Using genotype-environment associations to identify multilocus local adaptation

Brenna R. Forester[1], Jesse R. Lasky[2], Helene H. Wagner[3], Dean L. Urban[1]

1 – Duke University, Nicholas School of the Environment, Durham, NC 27708, USA.
2 – Pennsylvania State University, Department of Biology, University Park, PA 16802, USA.
3 – University of Toronto Mississauga, Department of Biology, Mississauga, ON, Canada

Keywords: constrained ordination, landscape genomics, natural selection, random forest, redundancy analysis, simulations

Corresponding Author: Brenna R. Forester, Box 90328, Duke University, Durham, NC 27708, Phone: (919) 613-8069, Fax: (919) 684-8741, Email: brenna.forester@gmail.com

Running title: Detecting multilocus selection

Abstract

Identifying loci under selection can provide insight into the mechanisms underlying local adaptation and inform management decisions for agricultural, natural resources, and conservation applications. Genotype-environment association (GEA) methods, which identify adaptive loci based on associations between genetic data and environmental variables, are particularly promising for distinguishing these loci. Univariate statistical methods have dominated GEA, despite the high dimensional nature of genomic data sets. Multivariate and machine learning methods, which can analyze many loci simultaneously, may be better suited to these large data sets since they consider how groups of markers covary in response to environmental predictors. These methods may also be more effective at detecting important adaptive processes, such as selection on standing genetic variation, that result in weak, multilocus signatures. Here we evaluate four of these methods, as well as a popular univariate approach, using published simulations of multilocus selection. We found that the machine learning method, Random Forest, performed poorly as a GEA. The univariate approach performed better, but had low detection rates for loci under weak selection. Constrained ordinations showed a superior combination of low false positive and high true positive rates across all levels of selection. These results were robust across demographic history, sampling designs, and sample sizes. Although further testing is needed on more complex genetic architectures, this study indicates that constrained ordinations are an effective means of detecting adaptive processes that result in weak, multilocus molecular signatures, providing a powerful tool for investigating the genetic basis of local adaptation and improving management actions.

Introduction

Analyzing genome scan data for loci underlying local adaptation has become common practice in evolutionary and ecological studies (Hoban *et al.* 2016). These analyses can provide insight into the mechanisms of local adaptation and inform management decisions for agricultural, natural resources, and conservation applications. Genotype-environment association (GEA) approaches are particularly promising for detecting these loci (Rellstab *et al.* 2015). Unlike differentiation outlier methods, which identify loci with strong allele frequency differences among populations, GEA approaches identify adaptive loci based on associations between genetic data and a set of environmental variables hypothesized to drive selection. Benefits of GEA include the option of using individual-based (as opposed to population-based) sampling and the ability to make explicit links to the ecology of organisms by including relevant predictors. The inclusion of predictors can also improve power and allows for the detection of selective events that do not produce high genetic differentiation among populations (De Mita *et al.* 2013; de Villemereuil *et al.* 2014; Rellstab *et al.* 2015).

Univariate statistical methods have dominated GEA since their first appearance (Mitton *et al.* 1977). These methods test one locus and one predictor variable at a time, and include generalized linear models (e.g. Joost *et al.* 2007; Stucki *et al.* 2016), variations on linear mixed effects models (e.g. Coop *et al.* 2010; Frichot *et al.* 2013; Yoder *et al.* 2014; Lasky *et al.* 2014), and non-parametric approaches (e.g. partial Mantel, Hancock *et al.* 2011). While these methods perform well, they can produce elevated false positive rates in the absence of correction for multiple comparisons. Corrections such as Bonferroni can be overly conservative (potentially removing true positive detections), while alternative correction methods, such as false discovery rate (FDR, Benjamini & Hochberg 1995), rely on an assumption of a null distribution of p-values, which may often be violated for empirical data sets. While none of these issues should discourage the use of univariate methods (though corrections should be chosen carefully, see François *et al.* (2016) for an excellent overview), other analytical approaches may be better suited to the high dimensionality of modern genomic data sets.

In particular, multivariate and machine learning approaches, which analyze many loci simultaneously, are well suited to data sets comprising hundreds of individuals sampled at many thousands of genetic markers. Compared to univariate methods, these approaches are thought to more effectively detect multilocus selection since they consider how groups of markers covary in response to environmental predictors (Rellstab et al. 2015). This is important because many adaptive processes are expected to result in weak, multilocus molecular signatures. These processes include selection on standing genetic variation, recent/contemporary selection that has not yet led to allele fixation, conditional neutrality, and the quantitative basis of many adaptive traits (Yeaman & Whitlock 2011; Le Corre & Kremer 2012; Savolainen *et al.* 2013; Tiffin & Ross-Ibarra 2014). Identifying the relevant patterns (e.g. coordinated shifts in allele frequencies across many loci) that underlie these adaptive processes is essential to both improving our understanding of the genetic basis of local adaptation, and conserving the evolutionary potential of species threatened by anthropogenic effects such as habitat fragmentation and climate change (Savolainen *et al.* 2013; Harrisson *et al.* 2014). While multivariate and machine learning methods may, in theory, be better suited to detecting these shared patterns of response, they have

not yet been tested on common data sets simulating multilocus adaptation, limiting confidence in their effectiveness on empirical data.

Here we evaluate a set of these methods, as well as a popular univariate approach, using published simulations of multilocus selection (Lotterhos & Whitlock 2014, 2015). We assess detection rates across methods using a common rank-based metric, and also present results based on cutoffs used in empirical studies. We then evaluate whether explicit correction for population structure improves performance. We follow up with a test of two of these methods on their ability to detect weak multilocus selection, as well as an assessment of two approaches to combining detections across different tests. We find that constrained ordinations maintain the best balance of true and false positive rates across a range of demographies, sampling designs, sample sizes, and selection levels.

Methods

Statistical approaches to GEA:

Multivariate statistical techniques, including ordinations such as principal components analysis (PCA), have been used to analyze genetic data for over fifty years (Cavalli-Sforza 1966). Indirect ordinations like PCA (which do not use predictors) use patterns of association within genetic data to find orthogonal axes that fully decompose the genetic variance. Constrained ordinations extend this analysis by restricting these axes to combinations of supplied predictors (Jombart *et al.* 2009; Legendre & Legendre 2012). When used as a GEA, a constrained ordination is essentially finding sets of loci that covary with multivariate environmental patterns. By contrast, a univariate GEA is testing for single locus relationships with single environmental predictors. The use of constrained ordinations in GEA goes back as far as Mulley *et al.* (1979), with more recent applications to genomic data sets in Lasky *et al.* (2012) and Forester *et al.* (2016). In this analysis, we test two promising constrained ordinations, redundancy analysis (RDA) and distance-based redundancy analysis (dbRDA). We also test an extension of RDA that uses a preliminary step of summarizing the genetic data into sets of covarying markers (Bourret *et al.* 2014). We do not include canonical correspondence analysis, a constrained ordination that is best suited to modeling unimodal responses, although this method has been used to analyze microsatellite data sets (e.g. Angers *et al.* 1999; Grivet *et al.* 2008).

Random Forest (RF) is a machine learning algorithm that is designed to identify structure in complex data and generate accurate predictive models. It is based on classification and regression trees (CART), which recursively partition data into response groups based on splits in predictors variables. CART models can capture interactions, contingencies, and nonlinear relationships among variables, differentiating them from linear models (De'ath & Fabricius 2000). RF reduces some of the problems associated with CART models (e.g. overfitting and instability) by building a "forest" of classification or regression trees with two layers of stochasticity: random bootstrap sampling of the data, and random subsetting of predictors at each node (Breiman 2001). This provides a built-in assessment of predictive accuracy (based on data left out of the bootstrap sample) and variable importance (based on the change in accuracy when variables are permuted). For GEA, variable importance is the focal statistic, where the predictor variables used at each split in the tree are molecular markers, and the goal is to sort individuals

into groups based on an environmental category (classification) or to predict an environmental response (regression). Markers with high variable importance are best able to sort individuals or predict responses. RF has been used in a number of recent GEA and GWAS studies (e.g. Holliday *et al.* 2012; Brieuc *et al.* 2015; Pavey *et al.* 2015; Laporte *et al.* 2016), but has not yet been tested in a GEA simulation framework. Finally, we compare these multivariate and machine-learning methods to a popular univariate method, latent factor mixed models (LFMM, Frichot *et al.* 2013).

Constrained ordinations:

We tested RDA and dbRDA as implemented by Forester et al. (2016). RDA is a two-step process in which genetic and environmental data are analyzed using multivariate linear regression, producing a matrix of fitted values. Then PCA of the fitted values is used to produce canonical axes, which are linear combinations of the predictors. We scaled genotypes for RDA. Distance-based redundancy analysis is similar to RDA but allows for the use of non-Euclidian dissimilarity indices. Whereas RDA can be loosely considered as a PCA constrained by predictors, dbRDA is analogous to a constrained principal coordinate analysis (PCoA, or a PCA on a non-Euclidean dissimilarity matrix). For dbRDA, we calculated the distance matrix using Bray-Curtis dissimilarity (Bray & Curtis 1957), which quantifies the dissimilarity among individuals based on their multilocus genotypes. For both methods, SNPs are modeled as a function of predictor variables, producing as many constrained axes as predictors. We identified outlier loci on each constrained ordination axis based on their "locus score", which represent the coordinates/loading of each locus in the ordination space. We use *rda* for RDA and *capscale* for dbRDA in the vegan, v. 2.3-5 package (Oksanen *et al.* 2013) in R v. 3.2.3 (R Development Core Team 2015) for this and all subsequent analyses.

Redundancy analysis of components:

This method, described by Bourret *et al.* (2014), differs from the approach described above in using a preliminary step that summarizes the genotypes into sets of covarying markers, which are then used as the response in RDA. The idea is to identify from these sets of covarying loci only the groups that are most strongly correlated with environmental predictors. We began by ordinating SNPs into principal components (PCs) using *prcomp* in R on the scaled data, producing as many axes as individuals. Following Bourret et al. (2014), we used parallel analysis (Horn 1965) to determine how many PCs to retain. Parallel analysis is a Monte Carlo approach in which the eigenvalues of the observed components are compared to eigenvalues from simulated data sets that have the same size as the original data. We used 1,000 random data sets and retained components with eigenvalues greater than the 99th percentile of the eigenvalues of the simulated data, using the hornpa package, v. 1.0 (Huang 2015).

Next, we applied a varimax rotation to the PC axes, which maximizes the correlation between the axes and the original variables (in this case, the SNPs). Note that once a rotation is applied to the PC axes, they are no longer "principal" components (i.e. axes associated with an eigenvalue/variance), but simply components. We then used the retained components as dependent variables in RDA, with environmental variables used as predictors. Next, components that were significantly correlated with at least one of the two constrained axes were retained.

Significance was based on a cutoff (alpha = 0.05) corrected for sample sizes using a Fisher transformation as in Bourret et al. (2014). Finally, SNPs were correlated with these retained components to determine outliers. We call this approach redundancy analysis of components (cRDA).

Random Forest:

The Random Forest approach implemented here builds off of work by Goldstein et al. (2010), Holliday et al. (2012), and Brieuc et al. (2015). This three-step approach is implemented separately for each predictor variable. The variables used in this study were continuous, so RF models were built as regression trees. For categorical predictors (e.g. soil type) classification trees would be used, which require a different parameterization (important recommendations for this case are provided in Goldstein et al. 2010).

First, we tuned the two main RF parameters, the number of trees (*ntrees*) and the number of predictors sampled per node (*mtry*). We tested a range of values for *ntrees* in a subset of the simulations, and found that 10,000 trees were sufficient to stabilize variable importance (note that variable importance requires a larger number of trees for convergence than error rates, Goldstein et al. 2010). We used the default value of *mtry* for regression (number of predictors/3, equivalent to ~3,330 SNPs in this case) after checking that increasing *mtry* did not substantially change variable importance or the percent variance explained. In a GEA/GWAS context, larger values of *mtry* reduce error rates, improve variable importance estimates, and lead to greater model stability (Goldstein *et al.* 2010).

Because RF is a stochastic algorithm, it is best to use multiple runs, particularly when variable importance is the parameter of interest (Goldstein *et al.* 2010). We begin by building three full RF models using all SNPs as predictors, saving variable importance as mean decrease in accuracy for each model. Next, we sampled variable importance from each run with a range of cutoffs, pulling the most important 0.5%, 1.0%, 1.5%, and 2.0% of loci. These values correspond to approximately 50/100/150/200 loci that have the highest variable importance. For each cutoff, we then created three additional RF models, using the average percent variance explained across runs to determine the best starting number of important loci for step 3. This step removes clearly unimportant loci from further consideration (i.e. “sparsity pruning”, Goldstein et al. 2010).

Third, we doubled the best starting number of loci from step 2; this is meant to accommodate loci that may have low marginal effects (Goldstein et al. 2010). We then built three RF models with these loci, and recorded the mean variance explained. We removed the least important locus in each model, and recalculated the RF models and mean variance explained. This procedure continues until two loci remain. The set of loci that explain the most variance are the final candidates. Candidates are then combined across runs to identify outliers. Locus rankings used average variable importance for each locus across the three runs.

Latent factor mixed models:

Latent factor mixed models are hierarchical Bayesian mixed models that account for population structure using latent factors (*K*), which are similar to principal components (Frichot *et al.* 2013). We tested values of *K* ranging from one to 25 using a sparse nonnegative matrix factorization

algorithm (Frichot *et al.* 2014), implemented as function *snmf* in the package LEA, v. 1.2.0 (Frichot & François 2015). We plotted the cross-entropy values and selected *K* based on the inflection point in these plots; when the inflection point was not clear, we used the value where additional cross-entropy loss was minimal.

We parameterized LFMM models with this best estimate of *K*, and ran each model ten times with 5,000 iterations and a burnin of 2,500. We used the median of the squared z-scores to rank loci and calculate a genomic inflation factor (GIF) to assess model fit (Frichot & François 2015; François *et al.* 2016). The GIF is used to correct for inflation of z-scores at each locus, which can occur when population structure or other confounding factors are not sufficiently accounted for in the model (François *et al.* 2016). The GIF is calculated by dividing the median of the squared z-scores by the median of the chi-squared distribution. We used the LEA and qvalue, v. 2.2.2 (Storey *et al.* 2015) packages in R. Full K and GIF results are presented in Table S1.

Correction for population structure:

To determine if explicit modeling of population structure improved the performance of ordinations and RF, we repeated those analyses after accounting for population structure using spatial eigenvectors (for RDA, dbRDA, and cRDA) and regression with ancestry coefficients (for RF). The spatial eigenvector procedure uses Moran eigenvector maps (MEM) as spatial predictors in partial RDA and dbRDA analysis. MEMs provide a decomposition of the spatial relationships among sampled locations based on a spatial weighting matrix (Dray *et al.* 2006). We used spatial filtering to determine which MEMs to include in the partial analyses (Dray *et al.* 2012). Briefly, this procedure begins by applying a principal coordinate analysis (PCoA) to the genetic distance matrix, which we calculated using Bray-Curtis dissimilarity. We used the broken-stick criterion (Legendre & Legendre 2012) to determine how many genetic PCoA axes to retain. Retained axes were used as the response in a full RDA, where the predictors included all MEMs. Forward selection (Blanchet *et al.* 2008) was used to reduce the number of MEMs, using the full RDA adjusted R^2 statistic as the threshold. Finally, retained MEMs that were significantly correlated with environmental predictors were removed (alpha = 0.05/number of MEMs). The final set of significant MEMs were used as conditioning variables in RDA and dbRDA. We used the spdep, v. 0.6-9 (Bivand *et al.* 2013) and adespatial, v. 0.0-7 (Dray *et al.* 2016) packages to calculate MEMs.

For RF, we followed Brieuc et al. (2015) and used individual ancestry coefficients to correct both allele counts and environmental variables. We used function *snmf* to estimate individual ancestry coefficients, running five replicates using the best estimate of *K*, and extracting individual ancestry coefficients from the replicate with the lowest cross-entropy. For genotypes, we used the residuals from logistic regression of SNP counts against ancestry coefficients. For environmental variables, we used the residuals from linear models of the variables against ancestry coefficients. These residuals were used as inputs into the RF framework described above.

Simulations:

We used a subset of simulations published by Lotterhos & Whitlock (2014, 2015). Briefly, four demographic histories are represented in these data, each with three replicated environmental surfaces (Fig. S1): an equilibrium island model (IM), equilibrium isolation by distance (IBD), and nonequilibrium isolation by distance with expansion from one (1R) or two (2R) refugia. In all cases, demography was independent of selection strength, which is analogous to simulating soft selection (Lotterhos & Whitlock 2014). Haploid, biallelic SNPs were simulated independently, with 9,900 neutral loci and 100 under selection. The mean of the environmental/habitat parameter had a selection coefficient equal to zero and represented the background across which selective habitat was patchily distributed (Fig. S1). Selection coefficients represent a proportional increase in fitness of alleles in response to habitat, where selection is increasingly positive as the environmental value increases from the mean, and increasingly negative as the value decreases from the mean (Lotterhos & Whitlock 2014, Fig. S1). This landscape emulates a weak cline, with a north-south trend in the selection surface. Of the 100 adaptive loci, most were under weak selection. For the IBD scenarios, selection coefficients were 0.001 for 40 loci, 0.005 for 30 loci, 0.01 for 20 loci, and 0.1 for 10 loci. For the 1R, 2R, and IM scenario, selection coefficients were 0.005 for 50 loci, 0.01 for 33 loci, and 0.1 for 17 loci. Note that realized selection varied across demographies, so results across demographic histories are not directly comparable (Lotterhos & Whitlock 2015).

We used the following sampling strategies and sample sizes from Lotterhos & Whitlock (2015): random, paired, and transect strategies, with 90 demes sampled, and 6 or 20 individuals sampled per deme. Overall, 72 simulations were used for testing. We assessed trend in neutral loci using linear models of allele frequencies within demes as a function of coordinates. We evaluated the strength of local adaptation using linear models of allele frequencies within demes as a function of environment.

The original simulation data assigned individual genotypes in a non-random fashion within populations. Because we were conducting individual-based analyses, we randomized allele counts for SNPs among individuals, within populations (K. Lotterhos, pers. comm.). We prepared two environmental predictors: habitat, which imposed a continuous selective gradient on the non-neutral loci, and the value for the x-coordinate of each population. We included the x-coordinate as a spurious predictor, analogous to an environmental variable hypothesized to drive selection that covaries with longitude. We scaled both variables prior to use. We did not use the y-coordinate as a second spurious predictor because it was highly correlated with habitat ($r > 0.7$) in the majority of simulations (Table S2).

Evaluation statistics:

In order to equitably compare the output from these methods, we used locus rankings to calculate the number of correct detections out of the number of selected loci in each simulation (i.e. a common cutoff for all methods). We ranked loci based on the relevant (scaled) test statistics across both predictors, i.e. loadings and correlations for ordinations, variable importance for RF, and z-scores for LFMM. For example, in a simulation with 100 loci under selection and 90 true

positive detections in the top 100 ranked loci, the true positive rate (TPR) would be 90/100, while the false positive rate (FPR) would be 10/100.

Since cutoffs (e.g. thresholds for statistical significance) are frequently used in empirical analyses for null hypothesis testing, we also provide detection results for commonly used cutoffs. We calculated a cutoff TPR as the number of correct positive detections out of the number possible. The cutoff FPR was the number of incorrect positive detections out of 9900 possible. For the main text, we present results from the "best" cutoff for each method; full results for all cutoffs tested are presented in the Supplemental Information. For constrained ordinations (RDA and dbRDA) we identified outliers as SNPs with a locus score +/- 2.5 and 3 SD from the mean score of each constrained axis. For cRDA, we used cutoffs for SNP-component correlations of alpha = 0.05, 0.01, and 0.001, corrected for sample sizes using a Fisher transformation as in Bourret et al. (2014). For LFMM, we compared two Bonferroni-corrected cutoffs (0.05 and 0.01) and a FDR cutoff of 0.1.

For both ranked and cutoff evaluation statistics, we calculated TPRs separately for different selection coefficients. In all cases, detection rates were averaged across the three replicate environments. Note that the number of selected loci ranged from 89-100, since some loci were removed by the original simulation authors due to low heterozygosity (Lotterhos & Whitlock 2015).

Weak selection:

We compared RDA and cRDA for their ability to detect signals of weak selection (s = 0.005 and s = 0.001). All tests were performed as described above, with no additional corrections for population structure, after removing loci under strong (s = 0.1) and moderate (s = 0.01) selection from the simulation data sets. The number of loci under selection in these cases varied from 43 to 76.

Results

Population corrections:

We found that explicitly accounting for population structure did not improve the performance of ordinations and was detrimental to the performance of RF. Spatial filtering had very little to no impact on TPRs and FPRs of ordination methods (Table S3 and S4). No corrections were applied to IM scenarios for ordination methods, due to low spatial structure (i.e. no PCoA axes were retained based on the broken-stick criterion). Regression of ancestry coefficients on RF inputs dramatically reduced TPRs (Table S3 and S4). All results presented here do not use population corrections; full results for runs with correction are presented in Table S3 and S4.

Ranked results:

The three ordinations performed comparably in terms of locus rankings, and tended to outperform RF and LFMM (Fig. 1). Ordinations performed best in IBD, 1R, and 2R demographies. Ordination results were relatively insensitive to sample size and sampling design

(with the IM random sample being the exception, with lower TPRs). Within ordination techniques, RDA and cRDA had slightly higher detection rates compared to dbRDA. RF had very low TPRs across all simulations. LFMM was more sensitive than ordinations to demography, sampling design, and sample size. Detection rates for LFMM were better with smaller sample sizes (6 individuals/deme), and generally higher for the paired sampling design.

All methods performed well on loci under strong selection, with all methods but RF detecting 100% of these loci (Figs. 2 and S2). Detection rates for loci under moderate and weak selection were comparable across ordination methods, with RDA and cRDA having the overall highest detection. RF had very low detection rates for moderate and weakly selected loci, while LFMM had lower detection rates than ordinations in non-equilibrium demographies. For ordinations, selection level detection rates were mostly comparable across sample sizes, except IM, where detection was better with the larger sample size. RF and LFMM had better detection with smaller sample sizes.

Cutoff results:

The best performing cutoffs were: RDA/dbRDA, +/- 3 SD; cRDA, alpha = 0.001; and LFMM, FDR = 0.1. Full cutoff results are presented in the Supplementary Information (Fig. S5, S6, and S7). Cutoff TPRs (Fig. 3) generally reflected ranked TPRs (Fig. 1), with ordinations performing best in most cases, RF having low detection rates overall, and LFMM performing well depending on the scenario. FPRs were low for all methods except cRDA (IBD, 1R, and 2R demographies). Selection level detection rates using cutoffs were generally higher than ranked results for cRDA, RF, and LFMM (Fig. S3 and S4).

Weak selection:

We compared RDA and cRDA for their ability to detect only weak loci in the simulations (Fig. 4). Using locus rankings, RDA had more consistent performance across scenarios, and had overall higher TPRs and lower FPRs compared with cRDA. cRDA had low detection rates in the 1R demography with the larger sample size, and in the IM demography, regardless of sample size (no detections at all with 6 individuals/deme). Using cutoffs, RDA had more consistent performance across all scenarios. Detection was better using cRDA in the 1R demography when sampling 6 individuals per deme, but was much worse when sampling 20 individuals per deme. Overall, cRDA had the same detection problems noted with the ranked results, in addition to high FPRs under 1R, 2R, and IBD demographies.

Combining detections:

We compared the univariate LFMM and multivariate RDA cutoff results for overlap and differences in their detections (Fig. 5). The methods had greater commonality in the loci they correctly identified as TPs than in the loci they incorrectly identified (FPs), indicating that mutual detections could be an effective way of reducing FPRs. In some cases, however, RDA detected a large number of selected loci that were not identified by LFMM (Fig. 5, second column), indicating that power would be lost when using only overlapping results. Significant

TP contributions from LFMM were found only in the IM demography. Overall, LFMM had greater numbers of unique FPs compared to RDA. Few unique TP detections and many unique FP detections limit the utility of combining LFMM and RDA.

Discussion

Multivariate and machine learning genotype-environment association (GEA) methods have been noted for their ability to detect multilocus selection (Rellstab *et al.* 2015; Hoban *et al.* 2016), although there has been no controlled assessment of the effectiveness of these methods in detecting multilocus selection to date. Since these approaches are increasingly being used in empirical analyses (e.g. Bourret et al. 2014; Brieuc et al. 2015; Pavey et al. 2015; Hecht et al. 2015; Laporte et al. 2016; Brauer et al. 2016), it is important that these claims are evaluated to ensure that the most effective GEA methods are being used, and that their results are being appropriately interpreted.

Here we compare a suite of GEA approaches in a simulation framework to assess their ability to correctly detect multilocus selection under different demographic and sampling scenarios. We found that constrained ordinations had the best overall performance across the demographies, sampling designs, sample sizes, and selection levels tested here. The univariate LFMM method also performed well, though power was scenario-dependent. Random Forest, by contrast, had very low detection rates overall. In the following sections we address these methods by category and discuss reasons for method performance and provide suggestions for their use on empirical data sets.

Constrained ordinations:

The three constrained ordination methods all performed well (Fig. 1 and 2). They were relatively insensitive to sample size (6 vs 20 individuals sampled per deme), in agreement with Xuereb *et al.* (In review) who found that reducing sampling from 500 to 100 individuals had only moderate effects on TPRs for RDA and dbRDA and no effect on FPRs. The one exception was the IM demography, where larger sample sizes consistently improved TPRs, as previously noted by De Mita *et al*. (2013) and Lotterhos & Whitlock (2015) for univariate GEAs. Power was lowest in the IM demography, which is typified by a lack of spatial autocorrelation in allele frequencies and a reduced signal of local adaptation (Table S2), making detection more difficult. Detection rates were highest for IBD, followed by the 2R and 1R demographies. All three methods were relatively insensitive to sampling design, with transects performing slightly better in 1R and random sampling performing worst in IM. Otherwise results were consistent across designs, in contrast to the univariate GEAs tested by Lotterhos and Whitlock (2015), most of which had higher power with the paired sampling strategy. Ordinations are likely less sensitive to sampling design since they take advantage of covarying signals of selection, making them more robust to sampling that does not maximize environmental differentiation (e.g. random or transect designs). All methods performed similarly in terms of detection rates across selection strengths. As expected, weak selection was more difficult to detect than moderate or strong selection, except for IBD, where detection levels were high regardless of selection (Fig. 2, and S2-S4).

High TPRs were maintained when using cutoffs for all three ordination methods. False positives were universally low for RDA and dbRDA. By contrast, cRDA showed high FPRs for all demographies except IM, tempering its slightly higher TPRs. These higher FPRs are a consequence of using component axes as predictors. Across all scenarios and sample sizes, cRDA detected component 1, 2, or both as significantly associated with the constrained RDA axes (Table S5). Most selected loci load on these components (keeping TPRs high), but neutral markers also load on these axes, especially in cases where there are strong trends in neutral loci (i.e. maximum trends in neutral markers reflect FPRs for cRDA, Table S2, Fig. 3). Given these results, we hypothesized that it might be challenging for cRDA to detect weak selection in the absence of a covarying signal from loci with stronger selection coefficients. If the selection signature is weak, it may load on a lower-level component axis (i.e. an axis that explains less of the genetic variance), or it may load on higher-level axes, but fail to be significantly associated with the constrained axes. Note that although cRDA contains a step to reduce the number of components, parallel analysis resulted in retention of all axes in every simulation tested here (Table S5). This meant that cRDA could search for the signal of selection across all possible components.

When tested on simulations with loci under weak selection only, RDA, which uses the genotype matrix directly, maintained similar detection patterns as in the full data set, indicating that selection signals can be detected with this method in the absence of loci under strong selection (Fig. 4). Using a cutoff, RDA maintained very low FPRs across all simulation scenarios and sample sizes. By contrast, cRDA detection was more variable, ranging from comparable detection rates with the full data set, to no/poor detections under certain demographies and sample sizes (Fig. 4). In these latter cases, poor performance is reflected in the component axes detected as significant (Table S5); instead of identifying the signal in the first few axes, a variable set of lower-variance axes are detected (or none are detected at all). This indicates that the method is not able to "find" the selected signal in the component axes in cases where that signal is not driven by strong selection. This result, in addition to higher FPRs for cRDA, builds a case for using the genotype matrix directly with a constrained ordination such as RDA or dbRDA, as opposed to a preliminary step of data conversion with PCA.

RDA plots illustrate how loci under selection can be distinguished from neutral loci using constrained ordinations (Fig. 6). RDA shows a negative relationship between habitat and the selected loci, and is clearly able to distinguish the signal of selection from the spurious x-coordinate predictor. Depending on where the cutoff is placed (i.e. how many deviations from the mean score), false negatives can be seen in the IM, 1R, and 2R demographies as the pink selected loci that are not well-differentiated from the "cloud" of gray neutral loci. This is particularly noticeable in the IM demography, where many of the loci under weak selection do not differentiate from the neutral signal. Data from natural systems likely lie somewhere among these demographic extremes, and successful differentiation in the presence of IBD and non-equilibrium conditions indicate that ordinations should work well across a range of natural systems.

Finally, our results suggest that additional correction for population structure is not needed for these methods, at least within the range of population structure present here (Table S3 and S4). This indicates that constrained ordinations are effectively accounting for the joint action

of selection (modeled on the constrained axes) and demography (residual variance not explained by environment that is modeled on the unconstrained axes). Biplots of the first two unconstrained (PC) axes (Fig. S9) and screeplots of the variance explained by the first 15 unconstrained axes (Fig. S10) reflect how demographies with different levels of population structure are modeled in the unconstrained axes by RDA. Testing these methods in simulation scenarios with more significant population structure would be a helpful follow-up to confirm the generality of these results.

Random Forest:

Random Forest performed very poorly in detecting loci under moderate and weak selection. These results indicate that RF is not a good GEA approach for large genomic data sets. Poor performance is caused by the sparsity of the genotype matrix (i.e. most SNPs are not under selection), which results in detection that is dominated by strongly selected loci (i.e. loci with strong marginal effects, Fig. 2). This issue has been documented in other simulation and empirical studies (Goldstein *et al.* 2010; Winham *et al.* 2012; Wright *et al.* 2016) and indicates that RF is not suited to identifying weak multilocus selection or interaction effects in these cases. Empirical studies that have used RF as a GEA have likely identified a subset of loci under strong selection, but are unlikely to have identified loci underlying more complex genetic architectures. Note that the amount of environmental variance explained by the RF model can be high (i.e. overall percent variance explained by the detected SNPs), while still failing to identify most of the loci under selection (Table S6). Removing strong associations from the genotypic matrix can potentially help with the detection of weaker effects (Goldstein *et al.* 2010), but this approach has not been tested on large matrices. Combined with the computational burden of this method (taking 10-14 days for the larger data sets), as well as the availability of fast and accurate alternatives such as RDA (which takes ~3 minutes on the same data), it is clear that RF is not a viable option for GEA analysis of genomic data.

Random Forest does hold promise for the detection of interaction effects in much smaller data sets (e.g. tens of loci, Holliday *et al*. 2012). However, this is an area of active research, and the capacity of RF models in their current form to both capture and identify SNP interactions has been disputed (Winham *et al.* 2012; Wright *et al.* 2016). New modifications of RF models are being developed to more effectively identify interaction effects (e.g. Li et al. 2016), but these models are computationally demanding and are not designed for large data sets. Overall, extensions of RF show potential for identifying more complex genetic architectures, but caution is warranted in using them on empirical data prior to rigorous testing on realistic simulation scenarios.

Latent factor mixed models:

The univariate LFMM method performed well, especially considering that it does not take advantage of covariation in allele frequencies to detect loci. Still, as expected, detection rates were lower overall for loci under moderate and weak selection when compared with ordinations. This is in agreement with LFMM results from de Villemereuil *et al.* (2014), who also found a reduction in power when detecting polygenic selection. Additionally, the performance of LFMM

was more dependent on sampling design, sample size, and demography than ordinations. Our results clearly demonstrated that Bonferroni corrections are too conservative for LFMM (Fig. S7), and that FDR-based approaches for multiple testing are much better suited to genomic data when the GIF indicates the test is well-calibrated (Table S1).

Should results from different tests be combined?

A common approach in local adaptation studies is to run multiple tests (GEA only, or a combination of GEA and differentiation methods) and look for duplicate detections across methods. This ad hoc approach is thought to increase confidence in TPRs, while minimizing FPRs. The problem with this approach is that it can bias detection toward strong selective sweeps to the exclusion of other adaptive mechanisms which may be equally important in shaping phenotypic variation (Le Corre & Kremer 2012; François *et al.* 2016). If the goal is to detect other forms of selection such as recent selection or selection on standing genetic variation, this approach will not be effective since most methods are unlikely to detect these weak signals.

This issue is illustrated by using two different combinations of RDA and LFMM detections: keeping only mutual detections and keeping all detections. Agreement on TPs is high, while agreement on FPs is low (Fig. 5 first column). Keeping only loci detected by both RDA and LFMM may therefore seem to be an effective way to reduce FPRs while maintaining good TPRs. However, depending on the scenario, RDA has a large number of true positive detections that are unique to that method (Fig. 5, second column). These unique TPs, all of which are under moderate and weak selection (Fig. 2), would be discarded using a duplicates-only criterion, limiting our inference to those loci with the strongest adaptive signal. This effect was also noted by Lotterhos & Whitlock (2015) when looking at detection overlap in the methods tested in their analysis. Alternatively, keeping all detections from both methods would dramatically increase FPRs, while providing very little improvement in TPRs since multiple unique detections by LFMM are found only in the IM demography (Fig. 5, third column).

The decision of whether and how to combine results from different tests will be specific to the study questions, the tolerance for false negative and false positive detections, and the capacity for follow-up analyses on detected markers. For example, if the goal is to detect loci with strong effects while keeping false positive rates as low as possible, running multiple GEA and/or differentiation-based methods and considering only duplicate detections could be a suitable strategy. However, if the goal is to detect selection on standing genetic variation or a recent selection event, combining detections from multiple tests would be too conservative. In this case, the best approach would be to use a single GEA method, such as RDA, that can effectively detect covarying signals arising from multilocus selection, while being robust to selection strength, sampling design, and sample size.

Conclusions and recommendations:

Random Forest performed very poorly as a GEA method on the simulations tested here. TPRs were limited due to strong marginal effects created by the subset of loci under strong selection. Still, RF may be useful for follow-up analyses of the genomic architecture of smaller sets of candidate loci. However, the effectiveness of RF for identifying interactions and other complex

genetic architectures is currently disputed, and practitioners should proceed cautiously until new extensions of RF are rigorously tested under realistic simulation scenarios. The univariate method we tested, LFMM, performed well, but was more sensitive to sampling designs and sample sizes than RDA and dbRDA. Additionally, since this method cannot detect covarying signals of selection, overall detection of loci under moderate and weak selection was reduced.

We found that constrained ordinations, especially RDA, show a superior combination of low FPRs and high TPRs across weak, moderate, and strong multilocus selection. These results were robust across the demographic histories, sampling designs, and sample sizes tested here. Additionally, RDA outperformed an alternative ordination-based approach, cRDA, especially (and importantly) when the multilocus selection signature was completely derived from loci under weak selection. It is important to note that constrained ordinations require complete data sets (no missing values). Fortunately, recent work has indicated that RDA and dbRDA are robust to even high levels (50%) of randomly missing data when using simple imputation methods such as the mean value across individuals (Xuereb *et al.* In review). Additionally, RDA and dbRDA can be used on both individual and population-based samples. It will be important to continue testing these promising methods in simulation frameworks that include genetic architectures that are more complex than the multilocus selection response modeled here. This includes locus interaction effects (i.e. epistasis) and more complex polygenic architectures. However, this study indicates that constrained ordinations are an effective means of detecting adaptive processes that result in weak, multilocus molecular signatures, providing a powerful tool for investigating the genetic basis of local adaptation and informing management actions to conserve the evolutionary potential of species of agricultural, forestry, fisheries, and conservation concern.

Acknowledgements

Thank you to Katie E. Lotterhos for sharing their simulation data (Lotterhos & Whitlock 2015) and for additional spatial coordinate data and advice. Tom Milledge with Duke Resource Computing provided invaluable assistance with the Duke Compute Cluster. We also thank Olivier François for helpful advice with LFMM.

References

Angers B, Magnan P, Plante M, Bernatchez L (1999) Canonical correspondence analysis for estimating spatial and environmental effects on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). *Molecular Ecology*, **8**, 1043–1053.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate - a practical and powerful approach. *Journal of the Royal Statistical Society Series B-Methodological*, **57**, 289–300.

Bivand R, Hauke J, Kossowski T (2013) Computing the Jacobian in Gaussian spatial autoregressive models: an illustrated comparison of available methods. *Geographical Analysis*, **45**, 150–179.

Blanchet FG, Legendre P, Borcard D (2008) Forward selection of explanatory variables. *Ecology*, **89**, 2623–2632.

Bourret V, Dionne M, Bernatchez L (2014) Detecting genotypic changes associated with selective mortality at sea in Atlantic salmon: polygenic multilocus analysis surpasses genome scan. *Molecular Ecology*, **23**, 4444–4457.

Brauer CJ, Hammer MP, Beheregaray LB (2016) Riverscape genomics of a threatened fish across a hydroclimatically heterogeneous river basin. *Molecular Ecology*, **25**, 5093–5113.

Bray JR, Curtis JT (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**, 325–349.

Breiman L (2001) Random Forests. *Machine Learning*, **45**, 5–32.

Brieuc MSO, Ono K, Drinan DP, Naish KA (2015) Integration of Random Forest with population-based outlier analyses provides insight on the genomic basis and evolution of run timing in Chinook salmon (*Oncorhynchus tshawytscha*). *Molecular Ecology*, **24**, 2729–2746.

Cavalli-Sforza LL (1966) Population structure and human evolution. *Proceedings of the Royal Society B: Biological Sciences*, **164**, 362–379.

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

De Mita S, Thuillet A-C, 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.

De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178–3192.

Dray S, Blanchet G, Borcard D *et al.* (2016) *adespatial: Multivariate multiscale spatial analysis*. R package version 0.0-7.

Dray S, Legendre P, Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, **196**, 483–493.

Dray S, Pélissier R, Couteron P *et al.* (2012) Community ecology in the age of multivariate multiscale spatial analysis. *Ecological Monographs*, **82**, 257–275.

Forester BR, Jones MR, Joost S, Landguth EL, Lasky JR (2016) Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Molecular Ecology*, **25**, 104–120.

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics*, **196**, 973–983.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF (2010) An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC Genetics*, **11**, 1–13.

Grivet D, Sork VL, Westfall RD, Davis FW (2008) Conserving the evolutionary potential of California valley oak (*Quercus lobata* Née): a multivariate genetic approach to conservation planning. *Molecular Ecology*, **17**, 139–156.

Hancock AM, Brachi B, Faure N *et al.* (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **334**, 83–86.

Harrisson KA, Pavlova A, Telonis-Scott M, Sunnucks P (2014) Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, **7**, 1008–1025.

Hecht BC, Matala AP, Hess JE, Narum SR (2015) Environmental adaptation in Chinook salmon (*Oncorhynchus tshawytscha*) throughout their North American range. *Molecular Ecology*, **24**, 5573–5595.

Hoban S, Kelley JL, Lotterhos KE *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist*, **188**, 379–397.

Holliday JA, Wang T, Aitken S (2012) Predicting adaptive phenotypes from multilocus genotypes in Sitka Spruce (*Picea sitchensis*) using Random Forest. *G3: Genes|Genomes|Genetics*, **2**, 1085–1093.

Horn JL (1965) A rationale and test for the number of factors in factor analysis. *Psychometrika*, **30**, 179–185.

Huang F (2015) *hornpa: Horn's (1965) Test to Determine the Number of Components/Factors.* R package version 1.0.

Jombart T, Pontier D, Dufour A-B (2009) Genetic markers in the playground of multivariate analysis. *Heredity*, **102**, 330–341.

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.

Laporte M, Pavey SA, Rougeux C *et al.* (2016) RAD sequencing reveals within-generation polygenic selection in response to anthropogenic organic and metal contamination in North Atlantic Eels. *Molecular Ecology*, **25**, 219–237.

Lasky JR, Des Marais DL, McKay JK *et al.* (2012) Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Molecular Ecology*, **21**, 5512–5529.

Lasky JR, Marais D, L D *et al.* (2014) Natural variation in abiotic stress responsive gene expression and local adaptation to climate in *Arabidopsis thaliana*. *Molecular Biology and Evolution*, **31**, 2283–2296.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

Legendre P, Legendre L (2012) *Numerical Ecology*. Elsevier, Amsterdam, The Netherlands.

Li J, Malley JD, Andrew AS, Karagas MR, Moore JH (2016) Detecting gene-gene interactions using a permutation-based random forest method. *BioData Mining*, **9**, 14.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Mitton JB, Linhart YB, Hamrick JL, Beckman JS (1977) Observations on the genetic structure and mating system of Ponderosa Pine in the Colorado front range. *Theoretical and Applied Genetics*, **51**, 5–13.

Mulley JC, James JW, Barker JSF (1979) Allozyme genotype-environment relationships in natural populations of *Drosophila buzzatii*. *Biochemical Genetics*, **17**, 105–126.

Oksanen J, Blanchet FG, Kindt R *et al.* (2013) *vegan: Community Ecology Package*.

Pavey SA, Gaudin J, Normandeau E *et al.* (2015) RAD sequencing highlights polygenic discrimination of habitat ecotypes in the panmictic American Eel. *Current Biology*, **25**, 1666–1671.

R Development Core Team (2015) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–4370.

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

Storey JD, Bass AJ, Dabney A, Robinson D (2015) *qvalue: Q-value estimation for false discovery rate control*. R package version 2.2.2.

Stucki S, Orozco-terWengel P, Forester BR *et al.* (2016) High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*.

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 É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

Winham SJ, Colby CL, Freimuth RR *et al.* (2012) SNP interaction detection with Random Forests in high-dimensional genetic data. *BMC Bioinformatics*, **13**, 164.

Wright MN, Ziegler A, König IR (2016) Do little interactions get lost in dark random forests? *BMC Bioinformatics*, **17**, 145.

Xuereb A, Stahlke A, Bermingham M *et al.* (In review) Missing data have limited effects on the performance of genotype-environment association methods. *Molecular Ecology*.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration–selection balance. *Evolution*, **65**, 1897–1911.

Yoder JB, Stanton-Geddes J, Zhou P *et al.* (2014) Genomic signature of adaptation to climate in *Medicago truncatula*. *Genetics*, **196**, 1263–1275.

Data accessibility

Simulation data from Lotterhos & Whitlock (2015): Dryad: doi:10.5061/dryad.mh67v
Supporting simulation data (coordinate files) for Lotterhos & Whitlock (2015) data provided by Wagner et al. 2017: Dryad: doi:10.5061/dryad.b12kk

Author contributions

BRF and DLU conceived the study. BRF performed the analyses and wrote the manuscript. HHW contributed code. JRL, HHW, and DLU helped interpret the results and write the manuscript.

Supporting information

Forester_etal_Rcode.zip: Contains R code for data preparation and all of the methods tested.

Forester_etal_SI.pdf: Supplemental Figures S1-S10 and Tables S1-S6.

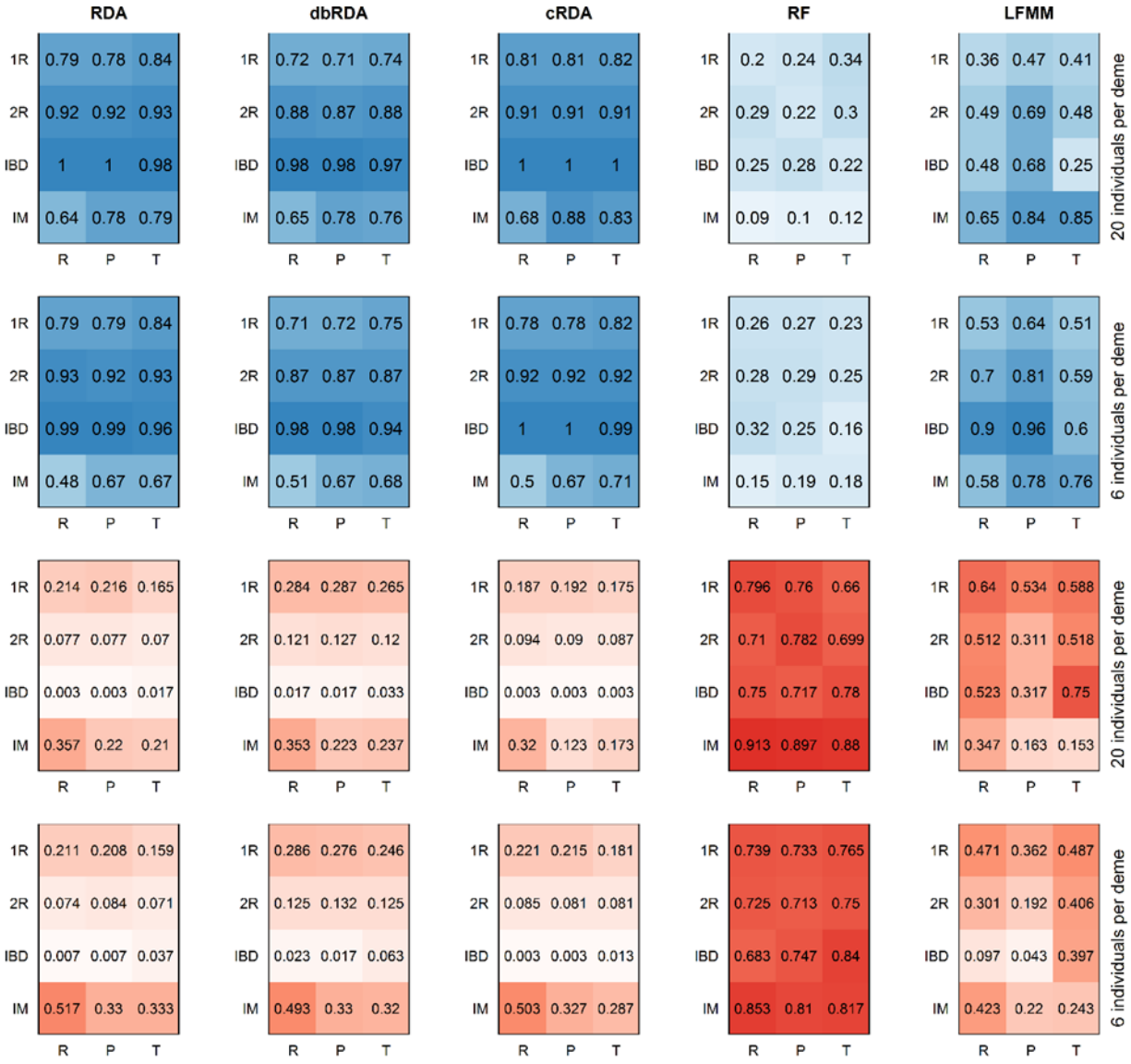


Figure 1. Average true positive (blue) and false positive (red) rates from five methods (columns) using locus rankings (i.e. number of positive detections out of number of loci under selection). Each method shows results for different sampling strategies (R = random, P = pairs, T = transects), demographies (1R and 2R = refugial expansion, IBD = equilibrium isolation by distance, IM = equilibrium island model), and sample sizes (rows).

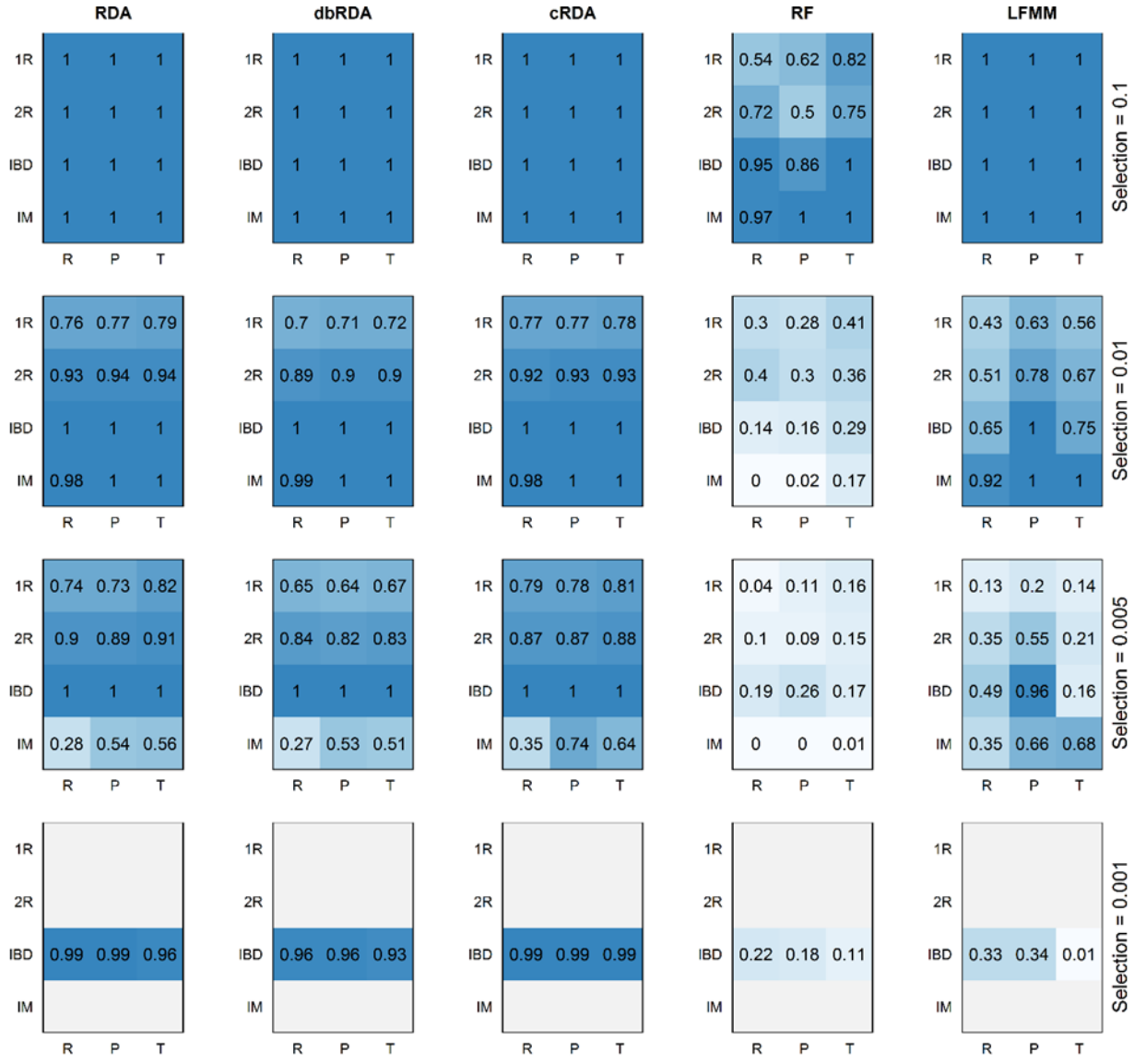


Figure 2. Average true positive rates for different levels of selection (rows) from five methods (columns) using locus rankings and a sample size of 20 individuals per deme. Each method shows results for different sampling strategies (R = random, P = pairs, T = transects) and demographies (1R and 2R = refugial expansion, IBD = equilibrium isolation by distance, IM = equilibrium island model). Only the IBD demography included very weak selection (s=0.001).

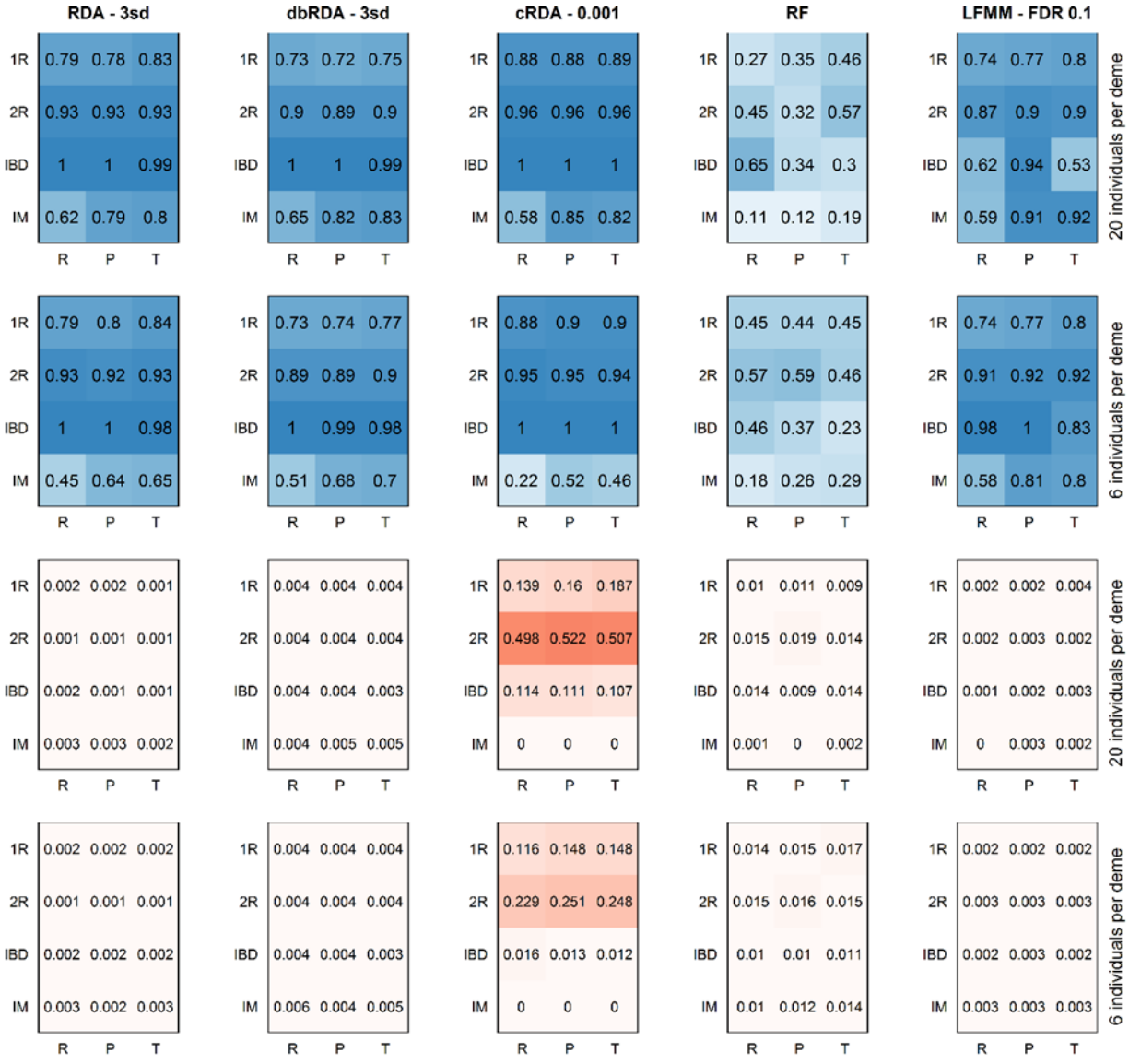


Figure 3. Average true positive (blue) and false positive (red) rates from five methods (columns) using the best cutoff for each method. Each method shows results for different sampling strategies (R = random, P = pairs, T = transects), demographies (1R and 2R = refugial expansion, IBD = equilibrium isolation by distance, IM = equilibrium island model), and sample sizes (rows).

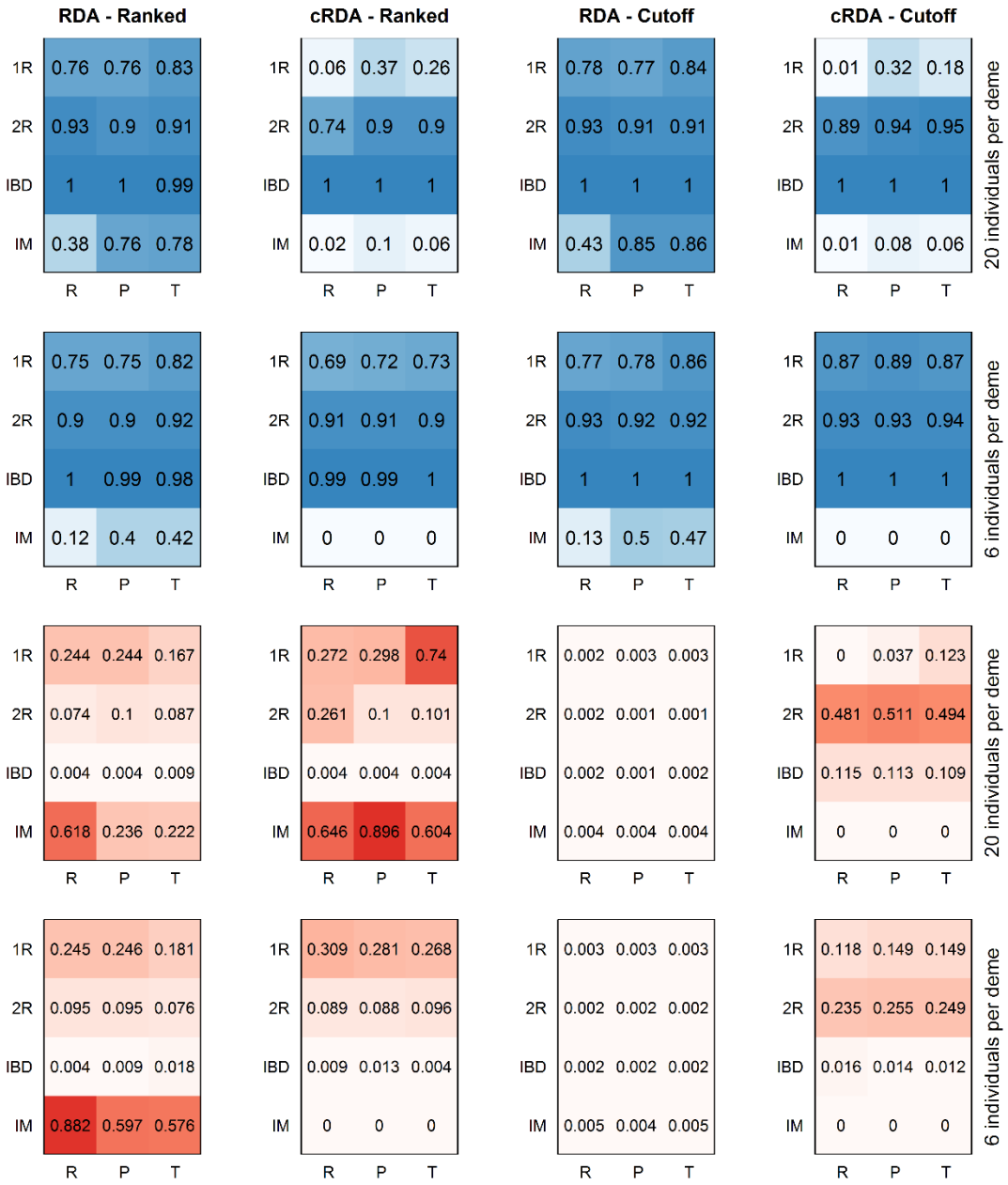
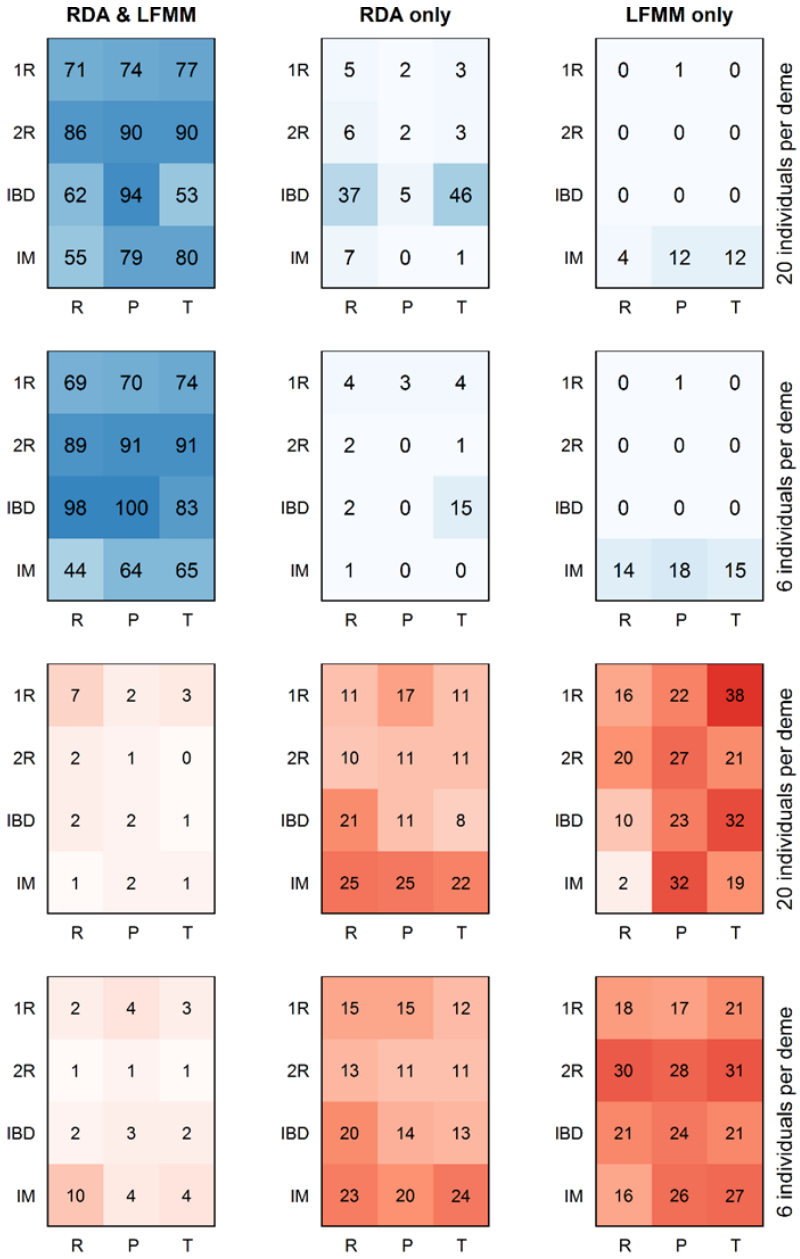


Figure 4. Average true positive (blue) and false positive (red) rates for RDA and cRDA (columns) on simulations with weak selection only. The first two columns show results for locus rankings, while the third and fourth columns show results for the best cutoff for each method. Results are presented for different sampling strategies (R = random, P = pairs, T = transects), demographies (1R and 2R = refugial expansion, IBD = equilibrium isolation by distance, IM = equilibrium island model), and sample sizes (rows).



765 **Figure 5.** Average counts of true positive (blue) and false positive (red) detections for two methods, RDA and LFMM. The first column shows the average number of loci detected by both methods. The second and third columns show the average number of detections that are unique to RDA and LFMM, respectively. Results are presented for different sampling strategies (R = random, P = pairs, T = transects), demographies (1R and 2R = refugial expansion, IBD = 770 equilibrium isolation by distance, IM = equilibrium island model), and sample sizes (rows).

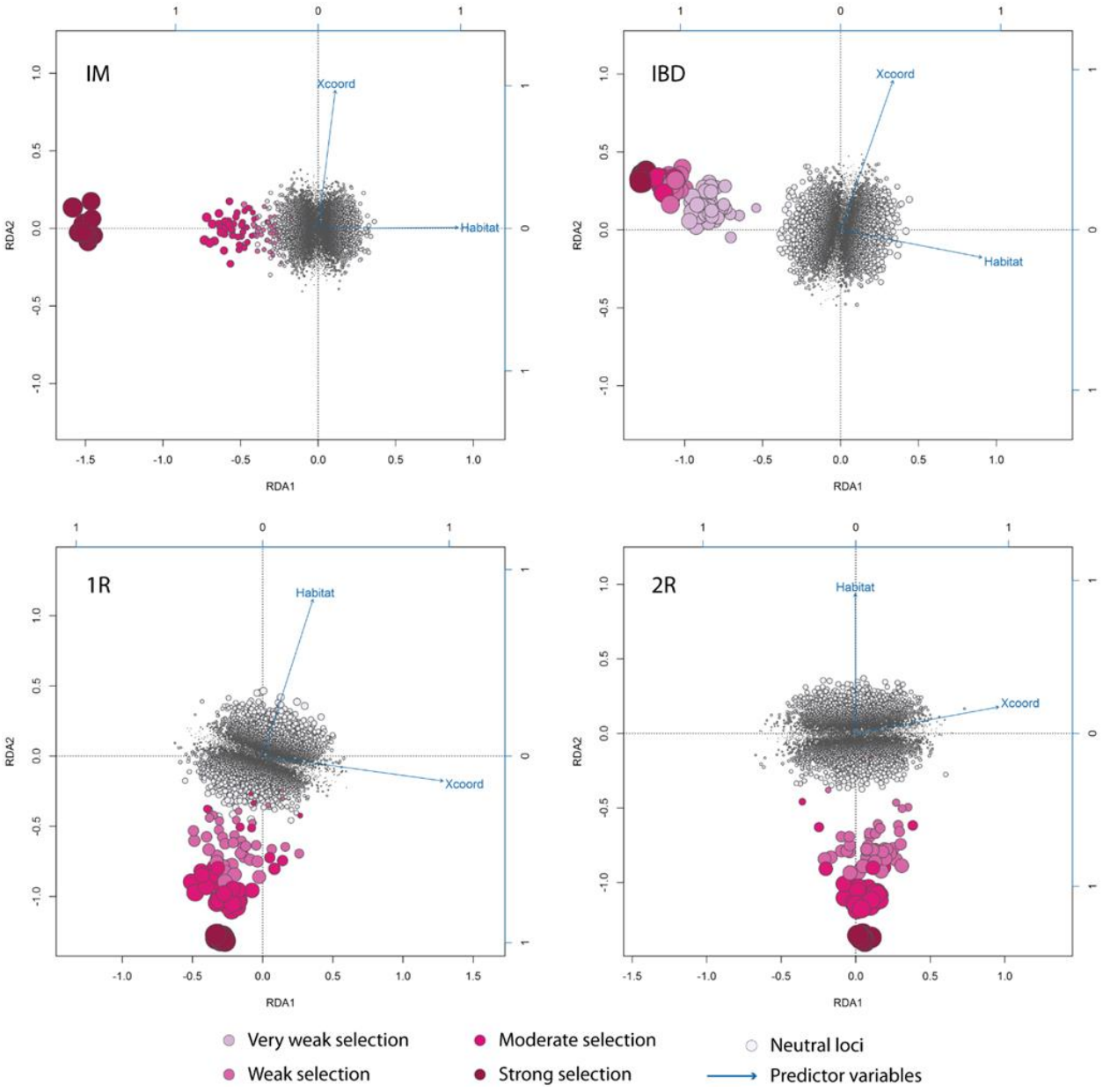


Figure 6. Redundancy analysis plots showing loci with point size scaled by their correlation with the driving environmental variable ("Habitat"), and correlation of predictor variables with the constrained RDA axes (arrows). Plots are shown for an equilibrium island model (IM), equilibrium isolation by distance model (IBD), and non-equilibrium one- and two- refugial expansion models (1R and 2R) for paired sampling (6 individuals/deme) on environmental surface "453". Scaling by correlation with "Xcoord" is provided for comparison in Fig. S8.