Supplementary Material

BlockFe$_{ST}$: Bayesian calculation of region-specific F_{ST} to detect local adaptation

Bastian Pfeifer and Martin J. Lercher

Supplementary Text: Simulations

To validate our approach and to compare it to alternative measurements, we performed extensive simulations using the MSMS program (Ewing and Hermisson 2010), an extension of Hudson's ms (Hudson 2002) that allows coalescent simulations for a structured population under selection.

Based on an *island model (background F_{ST} is identical between populations)* and a *divergence model (background pairwise F_{ST} differs between populations)*, we used MSMS to sample 950 loci under a neutral scenario and 50 loci under population-specific directional selection with a fixed number of segregating sites (s=20). We repeated the simulations with different recombination rates, start of selection, and varying background population history (*different coalescent times*). We also performed this analysis with variable mutation rates across the genome to show that *BlockFe*$_{ST}$ is also appropriate for gene scans where the number of SNPs generally differ between loci (Suppl. Figure S2).

We assessed the performance of *BlockFe*$_{ST}$ and compared it to those of BayeScan (Foll and Gaggiotti 2008) applied to individual SNPs, Hudson's F_{ST} (Hudson 1992), and the recently published PCA method implemented in the R-package pcadapt (Luu, Bazin and Blum 2017). In case of pcadapt and BayeScan, we used the P-values and calculated the sum of logs for each region to compare it to F_{ST} and *BlockFe*$_{ST}$.

To benchmark the ability of the different methods to detect positive selection, we plotted receiver-operator-characteristic (ROC) curves, which plot the fraction of true positives (sensitivity) versus the fraction of false positives (1 - specificity) at different cutoff values for the parameter used for discrimination. The area under this curve (AUC) is a cutoff-independent

measure of accuracy and was calculated with the R package pROC (Robin *et al.* 2011). In addition, as a measure of precision, we calculated the fraction of loci under positive selection correctly identified by the 5% most extreme values. We also tested *BlockFe*$_{ST}$ regarding computational speed (Suppl. Table S1).

Island model

We assume a population with an effective population size of N_e=10,000 that split into two subpopulations 4,000 ($4N_e$ x 0.1*)* generations ago. After the splitting event, there is no migration between the two populations. We use a fixed number of segregating sites (s=20) and set the sample size to 20 in each population. Directional selection is introduced 4,000 generations ago in both populations, with a selection strength of 0.01. The initial frequency of the beneficial allele is 0.01 in each population. To ensure that the selected allele does not get lost we switch on the -SFC parameter. The MSMS calls are:

Neutral:

msms 40 950 -s 20 -N 10000 -I 2 20 20 0 -ej 0.1 1 2

Positive selection:

msms 40 50 -s 20 -N 10000 -I 2 20 20 0 -ej 0.1 1 2 -SAA 200 -SaA 1 -SI 0.1 2 0.01 0.01 -SFC

For the *island model*, we report that *BlockFe*$_{ST}$ outperforms the alternative methods in almost all cases (Suppl. Figures S1-S4), especially when the signal of selection is not yet eroded by recombination (Suppl. Figures S1-S2). Even when neutral patterns are concatenated to the selected regions, *BlockFe*$_{ST}$ performs well as long as positive selection is the major signal in that region (Suppl. Figure S4). Surprisingly, the computationally simple moment estimator F_{ST} competes with the alternative method *pcadapt*.

Island model: Balancing selection

To test the ability of BayeScan, pcadapt, F_{ST}, and *BlockFe*$_{ST}$ to detect balancing selection, we introduce balancing selection 36,000 ($4N_e$ *0.9)* generations ago for the alternative model. The splitting event of the two populations is still set to be 4,000 ($4N_e$ x 0.1) generations ago. The MSMS calls are:

Neutral:
msms 40 950 -s 20 -N 10000 -I 2 20 20 0 -ej 0.1 1 2

Balancing selection:
msms 40 50 -s 20 -N 10000 -I 2 20 20 0 -ej 0.1 1 2 -SAA 1 **-SaA 200 -SI 0.9** 2 0.01 0.01 -SFC

In this simulation set-up, pcadapt, originally developed mainly to detect directional selection, is clearly the weakest method to detect balancing selection (Suppl. Figure S5). BayeScan, F_{ST}, and *BlockFe*$_{ST}$ show very similar AUC values, while the *BlockFe*$_{ST}$ results additionally indicate a high precision in the detection of outlier loci subject to balancing directional selection.

Divergence model

For the divergence model, we assume a population with an effective population size of N_e=10,000 that split into two subpopulation 8,000 ($4N_e$ x 0.2) generations ago. The next split occurs 4,000 ($4N_e$ x 0.1) generations ago. There is no migration between the two populations and the ancestral population. We use a fixed number of segregating sites (s=20) and set the sample size to 20 in each population. Directional selection is introduced 4,000 ($4N_e$ x 0.1) generations ago in one of the recently split populations as well as in the ancestral population, with a selection strength of 0.01. The initial frequency of the beneficial allele is 0.01. To ensure that the selected allele does not get lost we switch on the -SFC parameter. The MSMS calls are:

Neutral model:

msms 60 950 -s 20 -N 10000 -I 3 20 20 20 0 -ej 0.1 2 1 -ej x 3 1

Positive selection:

msms 60 50 -s 20 -N 10000 -I 3 20 20 20 0 -ej 0.1 2 1 -ej x 3 1 -SAA 200 -SaA 1 -SI 0.1 3 0 0.01 0.01 -SFC

When selection is strong, *BlockFe*$_{ST}$ is the best method under the divergence model judged by the AUC values (Suppl. Figures S6-S9). The power (precision) values, however, suggest slightly better results for pcadapt. The results indicate that *BlockFe*$_{ST}$ performs comparably well up to 20% neutrality eroding the signal of selection in the affected region (Suppl. Figure S9).

BlockFe$_{ST}$ Usage

```
# install the the PopGenome and BlockFeST package within R
install.packages('BlockFeST')
install.packages('PopGenome')
# Read in some data with PopGenome
genome = readData('FASTA')
# set the populations
pop1 = c('ind1','ind2','ind3')
pop2 = c('ind4','ind5','ind6')
genome = set.populations(genome, list(pop1,pop2))
# Extract SNP information from the genome class object
snps = getBayes(genome, snps=TRUE)
# Start BlockFeST
BlockFeST.result = BlockFeST(snps)
# Get the alpha values from the BlockFeST.result object
mean_alpha = BlockFeST.result@alpha
var_alpha  = BlockFeST.result@var_alpha
# Generate samples and calculate the empirical P-values
q = 0.95
iter = 1000
P_values = numeric(length(mean_alpha))
inc = numeric(length(mean_alpha))
for (x in 1:iter){
     samples1 = rnorm(rep(1,length(mean_alpha)),
                          mean(alpha),sqrt(var_alpha))
     quantile = quantile(samples1, q)
     samples2 = rnorm(rep(1,length(mean_alpha)),
                          mean(alpha),sqrt(var_alpha))
     inc     = inc + (samples2>quantile)
}
P_values = inc/iter
plot(BlockFeST.result@fst, P_values)
```

References

Ewing G, Hermisson J 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26: 2064-2065. doi: 10.1093/bioinformatics/btq322

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. doi: 10.1534/genetics.108.092221

Hudson RR 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337-338.

Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992) Estimating of levels of gene flow from DNA sequence data, Genetics, 13, 583-589.

Luu, K., Bazin, E. and Blum, M. G. B. (2017), *pcadapt*: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour, 17: 67–77. doi:10.1111/1755-0998.12592

Pfeifer B, Wittelsbuerger U, Ramos-Onsins SE, Lercher MJ. 2014. PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. 31:1929–1936

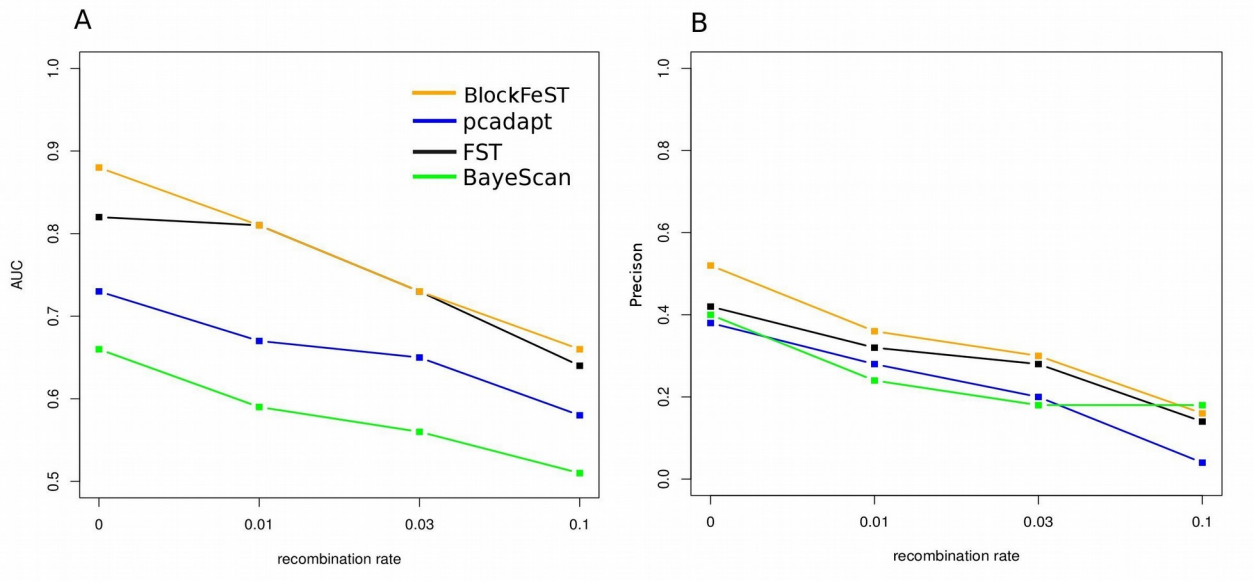
Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12: 77. doi: 10.1186/1471-2105-12-77

Supplementary Tables

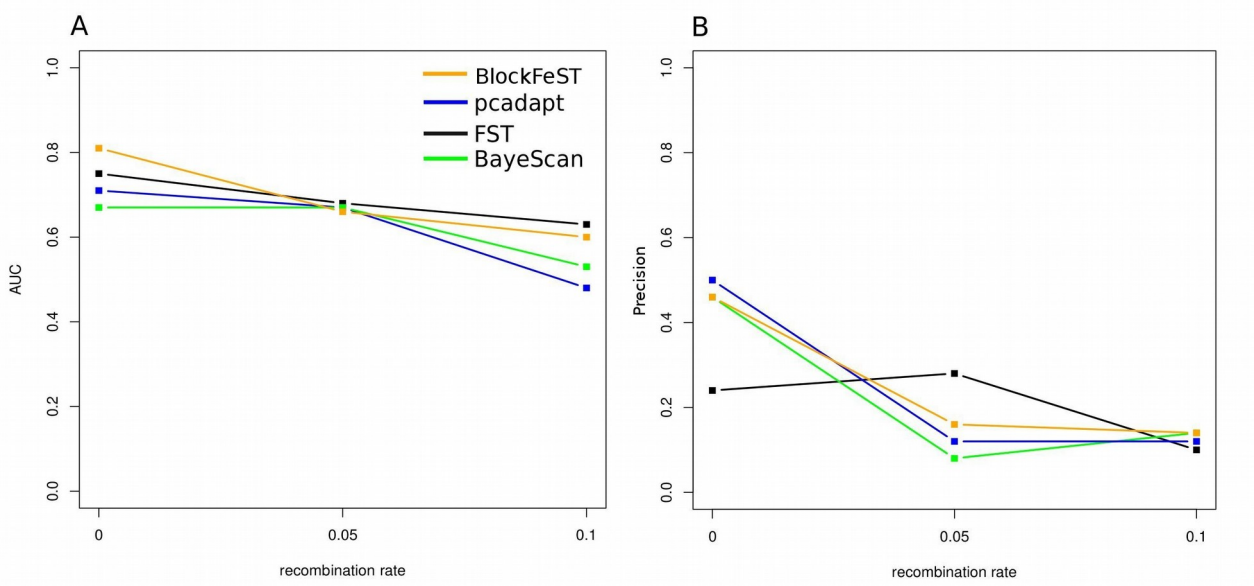
Table S1: *Computational speed of processing 20.000 SNPs and 3 populations including 20 individuals each. Hardware architecture: Intel® Core™ i3-2130 CPU @ 3.40GHz × 4*

Method	BayeScan	*BlockFe*$_{ST}$	F_{ST}	pcadapt
CPUs	4	1	1	1
Elapsed time	6.28 h	2.75 h	1.37 s	0.79 s

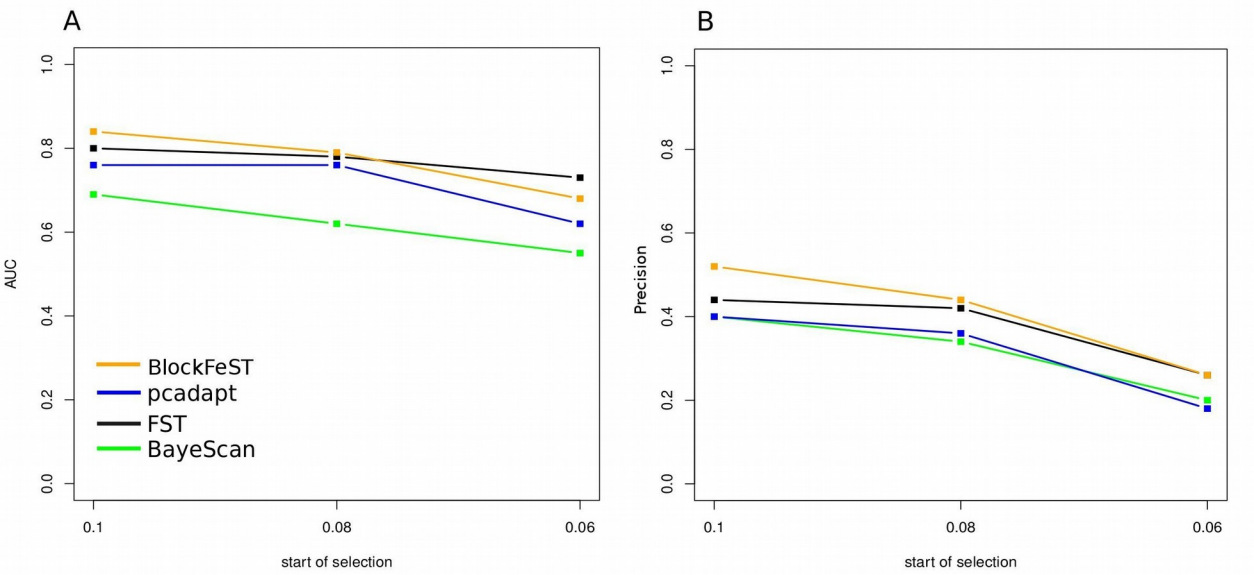
Supplementary Figures



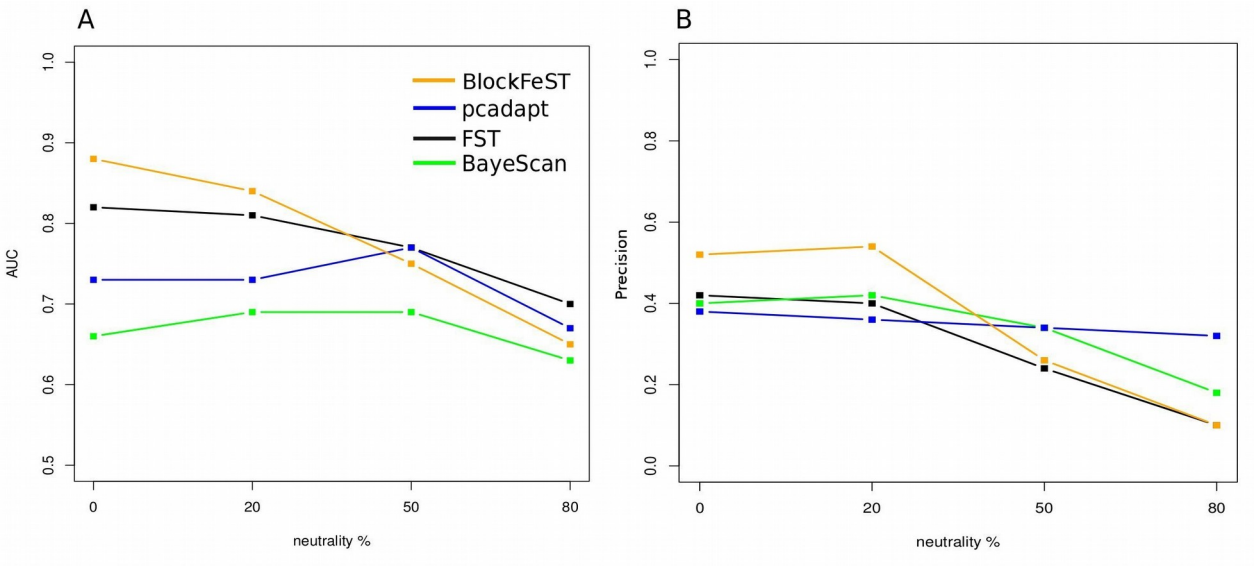
Supplementary Figure S1: Island model. *On the power to detect positive selection in case of increasing recombination rates.* Precision and AUC values were calculated for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$.



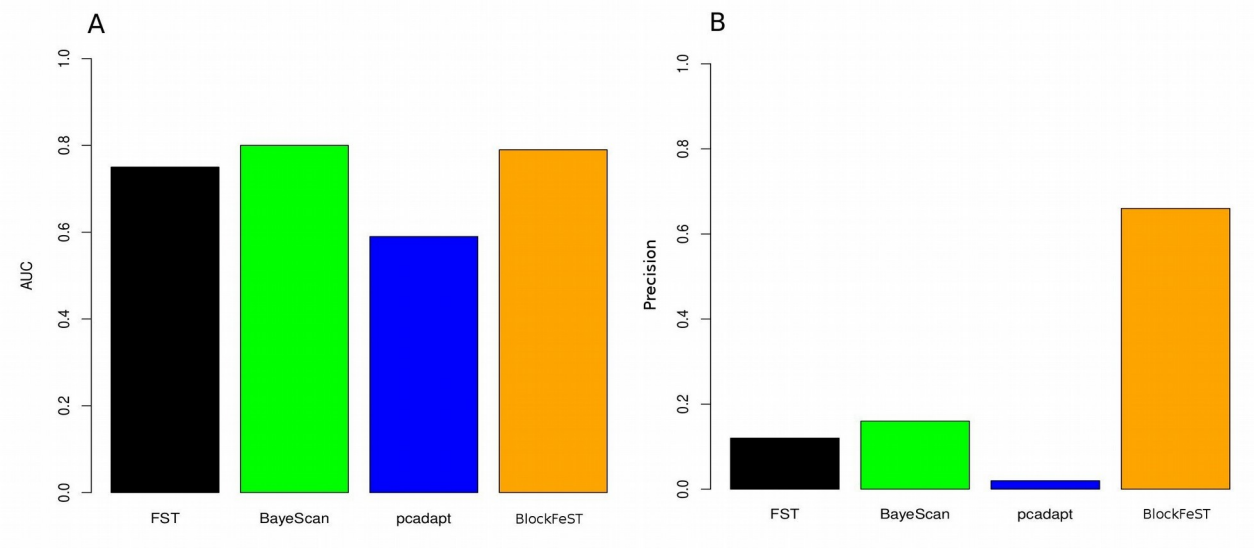
Supplementary Figure S2: Island model. *On the power to detect positive selection with increasing recombination rates and varying mutation rates across the genome*. For this specific scenario, we randomly sample *theta* (mutation rate) values out of [2,...,10] for the neutral loci and the loci under selection, resulting in a different amount of SNPs across regions. We show results based on the moment-based Hudson F_{ST} values, the maximal *P*-value in case of BayeScan, the max -$\log_{10}$ *P*-value of pcadapt, and the empirical *P*-values from *BlockFe*$_{ST}$.



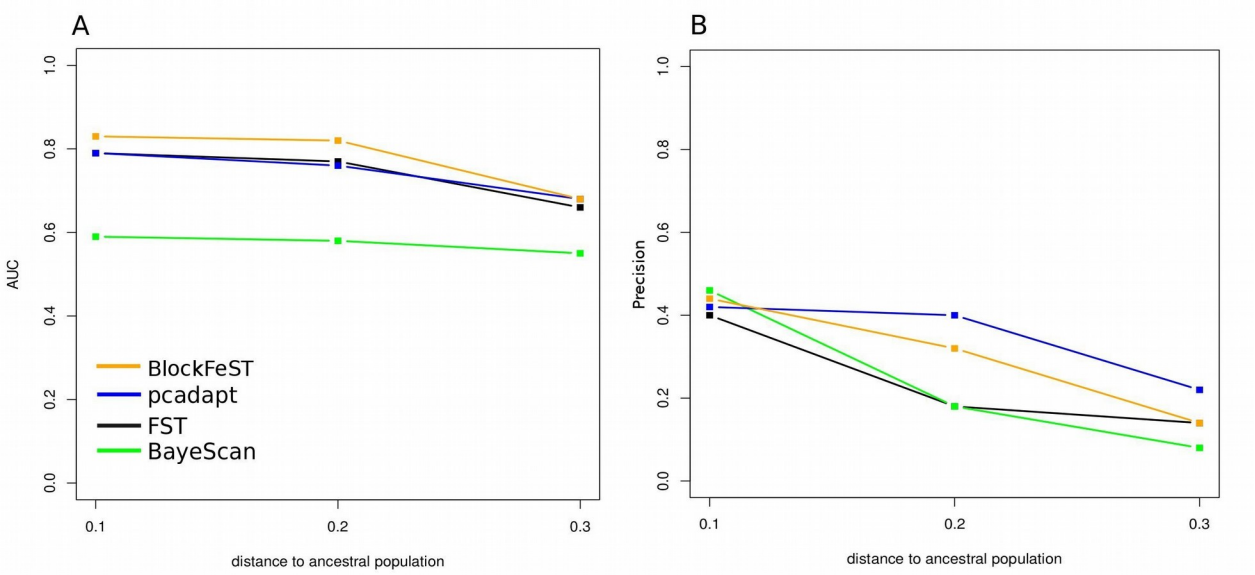
Supplementary Figure S3: Island model. *The effect of the start of selection*. We show AUC and Precision values for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$. The start of selection is 4,000 (0.1x4N_e), 3,200 (.08x4N_e) and 2,400 (0.06x4N_e) generations ago.



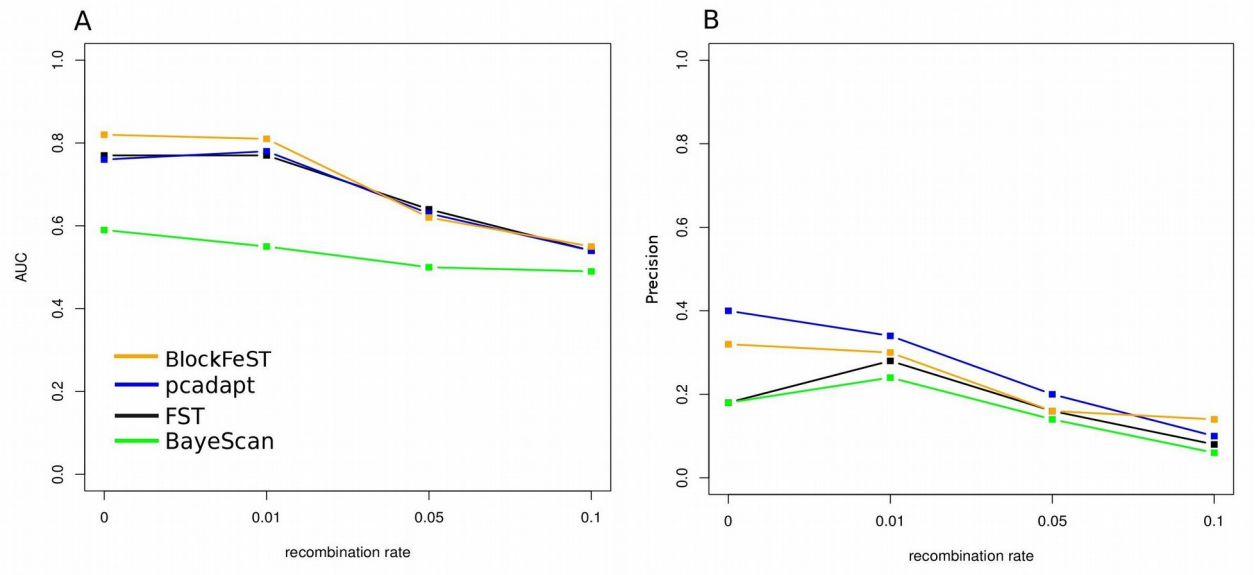
Supplementary Figure S4: Island model. *The effect of surrounding non-selected sequence.* We randomly sampled neutral SNPs out of the entire dataset and concatenate those to the ends of the selected regions. We show AUC and Precision values for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$.



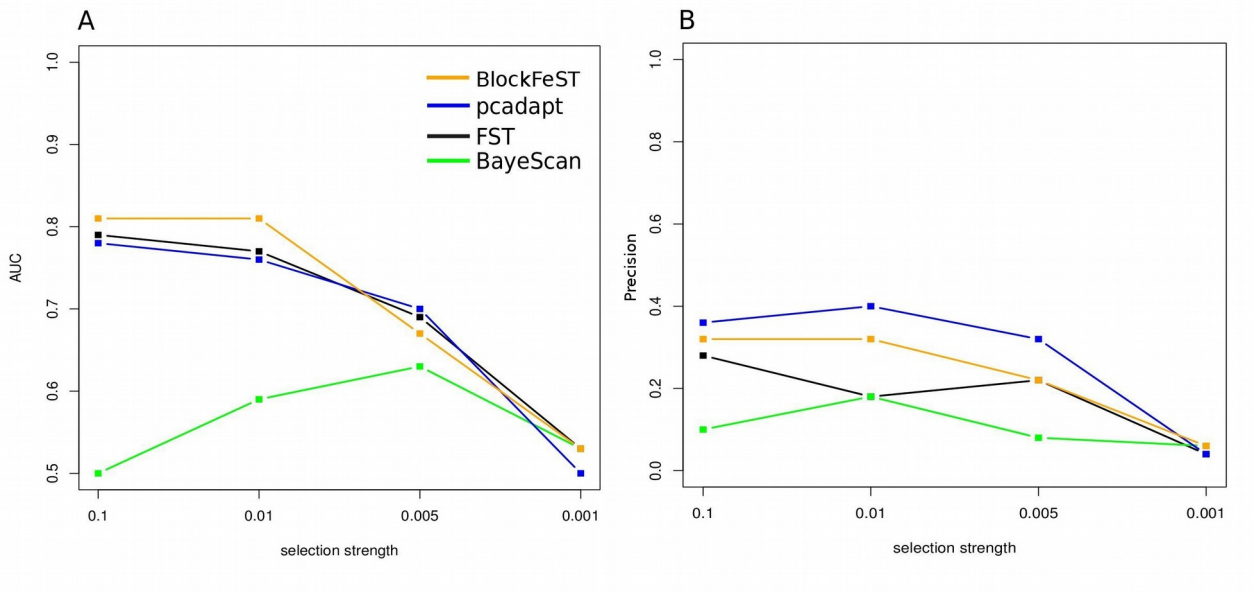
Supplementary Figure S5: Island model. *Balancing selection.* To test the methods' performance in detecting balancing selecting, we used the same neutral island model as before. Balancing selection is introduced 36,000 ($4N_e$ x0.9) generations ago, with a selection strength of 0.01 for the beneficial heterozygote alleles. We show AUC and Precision values for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$. In case of *BlockFe*$_{ST}$ and F_{ST}, loci subject to balancing selection are connected to low values and thus we used the 50 lowest values to calculate the Precision (panel B).



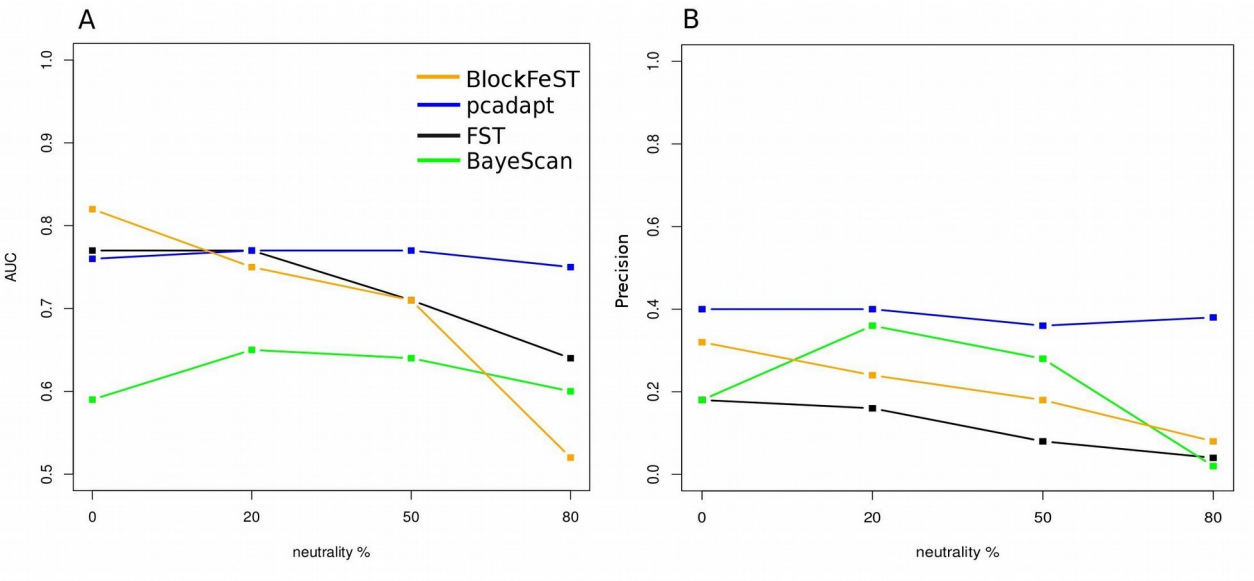
Supplementary Figure S6: Divergence model. *Background population history.* We show AUC and Precision values for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$. The ancestral population splits 4.000 ($4N_e$ x0.1) , 8.000 ($4N_e$x0.2), and 12.000 ($4N_e$ x0.3) generations ago.



Supplementary Figure S7: Divergence model. *The effect of recombination*. We show AUC and Precision values for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$. The ancestral population splits $4N_e$ = 8,000 generations ago.



Supplementary Figure S8: Divergence model. *The effect of selection strength.* We show AUC and Precision values for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$. The ancestral population splits $4N_e$ = 8,000 generations ago.



Supplementary Figure S9: Divergence model. *The effect of surrounding non-selected sequence.* We randomly sampled neutral SNPs out of the entire dataset and concatenate those to the ends of the selected regions. We show AUC and Precision values for the moment based Hudson F_{ST} values, the sum of log posterior *P*-values of BayeScan, the sum of log *P*-values of pcadapt, and the empirical *P*-values of *BlockFe*$_{ST}$. The ancestral population splits $4N_e$ = 8,000 generations ago.