Statistical Problems in Livestock Population Genomics

H. Simianer[1], Y. Ma[2], S. Qanbari[1]
[1]Georg-August-University Goettingen, Germany, [2]China Agricultural University, Beijing, China

ABSTRACT: Identifying signatures of recent or ongoing selection is of high relevance in livestock population genomics. From a statistical perspective, determining a proper testing procedure and combining various test statistics is challenging. Based on an extensive simulation study we discuss the statistical properties of eight different selection signature statistics. It is demonstrated, that a reasonable power to detect selection signatures requires high density marker information as obtained from sequencing, while small sample sizes are acceptable. We suggest a novel principal component based combination of different statistics, which yields a statistic with similar power as the best single statistic but with an improved positional resolution. An accurate and comprehensive set of selection signatures will be the basis for a better understanding of the forces driving artificial selection and will help to design more efficient livestock breeding programs.
Keywords: selection signatures; statistical testing; statistical power

Introduction

The term "population genomics" was first introduced by Gulcher and Stefansson (1998) and is used for the analysis of genetic variation on a whole genome basis within or across populations. Certain characteristics of populations (like effective population size or population differentiation) can be inferred using population genomic approaches. Another important field of application is the detection of so-called "selection signatures", well defined patterns in the genome generated by selective forces. In the livestock breeding context it is hoped that identification of selection signatures leads to a better understanding of how selection operates on the genomic scale, which in turn may help to create more efficient selection schemes.

Detecting selection signatures poses a number of statistical challenges. In this contribution, we address some important statistical problems in this context, review suggested solutions, and finally illustrate some of these problems on the basis of an extended simulation data set.

Statistical problems

Selection signature statistics. Selection signatures in most cases are computed from genomic data only, i.e. from high-throughput genotyping data or from whole genome sequence data. Usually the entire genome is systematically scanned for such signatures. Selection produces the following basic signals in the proximity of a selected locus (i) the allele frequency spectrum is shifted towards extreme (high or low) frequencies, (ii) there is an excess of homozygous genotypes, (iii) long haplotypes exist with high frequency, and (iv) local population differentiation is extreme. All selection signature statistics that have been suggested pick up one or a combination of these signals. In most cases, a whole genome screen is conducted in that for each single locus (mostly SNPs) a value of the chosen test statistic is calculated. In some cases, point-wise statistics are strongly masked by random noise, so that moving, sliding, or creeping windows approaches are used to smoothen the picture and remove the noise.

In any case, this results in a vector of test statistic values, the length of the vector being the number of loci which is typically in the order of $10^5 - 10^6$ (for SNP array based analyses) or $>10^7$ for sequence-based analyses. Notably we must assume a strong autocorrelation structure in this vector, since neighboring SNPs are known to be in linkage disequilibrium (see e.g. Qanbari et al. (2010)) and hence statistics calculated at these loci are expected to be correlated as well. Such a vector can be available for a single or for multiple test statistics, in the latter case it must be assumed that different test statistics might be correlated, since many of the suggested statistics reflect the same basic signal.

Assuming the situation described, the major statistical challenge is to identify 'significant' signatures in some methodological sound procedure. We will first review the different alternatives and approaches suggested for this task and will discuss their strengths and weaknesses. We will argue that results of a single test alone might not offer a sufficient fundament for clear and reproducible results. Therefore we will present suggestions made in the literature of combining the outcome of several tests and discuss their strengths and limitations. We will then suggest a simple alternative method of combining several test outcomes in one or several complementary combined tests. This approach will be illustrated with an extensive simulation study considering a large variety of selection scenarios to which eight different selection signature test statistics with very different performance profiles will be applied. We will show that the combined test leads to a better positional resolution of the selected region and in most cases has similar power as the locally most powerful single test (which may differ across scenarios). As a by-product, it will be demonstrated that selection signature analysis profits most from high-density genotypes (ideally obtained from whole genome re-sequencing), while good results can already be obtained with a rather moderate sample size. This finding suggests that the lack of consistency and reproducibility of the results of the first genome-wide scans for selection signatures in livestock populations may be caused by the insufficient marker density of the medium-density SNP arrays used to generate the data.

Statistical testing. A statistical test in the classical sense is based on the probability p, that an observed value of the chosen test statistic calculated from the data at hand can emerge under the assumption of the null hypothesis, in the present case assuming that the population studied was

not under selection (in general or at the locus considered). This requires that the distribution of the test statistic under the null hypothesis is available. There are different approaches to determine this distribution:

(i) The test statistic has a theoretically known distribution.

(ii) *The distribution can be determined empirically, e.g. by permutation of the data.* This is a widely used strategy in situations where the association of two types of observations (e.g. genotypes and phenotypes) is tested, as e.g. in QTL mapping. However, this principle is difficult to adopt for selection signature analysis where we usually just have genotype observations. Permutation tests have been suggested for haplotype-based analyses (e.g. Qanbari et al. (2012)) where permutation breaks up the sequential patterns, or in between population analyses (Gianola et al. (2010)), where random assignment of individuals to populations can mimic the case of no differential selection.

(iii) *The distribution can be obtained by simulation of the data under the null hypothesis.* For this, it is necessary to simulate data using a realistic demographic model by forward or backward (coalescent based) simulation. While such simulations have been applied in human genetics studies (e.g. Grossman et al. (2010)) phylogenetic patterns in most farm animals are rather complex and little understood so that calibrated demographies are not available. For this and other reasons, Woolliams and Corbin (2012) conclude that backward simulation via coalescence theory, although computationally attractive, may be of limited use in a farm animal context.

(iv) The distribution is derived under the assumption that the vast majority of loci is not under selection. It is a sensible assumption that the majority of loci are not under selection and just a few are. In this case the majority of SNPs can be used as an unselected 'control' (or unselected population forming the basis to derive the empirical distribution of the test statistic under the H_o) in contrast to the few selected ones forming signals (Gianola et al. (2010)).

Very close to this last suggestion is the widely used approach to identify the most extreme observed values of the test statistics as selection signals. This is sometimes called an 'outlier' statistic, which is somewhat misleading, since in most cases just extreme values are reported, regardless of the fact whether these values are from the same distribution as the bulk of observations or from a distinct one – only in the latter case would the term 'outlier' be appropriate.

In many studies the top 1 or 0.1 per cent of the test statistics are taken as 'significant' and often, their quantile value is reported as 'p-value'. From a statistical point of view this is misleading, since the reported values do not reflect a probability under the null hypothesis. An obvious argument demonstrating the shakiness this approach is that there always will be a set of loci with values of the test statistic in the top 1 per cent, so even if there is no selection at all, the method will report 'significant' results.

It should be noted, that not all selection signature studies refer to classical frequentistic genetics with a respective definition of error probabilities referring to a clearly defined null hypothesis, but some approaches have been suggested in a Bayesian framework applying Bayes factors to assess significance (e.g. Nielsen et al. (2009); Grossman et al. (2010)).

The concept described above implicitly assumes that the all demographic processes (such as drift, bottlenecks, fluctuating population sizes, stratification) affecting the population as a whole will create a typical set of genomic patterns across the genome. By looking for patterns strongly deviating from these overall patterns, i.e. using the majority of SNPs as sort of 'control', locus specific effects (such as selection) will be detected.

However, it was argued (Biswas and Akey (2006)) that locus-specific deviations can also be generated by non-locus specific forces: imagine for instance a population having recently gone through a very severe bottleneck. By chance (random genetic drift), only gametes with a specific haplotype in a certain region may survive this bottleneck. In the subsequent generations the population might expand again, but since there is only a single haplotype available, only this haplotype will be present until it is eventually broken down by mutations or admixture with introgressed genetics. A haplotype-based approach for selection signature detection, such as the widely used |iHS| statistic (Voight (2006)) will detect a high frequency long-range haplotype in such a data set and will interpret it as an indication of locus-specific selection in the respective region, although the signature has been created through random drift alone without any directional selection at all. It is well known, that severe bottlenecks have appeared in the recent demographic history of many livestock populations, thus it must be kept in mind that the mechanisms described (and similar ones) may lead to false positive signals, and then interpretation of results should be carried out with great caution.

Simulation study. The program msms (Ewing and Hermisson (2010)) was used to simulate population genetics datasets under a neutral model and a single locus selection model, respectively. Each simulation scenario represents a 10 Mb genomic fragment with a constant recombination rate (1cM/Mb). In order to apply both within and between population selection signature statistics one half of the population considered remained unselected while the other half of the population was selected in each selection scenario. In the selection case a selected allele was positioned in the center of the fragment considered and in the reference scenario the selection coefficient was $s = 0.02$. Data for analysis were sampled when the frequency of the selected allele reached a predefined value $p = 0.8$. Selection signature statistics then were computed for sample sizes $N = 50$ gametes in each selected or unselected subpopulation, and the average marker distance was $d = 2.5$ kb. Starting from this reference scenario each parameter was varied over a range of values listed in Table 1 while all other parameters were kept at the reference setting.

Table 1: Parameter settings varied in the simulated selection scenarios. Reference values are underlined.

Parameter	Range of values
Selection coefficient s	0.005, 0.01, <u>0.02</u>, 0.04, 0.08
Allele frequency p	0.2, 0.4, 0.6, <u>0.8</u>, 1.0
Sample size N	10, 30, <u>50</u>, 70, 90
Marker distance d	0.1, 0.5, <u>2.5</u>, 12.5, 62.5 kb

Eight different selection signature statistics were computed for each SNP: three statistics testing for divergent selection: F_{ST} (Gianola et al. (2010)), XPEHH (Sabeti et al. (2007)), XPCLR (Chen et al. (2010)), four statistics looking for a divergent allele frequency spectrum: Tajima's D (Tajima (1989)), Fu&Li D, Fu&Li F (Fu and Li (1993)) and CLR (Nielsen et al. (2005)), and |iHS| (Voight et al. (2006)) looking for high frequency conserved haplotypes. All statistics were orientated such that a high value indicates presence of selection.

To obtain the empirical distribution of the test statistics, one hundred simulations were run in which no selection was assumed in both populations, and the maximum observed value of each test statistic in each run was stored. The value cutting of the upper 1 per cent quantile of each statistic was used as an empirical significance threshold value.

To assess the power of each statistic, one hundred replicates were simulated under the corresponding selection scenario. A selection signature was assumed to be detected if at least one SNP within a 500 kb window around the selected locus exceeded the empirical significance threshold. This window size was determined by the extent of linkage disequilibrium in the simulated population. The percentage of detected signatures among all replicates is reported as the power.

Results and Discussion

Simulation scenarios. In the reference scenario, there is a clear separation between the methods considered regarding the power. Three methods (XPEHH, |iHS|, and CLR) have a power > 80%, while for all other methods the power is < 20%. The impact of the four parameters varied is depicted in **Figure 1**.

Regarding the frequency of the selected allele (**Figure 1A**), |iHS| appears to be most powerful for ongoing selection processes where the target allele has a medium to high frequency ($0.4 < p < 0.8$). However, at fixation ($p = 1$) |iHS| has limited power (~40%), while XPEHH and CLR have 100% power. All other statistics are hardly affected by the allele frequency level.

With a marker interval d = 62.5 kb, which is approximately the resolution obtained when genotyping mammals with 50k SNP arrays, all methods have a power < 30% (Figure 1B). This may explain the low reproducibility of the results of some of the first selection signature studies with farm animal data (Qanbari and Simianer (2014)).

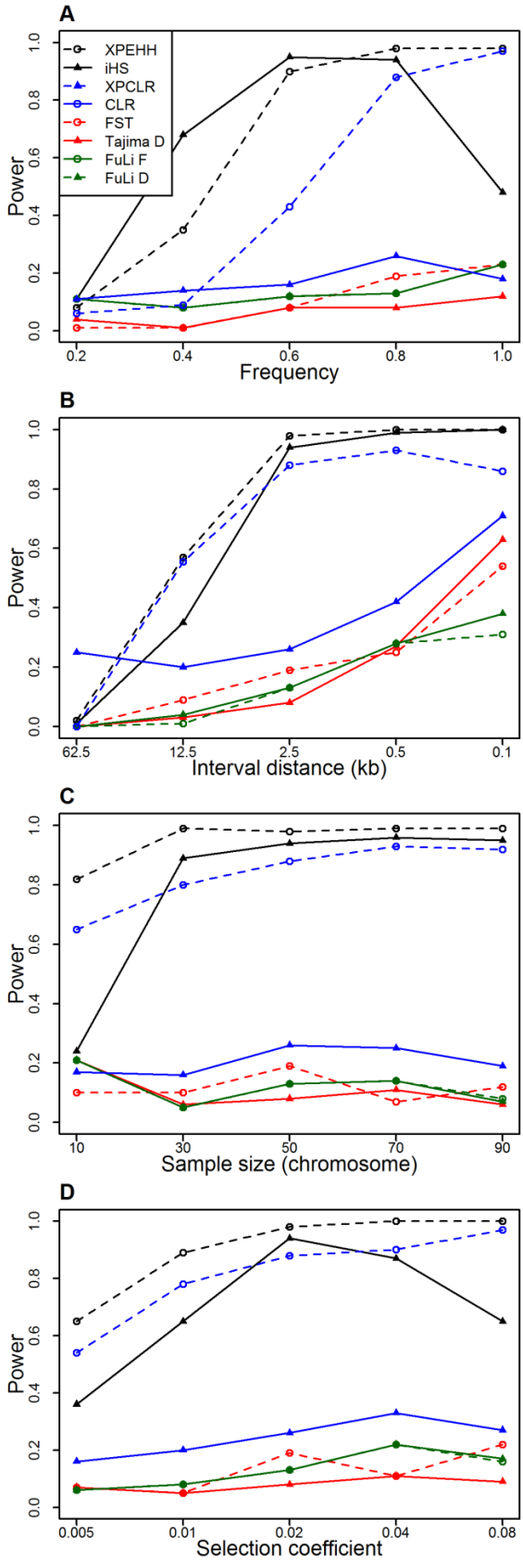

Figure 1: Power of eight different selection signature test statistics when varying four different parameters: (A) Frequency of the selected allele; (B) Marker interval distance; (C) Sample Size; (D) Selection coefficient.

Higher resolutions (note that in **Figure 1B** the scale of the x-axis is exponential) quickly lead to better results for the three 'high power methods' (XPEHH, |iHS|, and CLR), but in general all methods show an improved performance with resolution d = 0.1 kb, which is approximately what is obtained in whole genome sequencing. In this situation, the power of XPCLR comes close to the three top methods.

Regarding the impact of sample size (**Figure 1C**) it appears that a rather limited value (N = 30 gametes, equivalent to 15 diploid individuals) is sufficient to reach reasonable results with the three 'high power methods'. In contrast, the performance of the other methods does not profit from increased sample size (at least within the rather limited range taken into consideration). It should be noted that today in some farm animal applications much larger samples (thousands of gametes) are available, while the present study cannot provide any insight into the power of the methods considered in such a setting.

Finally, it is shown that the power of XPEHH and CLR monotonically increases with increasing selection coefficient (**Figure 1D**), while |iHS| has highest power with an intermediate (s = 0.02) selection coefficient, but the power erodes both with stronger and weaker selection. Most of the other methods show a slight increase of power with increasing strength of selection, but overall the power of those methods stays at a low level. In general it is difficult to judge which of the simulated selection coefficients reflects selection intensities of practical relevance in livestock populations, because a wide range of selection intensities is applied. While selection for some of the main production traits is very intense, leading to up to 1 per cent improvement through genetic progress per year (Hill (2010)), selection for some functional traits, such as fertility or disease resistance, is weak, but has operated over long periods, even as 'natural selection' prior to the actual domestication event.

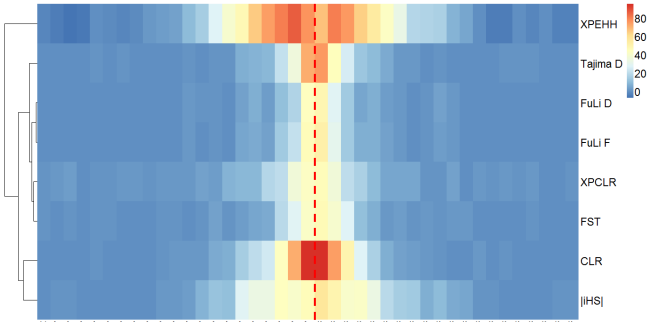

Figure 2: Heat map of the empirical power (in per cent) of eight different selection signature test statistics in 50 kb intervals. The simulated scenario was s = 0.02, N = 50, d = 0.1 kb and p = 1.0 (for |iHS| p = 0.8). The red dashed line indicates the position of the SNP under selection. The clustering of the test statistics is indicated on the left margin.

A further aspect that deserves consideration is the positional resolution of the selection signature statistic. **Figure 2** shows the power of the eight statistics in a scenario with maximum marker density (d = 0.1kb) reported for intervals of 50 kb. It becomes evident that for most statistics the highest power is focused around the selected position, while especially for XPEHH and |iHS| the region of highest power is quite broad, indicating that positional resolution is limited. The |iHS| statistic was considered for a final frequency of the selected allele of p = 0.8, because under fixation (p = 1) this statistic has a massive loss of power (cf. **Figure 1A**).

Combining test statistics. The eight test statistics considered in our simulation study are just a subset of all selection signature statistics suggested in the literature. Many of these statistics reflect the same phenomenon, as e.g. REHH (Sabeti et al. (2002)) and |iHS| are derived from the basic EHH (Sabeti et al. (2002)) statistic, trying to correct EHH for some local genetic pattern like a deviated recombination activity. Hence, those statistics are similar and highly correlated among each other. In other cases, statistics reflect very different signals, such as a deviated allele frequency spectrum within a population (CLR), high frequency long range haplotypes in a population (|iHS|) and extreme local divergence between populations (F_{ST}).

Grossman et al. (2010) have suggested combining various signals into a composite of signals (termed CMS) mainly to improve the resolution of the detected selection signatures. A keystone of this approach is the ability to simulate data according to calibrated demographic models using the coalescent approach. For most livestock species the actual demography is largely unknown and, if it was known, would probably be hardly suited for simulation using a coalescent approach. Beyond that, the general applicability of coalescent theory in livestock genomics was questioned by Woolliams and Corbin (2012).

Utsunomiya et al. (2013) suggested merging different genome wide scan statistics by combining p-values using a method suggested by Whitlock (2005). This method was successfully applied to detect selection signatures in beef cattle. However, it should be noted that some of the underlying assumptions of the method are hardly met, since the p-values of some of the single tests are not p-values in the classical statistical sense, but reflect quantile values from the empirical distribution of test statistic values under selection. Beyond this Utsunomiya et al. (2013) combined largely uncorrelated scan statistics.

In our case, statistics are partly highly correlated. We calculated the correlation matrix for all 8 test statistics from the simulated data under the null hypothesis. We found that e.g. the pairwise correlations between Tajima's D, Fu&Li D and Fu&Li F are all above 0.6 and the correlation between XPCLR and FST is 0.26, while for instance the absolute correlation between |iHS| and all other statistics does not exceed 0.06.

We thus suggest an approach that has the potential to combine several statistics that reflect the same selection signal, but is also able to reveal several combined statistics for different types of selection mechanisms. For this, we did a principal component analysis (PCA) of the correlation matrix of the eight test statistics. **Table 2** reports the loadings and the proportion of variance explained by the first five principal components (PC1 to PC5). Not surprisingly, PC1 combines the signal of the three highly correlated

statistics Tajima's D, Fu&Li D and Fu&Li F. This component explains 33 per cent of the overall variance in the correlation matrix. It should be noted that PCs are constructed in an orthogonal way, so that they pick up complementary information and the resulting linear combinations are uncorrelated. Also, the ranking by proportion of variance explained does not tell anything about the usefulness or power of the respective PC, since a combination of highly correlated low power statistics will probably not lead to a high power PC. PC5, which later will be reported to be the most useful one, still explains 11 per cent of the overall variance and gives most weight to the high-power methods XPEHH, |iHS| and CLR.

Table 2: Loadings and proportion of variance explained by the first five principal components (PC1 to PC5) derived from the correlation matrix of the eight selection signature statistics under the null hypothesis. Loadings with absolute values ≥ 0.33 are underlined.

	PC1	PC2	PC3	PC4	PC5
XPEHH	-0.03	0.38	0.40	0.53	0.63
XPCLR	0.02	0.67	-0.07	-0.05	-0.14
|iHS|	0.00	-0.01	-0.89	0.28	0.33
CLR	0.20	0.03	0.01	-0.74	0.63
Tajima D	0.52	-0.01	-0.08	0.03	0.05
FuLi D	0.57	-0.02	0.06	0.13	-0.12
FuLi F	0.60	-0.02	0.04	0.12	-0.09
FST	0.02	0.64	-0.17	-0.22	-0.25
Variance Explained	33%	16%	13%	12%	11%

We then formed new test statistics PC1 to PC5 by multiplying the loading of the respective test statistic by its value obtained in the simulation study and summing up over all eight tests. Scaling was appropriately taken into account (since the PCA was done based on the correlation matrix and not on the covariance matrix) and resulting combined test statistics again were orientated in such a way, that high values indicate selection.

It was observed that across all scenarios PC5, which combines the high-power methods XPEHH and |iHS| together with the CLR statistic, has similar power as the locally most powerful single test (which may differ across scenarios). **Figure 2** depicts this for variable sample size.

Applying PC5 to screen for selection signatures in the simulated genome region shows that a composite of multiple test statistics provides a better positional resolution at the selected locus and reduces the stochastic noise in non-selected regions. This is demonstrated in **Figure 3**, where results for one replicate of the simulated selection scheme under reference assumptions are depicted.

It should be noted that most of the eight single selection signature statistics produce a high value of the test statistic at the position of the selected SNP. The only single statistic producing a strong and unique signal at the correct position in this case is XPEHH. For the majority of statistics considered, though, these signals at the selected position are not very focused and overlap with false positive peaks in non-selected positions.

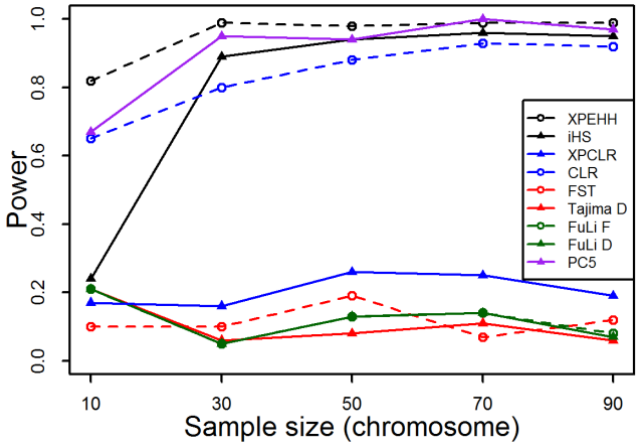

Figure 2: Power of eight different selection signature test statistics and the combined statistic PC5 with variable sample size

The combined statistic PC5 is shown to produce a very strong and focused signal in the selected region with hardly any false positive signals at other positions. The range of the signal is very focused around the selected SNP and clearly outperforms the signals at non-selected positions. This high resolution will be helpful when annotating the genes associated with the selection signature.

Conclusions

The presented results illustrate that different test statistics behave differently in different scenarios. Most remarkable is the clear evidence for the usefulness of high density markers in selection signature analysis, suggesting that whenever possible such studies should be based on sequence data, while results obtained with low to medium density SNP arrays appear to be of limited reliability. We suggest a simple and straightforward way of combining different correlated or independent test statistics which is shown to be efficient in mapping selection signatures with high power and positional resolution. Selection signature analysis is a relatively novel and highly promising approach in livestock population genomics, the first chromosome-wide screen for selection signatures being reported by Hayes et al. (2008). Important statistical challenges such as the problem of hard vs. soft sweeps (Hermisson and Pennings (2005)) and the need of studying selection on whole pathways rather than on single SNPs (Amato et al. (2009)) could not be addressed here but certainly are highly relevant in the context of livestock population genomics. An accurate and comprehensive set of selection signatures will be the basis for a better understanding of the forces driving artificial selection and will help to design more efficient livestock breeding programs.

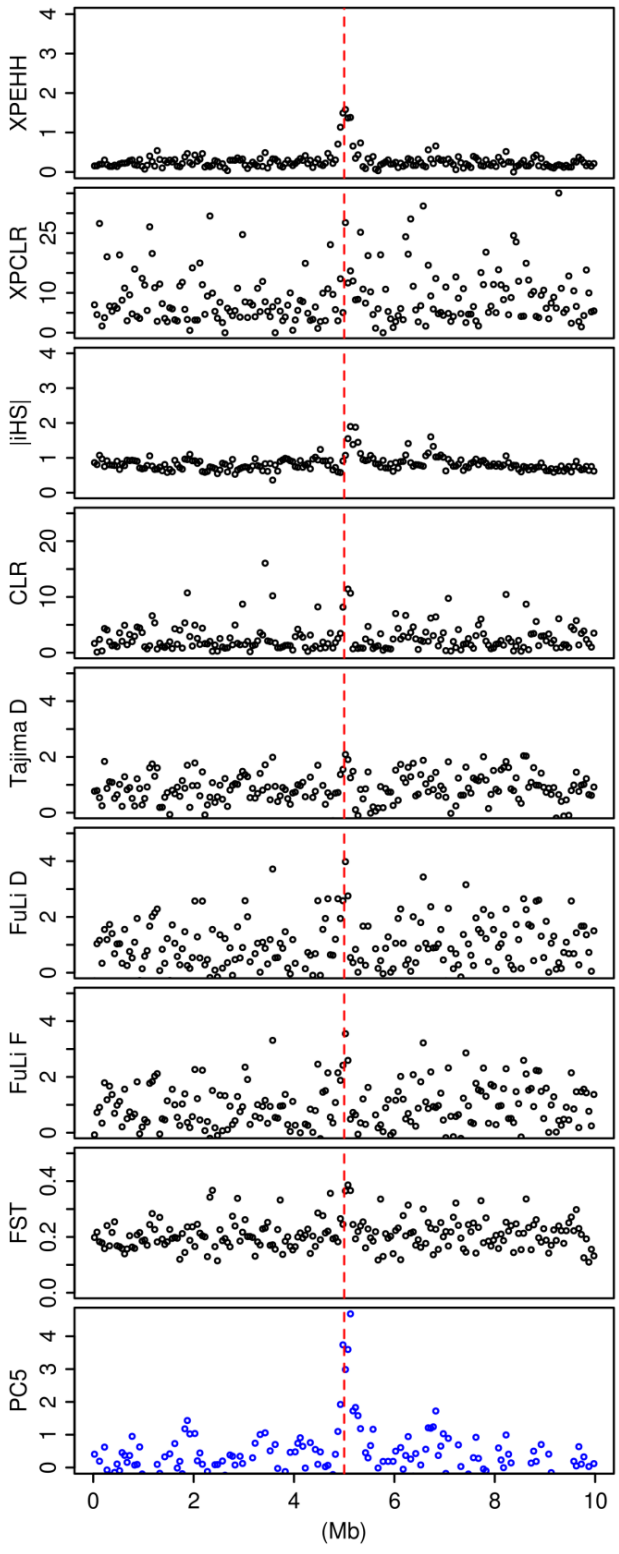

Figure 3: Observed values of the eight test statistics and PC5 in one replicate of the simulated reference scenario under selection. The red dashed line indicates the position of the SNP under selection.

Acknowledgements

Parts of this research were conducted within the AgroClustEr ‘‘Synbreed - Synergistic plant and animal breeding’’ (FKZ 0315528C) funded by the German Federal Ministry of Education and Research (BMBF) in association with the research training group ‘‘Scaling problems in statistics’’ (RTG 1644) funded by the Deutsche Forschungsgemeinschaft (DFG). Y. Ma acknowledges financial support by the China Scholarship Council (CSC).

Literature Cited

Amato, R., Pinelli, M., Monticelli, A. et al. (2009). PLoS ONE 4: e7927.

Biswas, S., Akey, J.M. (2006). Trends in Genetics 22: 437 – 446.

Chen, H., Patterson, N., and Reich D. (2010). Genome Res. 20: 393–402.

Ewing, G., Hermisson, J. (2010). Bioinformatics 26: 2064-2065.

Fu, Y. X., Li, W. H. (1993). Genetics, 133: 693-709.

Gianola, D., Simianer, H., Qanbari, S. (2010). Genet Res (Camb) 92:141-55.

Grossman, S.R., Shylakhter, I., Karlsson, E. L. et al. (2010). Science 327: 883-886.

Gulcher, J, Stefansson, K. (1998). Clin Chem Lab Med. 36:523-7.

Hayes, B. J., Lien, S., Nilsen, H. et al. (2008). Animal Genetics 39: 105-111.

Hermisson, J., Pennings, P. S. (2005). Genetics 169: 2335-2352

Hill, W. G. (2010). Phil. Trans. R. Soc. B 365: 73–85.

Nielsen, E. E., Hemmer-Hansen, J., Poulsen, N. A. et al. (2009). BMC Evol. Biol. 9: 276-287.

Nielsen, R.., Williamson, S., Kim, Y. et al. (2005). Genome Res. 15: 1566-1575.

Qanbari, S., Pimentel, E. C. G., Tetens, J. et al. (2010). Animal Genetics 41: 346-356.

Qanbari, S., Simianer, H. (2014). Livest. Sci., in press.

Qanbari, S., Strom, T. M., Haberer, G. et al. (2012). PLoS ONE 7: e49525.

Sabeti, P. C., Reich, D. E., Higgins, J.M. et al. (2002). Nature 419: 832–837.

Sabeti, P. C., Varilly, P., Fry, B. et al. (2007). Nature 449: 913-918.

Tajima, F. (1989). Genetics 123: 585-595.

Utsunomiya, Y. T., Pérez O’Brien, A. M., Sonstegard, T. S., et al. (2013). PLoS ONE 8: e64280.

Voight, B. F., Kudaravalli, S., Wen, X. et al. (2006). PLoS Biol: dol: 10.1371/journal.pbio.0040072

Whitlock, M. C. (2005). J. Evol. Biol. 18: 1368–1373.

Woolliams, J., Corbin, L. (2012). J. Anim. Breed. Genet. 129: 255–256.