Identification of Signatures of Selection in Dairy Cattle from Next-Generation Sequencing Data

Daniel Fischer[1], F. Panitz[2], A. Bagnato[3], E. Santus[4], J. Vilkki[1], M.A. Dolezal[3,5]

[1] MTT Agrifood Research Finland, Biotechnology and Food Research, Finland

[2] Aarhus University, Department of Molecular Biology and Genetics, Denmark

[3] Università degli Studi di Milano, Italy

[4] Associazione Nazionale Allevatori Razza Bruna, Italy

[5] University of Veterinary Medicine Vienna, Institute of Population Genetics, Department of Biomedical Sciences, Austria

Cardiff, 17th June 2014

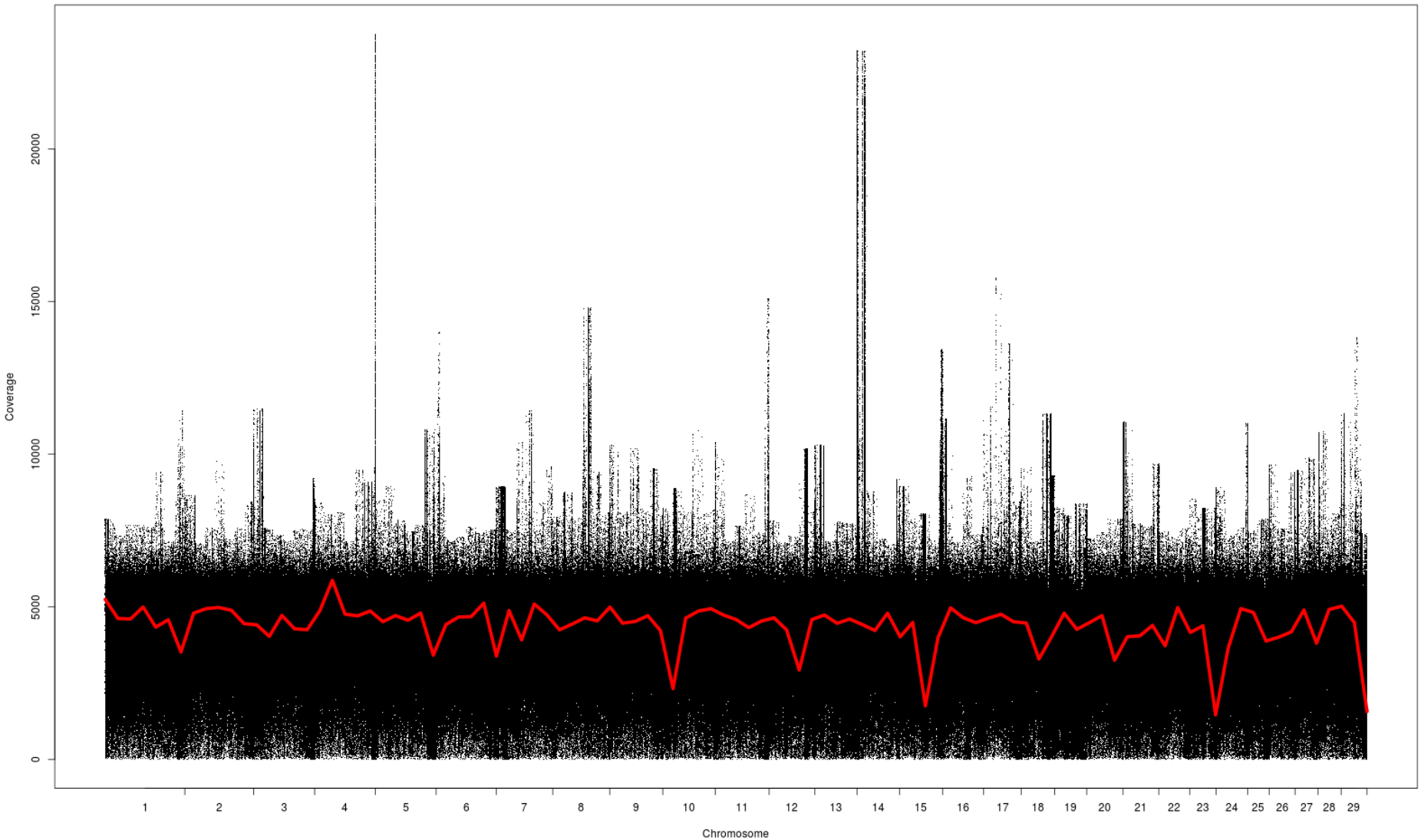
Signatures of Selection

- Selection leaves distinct traces in the genome, e.g.
  1. Reduced local variability
  2. Deviations in the Site Frequency Spectrum and
  3. Increased linkage disequilibrium and long haplotypes at high frequency
- Positions under selection are likely to be of functional importance!

The Dataset

- About 28.1 Mio autosome-wide SNPs called from Illumina paired-end NGS whole genome re-sequencing (20x).
- 20 Brown Swiss (BSW) bulls
- 17 Finnish Ayrshire (FAY) bulls
- Available from FP7 funded project QUANTOMICS
- SNPs called with SAMtools as part of the 1000 Bulls consortium
- Beagle v3 to reduce false positive calls and to phase data

Autosomal-wide coverage of multi-sample call dataset



Folded Minor Allele Frequency Spectrum

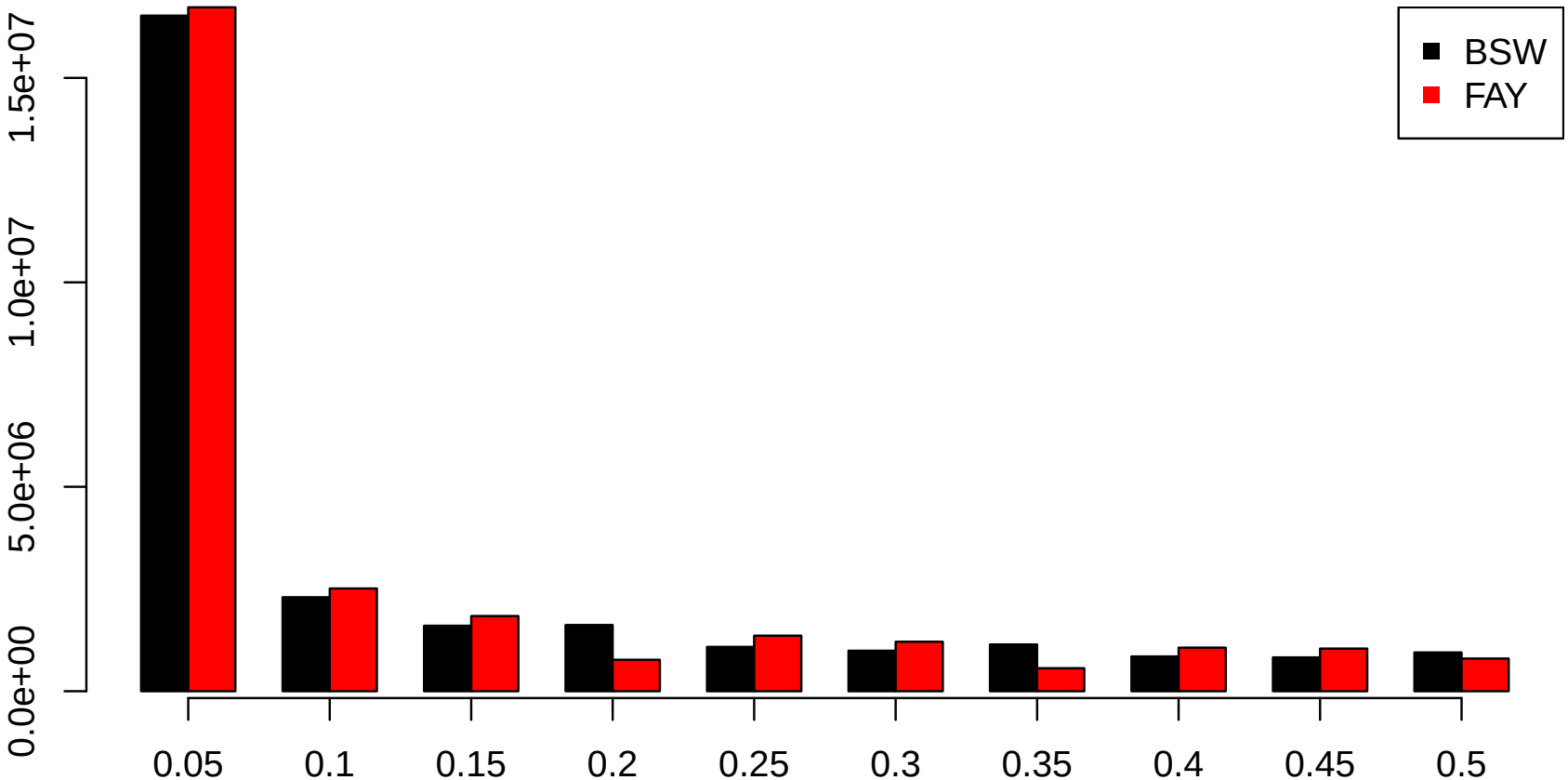


Figure 1 : Folded MAF

Signatures of Selection

We applied five different methods to detect signatures of selection:

1. Nucleotide Diversity Pi
2. Tajima D
3. Composite Likelihood Ratio test
4. Integrated Haplotype Scores
5. Fixation Index F_{st}

Notation

- We consider a set of n nucleotide sequences $\mathbb{X} = (x_1, x_2, \ldots, x_n)$
- The length of sequence x_i (in bases) is l_i with $i = 1, 2, \ldots, n$ and $l_{i,j}$ is the common length of two sequences i und j
- The relative frequency f_k with $k = 1, \ldots, m$ and $m \leq n$ is the relative abundance of sequences in $\mathbb{X}$. (Example: Let $n = 5$ and $x_1 = x_2 = x_4 \neq x_3 \neq x_5$, then $f_1 = \frac{3}{5}, f_2 = f_3 = \frac{1}{5}$)
- Lets S be the amount of segregating sites in $\mathbb{X}$ with $S \in [0, \max(l_{i,j})]$

Pi

- Pi is a statistic that measures the 'Nucleotide Diversity'.
- We calculated Pi for a stepping sequence length of 1000bp.
- The basic formula to calculate the statistic π is

$$\pi = \frac{m}{m-1} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} f_i f_j p_{ij}$$

  with p_{ij} being the ratio of nucleotide-wise differences between sequence i and j. Let $d_{i,j}$ be the amount of pairwise differences then is $p_{ij} = d_{ij}/l_{ij}$
- Hence, a large π indicates big differences between sequences in $\mathbb{X}$

Tajima's D

- Tajima's D is a statistic with the purpose to distiguish between a DNA sequence that evolves neutraly or what is possibly under a certain directional selection.
- The value of D also reacts to other factors, like demographical development and hence drawing overhasty conclusions might leed to wrong results!

- The mathematical formula for Tajima's D is

$$D = \frac{\pi - S/a_1}{\sqrt{V}}$$

  with a_1 being the harmonic series $\sum_{i=1}^{n-1} \frac{1}{i}$ and $V = \text{Var}[\pi - S/a_1]$
- Negative Tajima's D: Excess of rare variants (directional selection, positive or purifying; or population expansion)
- Positive Tajima's D: Excess of intermediate frequency variants (balancing or overdominant selection, population size reduction)

Composite Likelihood Ratio test (CLR) - SweeD tool

- A selective sweep is accompanied by the elimination of variation in the neighborhood of a benefitial mutation
- Affecting the local site frequency spectrum (=distribution of the expected number of polymorphic sites, SFS)
- The SFS is then used to detect selective sweeps, by calculating a CLR with a neutral model (=no sweeps) in the denominator and a model that allows Sweeps in the numerator, providing then a Likelihood for the presence of a sweep.
- Hence, here regions with large likelihoods are of special interest.
- Highest power for fixed sweeps

Integrated haplotype score - selscan tool

- iHS looks for exceptionally long high frequency haplotypes, highest power for ongoing sweeps
- Extended haplotype homozygosity is calculated for each core SNPs coded 0/1 (selscan is blind with respect to ancestral/derived state) for that are then iHH_0 and iHH_1 calculated ($=$ the integrated haplotype homozygosity).
- The (unstandarized) iHS values at a certain position are then

$$\ln\left(\frac{iHH_1}{iHH_0}\right)$$

Fixation Index F_{st}

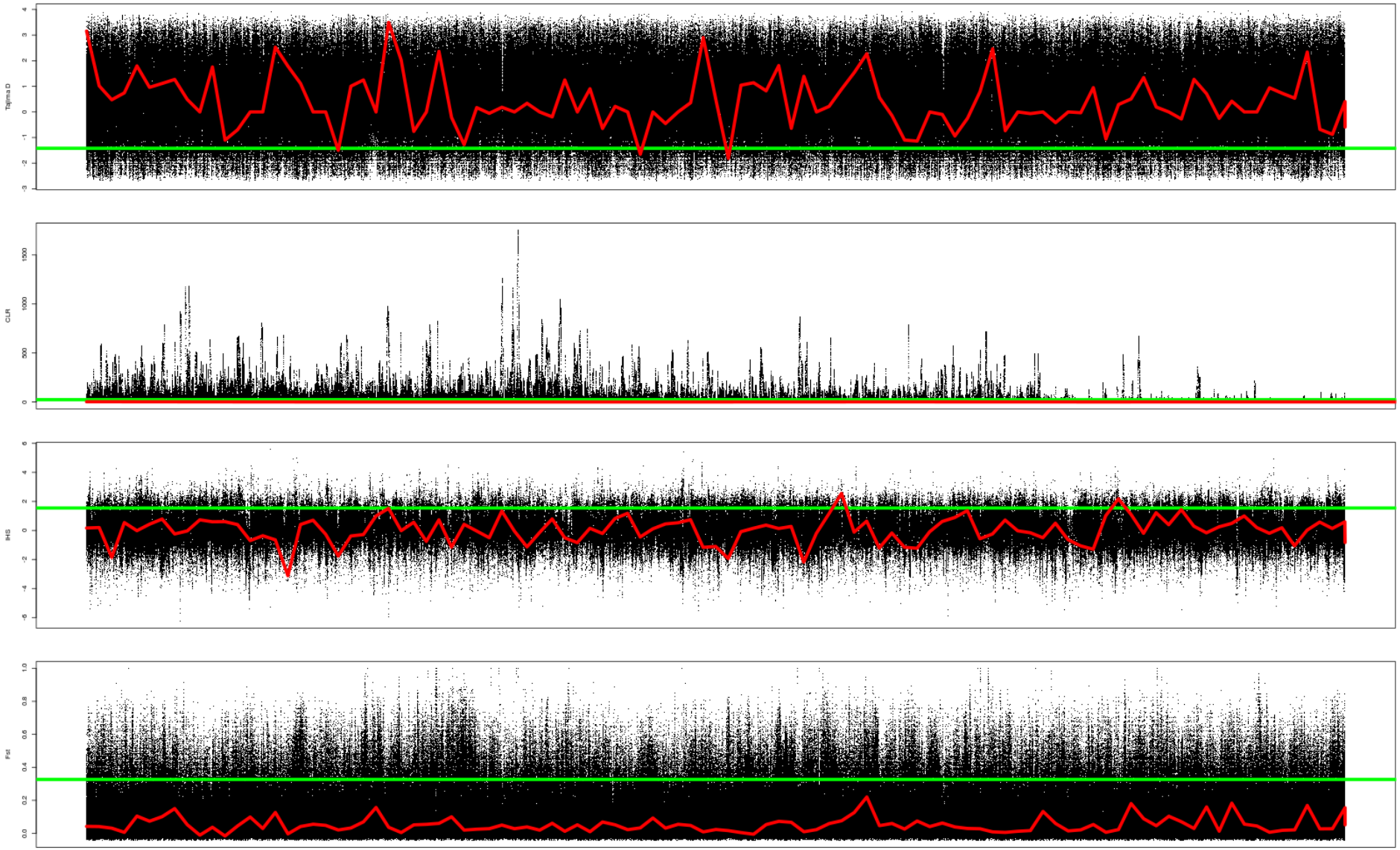
- The fixation index F_{st} is the only two-sample statistic we calculated.
- Having two subpopulations, one can calculate F_{st} from their average number of pairwise differences:

$$F_{st} = \frac{\pi_B - \pi_W}{\pi_B}$$

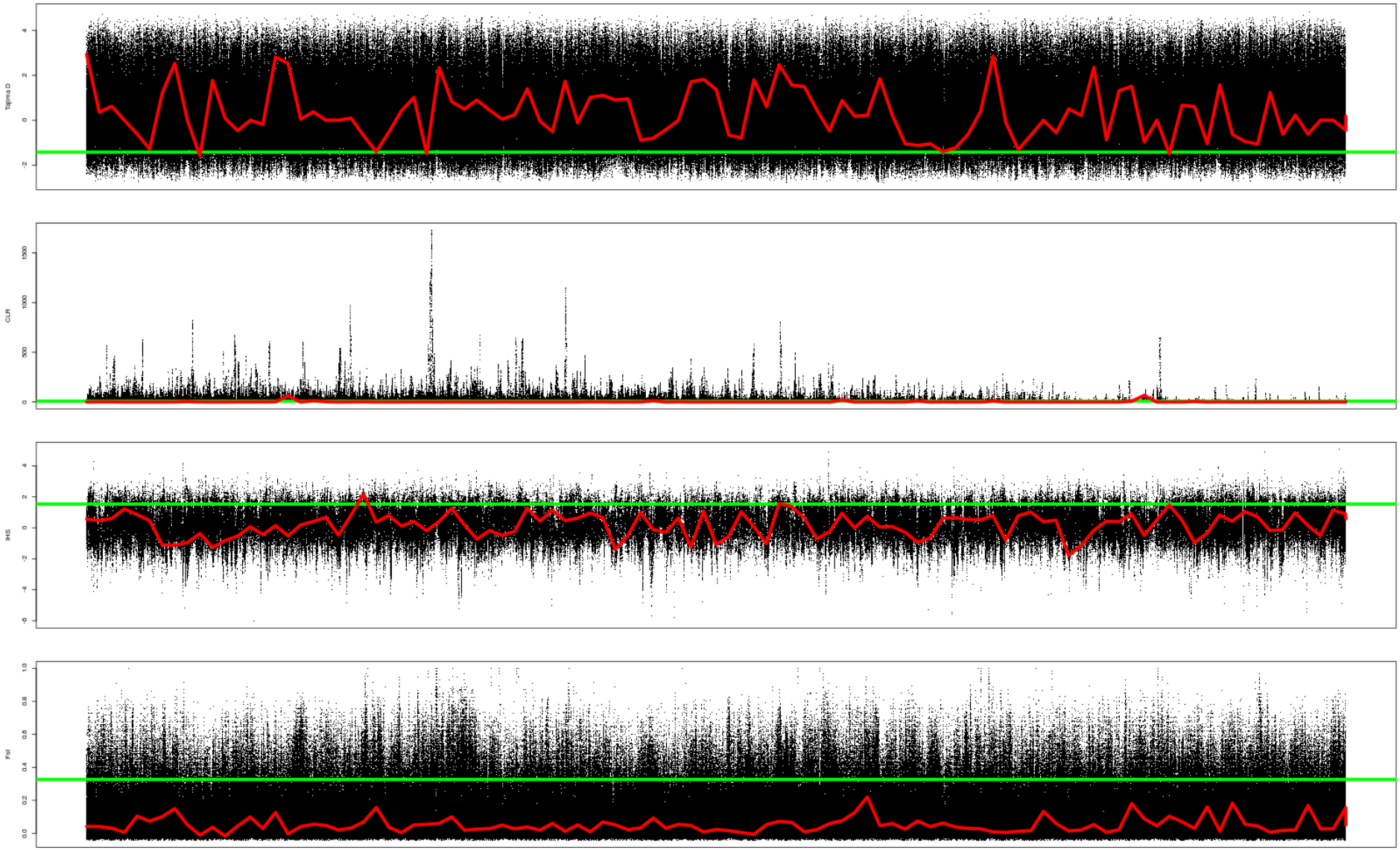
with π_B being the pairwise differences between the subpopulations and π_W the value for within subpopulation.

- F_{st} values close to zero indicate that the between and within differences are in the same level of quantity, whereas values close to one indicate that the 'within' differences are smaller compared to the 'between' differences

Visualization of the statistics - FAY



Visualization of the statistics - BSW



Histograms of the statistic values

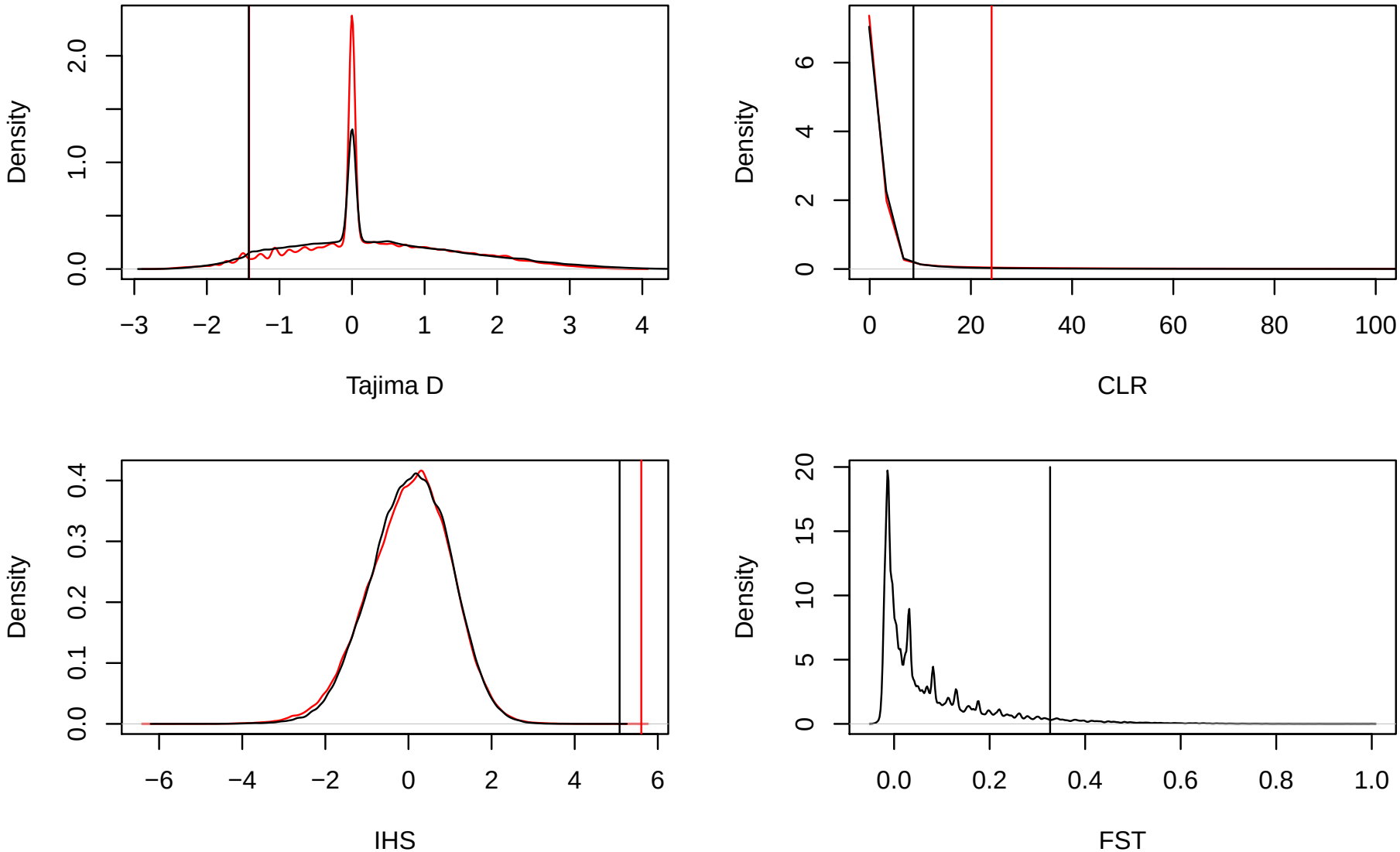


Figure 2 : Density plots of calculated statistics

Intersection with Annotation

- We intersected putative sweep regions with Ensembl v75 bovine gene set.
- For now we took the most extreme test statistics for each test and breed 5%.

Identified Regions and Genes - I

Statistic	Gen-hits	Bases (% of Genome)
Tajima D (FAY)	7252	10.96%
Tajima D (BSW)	6741	10.74%
CLR (FAY)	2581	3.49%
CLR (BSW)	4462	6.83%
iHS (FAY)	8289	13.25%
iHS (BSW)	7413	12.18%
Fst	9974	16.49%

Outlook

- Distiguish selection from demography
- GO pathway analysis