. another in heterogeneous regions

Commented [TB13R12]: Sounds like a good plan. I'll adopt that throughout

Commented [MCW14]: "exact same" is one of those phrases that copyeditors hate, because it is redundant. "same" means "identical"

Commented [MOU15]: Not sure how to reword – SNPs don't have frequencies, alleles do, but it's the SNP that is most informative (individual alleles are not informative on their own).

Commented [TB16R15]: I disagree on this one. In the terminology I tend to use a SNP does have a frequency. A SNP is not the presence of a polymorphism at a particular site, but a specific deviation from the reference genome. If we say allele frequency without invoking SNPs, we could be talking about multiple different alleles at a locus like microsats or something. I think talking in terms of SNPs makes it more concrete.

I changed it, but I am grumbling.

Commented [MCW17R15]: Sorry, I disagree, Tom. A SNp is a polymorphism (it's in the name). It's a polymorphism regardless of whether there is a reference genome or not.

Commented [TB18R15]: Hurumph

Commented [MCW19]: Those with the highest allele frequencies are actually LEAST informative.

Commented [TB20R19]: Of course, I'm just getting tangled switching between the two.

populations across a species' range. Using our simulated data, we compare the performance of WZA to Kendall's τ as well as *BayPass* (Gautier 2015), as it is a widely used approach that corrects for population structure in GEA studies. Additionally, we compare WZA to another window-based GEA approach that was 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 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. 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). In the WZA, we use $\overline{p}q$, a marker's expected heterozygosity, to determine these weights. 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 lower values of $\overline{p}q$ will have a greater relative error in estimates of local allele frequency than will sites with higher $\overline{p}q$, causing greater relative error and bias in estimates of the correlation between allele frequency and an environmental variable. In order to capture this effect approximately we use the same weights as used by Weir and Cockerham (1984) (i.e., $\overline{p}q$) to combine estimates of F_{ST} across sites.

We combine information 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. We calculate $Z_{W,k}$ for genomic region *k*, which contains *n* SNPs as

$$Z_{W,k} = \frac{\sum_{i=1}^{n} p_i\, q_i z_i}{\sqrt{\sum_{i=1}^{n} (p_i\, q_i)^2}}, \quad (1)$$

where 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 using the *qnorm* function in the R programming language, for example.

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 about 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]). To calculate empirical *p*-values, 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 analyzed). 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*-

Commented [SY21]: Revisit in light of Mike's derivation

Commented [TB22R21]: Done

Commented [SY23]: This implies that our weights are different than pq, so maybe state that these weights ARE pq

Commented [SY24]: Should we comment somewhere on unbalanced designs, where different populations have different numbers of individuals? This came up as a question from someone in my lab about how to use it, which I think made sense to them when I explained it but points out a potential source of misunderstanding. Not sure – it seems obvious to me as is!

Commented [TB25R24]: I don't think that we really want to open that can of worms. I think unbalanced design is something that would negatively influence all GEA approaches, not just the WZA.

I think that it's highly possible that a reviewer might ask for a more complete comparison of various GEA methods. I'm speaking with Katie Lotterhos next week and I'll get her opinion on that. Either way, I think that it's best to leave that can of worms closed!

Commented [MCW26]: should we need to clarify here that this is smaller for either sign of correlation?

Commented [TB27R26]: I don't think that it's necessary, but happy to add it if you think so too Sam

SY: I think it's fine as-is – would feel clunky to explain

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 the empirical *p*-values 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.

When we apply the WZA in this study, we used two different statistics as input: empirical *p*-values calculated from the genome-wide distribution of parametric *p*-values from Kendall's τ correlating the local environmental variable and local allele frequency (referred to as WZAτ), and empirical *p*-values calculated from the genome-wide distribution of Bayes factors as obtained using the *BayPass* program (referred to as WZA_{BP}; see below). Note that Lotterhos (2019) identified Spearman's ρ as having among the highest power of the GEA analyses that they had tested. We used Kendall's τ as it calculates accurate *p*-values in the presence of tied datapoints.

Commented [MOU28]: Do we need to mention that this may also be better if the test used for the p-values conducts a structure correction already?

Commented [TB29R28]: I don't know. We did use Bayes Factors which are corrected for structure (i.e. BayPass) so I'm not sure how saying that in a way that doesn't just say BayPass is not up to snuff

Commented [TB30R28]: I also now mention this a bit more in the Discussion

Commented [MCW31R28]: I think leave it in the simpler sate here withouthtis

Commented [SY32R28]: I think it's fine as-is

Commented [SY33]: Perhaps move the sentence "A feature of the WZA…" here

Commented [TB34R33]: Good idea. Done

Commented [SY35]: I replaced "statistics" here with tests, because it's actually less flexible than the top candidate test in this regard – WZA only works with p-values. Top candidate test can be done on FST, p-values, whatever.

I think it might make sense to move this to after the section on empirical p-values, because once you convert a given test statistic to an empirical p-value, this test can be applied to anything – but it's not clear yet that this is the case here.

Commented [TB36R35]: Done

Commented [MCW37]: should we need to clarify here that this is smaller for either sign of correlation?

Commented [TB38R37]: I don't think that it's necessary, but happy to add it if you think so too Sam

SY: I think it's fine as-is – would feel clunky to explain

Materials and Methods

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 so as to model a large population. We based our choice of population genetic parameters on estimates for conifer 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 simulated. All simulations were performed in *SLiM* v3.4 (Messer and Haller 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

$$m = \frac{\frac{1}{F_S T} - 1}{4N_d 196}$$

(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 uniformly distributed onto 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 = 48.37, 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 randomised 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, 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 models of either directional or stabilizing selection. In both cases, there were 12 causal genes distributed evenly across four simulated chromosomes that potentially contributed to local adaptation. With directional selection, mutations affecting fitness could only occur at a single nucleotide position in the center of the 12 potentially selected genes. Directionally selected mutations had a spatially antagonistic effect on fitness. In deme *d* with phenotypic optimum θ_d, the fitness of a selected allele was calculated as $1 + s_a\theta_d$ for an individual homozygous for a locally beneficial allele (selected alleles were semi-dominant). The fitness affecting alleles had a mutation rate of 3×10^{-7} in simulations modelling directional selection and a fixed s_a = 0.003 (see *Appendix*).

Under stabilizing selection, the mutations that occurred in the 12 genes had a normal distribution of phenotypic effects, with variance $\sigma_a^2 = 0.5$. Phenotype-affecting mutations occurred at a rate of 10^{-10} per base-pair in the 12 genes, and could occur at any of the 10,000 sites within a given gene. An individual's phenotype was calculated as the sum

Commented [SY39]: Seemed a bit clunky to say that a base-pair has a recombination rate, as recombination can only happen between bases

Commented [TB40R39]: That's a pretty common way of defining recombination rates, I think - I take your point though. Your edit is a lot cleaner

Commented [SY41]: Could be deleted to save space

Commented [TB42R41]: Done

Commented [MCW43]: Figure 1 seems to be missing in this file. I remain confused about why the grid is not made up of squares in the BC map. The grid seems like a series of tall thin rectangles. Also, the full map of BC seems squished, but in the different direction. Is the graph wrong or is there something deeper that needs to be explained?

Commented [TB44R43]: This is to do with our damn spherical earth and map projection stuff. At the equator it would be a square.

You've probably done high school geometry more recently than me helping Wilson, but here's what I remember:

1 degree of latitude (in km) = ~69 miles.

1 degree of longitude (in km) this far North (48th parallel) is approximately:

Miles * cosine(longitude in degrees) = 69 * Cos(48) = 45. So an area of 1 degree longitude x 1 degree latitude area up here is a 45x69mile rectangle. I used km to fit in with the Canadians.

I don't think there's anything deep that needs to be explained – we are not actually simulating real space. I used the DD0 map as a way to get something like a realistic distribution of spatial variation in an environmental variable. It's a bit funky to use the rectangular map, but I found that it maximized the variation in DD0 across BC

Commented [MCW45R43]: Sounds good. Calling it a rectangular grid helps.

of the effects of all phenotype-affecting mutations. We calculated an individual's fitness using the standard expression for Gaussian stabilizing selection,

$$W_{z_{i,j}} = exp\left[\frac{-(f_{i,j}-\theta_d)^2}{2V_s}\right],$$

where f_i is the phenotype of the i^{th} individual in environment j and V_s is the variance of the Gaussian fitness function (Walsh and Lynch 2018). We set $V_s = 196$ so that there was a 40% fitness difference between individuals perfectly adapted to the two extremes of the distribution of phenotypic optima. This was motivated by empirical studies of local adaptation that have demonstrated such fitness differences in numerous species (Hereford 2009; Bontrager et al. 2020); see *Appendix*.

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 (when modelling directional selection), or with a phenotypic optimum of 0 for all demes (when modelling stabilizing selection). 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. At the end of each simulations, neutral mutations were added at a rate of 10^{-8} using *PySLiM* (*https://pyslim.readthedocs.io/en/latest/*). For each combination of map and mode of selection, we performed 20 replicate simulations.

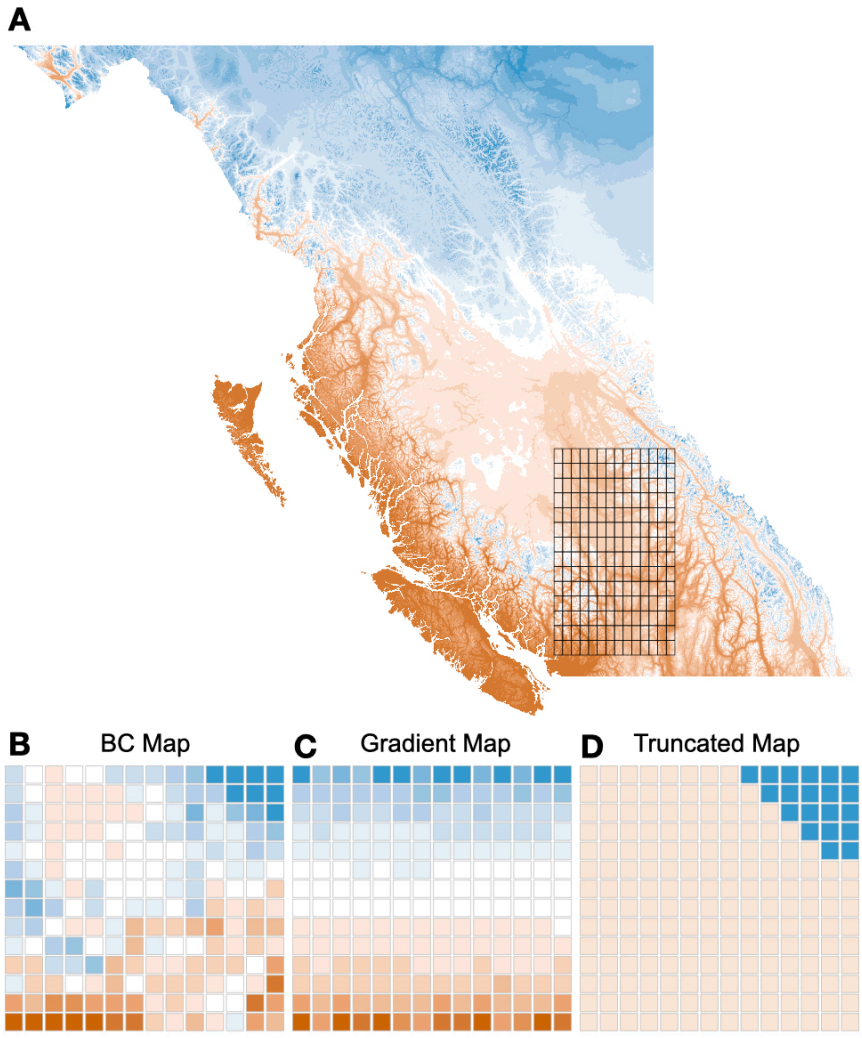


Figure 1 A) Degree days below zero across British Columbia, the overlain grid in A shows the locations we used to construct phenotypes for our simulated populations. B) A discretized map of DD0 in Southern British Columbia, we refer to the map in B as the *BC map*. C) A 1-dimensional gradient of phenotypic 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 genes contribute to local adaptation and which do not in our simulated data. While there were only 12 genes that were allowed to affect fitness in our simulations, not all of those need be used by the evolving population to result in local adaptation. As described above, our simulations incorporated a stochastic mutation model so from replicate to replicate the genes that contributed to local adaptation varied and, in the case of stabilizing selection, so did the effect size of the alleles in those genes.

For simulations modelling directional selection, we identified locally adapted genes based on the mean fitness of their alleles. For a given gene containing directionally selected alleles, our measure of local adaptation was the covariance between the mean fitness contributed by the selected allele in each population and the environment.

For simulations modelling stabilizing selection, we identified locally adapted genes based on the covariance of the environment and the phenotypic effects of their alleles, summed across all variant sites within each gene. For a given gene, we summed the additive phenotypic effects of all non-neutral variants and took the average for each population. Our measure of local adaptation for each gene was the covariance between that average additive phenotypic effect and environmental variation (we refer to this as *Cov(Phen, Env)*).

For both selection regimes, we defined locally adapted genes as those with a covariance between environment and allelic effect (in fitness or phenotypic terms) greater than 0.005. When assuming 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. In our simulations assuming stabilizing selection, individuals' and population mean phenotypes closely matched the phenotypic optima of their local environment (Figure S2). The average numbers of genes contributing to local adaptation in individual replicates in these simulations were 7.15, 6.45 and 5.35 for the *BC map*, the Gradient map and the Truncated map, respectively. However, when analyzing stabilizing selection simulations, we calculated the proportion of the total *Cov(Phen, env)* explained by a particular set of genes rather the number of true positives.

Analysis of simulation data

We performed GEA on our simulated data using either Kendall's τ-b (hereafter Kendall's τ), a rank correlation that does not model population structure, or *BayPass*, which corrects for a population covariance matrix (Gautier 2015). For all analyses, except where specified, we analyzed data for a set of 40 randomly selected demes and sampled 50 individuals from each to estimate allele frequencies. We sampled individuals from the same set of demes for all analyses, 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

Commented [TB46]: Based on the number of comments from Sam, I'd say that this section was (and maybe still is) pretty confusing.

I redrafted those paragraphs

Commented [SY47]: Adaptive implies potential, adapted means this potential is currently realized

Commented [TB48R47]: Ah, that clarifies your edits. I've been operating with Adaptive == Adapted.

Commented [SY49]: I feel like I'm not representing what you're meaning with the edits here, but this could be further clarified.

Commented [TB50R49]: Yep, things got a little turned around so I redrafted.

Commented [SY51]: Not sure – is this what you did? Wasn't clear how multiple variant sites were being treated

Commented [TB52R51]: Yeah, I've also struggled to write this clearly. That's why I leaned on the matrix definition. In the stabilizing selection sims there were potentially many variants in a gene. So, I calculated the average phenotypic effect of the variants present in a gene.

Mike had a snappy way of saying this that I've tried to capture in my edits. .

Commented [MCW53]: Why this value

Commented [TB54R53]: This was based on the distribution of cov(fitness,env) that I saw in the data

BayPass following the "worked example" in section 5.1.2 of the manual provided with the software.

We used three different methods to summarize the GEA results for each gene in a given simulation replicate: a single SNP-based approach, the WZA and the top-candidate method developed by Yeaman et al. (2016). For all three tests, we used either the *p*-values from Kendall's τ or *BayPass.*

- For the implementation of the single SNP-based approach, the SNPs with the most extreme test statistic (i.e. smallest *p*-value or largest Bayes factor) for each gene were 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). 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 method asks whether there is a significant excess of "outlier" SNPs in a region compared to what one would expect given the genome wide distribution. The number of outliers in a given genomic region is compared to the expected number of outliers based on the genome-wide proportion of SNPs that are outliers, using a binomial test. The *p*-value from the binomial test is used as a continuous index.
- For the implementation of the WZA, we converted the *p*-values (from Kendall's τ) or Bayes factors (from *BayPass*), into empirical *p*-values. For each of the n SNPs present in a gene, empirical *p*-values were converted into z scores and used to calculate WZA scores using Equation 1.

We examined effect of variation in recombination 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.

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 adjacent sets of *g* SNPs and retained the maximum WZA score for all sets of SNPs in the gene.

Tree sequences were manipulated using the *tskit* package. Mutations were added to trees using the *msprime* (REF), *tskit* and *PySLiM* workflow (version). F_{ST} and r^2 (an estimator of linkage disequilibrium) were calculated using custom Python scripts that

Commented [SY55]: I'm fine with deleting this sentence

Commented [TB56R55]: I like this sentence as it tells the reader that that's how we (perhaps only I) view single-SNP analyses in empirical studies

Commented [TB57]: This is not quite how I think of this. I don't view the SNP-based approach as a windowed statistic – thought maybe I ought to. The way I view it, it's like saying what are the top 5% of SNPs doing, after accounting for linkage among hits? Happy to be disagreed with though.

Plus, I don't really know of studies that would be appropriate here. Please add some if you know of some.

Commented [TB58R57]: Note that the thing that I was flagging has been deleted so this comment is just here for historical purposes.

Commented [MOU59]: No need to go over the differences in implementation, I'd just rephrase this and say we used a simplified version and discuss how you used it – not necessary to go into the byzantine way I first used it

invoked the *scikit-allel* package (REF).

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 (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 *github/TBooker/GEA/WZA*. Analyses were performed using a combination of R and Python. All plots were made using *ggplot2* (REF). Tree-sequence files for the simulated populations are available at Dryad and all processed GEA files are available on (SomeCoolLocation).

Commented [MCW60]: Why Spearman's instead of Kendall's like the rest of the paper?

Commented [TB61R60]: Because that's the data that is available in the Dryad repo.

Commented [MCW62]: Should we make a R package for WZA?

Commented [TB63R62]: Maybe a vignette? That way we can walk the user through the steps like Forester did for their RDA vignette.

Commented [MOU64]: Dryad also?

Commented [TB65R64]: I guess so

Results

The statistical properties of the WZA

To assess the statistical properties of the WZA, we first performed GEA analyses on populations that were evolving neutrally. Figure 2A shows the distribution of WZAτ scores for stepping-stone populations simulated under the *BC Map*. The null expectation for WZA scores 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 thicker right-hand tail than expected under the normal distribution.

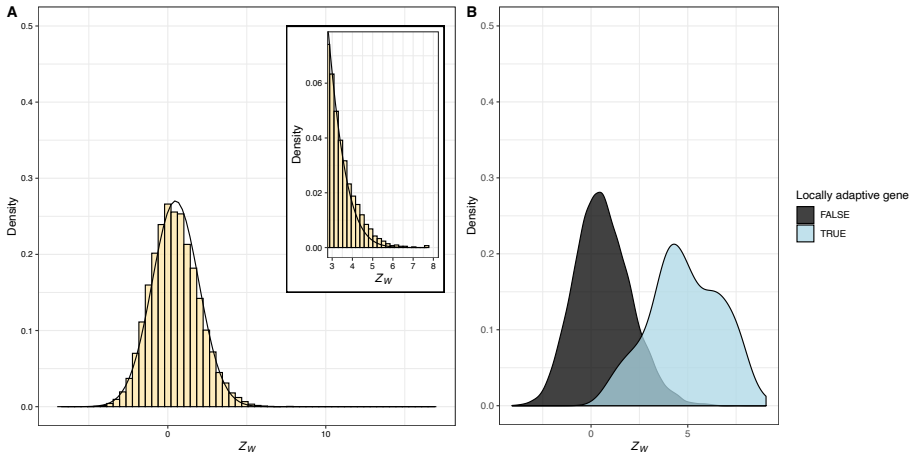


Figure 2. The distribution of WZA scores under neutrality and a model of local adaptation. A) A histogram of WZAτ scores under strict neutrality across a set of 20 replicate simulations, inset is a close-up view of the upper tail of the distribution of Z_W scores. B) A density plot showing the separation 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 WZA scores. To demonstrate this, we re-calculated WZA scores, 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 20 simulation replicates had mean a

Commented [SY66]: Is it "the WZA" or just "WZA" . I prefer "the WZA" as the article would be naturally used if you said each word in the acronym. But I'm fine with out the "the" if you prefer

I think it's better not to talk about efficacy here – that is really discussed below so I suggested an alternate heading

Commented [TB67R66]: Yeah, we need to decide on the whole "WZA" or "the WZA" thing. I'd probably agree with you Sam that "the WZA" is a bit more grammatical. I tend to sound out what I'm writing in my head as I'm writing so I've found myself often putting "the WZA" in things.

As an example, I think it is consistent to say both o the following things:
"We applied the WZA to our data"
And
"The distribution of WZA scores was …"

If you say ""The distribution of *the* WZA scores were …" it sounds odd.

Commented [SY68]: Seemed premature to explain this here. It's explained more clearly below

Commented [TB69R68]: I'm confused? This paragraph is saying that the distributions are non-normal and the next paragraph says why. I think this fits here.

Commented [MOU70]: Does the bc-map uncorrected really have such low Z-scores? (panel B) Something looks fishy here.

Commented [TB71R70]: This is now A, but yes. This is for neutrally simulations, so there are no selected ones here. The expectation is that the mean would be 0 with sd = 1 – it's a little off of that, but not by much as I mention in the text. So I'd expect the z-scores to be fairly low

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.

When evolution includes selection, WZA can often clearly distinguish regions of the genome containing loci that contribute to local adaptation from those that do not. Figure 2B shows clear 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 distributions of WZAτ scores for locally adapted genes when modelling stabilizing selection was broader than when modelling directional selection (Figure S5), consistent with differences in the distributions of effect size for the genes involved in local adaptation under the two selection models (Figure S6). 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.

Comparison of the WZA with other GEA approaches

We compared WZA to two other methods for identifying genomic regions that contribute to local adaptation from GEA data (Figure 3). To assess the performance of the different methods, we examined the top 1, 2, 3,... 50 genes in terms of WZAτ scores, $-log_{10}(p$-values) from the top-candidate method, or the single SNP Kendall's τ approach and calculated the proportion of all true positives that were identified in each case. In our simulations, there were 1,000 genes in total with around 6 locally adapted genes in each replicate (see Methods). For visualization purposes, we include Figure S7, which shows the $-log_{10}(p$-values) from Kendall's τ represented as a Manhattan plot for individual simulation replicates, WZAτ and top-candidate scores calculated from those data and the proportion of true positives detected using the three different analysis methods.

Commented [MCW72]: Seems more like a topic for the Discussion

Commented [TB73R72]: I don't know if I'd want to open this can of worms in the discussion …

Commented [SY74R72]: Seeing as we're not advocating any tests assuming a standard normal distribution of Z's, it seems not worth going into in detail – the paper is long already!

Commented [MOU75]: This is the good stuff…should be right at the beginning of this section. Caveats and supp mat refs could be moved afterwards.

Figure 3 compares the performance of the GEA methods across the three different maps of environmental variation that we simulated. Empirical GEA studies may vary substantially in terms of how many populations or demes are sampled. For each of the three maps we simulated, we analyzed samples of 10, 20 or 40 demes where allele frequencies were estimated from 50 individuals sampled in each location; Figure SF shows the specific demes we sampled in each case. Figure 3 shows that WZAτ substantially outperformed both the top-candidate and single SNP-based Kendall's τ analyses in many cases. When analyzing simulations that used the *BC* map or the *Truncated* map, WZAτ always outperformed the top-candidate and SNP-based methods, but particularly so when fewer demes were sampled (Figure 3). When simulations assumed the *Gradient* map, WZAτ outperformed the other GEA methods when the sample was restricted to 10 demes, but with larger samples, the tests were more similar (Figure 3). This suggests that WZAτ is a powerful method for identifying regions of the genome that contribute to local adaptation in empirical analyses, but particularly so when they are performed on small samples.

An additional source of variation in GEA studies comes from the number of individuals sampled in each location. We also examined the effect that reduced sampling of individuals within each deme had on the performance of the methods. Figure S8 shows that the WZA outperforms the top-candidate and SNP-based methods when a small number of individuals is used to estimate allele frequencies. Note that this is not strictly a test of how well pooled-seq will perform with small sample sizes, however. With small numbers of individuals in sequencing pools, differential amounts of DNA from each individual may add to error in allele frequency estimation (Schlötterer et al. 2014).

Commented [SY76]: This stuff all repeats what was said in the methods and obscures the main point, which comes in the next sentence ("Figure 3 shows that…").

My preference here would be to start this paragraph off with the sentence "Figure 3 shows…" and just eliminate the sentences highlighted here. Is there anything that is said here that someone couldn't figure out pretty easily by going to the methods? I would suggest moving any such details to the method and getting right into the results

Commented [TB77R76]: As I said on the Zoom call this morning, I think that this is a style difference. I like having the lead in and justification in the Results and leaving the methods more bare. We don't detail the whole 1, 2, 3,… 50 genes thing in the methods so I think that it needs to be said somewhere and it makes sense to me for that to be here where we point to Figure S7 .

I removed some gunk from this paragraph though.

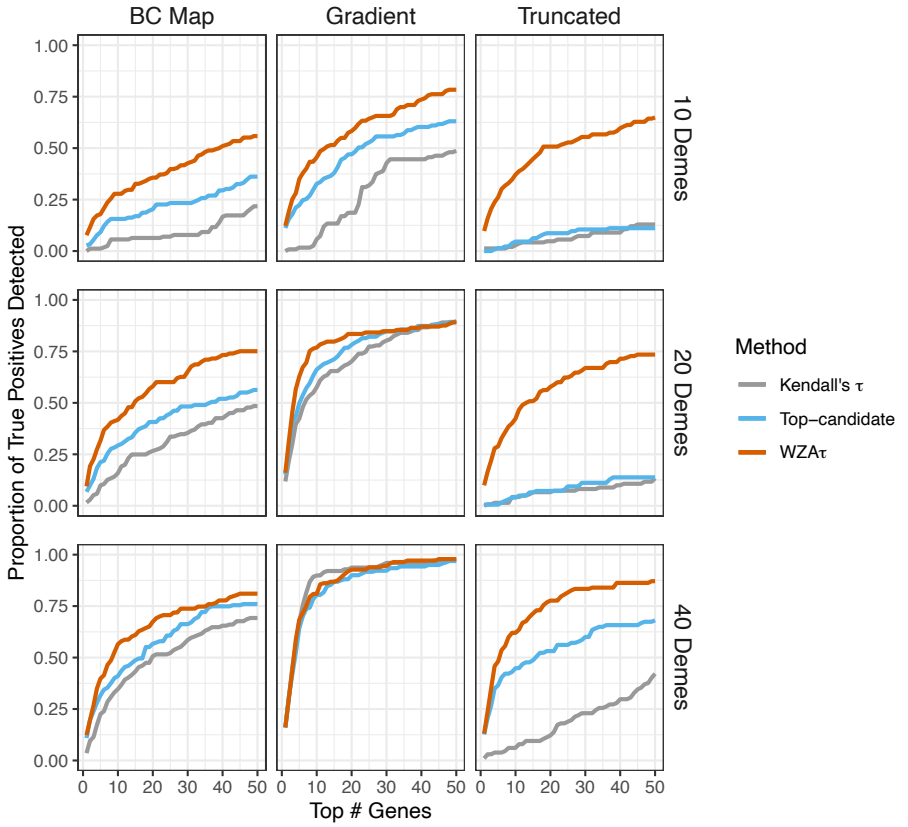


Figure 3 A comparison of three methods to identify outliers in GEA analysis conducted on simulations modelling local adaptation via directional selection. Plots show the proportion of true positives identified by examining the genes with the top ranked scores across the genome for the three GEA methods. The rows of the plot show results obtained from samples of 10, 20 or 40 demes as indicated by the labels on the right-hand side. Lines represent the means of 20 simulation replicates.

Effects of population structure correction

In each of the maps of environmental variation that we simulated, there was a strong correlation between environmental variables and gene flow. Environmental variation in each map was autocorrelated along a major axis: the diagonal axis from the bottom-left corner to the top right-corner in the case of the *BC* map (Figure 1B), the vertical axis in the case of the *Gradient* map (Figure 1C), and the top-right corner versus the rest of the

Commented [MCW78]: Change legend so order is WZAt, Top-candidae, Kendall's t, to match usual order in gaph (and to highlight WZA)

Change x-axis label to Number of top ranked genes (or something that gets rid of the #)

Commented [TB79R78]: Will do

Commented [SY80]: Great figure!

Commented [TB81R80]: Hopefully some people find it convincing!

Commented [SY82]: This section reads much more like a results section – no extensive methods recap, just getting right into it…nice!

landscape in the case of the *Truncated* map (Figure 1D). There was also a strong pattern of isolation-by-distance in our simulated populations (Figure S1). These two factors may make it difficult to identify genes involved in local adaptation in GEA studies (Meirmans 2012).

We compared the performance of the WZA to a widely adopted method for performing GEA that corrects for the confounding effects of population structure, *BayPass* (Gautier 2015). In all cases, WZA performed as well, or better than, *BayPass* (Figure 4). WZA performed much better than *BayPass* when selection was directional, but WZA was also significantly more likely to identify the genes underlying local adaptation with stabilizing selection.

Notably, even the single SNP analyses based on Kendall's τ in most cases outperformed BayPass, even though the Kendall's τ analysis did not adjust for spatial population structure. (The exception was the case with stabilizing selection on the *Truncated* map.) The discriminatory power of GEAs does not seem to be improved consistently by careful accounting of the underlying pattern of genetic structure.

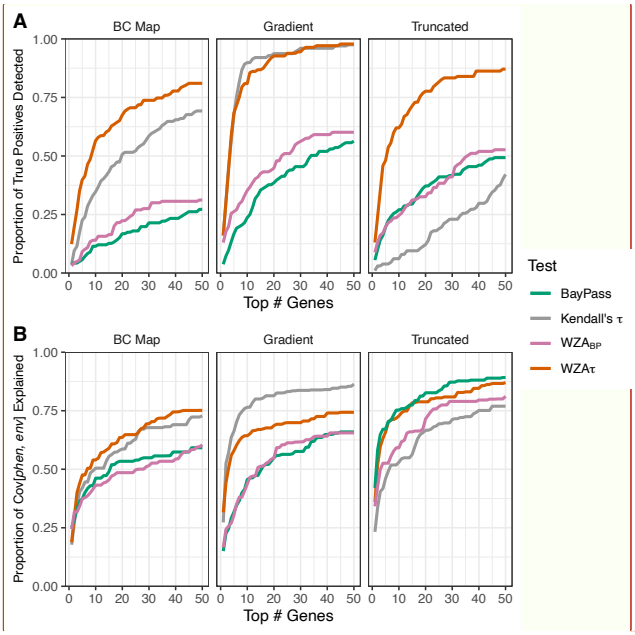


Figure 4 The performance of population structure correction. A) Results for simulations modelling directional selection and b) results for simulations modelling stabilizing selection. Lines represent the mean of 20 simulation replicates.

Commented [MCW83]: Make legend order match typical order in graphs top to bottom

Change x-axis label to match whatever used in new Fig 3.

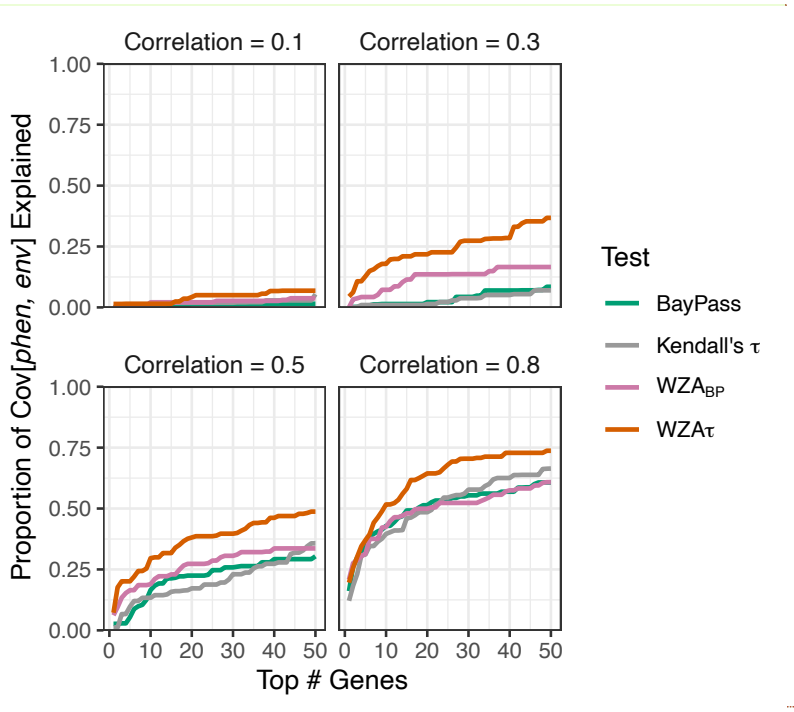
Commented [TB84R83]: Will do, but I won't get to that before I the paper to you, Sam

Commented [TB85R83]: Done

Commented [SY86]: Great figure! I really like the addition of WZA_bp...shows more clearly the effect of adding window-based stuff compared to structure correction effect.

The performance of WZA when environmental variables are weakly correlated with selection pressure

In the previous section, we conducted GEA assuming perfect knowledge of the phenotypic optima in each sampled deme. However, in empirical GEA analyses researchers will be probably always be limited to studying environmental variables that are imperfect proxies for historical selection patterns. Additionally, environmental variables are often obtained via interpolation and/or may be measured with error. These factors mean that the "E" in GEA will probably always be imperfectly correlated with historical selection. The strength of that correlation will, of course, determine power in GEA studies. Using the simulations modelling local adaptation on the *BC* map via stabilizing selection, we compared the performance of WZA against the single-SNP GEA methods when the measured environment is imperfectly correlated with the phenotypic optima.



Commented [TB87]: I'm not crazy about this replacement title. In my view it is not just measurement error. If we use mean annual temp. as a variable for analyzing doug-fir, for example, most records only go back to 1970 or so. That's only about 1 or 2 generations for that species, so it seems a bit handwavey to me to say that the only difference between MAT and selection is measurement error.

Commented [MCW88R87]: While what you say is true, it is not the issue we investigate in this section. I think this title (or something like it) captures what is being looked at here.

Commented [SY89]: this feels like more explanation than is needed. A single sentence to motivate this section is enough. Maybe something like this?:

As environmental variables used in empirical GEA studies are likely imperfect proxies of the true effect of selection, we examined how this might affect the power of the methods.

Commented [TB90R89]: I see what you're saying, but there's nowhere else in the MS that says this stuff so I'd prefer to keep it here. Maybe this point is screamingly obvious to everyone except me, though.

Commented [MCW91]: Change legend order and x-axis label

Commented [TB92R91]: Done

Figure 5 The proportion of true positives recovered when the measured environment is imperfectly correlated with phenotypic optima. The correlation between environment and selection pressure is shown above each panel. Results are shown assuming the *BC Map* and the model of stabilising selection. Line indicate the means from 20 simulation replicates, and each is based on samples of 50 individuals from each of 40 demes.

We found that the WZA outperformed single SNP approaches (Kendall's τ or BayPass) when the measured environment was not perfectly correlated with phenotypic optima. We analyzed a sample of 40 demes from the population with 50 individuals taken in each location (Figure S2) but added random noise to the phenotypic optima from these locations to simulate environmental variables that were variably correlated with selection pressures and with which to conduct GEA. As might be expected, when the correlation between the measured environment and phenotypic optima was very weak (i.e., a correlation of 0.1), very few true positives were present in the top 50 genes under any of the methods we used, and those genes present only accounted for a small proportion of the covariance between phenotype and environment (Figure 5). With a correlation of 0.3 between the measured environment and true selection, WZAτ outperformed WZA_{BP} and the single-SNP approaches (Figure 5). With a correlation of 0.5 or 0.8 between the measured environment and phenotypic optima WZAτ outperformed all other methods, with only relatively small differences in performance between WZA_{BP} and the single-SNP approaches. Overall, this result suggests that WZAτ outperforms the single-SNP approaches when the measurement of the environment is a poor proxy for historical selection.

The width of analysis windows and recombination rate variation

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 S9). This is the same effect we described in a previous paper focusing on F_{ST} (Booker et al. 2020).

Application of the WZA to data from lodgepole pine

We re-analyzed a previously published (Yeaman et al. 2016) lodgepole pine (*Pinus contorta*) dataset comparing the WZA to the top-candidate method, which had been

Commented [MCW93]: Thissection doesn't read like a results section. Perhaps shorten it to focus on "Effects of recombination rate on distribution of WZA scores"? and save comments for discussion

Commented [SY94R93]: Discuss on Thursday?

Commented [TB95R93]: Done

Commented [MOU96]: Self citation is entirely appropriate here!

Commented [TB97R96]: Well I'll help myself to one then

developed for the original study. Overall, the WZA and top candidate statistic were broadly correlated and identified many of the same genes as the most strongly associated loci, but also differed in important ways. Across the lodgepole pine genome, there was a mean WZA score of 0.013 with a standard deviation $\sigma = 1.67$, and a fat right-hand tail (Figure S11). Figure 6A shows the relationship between WZA scores and the $-log_{10}(p\text{-value})$ from the top-candidate method, which were positively correlated (Kendall's τ = 0.245, *p*-value < 10^{-16}). When many of the SNPs in a gene had strongly associated statistics, both methods would tend to yield high scores (Figure 6D-E). When there were many SNPs with marginally significant empirical *p*-values (*i.e.* 0.05 < *p* < 0.10) at relatively high frequencies, the WZA method would tend to yield a high score but the top candidate method would not (Figure 6B). By contrast, if the most strongly associated SNPs tended to have low allele frequencies, the top candidate method would tend to yield a high score but the WZA would not (Figure 6C). There were several genes that had WZA scores greater than 10 (approximately 6σ), but very modest top-candidate scores (Figure 6A). Figure 6B shows that for one such region, there were several SNPs with high mean allele frequency that have small *p*-values. This particular region had a high score from the top-candidate method. Conversely, Figure 6C shows a region that only had a $Z_W \approx 5$, but an extreme score from the top-candidate method. In this case, there were numerous SNPs that passed the top-candidate outlier threshold, but they were mostly at low allele frequency. Figures 6C-D show the relationship between allele frequency and the empirical *p*-value for SNPs present in two genes that had extreme scores from both the top-candidate method and the WZA.

Commented [MOU98]: Perhaps add a single sentence motivating this section as a topic sentence for the paragraph. Maybe something like this:

Overall, the WZA and top candidate statistic were broadly correlated and identified many of the same genes as the most strongly associated loci, but also differed in important ways.

Commented [MCW99]: The caption of this figure doesn't explain why three of the dots in part A are large and orange

Commented [TB100R99]: Done

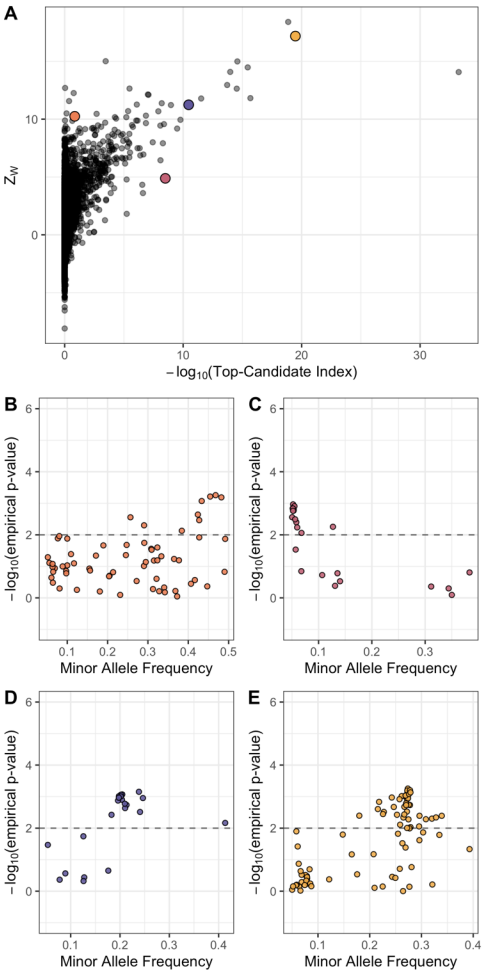


Figure 6 The WZA applied to GEA results on Lodgepole Pine for degree days below 0 (DD0). A) Z_w scores compared to scores from the top-candidate method for each of the genes analyzed by Yeaman et al. (2016). Panels B-E show the results for $-log_{10}$(p-values) for Spearman's ρ applied to individual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

Commented [MOU101]: Maybe change x-axis of panel A to "Top candidate score" or "index", so that it's not mis-represented as a true p-value.

Commented [TB102R101]: Yep, will do

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 usually more powerful than looking at individual sites in isolation, particularly when working with small samples or when the environmental variation being analyzed is only weakly correlated with selection (Figures 3 and 5). The WZA outperformed the other window-based method we examined, the top-candidate method (Figure 3). 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. Indeed, we found that when we had perfect knowledge of allele frequencies in all locations, the SNP-based GEA always outperformed or matched the WZA and top-candidate methods (Figure S13). However, such a situation is unrealistic, and empirical GEA studies will likely always be limited to finite samples from 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.

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 WZA was often better (Figure 3). 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 an arbitrary significance threshold. This is undesirable,

Commented [MCW103]: I think the real competition and point of comparison here is not top-candidate (as the first few paragraph now would be read) but sigle locus approaches. I think it would be better to start about he advantages of WZA ver single SNP approaches, then segue to why WZA is good for window-based analyses.

Commented [TB104R103]: I've tried to do that in this revision

Commented [MOU105]: Move to discussion

Commented [TB106R105]: Moved and expanded slightly.

Commented [MOU107]: This paragraph is good, but seems like a tangent – try to tie it back to the WZA/window-based methods – local adaptation should result in particularly pronounced LD

Commented [TB108R107]: I attempted to do that with the topic sentence in the previous paragraph.

Commented [SY109R107]: Reads much better now

Commented [MOU110]: I wouldn't say this – it just assumes that a fraction are extreme, but not necessarily causal.

I'd focus instead here on how when you use an arbitrary cutoff for significance, you can only detect adaptation where SNPs are actually rising above that cutoff – this restriction is not present in the WZA as there is no cutoff specified. This allows the WZA to detect genes that are extreme but have few individual SNPs that are extreme (panel B in the lodgepole figure)

Commented [TB111R110]: I reworded this section to address your concerns

Commented [SY112R110]: Reads much better now

however, because genuine genotype-environment correlations may be very weak and GEA may simply be an underpowered approach to identify alleles that contribute to local adaptation. If there were no detectable signal of local adaptation, ascribing significance to a fraction of the genome may lead to false positives. 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^{-10}. 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 edge of an expanding population. Allelic surfing can leave heterogeneous patterns of variation across a species range leaving signals across the genome that may resemble local adaptation (Novembre and Di Rienzo 2009; Klopfstein, Currat, and Excoffier 2006).

When performing a genome-scan using a windowed approach a question that inevitably arises is, how to choose the width of analysis windows? If analysis windows were too narrow, there may be little benefit in using a windowed approach over a single-SNP approach. In all the results presented above, 10,000bp analysis windows were used for the WZA. We found that the performance of the WZA when analysis windows that were narrower than 10,000bp was intermediate between the 10,000bp case and the single-SNP approach (Figure S12). Of course, if analysis windows were too wide, the signal oflocal adaptation may be diluted and the WZA would have little power. It seems like the ideal width for analysis windows would be informed by the the pattern of recombination rate variation, LD decay and SNP density across a species genome. In practice, it may be useful to perform the WZA on groups of SNPs, such as genes as in the Yeaman et al (2016) study. Future study is required to determine the optimal size for analysis windows.

A striking result from our comparison of the various GEA methods we tested in this study was the low power of *BayPass* compared to Kendall's τ (Figure 4). As mentioned in the Introduction, Lotterhos (2019) obtained a similar result in a previous study, though they had used Spearman's ρ rather than Kendall's τ. This presumably occurs because genome-wide population genetic structure is oriented along a similar spatial axis as adaptation, and the correction in *BayPass* therefore causes 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 is 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

Commented [MCW113]: Im not sure surfing is the prime problem compared to isolation by distance or patterns of shared history of sub-populations.

Commented [TB114R113]: Because ibd or shared history will shape the genome-wide distribution in a particular way it gets ironed out by the empirical p-value approach. Expansion and surfing can produce pockets of the genome with a patchy pattern of LD that looks like selection, on top of the usual IBD issues.

Commented [SY115R113]: I think this looks good as is – it's an example of one such process

Commented [TB116]: Here's an attempt at the window size paragraph. I'm not crazy about it, so please feel free to hack away at it.

I feel strongly that we should submit what we have soon but prepare for the possibility of a reviewer asking for the future direction I mention.

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. 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.

Perhaps more striking was how underpowered these GEA methods were at identifying the genes involved in local adaptation. In our simulations, around 6 locally adaptive genes established in each replicate in each of the cases we tested. When analyzing our simulations, we examined the true positives present in the top 1, 2, 3, …, 50 genes, but in most cases, the proportion of all true positives identified did not reach 1.0 (Figures 3-5), indicating high false discovery rates. Each simulation replicate included 1,000 genes, so the top 50 represents the 95th percentile of the genome-wide distribution. Examining the upper percentiles of the empirical distribution of GEA scores is an approach taken in empirical analyses (e.g. Shi et al. 2021; Leigh et al. 2021), though it would perhaps be preferable to have a threshold that was applied to an appropriate null distribution. One drawback of the WZA is that since the distribution of WZA scores was non-normal even under neutrality (Figure 2), we cannot compute a parametric *p*-value for each analysis window tested. *BayPass*, on the other hand, returns a Bayes factor for each analyzed SNP, the log-transformed ratio of the likelihoods under the alternate and null hypotheses. A general rule of thumb for the interpretation of Bayes factors is that BFs > 20 are considered strong evidence against the null hypothesis (i.e., Jeffrey's rule). Overall, applying a stringent Bayes factor threshold to *BayPass* may result in a test with low false positive rates, the WZA may provide a more sensitive test at the cost of specificity.

Ultimately, performing GEA analyses using analysis windows is an attempt to leverage information from closely linked sites. With the advent of methods for reconstructing ancestral recombination graphs from population genomic data (Hejase et al. 2020), perhaps a GEA method could be developed that explicitly analyzes inferred genealogies rather than individual markers in a manner similar to regression of phenotypes on genealogies proposed by Ralph et al. (2020). Such a method would require large numbers of individuals with phased genome sequences, which may now be feasible given recent technological advances (Meier et al. 2021).

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

Commented [SY117]: I took a crack at this – feel free to re-write!

As mentioned in my email – I'm not actually sure if convering to empirical p is better? Do we show this anywhere? I stopped writing here because I'm not actually sure I agree with what I was writing!

Commented [TB118R117]: No we didn't show this anywhere. The parametric p is where we started, but I abandoned that line of thinking after we adopted the empirical p-value approach. The empirical p way is more powerful and better behaved. I've added two additional figures to illustrate this.

Commented [MCW119]: Unpack this for the reader

Commented [TB120R119]: This is expanded on in the WZA description section

Commented [TB121R119]: We talked a while ago about how one of you could draft a little something to cover this point. Every time I try I write something verbose and clunky.

Commented [TB122]: I am actually pretty skeptical that this can be done. Also, it may be perceived as a dig at the BayPass/BayEnv approach, which I don't want to do.

Commented [SY123]: I'm not sure what this means exactly – how does this differ from upper percentiles? Or do you mean a threshold based on some assumed demographic model? Cause I really don't like that approach.

Commented [TB124R123]: What I'm trying to get at is that looking at the top 5% is not great because there is always a top 5%. If you had a good null model (which we do not have) you could calculate a p-value and determine significance on that.
In the paragraph above we say that it would be good to have a model that gave a parametric *p*-value that controlled for population structure. If you had such a p-value you could use crazy deviations from the null to determine significanc… [1]

Commented [MOU125]: This all seems like a tangent to the main point: … [2]

Commented [TB126R125]: Yeah, I suppose it's a tangent to the point about power, but I think it's important to … [3]

Commented [SY127R125]: I've added something above that does this now but needs finishing off

Commented [TB128R125]: I strongly suspect that there is a

Commented [MCW129]: I think this might be dropped – it is speculative and not directly connected to the method we develop here.

Commented [TB130R129]: One of the things we are advocating with the WZA is that people should try and … [4]

Commented [SY131R129]: I think it's good to connect to where the field may be going – I'd suggest keeping it

Commented [TB132R129]:

Acknowledgements

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 Finlay Booker for moral support throughout the course of this project. TRB is supported by funding from Genome Canada, Genome Alberta and NSERC Discovery Grants awarded to MCW and SY. SY is supported by an AIHS research chair and NSERC Discovery Grant. MCW is supported by an NSERC Discovery Grant. 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, Sally N, and Michael C Whitlock. 2013. “Assisted Gene Flow to Facilitate Local Adaptation to Climate Change.” *Annu. Rev. Ecol. Evol. Syst.* 44 (1): 367–88.

Barton, Nick, and Bengt Olle Bengtsson. 1986. “The barrier to genetic exchange between hybridising populations.” *Heredity* 57 (3): 357–76. *https://doi.org/10.1038/hdy.1986.135*.

Bhatia, Gaurav, Nick Patterson, Sriram Sankararaman, and Alkes L. Price. 2013. “Estimating and interpreting FST: The impact of rare variants.” *Genome Research* 23 (9): 1514–21. *https://doi.org/10.1101/gr.154831.113*.

Bontrager, M., C. D. Muir, C. Mahony, D. E. Gamble, R. M. Germain, A. L. Hargreaves, E. J. Kleynhans, K. A. Thompson, and A. L. Angert. 2020. “Climate warming weakens local adaptation.” bioRxiv. *https://doi.org/10.1101/2020.11.01.364349*.

Booker, Tom R., Sam Yeaman, and Michael C. Whitlock. 2020. “Variation in recombination rate affects detection of outliers in genome scans under neutrality.” *Molecular Ecology* 29 (22): 4274–9. *https://doi.org/10.1111/mec.15501*.

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

Coop, Graham, David Witonsky, Anna Di Rienzo, and Jonathan K. Pritchard. 2010. “Using environmental correlations to identify loci underlying local adaptation.” *Genetics* 185 (4): 1411–23. *https://doi.org/10.1534/genetics.110.114819*.

Forester, Brenna R., Matthew R. Jones, Stéphane Joost, Erin L. Landguth, and Jesse R. Lasky. 2016. “Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes.” *Molecular Ecology* 25 (1): 104–20. *https://doi.org/10.1111/mec.13476*.

Forester, Brenna R., Jesse R. Lasky, Helene H. Wagner, and Dean L. Urban. 2018. “Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations.” *Molecular Ecology* 27 (9): 2215–33. *https://doi.org/10.1111/mec.14584*.

Frichot, Eric, and Olivier François. 2015. “LEA: An R package for landscape and ecological association studies.” Edited by Brian O’Meara. *Methods in Ecology and Evolution* 6 (8): 925–29. *https://doi.org/10.1111/2041-210X.12382*.

Frichot, Eric, Sean D. Schoville, Guillaume Bouchard, and Olivier François. 2013. “Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models.” *Molecular Biology and Evolution* 30 (7): 1687–99. *https://doi.org/10.1093/molbev/mst063*.

781 Gautier, Mathieu. 2015. “Genome-wide scan for adaptive divergence and association
782 with population-specific covariates.” *Genetics* 201 (4): 1555–79.
783 *https://doi.org/10.1534/genetics.115.181453*.

784 Haldane, J. B. S. 1948. “The theory of a cline.” *Journal of Genetics* 48 (3): 277–84.
785 *https://doi.org/10.1007/BF02986626*.

786 Haller, Benjamin C, Jared Galloway, Jerome Kelleher, Philipp W Messer, and Peter L
787 Ralph. 2019. “Tree-sequence recording in SLiM opens new horizons for forward-time
788 simulation of whole genomes.” *Mol. Ecol. Resour.* 19 (2): 552–66.

789 Hancock, Angela M., Benjamin Brachi, Nathalie Faure, Matthew W. Horton, Lucien B.
790 Jarymowycz, F. Gianluca Sperone, Chris Toomajian, Fabrice Roux, and Joy Bergelson.
791 2011. “Adaptation to climate across the Arabidopsis thaliana genome.” *Science* 334
792 (6052): 83–86. *https://doi.org/10.1126/science.1209244*.

793 Hejase, Hussein A., Noah Dukler, and Adam Siepel. 2020. “From Summary Statistics to
794 Gene Trees: Methods for Inferring Positive Selection.” Elsevier Ltd.
795 *https://doi.org/10.1016/j.tig.2019.12.008*.

796 Hereford, Joe. 2009. “A quantitative survey of local adaptation and fitness trade-offs.”
797 The University of Chicago Press. *https://doi.org/10.1086/597611*.

798 Hoban, Sean, Joanna L. Kelley, Katie E. Lotterhos, Michael F. Antolin, Gideon
799 Bradburd, David B. Lowry, Mary L. Poss, Laura K. Reed, Andrew Storfer, and Michael
800 C. Whitlock. 2016. “Finding the genomic basis of local adaptation: Pitfalls, practical
801 solutions, and future directions.” *American Naturalist* 188 (4): 379–97.
802 *https://doi.org/10.1086/688018*.

803 Kirkpatrick, Mark, and Nick Barton. 2006. “Chromosome inversions, local adaptation
804 and speciation.” *Genetics* 173 (1): 419–34. *https://doi.org/10.1534/genetics.105.047985*.

805 Klopfstein, Seraina, Mathias Currat, and Laurent Excoffier. 2006. “The fate of mutations
806 surfing on the wave of a range expansion.” *Molecular Biology and Evolution*.
807 *https://doi.org/10.1093/molbev/msj057*.

808 Legendre, P., and L. Legendre. 2012. *Numerical Ecology, Volume 24*. 3rd Englis.
809 Elsevier. *https://www.elsevier.com/books/numerical-ecology/legendre/978-0-444-53868-*
810 *0*.

811 Lotterhos, Katie E. 2019. “The Effect of Neutral Recombination Variation on Genome
812 Scans for Selection.” *G3* 9 (6): 1851–67.

813 Meirmans, Patrick G. 2012. “The trouble with isolation by distance.” *Molecular Ecology*
814 21 (12): 2839–46. *https://doi.org/10.1111/j.1365-294X.2012.05578.x*.

Mimura, M., and S. N. Aitken. 2007. “Adaptive gradients and isolation-by-distance with postglacial migration in Picea sitchensis.” *Heredity* 99 (2): 224–32. *https://doi.org/10.1038/sj.hdy.6800987*.

Morales, Hernán E., Rui Faria, Kerstin Johannesson, Tomas Larsson, Marina Panova, Anja M. Westram, and Roger K. Butlin. 2019. “Genomic architecture of parallel ecological divergence: Beyond a single environmental contrast.” *Science Advances* 5 (12): eaav9963. *https://doi.org/10.1126/sciadv.aav9963*.

Mosca, Elena, Santiago C. González-Martínez, and David B. Neale. 2014. “Environmental versus geographical determinants of genetic structure in two subalpine conifers.” *New Phytologist* 201 (1): 180–92. *https://doi.org/10.1111/nph.12476*.

Noor, M. A. F., K. L. Gratos, L. A. Bertucci, and J. Reiland. 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. *https://doi.org/10.1073/pnas.221274498*.

Novembre, John, and Anna Di Rienzo. 2009. “Spatial patterns of variation due to natural selection in humans.” Nat Rev Genet. *https://doi.org/10.1038/nrg2632*.

Pavy, N., M. C. Namroud, F. Gagnon, N. Isabel, and J. Bousquet. 2012. “The heterogeneous levels of linkage disequilibrium in white spruce genes and comparative analysis with other conifers.” *Heredity* 108 (3): 273–84. *https://doi.org/10.1038/hdy.2011.72*.

Ralph, Peter, Kevin Thornton, and Jerome Kelleher. 2020. “Efficiently summarizing relationships in large samples: A general duality between statistics of genealogies and genomes.” *Genetics* 215 (3): 779–97. *https://doi.org/10.1534/genetics.120.303253*.

Rieseberg, Loren H. 2001. “Chromosomal rearrangements and speciation.” Elsevier. *https://doi.org/10.1016/S0169-5347(01)02187-5*.

Sakamoto, Takahiro, and Hideki Innan. 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. *https://doi.org/10.1534/genetics.119.302311*.

Schlötterer, Christian, Raymond Tobler, Robert Kofler, and Viola Nolte. 2014. “Sequencing pools of individuals-mining genome-wide polymorphism data without big funding.” Nature Publishing Group. *https://doi.org/10.1038/nrg3803*.

Stapley, Jessica, Philine G. D. Feulner, Susan E. Johnston, Anna W. Santure, and Carole M. Smadja. 2017. “Variation in recombination frequency and distribution across eukaryotes: Patterns and processes.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1736). *https://doi.org/10.1098/rstb.2016.0455*.

Todesco, Marco, Gregory L. Owens, Natalia Bercovich, Jean Sébastien Légaré, Shaghayegh Soudi, Dylan O. Burge, Kaichi Huang, et al. 2020. “Massive haplotypes

underlie ecotypic differentiation in sunflowers." *Nature* 584 (7822): 602–7. https://doi.org/10.1038/s41586-020-2467-6.

Walsh, B., and M. Lynch. 2018. *Evolution and Selection of Quantitative Traits*. Oxford University Press. https://global.oup.com/academic/product/evolution-and-selection-of-quantitative-traits-9780198830870?cc=ca&lang=en&.

Wang, Tongli, Andreas Hamann, Dave Spittlehouse, and Carlos Carroll. 2016. "Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America." Edited by Inés Álvarez. *PLOS ONE* 11 (6): e0156720. https://doi.org/10.1371/journal.pone.0156720.

Weir, Bruce S, and C Clark Cockerham. 1984. "Estimating F-statistics for the analysis of population structure." *Evolution* 38 (6): 1358–70.

Whitlock, M. C. 2005. "Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach." *Journal of Evolutionary Biology* 18 (5): 1368–73. https://doi.org/10.1111/j.1420-9101.2005.00917.x.

Yeaman, Sam. 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. https://doi.org/10.1073/pnas.1219381110.

Yeaman, Sam, Aleeza C. Gerstein, Kathryn A. Hodgins, and Michael C. Whitlock. 2018. "Quantifying how constraints limit the diversity of viable routes to adaptation." *PLoS Genetics* 14 (10): e1007717. https://doi.org/10.1371/journal.pgen.1007717.

Yeaman, Sam, Kathryn A. Hodgins, Katie E. Lotterhos, Haktan Suren, Simon Nadeau, Jon C. Degner, Kristin A. Nurkowski, et al. 2016. "Convergent local adaptation to climate in distantly related conifers." *Science* 353 (6306): 1431–3. https://doi.org/10.1126/science.aaf7812.

Zhou, Xiang, Peter Carbonetto, and Matthew Stephens. 2013. "Polygenic Modeling with Bayesian Sparse Linear Mixed Models." *PLoS Genetics* 9 (2): 1003264. https://doi.org/10.1371/journal.pgen.1003264.

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 with a minor allele frequency threshold greater than 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, $4N_d\mu_{neu} = 4.0 \times 10^{-8}$.

There are no estimates available of the mutation rate for locally adaptive alleles. As such, we had no empirical estimates to base our simulations on. Instead, 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 an average of 6.XX locally adaptive genes establishing. For stabilizing selection, a mutation rate of $\mu_{alpha} = 1 \times 10^{-8}$, 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 benficial allele. In the case of stabilizing selection, all 1,000bp 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 $\frac{0.005}{4\times0.5\times10^{-8}} = 2.5 \times 10^6$, 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.

When modelling directional selection, 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.

As stated the main text, for stabilizing selection simulations we chose $V_s = 192$ as this gave a maximum of 50% difference in fitness between individuals grown in home-like conditions versus away-like conditions.

Migration rate

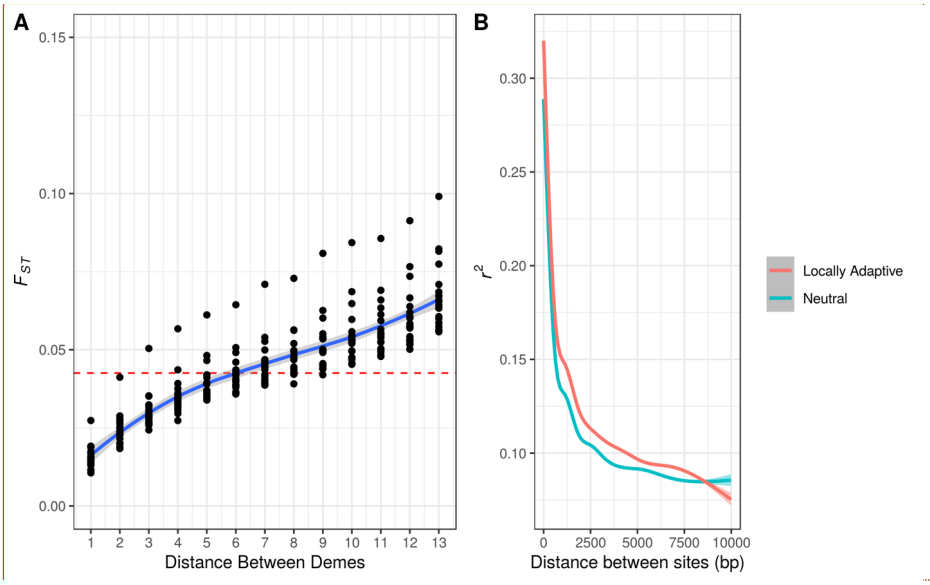
We wanted to model populations with F_{ST} across the metapopulation of approximately 0.05, as has been reported for widely distributed conifer species such as lodgepole pine and interior spruce (Yeaman et al. 2016). For the stepping-stone simulations, we chose a migration rate of $\frac{7.5}{2N_d}$ 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.

Commented [MCW133]: I don't understand this phrase – we don't used the maximum difference at all

Commented [TB134R133]: Yes we do. This is how we got to the selection parameters. The Hereford and Bontrager studies summarise a whole bunch of studies and show that local adaptation often gives rise to about 50% difference in fitness between localities. Motivated by that, I set the strength of directional or purifying selection such that the best possible genotype versus the worst possible genotype had an approximate 50% difference in fitness.

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}



963

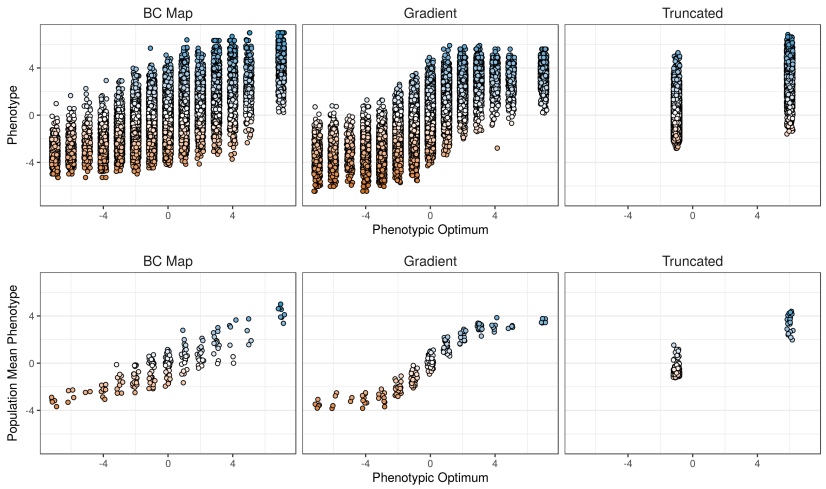
964 **Figure S1** Summary statistics from neutral simulations. A) F_{ST} between pairs of demes
965 in stepping-stone populations. The average across replicates is 0.042. B) LOESS
966 smoothed LD, as measured by r^2, between pairs of SNPs in genes that are either
967 evolving neutrally are locally adaptation as indicated by the color. Smoothing was
968 performed using the ggplot2 package in R.

969

Commented [MCW135]: The x-axis labels are inconsistent in use of case.

Commented [TB136R135]: Will fix

Commented [MCW137]: Missing number



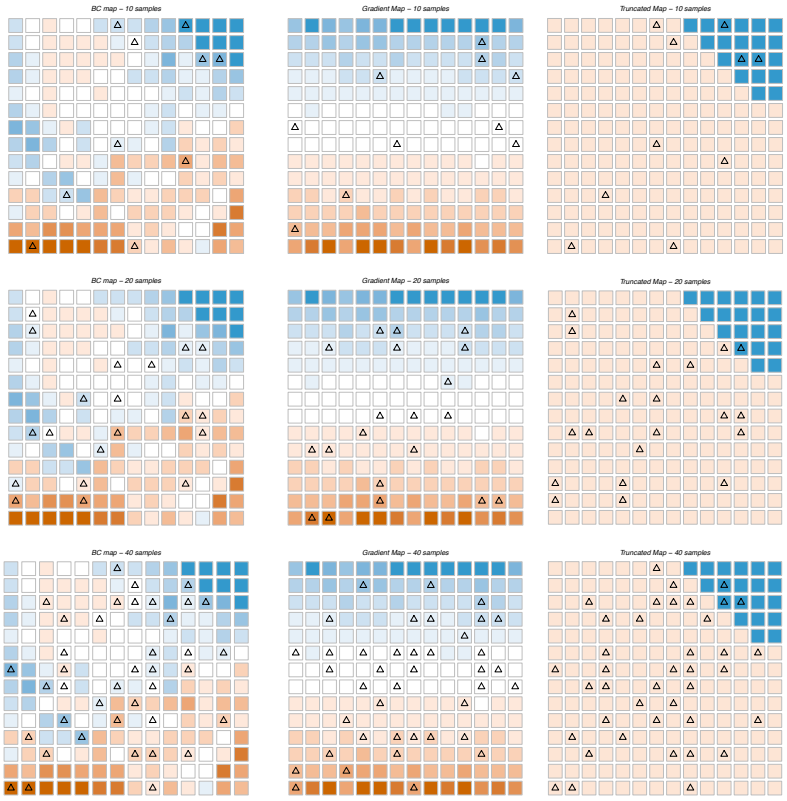
970

971 **Figure S2** Individual and population mean phenotypes observed in representative
972 simulations for each of the environment maps simulated. A small amount of horizontal
973 jitter was added to points for visualization purposes. Colors are for visualization
974 purposes only.

975

Commented [MCW138]: Say what the colors represnet

Commented [TB139R138]: They are to make the plot look pretty, they correspond to phenotype.



976

977 **Figure S3** Locations of sampled demes on the maps of environmental variation we
978 assumed in the simulations. Triangles indicate the locations where individuals were
979 sampled in each case. Colors represent the optimal phenotype in each population the
980 same as Figure 1 in the main text.

981

Commented [MCW140]: Say what the colors mean

Commented [TB141R140]: done

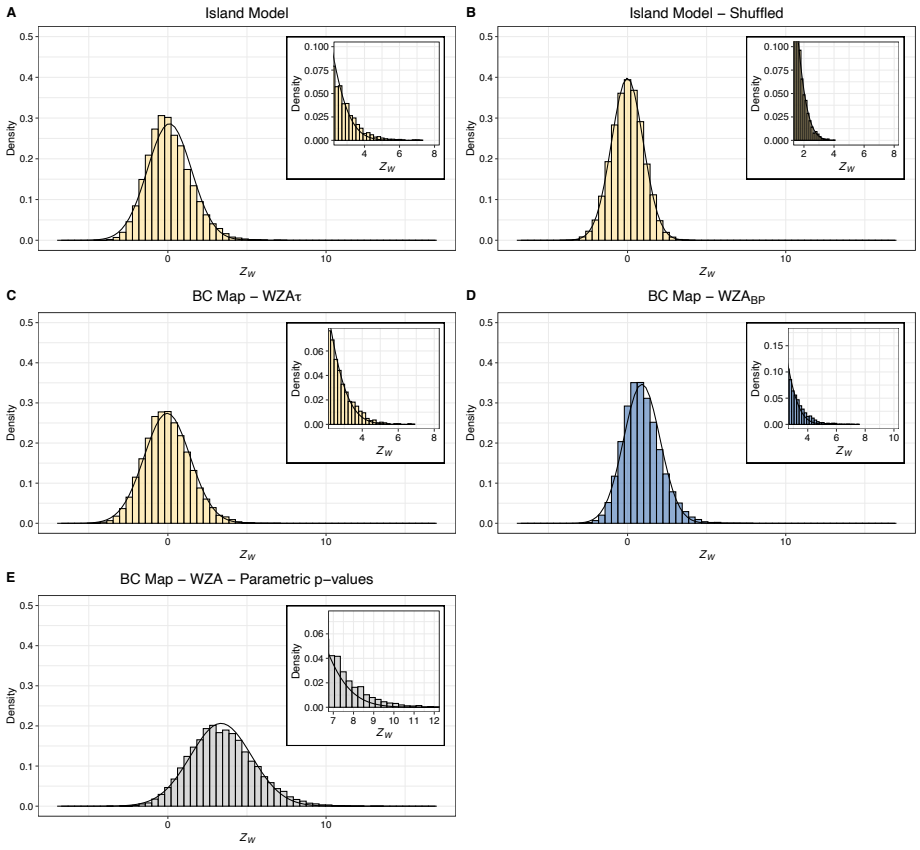


Figure S4 The distribution of WZA scores from neutral simulations with details of the right tail in the insets. Overlaid on each panel is the normal distribution fitted to each dataset. In all cases, results from 20 simulation replicates are plotted together.

Commented [SY142]: It might be helpful to show another panel that gives the WZA-tau on the raw p-values, not empirical p.

Commented [TB143R142]: Added

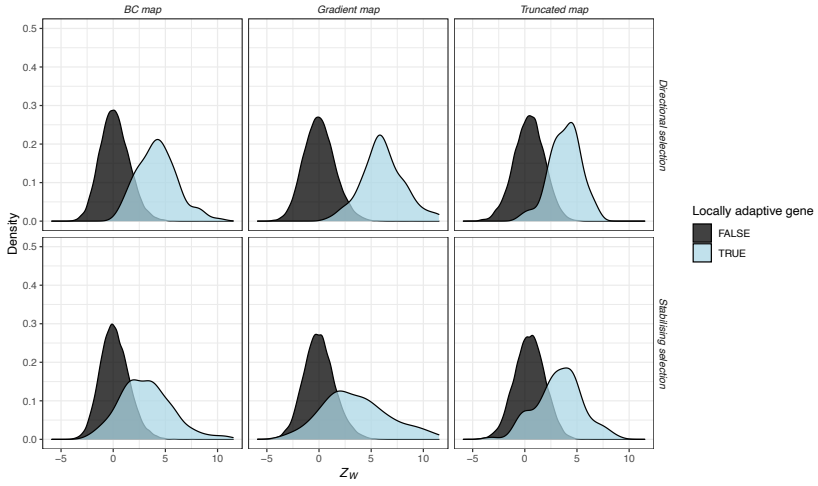
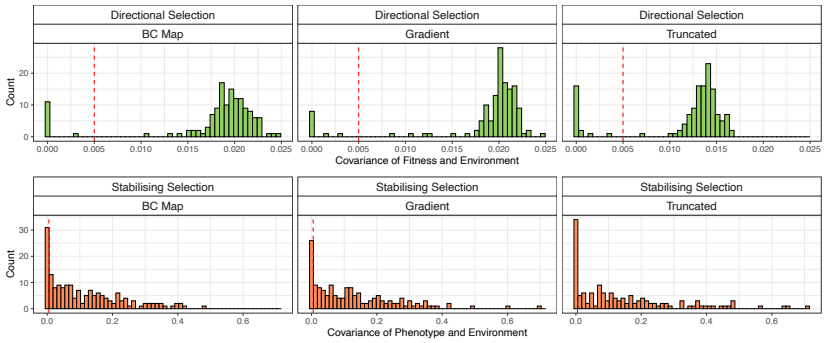


Figure S5 The distribution of WZA scores from simulations of local adaptation. Note, the plot does not indicate the relative frequency of genes that are or are not locally adaptive.



992

993 **Figure S6** The effect size distribution from simulations of local adaptation. The vertical
994 line indicates the threshold we applied to the simulated data to classify genes as locally
995 adaptive or not.

996

Commented [MCW144]: Be more specifc about meaning of Cov

Commented [TB145R144]: I'll return to this once we've finalized the section describing the quantity itself!

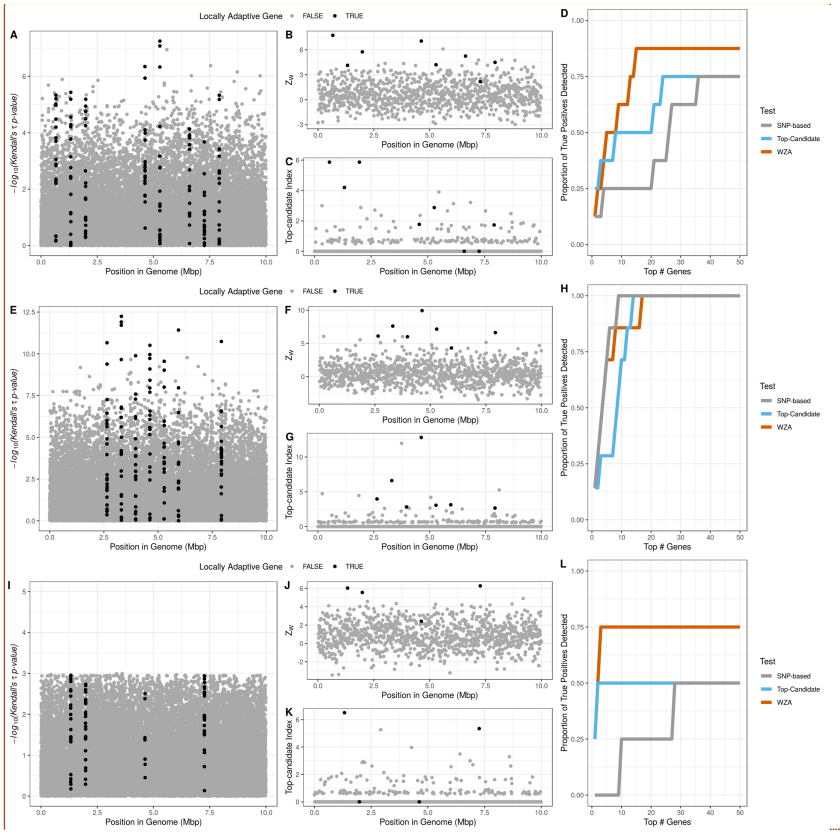


Figure S7 Plots demonstrating the genomic landscape of genotype-environment correlations for a single replicate for each of the three maps of environmental variation we simulated. From top to bottom, the three rows correspond to the *BC Map* (panels A-D), the gradient map (panels E-H) and the truncated map (panels I-L), respectively. The leftmost panel in each row shows the Manhattan plot of $-log_{10}$(p-values) from Kendall's τ (panels A, E and I). The central panels in each row show the distribution of Z_W scores from the WZA across the genome (B, F and J) and the distribution of results from the top-candidate method (C, G and K). The rightmost panels show the proportion of locally adapted genes identified using the three different tests for an increasing number of genes in the search effort. Results are shown for directional selection simulations. Note that only SNPs with a minor allele frequency > 0.05 are shown in panels (A, E and I).

Commented [MCW146]: Font size far too small in all cases

Commented [MCW147R146]:

Commented [TB148R146]: I'll make the text larger on these.

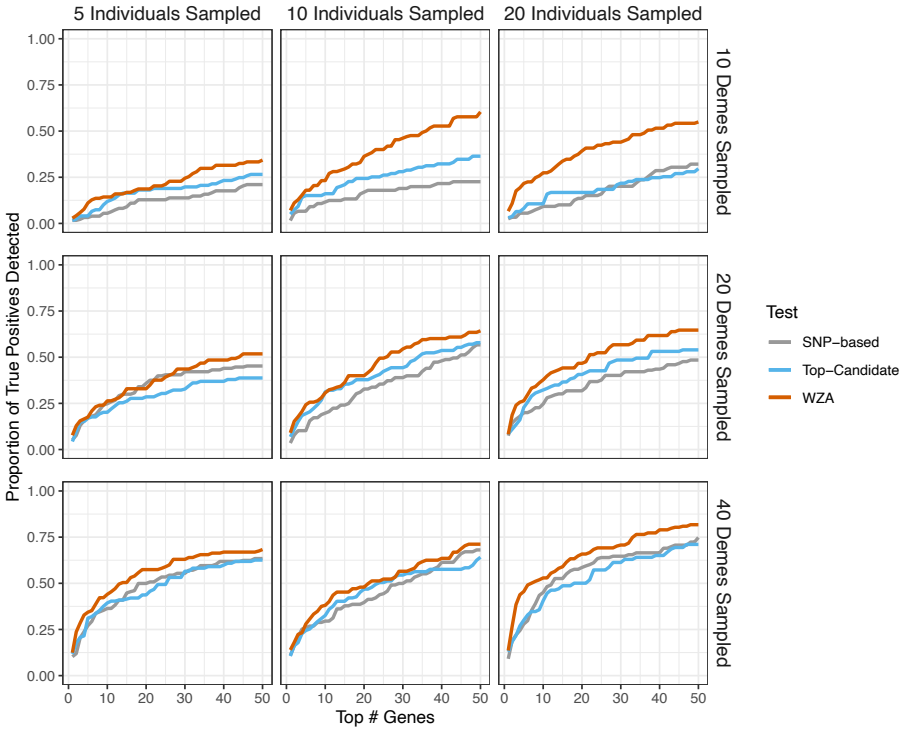


Figure S8 Comparison of the WZA, the top-candidate and the single-SNP approaches with varying numbers of individuals sampled per deme. Simulations shown used the *BC map* and directional selection. Lines represent the mean of 20 simulation replicates.

Commented [MCW149]: Same comments as main text figure about legend order and x-axis label

Commented [TB150R149]: I'll use the same order as In the main text.

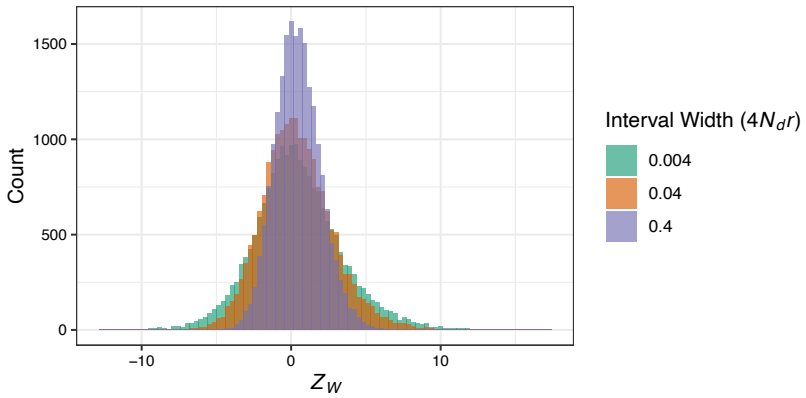


Figure S9 The distribution of Z_W scores under different recombination rates.

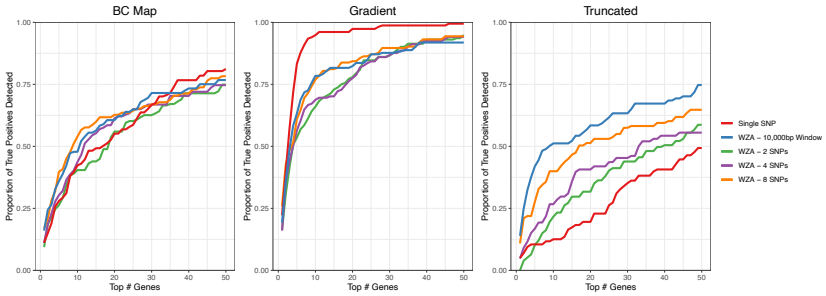


Figure S10 Comparing the performance of the WZA genes identified using the WZA, using analysis windows analyzing a fixed number of SNPs. Lines represent the means of 20 replicates.

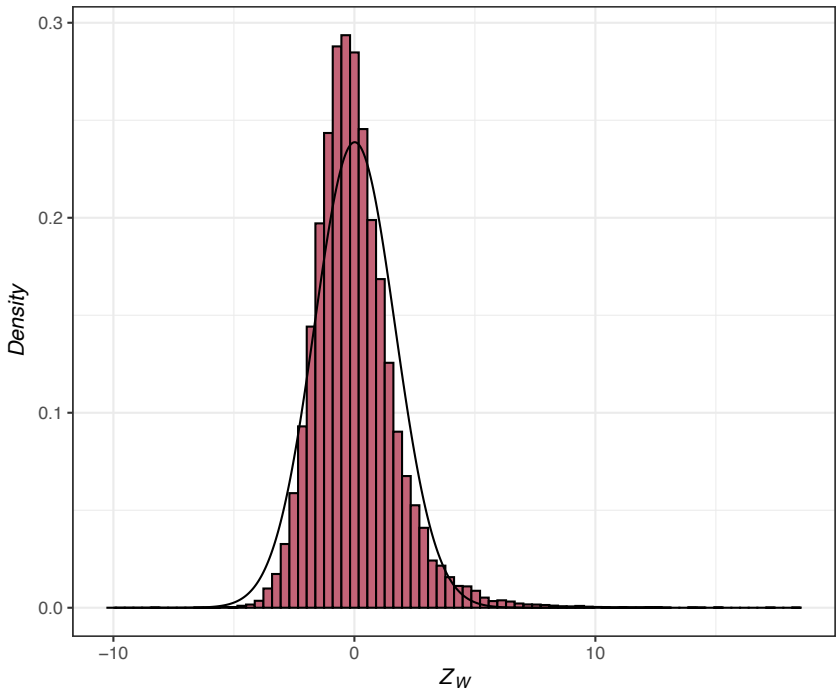


Figure S11 The distribution of A) degree days < 0 (DD0) across the populations of P. contorta sampled by Yeaman et al (2016) and B) Z_W scores for the GEA on DD0. Note that the DD0 values in A) are unscaled. In B) the curve shows a normal distribution fitted to the data.

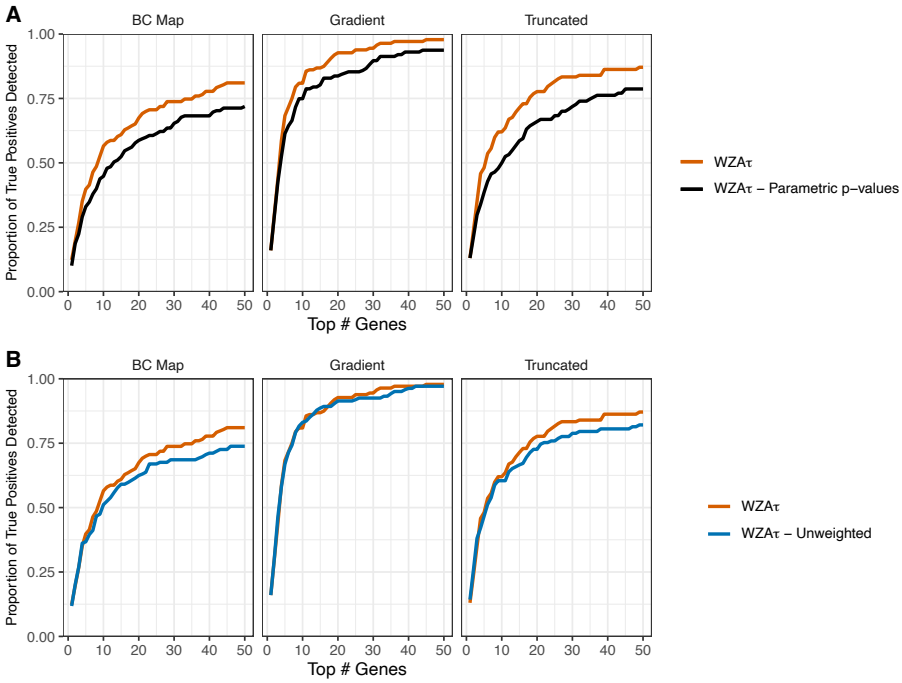

A
BC Map
Gradient
Truncated
Proportion of True Positives Detected
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
50
Top # Genes
WZAτ
WZAτ – Parametric p–values
B
BC Map
Gradient
Truncated
Proportion of True Positives Detected
Top # Genes
WZAτ
WZAτ – Unweighted

Figure S12

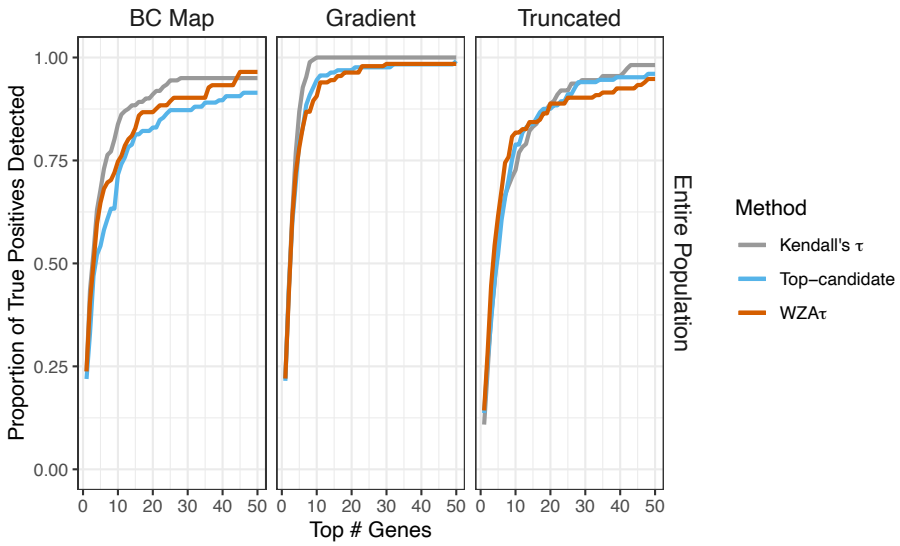


Figure S13 A comparison of three methods to identify the genetic basis of local adaptation when one has complete information on all aspects of the metapopulation.

Page 26: [1] Commented [TB124R123] Tom Booker 03/06/2021 14:28:00

What I'm trying to get at is that looking at the top 5% is not great because there is always a top 5%. If you had a good null model (which we do not have) you could calculate a p-value and determine significance on that.

In the paragraph above we say that it would be good to have a model that gave a parametric *p*-value that controlled for population structure. If you had such a p-value you could use crazy deviations from the null to determine significance

Page 26: [2] Commented [MOU125] Microsoft Office User 11/05/2021 11:21:00

This all seems like a tangent to the main point:

When structure correction is conducted in datasets where population structure covaries with the environment driving selection, the power to detect true positives is reduced – we don't actually prove that this is what's happening, but it should be discussed as the most plausible explanation

Page 26: [3] Commented [TB126R125] Tom Booker 21/05/2021 12:01:00

Yeah, I suppose it's a tangent to the point about power, but I think it's important to emphasize that the way someone else might analyse the data

Page 26: [4] Commented [TB130R129] Tom Booker 01/06/2021 13:53:00

One of the things we are advocating with the WZA is that people should try and build tree-thinking into their GEAs. So I disagree here – I like a little bit of speculation and future directions in a Discussion.