MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 1031–1046 doi: 10.1111/mec.13100

The relative power of genome scans to detect local adaptation depends on sampling design and statistical method

KATIE E. LOTTERHOS[1] and MICHAEL C. WHITLOCK

Department of Zoology, University of British Columbia, 6270 University Blvd., Vancouver, BC, V6T 1Z4, Canada

Abstract

Although genome scans have become a popular approach towards understanding the genetic basis of local adaptation, the field still does not have a firm grasp on how sampling design and demographic history affect the performance of genome scans on complex landscapes. To explore these issues, we compared 20 different sampling designs in equilibrium (i.e. island model and isolation by distance) and nonequilibrium (i.e. range expansion from one or two refugia) demographic histories in spatially heterogeneous environments. We simulated spatially complex landscapes, which allowed us to exploit local maxima and minima in the environment in 'pair' and 'transect' sampling strategies. We compared F_{ST} outlier and genetic–environment association (GEA) methods for each of two approaches that control for population structure: with a covariance matrix or with latent factors. We show that while the relative power of two methods in the same category (F_{ST} or GEA) depended largely on the number of individuals sampled, overall GEA tests had higher power in the island model and F_{ST} had higher power under isolation by distance. In the refugia models, however, these methods varied in their power to detect local adaptation at weakly selected loci. At weakly selected loci, paired sampling designs had equal or higher power than transect or random designs to detect local adaptation. Our results can inform sampling designs for studies of local adaptation and have important implications for the interpretation of genome scans based on landscape data.

Keywords: adaptation, genome scans, landscape genetics, range expansion, sampling design, spatial statistics

Received 4 September 2014; accepted 14 January 2015

Introduction

Genome scans have become a feasible way to quickly identify candidate genes that are targets of natural selection in both model and nonmodel species. This 'reverse ecology' approach (*sensu* Li *et al.* 2008) seeks to directly uncover the genetic basis of adaptation from population genetic patterns, rather than to identify the specific phenotypic traits that are acted upon by natural selection. Although adaptive genetic variation evolves as a result of environmental or ecological factors imposing selective pressure over a heterogeneous landscape, until recently many genome scan methods have not incorporated information on sampling location or the environment (Schoville *et al.* 2012). In addition, relatively little attention has been paid to sampling design: how sampling locations are chosen on the landscape ('sampling strategy') and how individuals are allocated among those samples ('sample size').

One of the most widely used genome scan methods is the F_{ST} outlier test. F_{ST} is a measure of genetic differentiation among sampled populations, and outliers—unusually large values of F_{ST}—are candidates for divergent selection among populations. Since originally

Correspondence: Katie E Lotterhos, Fax: (336) 758-6008; E-mail: lotterke@wfu.edu

[1]Present address: Department of Biological Science, Wake Forest University, Box 7325 Reynolda Station, Winston Salem, NC 27109, USA

presented by Lewontin & Krakauer in 1973; a number of outlier tests have been developed and used for the discovery of candidate genes (Beaumont & Nichols 1996; Vitalis *et al.* 2001; Beaumont & Balding 2004; Foll & Gaggiotti 2008; Excoffier *et al.* 2009; Bonhomme *et al.* 2010; Günther & Coop 2013; Duforet-Frebourg *et al.* 2014), including some tests which do not use F_{ST} per se but other measures of differentiation (e.g. X^TX from Günther & Coop 2013). Many of the widely used methods, however, do not appropriately control for the population structure that results from more realistic demographic histories such as range expansion (Lotterhos & Whitlock 2014). Note that while samples may correspond to environmental or other phenotypic differences among populations, that information is not incorporated into the test statistic.

While a benefit of outlier tests based on genetic differentiation is that they may uncover loci subject to any selective force, this is also a limitation because they cannot test specific hypotheses about the nature of selection. Recently, a number of methods have been developed that take explicitly into account the relationship between genotypes and the environment, which we refer to collectively here as genetic–environment associations (GEA tests, Hedrick *et al.* 1976). An early analytical approach to test for GEAs employed a logistic regression to provide a measure of association between allele frequencies and environmental parameters (Joost *et al.* 2007). However, Meirmans (2012) and De Mita *et al.* (2013) pointed out that this method suffered from high false-positive rates because it did not explicitly control for population structure in the test statistic. Because populations on a landscape typically exhibit some degree of isolation by distance, spatial autocorrelation in allele frequencies can cause associations between gene frequencies and the environment to occur by chance (Meirmans 2012). Thus, as in the case of F_{ST} outlier tests, it is important to control for population structure in the calculation of the GEA test statistic.

Currently, there are two emerging approaches towards controlling for population structure in GEA and outlier tests (for F_{ST} or other differentiation measures). In the first approach, a set of putatively neutral loci are used to first estimate the neutral population structure; then, the neutral population structure is controlled for in the calculation of the test statistic. This is the approach implemented in the program Bayenv2, which first estimates a covariance matrix among populations and then accounts for that covariance in the outlier or GEA test (Coop *et al.* 2010; Günther & Coop 2013). In the second emerging approach, population structure is controlled for by 'latent factors', which are estimated by considering the statistical model and the data simultaneously (Frichot *et al.* 2013; Duforet-Frebourg *et al.* 2014). Latent factors are closely related to principal components, and each factor describes a different aspect of the population structure. This approach is implemented in a GEA test in the program LFMM (Frichot *et al.* 2013) and an outlier test in the program PCAdapt (Duforet-Frebourg *et al.* 2014).

Despite the plethora of new methods being developed to analyse landscape genomic data, relatively little attention has been paid to sampling design for detecting local adaptation. Although a few studies have compared GEA methods to F_{ST} outlier tests with simulated data (De Mita *et al.* 2013; Jones *et al.* 2013; de Villemereuil *et al.* 2014), only one of these studies compared different sampling designs on a landscape. De Mita *et al.* (2013) used individual-based simulations to model selection on a 10×10 grid to a linear cline. They found that in the island model, sampling designs that allocated individuals to many populations (with few individuals per population) tended to have lower false-positive rates and higher power for most statistics (De Mita *et al.* 2013).

In nature, however, populations are adapting to environments that are spatially heterogeneous at different spatial scales. In principle, one could exploit environmental gradients at different spatial scales—such as in the form of multiple paired samples or transects—to improve the power of genome scans. If two populations are geographically close but experience substantially different selective environments, the expected differentiation due to nonselective reasons would be small, while selection could still be very effective. As a result, looking at pairs of nearby populations permits the measurement of the differentiation due to selective forces without the confounding possibility of differentiation due to neutral demographic history. Both GEA and F_{ST} outlier tests depend on finding loci that are more differentiated than expected for neutral loci. To find the starkest differences between the patterns of selected and neutral genes, we hypothesize that power can be increased by comparing populations that are physically near to each other (and therefore genetically similar for neutral genes) and yet in different environments (and therefore differentiated by selection). This idea has not yet been formally evaluated.

In this study, we took a comprehensive approach to formally evaluate how both sampling design and demographic history affect the power of GEA and F_{ST} outlier tests, for the two major ways of controlling for population structure (with a covariance matrix or with latent factors). In a previous study, we showed that the X^TX approach from Bayenv was one of the most powerful measures of genetic differentiation to detect local adaptation; as a result, in this study, we focus on this measure in lieu of other more widely used methods such as

BayeScan or fdist. We also add tests of a new outlier approach PCAdapt. In contrast to previous studies, we simulated large populations on complex environmental landscapes and focused on how sampling design affected the power to detect local adaptation. Our sampling designs differed in the number of individuals sampled ('sample size') and the location of populations with respect to the environment ('sampling strategy', such as paired, random or transects).

Not all of the methods we compared report the statistical support for concluding that a locus is affected by local adaptation in the same way; indeed, the Bayenv method gives no absolute standard for making these conclusions at all. To make fair comparisons, we compared the power of all programs on a common scale by creating empirical *P*-values from known neutral loci. In reality, such known neutral loci will not be unambiguously available to researchers, and therefore, the false-positive and false-negative rates that we report are likely underestimates of the performance of these methods with real data. This approach allowed us to focus on how sampling design and statistical analysis affected the ability of the methods to differentiate selected and neutral loci in a best-case scenario. This study demonstrates how one may gain or lose power to detect a true outlier depending on sampling and how the test statistic controls for population structure.

We found that GEA tests tended to have more power in the island model and that outlier tests had more power under isolation by distance—but the power of a particular statistic in a particular demography was typically dependent on the details of the sampling design. Overall, however, paired sampling strategies tend to have power greater than or equal to that of random or transect sampling strategies.

Methods

Landscape simulations

We used the landscape simulator previously described in Lotterhos & Whitlock (2014) to create replicated data from a variety of demographic models. Briefly, the simulator uses forward-in-time recurrence equations to model the evolution of independent haploid SNPs on a quasi-continuous square landscape. We modelled four demographic histories that resulted in the same overall neutral F_{ST} (equal to approximately 0.05) for each demography, but the distribution of F_{ST}'s around that mean was determined by demographic history. The four demographic histories we simulated were as follows: equilibrium isolation by distance (IBD-eq), nonequilibrium isolation by distance due to expansion from one (1R) or two (2R) refugia, and the island model at equilibrium (IM, Lotterhos & Whitlock 2014).

Parameters for each demography were chosen to give the same overall neutral $F_{ST} = 0.05$. For IBD and the refugia cases, the landscape size was 360 × 360 demes. Migration was determined by a discretized version of a Gaussian dispersal kernel with standard deviation $\sigma = 1.3$ demes, and carrying capacity per deme (K) equalled 16, 71, and 124 for IBD, 1R and 2R, respectively. The starting location for the refugia for 1R was centred at $x,y = [180, 0]$ and for 2R at [280, 0] and [0, 70]. IBD was run until equilibrium at 10 000 generations, but 1R and 2R were only run for 1000 generations, to mimic the many natural populations whose ranges have expanded since the last glacial maximum (Hewitt 2000). In the IM, migration among all demes was equal irrespective of location on the landscape: the migration rate was 0.01, the carrying capacity per deme (K) equalled 960, and the simulation was run until equilibrium at 5000 generations. The landscape size in the island model was 72 × 72 demes (deme size was chosen to equal neighbourhood size in the IBD model, see Lotterhos & Whitlock 2014). In all demographic histories, individuals that dispersed off the edges of the landscape had zero fitness.

We modelled a total of 3 replicate data sets per scenario consisting of 9900 neutral loci and 100 selected loci. For each replicate data set, an environment was randomly generated based on a spherical model (a function that describes the decay of covariance or autocorrelation in the environment with distance) using the gstat function in R. The gstat function was used to generate a random selective landscape that had a variogram (the function of variance of the difference in a pair of points as a function of distance) with a given shape. This allows different patterns of longitudinal, latitudinal or distance-based changes in the environment at different grains (details in Lotterhos & Whitlock 2014). All selected loci adapted to this landscape. Landscapes could be described as 'weak clines': over the entire landscape, there was a north–south trend in the environments, but at smaller spatial scales, there were nearby sites that differed substantially in their environments (Fig. 1).

Selected loci evolved assuming a linear relationship between the selection parameter on the landscape (s_L) and the environment (Lotterhos & Whitlock 2014). Demographic dynamics were independent of the strength of selection. The selection coefficient describing the proportional increase in fitness of one allele over the other in a deme was calculated as the product of half the number of standard deviations from the environmental mean times s_L. In other words, the selection parameter (s_L) indicates the strength of selection in

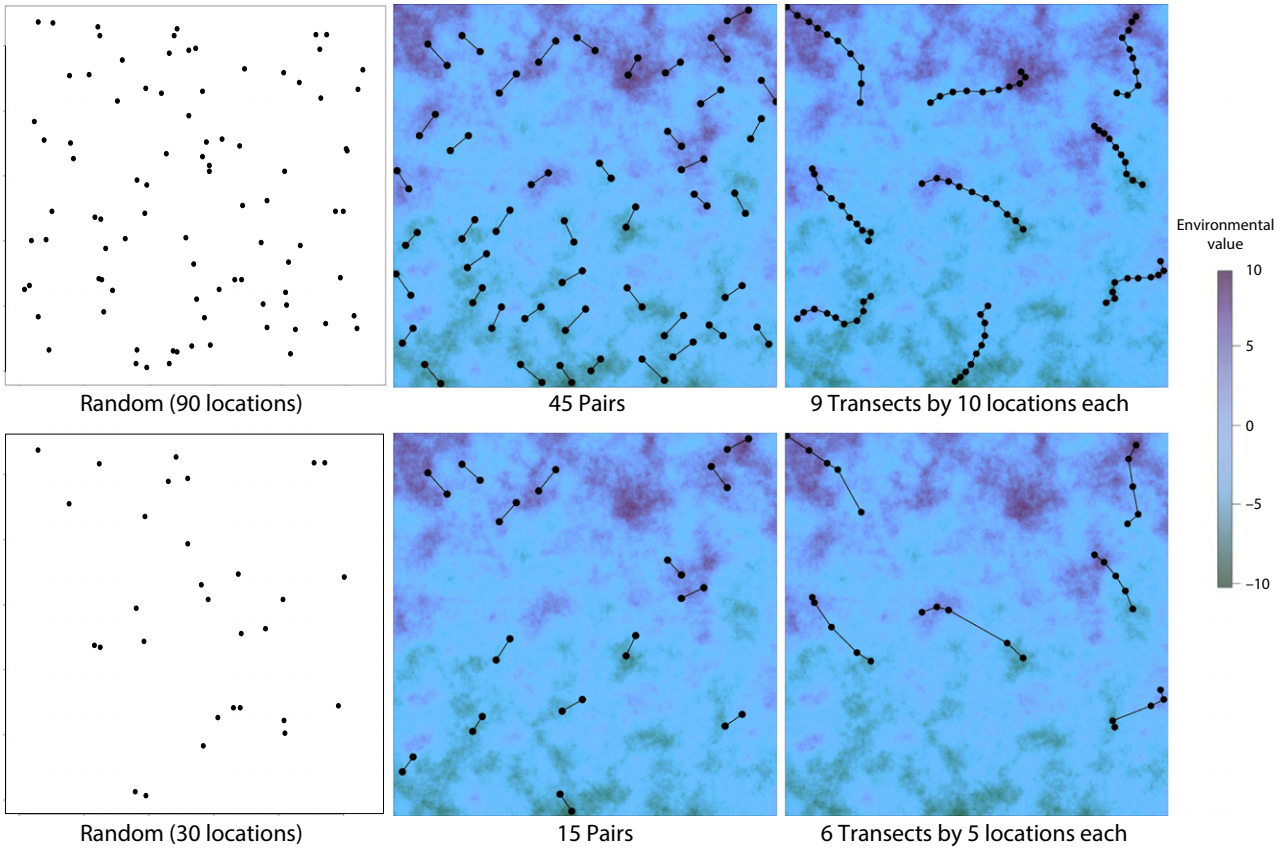

Fig. 1 Examples of sampling schemes on the same simulated landscape, where the colours represent values of the environment. Random samples were chosen without regard to the environment. Pairs and transects were chosen based on geographical proximity but with contrasting environments, as described in the main text. Initially, 90 locations were chosen for each design (top row). Locations were randomly chosen from the set of 90 to make data sets with fewer locations sampled (bottom row).

favour of the focal allele in an environment that is two positive standard deviations above the mean environment, and it is also the strength of selection *against* the focal allele in an environment that is two standard deviations below the environmental mean. Thus, s_L represents the near-maximum (or $-s_L$ the near-minimum) selection coefficient on the landscape.

The efficacy of selection was not equivalent in the four demographic histories (Fig. 2 left column, Lotterhos & Whitlock 2014). All selected loci showed response to selection in the IBD, but this was not the case for the other demographic histories. Specifically, loci with $s_L = 0.001$ did not show a response to selection in the IM, 1R and 2R scenarios, and so we did not include these loci in the data set because we were only interested in the power to detect loci which showed a response to selection (Fig. 2 left column). A total of 100 selected loci were modelled for each replicate. For IBD, the data set included four strengths of selection in each demography at the following percentages: $s_L = 0.001$ (40% of the loci), $s_L = 0.005$ (30%), $s_L = 0.01$ (20%) and $s_L = 0.1$ (10%). For IM, 1R and 2R, the data set included three strengths of selection at the following percentages: $s_L = 0.005$ (50%), $s_L = 0.01$ (33%) and $s_L = 0.1$ (17%). Note that even with these criteria, which were intended to make results comparable across cases, the response to selection is not equivalent among demographic histories (Fig. 2 left column). Thus, differences among methods should be interpreted by comparing patterns within demographic histories. Differences among demographic histories should be compared within a method.

Sampling designs

For each replicate simulation, we subsampled the landscape for 20 sampling designs, for a total of 240 data sets (Table 1, 20 designs * 4 demographic histories * 3 replicates). Sampling strategies were divided into three major categories: random, paired and transects. For random designs, 90 locations were chosen from random x and y coordinates on the landscape. For paired designs, sampling locations were chosen in pairs. The first point

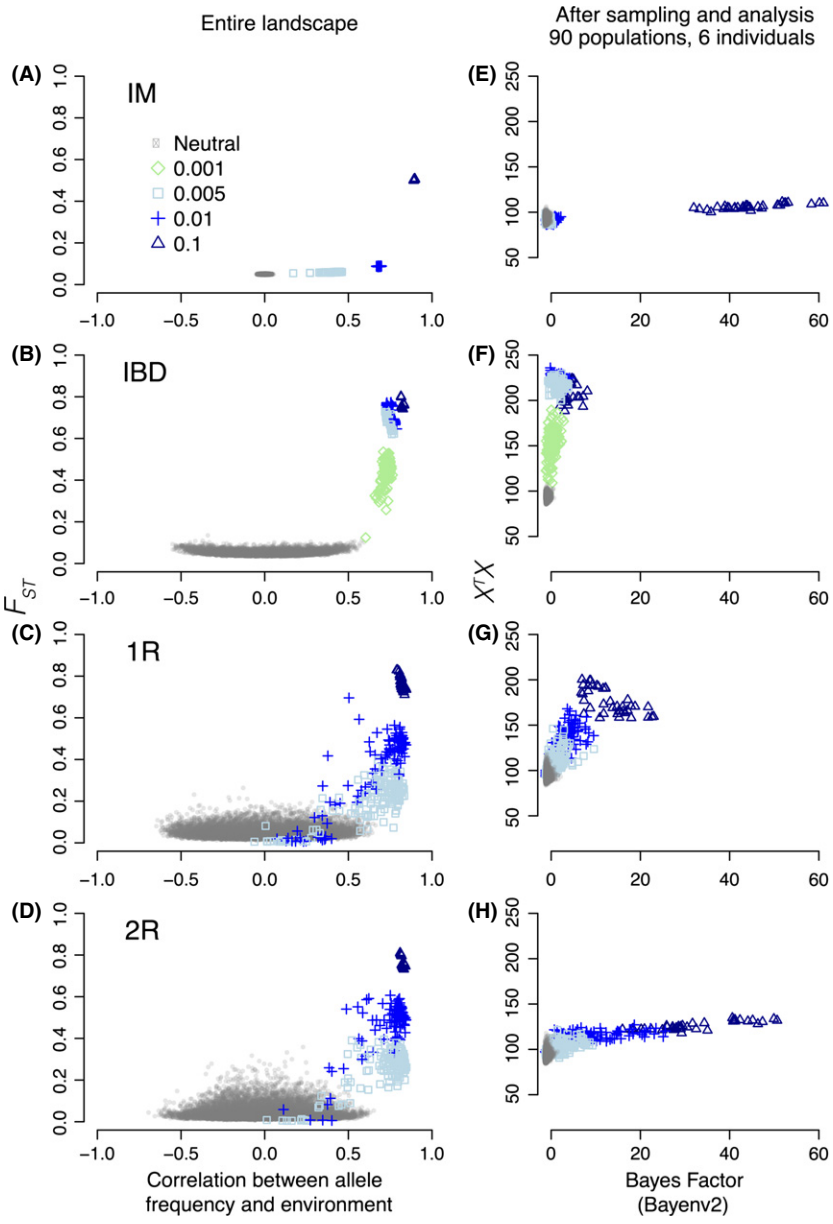

Fig. 2 A comparison between the distribution of statistics calculated using the true allele frequencies and all demes on the landscape (left column), and the distribution of analogous statistics after sampling and analysis (right column). Demographic histories are shown in rows (IM, island model; IBD, isolation by distance; 1R, expansion from one refuge; 2R, expansion from two refugia). In both columns, neutral loci are grey dots and selected loci are coloured according to the strength of selection on the landscape (s_L, as described in Methods). In the left column, true allele frequencies from all demes were used to calculate Weir and Cockerham's 1984 F_{ST} (y-axis) and the correlation between true allele frequencies and the environment (x-axis, note that for the purposes of plotting, selected loci were coded to have positive associations). In the right column, the sampling scheme used was 90 random populations with 6 individuals per population, and Bayenv2.0 was used to calculate X^TX (an F_{ST} analog) and Bayes factors representing the strength of association between sampled allele frequencies and the environment, after controlling for population structure. Note that the relative power of these statistics can change with total sample size (Fig. S1, Supporting information).

in the pair was randomly chosen on the landscape (with the criteria that it had to be more than 20 demes from another pair). The second point in the pair had to meet the following criteria: at least 1.5 standard deviations difference in the environment, between 10 and 20 demes away from its pair and at least 20 demes away from another pair. This process was repeated until 45 pairs were obtained (90 locations, Fig. 1). For transect designs, the landscape was divided into 9 sections (3 × 3), and in each section, one transect was determined along 10 random locations between a high and low value of the environment in that section (for a total of 90 locations on the landscape, Fig. 1). A single transect was 50–100 demes long (Fig. 1).

Once a complete set of random, paired and transect sites were chosen, these were randomly subsampled to create data sets with fewer locations (30 or 60 locations on the landscape, Fig. 1, bottom row). For transects, the populations with the most extreme environmental values on that transect were retained for subsampling. In each data set, locations were randomly sampled for individual genotypes using a binomial sampler and the allele frequency at that location. Note that our experimental design allows us to compare the effects of design (random, pair, transect), number of locations (30, 60, 90), number of individuals per location (6, 20, 30, 60) and total number of individuals collected (540, 600, 1200, 1800) (Table 1).

Table 1 Sampling designs for each data set. The random locations used were the same for each replicate environment, but the pairs and transects were specifically chosen based on values of the environment

	6 individuals per location	20 individuals per location	30 individuals per location	60 individuals per location
30 locations		***600 individuals*** 30 random locations 15 pairs 3 transects × 10 each 5 transects × 6 each		***1800 individuals*** 30 random locations 15 pairs 3 transects × 10 each 5 transects × 6 each
60 locations		***1200 individuals*** 60 random locations 30 pairs 6 transects × 10 each	***1800 individuals*** 60 random locations 30 pairs 6 transects × 10 each	
90 locations	***540 individuals*** 90 random locations 45 pairs 9 transects × 10 each	***1800 individuals*** 90 random locations 45 pairs 9 transects × 10 each		

Statistics

We evaluated power for two general ways to control for population structure: with a covariance matrix (Günther & Coop 2013) or with latent factors (Frichot *et al.* 2013; Duforet-Frebourg *et al.* 2014) (Table 2). We evaluated two different outlier approaches and three GEA methods in total (Table 2). Power for all methods was evaluated based on empirical *P*-values, which gives similar false-positive rates among methods (as described below). As F_{ST} analogs, we used X^TX from Bayenv2 (Günther & Coop 2013) and Bayes factors from PCAdapt (Duforet-Frebourg *et al.* 2014). Previously, we showed that based on empirical *P*-values, X^TX has similar or superior power to other genome scans for F_{ST} (Lotterhos & Whitlock 2014). Since that time, PCAdapt has been developed and shows promise, so we included it here. For GEA tests, we evaluated two statistics from Bayenv2 (Bayes factors and Spearman's ρ) and the $|z|$-score statistic from LFMM (Frichot *et al.* 2013). These statistics are explained in more detail below.

Statistics based on controlling for demographic history with a covariance matrix. Bayenv2.0 corrects for evolutionary nonindependence among samples with the variance–covariance matrix in allele frequencies (Günther & Coop 2013). From Bayenv2.0, we evaluated the power of X^TX (an analog of F_{ST}), log-transformed Bayes factors (a value that measures the strength of evidence in favour of a linear relationship between allele frequencies and the environment after population structure is controlled for) and Spearman's ρ (a GEA statistic that measures the nonparametric correlation between allele frequencies and the environment after population structure is controlled for). Each statistic was converted to empirical *P*-values based on known neutral loci, and then, empirical *P*-values were converted to *q*-values for the assessment of significance. Bayenv2.0 is implemented in two steps: in the first step, the variance–covariance matrix is calculated from allelic data. We used the variance–covariance matrix from the final run of the MCMC after 10^5 iterations. In the second step, the variance–covariance matrix is used to control for evolutionary history in the calculation of the statistics for each SNP (again using 10^5 iterations of the MCMC). Although it is possible to estimate the covariance matrix in Bayenv2.0 with just neutral loci, we used all loci to estimate the covariance matrix so that results would be comparable to the latent factor models, which do not have this advantage.

Statistics based on controlling for demographic history with latent factors. Latent factor models require the user to specify the number of latent factors, which should reflect to some approximation the number of populations in the data. LFMM is a GEA method that estimates allele–environment correlations while simultaneously correcting for population structure with latent factors (Frichot *et al.* 2013). These types of factor models are closely related to principal components approaches. For each locus, LFMM outputs a $|z|$-score for the regression coefficient and converts the $|z|$-score into a probability based on a Gaussian distribution. To implement LFMM, the number of latent factors (K) must be specified by the user. As recommended by the authors, we based our choice of K on a Tracy–Widom test (Patterson *et al.* 2006). We implemented LFMM with a 200-iteration burn-in and 1200 iterations of the Gibbs Sampler. For comparison to other programs, we used empirical *P*-values based on the $|z|$-score and converted them to *q*-values as previously described.

PCAdapt is an outlier test that also infers population structure with latent factors, but jointly determines population structure and outlier loci (Duforet-Frebourg

Table 2 : Three GEA statistics and two genetic differentiation outlier measures compared in this study. 'Method of controlling for population structure' indicates how evolutionary non-independence is controlled for in the test statistic. GEA: genetic–environment association

	Method of controlling for population structure	
	Covariance matrix (Bayenv2, Günther & Coop 2013)	Latent factors
GEA test	Bayes factors Spearman's ρ	| *z* |-scores (LFMM, Frichot *et al.* 2013)
F_{ST} outlier	X^TX	Bayes factors (PCAdapt, Duforet-Frebourg *et al.* 2014)

et al. 2014). The method uses a hierarchical Bayesian model that seeks for loci atypically explained by one of the K factors. Also, the method allows the user to optionally scale the SNPs so that they all have a variance equal to one, which may decrease the false-positive rate in certain demographic histories ('scaling' parameter) (Duforet-Frebourg *et al.* 2014). The program outputs a log-transformed Bayes factor for each locus, which is a measure of evidence in favour of selection over neutrality across all factors. We ran PCAdapt with a burn-in of 200 steps and 10 000 steps in the MCMC, and $K = 10$ with scaling. Lower latent factors (i.e. $Z = 1$) should explain the main features of population structure, while higher latent factors (i.e. $Z = 10$) should explain finer scale aspects of population structure. For comparison to other programs, we used empirical P-values based on the Bayes factor and converted them to q-values as previously described.

Using empirical P-values to compare methods and sampling designs

Our goal in this study was to compare the ability of different sampling designs and analysis methods to separate selected loci from neutral loci. For the comparison of power (the number of true-positive selected loci divided by the total number of selected loci in the data set) among all programs, we calculated empirical P-values for each statistic. An empirical P-value uses the test statistics of all known neutral loci (e.g. X^TX or a Bayes factor) in the data set to generate a null distribution. Using this null distribution, P-values for each possible selected locus can be calculated from its cumulative frequency in this null distribution [thus, selected loci that fall outside the neutral distribution have an empirical P-value of 1/(number of neutral loci)]. The rationale for using the empirical P-value approach is because both sampling and the statistical method will affect the distribution of selected loci relative to neutral loci. This is illustrated by comparing the left column in Fig. 2 (the true distribution of statistics across the entire landscape) to the right column in Fig. 2 (the distribution of two analogous statistics after sampling and analysis, shown for all sampling designs).

The advantage of the empirical P-value approach is that it reveals the relative distribution of neutral to selected loci of each test statistic. Note that we do not evaluate false-positive rates in this study, because they are not very meaningful for empirical P-values. Previously, we showed that evaluating false-positive rates based on output P-values or Bayes factor cut-offs resulted in very high false-positive rates—even though all statistics had similar power to separate neutral and selected loci based on empirical P-values (Lotterhos & Whitlock 2014). We found that latent factor models suffered a similar problem based on author recommendations (K. E. Lotterhos, *in prep*). We chose to focus on empirical P-values because there is no given criterion for rejection in Bayenv2, and also because statistics would be compared on equal grounds. Our comparisons based on empirical P-values (i.e. calculated based on a set of truly neutral loci) represents a best-case scenario that eliminates several sources of both false-positive and false-negative errors. As such, it is useful for comparing the information content of various sampling designs, and this approach shows the relative information content of the test statistics calculated by the different methods. It does not, however, capture the true error rates of these methods, which in any case do not always provide precise criteria for decision-making.

Thus, in comparing statistics, our results focus mainly on the comparison of the maximum possible power of the program and the sampling design to discriminate selected loci from neutral loci. We converted empirical P-values into q-values using the q-value package in R (Storey & Tibshirani 2003). For all of the following analyses where q-values were used to assess significance, significance was based on a q-value cut-off of 0.01.

In practice, it will be impossible to create empirical P-values as reliable as those used in this study because noncoding regions may be under selection (Guttman *et al.* 2009), and because of linkage between neutral and selected loci (Charlesworth *et al.* 1997). In comparing methods and sampling designs, however, empirical P-values are a useful way to measure their power on equal grounds.

Results

Effect of sampling and statistical analysis on the distribution of test statistics

The distribution of allele–environment correlations vs. F_{ST}, using the true allele frequencies for all 129 600

demes on the landscape, is shown on the left column of Fig. 2 (coloured by the strength of selection on the landscape, s_L). These plots correspond to the distribution of test statistics before sampling and analysis. The IM had a very clear separation between neutral and selected loci along the allele–environment correlation axis, but little separation in F_{ST} for $s < m$ (the latter is expected by theory). The IBD and refugia scenarios had clear separation between neutral and selected loci along both axes (Fig. 2B–D), although the variance in the allele–environment correlation for neutral loci was much larger than that for the IM.

The distribution of Bayes factors vs. X^TX from Bayenv2 is shown on the right side of Fig. 2 for the 90 random populations and 6 individuals/population—this corresponds to the distribution of test statistics after sampling and analysis. The variance of neutral Bayes factors was much smaller than it was before analysis, due to the correction for population structure (compare spread of grey dots along the *x*-axis on right side of Fig. 2 to the left side). Yet, at the same time, there was less statistical separation between neutral and selected loci along this axis for the refugia scenarios, and to a greater extent for IBD.

Although much of the variation in the distribution of statistics after analysis was due to the strength of selection (Fig. 2), total sample size (total number of individuals) could also substantially change the distribution of these test statistics (Fig. S1, Supporting information). For X^TX, power increased substantially with total sample size in the IM and 2R models, but increasing total sample size had little effect on the power of Bayes factors (Fig. S1, Supporting information). Generally, however, total sample size greatly affected power for all statistics. Power increased when the number of individuals sampled per location was increased (Fig. S2, Supporting information) or the number of locations sampled was increased (Fig. S3, Supporting information)—although in some cases, the gain in power was only moderate compared to the gain in power from using a paired sampling strategy (see below).

Whether the allocation of individuals into populations affected power also depended on total sample size and the statistic employed. For X^TX, allocating 540 individuals into 90 random populations had higher power than allocating 600 individuals into 30 random populations—but allocation had little effect on the distribution of Bayes factors (Fig. S1, Supporting information). However, for a larger total sample size of 1800 haploids (approximately equivalent to 900 diploids), we found little benefit of changing the allocation of individuals into populations (i.e. more individuals in fewer populations or fewer individuals into more populations, Fig. S4, Supporting information). With this large sample size, power was improved when individuals were allocated into more populations in the IM (a result also noted by De Mita *et al.* 2013 for the IM), but for the other demographic histories, there was no clear benefit in the way individuals were allocated (Fig. S4, Supporting information).

Relative power of GEA tests depends on total sample size

We used the true-positive rates of Bayes factors from Bayenv2.0 (a GEA method) as a standard baseline to compare the true-positive rates to the other two GEA approaches. These 'power–power' plots allow easier visualization of the context-dependent performance of different statistics. We compared the power of statistics for a sampling design of 90 locations and 6 individuals per location (total 540) to 90 locations and 20 individuals per location (total 1800) (Some of the differences in power within this design came from differences in the random, pairs and transects).

For both sampling designs, Spearman's ρ had lower or equal power than Bayes factors in Bayenv2.0 (Fig. 3, top row). Spearman's ρ generally had lower power than all the statistics we evaluated, and it is not considered further in the other comparisons.

However, the sampling design × demography interactions became evident when comparing the $|z|$-scores from LFMM with Bayes factors from Bayenv2.0 (Fig. 3, bottom row). When 90 locations and 6 individuals were sampled, LFMM had higher power for the IM, but Bayes factors had slightly higher power for 2R (Fig. 3C). But when the number of individuals per location was increased, LFMM had higher power for IBD, and the two statistics had similar power for the other demographic histories (Fig. 3D). Overall, there was relative increase in power for LFMM with total sample size, which came from an increase in the power of LFMM and little change in the power of Bayenv2.0.

Relative power of F_{ST} outlier tests depends on sample size

In the same manner as above, we compared PCAdapt to X^TX from Bayenv2.0 (both F_{ST} analogs, Fig. 4) and observed again sampling design × demography interactions. When 90 locations and 6 individuals were sampled, PCAdapt had higher power for 2R and X^TX had higher power for 1R (Fig. 4a). When sample size per population was increased to 20, PCAdapt had higher power for the IM and X^TX had higher power for 1R and 2R (Fig. 4b).

Relative power of GEA vs. F_{ST} outlier tests depends on demographic history

The overlap among methods for the detection of locally adapted loci is summarized in a Venn diagram in Fig. 5, which represents the average number of loci found by each statistic over all sampling designs (out of a total of approximately 100 selected loci in each design—note that in the refugia models, a few selected loci were excluded from analysis because of low polymorphism). LFMM found most of the detectable selected loci in the IM, while either X^TX or PCAdapt found most of the selected loci for IBD (Fig. 5). In the refugia scenarios, all four statistics on average found about a third to a half of the selected loci, but the remainder of the selected loci were spread among the methods (Fig. 5).

We explored the variation in power among methods to detect weak selection on the landscape (s_L = 0.005) compared to strong selection on the landscape (s_L = 0.1) (Table 3). For loci under weak selection, very few were found by all four methods (0%–36% across demographic histories), and most that were detected were found by two methods or less (11–50%, Table 3). For loci under strong selection, the majority were found by all four methods (84–97%, Table 3).

Power of sampling strategies: random, transects or pairs?

Next, we compared random, transect and paired sampling strategies for otherwise equivalent designs. When 90 locations and 6 individuals per location (total 540) were sampled, the paired strategy generally had equal or higher power than the random or transect strategies (Fig. 6). Interestingly, pairs increased power the most for GEA methods under IBD (Fig. 6B, C), which was the demography in which GEA approaches had generally low power (Fig. 5). The trend for pairs to have equal or higher power persisted for larger total sample sizes (not shown), with the exception that the paired strategy did not show a clear benefit with PCAdapt (Fig. 6D).

To further explore the source of increased power, we plotted the power to detect loci that evolved to different strengths of selection on the landscape (s_L). For Bayes factors from Bayenv2.0, paired sampling increased the power to detect loci under weak-to-moderate selection (Fig. 7A–B), while sampling strategy did not matter for loci under strong selection because they were always detected (Fig. 7C). This trend for paired sampling designs to increase the power to detect loci that evolved under weak selection was also true for LFMM (especially in IBD, Fig. S5 left column, Supporting information) and X^TX (especially in 1R and 2R, Fig. S5 middle column, Supporting information)—but not for PCAdapt, whose power remained relatively unaffected by the location of populations (Fig. S5 right column, Supporting information).

Discussion

In this study, we found complicated trade-offs between sampling designs for F_{ST} outlier and GEA genome

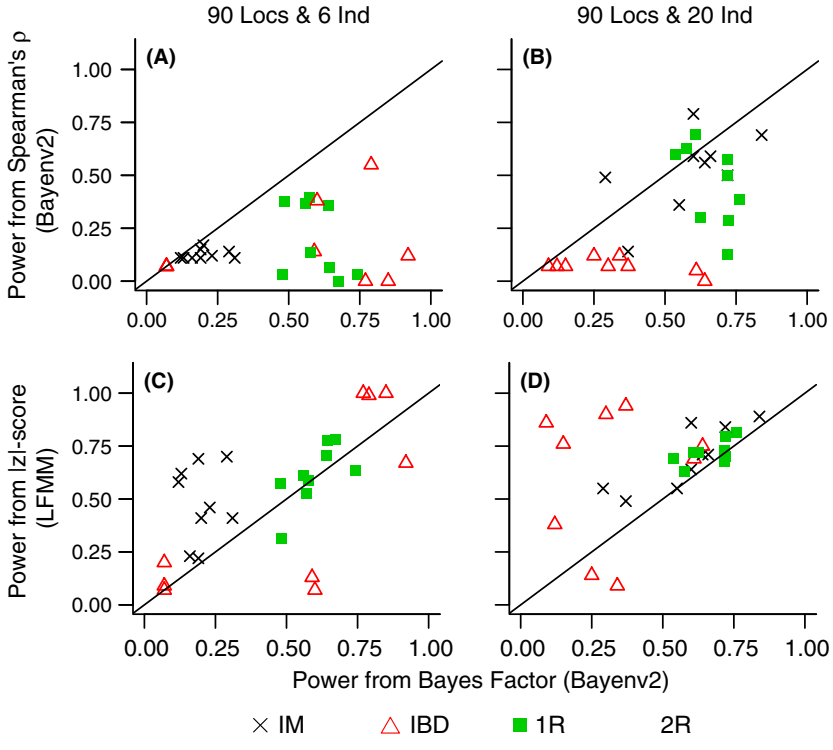

Fig. 3 Comparison of true-positive rates (power) between the three genetic–environment association statistics evaluated in this study (in rows), for two total sample sizes (in columns, 540 haploids on the left and 1800 haploids on the right), and different demographic histories (IM, island model; IBD, isolation by distance; 1R, expansion from one refuge; 2R, expansion from two refugia). The top row shows the relative power of Bayes factors to Spearman's ρ from Bayenv2.0. Spearman's ρ generally had low power compared to all statistics in our simulations. The bottom row shows the relative power of Bayes factors from Bayenv2.0 and |z|-scores from LFMM. Points are coloured by demography; some of the variation within a demography comes from the way populations were distributed across a landscape (random, paired or transect).

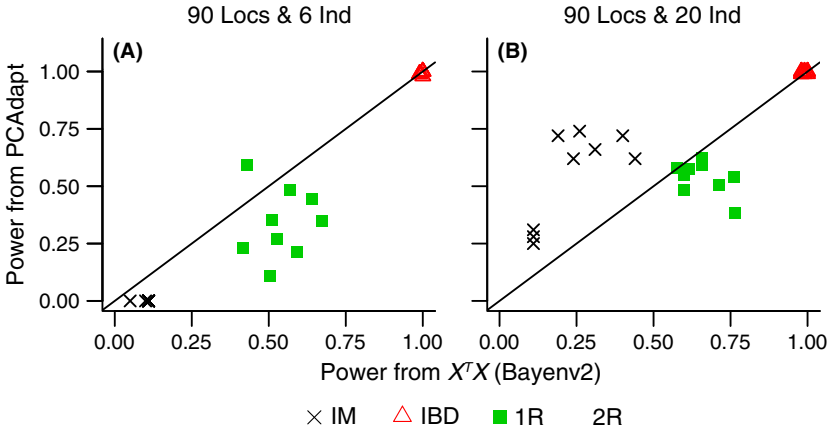

Fig. 4 Comparison of true-positive rates (power) between two F_{ST} outlier tests: X^TX from Bayenv2.0 and Bayes factors for PCAdapt ($K = 10$ with scaling). (A) 90 locations with 6 individuals each (540 haploids); (B) 90 locations with 20 individuals each (1800 haploids). Points are coloured by demography; some of the variation within a demography comes from the way populations were distributed across a landscape (random, paired or transect). IM, island model; IBD, isolation by distance; 1R, expansion from one refuge; 2R, expansion from two refugia.

scans on a landscape. Because we used empirical *P*-values based on known neutral loci, our results reflect the degree to which a method and sampling design are able to discriminate selected loci from a distribution of neutral loci. Although we found that a paired sampling strategy was often more powerful than a random or transect strategy, we also found that the relative power of any two methods depended on the total number of individuals sampled and demographic history.

Advantages and disadvantages of GEA and outlier tests

Genetic–environment association and outlier tests have different advantages. One advantage of GEA methods is that they can pinpoint the environmental axis that is driving selection—but results should still be interpreted with caution because there are many abiotic and biotic variables that are autocorrelated on the landscape. However, the selective environment is not always known. An advantage of outlier tests is that they may uncover selected loci without knowledge of the selective environment or phenotype under selection. We compared GEA approaches to outlier tests within a method and sampling design, which reflected how demographic history affected the power of these two types of genome scans.

Using a comprehensive set of simulations, De Mita *et al.* (2013) concluded that GEA tests were generally more powerful than F_{ST} outlier tests. In our simulations, however, we found that GEA methods were only the most powerful method in the island model, which was consistent with the strong separation between neutral and selected loci along this axis before sampling and analysis. Under isolation by distance, GEA tests had surprisingly low power—despite a response to selection on the landscape. In the IBD, the correlation between neutral allele frequencies and the environment had much higher variance than they did under the IM. By correcting for population structure in IBD, the GEA methods that we tested lost a surprising amount of power. Such a loss of power has also been noted in hierarchical coalescent simulations with strong correlations between demography and environmental variables (de Villemereuil *et al.* 2014). Interestingly, GEA approaches did not suffer such a significant loss of power in the range expansion cases compared to IBD, even though in the refugia cases, there was similar variance in the allele–environment correlation for neutral loci and greater variance for selected loci. This suggests there is some spatial aspect to the pattern of neutral population structure that is more accurately estimated and/or controlled for in the range expansion cases, which had stronger patterns of neutral structure than IBD.

While the importance of controlling for population structure in GEA and F_{ST} outlier tests is well recognized, for some methods, power may be limited when neutral population structure corresponds with the environmental axis of interest. Such may be the case when neutral gene flow is higher among similar environments ('isolation by environment'), which may occur by assortative mating or prezygotic isolating mechanisms (such as timing of reproduction in the environment, e.g. mismatches in flowering time), by postzygotic selection against or genetic incompatibilities among immigrants or hybrids, or via nonadaptive processes that lead to nonrandom gene flow (Bierne *et al.* 2011; Sexton *et al.* 2013; Shafer & Wolf 2013). In these cases, the power of GEA methods may be relatively low in comparison with F_{ST} outlier tests.

In the refugia scenarios, both GEA and F_{ST} outlier tests could detect loci that evolved under strong selection, but these types of methods had differing abilities to detect loci that evolved under weak selection. These results are encouraging because they indicate that the common practice of identifying outliers that overlap among different methods may reduce

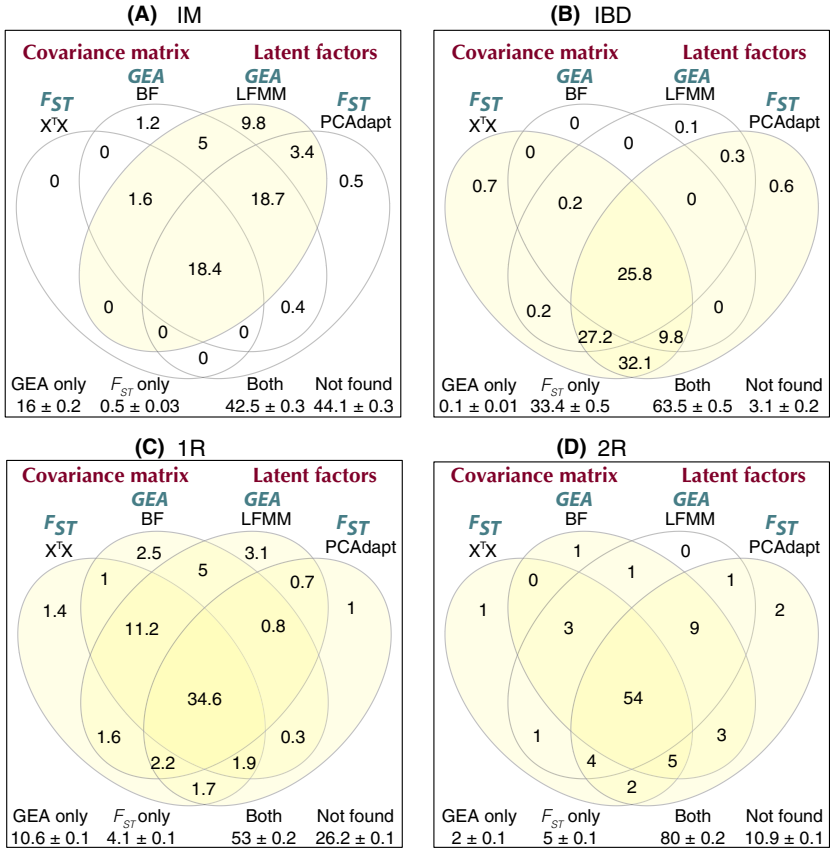

Fig. 5 Venn diagrams representing the overlap in the average number of selected loci (out of 100) found by each method, averaged over all sampling designs: (A) island model, (B) isolation by distance, (C) expansion from one refuge, (D) expansion from two refugia. X^TX and *BF* (Bayes factors from Bayenv2.0) correct for population structure with a covariance matrix, while LFMM and PCAdapt use latent factors. In cases when one or two methods outperformed the others, those methods are highlighted. The numbers at the bottom summarize the average number of loci found by either GEAs, F_{ST} analogs, both or neither.

Table 3 Comparison of the overlap in statistics to detect loci under weak selection (s_L = 0.005) compared to strong selection (s_L = 0.1)

Category	Strength of selection	IM	IBD	1R	2R
% loci found by two methods or less	weak	11%	50%	49%	20%
	strong	1%	0	0	0
% loci found by three methods	weak	12%	39%	2%	28%
	strong	15%	3%	1%	4%
% loci found by all four methods	weak	0%	7%	14%	36%
	strong	84%	97%	93%	96%
Not found	weak	77%	4%	35%	16%
	strong	0%	0	6%	0

false-positive rates. They are also discouraging, however, because they demonstrate that this common practice will miss weakly selected loci, which are probably a majority of the loci under selection. Although power can be increased by considering outliers shown by at least one method, this might come at the cost of increased false-positive rate because *P*-values or arbitrary Bayes factor cut-offs from these methods are unreliable (Lotterhos & Whitlock 2014, K. E. Lotterhos, *in prep.*).

Using a covariance matrix vs. latent factors to control for neutral structure

Although the power of various methods has been previously explored (De Mita *et al.* 2013; Jones *et al.* 2013; de Villemereuil *et al.* 2014), this study is the first to document demography × sample size interactions in the relative power of methods based on a covariance matrix vs. latent factors to control for neutral population structure. We found that Spearman's ρ, the nonparametric

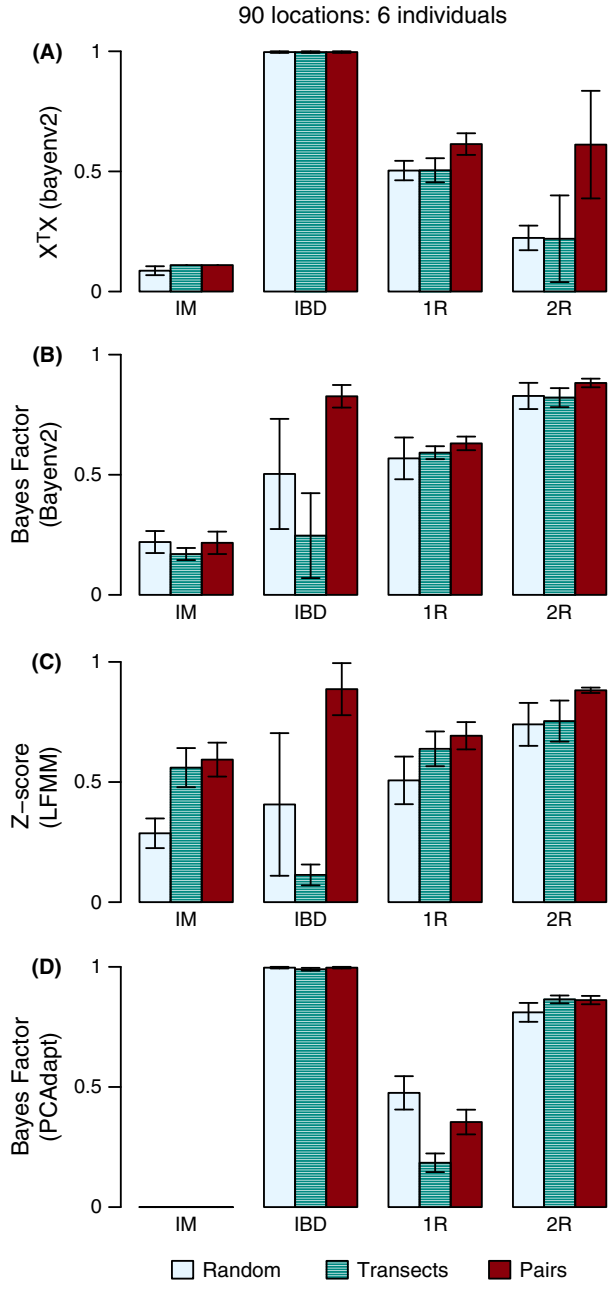

Fig. 6 Comparison of random, transect and paired sampling designs for 90 locations with 6 haploid individuals each (540 total), for four of the statistics evaluated in this study.

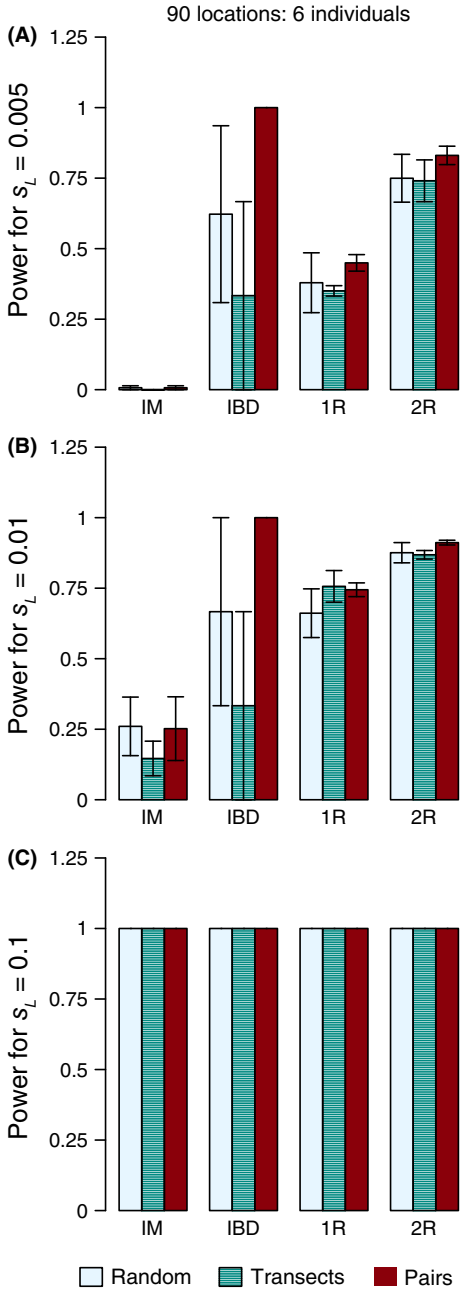

Fig. 7 Power to detect the landscape selection parameter (s_L) for Bayes factors from Bayenv2.0 as a function of sampling design (random, transect or pair) and demographic history. Analogous plots for the other statistics are shown in Fig. S5 (Supporting information).

statistic, had lower power than the other methods we tested. It remains to be determined whether this is a general phenomenon or whether this is result of the spatial pattern of selection in our simulations. For the other statistics, as our comparisons were based on the same data sets, demography × sample size interactions were a result of differences in the statistical model employed. We found the change in relative power with total sample size usually only occurred in particular demographic histories. Typically, the power of latent factor models increased with the number of individuals per population, which may have occurred because in Bayenv2.0, populations are the units of analysis, whereas LFMM is individual based. Even though in many scenarios, these statistics performed similarly based on empirical *P*-values, the false-positive rates for

the latent factor models were undesirably high based on author recommendations (K. E. Lotterhos, *in prep*).

These demography × sample size interactions will make it difficult for the empirical biologist to choose among methods. Demographic history of a population is rarely known with accuracy, and in the rare case when it is accurately known, the effect of the number of individuals sampled on the power of different methods would have to be evaluated with simulations.

A paired sampling strategy can increase power to detect weak selection

Our study was the first to compare random, paired and transect sampling strategies for their power to find genes responsible for local adaptation. Pairs had the highest power to detect genes with weak selection because they were designed to maximize the difference in adaptive environment while minimizing the differences in evolutionary history. Although transects were designed under a similar rationale, some samples in transects were inefficiently used because the power came mainly from the samples in the transect that have the greater difference in adaptive environment, with little information coming from the populations in the middle of the transect with intermediate environments. Additionally, transects were designed at a slightly larger spatial scale than pairs, which may have affected their power. At a slightly larger scale, more neutral divergence is expected, but at the same time, migration is lower and selection is more efficient. Thus, choosing the appropriate spatial scale will be critical part of employing paired sampling designs.

In populations adapting to spatially heterogeneous environments and with complicated patterns of dispersal, the spatial scale of local adaptation is often unknown. This uncertainty has consequences for paired sampling strategies if pairs are chosen with respect to the environment and not with respect to the scale of migration–selection balance. Such was the case in our simulations, in which pairs were chosen with respect to the environment: individual pairs sometimes had low power to detect local adaptation (results not shown). Paired sampling designs based on the scale of local adaptation from common garden or reciprocal transplant experiments (reviewed in Hereford 2009) will have higher reliability than designs based on values of the environment alone. A new Bayesian hierarchical method may be useful in the analysis of some paired designs (Foll *et al.* 2014), but caution should be used in applying this method because isolation-by-distance scenarios (such as those simulated here) may violate the assumptions of the hierarchical model. Additionally, transect designs are still useful for answering questions about clines in allele frequencies and the spatial scale of local adaptation. How an investigator chooses to distribute populations on the landscape will depend on the question.

In general, however, total sample size influenced power much more than the distribution of populations on the landscape. Landscape genomic studies employing the methods evaluated in this study should strive to maximize the overall sample size.

Other limitations to the genome scan approaches

As described in the methods, in practice, it will be impossible to create empirical *P*-values as reliable as those used in this study because in natural systems, neutral loci cannot be reliably identified. Some recently developed methods require random sets of SNPs to be used as a null distribution to generate *P*-values (Berg & Coop 2014), which is analogous to the empirical *P*-value approach employed in this study. Such null distributions may be biased by the inclusion of putatively neutral regions, such as noncoding regions, which may be experiencing spatially heterogeneous selection (Guttman *et al.* 2009). As such, using a list of putatively neutral loci from a real organism will likely create a null distribution that is conservative, because the 'neutral' distribution would include the effects of some selection.

The error rates of these methods also can be affected by linkage between the studied markers and other loci that experience selection. Such linkage can generate false positives, either because of background selection or genetic hitchhiking. With background selection, purifying selection on linked sites can reduce the effective size of the focal nucleotide (Charlesworth 1994) and therefore increase differentiation. Such background selection is presumably stronger in coding regions, and therefore, a list of noncoding loci may provide a biased comparison for variation in coding regions. Such an effect would further increase the false-positive rate of all methods, especially methods that use the empirical *P*-value approach.

Genetic hitchhiking, where a neutral or weakly selected allele changes in frequency because of linkage to another allele that is experiencing selection, can also be a source of false positives. A neutral site may differentiate in frequency through indirect selection caused by variation at a nearby site. The extent to which hitchhiking affects the results of genome scans will depend on several evolutionary parameters including recombination rate, population size, the strength of selection, whether selection acts on a new mutations or on standing variation, how recent selection is, the number of loci and their effect sizes, as well as marker density across the genome and analytical method (Hermisson &

Pennings 2005; Oleksyk *et al.* 2010; Pritchard & Di Rienzo 2010; Tiffin & Ross-Ibarra 2014). Many natural populations of large effective size are characterized by rapid decay of linkage in the genome (e.g. long-lived trees with large N_e, Neale & Savolainen 2004), so it may be possible in these species to narrow the genomic window to the locus under selection, provided that marker density is high enough (Tiffin & Ross-Ibarra 2014). On the other hand, some species exhibit extensive linkage disequilibrium, which may decrease genomic resolution (i.e. Hohenlohe *et al.* 2012). Linkage disequilibrium, however, may be helpful at lower marker densities because it increases the probability of identifying regions under selection. Our results show there are challenges to the application of these methods even in the case without linkage to other selected loci; in reality, the false-positive rates will be greater than we show here because of the effects of linkage with other sites.

Genome scans such as GEA and F_{ST} outlier tests are just the first step to discovery, and they should be followed by more detailed analyses. These summary statistics throw away some important information, such as the class of SNP (coding vs. noncoding, amino acid substitution, annotation, etc.). When haplotype data are available or can be reconstructed (e.g. Stephens *et al.* 2001), statistics that describe substitution rates or the local extent of linkage disequilibrium can be powerful ways at detecting selection in certain biological scenarios (reviewed in Hohenlohe *et al.* 2010; Oleksyk *et al.* 2010). Genome scans can provide a list of candidate genes that are worthy of further functional, experimental and genomic study; understanding and improving the true power of these methods will help in assessing which and how many loci should be targeted for further study.

Concluding statements

Discovering the genes responsible for local adaptation is difficult. We have shown that a simple modification of the sampling design, where sampling locations are chosen in pairs that are geographically close but selectively diverged, can increase the power to detect locally selected loci. In some cases, the power gained from a paired sampling strategy may exceed the power gained by increasing the sample size. This increase in power may be intuitive, because within such pairs, neutral loci will not be greatly differentiated, but selected loci may be. It is this contrast between neutral and selected loci that determines the power of these tests.

Differentiation outlier tests and genetic–environment associations can detect some of the loci responsible for large amounts of local adaptation. In general, genetic–environment association tests that account for idiosyncratic patterns of neutral genetic relatedness, such as those considered here, have the most power to find such genes. However, which approach (genetic differentiation outliers vs. genetic–environment associations) is best depends on the pattern of demographic history for the populations and, to a lesser extent, the sampling design. Different methods often find different genes in empirical studies. We see that finding different loci under selection with different statistical frameworks is to be expected.

Acknowledgements

The editor and two reviewers made thoughtful comments that greatly improved the quality of this manuscript. We would also like to thank Loren Reisberg, Sam Yeaman, Sally Aitken and the Whitlock and Otto laboratory groups for useful discussions, comments and feedback on the ideas presented in this manuscript. This research is part of the AdapTree Project, funded by Genome Canada, Genome BC, Alberta Innovates Bio Solutions, the Forest Genetics Council of British Columbia, the BC Ministry of Forests, Lands and Natural Resources Operations, Virginia Tech, the University of British Columbia and the University of California, Davis, and it has been partially supported by a Discovery Grant from the National Sciences and Engineering Research Council (Canada).

References

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society London B: Biological Science*, **263**, 1619–1626.

Berg JJ, Coop G (2014) A Population Genetic Signal of Polygenic Adaptation. *PLoS Genetics*, **10**, e1004412.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, **20**, 2044–2072.

Bonhomme M, Chevalet C, Servin B *et al.* (2010) Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics*, **186**, 241–262.

Charlesworth B (1994) The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetics Research*, **63**, 213–227.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genetics Research*, **70**, 155–174.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

De Mita S, Thuillet AC, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, doi:10.1093/molbev/msu182.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.
Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.
Foll M, Gaggiotti OE, Daub JT, Excoffier L (2014) Hierarchical Bayesian model of population structure reveals convergent adaptation to high altitude in human populations. *ArXive*, arXiv:1402.4348v1.
Frichot E, Schoville SD, Bouchard G, Francois O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.
Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.
Guttman M, Amit I, Garber M *et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics*, **7**, 1–32.
Hereford J (2009) A quantitative survey of local adaptation and fitness trade-offs. *American Naturalist*, **173**, 579–588.
Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.
Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907–913.
Hohenlohe PA, Phillips PC, Cresko WA (2010) Using population genomics to detect selection in natural populations: key concepts and methodological considerations. *International Journal of Plant Sciences*, **171**, 1059–1071.
Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, **367**, 395–408.
Jones MR, Forester BR, Teufel AI *et al.* (2013) Integrating landscape genomics and spatially explicit approaches to detect loci under selection in clinal populations. *Evolution*, **67**, 3455–3468.
Joost S, Bonin A, Bruford MW *et al.* (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.
Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of theory of selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.
Li YF, Costello JC, Holloway AK, Hahn MW (2008) "Reverse ecology" and the power of population genomics. *Evolution*, **62**, 2984–2994.
Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, **23**, 2178–2192.
Lotterhos KE, Whitlock MC (2015) Data from: the relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Dryad*, doi:10.5061/dryad.mh67v.
Meirmans PG (2012) The trouble with isolation by distance. *Molecular Ecology*, **21**(12), 2839–2846.
Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. *Trends in Plant Science*, **9**, 325–330.
Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **365**, 185–205.
Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.
Pritchard JK, Di Rienzo A (2010) Adaptation–not by sweeps alone. *Nature Reviews Genetics*, **11**, 665–667.
Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution, and Systematics*, **43**, 23–43.
Sexton JP, Hangartner SB, Hoffmann AA (2013) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*, **68**, 1–15.
Shafer AB, Wolf JB (2013) Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. *Ecology Letters*, **16**, 940–950.
Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, **68**, 978–989.
Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440–9445.
Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, **29**, 673–680.
de Villemereuil P, Frichot E, Bazin E, Francois O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.
Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

K.E.L. and M.C.W. formulated the landscape simulator, the study questions, and the experimental design. K.E.L. implemented the simulations and analyses.

Data accessibility

The simulated data sets and summary statistics are archived on Dryad: doi:10.5061/dryad.mh67v.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 This graph corresponds to the right column of Fig. 2 in the main text, but with all sample sizes shown shown for the random sampling scheme.

Fig. S2 Effect of increasing the number of individuals sampled per location (averaged over random, paired, and transect designs).

Fig. S3 Effect of increasing the number locations sampled on the landscape for a sample size of 20 per population (averaged over random, transect, and paired designs).

Fig. S4 For the same total sample size (1800 individuals), we evaluated whether power depended on allocating individuals into more locations, or into fewer locations with more individuals per location.

Fig. S5 Power of random, transect, and paired designs to detect the landscape selection param-eter (S_L, by rows) for three statistics (by columns): $|z|$ scores from LFMM, X^TX from Bayenv2.0, and Bayes factors from PCAdapt.