Supplementary Material

The effect of neutral recombination variation on genome scans for selection

KE Lotterhos

List of Figures

Supplemental Methods

Selective sweep - integrated haplotype score (iHS). The integrated haplotype score, or iHS, measures local haplotype diversity within a single population, and is designed to detect recent hard selective sweeps that reduce local haplotype diversity (Voight et al. 2006). iHS is a measure of the amount of extended haplotype homozygosity (Sabeti et al. 2002) at a given SNP along the ancestral allele relative to the derived allele (Voight et al. 2006). iHS was calculated with scikit-allel v1.1.10.

Selective sweep - H_{12}. The test statistic H_{12} estimates haplotype homozygosity by combining the frequencies of two most frequent haplotypes into a single frequency and adding it to the total haplotype homozygosity (Messer and Petrov 2013; Garud et al. 2015). The H_{12} statistic increases with strong and recent adaptation (Schlamp et al. 2016), and has similar sensitivity for both hard and soft sweeps, as long as the latter only comprise a few frequency components (Messer and Petrov 2013; Garud et al. 2015). H_{12} was calculated with scikit-allel v1.1.10.

Selective sweep - H_2/H_1. The test statistic H_2/H_1 compares the haplotype homozygosity using all but the most frequent haplotype to the total haplotype homozygosity (Messer and Petrov 2013; Garud et al. 2015). This value is expected to be small for hard sweeps (when H_1 is large because one haplotype reaches high frequency in the sample) and large for soft sweeps (because multiple adaptive alleles arise on different haplotypes simultaneously, resulting in two or more common haplotypes at similar frequency) (Messer and Petrov 2013; Garud et al. 2015). H_2/H_1 was calculated with scikit-allel v1.1.10. Because the simulations only included hard sweeps, this statistic was transformed to $-log_{10}(H_2/H_1)$ for performance evaluation.

Differentiation outlier - OutFLANK (F_{ST}). OutFLANK seeks to identify loci with larger F_{ST} than expected by neutrality (Whitlock and Lotterhos 2015). The program assumes that the neutral F_{ST} is chi-squared distributed, and the neutral parameterization uses maximum likelihood to estimate the mean F_{ST} and degrees of freedom for neutral loci by trimming off the tails of the F_{ST} distribution. These parameters are then used to calculate P-values for all loci. Individuals were grouped into populations based on proximity to each other on the landscape (Figure 1). The nave approach was evaluated using the $-log_{10}$ P-values that resulted from running the algorithm on all SNPs. The best practice was evaluated using the $-log_{10}$ P-values that resulted from running the algorithm in two steps: first, using a quasi-independent set of thinned SNPs to estimate the neutral mean F_{ST} and degrees of freedom, and then second, using these estimates to parameterize the chi-square distribution for obtaining P-values for all SNPs. In both cases, default values were used for trimming (H_e ¿ 0.1, and left and right trim fractions of 0.05). The algorithm was implemented with the R package OutFLANK v0.2.

Differentiation outlier - PCAdapt. PCAdapt seeks to identify loci that have outlier loadings along PC axes that describe population structure, and does not require individuals to be assigned to populations (Duforet-Frebourg et al. 2014; Luu et al. 2017). For each SNP, the procedure computes a Mahalanobis distance on a vector of z-scores that corresponds to the z-scores obtained when regressing a SNP by the K PCs. The neutral parameterization in this case are the PC axes that describe population genetic structure. P-values for each locus are obtained from a chi-squared distribution with K degrees of freedom. I chose the best value of K based on a scree plot according to Cattells rule (Cattell 1966) and determined $K = 3$. The algorithm was implemented with the 'snp_pcadapt' function in the R packages bigsnpr v 0.2.1 and bigstatsr v 0.2.3 (Priv et al. 2018). The nave approach was evaluated using the $-log_{10}$ P-values that resulted from running the algorithm on all SNPs with MAF > 0.01. The best practice was implemented by using the quasi-independent thinned set of SNPs to estimate the PC axes, and then using these obtain $-log_{10}$ P-values from the regression for all SNPs with MAF > 0.01 (note that the method used for SNP thinning was more comprehensive than that implemented with the LD.clumping option in the new version 4.0 of the R package pcadapt).

Differentiation outlier - BayPass (X^TX). BayPass (v2.1) implements the model of Bayenv2 (Gnther and Coop 2013) as well as some extensions (Gautier 2015). The X^TX measures the degree to which loci are differentiated among populations, with neutral parameterization described by the covariance in allele frequencies among populations. BayPass was implemented with the standard model with default settings. The nave approach was evaluated using the X^TX values that resulted from running the algorithm on all SNPs with MAF > 0.01. The best practice was implemented by using the quasi-independent thinned set of SNPs to estimate the covariance matrix, and then using these to calculate X^TX for all SNPs with MAF > 0.01.

GWAS - LFMM ridge and LFMM lasso. Ridge regression and lasso are two different methods for adjusting for coefficient inflation in large data sets with collinear predictor variables (Caye et al. 2018). Ridge regression adjusts for coefficient inflation by minimizing the residual sum of squares of predictors with a ridge penalty, which adjusts all model coefficients by a shrinkage term (thus either all predictors are included in the final model, or none are). Lasso adjusts for coefficient inflation by minimizing the residual sum of squares with a lasso penalty, which is equal to the sum of absolute value of model coefficients (thus, the resulting coefficients for some predictors could be near zero, indicating they are dropped from the model). I implemented both types of regression in the R package lfmm v 2.0, with the functions 'lfmm_ridge' and 'lfmm_lasso' (Caye et al. 2018), with $K = 3$ for consistency with other methods.

GEA- latent-factor mixed model. A genetic-environment association (genotype as a function of population-mean environment) was implemented with a latent factor mixed model (LFMM) with the function lfmm in the R package LEA v 1.8.1 (Frichot and Franois 2015). The GEA was modeled as a Bayesian hierarchical model in which the LFMM parameters were estimated with a Gibbs sampler algorithm (Frichot et al. 2013). The number of latent factors was chosen by the cross-entropy criterion, which approximates the number of ancestral populations (and is closely linked to the number of principal components that explain variation in the genomic data, Frichot and Franois 2015). Based on this criterion and for consistency with pcadapt, I chose $K = 3$. The algorithm calculates z-scores for each locus and the corresponding P-values, which were averaged over

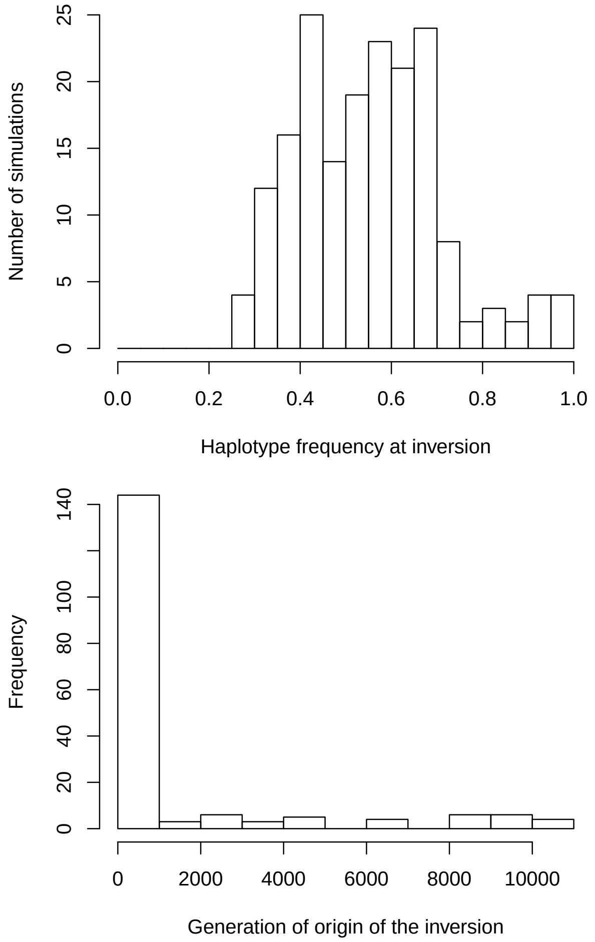
3 replicate runs. Similar to the GWAS models, the latent factors and model coefficients are estimated jointly, and the algorithm can only be run on all SNPs at the same time.

GEA - BayPass (Bayes Factor). BayPass also allows to test for associations with population-specific covariates, and for each locus calculates a Bayes Factor (BF) that reflects the degree of support for the hypothesis that the association does not equal zero, compared to the hypothesis that the association equals 0. We ran BayPass with the covariate being the population-mean environment (GEA). The nave approach and best practice were implemented as described for X^TX above, except the evaluation was done on $log_{10}(BF)$ instead of X^TX.

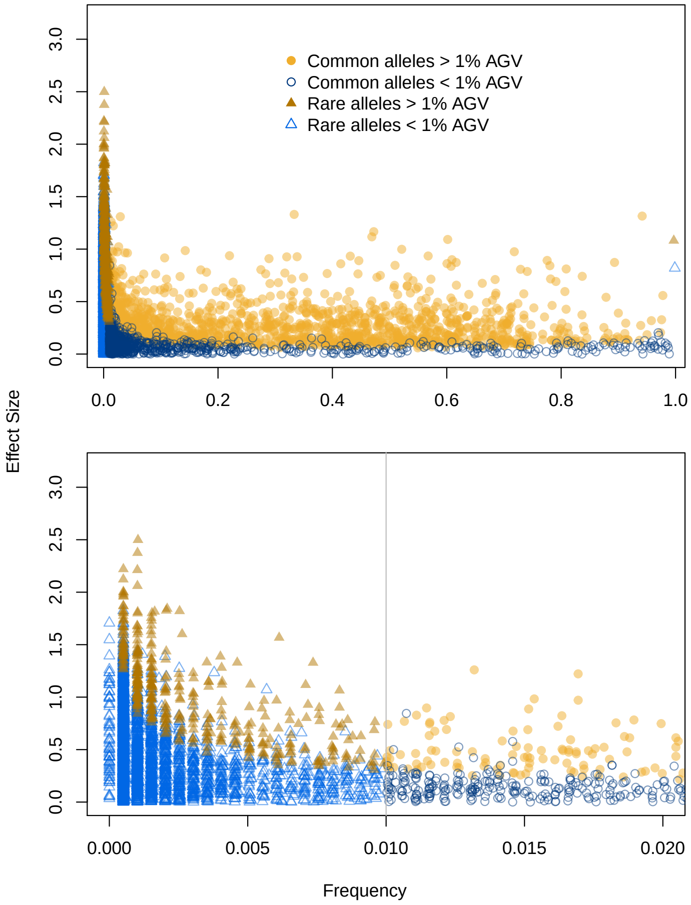
GEA - Spearmans ρ. The nonparametric rank-order association was calculated with Spearmans ρ between genotype and environment for each SNP. This measure does not correct for population structure.

GEA - Redundancy analysis. Redundancy analysis (RDA) is a method to extract and summarize the variation in a set of response variables that can be explained by a set of explanatory variables (Legendre and Legendre 2012). In this case, the response variables are the SNPs and the explanatory variable is the environment. Each SNP receives a loading on the constrained axis (e.g., the environmental predictor) and a loading on many unconstrained axes (e.g., principal components that capture structure in the SNP data). The loading on the constrained axis was used to evaluate performance of the RDA. RDA was implemented following the recommendations of Forester et al. (2018) using the rda function in the R package vegan v2.5.2. Note that RDA can only be performed on one set of SNPs as implemented in this function, so all SNPs were used for this analysis. Note that RDA does not have a neutral parameterization or correction for population structure in the association test, but Forester et al. (2018) found that including a structure correction in the RDA actually reduced power.

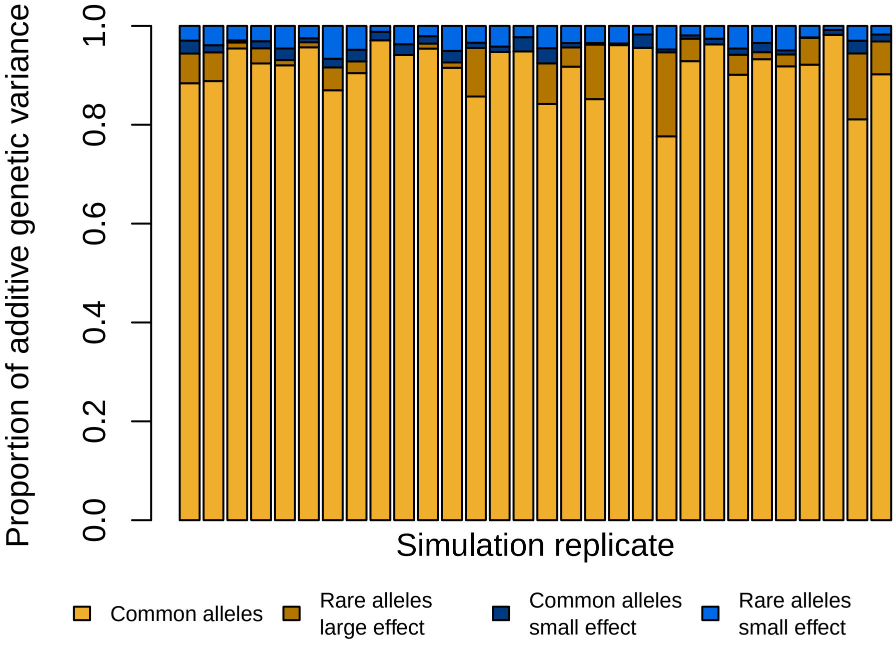
1. Supplementary Figures



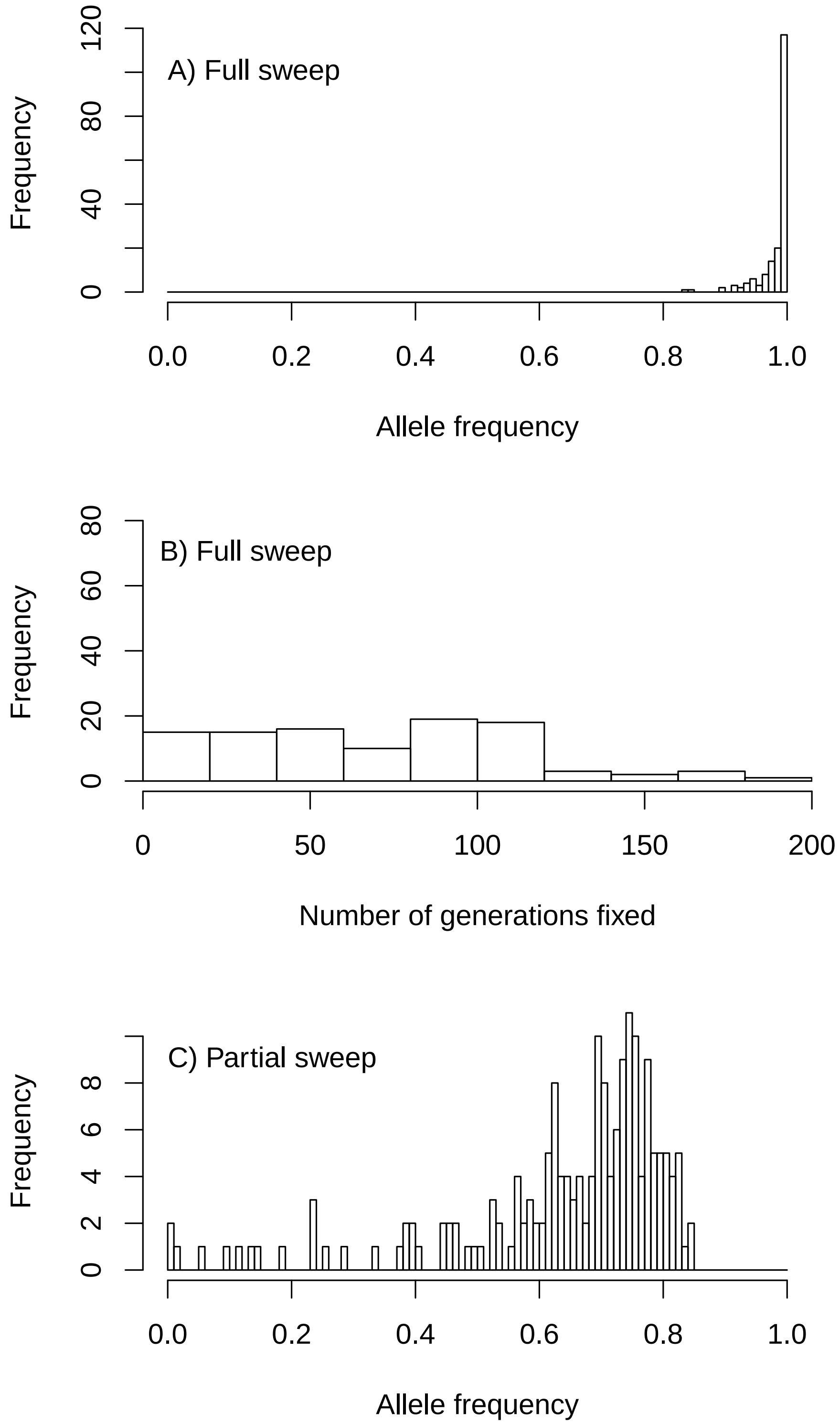
Supplementary Figure 1. The distribution of minor haplotype frequency at the inversion (top) and number of generations since the origin of the inversion (bottom).



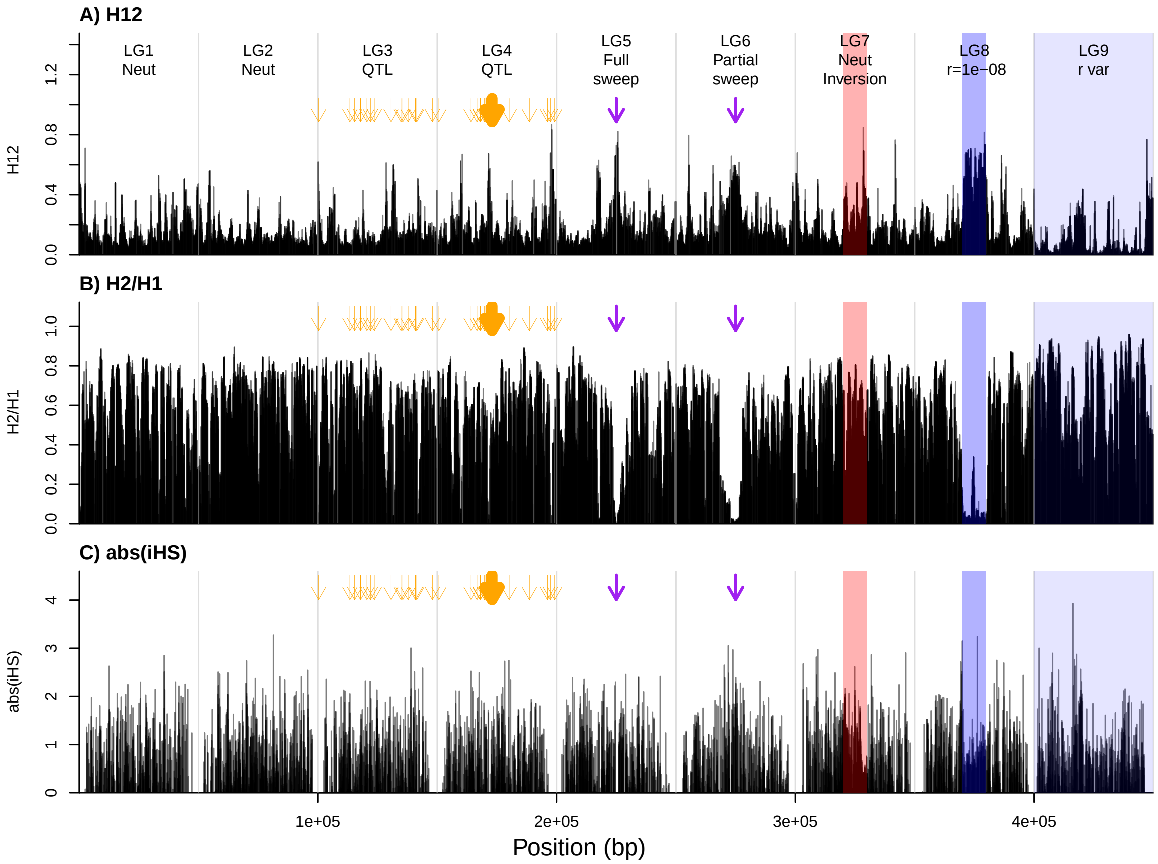
Supplementary Figure 2. The distribution of frequencies and effect sizes for different categories of causal loci across all replicate simulations. Rare alleles were categorized as loci with less that a frequency of 1%, while common alleles had a frequency of greater than 1%. Loci were further categorized whether they explained greater or less than 1% of the additive genetic variance. The bottom panel is zoomed in on alleles with low minor allele frequency of the derived allele (left side of top plot).



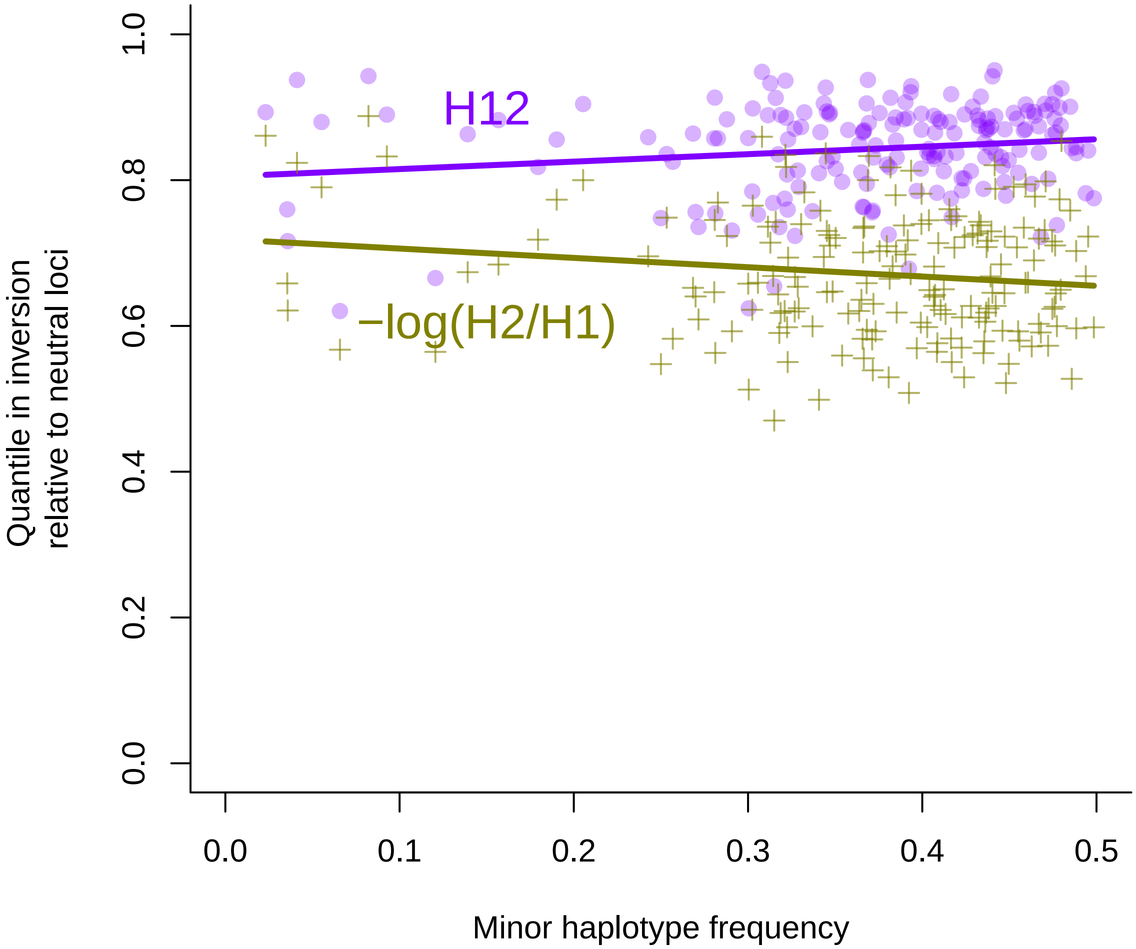
Supplementary Figure 3. The proportion of additive genetic variance explained by different categories of loci across a subset of replicate simulations. For definition of categories see main text.



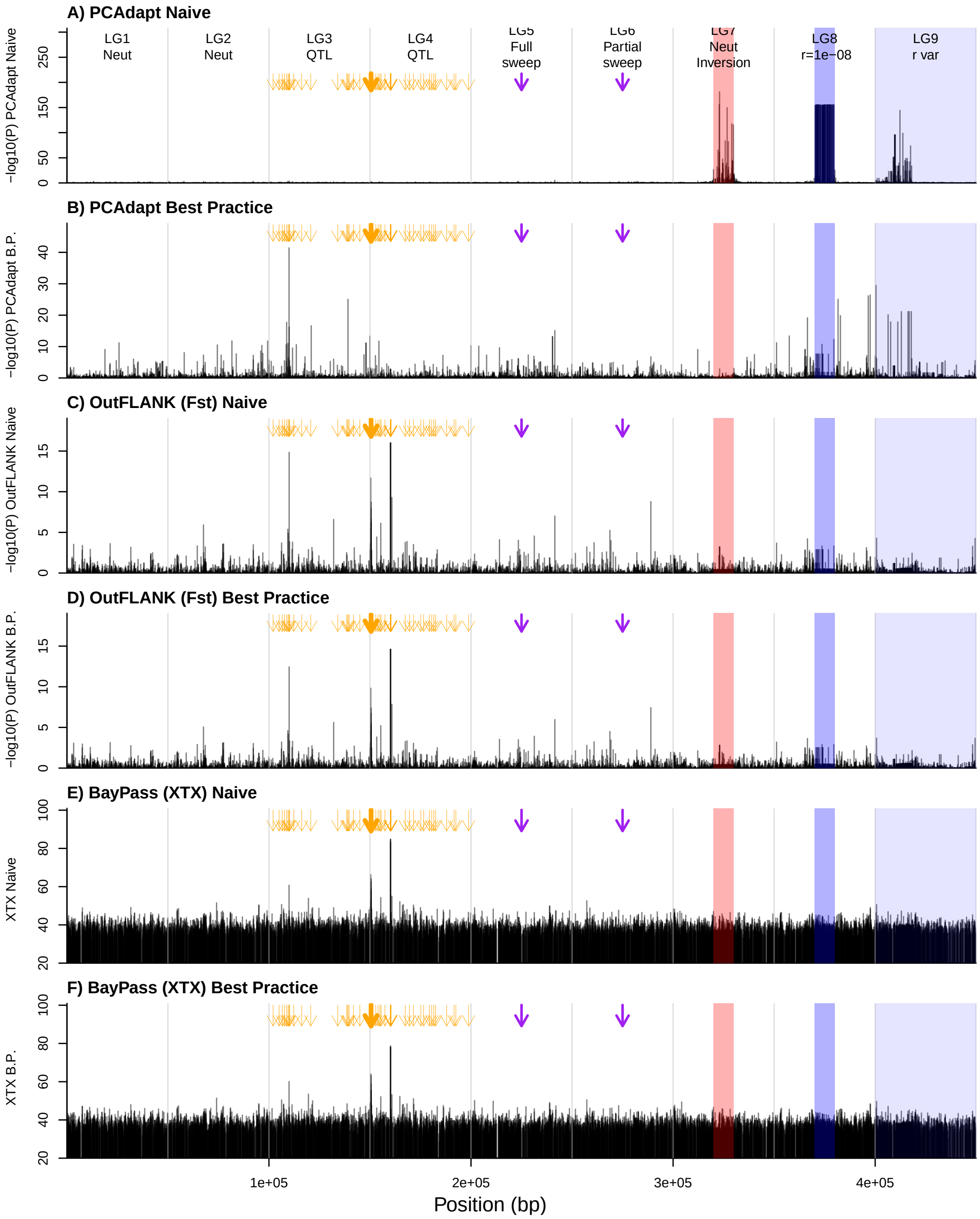
Supplementary Figure 4. Distributions of (A) the final allele frequency of the full sweep simulated on LG-5, (B) for fixed mutations, the number of generations since fixation for sweep mutations simulated on LG-5, and (C) the final allele frequency of the partial sweep simulated on LG-6.



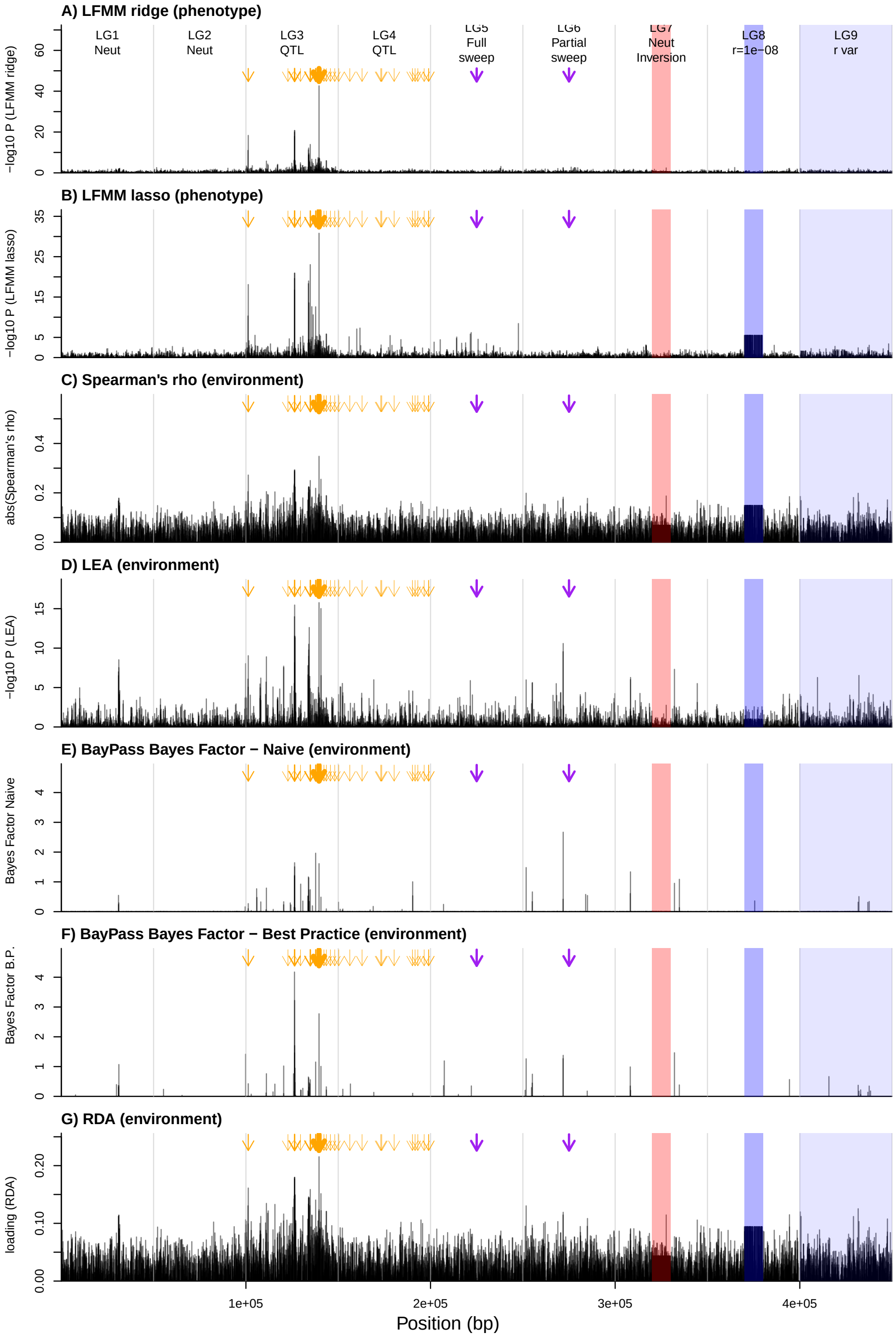
Supplementary Figure 5. Manhattan plot for selective sweep methods. The purple arrows show the locations of the hard sweeps. The orange arrows show the locations of the QTNs, and the arrow thickness is proportional to the percent of the additive genetic variance explained by each QTN.



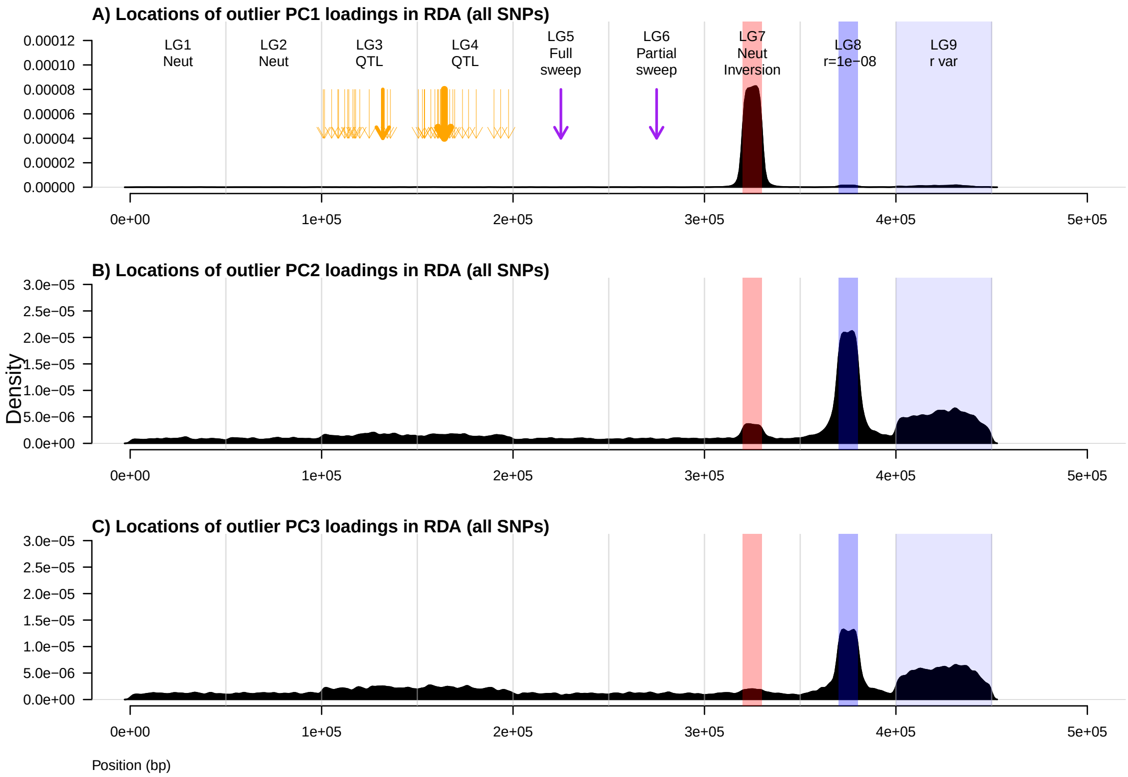
Supplementary Figure 6. Scatterplot of the empirical quantile of each statistic (relative to neutral loci) as a function of minor haplotype frequency at the inversion. Lines show best fit linear model.



Supplementary Figure 7. Manhattan plots for differentiation outlier methods from one replicate simulation. The orange arrows show the locations of the QTNs, and the arrow thickness is proportional to the percent of the additive genetic variance explained by each QTN. The purple arrows show the locations of the selective sweeps. The "naive" approach used all SNPs for neutral parameterization, while the "best practice" used a set of SNPs that had been thinned for LD for neutral parameterization.



Supplementary Figure 8. Manhattan plots for association methods for one replicate simulation. The orange arrows show the locations of the QTNs, and the arrow thickness is proportional to the percent of the additive genetic variance explained by each QTN. The purple arrow shows the location of the selective sweep. The "naive" approach used all SNPs to estimate the neutral population structure, while the "best practice" used a set of SNPs that had been thinned for LD.



Supplementary Figure 9. Loadings of loci onto the unconstrained axes (e.g., PC axes) from the Redundancy Analysis (RDA) across all replicate simulations. Each panel shows the frequency distribution of genomic locations that had outlier PC loadings (95%) for that scenario. A) Loadings on the first unconstrained axis. B) Loadings on the second unconstrained axis. C) Loadings on the third unconstrained axis.

Caye K., B. Jumentier, and O. Francois, 2018 LFMM 2.0: Latent factor models for confounder adjustment in genome and epigenome-wide association studies. BioarXiv. https://doi.org/10.1101/255893

Duforet-Frebourg N., E. Bazin, and M. G. B. Blum, 2014 Genome scans for detecting footprints of local adaptation using a Bayesian factor model. Mol. Biol. Evol. 31: 24832495.

Forester B. R., J. R. Lasky, H. H. Wagner, and D. L. Urban, 2018 Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. Mol. Ecol. 27: 22152233.

Frichot E., S. D. Schoville, G. Bouchard, and O. Franois, 2013 Testing for associations between loci and environmental gradients using latent factor mixed models. Mol. Biol. Evol. 30: 16871699.

Frichot E., and O. Franois, 2015 LEA: An R package for landscape and ecological association studies. Methods Ecol. Evol. 6: 925929.

Garud N. R., P. W. Messer, E. O. Buzbas, and D. A. Petrov, 2015 Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet. 11: e1005004.

Gautier M., 2015 Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics 201: 15551579.

Gnther T., and G. Coop, 2013 Robust identification of local adaptation from allele frequencies. Genetics 195: 205220.

Legendre P., and L. F. J. Legendre, 2012 Numerical Ecology. Elsevier.

Luu K., E. Bazin, and M. G. B. Blum, 2017 pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol. Ecol. Resour. 17: 6777.

Messer P. W., and D. A. Petrov, 2013 Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. 28: 659669.

Prive' F., H. Aschard, A. Ziyatdinov, and M. G. B. Blum, 2018 Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty185

Sabeti P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, et al., 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832837.

Schlamp F., J. van der Made, R. Stambler, L. Chesebrough, A. R. Boyko, et al., 2016 Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. Mol. Ecol. 25: 342356.

Voight B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. PLoS Biol. 4: e72.

Whitlock M. C., and K. E. Lotterhos, 2015 Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST. Am. Nat. 186 Suppl 1: S2436.