Detection of selection signatures within and between cattle populations

H. Simianer[*], S. Qanbari and D. Gianola[†]

Introduction

Genomics techniques provide opportunities to better understand the genetic background of important trait complexes in farm animals. The major approach towards this goal is the localisation of quantitative trait loci (QTL) via linkage or association mapping. Both approaches are based on linking the phenotypic variation to genomic patterns as reflected by marker genotypes. Despite the fact that massive funding has gone into QTL mapping projects in farm animals, the number of fully characterised functional mutations so far is disappointingly low (Ron and Weller 2007).

With the availability of high throughput genotyping platforms for single nucleotide polymorphisms (SNPs) in farm animals novel approaches towards understanding genetic mechanisms in selected populations get within reach. A common feature of farm animal populations is that they are subject to a selection pressure which clearly exceeds the level of selective forces in natural selection. It is of interest to identify regions in the genome that show a measurable reaction to selection pressure. Interestingly, such regions can be detected by analysing patterns in the genomic data alone and do not require phenotypic information. Once such signals are identified and mapped, it is possible to annotate the functional genomic background of those regions, leading to a better understanding of how selection works in different trait complexes. At the same time, understanding the actual genomic mechanisms of selection in the history of a population provides a basis for developing more efficient genomic selection strategies to be applied in the future.

In this paper we will review some major concepts for detecting signatures of directional selection based on high throughput genotypes both within and between populations. We will illustrate these methods with examples from recent large scale studies in cattle populations. Finally we will discuss the strengths and weaknesses of the suggested approaches and will address the main methodological challenges.

Methods

Within breed analyses. For within breed analyses a number of single-locus based methods analysing the site frequency spectrum are well established, such as Tajima's *D* (Tajima 1989) or Fay and Wu's *H* (Fay and Wu 2000). However, we will concentrate here on methods linking haplotype frequency and length. A novel mutation under positive selection pressure will increase rapidly in frequency, so that the surrounding conserved haplotype is long, which is called a 'selective sweep' (Maynard Smith and Haigh 1974). As selection carries an allele on a specific haplotype to higher frequency faster than the rate at which it is broken down by recombination, high frequency haplotypes will be longer than expected under neu-

[*] Georg-August University Göttingen, Animal Breeding and Genetics Group, 37075 Göttingen, Germany
[†] University of Wisconsin-Madison, Department of Animal Sciences, Madison WI 53706, USA

trality. This is the basic concept of the *Extended Haplotype Homozygosity* (EHH) statistic as introduced by Sabeti *et al.* (2002). To evaluate the decay of haplotype homozygosity with increasing distance from a core haplotype of interest we performed the EHH test as implemented in the software *Sweep v.1.1* (Sabeti *et al.* 2002). The test is based on the contrast of a core haplotype with both high frequency and extended homozygosity with other core haplotypes at the same locus. EHH is the probability that two randomly chosen haplotypes carrying the candidate core haplotype are homozygous for the entire interval spanning the core region to a given adjacent locus. The EHH of a tested core haplotype *t* is

$$EHH_t = \sum_{i=1}^{s} \binom{e_{ti}}{2} \Big/ \binom{c_t}{2}$$

where c_t is the number of samples of a particular core haplotype *t*, e_{ti} is the number of samples of a particular extended haplotype *i*, and *s* is the number of unique extended haplotypes.

Recombination rates vary along the genome, as was e.g. demonstrated by Simianer *et al.* (1997) for the bovine genome. This raises the possibility that a larger EHH statistic may rather be due to low recombination rates in a particular region and not necessarily to recent positive selection. To account for this, Sabeti *et al.* (2002) suggested the *Relative Extended Haplotype Homozygosity* (REHH) statistic which corrects the observed EHH value at one core haplotype for the average level of EHH values for all relevant core haplotypes in the same region.

REHH statistics make only use of genomic patterns as displayed by marker configurations. Neither phenotypes nor external knowledge on e.g. novel mutations (relative to wild ancestors) are used. SNPs or core haplotypes under positive selection pressure are only indirectly identified by having a high frequency and a long surrounding haplotype stretch at the same time. As an alternative, Voight *et al.* (2006) developed the *integrated Haplotype Score* (|iHS|) based on the comparison of EHH between derived and ancestral alleles within a population. Ancestral alleles are those variants which are present in the wild ancestor or an outgroup. Compared to REHH the |iHS| statistic uses additional external information and thus should be more robust and powerful.

Between breed analyses. Rather different philosophies are used to identify regions under selection based on multi-population data. Here, the assumption is that selection causes an increased differentiation on the genomic level if breeds are selected divergently. A widely used single locus statistic for differentiation is F_{ST} measuring relatedness between pairs of alleles within a sub-population relative to that in an entire population (Wright 1951; Cockerham 1969; Weir and Hill 2002). Equivalently, F_{ST} can be interpreted as a measure of dispersion of gene frequencies among groups relative to the variation expected in the population from which such groups derived. Gianola *et al.* (2010) suggested a simple Bayesian model for drawing samples from posterior distributions of locus-specific F_{ST} values. Following Jeffreys' rule (Bernardo and Smith 1994) a *Beta(0.5, 0.5)* distribution is assigned as a reference prior to all loci in all populations. This leads to the joint posterior density of all allele frequencies

$$g(\mathbf{p}|Data) = \prod_{i=1}^{N} \prod_{l=1}^{L} Beta(n_{il1} + 0.5, n_{il2} + 0.5)$$

where $\mathbf{p}$ is the vector of allele frequencies, n_{ilk} is the number of observed alleles $k = 1,2$ at the biallelic locus $1 \le l \le L$ in group $1 \le i \le N$ and $Data$ is the set of observed allele realisations in the sample. Samples can be drawn from this posterior distribution from which Monte Carlo estimates of features of the posterior distribution, like posterior means, variances, or credibility intervals can be calculated.

Based on this it is also possible to draw samples from the posterior distribution of F_{ST} values. Let $p_{il}^{(s)}, s = 1,2,\cdots,S$ be samples from the posterior distribution of allele 1 at locus l in group i. Then, a draw from the posterior distribution of F_{ST_l} is given by

$$F_{ST_l}^{(s)} = \frac{N\sum_{i=1}^{N}\left(p_{il}^{(s)}\right)^2 - \left(\sum_{i=1}^{N} p_{il}^{(s)}\right)^2}{N\sum_{i=1}^{N} p_{il}^{(s)} - \left(\sum_{i=1}^{N} p_{il}^{(s)}\right)^2}$$

which is a random variable with support in (0,1). From this it is rather straightforward to assess features of the posterior distribution of F_{ST} values.

Breed differentiation by reproductive separation and genetic drift is a process which is affecting all loci in the genome in a similar fashion. Therefore one would expect, in the absence of directional selection, that all F_{ST} values across the genome originate from the same stochastic process and hence share a single distribution. If, however, breed differentiation is not based on drift alone, but directional selection is active and operates differently on different genomic regions, one would expect a larger degree of differentiation in some regions compared to the average or background level of differentiation caused by drift. Hence, F_{ST} values in different regions are expected to stem from different distributions, with higher F_{ST} values reflecting divergent selection and lower F_{ST} values reflecting balancing selection, respectively. This can be assessed by clustering a set of F_{ST} values (in this case, posterior means) from a multi locus analysis into data driven groups.

Applications to cattle data

Within breed analyses. Qanbari *et al.* (2010a) performed a whole genome scan for selection signatures in the Holstein genome based on 810 individuals genotyped with the Illumina Bovine SNP50 BeadChip. After filtering, 40'854 autosomal SNPs covering 2544.1 Mbp of the genome could be used. A total of 3741 core regions spanning 472.1 Mbp (18.55%) of the genome were identified. A total of 28'323 EHH tests with an average of 7.57 tests per core region were calculated, reflecting multiple core haplotypes per region and tests in both (3' and 5') directions. Figure 1 displays box plots of the distribution of -$\log_{10}$ (p-values) within bins of core haplotype frequency. It is evident that the extreme outliers primarily appear with moderate haplotype frequencies.

Extreme REHH values were found on all chromosomes with substantial clusters on chromosomes 2, 10, and 20. Analysis of a set of 10 candidate regions, which a priori were assumed to be under selection, revealed that 5 of them (DGAT1, Casein Cluster, Growth Hormone Receptor, Somatostatin and Leptin Receptor) showed extreme REHH values. Results are also in agreement with the findings of Hayes *et al.* (2009) who suggested signatures of selection

in the vicinity of GHR and DGAT1 genes as revealed by allele frequency differences. The long range linkage disequilibrium (LD) consistency observed in this study is also in coincidence with the reports of Grisart *et al*. (2001) and Marques *et al*. (2008) who used EHH plots to evaluate extended long range LD around DGAT1, especially the second most frequent core haplotype in the DGAT1 region which was shown to be in complete LD with the causative DGAT1 mutation (Qanbari, 2010a). The annotation of the most extreme selection signals revealed that the affected regions are mostly associated with physiological pathways related to growth and lactation, but also genes affecting female and male fertility were strongly represented. This comprised genes affecting spermatogenesis which were also found to be under positive selection in the human genome (Wyckoff *et al.* 2000).

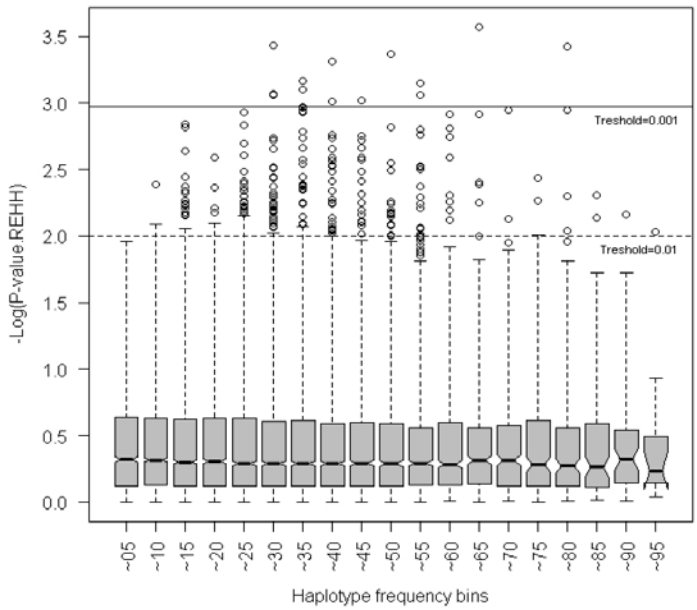

Figure 1: Box plot of the distribution of P-values in core haplotype frequency bins of 5% (Qanbari *et al.*, 2010a).

In a follow-up study Qanbari *et al.* (2010b) applied the |iHS| test to a sample of 2091 German Holstein Friesian animals, again genotyped with the Illumina Bovine SNP50 BeadChip. Ancestral alleles were taken from Matukumalli *et al.* (2009). The 39'474 |iHS| values were combined into 5055 non-overlapping sliding windows of 500 kb length, results are depicted in Figure 2. The interval with an outlier value (Mb 57.25–57.75 on chromosome 18) contains the Sialic acid binding Ig-like lectin 5 gene and the Zinc finger protein 577 gene which recently were reported as candidates to have large effects on calving ease, several conformation traits, longevity, and total merit in Holstein cattle (Cole *et al.* 2009). The overall list of candidate regions identified through the |iHS| test comprised genes involved in the biological processes such as anatomical structure development, muscle development, carbohydrate and lipid metabolism, spermatogenesis and fertilisation. These results generally are consistent with the observations of Flori *et al.* (2009). Evidence for positive selection in the genomic

region surrounding muscle related genes has been also reported in racing horses (Gu *et al.*, 2009).

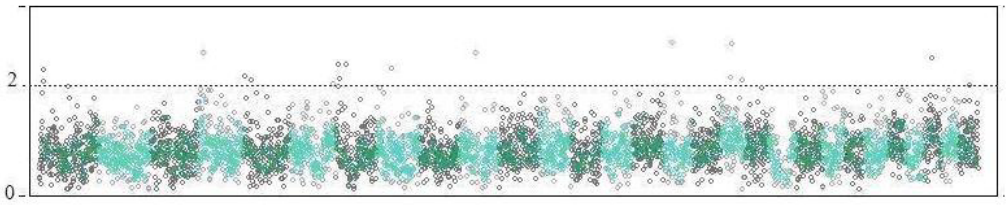

Figure 2: Genome wide distribution of |iHS| values in the Holstein genome ordered by chromosome. Each dot represents the average |iHS| values for a 500 kb window (Qanbari *et al.*, 2010b).

Between breed analyses. In the same study, Qanbari *et al.* (2010b) also performed an F_{ST}-based analysis following the Bayesian model as suggested by Gianola *et al.* (2010). The study was based on 2091 German Holstein Friesian, 277 German Brown Swiss, 102 Canadian Angus and 43 Canadian Piedmontese animals. All animals were genotyped with the Illumina Bovine SNP50K BeadChip, and in this case all autosomal SNPs were used which had a MAF > 0.001% in at least one of the populations. In total, 2514 animals and 40'595 SNPs were used. 2000 samples of the joint posterior density of allele frequencies were drawn, resulting in 2000 draws from the posterior distribution of F_{ST} values per locus. The posterior mean ± standard errors of pairwise F_{ST} values between breeds are given in Table 1. Clearly the amount of differentiation within dairy (Holstein, Brown Swiss) and within beef (Angus, Piedemontese) breeds, respectively, is substantially smaller and less variable than the differentiation between dairy and beef breeds, respectively.

Table 1: Posterior means ± standard errors of the pairwise F_{ST} values between breeds (Qanbari *et al.*, 2010b)

	Brown Swiss	Angus	Piedemontese
Holstein	0.057 ± 0.076	0.274 ± 0.315	0.270 ± 0.313
Brown Swiss		0.290 ± 0.334	0.281 ± 0.335
Angus			0.022 ± 0.041

Figure 3 shows the posterior density of the distribution of posterior means of F_{ST} values over all loci between dairy and beef breeds. The histogram illustrates well that a majority of the density is centered around ~0.05, while a substantial proportion of the posterior density covers the entire parameter space. A mixture model of posterior means analysis (over loci) with the Flexmix routine in R (Leisch 2004) reveals a mixture of two normals. Strong evidence for selection in the region of the GHR gene on BTA20 is consistent with the findings of Flori *et al.* (2009) and Hayes *et al.* (2009), the latter reporting it for an Angus vs. Holstein breed

comparison. Two regions on BTA2 and BTA5 in the vicinity of ZRANB3, R3HDM1 and WIF1 genes known to affect feed efficiency and mammalian mesoderm segmentation, respectively (The Bovine HapMap consortium 2009), also matched to the outlier F_{ST} windows in our study. Extreme peaks were frequent in presumed gene deserts, which may reflect selection acting on uncharacterised regulatory regions or simply fixation of non-coding DNA by genetic drift. This observation is consistent with the reports of Flori *et al.* (2009) and Gu *et al.* (2009) which detected signals of divergent selection in genome regions with poor gene content in genome wide analyses of cattle and thoroughbred horse, respectively, using the F_{ST} statistic. Thus, these results in combination with the observations from Voight *et al.* (2006), Carlson *et al.* (2005) and Wang *et al.* (2006) on human population data suggest that non-coding regions may have been an important substrate for adaptive evolution.

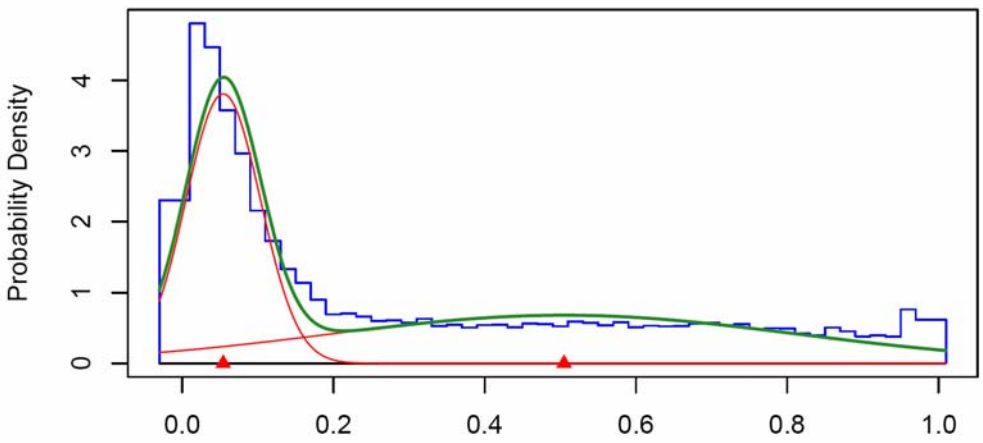

Figure 3: Histogram of the distribution of posterior mean over loci of F_{ST} values between dairy and beef breeds and densities of the underlying mixture of two normals (green) and the respective components (red)

Discussion

All methods considered use different indications of selection on the genomic level. Therefore, a moderate inconsistency of results across different approaches is expected. Beyond this methodological reason, a lack of power due to limited sample size and marker density may also contribute to discordant results. If, say, power is only 50 per cent for finding a selection signature in one location with either method A or method B, the probability of finding the signature with both method A and B is only 25 per cent. In the comparison across breeds, selection for a certain trait may result in different genomic set-ups, again leading to inconsistent results. If, say, selection acts on a phenotype which is a result of a complex pathway, the process may be optimised by modifications of different steps in the pathway in different selected populations. Again, discordant signals will be detected in such a case. While the reported studies in the cattle genome only analysed autosomal markers, X-linked genes were found to be overrepresented in studies of selection signatures in the genome of primates (Nielsen *et al.* 2005). This is attributed to the large number of sperm- and testis-associated

genes on the chromosome, as well as to the hemizygosity of the non-pseudoautosomal part of the X-chromosome in males, exposing recessive alleles to an increased selective pressure (Schaffner 2004).

A general drawback of the methods described is that no rigorous statistical testing procedure is available and implemented. The methods rather depend on outlier detection, assuming that the most extreme values of the statistic discovered in a genome-wide scan, relative to the average value of the test statistic over all locations, likely represent regions that were under highest selection pressure. In general, selection signature methods are mainly used either to identify candidate regions for further molecular-genetic characterisation (for a review in the human genetics context see Sabeti *et al.* 2006), or to assess putative selection effects in candidate genes derived from biological knowledge and/or from mapping studies in the same material (see e.g. Nilsen *et al.* 2009). For a rigorous statistical test of a selection signature at a specific locus, the distribution of the test statistic under the null hypothesis (no signature at this locus) is required. This distribution can either be derived theoretically or generated empirically, e.g. by a permutation of the data (Churchill and Doerge 1994) or based on backward simulation procedures like the coalescent (Schaffner *et al.* 2005). The suggested clustering of F_{ST}-values (Gianola *et al.* 2010) in the genome-wide scan may be a step towards a detection of genomic regions with different characteristics caused by locus-specific forces, one of which may be an increased sensitivity towards selection. A proper testing strategy also should account for the multiple testing problem, which is notorious in the context of whole genome scans, and the problem of a potential ascertainment bias due to SNP selection.

Other methodological issues which are of special concern in farm animals are (i) the small effective population size of farm animal populations compared to human populations, (ii) the possible admixture of populations and (iii) the available marker density, which is one or two orders of magnitudes away from what is available in human genetics.

Conclusions

Whole genome scans for selection signatures in farm animals are a novel approach which became feasible only recently with the availability of high throughput SNP genotyping. Despite a number of open methodological questions which are addressed in this contribution, the concept is promising and early results are highly encouraging. With larger data sets and higher marker densities becoming available in the near future, these approaches will provide a much better insight on the biological mechanisms underlying natural and artificial selection in farm animals. This will provide the basis for the design of more efficient selection strategies, but also will help to understand biological limits and constraints.

Acknowledgements

This study is part of the project FUGATO-plus GenoTrack and was financially supported by the German Ministry of Education and Research, BMBF, the Förderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven. SQ thanks the H. Wilhelm Schaumann Stiftung Hamburg for financial support. DG was partially supported by the Wisconsin Agriculture Experiment Station and by the Alexander von Humboldt Foundation.

References

Bernardo, J. M., and Smith, A. F. M. (1994). Bayesian Theory. Chichester: *Wiley*.
Carlson, C. S., Thomas, D. J., Eberle, M. A. *et al*. (2005). *Genome Research*, 15:1553–1565.
Churchill G. A., and Doerge R. W. (1994). *Genetics*, 138:963-971.
Cockerham, C. C. (1969). *Evolution*, 23:72-84.
Cole, J. B., VanRaden, P. M., O'Connell, J. R. *et al*. (2009). *J. Dairy Sci.*, 92:2931–2946.
Fay, J. C., and Wu, C. I. (2000). *Genetics*, 155:1405–1413.
Flori, L., Fritz, S., Jaffrezic, F. *et al*. (2009). *PLoS One*, 4:e6595.
Gianola, D., Simianer H. and Qanbari S. (2010). A Two-step Method for Detecting Selection Signatures Using Genetic Markers. *Genetical Research*, submitted.
Grisart B., Coppieters W., Farnir F. *et al*. (2001). *Genome Research*, 12:222–231.
Gu, J., Orr, N., Park, S. D. *et al*. (2009). *PLoS ONE*, 4:e5767.
Hayes, B. J., Chamberlain A. J., Maceachern S. *et al*. (2009). *Animal Genetics*, 40:176-184.
Leisch, F. (2004). *Journal of Statistical Software*, 11:1-18.
Marques E., Schnabel R., Stothard P. *et al*. (2008). *BMC Genetics*, 9:45.
Matukumalli, L. K., Lawley, C. T., Schnabel, R. D. *et al*. (2009). *PLoS ONE*, 4:e5350.
Maynard Smith, J., and Haigh, J. (1974). *Genetical Research*, 23:23–35.
Nielsen R., Bustamante C., Clark A. G. *et al*. (2005). *PLoS Biol*, 3:e170.
Nilsen, H., Olsen, H. G., Hayes, B. *et al*. (2009) *Anim. Genet.*, 40:701-712.
Qanbari, S., Pimentel, E., Tetens, J. *et al*. (2010a) *Animal Genetics*, doi:10.1111/j.1365-2052.2009.02016.x.
Qanbari, S., Gianola, D., Hayes, B. *et al*. (2010b). Application of Site and Haplotype-Frequency Based Approaches for Detecting Selection Signatures in Cattle. *BMC Genomics*, submitted.
Ron M., and Weller J. I. (2007). *Anim Genet.*, 38:429-39.
Sabeti, P. C., Reich, D. E., Higgins, J. M. *et al*. (2002). *Nature*, 419:832–837.
Sabeti P. C., Schaffner S. F., Fry, B. *et al*. (2006) *Science*, 312,1614-1620.
Sabeti, P. C., Varilly, P., Fry, B. *et al*. (2007). *Nature*, 449:913-918.
Schaffner S. F. *(*2004*.) Nat Rev Genet*, *5*:*43*–51.
Schaffner, S. F., Foo, C., Gabriel, S. *et al*. (2005). *Genome Res.*, 15:1576–1583.
Simianer, H., Szyda, J., Ramon, G., Lien, S. (1997). *Mammalian Genome*, 8:830-835.
Tajima, F. (1989). *Genetics*, 123:585–95.
The Bovine HapMap consortium (2009). *Science*, 324:528-532.
Voight, B. F., Kudaravalli, S., Wen, X. *et al*. (2006). *PLoS Biology*, 4:e72.
Wang, E. T., Kodama, G., Baldi, P. *et al*. (2006). *Proc. Natl. Acad. Sci. U.S.A.*, 103:135.
Weir, B. S. and Hill, W. G. (2002). *Annual Review of Genetics*, 36:721-750.
Wright, S. (1951). *Annals of Eugenics*, 15:323-354.
Wyckoff, G. J., Wang, W., and Wu, C. I. (2000). *Nature*, 403,304-309.