STUDY OF GENOMIC STRUCTURE AND SIGNATURES OF RECENT POSITIVE SELECTION IN CATTLE

Dissertation submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy (PhD) at the

Faculty of Agricultural Sciences,

Georg-August University, Göttingen

presented by

Saber Qanbari

born in Tabriz (Iran)

Göttingen, January 2009

D 7

Supervisor: Prof. Dr. Henner Simianer

Co-supervisor: Prof. Dr. Georg Thaller

Date of disputation: 25th January 2009

ACKNOWLEDGEMENTS

Most and above all, I would like to express my appreciation to my supervisor, Prof. Dr. Henner Simianer, for giving me an opportunity to work on this fascinating research topic and providing me the right guidance and support through the course of this research.

I am grateful to Prof. Dr. Georg Thaller for accepting the co-supervision of this thesis and serving on my referees committee.

My gratitude goes to the H. Wilhelm Schaumann Stiftung Hamburg for providing my PhD scholarship.

Thanks to Reza for all helps and Eduardo for sharing the office with me and accompanying me in daily lunch and trips beyond working.

With Enayat and Habib I have enjoyed pleasant and valuable friendship in Göttingen. I am grateful to them.

Flavio invited me to Guelph and provided me the opportunity to experience Canadian life style. During my stay in Guelph I enjoyed pleasing talks of Mehdi and Mohsen and gained enjoyable experiences like visiting Niagara Falls and CN-Tower, camping and canoeing with Mike and Anthony. I thank all of them.

Finally, throughout of my entire academic career, my family has been a constant source of inspiration. To them I am deeply grateful.

TABLE OF CONTENTS

SUMMARY

The knowledge of the extent and pattern of linkage disequilibrium (LD) is necessary for estimating the number of SNPs required for implementing association mapping studies as well as describing genomic structure of the bovine genome as a whole. In the first work of this study we used Illumina Bovine SNP50K BeadChip genotypes in a sample of German Holstein–Friesian cattle and developed a second generation of LD map statistics which has four times higher resolution compared to the maps available so far. These results revealed a lower level of LD for SNP pairs at distances ≤100 Kb than previously thought. The level of LD obtained in this study indicated that a denser SNP map would be beneficial to capture the LD information required for whole-genome fine mapping and genomic selection and to completely assess the pattern of LD across the genome.

Effective population size (Ne) was estimated based on the direct estimates of recombination rates from haplotype data and showed a persistent decline in about 100 individuals at the current generations. The impact of allele frequency in analyzing genome-wide LD was also explored in this part. Our observation revealed that minimizing the allele frequency difference between SNPs, reduces the influence of frequency on r^2 estimates and provides a useful metric for analyzing LD. The larger block size in Holstein cattle observed in this study indicates substantially greater LD in cattle than in human populations.

The second task of this thesis involved our attempts to find traces of decades of intensive artificial selection for traits of economically importance in modern cattle. In the first experiment we employed the recently described Extended Haplotype Homozygosity (EHH) test for tagging the genome wide footprints of positive selection in Holstein-Friesian cattle. This test uses the characteristics of haplotypes to detect selection by measuring the decay of haplotype homozygosity within a single population. To formally assess the significance of these results, we compared the combination of frequency and the Relative Extended Haplotype Homozygosity (REHH) value of each core haplotype with equally frequent haplotypes across the genome. A subset of the putative regions

showing the highest significance genome-wide was mapped. Regarding the fact that problems arising from multiple testing may have affected the results we performed a further validation by aligning the 12 regions of extreme REHH to the bovine genome (Btau 4.0) to verify any coincidence of the preliminary signals observed with important genomic regions. We found co-location of a panel of genes such as FABP3, CLPN3, SPERT, HTR2A5, ABCE1, BMP4 and PTGER2 and some others with putative regions. This panel represents a broad range of economically important traits such as milk yield and composition as well as reproductive and behavioral traits. We also reported high values of LD and a slower decay of haplotype homozygosity for some candidate regions harboring major genes related to dairy quality. The results of this study provided a genome wide map of selection footprints in Holstein genome.

In further experiments we exploited the variation among populations to explore the signatures of past selection. In this sense, we developed a new Bayesian approach for detecting differentiated loci based on F_{ST} and applied it to a set of geographically separated populations with identical or diverse breeding goals. This algorithm was able to deal with a large battery of marker information. Clustering the genome-wide estimates of F_{ST} values between Holstein and Brown Swiss versus Angus and Piedemontese breeds using Akaike's criterion recognized two groups, one representing putatively neutral loci, and the other possibly corresponding the genomic regions affected by selection.

We examined the potential of F_{ST} analysis in detecting selection signals by testing some candidate major genes in our data set. The results revealed F_{ST} values larger than expected ($P < 10\%$) for regions harboring the Casein cluster, GHR, STS, LP and IGF-1 genes which are supposed to be targets for artificial selection. However, we were not able to propose strong candidate genes on the basis of the gene content in the vicinity of extreme signals. As an explanation, we theorized that selection may work on genes that were not considered the primary targets of selection so far. Consistent with the previous reports our results mostly revealed gene deserts in the location of extreme peaks, which may reflect selection acting on uncharacterized regulatory region or simply fixation of non-coding DNA by genetic drift in the absence of any selection. Thus, these results in

combination with the observations on human population data suggest that non-coding regions have been an important substrate for adaptive evolution.

In a parallel analysis the integrated Haplotype Homozygosity Score (|iHS|) a derivation of EHH test, was applied for tracing on-going sweeps. After estimating |iHS| for each locus, we defined regions of the genome that may contain targets of positive selection as windows in the extreme of empirical distribution. This criterion resulted in 94 significant windows ($P \leq 0.05$). These results revealed significant enrichments for genes such as SPATA17, MGAT1, PGRMC2 and SRD5A2 in the region of clustered signals which belong to the number of functional categories relevant to reproduction including gamete generation, embryo development and spermatogenesis and genes in these categories may provide strong candidates for selection for fertility traits.

Another interesting observation is the presence of the genes like Actinin, Collagen and fibroblast activation protein as well as the gene responsible for developing the cartilage rudiments in the positively selected regions of beef cattle. These results suggest that selection for muscle related phenotypes play a major role in the shaping the beef cattle. These results generally are consistent with the previous reports and begin to suggest general themes about the types of genes that have been targets of positive selection in cattle genome.

Overall, based on the results of this study we conclude that high-resolution genome scans using dense markers are capable to identify outlier regions that potentially contain genes contributing to within and inter-breed phenotypic variation. Our results may be of future interest for identifying signatures of recent positive artificial selection between the cattle breeds or as additional evidence for any polymorphisms that show associations with beef or milk traits.

1st CHAPTER

INTRODUCTION

Preface

Most traits of economic importance in cattle are of complex and quantitative in nature. These traits are regulated by a combination of genes and environmental factors, which make it much more difficult to locate the genes controlling the trait of interest. Until recently, the genetic improvement has been achieved using conventional breeding programs which are based on the statistical evaluation of breeding values estimated from the phenotypes of an animal and its relatives. However, some of the traits cannot be improved very efficiently using the conventional breeding program for reasons such as low heritability of the traits, difficulty or expense in collecting phenotypes, or phenotypes collected later in life (Dekkers, 2004). Advanced genetic progress in such traits can be achieved by selection based on genetic markers (marker assisted selection; MAS). However, before the implementation of marker assisted selection, characterization of variants and their association with quantitative trait loci (QTL) in the genome of the respective breed is essential. Therefore, search for regions underlying the phenotypic variation of relevant traits is of great interest in breeding strategies which aim at using existing variation in those genes to select for superior individuals.

High-throughput genotyping

Since the trait-affecting gene is *a priori* unknown, all methods are base on neutral genetic markers. These are variations of the genome that can be genotyped at reasonable cost and time. Single nucleotide polymorphisms (SNPs) are the most abundant form of genomic variation and are defined as the single base pair position in DNA at which different sequences alternatives exist. SNPs are usually bi-allelic and, thus, show a low heterozygosity, but have the advantage of low mutation rates and low genotyping costs for large-scale genotyping through automation. The completion of the bovine genome sequence assembly (The Bovine Genome Consortium, 2009) formed a huge source of available SNP markers, suitable to carry out genome-wide studies (see below).

Currently, two major companies are providing fixed SNP panels for genome-wide SNP genotyping – Illumina and Affymetrix Inc. Both of these companies are offering very high throughput porcessing, high genotyping accuracy and low cost per SNP analysis. There are obvious advantages of having fixed SNP panels, including the possibility of combining datasets across laboratories and designing statistical methods for commonly used panels. SNPs are the marker of choice in general use today and in this study we use the bovine 50K Bead chip provided by Illumina Inc.

Search for genes underlying phenotypic variation

The search for the trait-affecting regions of the genome can be performed using either top-down or bottom-up genetic approaches (Ross-Ibarra *et al.,* 2007). In top-down methods (also called association mapping), researchers start with a phenotype of interest and dissect down to the underlying genetic basis. An association between a genetic variant and a phenotype would suggest that either the variability at that locus is the causative mutation underlying the QTL, or the variation is in linkage disequilibrium with the QTL. This approach usually requires positional cloning of QTL or association analyses targeting particular candidate genes identified based on homology to genes that are known to control the same, or similar, phenotypes in another species. Detection of such polymorphisms is an important prerequisite for marker assisted selection which will expedite genetic improvement of economically important traits. Although top-down approaches seem to be promising to dissect phenotypic variation in livestock populations, they are not without drawbacks. For example, positional cloning is both costly and labor-intensive, and such efforts have resulted in only a few successes in livestock systems (e.g., Grisart *et al.,* 2002; Van Laere *et al.,* 2003; Cohen-Zinder *et al.,* 2005). Moreover, while association mapping holds great promise when researchers have *a priori* knowledge of the genes that are likely to be regulating a trait of interest, such studies can produce a biased picture of the types of genes that are responsible for phenotypic evolution.

By contrast, bottom-up approaches involve the generation and statistical evaluation of population genomic data to identify likely targets of past selection. The main principles

of the population genomics approach to QTL mapping are that neutral loci across the genome will be similarly affected by genetic drift, demography, and evolutionary history of populations, while loci under selection will often behave differently and, therefore, reveal "outlier" patterns of variation. As such, functionally important genes can, at least in principle, be identified based on observed patterns of genetic variation even in the absence of information as to which trait(s) they regulate. Such bottom-up approaches provide a more or less unbiased view of the molecular basis of phenotypic evolution. The population genomics approach can also identify genes subjected to strong selection pressure and eventually fixed within breeds, and, in particular, genes involved in adaptation to extreme environments, disease resistance etc (Akey et al., 2002; Hayes *et al.,* 2009). Many of these traits, which are of great importance to the sustainability of animal breeding, are difficult or impossible to investigate by classic QTL mapping or association study approaches (Dekkers, 2004), often due to a lack of well defined phenotypes. Taking all of the above into account, it is clear that gene mapping strategies must be interpreted within the context of the genetic structure of the populations being studied.

Linkage disequilibrium

The basic factor influencing the outcome of statistical gene mapping strategies in animal species is the phenomenon of linkage disequilibrium (LD) or allelic association. An individual's chromosomal genotype consists of two haplotypes, one derived from the maternal gamete and the other from the paternal one. In a narrower sense, a haplotype is the particular combination of alleles that are inherited together as a unit (Figure 1).

Figure 1: A 3 SNP haplotype pair in a diploid individual.

LD refers to correlations among neighboring alleles reflecting haplotypes descended from single ancestral chromosomes (Reich *et al.,* 2001). Haplotypes are only disrupted by mutation and recombination in subsequent generations. Haplotypes therefore can be used as markers for tracking a variant allele in a population. Quantifying the extent of LD is the essential first step to determine how many markers are required to perform whole genome association studies. In addition, patterns of LD aid in exploring the different evolutionary forces that may have generated LD in certain regions of the genome (Ardlie *et al.,* 2002). Therefore, LD maps not only identify alleles that have undergone selection, but are also important for the design and application of association studies in cattle populations.

LD has been found to be variable both within and among loci and populations (Gabriel *et al.,* 2002; Pritchard and Przeworski, 2001). Since LD depends on the age of the SNP-creating mutations, the population history, genetic drift, the recombination fraction, gene conversion, admixture, hitchhiking, effective population size and selection (Ardlie *et al.,* 2002), it is highly variable even between close loci (see Chapter 2 for details).

A number of measures for the strength of LD have been proposed. To formally introduce pairwise LD measures, consider two bi-allelic loci A and B, possessing alleles A_1, A_2; B_1, B_2, respectively. Let p_{ij} denote the probability of haplotype (*i, j*), i.e. locus 1 exhibits the allele *i* and locus 2 the allele *j*. Let $p_{i.}$ and $p_{.j}$ denote the single frequencies of alleles *i* and *j* at loci 1 and 2, respectively. These probabilities can be arranged in a contingency table:

	1	2	$\sum$
1	p_{11}	p_{12}	$p_{1.}$
2	p_{21}	p_{22}	$p_{2.}$
$\sum$	$p_{.1}$	$p_{.2}$	1

Under linkage equilibrium, the expected haplotype frequencies are the product of the allele frequencies: $p_{ij} = p_i.p._j$.The deviation from the expectation for this particular haplotype is measured by:

$$D_{ij} = p_{ij} - p_i.p._j (i, j = 1,2)$$

For two bi-allelic loci, the absolute value of the deviation is the same for all four haplotypes: $D_{ij} = (-1)^{i+j} D$ where $D = p_{ij} - p_i.p._j$. Thus, the deviation for one haplotype describes the other three as well. However, linkage disequilibrium decays with time (*t*) and recombinational distance (*r*) according to the following formula:

$$D_t = (1-r)^t D_0$$

where D_0 is the extent of disequilibrium at some starting point and D_t is the extent of disequilibrium *t* generations later. Over time, recombination erodes linkage disequilibrium between alleles, which occurs more frequently between distantly located genes than between tightly linked genes. Therefore, *D* would be small between loci far apart from each other and would decrease with time as a result of recombination. Because of these dependencies, it has not been recommended to use *D* for measuring and comparing the level of LD but to use a standardized parameter (Ardlie *et al.,* 2002). The absolute value of *D'* (also called Lewontin's *D*') is calculated by dividing *D* by its maximum possible value, given the allele frequencies at the two loci (Lewontin, 1964).

$$D' = \begin{cases} \dfrac{D}{\min(p_1.p._2, p._1 p_2.)} : D > 0 \\ \dfrac{D}{\min(p_1.p._1, p._2 p_2.)} : D < 0 \end{cases}$$

When *D'* equals 1, this suggests that the two loci are in complete LD and there has been no recombination between them. When *D'* is less than 1, it means that the two loci have been separated by recombination. When *D'* equals 0, it signifies no LD.

Another measure of linkage disequilibrium is the square of the correlation coefficient (r^2) between marker alleles (Hill and Robertson, 1968). It is calculated as D divided by the product of the four allele frequencies at the two loci:

$$r^2 = \frac{D^2}{p_1 \cdot p_2 \cdot p_{\cdot 1} p_{\cdot 2}}$$

When r^2 is equal to one for two markers, it shows perfect linkage disequilibrium and one marker provides complete information about the other marker, making the other marker redundant (Ardlie *et al.,* 2002).

Although these measures are useful to assess pairwise LD, they cannot consider more than two loci and, thus, are blind to simultaneous associations between alleles of more than two loci. Furthermore, the measure *D'* is not suitable for differentiating different degrees of LD. It equals ±1 if at least one haplotype is missing (Ardlie *et al.,* 2002). Missing haplotypes are more probable for rare SNP alleles and for multiple SNP sequences than for short sequences of common SNPs. Also, the measure *D'* shows much more inflation than r^2 when small or moderate sample sizes are used (McRae *et al.,* 2002, Weiss and Clark 2002). In other words, for small to moderate sample sizes, estimates of *D'* can exhibit a considerable upward bias (Teare *et al.,* 2002, Terwilliger *et al.,* 2002). Even if *D'* is estimated to be below 1, it might be strongly biased. The strength of LD between a trait locus and a marker, measured by r^2, is indirectly proportional to the power of finding an association (Kruglyak, 1999; Pritchard and Przeworski, 2001; Teare *et al.,* 2002). As such, the decline of r^2 with distance determines how many markers are required in a genome scan to detect a QTL, which cannot be predicted by using *D'* (Hayes, 2007). Therefore, *D'* is rather an indicator for missing haplotypes, perhaps due to absent recombination events, than a reliable measure of LD. Early LD studies in cattle used the measure *D' (*e.g., Farnir *et al.,* 2001; Tenesa *et al.,* 2003), but r^2 has recently emerged as a measure of choice for comparing the extent of LD (Pritchard and Przeworski, 2001, Weiss and Clark, 2002).

Signatures of the positive selection

Detection of signatures of selection is an important tool to identify potential genes that might underlie economically important traits and which will improve our ability to link genetic variants to the phenotype of interest (Hayes *et al.,* 2008; The Bovine HapMap consortium, 2009). The modern cattle has been intensively selected during the last centuries, as such, it has achieved tremendous phenotypic changes over the past 40 years. Consequently, genomic regions controlling traits of economic importance are expected to exhibit footprints of selective breeding. However, it is unknown how selection has changed the Holstein genome and what genome changes are associated with the phenotypic changes. The advent of the bovine genome sequence and the flood of new polymorphism data that has come with it (Matukumalli *et al.,* 2009) has provided valuable new tools in the search for traces of the recent selection in the cattle genome (e.g., The Bovine HapMap consortium, 2009, Hayes, 2009).

To this end a number of statistical tests have been developed mostly by human geneticists to explore different aspects of how to infer deviations from what is expected with regard to genetic variability under a neutral model (e.g., Tajima, 1989; Fay and Wu, 2000; Sabeti *et al.,* 2002; Voight *et al.*, 2006; Akey *et al.,* 2002, among some others).

Although all statistics are based on neutral genomic variation, not all of them rely on the same kind of information. These methods can be categorized into two classes, named Class I, and II, according to the information used. Class I tests are based on the frequency spectrum of single mutations in the sample in contrast to class II estimators which principally are based on the haplotype distribution. From class I estimators, the most relevant tests are Tajima's *D* (Tajima, 1989); Fu and Li's *D, F, D** and *F** (Fu and Li, 1993); Fay and Wu's *H* (Fay and Wu, 2000), and *R2* (Ramos-Onsins and Rozas, 2002). Most of these estimators were designed for full-sequence data and not for genome wide collections of pre-ascertained SNPs that are currently available in some livestock species.

Some of the most easily distinguished traces left by the forces of selection are those left by selective sweeps. Selective sweeps occur when an allele becomes more frequent in a

population as a result of positive selection. As the positively selected allele increases in frequency, linked nearby alleles will do so, too, a phenomenon known as genetic hitchhiking (Smith and Haigh, 1974). A strong selective sweep will result in a region of the genome where the positively selected haplotype (of the selected variant and linked neighboring alleles) is at high frequency, thus leading to a reduced haplotype diversity in the region. Thus the occurrence of a selective sweep can be investigated by measuring LD or by observing if a haplotype is overrepresented in a population. This is the basic for Class II statistics. Within this class the most relevant statistics are F_S (Fu. 1997), Dh (Nei, 1987), B and Q (Wall, 1999), Z_{nS} (Kelly, 1997), Z_A and ZZ (Rozas *et al.,* 2001), populations differentiation index F_{ST} (Akey *et al.,* 2002), EHH (Sabeti *et al.,* 2002) and its extensions such as iHS (Voight *et al.,* 2006) and XPEHH (Sabeti *et al.,* 2007) estimators. Among the various statistics used for recognizing signals of positive selection from polymorphism data, the EHH, iHS and F_{ST} estimators are particularly useful and would be the methodologies of choice in this study.

Extended haplotype homozygosity (EHH) test

As a selective sweep carries an allele on a specific haplotype to high frequency faster than the rate at which it is broken down by recombination, high frequency haplotypes will be observed longer than expected under neutrality (Sabeti *et al.,* 2002). This phenomenon has been exploited in the 'Extended Haplotype Homozygosity' algorithm for detecting recent positive selection and may be useful in detecting more recent positive selection (see chapter 3 for details).

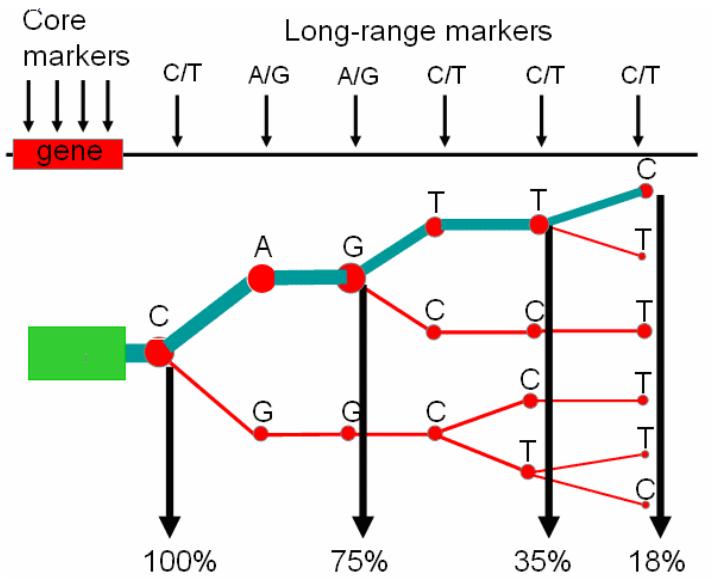

Figure 2. The decay in haplotype homozygosity as a function of distance from the mutation of interest. Haplotype homozygosity is defined as the probability, at any distance, that any two haplotypes that start out the same have all the same SNP genotypes (From presentation by: David Reich, Broad Institute).

Integrated haplotype homozygosity (iHS) test

This test uses the EHH statistic, which measures the decay of haplotype homozygosity, as a function of distance of haplotypes that carry a specified core allele at one end. In this concept, directional selection favoring a new mutation results in a rapid increase in the frequency of the selected allele along with the background haplotype on which the mutation arose. This phenomenon increases LD on the chromosomes which harbor the derived (selected) allele. Thus, this measure is most sensitive to a rapid increase in the frequency of the derived allele at a selected site, but the derived allele must have existed on few distinct backgrounds (haplotypes) prior to selection and must not have reached fixation yet (Voight et al. 2006; Sabeti et al. 2007) (see chapter 5 for details).

Population differentiation index (F_{ST})

One of the most widely used methods to detect differential selective pressures between populations is F_{ST}, a measure of the proportion of the genetic variance explained by differences among populations. Fst can be used to find genes under local selection by

comparing the F_{ST} value of a single locus against the genome-wide values (Akey *et al.,* 2002). Allele frequency differences between populations are mainly caused by genetic drift, that is, by the random process driven by demographic history. Drift affects all loci across the genome in a similar fashion. Loci under selection will often behave differently and, therefore, reveal "outlier" patterns of variation, loss of diversity (increase of diversity if the loci were under a balanced selection), and through hitchhiking effects selection will also influence linked markers, allowing the detection of a "selection signature" (outlier effects). This signal can often be detected by genotyping a large number of markers along a chromosome and identifying clusters of outliers (see chapters 4 and 5 for more details). A large number of studies have been published based on this principle, building genome-wide empirical distributions of F_{ST} based on increasing numbers of autosomal SNPs (Akey *et al.,* 2002; Hayes et al., 2009; among many others).

Scope of the thesis

The knowledge of the extent and pattern of LD is necessary for estimating the number of SNPs required for implementing association mapping studies as well as describing certain genomic regions. It provides also a better understanding of genomic structure from which we can make some tentative inferences about the bovine genome as a whole. As a first scope for this thesis we generate a new generation of high density LD map of Holstein cattle describing genetic structure based on genotyping thousands of SNPs. This issue is presented in chapter 2 of this thesis.

The second scope of this thesis looks for the traces of decades of intensive artificial selection for traits of economically importance in modern cattle and shedding light on possible selective events of genes involved which is obviously of great interest. These objectives would be accomplished by application of EHH and iHS statistics to develop genome-wide map of signatures of recent positive selection. These results are reported in chapter 3 and partly in chapter 5 of the thesis, respectively.

The F_{ST} measure of Wright (1951) will also be used to examine the differences in allele frequencies due to selection. To this purpose we develop a new and simple Bayesian algorithm for estimating a population differentiation index. Chapter 4 presents this algorithm in detail. Finally, we apply this new estimator to measure the population division among a set of cattle breeds with diverse breeding goals and compare the results with other statistics. The last results would be covered in chapter 5 followed by a general discussion in chapter 6.

References

Akey, J. M., Zhang, G., Zhang. K., Jin, L., Shriver. M. D. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12: 1805-1814.

Ardlie, K. G., Kruglyak, L., and Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3: 299-309.

Cohen, M., Seroussi, E., Larkin, D. M., Loor, J. J., Everts-van der Wind, A., Heon-Lee, J., Drackley, J. K., Band, M. R., Hernandez, A. G., Shani, M., Lewin, H. A., Weller, J. I., Ron, M. 2005 Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15: 936–44.

Dekkers J. C. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. Journal of Animal Science 82: 313-328.

Farnir, F., Coppieters, W., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., and Georges, M. 2000. Extensive Genome-wide Linkage Disequilibrium in Cattle. Genome Research 10: 220-227.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., Mni, M., Simon, P., Wagenaar, D., 2000. Extensive genome-wide linkage disequilibrium in cattle. Genome Research 10: 220-7.

Fay, J. C., and Wu C. I. 2000. Hitchhiking under positive Darwinian selection. Genetics 155: 1405–13.

Fu, Y. X. and Li, W. H. 1993. Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., Altshuler, D. 2002. The structure of haplotype blocks in the human genome. Science 296: 2225–2229.

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J. J., Kvasz, A., Mni, M., Simon, P., Frere, J. M., Coppieters, W., Georges, M. 2001. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12: 222–231.

Hayes, B. J., Lien, S., Nilsen, H., Olsen, H. G., Berg, P., Maceachern, S., Potter, S., Meuwissen, T. H. 2008. The origin of selection signatures on bovine chromosome 6. Animal Genetics 39: 105-111.

Hayes, B.J., Chamberlain, A. J., Maceachern, S., Savin, K., McPartlan, H., MacLeod, I., Sethuraman, L., Goddard, M. E. 2009. A genome map of divergent artificial selection between Bos Taurus dairy cattle and Bos taurus beef cattle. Animal Genetics 40: 176-184

Hill., W. G., and Robertson, A. 1966. The effect of linkage on limits to artificial selection. Genetical Research 8: 269–294.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22:139-144.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49–67.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P., Sonstegard, T. S., Van Tassell, C. P. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4: e5350

Maynard-Smith, J., and Haigh, J. 1974. The hitch-hiking effect from a favourable gene. Genetical Research 23: 23–35.

McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Crawford, A. M., Slate, J. 2002. Linkage disequilibrium in domestic sheep. Genetics 160: 1113–1122.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Pritchard, J. K., and Przeworski, M. 2001. Linkage disequilibrium in humans: models and data. American Journal of Human Genetics 69: 1–14.

Ramos-Onsins, S. E. and Rozas, J. 2002. Statistical properties of new neutrality tests against population growth. Molecular Biology and Evolution 19: 2092-2300.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., Lander, E. S. 2001. Linkage disequilibrium in the human genome. Nature 411: 199–204.

Ross-Ibarra, J., Morrell, P. L., and Gaut, B. S. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proceeding of the National Academy of Sciences of USA 104: 8641–8648.

Rozas, J., M. Gullaud, G. Blandin., and Aguade, M. 2001. DNA variation at the rp49 gene region of drosophila simulans: Evolutionary inferences from an unusual haplotype structure. Genetics 158: 1147-1155.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E. *et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Teare, M. D., Dunning, A. M., Durocher, F., Rennart, G., and Easton, D. F. 2002. Sampling distribution of summary linkage disequilibrium measures. Annals of Human Genetics 66: 223–233.

Tenesa, A., Knott, S. A., Ward, D., Smith, D., Williams, J. L., Visscher, P. M. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. Journal of Animal Science 81: 617–623.

Terwilliger, J. D., and Weiss, K. M. 1998. Linkage disequilibrium mapping of complex disease: fantasy or reality? Current Opinion in Biotechnology 9: 578–594.

Terwilliger, J. D., Haghighi, F., Hiekkalinna, T. S., Göring, H. H. 2002. A biased assessment of the use of SNPs in human complex traits. Current Opinion in Genetics & Development 12: 726–34.

The Bovine Genome Sequencing and Analysis Consortium. 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science 324: 522-528.

The Bovine HapMap consortium. 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324: 528-532

Van Laere, A. S., Nguyen, M., Braunschweig, M., Nezer, C., Collette, C., Moreau, L., Archibald, A.L., Haley, C. S., Buys, N., Tally, M., Andersson, G., Georges, M.,

Andersson, L. 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425: 832–836.

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

Wall, J. D. 1999. Recombination and the power of statistical tests of neutrality. Genetical Research 74: 65-79.

Weiss K.M, and Clark A.G. 2002. Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19-24.

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics 15: 323-54.

2nd CHAPTER

The Pattern of Linkage Disequilibrium in German Holstein Cattle

S. Qanbari[*], E. C. G. Pimentel[*], J. Tetens[§], G. Thaller[§], P. Lichtner[†], A.R. Sharifi[*] and H. Simianer[*]

[*] Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

[§] Institute of Animal Breeding and Animal Husbandry, Christian-Albrechts-University, 24098 Kiel, Germany

[†] Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Animal Genetics 2010, 41: 346–356

ABSTRACT

We used DNA samples of 810 German Holstein–Friesian cattle genotyped by the Illumina Bovine SNP50K BeadChip to analyze linkage disequilibrium (LD) structure. A panel of 40'854 (75.65%) markers was included into the final analysis. The pair-wise r^2 statistic of SNPs apart up to 5Mbp across the genome was estimated. A mean value of r^2=0.30 ± 0.32 was observed in pair-wise distances of <25 Kb and it dropped to 0.20 ± 0.24 at 50 to 75 Kb, which is nearly the average inter-marker space in this study. The proportion of SNPs in useful LD was 26% for the distance of 50 and 75 Kb between SNPs. We found a lower level of LD for SNP pairs at the distance ≤100 Kb than previously thought. Analysis revealed 712 haplo-blocks spanning 4.6 % of the genome. Mean and median block length were estimated as 164. ± 117.1 and 144 Kb, respectively. Analysis of effective population size based on the direct estimates of recombination rates from SNP data showed a decline in effective population size to 103 up to ~4 generation ago. The impact of allele frequency in analyzing genome-wide LD was also explored in this study. The observations revealed that minimizing the allele frequency difference between SNPs, reduces the influence of frequency on r^2 estimates. This study presents a second generation of LD map statistics for the Holstein genome which has four times higher resolution compared to the maps available so far.

INTRODUCTION

Linkage disequilibrium (LD) defined as the non random relationship between loci has recently been in the focus of attention. LD is the structural basis of 'Genomic Selection' programs (Meuwissen *et al*. 2001) and helps to determine the actual genes responsible for variation of economically important traits (Van Laere *et al*. 2003; Grisart *et al*. 2004) through association mapping. The feasibility and efficiency of these approaches depends strongly on the extent, distribution and structure of LD, which determine how many markers are required for a genome scan in the population under study (Khatkar *et al*. 2007). Moreover, for high-resolution association mapping, it is also necessary to identify block-like structures of haplotypes and a minimal set of polymorphisms (haplotype tagging SNPs; htSNPs) that capture the most common haplotypes of each block (Johnson *et al.* 2001; Dawson *et al.* 2002). Due to the variation in local recombination rates, mutation rates, and genetic hitchhiking the breakdown of LD is often discontinuous producing haploypic tracts across the genome (Ardlie *et al.* 2002; International Hapmap Consortium, 2005). Simianer *et al*. (1997) demonstrated that this variability is also prevalent in the bovine genome and recombination probabilities even differ between families. As a result, today's chromosomes comprise a mosaic of haplotype blocks derived from ancestral chromosome fragments (e.g., Khatkar *et al.* 2007) and shared discrete haplotype blocks and LD patterns can be observed even in apparently unrelated individuals and populations (Gautier *et al*. 2007; Marques *et al*. 2008). Identifying these continental tracts can provide haplotypes to be used as genetic markers and delimit regions where htSNPs can reasonably be defined. They could also provide information on the spacing of SNPs in association studies, i.e. where SNPs should be considered and where not. By adjusting for the differences in recombination rates across the genome haplotype blocks can also be used for identifying the signatures of recent positive selection (Sabeti *et al*. 2001).

With the availability of new technologies of SNP genotyping an increasing number of studies have aimed at quantifying LD characteristics in domestic animals, especially in cattle. Most of these studies used a low marker density or were done in limited regions of

the studied genomes. Farnir *et al*. (2000) performed the first whole-genome LD study to characterize the extent and pattern of LD based on the information of 284 microsatellite markers in Dutch Holstein cattle. Several subsequent studies have confirmed extensive LD in cattle (Khatkar *et al*. 2006a; Odani *et al*. 2006; McKay *et al*. 2007, Marques *et al*. 2008). They all describe an extensive LD and revealed that different measures of LD such as r^2 and D' yield different conclusions in terms of the extent of LD. Recently, Sargolzaei *et al*. (2008) and Kim & Kirkpatrik (2009) reported a genome-wide LD profile based on the Affymetrix 10K SNP array in Holstein population of North America. Khatkar *et al*. (2007) reported a comprehensive genome-wide profile of LD statistics and haploblock characteristics based on a panel of 15,036 single nucleotide polymorphisms (SNP) in Australian Holstein-Friesian cattle. The final average inter-marker spacing in their study was 251.8 Kb which is by the factor 5×10^{-3} less dense than the panel currently being used in LD analysis of human genome. However it is now known that BTAu_3.1 build used to physically locate SNPs in their study has inconsistencies with other independently built cattle maps (Marques *et al*. 2007, Snelling *et al*. 2007). More recently Villa-Angulo *et al*. (2009) used a panel of 31'857 SNPs generated by the Bovine HapMap Consortium to characterize a high-resolution haplotype block structure of 19 breeds of different geographic origin. They focused mainly on 101 high density regions spanning up to 7.6 Mb on three chromosomes 6, 14 and 25 with an average density of approximately one SNP per 4 Kb.

With the availability of larger-scale SNP data sets it has become possible to construct LD maps with higher resolution. In this study we use SNP data generated with the Illumina Bovine SNP50K BeadChip to create a second generation LD map of Holstein-Friesian cattle. We also explore some properities of r^2 as the most common measure of LD in this study.

MATERIALS AND METHODS

Data preparation and haplotype reconstruction

Semen or blood samples from 810 German Holstein–Friesian cattle including 469 bulls and 341 bull dams were used as the source of genomic DNA and were genotyped using the Illumina Bovine SNP50K BeadChip. This chip contains a total of 54'001 SNPs with a mean neighbor marker distance of 48.75 Kb. 1728 SNP loci were excluded because of unknown genomic position and 11 markers were monomorphic. For the purposes of this study, only autosomic SNPs with minor allelic frequencies (MAF) ≥ 0.05 were included in the LD analysis. The number of heterozygous loci was determined and used to estimate the average heterozygosity for all individuals. The allele frequencies, observed heterozygosity and expected heterozygosity for each SNP were determined.

For this analysis fully phased haplotype data were required. After the aforementioned filtering process we reconstructed haplotypes for each chromosome using default options in fastPHASE (Scheet & Stephens 2006).

Measure of LD

Several statistics have been used to measure the LD between a pair of loci. The two most common measures are the absolute value of D', and r^2, both derived from Lewontin's D (Lewontin 1964). We used r^2 which is generally accepted as the more robust and better interpretable LD parameter (Kruglyak 1999; Ardlie *et al*. 2002; Terwilliger *et al*. 2002)

Consider 2 loci, A and B, each locus having 2 alleles (denoted A_1, A_2; B_1, B_2, respectively). We denote f_{11}, f_{12}, f_{21}, and f_{22} as the frequencies of the haplotypes A_1B_1, A_1B_2, A_2B_1, and A_2B_2, respectively; f_{A1}, f_{A2}, f_{B1}, and f_{B2} are the frequencies of A_1, A_2, B_1, and B_2, respectively. Following Hill and Weir (1994),

$$r^2 = \frac{(f_{11}f_{22} - f_{12}f_{21})^2}{f_{A1}f_{A2}f_{B1}f_{B2}}$$

LD haplo-block partitioning

Existing block definition algorithms are based on two alternative methods: Either pairwise *D'* values above a lower limit are used to detect regions of little or no recombination (Gabriel *et al*. 2002; Daly *et al*. 2001; Wang *et al*. 2002), or blocks are defined by employing some haplotypic diversity criterion, where a small number of common haplotypes provide high chromosomal frequency coverage (Patil *et al*. 2001; Zhang *et al*. 2002, 2003; Anderson & Novembre 2001). For the purpose of this study we used the algorithm suggested by Gabriel *et al.* (2002) defining a pair of SNPs to be in ''strong LD'' if the upper 95% confidence bound of D' is between 0.7 and 0.98. Reconstructed haplotypes were inserted into HAPLOVIEW v4.1 (Barrett *et al.* 2005) to estimate LD statistics and constructing the blocking pattern as well as identifying haplotype tagging SNPs for all 29 autosomes.

Estimating effective population size using LD

According to Wright (1938) effective population size (N_e) is defined as "the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration". N_e provides useful information about the population evolution and improves the understanding and modeling of the genetic architecture underlying complex traits (Reich & Lander 2001). N_e can be estimated from LD data and the availability of dense markers has made this option feasible. Sved (1971) has formulized the relationship of LD and N_e in the absence of mutation as $r^2 = 1/ (4N_e c + 1)$ where c represents the linkage map distance in Morgan. If mutation is accounted for in the model, the expectation of r^2 is $1/ (4N_t c + 2)$, where N_t is the population size $1/2c$ generations ago. For more information we refer to Tenesa *et al*. (2007). In this study we assessed genetic distance c directly by estimating the recombination rates across the genome using PHASE v.2.1 (Li and Stephens 2003). To this purpose, random segments of 15 Mbp were selected on each autosome. The recombination model was applied based on 100 individuals and increasing the number of iterations of the final run 10 times to

obtain better estimates of uncertainty. The prior value for effective population size was set to 100. In order to save the computing time, we used known haplotypes with fragment sizes of 12 bp. An average of N_e over chromosomes was then calculated corresponding to the various times in the past. We inferred N_e for each autosomal chromosome at distance bins of <0.025, 0.025-0.05, 0.05-0.1, 0.1-0.5, 0.5-1, 1-5 and 5-15 cM. This range of linkage map distance infers the past effective population size up to 2000 generations ago.

RESULTS

Marker statistics and genetic diversity

A total of 40'854 (75.65%) markers passed the above filtering criteria and were included into the final analysis. This subset of markers covers 2544.1 Mbp of the genome with 62.27 ± 58.3 Kbp average adjacent marker spacing. The largest gap between SNPs (2081.5 Kbp) was located on chromosome 10. For the SNPs analyzed in this study, the average observed heterozygosity and mean MAF were estimated as 0.37 ± 0.12 and 0.28 ± 0.15, respectively. Figure 1 displays the distribution of the MAF of SNPs genotyped. The almost uniform distribution across frequency classes likely is due to the construction of the SNP array which was optimized with respect to a uniform SNP spacing and MAF distribution. The observed heterozygosity in the studied Holstein population averaged as 0.23.

Table 1. Genome wide summary of marker and haplotype blocks in the Holstein cattle

Chr	Initial (n)	Final (n)	Chr-Length (Mbp)	Linkage Map (cM)	Block (n)	Block_length (Kb)	Mean_BL ± SD (Kb)	BSNPs* (n)	htSNPs (n)	Max Gap (Kb)
1	3343	2641	161.1	154	57	9159	160.7 ± 112	267	2263	683.9
2	2764	2149	140.6	126	35	6181	176.6 ± 146	167	1876	651.9
3	2566	2037	127.9	128	52	7428	142.8 ± 102	216	1790	813.7
4	2541	1999	124.1	119	41	7173	175.0 ± 118	197	1759	889.7
5	2181	1718	125.8	135	30	6333	211.1 ± 147	149	1521	1050.5
6	2535	2044	122.5	134	46	7918	172.1 ± 119	225	1778	826.2
7	2294	1767	112.1	135	34	6919	203.5 ± 127	177	1519	657.0
8	2362	1849	116.9	128	42	7586	180.6 ± 108	196	1588	738.3
9	2036	1623	108.1	116	20	3879	194.0 ± 136	89	1469	760.8
10	2179	1713	106.2	118	40	4787	119.7 ± 93	166	1519	2081.5
11	2267	1813	110.2	130	27	4658	172.5 ± 104	126	1624	989.5
12	1683	1320	85.3	109	20	3102	155.1 ± 154	85	1190	788.7
13	1802	1396	84.3	105	32	4793	149.8 ± 93	136	1227	608.9
14	1722	1356	81.3	103	28	5402	192.9 ± 96	141	1166	576.0
15	1688	1365	84.6	109	19	3157	166.2 ± 116	81	1245	660.2
16	1606	1251	77.8	94	31	6443	207.8 ± 274	151	1087	1015.4
17	1585	1284	76.5	95	14	1971	140.8 ± 68	64	1170	840.4
18	1351	1100	66.1	84	12	1635	136.3 ± 54	53	1012	896.4
19	1378	1108	65.2	109	15	2876	191.7 ± 96	77	1006	553.1
20	1564	1252	75.7	82	23	3713	161.4 ± 89	102	1099	837.1
21	1419	1093	69.2	83	16	2279	142.4 ± 102	65	985	849.4
22	1299	1009	61.8	88	15	1723	114.9 ± 85	58	903	601.3
23	1083	871	53.3	80	7	1500	214.3 ± 150	35	805	476.3
24	1294	1013	64.9	78	13	1944	149.5 ± 113	56	916	527.3
25	987	810	44.0	68	15	1834	122.3 ± 94	68	752	589.9
26	1086	849	51.7	79	12	2136	178.0 ± 180	48	763	682.6
27	977	798	48.7	67	3	408	136.0 ± 46	13	748	1776.8
28	942	779	46.0	61	3	461	153.7 ± 156	12	740	470.6
29	1048	847	52.0	69	10	1461	146.1 ± 104	38	781	1505.8
Total	**51'582**	**40'854**	**2544.1**	**2986**	**712**	**118'859**	**164.4 ± 117**	**3258**	**36'301**	**2081.5**

Number of SNPs forming haplo-blocks

Pattern of haplotype blocks

Critical for association studies is the identification of haplotype blocks and the minimal set of htSNPs required to capture haplotype variation in a population sufficiently, which will reduce cost and effort. Table 1 presents a descriptive summary of genome wide marker and haplo-block distribution in the data set analyzed. A total of 712 haplo-blocks spanning 118'859 Kb (4.67 %) of the genome were detected. Mean and median block length were estimated as 164. ± 117.1 and 144 Kb, respectively, with a maximum of 1261 Kb. The distribution of haplotype block size is depicted in Figure 2.

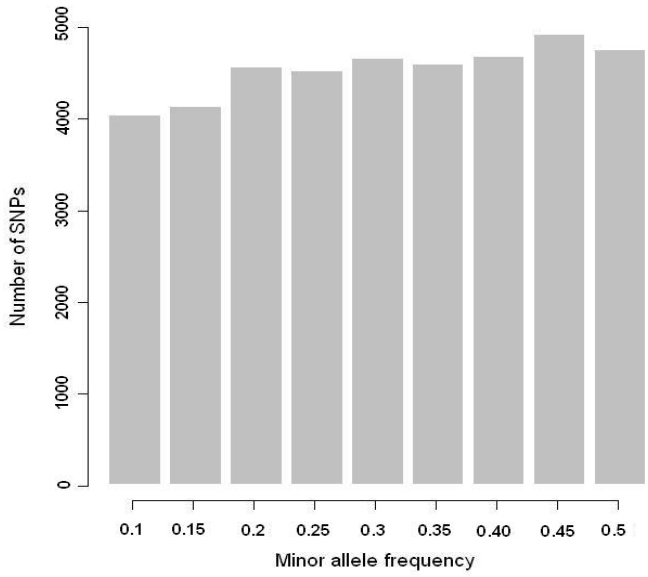

Figure 1. Minor allele frequency of SNPs.

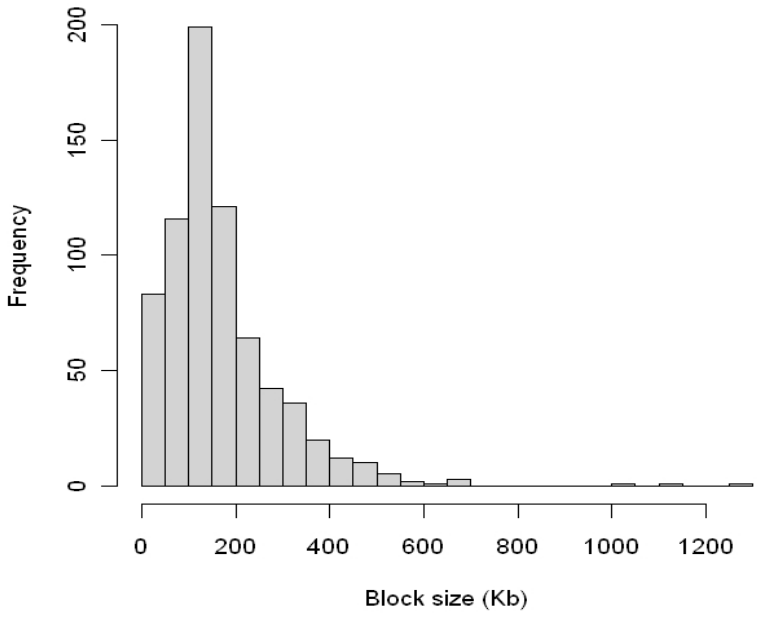

Figure 2. Distribution of haplo-block size in Holstein cattle genome.

Chromosome 1 having 57 blocks spanning 9159 Kb and Chromosome 27 with 3 blocks covering 408 Kb showed the longest and shortest haplotypic structures in the genome. In total, 3258 SNPs (7.97 % of all used SNPs) formed blocks with a range of 2-11 SNPs per tract. Using the tagger option incorporated in HAPLOVIEW, 36'301 SNPs (89% of all used SNPs) were tagged in the data set analyzed. These SNPs can tag either neighboring markers or a set of common haplotypes within an LD block. Figure 3 displays the distribution of htSNPs across the genome of studied population.

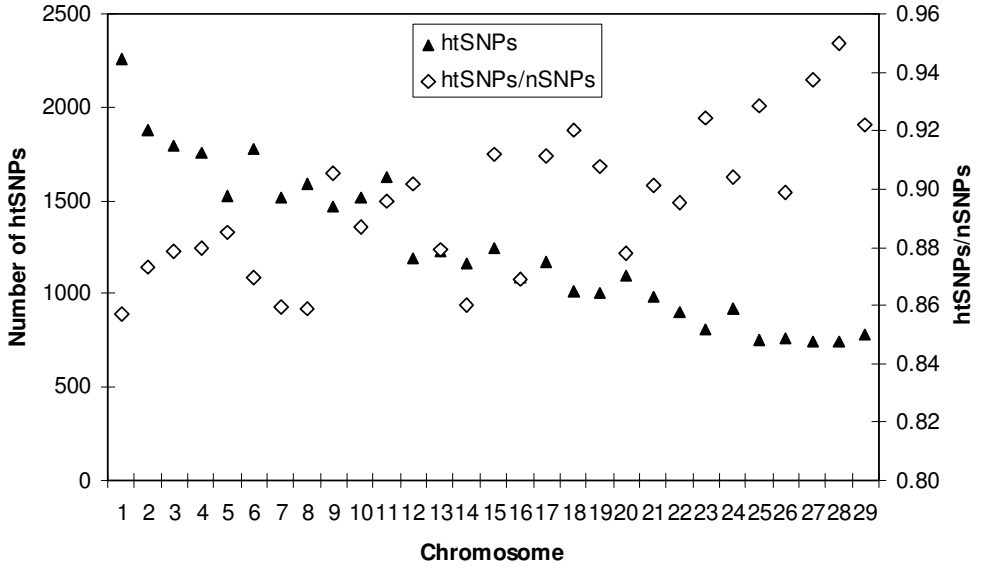

Figure 3. Distribution of htSNPs across the genome of Holstein population studied. Triangles displays the number of htSNPs for each chromosome and diamonds represents the ratio of htSNPs versus SNPs analyzed for each chromosome.

Extent of LD across the genome

All possible SNP pairs with distance ≤ 5 Mbp on the same chromosome produced 3'216'038 pair-wise LD values on the 29 bovine autosomes. In order to visualize the decay of LD and the proportion of pair markers in useful LD we stacked r^2 and plotted them as a function of inter-marker distance categories (<0.025, 0.025-0.05, 0.05-0.075, 0.075-0.12, 0.12-0.2, 0.2-0.5, 0.5-1.5, 1.5-3 and 3-5 (Mbp) (Figure 4). This genome-wide bar plot illustrates the rate at which LD decays with physical distance and forms the basis for comparison between studies. We observed an inverse relationship between LD and marker distance, confirming recent studies on r^2 measures in cattle. Overall, six cases of complete LD (r^2 = 1.0) were observed for the entire genome.

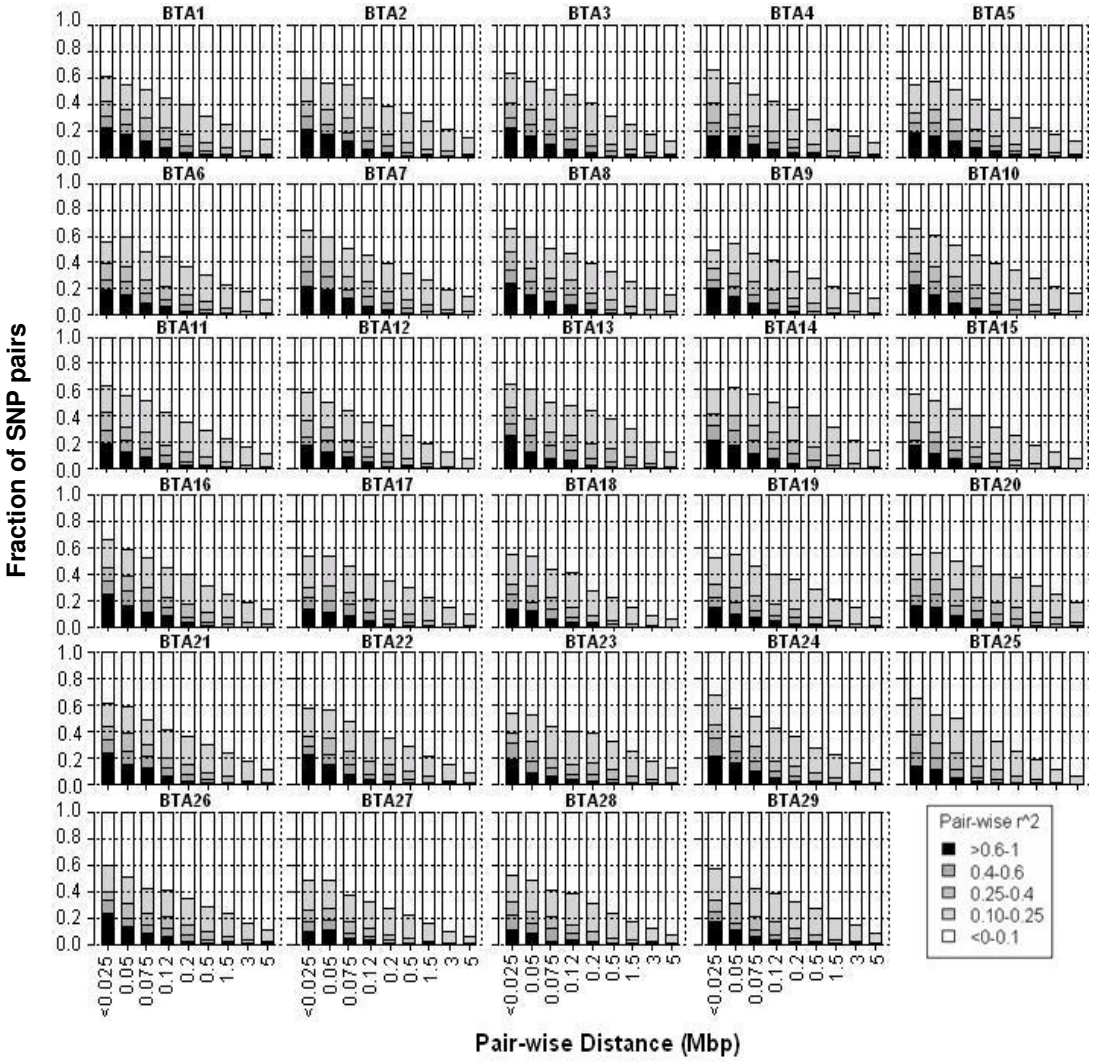

Figure 4. Level of LD decay as a function of distance between pairs of SNPs up to 5 Mbp for the entire genome.

The mean r^2 values and the proportion of SNP pairs that shows statistically significant LD for SNP pairs apart up to 5 Mbp are presented in Table 2. A mean value of r^2=0.30 ± 0.32 was observed in pair-wise distances of <25 Kb and it dropped to 0.20 ± 0.24 at 50 to 75 Kb, the interval which includes the average inter-marker space in this study. In contrast an overall mean value of r^2= 0.21 ± 0.26 was observed for SNPs less than 100 Kb apart from each other compared with r^2= 0.59 presented by Sargolzaei *et al*. (2008) for the north American Holstein cattle. The similar study by Kim & Kirkpatrik (2009) revealed strong LD (r^2 > 0.8) in genomic regions of

approximately 50 Kb or less which is much larger than the observation of this study (r^2= 0.29).

The threshold for useful LD that was chosen in this study is 0.25. With this threshold and considering that on average 1 cM is equivalent to 1 Mb, useful LD extended over 0.5–1.5 cM so that the proportion of SNP pairs in useful LD is above 5%. The proportion of SNPs in useful LD was 39% for the distance of 25 Kb or less between SNPs. This proportion drops to 0.26% for SNPs between 50 and 75 Kb apart from each other. Overall, for SNPs less than 100 Kb apart from each other the proportion of SNPs in useful LD was 0.29 %. This proportion was reported as 68.34% by Sargolzaei *et al*. (2008) even with a higher threshold as 0.3. However, the substantial LD estimated for SNP pairs more than 100 Kb apart (r^2= 0.14) is similar.

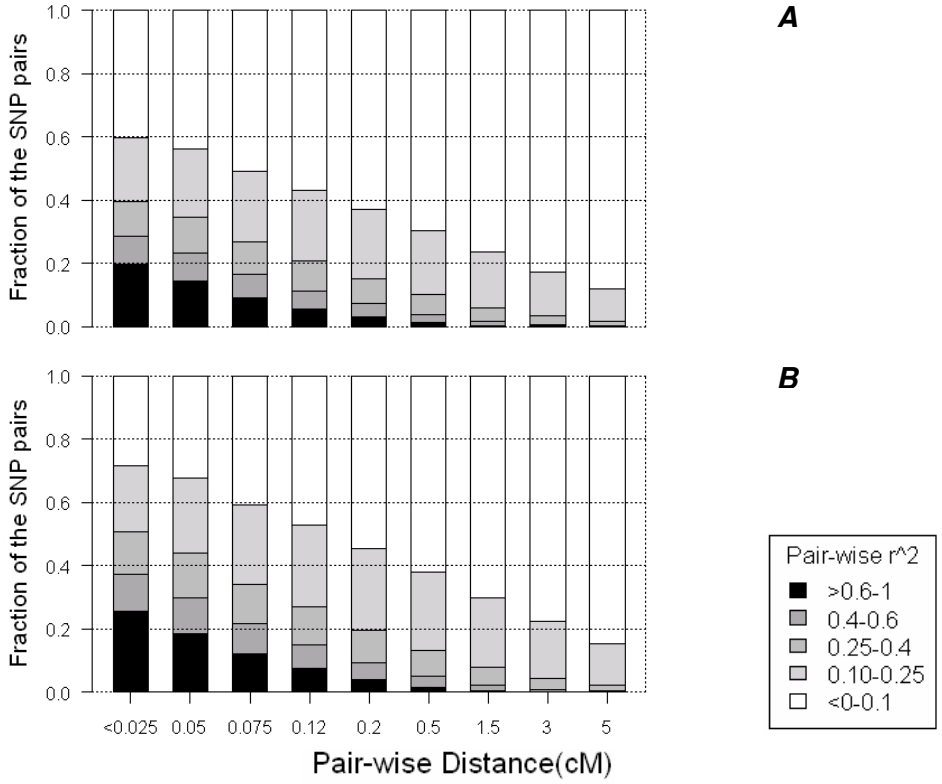

Figure 5. Comparison of fraction of marker pairs with different r^2 levels (<0.1, ≥0.25, ≥0.4, ≥0.6, and >0.6, depicted by different colors) for marker pairs in different distance bins maximum 5 Mbp. (*A*) SNP pairs of all 40'854 SNPs with MAF ≥5%; (*B*) consider only SNP pairs with MAF ≥0.15.

It is known that linkage disequilibrium between SNPs with a low minor allele frequency is biased upwards and high-frequency polymorphisms are preferable for accurate estimation of LD (Reich *et al.* 2001). In part this can be explained by statistical properties of the LD statistics (Dunning *et al*. 2000), but may also have an evolutionary interpretation because low frequency SNPs have a higher probability of having arisen recently (Nordborg & Tavaré 2002). Taking this into account we evaluated the decay of LD for the SNPs with MAF greater than 15% to elucidate its usefulness in terms of having SNP pairs in useful LD for genomic association analysis. We observed an increase of about 10% in frequency of SNP pairs representing useful LD for almost all physical distance bins up to 5 Mbp (Figure 5).

LD properties

The decay of LD measures with the physical distance is well documented. LD is expected to be a function of linkage distance in animal populations, at least for tightly linked loci. It is also reported that SNPs of divergent MAFs on average have different LD properties (Pritchard & Przeworski 2001). Figure 5 displays the decay of LD as a function of physical distance and absolute MAF difference (ΔMAF) between SNP pairs. It can be seen that pair-wise r^2 decreases with increasing distance and increasing ΔMAF. It is evident that the dependence of r^2 on distance is stronger than its dependence on ΔMAF. It is also shown that SNP pairs in short physical distance are more affected by ΔMAF. The magnitude of this dependency in the case of SNP pairs far from each other is negligible.

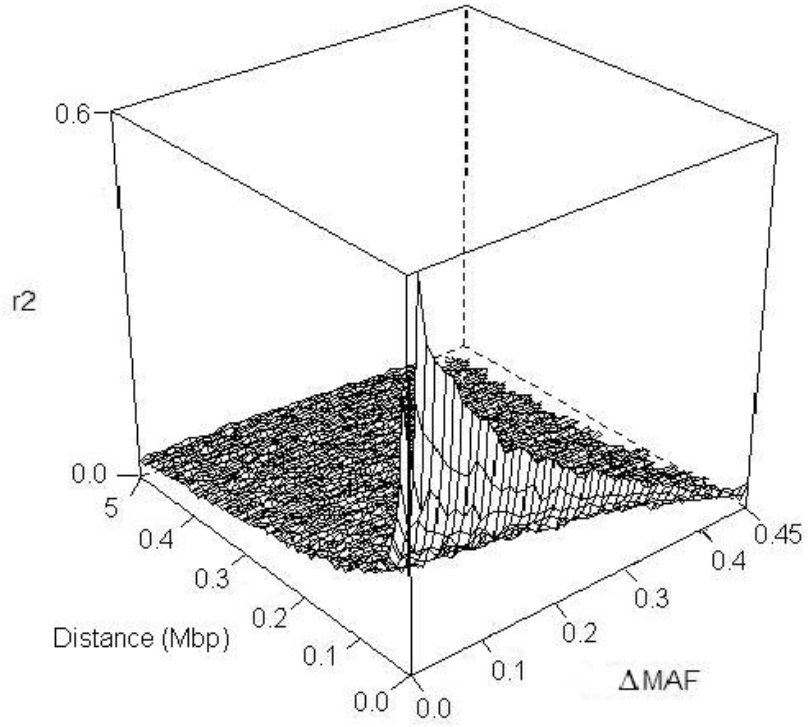

Figure 6. Three dimensional surface plot depicts the decay of LD vs. inter-marker distance and MAF interval

To explore the dependence of LD on allele frequency we calculated the average r^2 statistic within nine bins of physical distance between frequency-matched pairs of SNPs with $\Delta MAF \leq 10\%$ and compared results with the average r^2 between all SNP pairs (Table 2). Mean r^2 were higher between matched SNP pairs than between non-matched ones for all distance bins, with a difference of around 50% in the shortest distances. For the markers within a distance range of 50 to 75 Kb, the proportion of SNP pairs in useful LD increased from 26 to 39 %. We observed a higher extent of LD for frequency-matched vs. non-matched pairs of SNPs. As such, with frequency-matched pairs of SNPs, LD significantly extended up to the range of 1.5 to 3 Mbp.

Table 2. Frequency and mean r^2 estimated for SNP pairs in different distances compared with the frequency matched SNP pairs.

Distance (Mb)	**SNP Pairs (n)**		**Median r^2**		**Mean r^2±SD**		**Frequency r^2≥0.25 (%)**	
	All pairs	**ΔMAF≤0.1**	**All pairs**	**ΔMAF≤0.1**	**All pairs**	**ΔMAF≤0.1**	**All pairs**	**ΔMAF≤0.1**
<0.025	6002	4617	0.16	0.39	0.30± 0.32	0.45± 0.38	39	56
0.025-0.05	20108	12735	0.13	0.25	0.25± 0.28	0.38± 0.35	34	50
0.05-0.075	17938	8340	0.09	0.14	0.20± 0.24	0.29± 0.31	26	39
0.075-0.12	31833	10725	0.07	0.09	0.16± 0.20	0.22± 0.27	20	30
0.12-0.2	55778	12906	0.06	0.06	0.12± 0.16	0.16± 0.22	15	22
0.2-0.5	204584	28572	0.04	0.04	0.09± 0.12	0.11± 0.16	10	15
0.5-1.5	664447	52743	0.03	0.03	0.07± 0.09	0.08± 0.12	6	9
1.5-3	965989	35720	0.02	0.02	0.05± 0.07	0.06± 0.09	3	5
3-5	1249359	17384	0.02	0.02	0.04± 0.06	0.05± 0.07	1	3

In a further step, we plotted the r^2 versus minor allele frequencies of both loci (Figure 7). SNP pairs with highest MAF interval represent lowest r^2 and vice versa. Frequency-matched SNPs with moderate or low MAF values both result in the highest r^2 regions. However, there is a trend demonstrating a slight raise of LD for matched SNPs with moderate MAF comparing the matched SNPs with lower MAF. Therefore, it can be concluded that the frequency matched SNP pairs are less influenced when calculating pairwise r^2 values substantiating lower decay of LD for these loci.

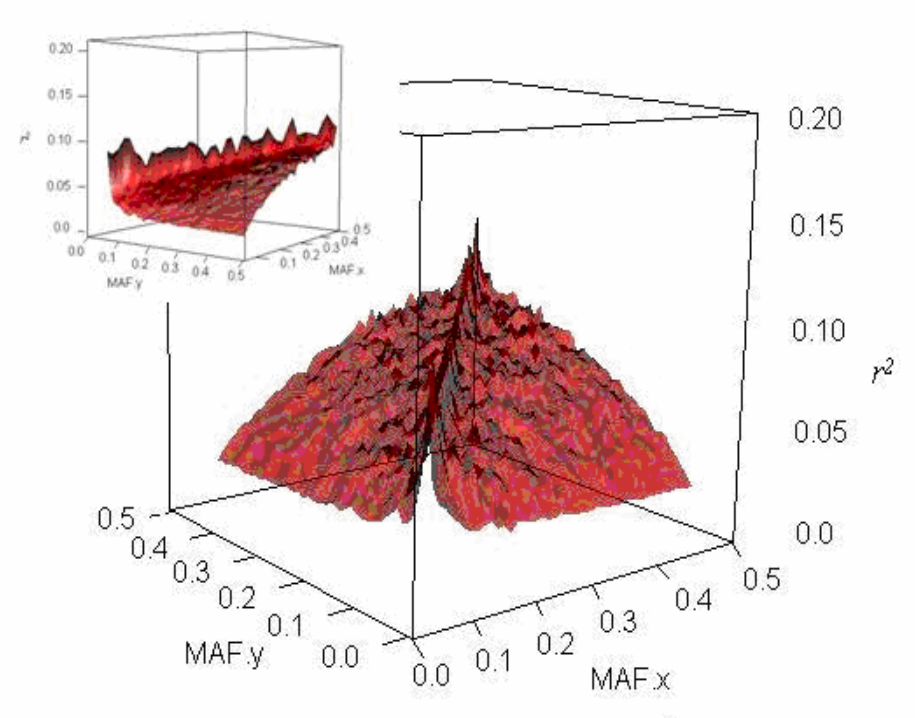

Figure 7. The prospective plot depicts the decay of LD with allele frequencies of SNP pairs. r^2 means were calculated for 45 bins of each 0.01 allele frequency.

Past effective population size

In most studies so far, genetic distance *c* was approximated by using physical distance directly (1Mbp~1cM) for the estimation of N_e (Gautier *et al*. 2007, Hayes *et al*. 2008; Kim & Kirkpatrik 2009). In this study we estimated the recombination rates directly from dense SNP data. Figure 8 displays the decay of LD as a function of recombination rate between pairs of SNPs. Recombination rates are not constant

within chromosomes and vary among regions. Overall, a correlation of -0.22 was observed between r^2 and recombination rate over all adjacent marker intervals analysed. LD values averaged in bins of estimated linkage distance were used to study the changes in effective population size of the population from 2000 generation ago up to the present.

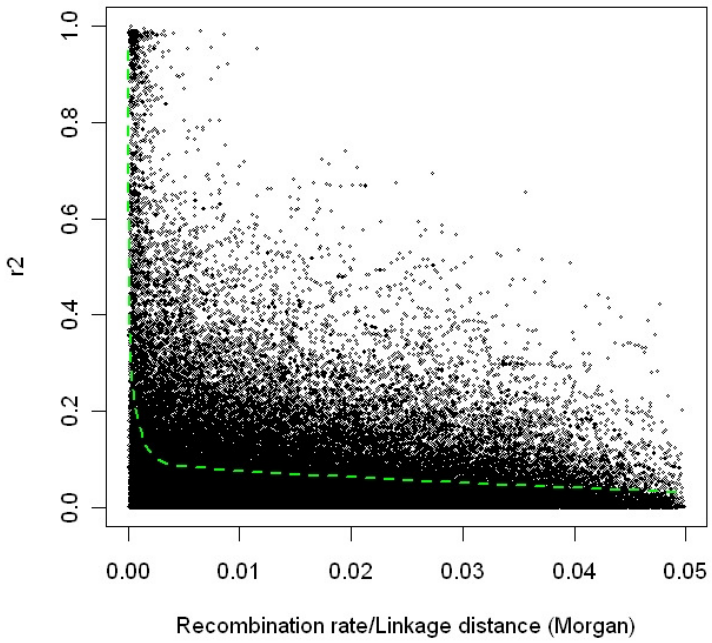

Figure 8. LD between SNP pairs was plotted on the estimates of recombination rate as a measure of linkage distance (M).

We compared the results with the estimates of *Ne* based on available cattle linkage map information (http://www.marc.usda.gov/genome/cattle/cattle.html). Given the known linkage and physical lengths of chromosomes (Table 1) we transformed the physical position to the approximate linkage distance between pairs of SNPs and averaged the estimates over chromosomes. While N_e was inferred as 1113 for 500 generations ago, estimates based on recombination rates show a decline in effective population size to 103 up to ~4 generation ago (Figure 9.A). This is close to the estimate ($N_e \leq 100$) in North American Holstein population based on the analysis of both LD (Kim & Kirkpatrik 2009) and inbreeding rate (Young & Seykora 1996). With the mutation included model it drops to the 56 which is close to the inbreeding

based estimates (N_e < 50) and (N_e = 52) in the Danish (Sorensen *et al*. 2005) and German (Koenig & Simianer 2006) Holstein populations, respectively.

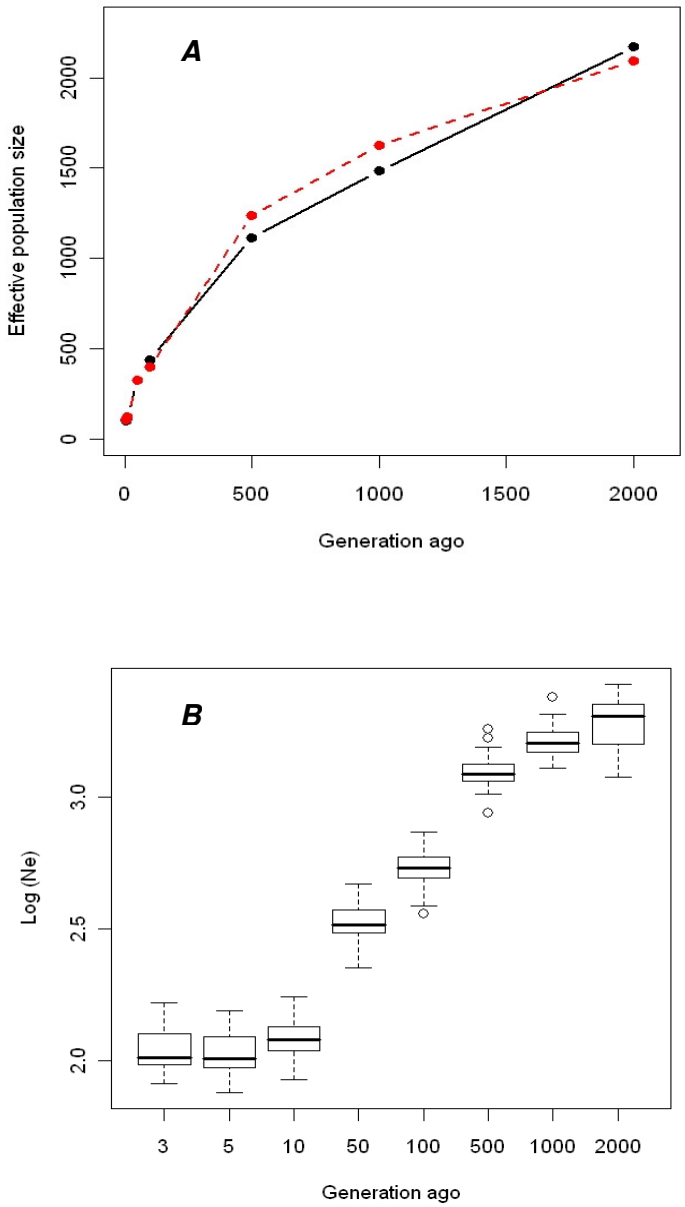

Figure 9. Estimated effective population size over the past generations from linkage disequilibrium data. (*A*) Dashed and solid lines represent *Ne* based on estimates of recombination rates and approximated linkage distances, respectively. (*B*) Boxplot represents the trend of $\log_{10}$ (*Ne*) over the past time and illustrates the divergence of inferred *Ne* among the chromosomes due to the variation in estimated recombination rates.

DISCUSSION

Linkage disequilibrium (LD) maps increase power and precision in association mapping, define optimal marker spacing and identify recombination hot-spots and regions influenced by natural selection. In this report we present an analysis of LD of 40'854 markers densely distributed across the entire bovine genome in a sample of German Holstein cattle. Although the principle of LD is fairly simple (i.e. the non-random segregation of markers in close proximity), the complex interplay between all confounding factors complicates the interpretation of LD results. Since LD depends on the age of the SNP-creating mutations, the demographic population history, genetic drift, the recombination fraction, directional selection, population stratification and other factors, it is highly variable even between close loci (Kruglyak 1999; Ardlie *et al*. 2002; Pritchard & Przeworski 2001). As a result, two markers that are very close together can exhibit a low level of LD, while markers that are more distant can show a higher than expected level of LD.

In this study we used pair-wise r^2 statistic up to 5Mbp across the bovine genome to estimate the extent of LD. The first reports on the extent of LD in cattle genome described a long range of LD (e.g., up to 20 cM) (Farnir *et al*. 2000; Tenesa *et al*. 2003). Further analyses with denser markers confirmed extensive LD, but in general found lower levels (Spelman & Coppieters 2006, Khatkar *et al*. 2006a). Recently, two genome wide studies based on 10K SNP data have revealed that the level of LD is less than previously thought (Sargolzaei *et al*. 2008; Kim & Kirkpatrik 2009). The results of this study demonstrate even less LD for SNP pairs at the distances ≤100 Kb.

It was suggested that LD within genes is higher than LD in inter-genic regions at least for tightly linked markers (Kim & Kirkpatrik 2009), hence the discrepancy observed may be due to a systematic difference of the selected set of SNPs. For the Illumina Bovine SNP50k BeadChip SNPs were mainly selected to evenly cover the entire genome while in other studies the SNPs were targeted to certain candidate regions. Especially for the use in genomic selection and whole genome association mapping without prior positional information the average LD over the entire genome is the quantity of interest, which was evaluated in our study. In general it is difficult to compare the level of LD obtained in different studies because of different sample

sizes, LD measures, marker types, marker densities, and recent and historical population demographics (Pritchard & Przeworski 2001).

The decay of LD in a genome determines the resolution of quantitative trait loci detection in association mapping studies and indicates the required marker density. It was shown that in indirect association studies, the sample size must be increased by roughly $1/r^2$ when compared with the sample size for detecting the causal mutation directly (Kruglyak 1999; Pritchard & Przeworski 2001). Meuwissen *et al*. (2001) simulated the required level of LD (r^2) for genomic selection to achieve an accuracy of 0.85 for genomic breeding values to be 0.2. Ardlie *et al*. (2002) defined high values of LD as $r^2 > 1/3$. In this study we assumed the threshold of useful LD to be 0.25. To achieve this level our results indicate that the SNP spacing should be ~35 Kb in future population wide studies with a whole-genome approach. This implies the use of more than 75,000 SNPs per individual, assuming that all SNPs are informative (with a MAF $\geq$ 0.05). According to the results of this study, the same power can be achieved by implementing a panel of 50'000 SNPs with moderate frequencies (e.g., MAF $\geq$ 0.15) which simultaneously improves the accuracy and magnitude of estimated LD between pairs of SNPs.

In this study, we examined the decay of LD as a function of physical distance. Despite the LD map showing a distinct decrease of LD values over increasing physical distance, the LD also showed extensive variability between genomic regions and chromosomes. This variation was probably due to recombination rates varying between and within chromosomes, heterozygosity and effects of selection.

The impact of allele frequency in analyzing genome-wide LD was also explored in this study. Our results demonstrate that the dependence of LD on the MAF interval of SNP pairs is stronger for SNPs in short distances. These results also reveal that the minimizing the allele frequency difference between SNPs, provides a more sensitive and useful metric for analyzing LD across the bovine genome. Although an entirely frequency-independent measure of LD is not possible (Lewinton 1988), frequency matching between SNP pairs removes one major source of statistical noise when assessing the LD structure.

There are several published studies reporting LD properties based on dense SNP markers in cattle. Khatkar *et al*. (2006a) pioneered exploiting dense SNPs in developing bovine LD maps by characterizing the LD profile for chromosome 6 in the Australian Holstein population. They used 220 SNPs and confirmed an extensive level of LD in Holstein cattle. Gautier *et al*. (2007) studied LD properties in cattle breeds of different origin and observed the haplotype blocks extended up to 700 Kb in some cattle breeds. Khatkar *et al*. (2007) developed a primary genome-wide LD map based on a panel of 9,195 informative SNPs reporting 727 blocks with three or more SNPs, (mean length=69.7 Kb) covering 2.18 % of the genome. In a similar study Marques *et al*. (2008) compared LD properties of chromosome 14 in Holstein and Angus cattle and reported 64 blocks (33 bp to 1126 Kbp). Recently, Kim & Kirkpatrik (2009) reported 119 haplo-blocks (with more than 4 SNPs) with a mean length of 26.2 Kb in a whole genome scan of Holstein cattle. It was suggested that as the number of markers increases more haplotype blocks will be identified. This was confirmed by Villa-Angulo *et al*. (2009) who reported the blocks of smaller size with an overall mean of 10.3 kb across 19 breeds. However, compared to the marker density used in the previous studies, the present study with more SNPs reporting 712 blocks doesn't follow this expectation. Similar to the LD differences observed, this could also be due to the use of a different set of markers, which are evenly distributed across the genome covering both genic and inter-genic regions. Although the number of blocks is not different, we observed a higher block coverage percentage compared to the Australian population.

The average extent of LD in the human genome has been extensively studied: it extends a few Kb up to 50 Kb but is highly variable, depending on the population and threshold used to measure LD. Most recently, Villa-Angulo *et al*. (2009) reported almost similar size of blocks as human for the bovine genome. However, it was limited to the high dense targeted regions of the genome only on three chromosomes. Compared to the results of human studies, average block sizes observed in the present study on the bovine genome are 20-30 times larger than similar haplotype blocks found in the human genome (Hinds *et al*. 2005). It must be noted that the marker density used in this study is about 100 times sparser than the one currently being used for the human genome. Hence, some of the long blocks observed in the present study

may break down to smaller tracts if the SNP density was increased. However, due to the smaller effective population size of cattle compared to the human population (Hayes *et al*. 2003) and relatively high inbreeding frequency, a greater level of LD and larger haplotype blocks are expected.

CONCLUSIONS

We presented a second generation of LD map statistics for the Holstein genome which has four times higher resolution compared to the maps available so far. We found a lower level of LD for SNP pairs at distances ≤100 Kb than previously reported. Assuming that $r^2 > 0.25$ is useful for association studies, the level of LD obtained in this study indicates that a denser SNP map would be beneficial to capture the LD information required for whole-genome fine mapping and genomic selection and to completely assess the pattern of LD across the genome. The results show that frequency matched SNP pairs reduce the dependence of r^2 on allele frequency and provide a useful metric for analyzing LD. The larger block size in Holstein cattle observed in this study indicates substantially greater LD in cattle than in human populations.

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.

References

Anderson, E., and Novembre, J. 2003. Finding haplotype block boundaries by using the minimum-description-length principle. American Journal of Human Genetics 73: 336–354

Ardlie, K. G., Kruglyak, L., and Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3: 299–309.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., Lander, E. S. 2001. High-resolution haplotype structure in the human genome. Nature Genetics 29: 229–232.

Dawson, E., Abecasis, G. R., Bumpstead, S. Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lõhmussaar, E., Zernant, J., Tõnisson, N., Remm, M., Mägi, R., Puurand, T, Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R., Dunham, I. 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature 418: 544–548.

Farnir, F., Coppiettiers, W., Arranz, J. J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D., Georges, M. 2000. Extensive genome-wide linkage disequilibrium in cattle. Genome Research 10: 220–227.

Gabrie, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., Altshuler, D. 2002. The structure of haplotype blocks in the human genome. Science 296: 2225–2229.

Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Foglio, M., Grohs, C., Boland, A., Garnier, J. G., Boichard, D., Lathrop, G. M., Gut, I. G., Eggen, A. 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. Genetics 177: 1059–1070.

Grisart B., Farnir F., Karim L. et al. (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proceedings of the National Academy of Sciences of the United States of America 101, 2398–2403.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research 13: 635–643.

Hayes, B J., Lien, S., Nilsen, H., Olsen, H. G., Berg, P., Maceachern, S., Potter, S., Meuwissen, T. H. E. 2008. The origin of selection signatures on bovine chromosome 6. Animal Genetics 39: 105-111.

Hinds D.A., Stuve L.L., Nilsen G.B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., Cox, D. R. 2005. Whole-genome patterns of common DNA variation in three human populations. Science. 307:1072–1079

Johnson G.C., Esposito L., Barratt B.J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., Todd, J. A. 2001. Haplotype tagging for the identification of common disease genes. Nature Genetics 29: 233–237

Khatkar, M. S., Collins, A., Cavanagh, J.A.L., Hawken, R. J., Hobbs, M., Zenger, K. R., Barris, W., McClintock, A. E., Thomson, P. C., Nicholas, F. W., Raadsma, H. W. 2006a. A first generation metric linkage disequilibrium map of bovine chromosome 6. Genetics 174: 79–85.

Khatkar, M. S., Zenger, K. R., Hobbs, M., Hawken, R. J., Cavanagh, J. A., Barris, W., McClintock, A. E., McClintock, S., Thomson, P. C., Tier, B., Nicholas, F, W., Raadsma, H. W. 2007. A primary assembly of a bovine haplotype block map based on a 15,036 single nucleotide polymorphism panel genotyped in Holstein Friesian cattle. Genetics 176: 763–772.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49–67.

Lewontin, R. C. 1988. On measures of gametic disequilibrium. Genetics 120: 849–852.

Marques, E., De Givry, S., Stothard, P., Murdoch, B., Wang, Z., Womack, J., Moore, S. S. 2007. A high resolution radiation hybrid map of bovine chromosome 14 identifies scaffold rearrangement in the latest bovine assembly. BMC Genomics 8: 254.

Marques, E., Schnabel, R., Stothard, P., Kolbehdari, D., Wang, Z., Taylor, J. F, Moore, S. S. 2008. High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. BMC Genetics 9: 45.

Mckay, S. D., Schnabel, R. D., Murdoch, B. M., Matukumalli, L. K., Aerts, J., Coppieters, W., Crews, D., Dias Neto, E., Gill, C. A., Gao, C., Mannen, H., Stothard, P., Wang, Z., Van Tassell, C. P., Williams, J. L., Taylor, J. F., Moore, S. S. 2007. Whole genome linkage disequilibrium maps in cattle. BMC Genetics 8: 74–85.

Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 1819-1829.

Odani, M., Narita, A., Watanabe, T. Yokouchi, K., Sugimoto, Y., Fujita, T., Oguni, T., Matsumoto, M., Sasaki, Y. 2006. Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. Animal Genetics 37: 1–6.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N. , Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., Cox, D. R. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294: 1719–1723

Pritchard, J. K., and Przeworski, M. 2001. Linkage disequilibrium in humans: models and data. American Journal of Human Genetics 69: 1–14.

Reich, D. E., and Lander, E. S. 2001. On the allelic spectrum of human disease. Trends in Genetics 17: 502–510.

Reich, D. E., Cargill, M., Bolk, S., Ireland. J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., Lander, E. S. 2001. Linkage disequilibrium in the human genome. Nature 411: 199–204.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., and Schaeffer, L. R. 2008. Extent of linkage disequlibrium in Holstein cattle in North America. Journal of Dairy Science 91: 2106–2117.

Scheet, P., and Stephens, M. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics 78: 629-644.

Simianer, H., Szyda, J., Ramon, G., Lien, S. 1997. Evidence for individual and between family variability of the recombination rate. Mammalian Genome 8: 830 - 835.

Koenig, S., and Simianer, H. 2006. Approaches to the management of inbreeding and relationship in the German Holstein dairy cattle population. Livestock Science 103: 40–53

Snelling, W. M., Chiu, R., Schein, J.E., et al. 2007. A physical map of the bovine genome. Genome Biology 8: R165.

Sorensen, A. C., Sorensen, M. K., and Berg, P. 2005. Inbreeding in Danish dairy cattle breeds. Journal of Dairy Science 88: 1865–1872.

Sved, J. A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology 2: 125–141.

Tenesa A., Navarro P., Hayes B.J. et al. (2007) Recent human effective population size estimated from linkage disequilibrium. Genome Research 17, 520–526.

Tenesa, A., Knott ,S. A., Ward, D., Smith, D., Williams, J. L., Visscher, P. M. 2003. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. Journal of Animal Science 81: 617–623.

Terwilliger, J. D., and Weiss, K. M. 1998. Linkage disequilibrium mapping of complex disease: fantasy or reality? Current Opinion in Biotechnology 9: 578–594.

The International HapMap Consortium, 2005. A haplotype map of the human genome. Nautre 437: 1299–1320.

Van Laere, A. S., Nguyen, M., Braunschweig, M. Nezer, C., Collette, C., Moreau, L., Archibald, A. L., Haley, C. S., Buys, N., Tally, M., Andersson, G., Georges, M., Andersson, L. 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425: 832–836.

Villa-Angulo, R., Matukumalli, L. K., Gill, C. A., Choi, J., Van Tassell, C. P., Grefenstette, J. J. 2009. High-resolution haplotype block structure in the cattle genome. BMC Genetics 10: 19.

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., Jin, L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. American Journal of Human Genetics 71: 1227–1334.

Wright, S. 1938. Size of population and breeding structure in relation to evolution. Science (Wash DC) 87: 430–431.

Young, C. W., and Seykora, A. J. 1996. Estimates of inbreeding and relationship among registered Holstein females in the United States. Journal of Dairy Science 79: 502–505.

Zhang, K., Deng, M., Chen, T., Waterman, M. S, Sun, F. 2002. A dynamic programming algorithm for haplotype block partitioning. Proceedings of the National Academy of Sciences of the United States of America 99: 7335–7339

Zhang, K., Sun, F., Waterman, M. S., and Chen, T. 2003. Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. American Journal of Human Genetics 73: 63–73.

3rd CHAPTER

A Genome-Wide Scan for Signatures of Recent Selection in Holstein Cattle

S. Qanbari[*], E. C. G. Pimentel[*], J. Tetens[§], G. Thaller[§], P. Lichtner[†], A.R. Sharifi[*] and H. Simianer[*]

[*] Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

[§] Institute of Animal Breeding and Animal Husbandry, Christian-Albrechts-University, 24098 Kiel, Germany

[†] Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

Animal Genetics 2010, 41: 377–389

ABSTRACT

The data from the newly available 50K SNP chip was used for tagging the genome wide footprints of positive selection in Holstein-Friesian cattle. To this purpose, we employed the recently described Extended Haplotype Homozygosity test, which detects selection by measuring the characteristics of haplotypes within a single population. To formally assess the significance of these results, we compared the combination of frequency and the Relative Extended Haplotype Homozygosity value of each core haplotype with equally frequent haplotypes across the genome. A subset of the putative regions showing highest significance in the genome wide EHH tests was mapped. We annotated genes to identify possible influence they have in beneficial traits using the Gene Ontology (GO) database. A panel of genes, including FABP3, CLPN3, SPERT, HTR2A5, ABCE1, BMP4 and PTGER2 was detected which overlapped with the most extreme P-values. This panel comprises some most interesting candidate genes and QTL representing a broad range of economically important traits such as milk yield and composition as well as reproductive and behavioral traits. We also report high values of LD and a slower decay of haplotype homozygosity for some candidate regions harboring major genes related to dairy quality. The results of this study provide a genome wide map of selection footprints in Holstein genome and can be used to better understand the mechanisms of selection in dairy cattle breeding.

INTRODUCTION

Recently, linkage disequilibrium (LD) has received considerable attention among livestock geneticists primarily to perform genome based selection (see e.g. Meuwissen *et al*. 2001) and to determine the actual genes responsible for variation of economically important traits (Pollinger *et al*. 2005; Daetwyler *et al*. 2008; Prasad *et al*, 2008; Hayes *et al*. 2008, 2009).

The search for genes underlying phenotypic variation can be performed in two different directions; (i) from phenotype to genome which is performed by LD based association mapping and may involve positional cloning of quantitative trait loci (QTL) or targeting particular candidate genes identified based on homology to known genes, and (ii) from genome to phenotype which involves the statistical evaluation of genomic data to identify likely targets of past selection. The latter approaches identify patterns of LD in or between populations which are incompatible with the hypothesis of genetic neutrality, and these patterns are called selection signatures.

Alleles under positive selection pressure encounter a fast increase in allele frequency. For a neutral mutation it will take many generations until the mutated allele has reached a high population frequency through drift. The LD in the vicinity of this locus will be degraded through recombination (Kimura 1983), so that frequent alleles in little LD with the neighbouring loci usually reflect old mutations. A novel mutation under positive selection pressure however will increase rapidly in frequency, so that the surrounding conserved haplotype is long, which is called a 'selective sweep' (e.g., Maynard Smith and Haigh 1974; Nielsen *et al*. 2005).

This is the background of the extended haplotype homozygosity (EHH) statistic suggested by Sabeti *et al*. (2002) for the detection of recent selection. To account for facts like variability of recombination (Simianer *et al*. 1997), Sabeti *et al*. (2002) proposed to use the contrast of the EHH statistic of one core haplotype vs. other haplotypes in the same position. Alternative methods for detecting selective sweeps from DNA sequence data were developed which include Tajima's *D* (Tajima 1989) and Fay and Wu's *H* test (Fay and Wu 2000) for selected mutations, measuring large allele-frequency differences among populations by F_{ST} (e.g., Akey *et al*. 2002) and

the integrated Haplotype Score (iHS; Voight *et al*. 2006), an extension of the EHH statistic (Sabeti *et al*. 2002). Both *D* and *H* tests were designed for full-sequence data and not for genome wide collections of pre-ascertained SNPs that are currently available in some livestock species. Recently, MacEachern et al. (2009) applied Fay and Wu's *H* test to examine the positive selection between Holstein and Angus cattle which represent opposite directional selection. They used a new metric to overcome the problems of ascertainment bias and observed significant deviations of allele frequency in two breeds. The iHS method is more powerful than *D* and *H* tests for selected mutations (Voight *et al*. 2006) but, to be applied properly, it requires the genotype of the selected mutation as well as a known ancestor allele.

Among the various statistics used for recognizing signals of positive selection from polymorphism data, the EHH test is particularly useful (Zang *et al*. 2006; Walsh *et al*. 2006). It detects selection by measuring the characteristics of haplotypes within a single population, is applied for putative core regions, and does not require definition of an ancestor allele genotype. Furthermore, it is designed to work with SNP rather than sequencing data, being less sensitive to ascertainment bias than other approaches (Tang *et al*. 2007).

Holstein-Friesian cattle has been intensively selected during the last centuries, especially so in the last decades after the implementation of progeny-test based breeding programs in the 1960s (Skjervold and Langholz 1964). Consequently, genomic regions controlling traits of economic importance are expected to exhibit signatures of selective breeding. With the availability of large-scale SNP data it has become possible to construct an LD map with higher resolution and to scan the genome for positions that may have been targets of recent positive selection in the Holstein-Friesian population. In this study we report the first results of such a systematic genome scan, in which (i) the region of known functional candidate genes (confirmed QTL) were checked for signatures of recent selection and (ii) positional candidate genes are reported in proximity to the genomic positions showing the most significant indications of selection.

MATERIALS AND METHODS

DNA samples and data preparation

Semen or blood samples were obtained from 810 German Holstein–Friesian cattle including 469 bulls and 341 bull dams. Genomic DNA was extracted applying a modified Miller protocol (Miller *et al.*, 1988) including dithiothreitol treatment for the semen samples. Genotyping was carried out using the Illumina Bovine SNP50 BeadChip (Matukumalli et al. 2009) containing a total of 54001 SNPs with a mean neighbor marker distance of 48.75 Kb. Markers were filtered to exclude loci assigned to unmapped contigs or unpositioned according to the latest reference assembly of the bovine genome Btau 4.0 (1728), monomorphic loci (11) and loci with a minor allele frequency (MAF) < 0.05 (10864).

Reconstruction of haplotypes and LD analysis

The subset of animals used in this study belongs to the total population of Holstein cattle chosen for the genomic selection program in Germany. We overlooked the probable effect of founders and admixture in the population demography due to the considerable number of bulls analyzed and assumed animals to be unrelated. For the analyses fully phased haplotype data were required. After the aforementioned filtering process we reconstructed haplotypes for every chromosome using default parameters in fastPHASE (Scheet and Stephens 2006). Reconstructed haplotypes were inserted into HAPLOVIEW v4.1 (Barrett *et al.* 2005) to estimate LD statistics based on pair-wise r^2 and constructing the blocking pattern in the candidate regions of interest for selection signature analysis.

Application of EHH test

According to natural selection theory, regions under positive selection have frequent alleles, existing on long range LD backgrounds. Accordingly, the “core region” is defined as the region of interest in the genome, presumably characterized by the strong LD among SNPs and involves a set of “core haplotypes”. For identifying core regions Sweep v.1.1 (Sabeti *et al.* 2002) implements the algorithm suggested by Gabriel *et al.* (2002) defining a pair of SNPs to be in strong LD if the upper 95%

confidence bound of D' is between 0.7 and 0.98. The program was set to select core regions with at least 3 SNPs.

To evaluate how LD decays across the genome we performed the EHH test (Sabeti *et al*. 2002). This test is based on the contrast of a core haplotype with a combination of 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 locus (Sabeti *et al*. 2002). The EHH of a tested core haplotype *t* is

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

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.

It has been observed in many experimental organisms (reviewed in Lichten and Goldman 1995; Petes 2001) that various chromosomal regions have higher (or lower) recombination rates than would be expected on the basis of the genome average recombination rate (~ 1 cM/Mb). Simianer *et al*. (1997) demonstrated that this variability is also prevalent in the bovine genome, and that recombination probabilities even differ between families. Regions with high (low) recombination fractions are called hot (cold) spots. Accordingly, the LD would be stronger in recombination cold spots than in recombination hot spots, which raises the possibility that a larger LD statistic may rather be due to low recombination rates in a particular region and not necessarily to recent positive selection. The 'Relative Extended Haplotype Homozygosity' (REHH) statistic proposed by Sabeti *et al*. (2002) corrects EHH for the variability in recombination rates. It is computed by $EHH_t / \overline{EHH}$, with $\overline{EHH}$ defined as the decay of EHH on all other core haplotypes combined and is calculated as:

$$\overline{EHH} = \frac{\sum_{j=1, j\neq t}^{n} \left[\sum_{i=1}^{s} \binom{e_i}{2} \right]}{\sum_{i=1, j\neq t}^{n} \binom{c_i}{2}}$$

where *n* is the number of different core haplotypes.

To determine the empirical significance of REHH values, we ordered haplotypes into 20 bins according to their frequency and compared the REHH for each common haplotype in a candidate region to all equally frequent haplotypes. P-values were obtained by log-transforming the REHH in the bin to achieve normality, and calculating the mean and the standard deviation. As such, core haplotypes with extreme REHH in the distribution considered as significant.

RESULTS

Marker and core haplotype statistics

A total of 41'398 (76.66%) markers passed the filtering criteria and, excluding chromosome X, 40'854 SNPs were included in the final analysis. This subset of markers covers 2544.1 Mbp of the genome (Btau 4.0 assembly) with 61.91 Kbp average adjacent marker spacing. For the SNPs analyzed in this study, the average minor allele frequency (MAF) was 0.28 ± 0.15. Table 1 presents a descriptive summary of genome wide marker and haplotype distribution in the data set. A total of 3741 core regions spanning 472'127.2 Kbp (18.55 %) of the genome were detected.

Table 1. Summary information of genome-wide marker and core region (CR) distribution in Holstein cattle.

Chr	SNP (n)	Chr Length (Mbp)	Mean Distance (Kb)	CR (n)	Mean CR Length (Kb)	[a]Coverage CR length (Kb)	Max CR Length (Kb)	[b]CR length/ Chr Length	[c]CR SNPs (n)	Max CR SNPs (n)	[d]CR SNPs/ SNP
1	2641	161.06	61.0	265	134.2 ±93	35372.8	795.7	0.22	946	11	0.36
2	2149	140.63	65.4	207	125.1 ±89	25901.7	696.3	0.18	733	9	0.34
3	2037	127.91	62.8	209	136.8 ±112	28593.4	908.5	0.22	746	7	0.37
4	1999	124.13	62.1	191	124.5 ±82	23646.5	490.2	0.19	683	9	0.34
5	1718	125.80	73.2	164	142.4 ±93	23214.8	523.5	0.18	579	11	0.34
6	2044	122.54	60.0	217	124.7 ±87	27075.6	517.7	0.22	784	12	0.38
7	1767	112.06	63.4	171	136.9 ±94	23281.9	610.6	0.21	630	10	0.36
8	1849	116.91	63.2	187	136.9 ±87	25598.6	566.6	0.22	675	9	0.37
9	1623	108.07	66.6	127	132.9 ±96	16879.7	588.0	0.16	455	8	0.28
10	1713	106.20	62.0	171	125.6 ±172	21488.6	2212.8	0.20	592	7	0.35
11	1813	110.17	60.8	174	120.1 ±89	20900.2	661.3	0.19	607	9	0.33
12	1320	85.28	64.6	104	126.5 ±102	13164.1	661.7	0.15	359	7	0.27
13	1396	84.34	60.4	138	121.9 ±70	16824.7	446.4	0.20	483	8	0.35
14	1356	81.32	60.0	127	131.8 ±89	16743.4	546.0	0.21	454	10	0.33
15	1365	84.60	62.0	114	122.8 ±87	14008.1	686.4	0.17	392	6	0.29
16	1251	77.82	62.2	127	136.5 ±151	17341.6	1331.7	0.22	473	14	0.38
17	1284	76.45	59.5	110	111.6 ±65	12276.7	378.8	0.16	373	8	0.29
18	1100	66.12	60.1	98	117.8 ±74	11548.2	689.7	0.17	332	6	0.30
19	1108	65.21	58.9	89	125.8 ±81	11202.8	586.8	0.17	309	9	0.28
20	1252	75.71	60.5	121	122.1 ±73	14776.7	483.4	0.20	417	8	0.33
21	1093	69.17	63.3	94	114.0 ±61	10720.8	349.6	0.15	315	6	0.29
22	1009	61.83	61.3	92	112.7 ±52	10373.6	279.5	0.17	316	6	0.31
23	871	53.33	61.2	62	105.8 ±71	6559	502.6	0.12	207	8	0.24
24	1013	64.95	64.1	85	125.0 ±88	10631.1	457.5	0.16	296	8	0.29
25	810	44.02	54.3	71	100.3 ±51	7121.5	273.9	0.16	242	6	0.30
26	849	51.73	60.9	70	124.6 ±96	8724.5	719.3	0.17	239	8	0.28
27	798	48.73	61.1	57	108.8 ±69	6204.6	436.4	0.13	184	5	0.23
28	779	46.00	59.0	41	112.4 ±65	4608.1	314.5	0.10	135	5	0.17
29	847	51.98	61.4	58	126.6 ±77	7343.9	399.7	0.14	195	6	0.23
Total	**40854**	**2544.07**	**61.91**	**3741**	**123.7 ±87**	**447827.2**	**2212.8**	**0.18**	**13151**	**14**	**0.32**

[a]Total length covered by core regions, [b]The proportion of total core regions length on chromosome length, [c]Number of SNPs forming core regions, [d]The proportion of total number of SNPs forming core regions on number of SNPs used

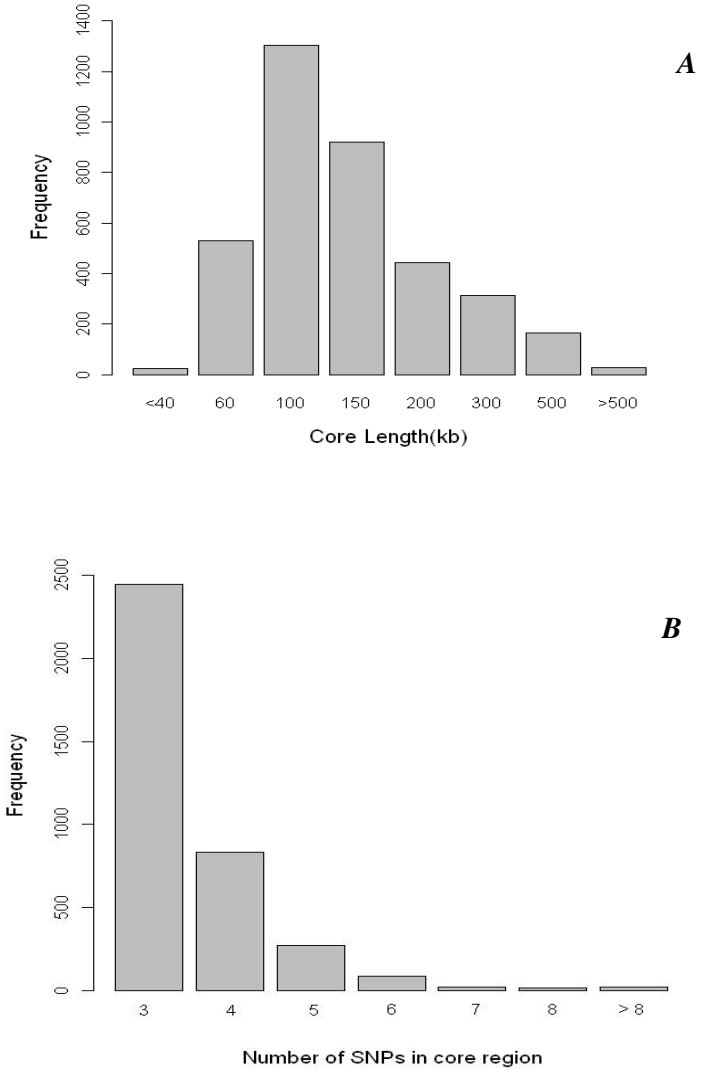

Figure 1. Distribution of the length of core regions (*A*) and the number of SNPs forming the core regions (*B*) in Holstein genome

Mean core region length was estimated as 123.7 Kb ± 87.2 Kb, with a maximum of 2212.8 Kb. There were 265 core regions spanning on 35'372.8 Kb in chromosome 1 and 41 core regions covering 4608.1 Kb of chromosome 28. These were the largest and smallest haplotypic structures in the genome. For each chromosome the proportion of length covered by core regions versus total length as well as the number of SNPs forming core regions versus the total number of SNPs are given in Table 1.

The distribution of the size of core regions is depicted in Figure 1. Overall, 13151 SNPs (31.19 %) participated in forming core regions with a range of 3-14 SNPs per tract.

EHH test in candidate genes

The first step of our analysis focused on ten genes or gene clusters which are well-known to be related to dairy qualities and therefore were assumed to be potentially under recent selection. Table 2 gives the names, details, and test statistics for the chosen panel. For those candidate genes we calculated REHH as a measure of LD surrounding a haplotype of interest. REHH values much greater than 1 indicate increased homozygosity of a haplotype compared with all other core haplotypes in the genome. REHH was calculated at 1 cM distance on both the upstream and downstream sides from a core for all the possible cores present. We chose this length because of the longer extent of LD in cattle compared to human, in which commonly the considered length is around 250 Kb (Sabeti *et al*. 2002, Yu *et al*. 2005).

The results of the EHH test for the Casein cluster shows that P-values for the core haplotype 1 (frequency = 47 percent) exceeded the 99th percentile when REHH was plotted against the haplotype frequency (Figure 2.A1 and 4). P-values calculated for core haplotype 1 in both upstream and downstream direction are both 0.01, which indicates a clear signal of recent selection. In the case of the DGAT1 gene, the second most frequent haplotype (frequency = 30 percent) showed the highest REHH in the core region when plotted up to 1 cM from the candidate region in the downstream direction. As shown in figures 2.A2 and 3, haplotype homozygosity extended up to 1cM only in downstream direction for this core region. This is due to the position of DGAT1 which is located at approximately 400Kb on BTA14 and also the lower LD observed in the upstream direction. This analysis also showed significant P-values for core haplotype 2 of the Leptin Receptor gene (LPR) and core haplotype 1 of the Somatostatin gene (SST) and approached significance for the Growth Hormone Receptor gene (GHR).

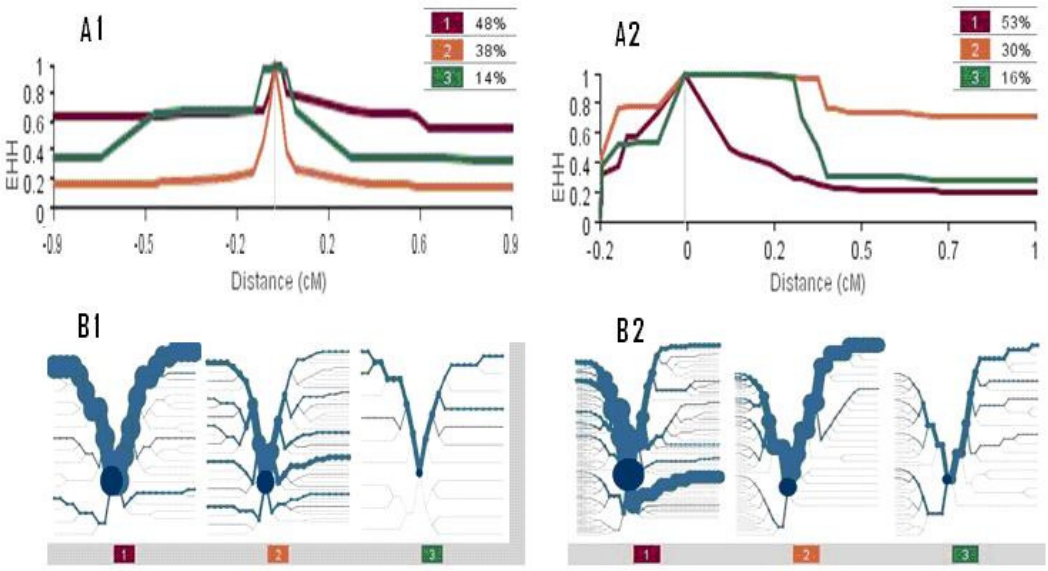

Figure 2. A1 and A2: EHH vs. distance plots for Casein cluster (1) and DGAT1 (2) core regions, showing decay of haplotype homozygosity as a function of distance for three most frequent haplotypes. Legends represent the core haplotype frequencies. B1 and B2: Haplotype bifurcation plots of three core haplotypes for Casein cluster and DGAT1 regions, respectively.

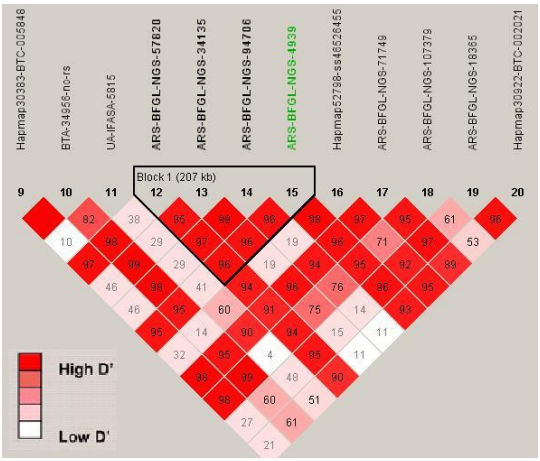

Figure 3. A graphical representation of pair-wise D' for DGAT1 region calculated and visualized using HAPLOVIEW. SNP shown in green represent the closest SNP to DGAT1 gene and is involved in the block structure of length 207 Kb.

Table 2. Summary Statistics of EHH Test for Selection Signature in candidate genes

Candidate Region	Chr	Closest SNP Name & Position (bp)	Core Position	Hap Freq (%)	EHH	REHH[a]	REHH[a] P-Value
DGAT1	14	ARS-BFGL-NGS-4939 443936	236533- 443936	H1: 53 H2: 30	- / 0.19 - / 0.70	- / 0.32 - / 3.47	- / 0.83 [b] - / **0.06**
Casein Cluster	6	Hapmap24184-BTC-070077 88391612	88350095- 88427760	H1: 48 H2: 38	0.63 / 0.54 0.15 / 0.13	3.69 / 3.61 0.24 / 0.24	**0.01 / 0.01** 0.95 / 0.95
GH	19	ARS-BFGL-NGS-73805 49652377	49523705- 49690250	H1: 31 H2: 24	0.21 / 0.17 0.41 / 0.38	0.42 / 0.35 1.35 / 1.60	0.86 / 0.90 0.92 / 0.94
GHR	20	UA-IFASA-8974 33908597	33908597- 34080608	H1: 54 H2: 25	0.72 / 0.86 0.19 / 0.24	1.62 / 1.76 0.25 / 0.28	**0.10 / 0.08** 0.98 / 0.97
SST	1	ARS-BFGL-NGS-38958 81376956	81283582- 81376956	H1: 34 H2: 30	0.76 / 0.84 0.27 / 0.44	3.16 / 2.44 0.49 / 0.72	**0.03 / 0.07** 0.80 / 0.62
IGF-1	5	ARS-BFGL-NGS-116459 71169823	71073539- 71381565	H1: 32 H2: 31	0.35 / 0.24 0.22 / 0.20	1.10 / 0.80 0.52 / 0.47	0.38 / 0.55 0.76 / 0.82
ABCG2	6	BTA-22850-no-rs 37374911	37135014- 37374911	H1: 35 H2: 21	0.19 / 0.18 0.29 / 0.25	0.50 / 0.45 1.10 / 0.95	0.76 / 0.79 0.53 / 0.61
Leptin	4	ARS-BFGL-NGS-34894 95715500	95715500- 95825044	H1: 79 H2: 9	0.14 / 0.15 0.33 / 0.34	0.37 / 0.38 2.25 / 2.22	0.45 / 0.42 0.39 / 0.40
LPR	3	ARS-BFGL-NGS-74572 85569203	85129366- 85176769	H1: 49 H2: 37	0.08 / 0.11 0.33 / 0.43	0.25 / 0.27 3.04 / 3.03	0.92 / 0.90 **0.04 / 0.04**
PIT-1	1	DPI-55 35756434	35713131- 36085241	H1: 32 H2: 20	0.22 / 0.24 0.18 / 0.20	0.65 / 0.62 0.40 / 0.35	0.67 / 0.69 0.92 / 0.94

[a] REHH and P-values are presented for upstream and downstream sides from each core haplotype, respectively

[b] As shown in figures 2.A2 and 3 haplotype homozygosities were extended up to 1cM only in the downstream direction for this core region

Whole genome screen for selection signatures

For all 3741 core regions, a total of 28'323 EHH tests with an average of 7.57 tests per core region were calculated. To find outlying core haplotypes we calculated REHH at 1 cM distance on both the upstream and downstream sides. Figure 4 shows the distribution of REHH values vs. haplotype frequencies. Corresponding P-values are indicated by the use of different symbols. Based on the selection signature theory, a beneficial allele undergoing positive selection is fixed or is going to be fixed in the population. Hence core haplotypes harboring these alleles would have a high frequency. Taking this into consideration, we skipped core haplotypes with frequency <25% and plotted the $-\log_{10}$ of the P-values associated with REHH against the chromosomal position to visualize the chromosomal distribution of outlaying core haplotypes (Figure 5). It is evident that these signals are non-uniformly distributed across chromosomes and chromosome segments, with a substantial overrepresentation on parts of chromosomes 10, 2, and 13.

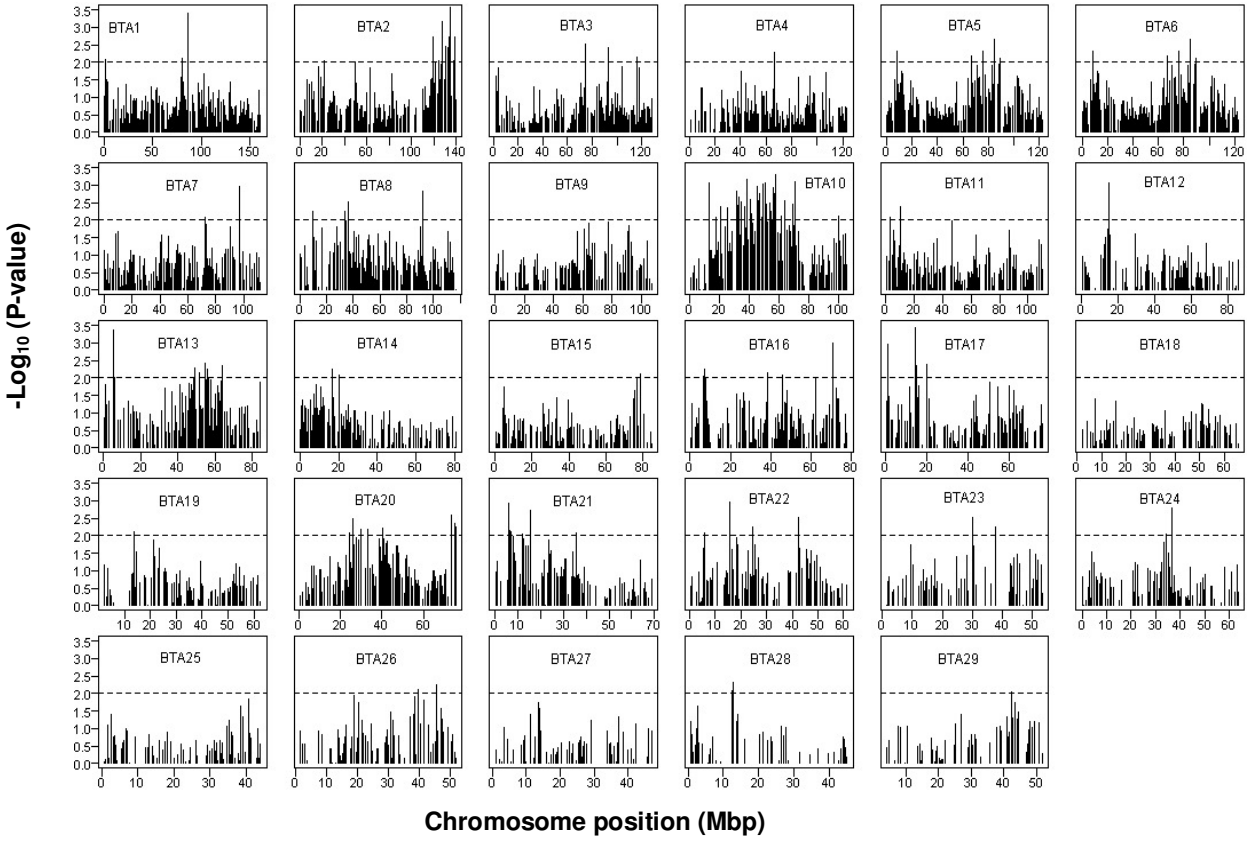

FIGURE 5. Genome wide map of P-values for core haplotypes with frequency ≥0.25. Dashed lines display the threshold level of 0.01.

Table 3 presents the genome wide statistics of the selection signature test including the number of tests and outlying core haplotypes for each chromosome. Of 12'435 tests on core haplotypes with frequency ≥ 0.25, in total 161 tests displayed outlying peaks on a threshold level of 0.01. Bovine chromosomes 6 and 14 which harbor known genes and QTL for several economically important traits (Stone *et al*. 1999; Mosig *et al*. 2001; MacNeil and Grosz 2002; Casas *et al*. 2003; Li *et al*. 2004; Ashwell *et al*. 2005; Nkrumah *et al*. 2007) showed 8 and 2 outliers, respectively. The number of peaks rises to 41 and 14, respectively, when the threshold is set to P<0.05.

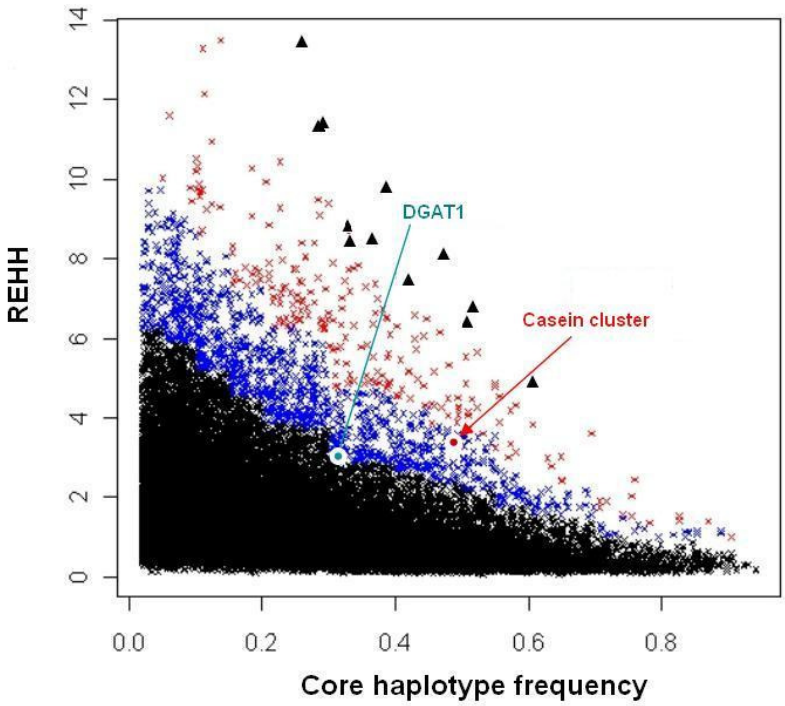

Figure 4. Distribution of REHH vs. haplotype frequencies in the Holstein genome. REHH was calculated at 1 cM distances in the upstream and downstream direction for all possible core haplotypes. Core haplotypes with P-values lower than 0.05 and 0.01 are presented in blue and red, respectively. The panel of 12 core haplotypes displaying the lowest P-values ($P < 0.001$) is represented by triangles. Values representing DGAT1 and the Casein cluster are indicated.

Table 3. A summary statistics of whole genome EHH tests

Chr	Tests on CH[a] (n)	Pvalue<0.05 (n)	Pvalue<0.01 (n)
1	865	17	4
2	678	58	17
3	686	15	3
4	626	11	1
5	531	37	8
6	695	41	8
7	540	13	2
8	605	41	6
9	430	14	0
10	552	123	45
11	590	15	2
12	335	9	2
13	476	51	8
14	439	14	2
15	390	8	2
16	423	29	6
17	365	25	6
18	340	3	0
19	311	5	1
20	400	51	11
21	336	31	7
22	300	27	7
23	213	18	4
24	290	8	4
25	248	5	0
26	234	14	2
27	195	5	0
28	146	5	2
29	196	9	1
Total	**12435**	**702**	**161**

[a]The number of tests on core haplotypes (both sides) with frequency ≥ 0.25

We examined the conformity of the distribution of Tukey's outliers with outlying core haplotypes defined on the threshold level of 0.01. Figure 6 displays box plots of the distribution of -log10 (p-values) within each bin of core haplotype frequency. In order to fit the distribution of -log10 (p-values), the threshold defining outliers (1%) displayed in the box plots were set to Q1-3*IQR and Q3+3*IQR, where IQR is the interquartile range and Q1 and Q3 are the first and third quartiles respectively. It is evident that the extreme outliers appear in the moderate bins of haplotype frequencies.

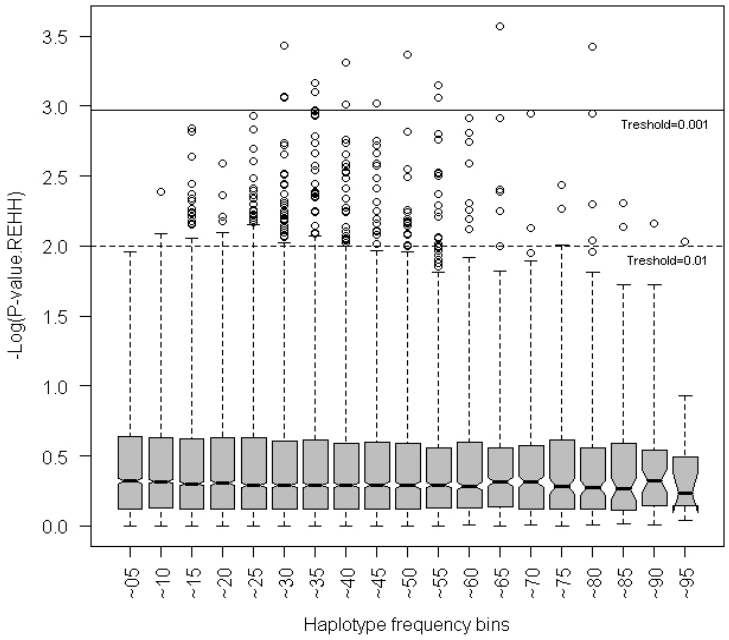

Figure 6. Box plot of the distribution of P-values in core haplotype frequency bins of 5% difference. Core haplotypes with P-values lower than 0.01 and 0.001 are separated with dashed and continuous threshold lines, respectively.

Mapping positively selected regions to genome annotations

A summary of statistics for 12 positively selected core regions presenting the lowest P-values of REHH test is shown in Table 4. Corresponding genes were identified using the map viewer option and aligning the core positions to the fourth draft of the bovine genome sequence assembly (Btau 4.0). We extended core regions in both directions up to 1cM as the length of the core domains. A subset of genes and EST regions were annotated for each core region. We screened this list for the most interesting candidate genes in each core region presenting top peaks. Interestingly, some regions overlapped with genes previously suggested being under selection. For example on chromosome 2q45, a core haplotype harboring the Fatty Acid-Binding Protein 3 muscle and heart (FABP3) gene showed a strong signature of selection (P-value <0.0006). FABP3 is involved in gene networks driving bovine milk fat

synthesis during the lactation cycle and plays a key role in the regulation of the channeling of fatty acids toward copious milk fat synthesis in bovine mammary (Bionaz and Loor 2008). Another strong signature of selection on chromosome 12 matches the SPERT (spermatid-associated protein) and 5-hydroxytryptamine (serotonin) receptor 2A (HTR2A) genes.

Table 4. Summary Statistics for 12 core haplotypes showing the lowest P-values of REHH test

Chr	Position	Core Length(Kb)	Hap Freq	EHH	REHH	REHH P-Value	Gene/EST (n)	Candidate Gene	Function	Reports in Bovine
2	127125963-127172772	46.81	0.33	0.61	8.66	0.00068	12	FABP3	Regulating of the channeling of fatty acids toward copious milk fat synthesis in bovine mammary	Bionaz and Loor (2008)
2	134666758-134761842	95.08	0.61	0.47	5.07	0.00027	16	HMGCL 3	Hydroxymethylglutaryl-coa lyase activity	Jiang *et al*. (2008)
								E2F2	Activating Transcription Factor-2 in skeletal growth control	Phyllis & Luvalle (2003)
10	13146429-13225603	79.17	0.28	0.97	11.21	0.00087	31	PTGER2	Prostaglandin E receptor activity	Arosh *et al*. (2003)
								LCTL	Lactase like protein	
10	38264640-38625718	361.08	0.51	0.98	6.74	0.00071	21	CAPN3	Calcium-dependent cysteine-type endopeptidase activity and protein binding	Barendse *et al*. (2008)
10	48942782-49031850	89.07	0.41	0.57	7.58	0.00095	12	RORA	Steroid hormone receptor activity, transcription factor activity and zinc ion binding	
10	51073231-51138335	65.1	0.51	0.80	6.45	0.00087	17	GCNT3	Transferase activity, transferring glycosyl groups	
								LIPC	LPL is a key enzyme in catabolism of plasma lipoprotein (TGs)	
10	57638141-57773467	135.33	0.39	0.96	9.76	0.00049	13	CYP19	Conversion of androgen to estrogen	
10	70455224-70552188	96.96	0.33	0.91	8.40	0.00079	21	BMP4	Development and functioning of follicles and oocyte maturation	Fatehi *et al*. (2005)
12	14556717-14658840	102.12	0.29	0.79	11.26	0.00085	11	SPERT	Spermatid-associated protein	Reist *et al*. (2003)
								5HTR2A	G-protein coupled receptor activity	
13	5082478-5148264	65.79	0.48	0.99	8.03	0.00043	1	BTBD3	Proteins with a bric--brac, tramtrack, broad-complex/Poxvirus zinc fingers domain plays role in DNA binding, regulation of gene transcription and organization of macromolecular structures	
16	70812261-71003946	191.69	0.36	0.79	8.42	0.00097	17	HSD11B1	KEGG pathway: Androgen and estrogen metabolism, C21-Steroid hormone metabolism	
								LPGAT1	Acyltransferase	
17	13973226-14208603	235.38	0.26	0.81	13.41	0.00037	7	ABCE1	Transmembrane proteins	

HTR2A 5 acts in serotoninergic pathways which are involved in economically important bovine gastrointestinal (GI) motility disorders such as displaced abomasum and cecal dilatation/dislocation (Reist *et al*. 2003). It was also suggested that variants of this gene are related with behavioral disorders in human (Khait *et al*. 2005) and aggressiveness in canine (Peremans *et al*. 2003). This point looks more interesting when we compare the temperament behavior of modern cattle breeds which have been bred during the last decades with native cattle breeds worldwide.

Table 5. A list of candidate genes located nearby the peak regions on chromosome 10

Gene	**Position (bp)**
PYGL phosphorylase, glycogen, liver	43,866,028 - 43,990,507
L2HGDH L-2-hydroxyglutarate dehydrogenase	43,275,277 - 43,326,310
TRIP4 thyroid hormone receptor interactor 4	45,955,859 - 46,007,749
LACTB lactamase, beta	47,246,369 - 47,264,257
CA12 carbonic anhydrase XII	46,990,214 - 47,052,877
BMP4 bone morphogenetic protein 4	67,159,768 - 68,659,191
CGRRF1 cell growth regulator with ring finger domain 1	68,781,739 - 68,809,111
CDKN3 cyclin-dependent kinase inhibitor 3	68,622,652 - 68,639,430
GCH1 GTP cyclohydrolase 1	69,125,665 - 69,182,772
SOCS4 suppressor of cytokine signaling 4	69,288,425 - 69,302,353
NAT12 N-acetyltransferase 12 (GCN5-related, putative	71,796,688 - 71,818,386
TCF12 transcription factor 12	54,015,561 - 54,205,361
GRINL1A glutamate receptor, ionotropic, N-methyl D-aspartate-like 1A	53,121,124 - 53,128,327
LIPC lipase, hepatic	52,220,965 - 52,415,726

We found an unexpected high number of outliers on chromosome 10. One of the core regions representing strong signal (P-value < 0.0007) harbors the Calpain3 (CPN3) gene (Barendse *et al*. 2008). Another strong peak (P < 0.0008) on chromosome 10 is associated with the bone morphogenetic protein4 (BMP4) gene, which is involved in the bone morphogenetic protein (BMP)-signaling system, present in bovine antral follicles, and plays a role in development and functioning of follicles (Fatehi *et al*. 2005). The other signal (P < 0.0008) observed on chromosome 10 is in vicinity of the prostaglandin E receptor 2-subtype EP2 (PTGER2) gene. EP2 is the major cAMP-generating PGE (2) receptor expressed and regulated in the bovine uterus during the estrous cycle and early pregnancy (Arosh *et al*. 2003). It should be noticed that EHH for 5 out of 6 core regions on chromosome 10 were estimated as >0.95 when plotted

up to 1 cM. EHH extended up to at least 2 cM in both directions and spanned a larger number of candidate genes which could have been the targets of recent artificial selection in these regions (Table 5).

We also explored three QTL databases available online (http://genomes.sapac.edu.au/bovineqtl/index.html, http://www.animalgenome.org/QTLdb/cattle.html, http://www.-vetsci.usyd.edu.au/reprogen/QTLMap/) to identify the overlaps of the outlying core regions with published QTL in dairy and beef cattle. Table 6 lists the traits, approximate position and reported population of the overlapping QTL for each core region. In the majority of cases we found an overlap between the core regions presenting top P-values and those that had previously been identified to be harboring beef or dairy QTL. An interesting feature of this comparison is that the majority of these QTL have been reported in Holstein populations.

Table 6. Reported QTL nearby the core regions with lowest P-values

Chr	Start – End (bp)	Trait	Position (cM)	Population	Reported Statistic		
					F-ratio	P.value	
2	127125963-127172772	Marbling Score	126	Multi-breed beef		Significant	MacNeil *et al.* (2002)
		Fat Yield	115-130	German Holstein		0.01	Harder *et al.* (2006)
		Birth Weight	115-128	Multi-breed beef		0.014	Grosz & MacNeil (2001)
2	134666758-134761842						
10	13146429-13225603	Milk Yield	11-20	German Holstein	2.3		Thomsen *et al*. (2001)
		Protein Percent	19	Israel Holstein		0.01	Mosig *et al*. (2001)
		Carcass Weight	0-30	Multi-breed beef	12.0		Casas *et al*. (2003)
		Marbling Score	0-28	Multi-breed beef	11.0		Casas *et al*. (2003)
10	38264640-38625718						
10	[48942782-51138335]	Non return rate	48	German Holstein		Significant	Kuhn *et al*. (2003)
		SCC	49	German Holstein		0.02	Kuhn *et al*. (2003)
		Body Depth	46	USA Holstein	3.06		Ashwell *et al*. (2005)
10	57638141-57773467	Protein Percent	55	Israel Holstein		0.02	Mosig *et al*. (2001)
10	70455224-70552188	Carcass Trait	60-79	Multi-breed beef		Suggestive	MacNeil *et al.* (2002)
		Udder Depth	68.1	Holstein		0.02	Biochard *et al.* (2003)
		Calving ease	73.1	US Holstein	21.78		Schnabel *et al.* (2005)
		Milk Yield	69.0	Canadian Holstein		0.05	Plante *et al*. (2001)
		SCC	73.9	US Holstein	13.18		Schnabel *et al.* (2005)
		Teat Placement	68.1	Holstein		0.02	Biochard *et al.* (2003)
		Udder Cleft	68.1	Holstein		0.02	Biochard *et al.* (2003)
		Teat Length	68.1	US Holstein	14.03		Schnabel *et al.* (2005)
		Protein Percent	73	Israel Holstein		0.03	Mosig *et al*. (2001)
12	14556717-14658840	Milk Yield	21	Finnish Ayrshire		0.006	Viitala *et al.* (2003)
		Protein Yield	21	Finnish Ayrshire		0.02	Viitala *et al.* (2003)
		Protein Percent	21	Israel Holstein		0.01	Mosig *et al*. (2001)
13	5082478-5148264	Dairy form	0-9	USA Holstein	2.82		Ashwell *et al*. (2005)
16	70812261-71003946	Hot Carcass Weight	54-77	Wagyu x Limousin		Significant	Alexander *et al*. (2007)
		Udder Depth	61-72	USA Holstein	3.28		Ashwell *et al*. (2005)
17	13973226-14208603	Rump Angle	0-30	Holstein, Normande, Montbeliarde		0.005	Biochard *et al.* (2003)

DISCUSSION

Holstein-Friesian cattle, the world's highest producing dairy animal, are believed to have been artificially selected since a few thousand years ago (Bradley & Cunningham 1998). Therefore, identifying the regions that have been subjected to selective breeding would facilitate the identification of genes related to traits of interest or biological relevance. A genome wide map of selection events will also help to better understand the mechanisms of selection in artificially selected populations. Unfortunately, robust inferences of recent positive selection from genomic data are difficult because of the confounding effects of population demographic history. For example, both positive selection and an increase in population size may lead to an excess of low-frequency alleles in a population relative to what is expected under a standard neutral model, i.e., a constant-size, randomly mating population in mutation-drift equilibrium (Akey *et al*. 2004). Therefore, rejection of the standard neutral model usually cannot be interpreted as unambiguous evidence for recent selection. In contrast to human populations, the strength of artificial selection is supposed to be much more pronounced than natural selection on fitness related traits. Therefore, it is reasonable to hypothesize that targets of artificial selection will be easier to find in domesticated livestock populations than in non-domesticated populations (Biswas and Akey 2006).

In this study we employed the long-range haplotype test, which detects selection by measuring the characteristics of haplotypes within a single population. We mapped a subset of the putative regions, identified by extreme P-values across the genome (Figure. 5) and used this information to annotate genes which may be under selection pressure. The identified genes reflect a series of pathways, like steroid metabolism, regulation of transcripts, transportation and other functional categories. For most genes associated with signals of selection a biological link to traits such as milk yield and composition, reproduction and behavior, which are known to be under selection, can be hypothesized. However these results need to be confirmed by further studies.

Applying the EHH test on our data revealed 161 regions exhibiting footprints of recent positive selection at a threshold level of 0.01. We observed that other

haplotypes present in this region display a shorter extent of homozygosity, indicating abundant historical recombination. Therefore, the long stretch of homozygosity observed in this region presumably is not simply due to a low local recombination rate but likely reflects the combination of strong and recent selective pressure, pushing beneficial mutations rapidly towards high frequency with long conserved haplotypes surrounding them. The test on whole genome of Holstein genome revealed a signal on positions 62.27 Mbp chromosome 2 which is close to the one reported by Barendese et al. (2009) and the Bovine HapMap Consortium (2009) to be related with feed efficiency traits in a set of cattle breeds. There are also a cluster of strong signals on the chromosome 6 (position 88.35 Mbp) and chromosome 25 (position 30.24 Mbp) confirming the signatures related to multiple beef traits (Barendese et al. 2009) and ZNF187 gene (HapMap Consortium 2009), respectively.

We examined the validity of EHH analysis by testing some candidate major genes in our data set. The results revealed a longer than expected range of LD in core regions harboring the Casein cluster, DGAT1, GHR, STS and LPR genes which are supposed to affect milk yield and composition traits in Holstein cattle. This observation is in agreement with results 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 LD consistency observed in this study is also in coincidence with the reports of Grisart *et al.* (2003) and Marques *et al.* (2008) who used EHH plots to evaluate extended long range LD around DGAT1. The long range of LD observed for the second frequent core haplotype in the DGAT1 region in previous reports is confirmed in our study. A substantial proportion of the analyzed candidate genes showed P-values ≤ 0.10 which supports the validity of our approach. However, some of the candidate genes such as ABCG2 did not meet our definition of positively selected genes but may have nonetheless been targets of selection (Hayes *et al.* 2008, 2009; The Bovine HapMap Consortioum 2009). Different hypothesis can be proposed to explain the incongruities. The disparity shown may have arisen because of a possible higher initial frequency of beneficial alleles (Innan and Kim 2004). Such an allele might e.g. be imported into a breed through crosses with other breeds. In such a case selection may have started from a moderate initial frequency, and beneficial alleles may be included in diverse haplotypes. The density of the markers is also

critical for the power of such studies and could be a source of discrepancy. Comparing the average marker spacing with mean core length and number of SNPs forming cores would results that the core regions were appeared where the marker density is greater than the average. This would imply that a new SNP chip with sufficient genome-wide marker density is required to efficiently identify core haplotypes. Furthermore, with a denser marker map a larger proportion than the 18.5 % of the mapped genome would be assigned to core regions. Although the effects of marker density on the distribution of REHH values is not clear (Zhang *et al*. 2006), a denser map would allow a more reliable and comprehensive screening of the genome for signatures of selection. The incongruities can also result by the complex genomic interactions, or lack of power given the sample size available for this study.

CONCLUSION

Our results provide a genome wide map of selection footprints in Holstein genome. Many of the regions showing top P-values seems to play important roles in economically interested traits in dairy cattle and can now serve as starting points for formulating biological hypotheses. We also reported high values of LD and a slower decay of haplotype homozygosity for some candidate regions harboring major genes related to dairy quality. Other candidate regions do not show such signals, which may be due either to statistical or to biological reasons. Additional studies are needed to confirm and refine our results. This may comprise within population studies with larger sample size and increased SNP density, comparative studies with geographically separated populations with identical or diverse breeding goals, and a detailed functional characterization of the candidate regions identified to be under recent directed selection.

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.

References

Alexander, L. J., Snelling, W. M., and Macneil, M. D. 2007. Quantitative trait loci with additive effects on growth and carcass traits in a Wagyu&Limousin F2 population. Animal Genetics 38: 413-416.

Akey, J. M., Zhang, G., Zhang. K., Jin, L., Shriver. M. D. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12: 1805-1814.

Akey, J., M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D. Nickerson, D. A., Kruglyak, L. 2004. Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes. PLoS Biology 2: e286.

Arosh, J. A., Banu, S. K., Chapdelaine, P., Emond, V., Kim, J. J., MacLaren, L. A., Fortier, M. A. 2003. Molecular Cloning and Characterization of Bovine Prostaglandin E2 Receptors EP2 and EP4: Expression and Regulation in Endometrium and Myometrium during the Estrous Cycle and Early Pregnancy. Endocrinology 144: 3076-3091.

Ashwell, M. S., Heyen, D. W., Weller, J. I., Ron, M., Sonstegard, T. S., Van Tassell, C. P., and Lewin H.A. 2005. Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle. Journal of Dairy Science 88: 4111–9.

Barendse, W., Harrison, B., Bunch, R. J., and Thomas, M. B. 2008. Variation at the Calpain 3 gene is associated with meat tenderness in zebu and composite breeds of cattle. BMC Genetics 9: 41.

Barendse, W., Harrison, B., Bunch, R. J., Thomas, M. B., and Turner, L. B. 2009. Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. BMC Genomics 10:178.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.

Boichard, D., Grohs, C., Bourgeois, F., Cerqueira, F., Faugeras, R. Neau, A., Rupp, R., Amigues, Y., Boscher, M. Y., Levéziel, H. 2003. Detection of genes influencing economic traits in three French dairy cattle breeds. Genetics Selection Evolution 35: 77-101.

Bionaz, M., and Loor, J. 2008. Gene networks driving bovine milk fat synthesis during the lactation cycle. BMC Genomics 9: 366.

Biswas, S. and Akey J. M. 2006. Genomic insights into positive selection. Trends in Genetics 22: 437-446.

Bradley, D. G., and Cunningham, E. P. 1998. Genetic aspects of domestication. In: The Genetics of Cattle (Ed. by R. Fries & A. Ruvinski), pp. 15–32. CAB International, Oxon, UK.

Casas, E., Shackelford, S. D., Keele, J. W., Koohmaraie, M., Smith, T. P., and Stone, R.T. 2003. Detection of quantitative trait loci for growth and carcass composition in cattle. Journal of Animal Science 81: 2976–83.

Cohen-Zinder, M., Seroussi, E., Larkin, D. M., Loor, J. J., Everts-van der Wind, A., Heon-Lee, J., Drackley, J. K., Band, M. R., Hernandez, A. G., Shani, M., Lewin, H. A., Weller, J. I., Ron, M. 2005 Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Research 15: 936–44.

Daetwyler, H. D., Schenkel, F. S., Sargolzaei, M. and Robinson, J. A. B. 2008. A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. Journal of Dairy Science 91: 3225-3236.

Fatehi A.N., Hurk R.V.D., Colenbrander B., Daemen A., Van Tol H. Monteiro, R. M., Roelen , B. A., Bevers, M. M. 2005. Expression of bone morphogenetic protein2 (BMP2), BMP4 and BMP receptors in the bovine ovary but absence of effects of BMP2 and BMP4 during IVM on bovine oocyte nuclear maturation and subsequent embryo development. Theriogenology 63: 872-889.

Fay, J. C., and Wu, C. I. 2000. Hitchhiking under positive Darwinian selection. Genetics 155: 1405–13.

Gabrie,l S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., Altshuler, D. 2002. The structure of haplotype blocks in the human genome. Science 296: 2225–2229.

Grisart B., Farnir F., Karim L. et al. (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proceedings of the National Academy of Sciences of the United States of America 101, 2398–2403.

Harder, B., Bennewitz, J., Reinsch, N., Thaller, G., Thomsen, H., Kühn, C., Schwerin, M., Erhardt, G., Förster, M., Reinhardt, F., Kalm, E. 2006. Mapping of quantitative trait loci for lactation persistency traits in German Holstein dairy cattle. Journal of Animal Breeding & Genetics 123: 89-96.

Hayes, B. J., Lien, S., Nilsen, H., Olsen, H.G., Berg, P. et al. 2008. The origin of selection signatures on bovine chromosome 6. Animal Genetics 39: 105-111.

Hayes, B. J., Chamberlain, A. J., Maceachern, S., Savin, K., Mcpartlan, H. et al. 2009. A genome map of divergent artificial selection between Bos Taurus dairy cattle and Bos taurus beef cattle. Animal Genetics 40: 176-84

Innan H. & Kim Y. (2004) Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America 101, 10667-10672.

Jiang, Z., Michal, J. J., Kunej, T., and Macneil., M. D. 2008. Associations Of Nucleus Encoded Mitochondrial Genes With Carcass, Meat And Eating Quality In Beef Cattle. Proceeding of Plant & Animal Genomes XVI Conference. January 12-16, San Diego, CA.

Khait, V. D., Huang, Y., Zalsman, G., Oquendo, M. A., Brent, D. A. Harkavy-Friedman, J. M, Mann, J. J. 2004. Association of Serotonin 5-HT2A Receptor Binding and the T102C Polymorphism in Depressed and Healthy Caucasian Subjects. Neuropsychopharmacology 30: 166-172.

Kühn, C.H., Bennewitz, J., Reinsch, N., Xu, N., Thomsen, H., Looft, C., Brockmann, G. A., Schwerin, M., Weimann, C., Hiendleder, S., Erhardt, G., Medjugorac, I., Förster, M., Brenig, B., Reinhardt, F., Reents, R., Russ, I., Averdunk, G., Blümel, J., Kalm, E. 2003 Quantitative trait loci mapping of functional traits in the German Holstein cattle population. Journal of Dairy Science 86: 360–68.

Li, C., Basarab, J., Snelling, W. M., Benkel, B., Kneeland, J., Murdoch, B., Hansen, C. and Moore, S. S. 2004. Identification and fine mapping of quantitative trait loci for backfat on bovine chromosomes 2, 5, 6, 19, 21 and 23 in a commercial line of Bos taurus. Journal of Animal Science 82: 967–972.

Lichten, M., and Goldman, A. S. H. 1995. Meiotic Recombination Hotspots. Annual Review of Genetics 29: 423-444.

MacEachern, S., Hayes, B., McEwan, J., and Goddard, M. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics 10: 181.

MacNeil, M. D., and Grosz, M. D. 2002. Genome-wide scans for QTL affecting carcass traits in Hereford × composite double backcross populations. Journal of Animal Science 80: 2316–24.

Marques, E., Schnabel, R., Stothard, P., Kolbehdari, D, Wang, Z., Taylor, J. F., Moore, S. S. 2008. High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. BMC Genetics 9: 45.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P., Sonstegard, T. S., Van

Tassell, C. P. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4: e5350

Maynard-Smith, J., and Haigh, J. 1974. The hitch-hiking effect from a favourable gene. Genetical Research 23: 23–35.

Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 1819-1829.

Miller, S. A., Dykes, D. D., Polesky, H. F. 2001. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Research 16: 1215.

Montgomery, G., Galloway, S., Davis, G. H., and Mcnatty, K. P. 2001 Genes controlling ovulation rate in sheep. Reproduction 121: 843-852.

Mosig, M. O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., and Friedmann A. 2001. A whole-genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics 157: 1683–1698.

Nielsen, R. 2005. Molecular Signatures of Natural Selection. Annual Review Genetics 39: 197-218.

Nkrumah, J. D., Sherman, E. L., Li, C., Marques, E., Crews, D. H., Bartusiak, R., Murdoch, B., Wang, Z., Basarab, J. A., and Moore, S. S. 2007. Primary genome scan to identify putative QTL for feedlot growth rate, feed intake and feed efficiency of beef cattle. Journal of Animal Science 85: 3170–3181.

Peremans, K., Audenaert, K., Coopman, F., Blanckaert, P., Jacobs, F. Otte, A, Verschooten, F., van Bree, H., van Heeringen, K., Mertens, J., Slegers, G., Dierckx, R. 2003. Estimates of regional cerebral blood flow and 5-HT2A receptor density in impulsive, aggressive dogs with 99mTc-ECD and 123I-5-I-R91150. European Journal of Nuclear Medicine and Molecular Imaging 30: 1538-1546.

Petes, T. D. 2001. Meiotic recombination hot spots and cold spots. Nature Review Genetics 2: 360-369.

Plante, Y., Gibson, J. P., Nadesalingam, J., Mehrabani-Yeganeh, H., Lefebvre, S. Vandervoort, G, Jansen, G. B. 2001. Detection of Quantitative Trait Loci Affecting Milk Production Traits on 10 Chromosomes in Holstein Cattle. Journal of Dairy Science 84: 1516-1524.

Pollinger, J. P., Bustamante, C. D., Fledel-Alon, A., Schmutz, S., Gray, M. M., and Wayne R. K. 2005. Selective sweep mapping of genes with large phenotypic effects. Genome Research 15: 1809–1819.

Prasad, A., Schnabel, R. D., McKay, S. D., Murdoch, B., Stothard, P., Kolbehdari, D., Wang, Z., Taylor, J. F., and Moore, S. S. 2009. Linkage disequilibrium and

signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. Animal Genetics 39: 597-605.

Reist, M., Pfaffl, M.W., Morel, C., Meylan, M., Hirsbrunner, Blum, J. W., Steiner, A. 2003. Quantitative mRNA analysis of eight bovine 5-HT receptor subtypes in brain, abomasum, and intestine by real-time RT-PCR. J. Receptor Signal Transduction 23, 271–287.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J, Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E. et al. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

Scheet, P., and Stephens, M. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. American Journal of Human Genetics 78: 629-644.

Schnabel, R. D., Taylor, J. F., and Ashwell, M. S. 2005. Whole-genome scan to detect QTL for milk production, conformation, fertility and functional traits in two US Holstein families. Animal Genetics 36: 408-416.

Simianer, H., Szyda, J., Ramon, G., Lien, S. 1997. Evidence for individual and between family variability of the recombination rate. Mammalian Genome 8: 830 - 835.

Skjervold, H., and Langholz, H. J. 1964. Factors affecting the optimum structure of A.I. breeding in dairy cattle. Z. Tierz. Zuechtungsbiology 80: 26–40.

Stone, R. T., Keele, J. W., Shackelford, S. D., Kappes, S. M., and Koohmaraie, M. 1999. A primary screen of the bovine genome for quantitative trait loci affecting carcass and growth traits. Journal of Animal Science 77: 1379–84.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–95.

Tang, K., Thornton, K. R., and Stoneking, M. 2007.A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biology 5: e171.

Thomsen, H., Reinsch, N., Xu, N., Looft, C., Grupe, S., Kuhn, C., Brockmann, G. A., Schwerin, M., Leyhe-Horn, B., Hiendleder, S., Erhardt, G., Medjugorac, I., Russ, I., Forster, M., Brenig, B., Reinhardt, F., Reents, R., Blumel, J., Averdunk, G., Kalm, E. 2000. A male bovine linkage map for the ADR granddaughter design. Journal of Animal Breeding & Genetics 117: 289-306.

The Bovine HapMap consortium, 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324: 528-532

Viitala, S. M., Schulman, N. F., De Koning, D. J., Elo, K., Kinos, R. Virta, A., Virta, J., Mäki-Tanila, A., Vilkki, J. H. 2003. Quantitative Trait Loci Affecting Milk Production Traits in Finnish Ayrshire Dairy Cattle. Journal of Dairy Science 86: 1828-1836.

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

Walsh, E., Sabeti, P., Hutcheson, H. B., Fry, B., Schaffner, S. F., de Bakker, P. I., Varilly, P., Palma, A. A., Roy, J., Cooper, R., Winkler, C., Zeng, Y., de The, G., Lander, E. S., O'Brien, S., Altshuler, D. 2006. Searching for signals of evolutionary selection in 168 genes related to immune function. Human Genetics 119: 92-102.

Yu, F., Sabeti, P. C., Hardenbo,l P., Fu, Q., Fry, B. Lu, X., Ghose, S., Vega, R., Perez, A., Pasternak, S., Leal, S. M,, Willis, T. D., Nelson, D. L., Belmont, J., Gibbs, R. A. 2005. Positive Selection of a Pre-Expansion CAG Repeat of the Human SCA2 Gene. PLoS Genetics 1: e41.

Zhang, C., Bailey, D. K., Awad, T., Liu, G., Xing, G., Cao, M., Valmeekam, V., Retief, J., Matsuzaki, H., Taub, M., Seielstad, M., Kennedy, G. C. 2006 A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22: 2122-2128.

4th CHAPTER

A Two-step Method for Detecting Selection Signatures Using Genetic Markers

Daniel Gianola[*, †, ‡], Henner Simianer[‡] and Saber Qanbari[‡]

[*] Department of Animal Sciences and Department of Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin 53706

[†]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway

[‡]Department of Animal Sciences, Georg-August-Universität, Göttingen, Germany

Genetics Research 2010, 92: 141-155

Summary

A two-step procedure is presented for analysis of θ (F_{ST}) statistics obtained for a battery of loci, which eventually leads to a clustered structure of values. The first step uses a simple Bayesian model for drawing samples from posterior distributions of θ-parameters but without constructing Markov chains. This step assigns a weakly informative prior to allelic frequencies and does not make any assumptions about evolutionary models. The second step regards samples from these posterior distributions as "data" and fits a sequence of finite mixture models, with the aim of identifying clusters of θ-statistics. Hopefully, these would reflect different types of processes and would assist in interpreting results. Procedures are illustrated with hypothetical data, and with published allelic frequency data for Type-II diabetes in three human populations, and for 12 isozyme loci in 12 populations of the argan tree in Morocco.

INTRODUCTION

The discovery of a massive number of single nucleotide polymorphisms (SNPs) in the genome of several species has enabled exploration of genome-wide signatures of selection via an assessment of variation in marker allele frequencies among populations (e.g., Holsinger and Weir, 2009). Several methods have been proposed for doing this, such as site frequency spectrum, linkage disequilibrium and population differentiation (Sabeti et al., 2006, Akey, 2009). Concerning population differentiation, a parameter $\theta = F_{ST}$, measuring relatedness between pairs of alleles within a sub-population relative to that in an entire population, has been used for this purpose (Wright, 1951, Cockerham, 1969, Weir and Hill, 2002), Lewontin and Krakauer (1973) and Robertson (1975) discuss related approaches. Equivalently, θ 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. For example, Akey et al. (2002) analyzed over 26,500 SNPs for which allele frequencies were available in three populations of humans. The θ parameter was estimated for every marker locus and the distribution of estimates over the entire

genome, and by chromosome, was examined. By referring these estimates to their empirical genome-wide distribution, 174 candidate genes were identified as possible targets of selection.

Holsinger and Weir (2009) provide an account of the logic of the procedure. Briefly, given a set of loci in a given species, a reasonable assumption is that all share the same demographic history and patterns of migration. If these loci are neutral and have similar mutation rates, members of this set can be conceivably regarded as exchangeable realizations of the same evolutionary process. Loci showing departures from the resulting distribution may serve as flags of genomic regions that have been under the influence of selection. Under the hypothesis of selective neutrality, the distribution (over loci) of estimates of θ is expected to be driven by genetic drift, assumed to affect all loci in a similar fashion. On the other hand, when selection operates on one or several loci (as in a multifactorial model for complex traits), markers that are within genes or in nearby locations will display large or small values of θ , the latter occurring when some sort of balancing selection takes place (Cavalli-Sforza, 1996). This opens an avenue for identification of regions associated with population differentiation, e.g., high versus low producing breeds of dairy cattle. Knowledge of such regions may be useful for enhancing the effectiveness of breeding programs via marker-assisted selection, or for tagging variants associated with disease or quantitative traits. While unusual values of θ may point to genomic locations where selection may have operated, there is arbitrariness with respect to characterizing the type of selection that might have occurred. Typically, loci are classified as either neutral, or subject to balancing selection (low values of θ), or favored by selection within some specific population or environment (large population differentiation, thus leading to large values of θ). If the values of θ arise from different evolutionary or artificial (such as in plant and animal breeding) processes, one would expect to observe a mixture of distributions leading to clusters representing the different kinds of mechanisms operating. There is no apparent reason why there should be only two or three such clusters, there may be several clusters harboring loci undergoing different types of selection processes. On the other hand, if θ values vary completely at random due to genetic drift, a single cluster is to be expected.

Statistical issues associated with inferring θ statistics have been discussed, e.g., by Weir and Cockerham (1984) and Weir and Hill (2002), with emphasis in methods of moments estimation, by Balding (2003) using maximum likelihood for beta-binomial and Dirichlet-multinomial distributions, and by Holsinger (1999), Beaumont and Balding (2004) and Guo et al. (2009) employing Bayesian procedures. None of these treatments have addressed the possible existence of a clustered structure.

The objective of this paper is to present a two-step procedure eventually leading to clusters of θ values. The first step, along the lines of Holsinger (1999), Balding (2003) and Beaumont and Balding (2004), uses a simple Bayesian structure for drawing samples from the posterior distributions of θ parameters but without constructing Markov chains. This step assigns a weakly informative prior to allelic frequencies and does not make any assumptions about evolutionary models. The second step regards samples from these posterior distributions as "data" and fits a sequence of finite mixture models, with the aim of identifying clusters of θ statistics. Hopefully, these would reflect different types of processes and would assist in interpreting results.

The paper is organized as follows. Section BACKGROUND reviews basic concepts. In ESTIMATION OF PARAMETERS the first step of the procedure is presented, contrasted with maximum likelihood, and illustrated with a hypothetical data set set and with data on type-II diabetes in three populations. CLUSTERING OF θ PARAMETERS describes the second step of the procedure, and illustrates it with a data set containing allelic frequencies for 12 polymorphic isozyme loci in 12 populations of the argan tree (Argania spinosa L. Skeels) of Morocco presented in Petit et al. (1998) and analyzed by Holsinger (1999). The paper concludes with a discussion of the proposed methodology.

BACKGROUND

Basic concepts

The stage is set by reviewing essentials of a treatment proposed by Cockerham (1969, 1973). Suppose that genetic markers (e.g., SNPs) are screened in a set of individuals

in each of R groups or populations, the latter viewed as drawn at random from some conceptual hyper-population from which such groups derive. Consider a bi-allelic locus (developments carry to multiple alleles as well) and let A_l and a_l be the two alleles at locus 1 (l = 1, 2, …, L), define p_l = Pr (A_l) to be the true, unobserved, frequency of allele A_l in the hyper-population, with 1- p_l = Pr (a_l) being the frequency of a_l. Cockerham (1969) defines a_l as any allele other than A_l and uses an indicator variable x to denote allelic state ("content"), such that

$$x_{rij,l} = \begin{cases} 1 \text{ if an allele } is\ A_l \\ 0 \quad \text{otherwise} \end{cases}$$

Here, r = 1, 2, …, R denotes group or replicate, i indicates an individual, j is an index for a within individual deviation, and l = 1, 2, ..., L is an indicator for locus. Even though $x_{rij,l}$ is a binary variable (so a linear model is questionable) Cockerham (1969) uses the linear decomposition

$$x_{rij,l} = p_l + a_{r,l} + b_{ri,l} + w_{rij,l}, \tag{1}$$

where p_l is as before and

$$a_{r,l} \sim (0, \sigma^2_{a,l}), b_{ri,l} \sim (0, \sigma^2_{b,l}), and \quad b_{ri,l} \sim (0, \sigma^2_{w,l})$$

Are mutually uncorrelated zero-mean random deviates, specific to locus l; the $\sigma^{2'}$s are variance components. Under the assumption that all alleles at locus l in the population have the same marginal distribution,

$$E(x_{rij,l}) = p_l$$

and

$$Var(x_{rij,l}) = p_l(1 - p_l) = p_l + a_{a,l} + b_{b,l} + w_{w,l} = \sigma^2_l$$

for l = 1, 2, ..., L. Decomposition (1) induces the following covariance structure between allelic content variables:

$$Cov\left(x_{rij,l}, x_{r'r'j',l}\right) = \begin{cases} \sigma_l^2 & \text{if } r = r', i = i', j = j' \\ \sigma_a^2 & \text{if } r = r', i \neq i', j \neq j' \\ \sigma_{a,l}^2 + \sigma_{b,l}^2 & \text{if } r = r', i = i', j \neq j' \\ Cov\left(a_r, a_{r'}\right) & \text{if replicates are correlated somehow} \end{cases}$$

A standard assumption is $Cov\left(a_r, a_{r'}\right) = 0$. The following correlations (all positive) follow.

- Pairs of alleles drawn at random from different individuals in the same group are correlated as

$$\rho_{a,l} = \frac{\sigma_{a,l}^2}{\sigma_{b,l}^2 + +\sigma_{b,l}^2 + \sigma_{w,l}^2} = \theta_l = F_{ST,l} \qquad (2)$$

so $0 \leq \theta \leq 1$ for all l:

- Pairs of alleles drawn within individuals over all replicates bear a correlation equal to

$$\rho_{ab,l} = \frac{\sigma_{a,l}^2 + \sigma_{b,l}^2}{\sigma_{b,l}^2 + +\sigma_{b,l}^2 + \sigma_{w,l}^2} = f_l = F_{IT,l}$$

where F is the total inbreeding coefficient, also known as F_{IT} (e.g., Weir and Hill, 2002).

- The correlation between alleles within individuals within the same replicate is

$$\rho_{b,l} = \frac{\sigma_{b,l}^2}{\sigma_{b,l}^2 + \sigma_{w,l}^2} = f_l = F_{IS,l}$$

which is the within sub-population inbreeding coefficient f.

It is easy to show that

$$\theta_l = \frac{F_{IT,l} - F_{IS,l}}{1 - F_{IS,l}} = F_{ST,l}$$

This expression satisfies

$$1 - F_{IT,l} = (1 - F_{IS,l})(1 - F_{ST,l})$$

indicating that a reduction in heterozygosity has as two sources: one that is due to population subdivision or Wahlund's effect, (1 - $F_{ST,l}$) , and a reduction within subpopulation or group caused by "local" inbreeding, (1 - $F_{IS,l}$).

Note that parameter F_{ST} given in (2) can also be written as

$$\theta_l = \frac{\sigma^2_{a,l}}{\sigma^2_{a,l} + \sigma^2_{b,l} + \sigma^2_{w,l}} = \frac{\sigma^2_{a,l}}{p_l(1 - p_l)}$$

Since $\sigma^2_{a,l}$ is the between-group variance in allelic content as per model (1), given R groups, a parametric representation of θ_l in terms of the unknown gene frequencies is

$$\theta_l = \frac{\dfrac{\sum_{r=1}^{R} (p_{r,l} - \bar{p}_l)^2}{R}}{\bar{p}_l - (1 - \bar{p}_l)}, \qquad (3)$$

where $\bar{p} = \sum_{r=1}^{R} p_{r,l} / R$ is the average (over groups) of the frequencies of allele A_l at locus l: Note that $\bar{p}_l$ is taken as an unweighted average, it does not seem sensible to express a parameter in terms of sample size (unless weights assigned to samples reflect true population sizes). Expressing θ_l explicitly in terms of the locus-specific gene frequencies yields

$$\theta_l = \frac{\sum_{r=1}^{R} p_{r.l}^2 - \frac{\left(\sum_{r=1}^{R} p_{r.l}\right)^2}{R}}{\left(\sum_{r=1}^{R} p_{r.l} - \frac{\left(\sum_{r=1}^{R} p_{r.l}\right)^2}{R}\right)}, \quad (4)$$

providing a mapping from the joint space of R allelic frequencies to the single dimensional space of θ_l, which resides in (0, 1).

In Cockerham's model (1), the random variables are the discrete allelic outcome (x) and the random deviates *a, b* and *w*. Their joint distribution is indexed by fixed parameters, one of which is the intra-class correlation θ_l given in (2). Now, if allelic frequencies for different populations are drawn at random from the same stochastic evolutionary process (e.g., as generated by random drift), θ_l becomes a random variable. Over loci, this defines the distribution of values of θ expected under neutrality assumptions, and the resulting process will depend on the distribution assumed for the allelic frequencies. From a Bayesian perspective, every unknown is a random variable and, since allelic frequencies are unknown, θ as given in (3) will posses a distribution, both *a priori* and *a posteriori*. In the first step of the method proposed in this paper, the posterior distribution of θ_l will result from assigning a vague prior to all allelic frequencies, corresponding in some sense to what could be termed as a fixed effects treatment from a frequentist perspective. The second step addresses the question of whether or not all 114 θ_l stem from the same distribution or from different distributions resulting from heterogeneity of the underlying stochastic processes. This makes the treatment proposed here different from those in, e.g., Holsinger (1999) or Balding (2003).

ESTIMATION OF PARAMETERS

Inferring gene frequencies

Gene frequencies can be inferred using a simple Bayesian approach. Suppose that nr individuals are genotyped in population *r*, so that the number of alleles screened at locus *l* is $2n_r = n_{r,Al} + n_{r,al}$, where $n_{r,Al}$ and $n_{r,al}$ are the observed numbers of copies of A_l and a_l, respectively.

A convenient assumption is that of mutual independence between the distributions of alleles at different loci (stronger than that of pairwise linkage equilibrium). Linkage disequilibrium is pervasive but the assumption made above facilitates matters and is widely used, e.g., by Corander et al. (2003). Let $p = (p1, p2, ..., pR)$' be an $RL \times 1$ vector of allelic frequencies for all *R* groups, where $p_r = (p_{r,1}, p_{r,2}, ..., p_{r,L})$' has order $L \times 1$: Under the mutual independence assumption, the likelihood conferred by the observed number of copies of alleles to the gene frequencies is

$$l(\mathbf{p}|DATA) = \prod_{r=1}^{R}\prod_{l=1}^{L} p_{r,l}^{n_{r,A_l}} (1-p_{r,l})^{n_{r,a_l}} \qquad (5)$$

The maximum likelihood estimator of $p_{r,l}$ is $\hat{p}_{r,l} = \frac{n_{r,A_l}}{2n_r}$ and its empirical variance is $Var(\hat{p}_{r,l}) = \frac{\hat{p}_{r,l}(1-\hat{p}_{r,l})}{2n_r}$. The maximum likelihood estimator is unbiased but unstable, and may take values at the boundaries of the parameter space in small samples. In a Bayesian treatment, allelic frequencies are assigned a prior distribution that might be homogeneous or heterogeneous over populations, chromosomes or genomic regions (e.g., coding versus non-coding regions). For example, Holsinger (1999, 2006) adopts a prior beta distribution, $Beta(p_l \mid a_l, b_l)$ (and interprets it as describing variation over populations) with parameters

$$a_l = \frac{1-\theta}{\theta} x_l,$$

and

$$b_l = \frac{1-\theta}{\theta}(1-x_l).$$

Here θ is common to all loci (i.e., the hypothesis of neutrality) and x_l is the mean allelic frequency at locus l (averaged over populations). Using properties of the beta distribution in the parametric definition of θ leads to

$$\frac{Var(p_l)}{E(p_l)[1-E(p_l)} = \frac{\dfrac{a_l b_l}{(a_l+b_l)^2(a_l+b_l+1)}}{\dfrac{a_l}{(a_l+b_l)}\cdot\dfrac{b_l}{(a_l+b_l)}} = \theta.$$

Then, the joint posterior distribution of all unknowns (all132 allelic frequencies, θ and vector $\mathbf{x} = \{x_l\}$) is

$$g(\mathrm{p},\theta,\mathrm{x}|DATA) \propto \prod_{r=1}^{R}\prod_{l=1}^{L} p_{r,l}^{n_{r,A_l}+\frac{1-\theta}{\theta}(1-x_l)-1}(1-p_{r,l})^{n_{r,a_l}+\frac{1-\theta}{\theta}(1-x_l)-1})g(\theta)g(\mathrm{x}).$$

Holsinger (1999) took $g(\theta) = Beta(1, 2)$ distribution as prior for θ and assumed that all x_l were identically distributed according to the uniform process $g(x_l) = U(0, 1)$: Given θ and $\mathbf{x}$, the allelic frequencies are conditionally independent with conditional posterior distributions

$$g(p_{r,l} \mid ELSE) = Beta\left(n_{r,A_l}+\frac{1-\theta}{\theta}x_l, n_{r,a_l}+\frac{1-\theta}{\theta}(1-x_l)\right);$$
$$r = 1,2,...,R; \quad l = 1,2,...,L$$

where *ELSE* means all parameters other than $p_{r,l}$ and the data observed: However, the conditional posterior distributions of θ and $\mathbf{x}$ are not recognizable, so an elaborate sampling scheme, e.g., one based on Markov chain Monte Carlo methods, must be tailored. Holsinger (1999) found that inferences were insensitive with respect to the choices of beta and uniform prior distributions for θ and elements of $\mathbf{x}$, respectively. However, it was assumed (as in a neutral model) that all loci share the same θ

parameter. This produces a mutual borrowing of information among loci (shrinking $p_{r,l}$ towards a common value), but the procedures is not explicit with respect to the existence of heterogeneity over loci due forces such as differential mutation or selective sweeps. As proposed by Beaumont and Balding (2004), one could estimate locus specific θ-values and refer these estimates to the posterior distribution of θ under the homogeneity value. In this manner, outliers could be found with respect to the "neutral" distribution, but this would not inform about the structure of any latent heterogeneity. Here, an alternative approach is used. Jeffreys rule (Bernardo and Smith, 1994, Sorensen and Gianola, 2002) is used to produce a reference prior, which is a $Beta\ (\frac{1}{2},\frac{1}{2})$ distribution assigned to all loci in all populations. This reference prior distribution is minimally informative in a well defined sense (Bernardo and Smith, 1994). Using Bayes theorem, the joint posterior density of all allelic frequencies is now

$$\begin{aligned} \mathrm{g(p \mid DATA)} &\propto \prod_{r=1}^{R}\prod_{l=1}^{L} p_{r,l}^{n_{r,A_l}+\frac{1}{2}-1}(1-p_{r,l})^{n_{r,a_l}+\frac{1}{2}-1} \\ &= \prod_{r=1}^{R}\prod_{l=1}^{L} Beta\left(n_{r,A_l}+\frac{1}{2}, n_{r,a_l}+\frac{1}{2}\right) \end{aligned} \tag{6}$$

Thus, allelic frequencies at different loci are mutually independent, a posteriori, with $p_{r,l}$ following a beta distribution with parameters $\alpha_{rl} = n_{r,A_L} + \frac{1}{2}$ and $\beta_{rl} = n_{r,a_l} + \frac{1}{2}$ Possible point estimates of allelic frequencies are the posterior mean

$$\vec{p}_{r,l} = \frac{n_{r,A_L} + \frac{1}{2}}{2n_r + 1}, \tag{7}$$

and the posterior mode

$$\vec{p}_{r,l} = \frac{n_{r,A_l} - \frac{1}{2}}{2n_r - 1}, \text{ for } n_{r,Al} \geq 1. \tag{8}$$

The variance of the posterior distribution of $p_{r,l}$ is

$$Var(pr,l \mid DATA) = \frac{\left(n_{r,A_l} + \frac{1}{2}\right)\left(n_{r,a_l} + \frac{1}{2}\right)}{(2n_r + 1)^2(2n_r + 2)} \quad (9)$$

Even though a weakly informative prior is used, differences exist with respect to maximum likelihood. To illustrate this point, consider a hypothetical example with 2 groups, *M* and *N*. Suppose that 100 individuals are genotyped in group *M* and that the observed number of Al alleles is 199, i.e., the locus is nearly fixed. The maximum likelihood estimate of $p_{M,l}$ is 0.995 and its estimated standard error is 4.99×10^{-3}; a calculation based on asymptotic normality (without truncation) yields that the probability of obtaining estimates larger than 1 is close to 0.16! Further, the probability of obtaining estimates between 0.9 and 0:995 is close to $\frac{1}{2}$. On the other hand, the posterior distribution of $p_{M,l}$ is *Beta* (199 + $\frac{1}{2}$, 1 + $\frac{1}{2}$). The posterior mean and posterior standard deviation are 0.993 (note some shrinkage away from the edge of the parameter space) and 6.06×10^{-3}, respectively, the posterior probability of the frequency being larger than 1 is exactly zero, and the probability that $p_{M,l}$ takes values between 0.9 and 0.995 is about 0.57. Figure 1 displays the posterior distribution of the allelic frequency obtained with Jeffrey's prior, overlaid against the normal approximation to the distribution of the maximum likelihood estimates. Clearly, the approach used makes a difference, even in a situation where allelic frequencies are estimated with reasonable precision, as indicated by the small standard error of the maximum likelihood estimate and the small posterior standard deviation in the Bayesian analysis (the coefficient of variation of the posterior distribution is less than 1%).

In the second population, *N*, 30 individuals are genotyped and 10 alleles are of the type A_l, the maximum likelihood estimate of $p_{N,l}$ is then $\frac{1}{6}$, much lower than in M, and its sampling variance is 2.31×10^{-3}. The posterior distribution of $p_{N,l}$ is *Beta* (10 + $\frac{1}{2}$, 50 + $\frac{1}{2}$). In *N*, the posterior density of $p_{N,l}$ and the normal approximation to the density of the distribution of the maximum likelihood estimator are very similar (not shown here).

Differences in allelic frequencies between populations *M* and *N* at the locus in question may be due to random drift or may suggest a signature of selection.

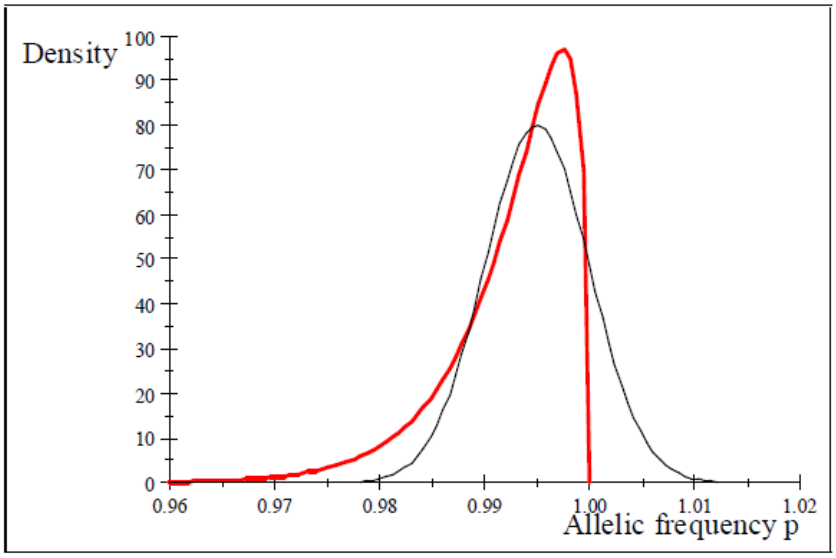

Figure 1. Posterior density (thick line) of the allelic frequency *p* at a locus for which 199 copies have been observed out of 200 alleles counted in hypothetical population M; the posterior distribution is Beta (199+ ½, 1+ ½). The thin line is the density of a normal approximation to the sampling distribution of the maximum 2 likelihood estimator.

Inferring θ by maximum likelihood

A likelihood-based estimate of θ can be obtained by replacing in (3) or (4) the unknown allelic frequencies by their maximum likelihood estimates. For the example of populations *M* and *N* above, the estimate is

$$\hat{\theta}_l = \frac{\sum_{r=1}^{2}\left(\hat{p}_{r,l} - \bar{\hat{p}}_l\right)^2}{2\bar{\hat{p}}_l\left(1-\bar{\hat{p}}_l\right)} \approx 0.7046$$

The sampling variance of the maximum likelihood estimator of θ_l can be approximated using a Taylor series expansion. As shown in Appendix A, the first derivative of θ_l with respect to the allelic frequency at locus *l* in group *r* is

$$\frac{\partial}{\partial p_{r,l}}\theta_l = \left[\frac{2(p_{r,l}-\bar{p}_l)}{\bar{p}_l{}^2-\bar{p}_l^2}-\frac{(1-2\bar{p}_l)}{\bar{p}_l(1-2\bar{p}_l)}\right]\frac{\theta_l}{R},$$

for $r=1,2,...,L,$ where $\bar{p}_l=\frac{\sum_{r=1}^{R}p_{r,l}}{R}$ is as before and $\bar{p}_l^2=\frac{\sum_{r=1}^{R}p_{r,l}^2}{R}$. Further, let $\hat{\nabla}=\left\{\frac{\partial}{\partial p_{r,l}}\theta_l\right\}_{p_{r,l}=\hat{p}_{r,l}}$, be an $RL\times 1$ vector of first derivatives evaluated at the maximum likelihood estimates of the allelic frequencies. Then, approximately

$$\hat{Var}(\hat{\theta}_l)\approx\hat{\nabla}'\hat{Var}(\hat{p})\hat{\nabla}=\sum_{r=1}^{R}\left\{\left[\frac{2(p_{r,l}-\bar{p}_l)}{\bar{p}_l{}^2-\bar{p}_l^2}-\frac{(1-2\bar{p}_l)}{\bar{p}_l(1-2\bar{p}_l)}\right]\frac{\theta_l}{R}\right\}^2_{p_{r,l}=\hat{p}_{r,l}}\frac{\hat{p}_{r,l}(1-\hat{p}_l)}{2n_r}$$

where $\hat{Var}(\hat{p})=Diag\left(\frac{\hat{p}_{r,l}(1-\hat{p}_l)}{2n_r}\right)$ is a diagonal matrix containing the estimates of the sampling variances of the maximum likelihood estimates of allelic frequencies $p_{r,l}$. For the hypothetical example, $Var(\hat{\theta})\approx 9.8265\times 10^{-5}$. The asymptotic normal approximation to the distribution of the estimates assigns nil probability to "estimates" outside of (0, 1); the probability of obtaining estimates of θ between 0.67 and 0.74 for this two-population situation is 0.9996.

Bayesian inference of θ

Consider now finding the posterior distribution of θ_l as defined in (4) and without making the assumption that the θs are realizations from the same stochastic process, i.e., without borrowing information across loci over and above the shrinkage of allelic frequencies produced by Jeffrey's prior.

The posterior distribution is analytically difficult to arrive at because θ_l is a non-linear function of gene frequencies in all R groups. However, since it is easy to obtain independent samples from each of the $Beta\ (n_{r,A_l}+\frac{1}{2}, n_{r,a_l}+\frac{1}{2})$ processes, Monte Carlo estimates of features of the posterior distribution of θ_l can be obtained without

using Markov chain Monte Carlo methods at all. Let $p_{r,l}^{(s)}, s = 1,2,...,S,$ be samples from the posterior (beta) distribution of $p_{r,l}$, the frequency of allele A_l at locus l. Then, a draw from the posterior distribution of θ_l is given by

$$\theta_l^{(s)} = \frac{\sum_{r=1}^{R}(p_{r,l}^{(s)} - \frac{\left(\sum_{r=1}^{R}(p_{r,l}^{(s)}\right)^2}{R}}{\left(\frac{R\sum_{r=1}^{R}p_{r,l}^{(s)} - \left(\sum_{r=1}^{R}(p_{r,l}^{(s)}\right)^2}{R}\right)} \qquad (10)$$

which is a random variable with support in (0, 1) (Holsinger, 2006). Then, from S samples, the mean, median, variance, etc., of the posterior distribution of θ_l can be estimated. Each θ_l (l = 1, 2, ..., L) will have a point estimate and an assessment of uncertainty, e.g., a credibility interval of size 95% given by the 2.5% and 97.5% percentiles of the corresponding posterior distribution estimated either from samples or from the normal theory approximation given in Appendix B.

In the hypothetical populations M and N the posterior distributions of the frequency of A_1 are *Beta* (199.5, 1.5) and *Beta* (10.5, 50.5), respectively. With draws denoted as $B^{(s)}$ (., .) , S samples from the posterior distribution of θ_l can be obtained as:

$$\theta_l^{(s)} = \frac{[B^{(s)}(199.5,1.5)]^2 + [B^{(s)}(10.5,50.5)]^2 - \frac{\{[B^{(s)}(199.5,1.5)] + [B^{(s)}(10.5,50.5)]\}^2}{2}}{[B^{(s)}(199.5,1.5)]^2 + [B^{(s)}(10.5,50.5)]^2 - \frac{\{[B^{(s)}(199.5,1.5)] + [B^{(s)}(10.5,50.5)]\}^2}{2}}$$

$; s = 1,2,..., S.$

To illustrate, 5000 samples were drawn from each of the two beta distributions, to form S= 5000 corresponding draws from the posterior distribution of θ_l. The mean and median were 0.6966 and 0.6972, respectively, the standard deviation was 0:070 and the range of values samples spanned from 0.4268 to 0.8883. The posterior density

of θ_l and the empirical cumulative distribution function are in Figures 2 and 3, respectively. Values of θ_l appearing with appreciable density range from about 0.5 to 0.9 (Figure 2), with small posterior probability assigned to values smaller than 0.6. (Figure 3).

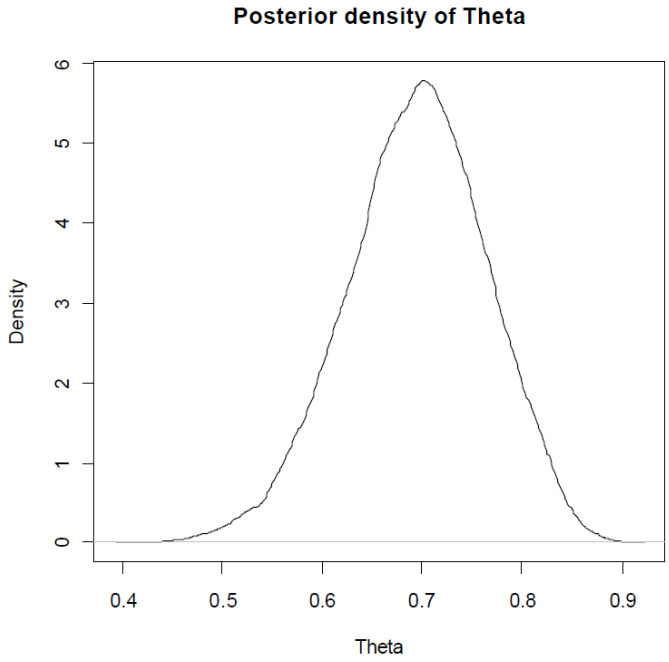

Figure 2. Posterior density of θ_l for the hypothetical example 5 of populations M and N

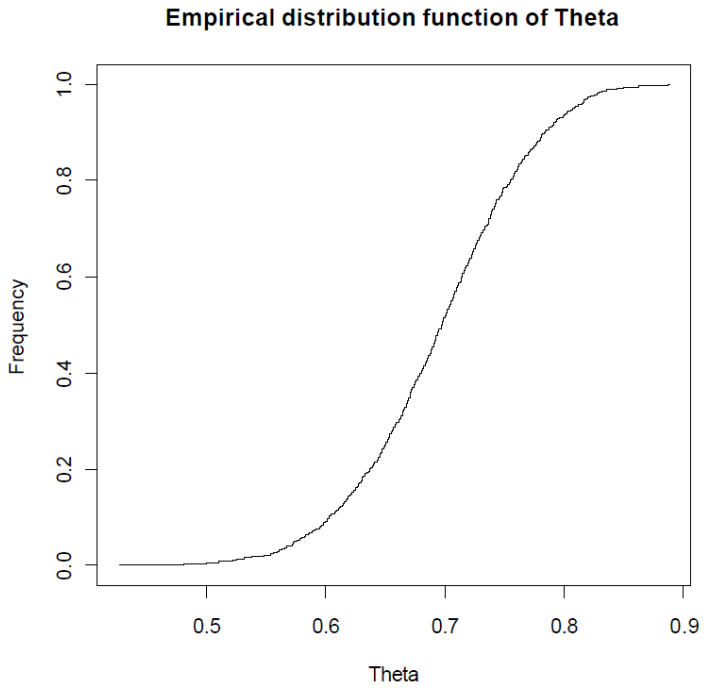

Figure 3. Empirical cumulative distribution function of θ_l for the hypothetical example of populations *M* and *N*.

A Bayesian "null" distribution for assessing sampling variation uncertainty

It is important to check whether or not posterior estimates of θ_l depart from what would be expected by chance alone. A posterior distribution consistent with expectations under a "null" model is formulated next. The θ_l statistics calculated from the "full" model above can then be referred to this null distribution. Note that the "null" distribution given below describes the uncertainty to be expected from drawing random samples from the same population, but not the variability to be expected due to genetic drift. If estimates of θ_l fall in this null distribution, this would indicate that the study lacks power to answer evolutionary questions in any meaningful manner.

A "null" distribution" is arrived at by stating that $p_{r,l} = p_l$ is the same random variable for all *R* populations. Under this assumption, the posterior distribution of the vector of gene frequencies (now of dimension L× 1) under the "null" model is

$$g(\mathbf{p}|DATA, \text{Null}) \propto \frac{\left[\prod_{r=1}^{R}\prod_{l=1}^{L} p_l^{n_{r,A_l}}(1-p_l)^{n_{r,a_l}}\right]\prod_{l=1}^{R} p_l^{\frac{1}{2}-1}(1-p_l)^{\frac{1}{2}-1}}{\prod_{l=1}^{L} Beta\left(\sum_{r=1}^{R} n_{r,A_l}+\frac{1}{2}, \sum_{r=1}^{R} n_{r,a_l}+\frac{1}{2}\right)} \quad (11)$$

Hence, allelic frequencies p_l are mutually independent, a posteriori, with $p_l|DATA$, Null being a beta distribution with parameters $\alpha_l = \sum_{r=1}^{R} n_{r,A_l} + \frac{1}{2}$ and $\beta_r = \sum_{r=1}^{R} n_{r,a_l} + \frac{1}{2}$. A draw from the posterior distribution of the F_{ST} statistic under this model takes the form

$$\theta_{l,Null}^{(s)} = \frac{\sum_{r=1}^{R}(p_l^{(r,s)} - \overline{p}_l^{(s)})^2}{\frac{R}{\overline{p}_l^{(s)}(1-\overline{p}_l^{(s)})}}, \quad (12)$$

Where $p_l^{(r,s)}$ is a draw from $Beta\left(\sum_{r=1}^{R} n_{r,A_l}+\frac{1}{2}, \sum_{r=1}^{R} n_{r,a_l}+\frac{1}{2}\right)$, with R such draws involved in a realization of $\theta_l^{(s)}$, and $\overline{p}_l^{(s)}$ is the average of the R draws. A set of samples from the posterior distribution of θ_l under the null model is obtained by repeating the sampling process S times. This distribution serves as a reference against which the θ_l statistics calculated from the "full" model can be compared. If the posterior mean of θ_l obtained from the "full" model falls outside of a high density area of the posterior distribution of θ in the null model, then the divergence between populations would be probably due to drift or selection (assuming mutation rates are constant over populations), but not due to chance alone.

For the example of populations M and N, $\sum_{r=1}^{R} n_{r,A_l} = 209$ and $\sum_{r=1}^{R} n_{r,a_l} = 51$. Figure 4 depicts the *Beta* (209.5, 51.5) distribution of the allelic frequency under the "null" model. Note that the maximum likelihood estimates of the allelic frequencies in the M and N populations, of 0.995 and $\frac{1}{6}$, respectively, are not assigned any appreciable density under this model. Upon drawing 5000 independent samples from the beta

distribution of the allelic frequency under the null model, 5000 draws for $\theta_{l,Null}^{(s)}$ were obtained by evaluating (12) for each of the samples. Draws ranged from 8.24×10^{-13} to 0.0503; the mean (standard deviation) was 0.0038(0.0053) and the posterior median was 0.0017.

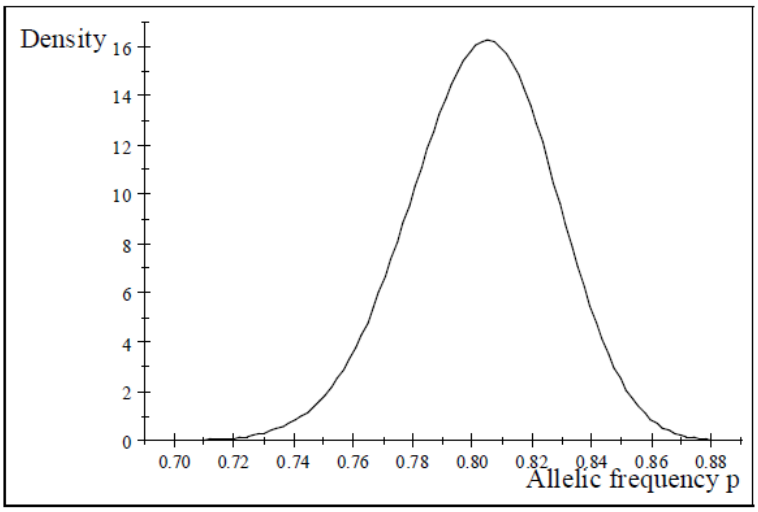

Figure 4. Posterior density of the allelic frequency p under a drift ("null") model for hypothetical populations M and N ; 209 copies of A_l are observed out of 260 alleles screened.

The posterior density of $\theta_{l,Null}$ was very sharp as shown in Figure 5. In the full model, the estimated posterior mean (standard deviation) of θ_l was 0.6966, which is unlikely to have been generated under the null distribution. This would make the locus a reasonable candidate for further examination.

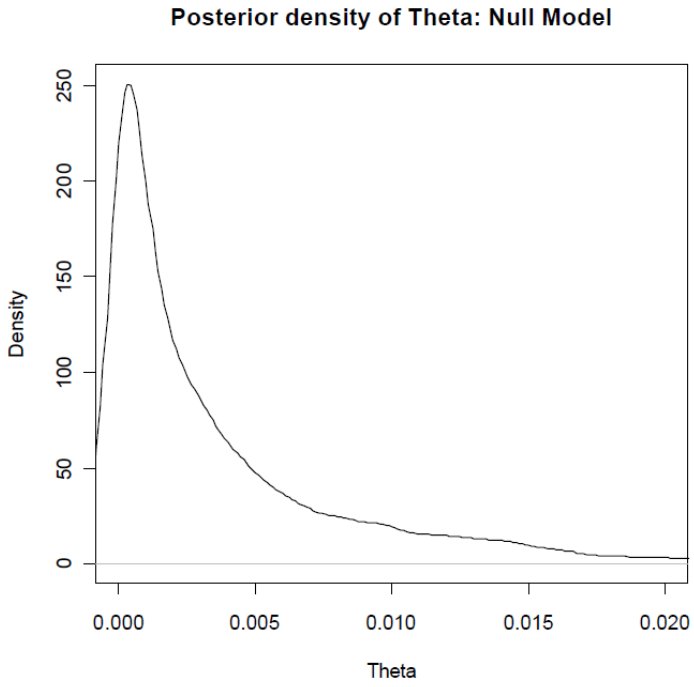

Figure 5. Posterior density of θ_l under the null model for the hypothetical example of populations *M* and *N*:

Illustration of sampling variation with candidate genes for type-II diabetes

The Bayesian method was applied to data pertaining to identification of candidate gene variants for type II diabetes in Polynesians (Myles et al., 2007). Prevalence of this disease is high in several Pacific populations, e.g., 40% of adults living in the island of Nauru. DNA samples were obtained from 23 Polynesians, 23 New Guineans and 19 Han Chinese from Beijing. Type II diabetes associated alleles were from 10 SNP loci having evidence of association. Estimated frequencies and θ l statistics are shown in page 587 of Myles et al. (2007). To illustrate the Bayesian procedure, data for the KCNJ11 locus was used, and susceptibility allele frequencies (Al in our notation) were 0.30, 0.25 and 0.34 in the three populations above, respectively. Their figures do not lead to an integer number of alleles, due to rounding error, so the number of observed Al alleles used here was set to 14 (Polynesians), 12 (New Guineans) and 13 (Han Chinese). Myles et al. (2007) employed an "unbiased

estimator" of θ 1 for calculating population pairwise differences, and their estimates were 0.003 (New Guinea- China), -0.024 (China-Polynesia) and -0.017 (New Guinea-Polynesia). Note the two negative estimates of a parameter that resides in (0, 1), standard errors or significance levels were not provided. Their analysis suggests that this locus is not associated with prevalence of the disease.

The posterior distributions of A_l were $Beta_{\text{Polynesians}}$ (14.5, 32.5), $Beta_{\text{New Guineans}}$ (12.5, 34.5) and $Beta_{\text{Han Chinese}}$ (13.5, 25.5). The number of samples drawn from each of these 3 posterior distributions was S = 1000, and 1000 draws from the posterior distribution of θ_{KCNJ11} were obtained by evaluation of (10). Values of θ_{KCNJ11} ranged from 2.423×10^{-5} to 0.1500, with an estimated posterior mean (standard deviation) of 0.019 (0.0193), this estimate is higher than that of Myles et al. (2007). The non-parametric estimate of the posterior density of θ_{KCNJ11} is shown in Figure 6, illustrating that the true value of the FST parameter is very likely below 0.10. The posterior inter-correlation structure between allelic frequencies and θ_{KCNJ11} in the full model was examined and, as expected, draws from the posterior distributions of allelic frequencies in the three populations were uncorrelated. Samples of θ_{KCNJ11} were positively correlated (0.55) with those for allelic frequency in Chinese Han, and the 95% confidence interval for the correlation was 0.51-0.60. However, draws for θ_{KCNJ11} were negatively correlated with allele frequencies in Polynesians (-0.07) and New Guineans (-0.39), the confidence intervals for these two correlations were (-0.13,-0.01) and (-0.44,-0.34), respectively.

For the "null" model, the 1000 samples from the posterior distribution of $\theta_{\text{KCNJ11,Null}}$ ranged from 3.62×10^{-6} to 0.1460, with the posterior mean (standard deviation) estimated at 0.002 (0.002); the posterior median was 0.002 as well. The posterior mean (standard deviation) estimate of θ_{KCNJ11} under the "full" model was 0.019, and it did not enter with high density in the "null" model (not shown).

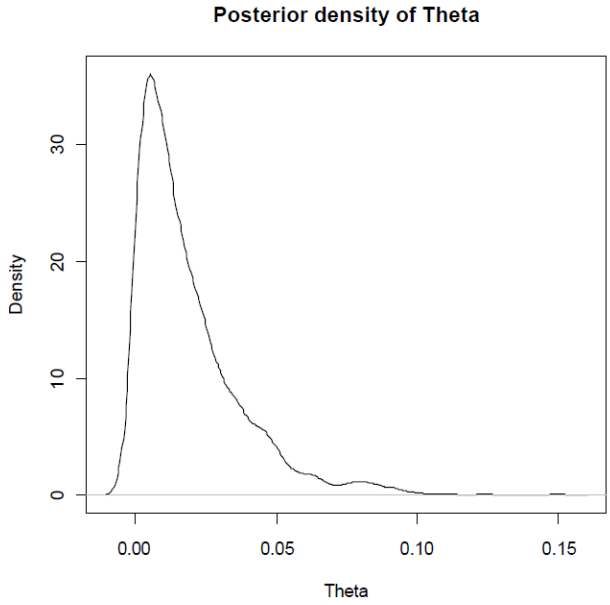

Figure 6. Density of the posterior distribution of θ_{KCNJ11} obtained from allelic frequencies in Myles et al. (2007).

Although variation in allelic frequency at locus KCNJ11 among the three populations departs from what would be expected from chance alone (statistical sampling), the observed θ value is very small. This may support the hypothesis that this locus may not be associated with differences in prevalence of type II diabetes, in agreement with Myles et al. (2007). Allelic frequencies were uncorrelated, as it should be, given that the three replicates were drawn from the same *Beta* (209.5, 51.5) distribution. The $\theta_{KCNJ11,Null}$ statistic was uncorrelated with allelic frequencies, and the correlations were -0.08, -0.11 and 0.03 in the three replicates, with all confidence intervals including 0.

CLUSTERING OF θ-PARAMETERS

The second step of the procedure consists of clustering a set of estimates of θ-values (in this case, posterior means) from a multi-locus analysis into data driven groups. The expectation is that these clusters might be representative of different processes taking place in the populations such as balancing or directional selection, neutrality or anything else. The method is illustrated with data from a study of Petit et al. (1998) in

which alleles were sampled for 12 isozyme loci of the *Argania* genus tree in each of 12 areas (populations) of Morocco. The data, given in page 847 of Petit et al. (1998), were modified as shown in Table 1. The modification consisted of treating all loci as bi-allelic by lumping alleles for loci with more than 2 variants into 2 classes. The number of individuals sampled per population ranged between 20 and 50, and the number of alleles per locus varied originally between 2 and 5. Note that, at some loci, one of the alleles was fixed in almost all populations. For example, for locus 3, the only population in which segregation was observed was TA.

Table 1. Allelic frequencies at 12 isozyme loci in each of 12 Argan tree populations, adapted from Petit et al. (1998) by making all loci bi-allelic. A1-A12 represent frequencies of the "A" allele at loci 1-12; No. A1-No. A12 are the observed number of copies of the alelles. The number of "a" alleles can be calculated from the number of individuals samples and the number of "A" alleles observed.

Population	AB	AD	AR	BS	GO	MI	OG	SI	TA	TE	TM	TT
No. Individuals	20	40	20	30	32	20	30	20	30	20	20	50
A1	0.525	0.512	0.475	0.467	0.047	0.475	0.517	0.575	0.517	0.425	0.55	0.52
No. A1	21	41	19	28	3	19	31	23	31	17	22	52
A2	0.4	0.438	0.55	0.917	0.688	0.525	0.467	0.825	0.483	0.925	0.475	0.51
No. A2	16	35	22	55	44	21	28	33	29	37	19	51
A3	1	1	1	0	1	1	1	1	0.75	1	1	1
No. A3	40	80	40	0	64	40	60	40	45	40	40	100
A4	0.525	0.375	0.45	0.517	0.922	0.525	1	0.7	0.467	0.575	0.5	0.52
No. A4	21	30	18	31	59	21	60	28	28	23	20	52
A5	0.475	0.463	0.475	1	1	1	1	1	0.817	1	1	0.51
No. A5	19	37	19	60	64	40	60	40	49	40	40	51
A6	0.85	0.538	0.9	0.533	0.922	0.575	0.55	0.75	0.517	0.525	0.55	0.53
No. A6	34	43	36	32	59	23	33	30	31	21	22	53
A7	1	1	1	0.567	0.922	0.9	1	1	0.967	1	1	1
No. A7	40	80	40	34	59	36	60	40	58	40	40	100
A8	1	1	1	1	1	1	1	1	1	1	0.575	0.97
No. A8	40	80	40	60	64	40	60	40	60	40	23	97
A9	1	0.937	1	1	0.312	1	1	1	1	1	1	1
No. A9	40	75	40	60	20	40	60	40	60	40	40	100
A10	0.925	0.5	0.525	0.625	0.475	0.5	0.55	0.4	0.575	0.5	0.475	0.5
No. A10	37	40	21	38	30	20	33	16	35	20	19	50
A11	0.6	0.7	0.575	0.5	0.6	0.525	1	0.375	0.625	0.475	0.55	0.47
No. A11	24	56	23	30	38	21	60	15	38	19	22	47
A12	1	1	0.85	0.6	0.875	0.775	1	0.875	1	1	1	0.87
No. A12	40	80	34	36	56	31	60	35	60	40	40	87

For each locus, 2000 samples were drawn from the beta posterior distributions of allelic frequencies. For example, the posterior distribution of $p_{AB,1}$ was Beta (21.5,

19.5). From these samples, 2000 draws from the posterior distribution of θ for each locus were formed as in (10). The posterior means were.

$$\begin{bmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & \theta_6 & \theta_7 & \theta_8 & \theta_9 & \theta_{10} & \theta_{11} & \theta_{12} \\ 0.098 & 0.168 & 0.791 & 0.166 & 0.393 & 0.137 & 0.299 & 0.382 & 0.593 & 0.095 & 0.122 & 0.190 \end{bmatrix}$$

so estimates of θ varied over loci from about 0.095 (locus 10) to 0.791 (locus 3), all these estimates did not enter into the corresponding "null" distributions. Boxplots of the posterior distributions of the θ parameters are in Figure 7. Visually, it is tempting to suggest four clusters: the first one would include locus 3, with the posterior mean of θ close to 0.79, the second cluster would include locus 9, with an estimate of θ of 0.59. The third cluster would include loci 5, 7 and 8 with estimates ranging between 0.30 and 0.39, and the fourth cluster would be represented by loci 1, 2, 4, 6, 10, 11, 12 having the lowest estimates of θ.

The existence of an underlying structure is suggested by the distribution of all 24000 samples, presented in Figure 8. In the left panel, a non-parametric density estimate was obtained from these samples treated as if all draws (2000 for each of the 12 loci) had been made from the same process, the densities in the middle and right panels correspond to the logit, i.e., $\log\left(\frac{\theta}{1-\theta}\right)$, and Gompit, -log(-log (θ)), transforms of the sampled θ values, respectively. The three densities suggest that θ values cluster around 3, perhaps 4, modes.

The structure was explored more formally by fitting a sequence of finite mixture models to the means of the posterior distribution of the θ-values for each of the 12 loci. These posterior means are independent (under the assumptions made for the allelic frequency models) but not identically distributed, since they are estimated with different precision, due to unequal numbers of individuals sampled and varying allelic frequencies. The distributions of θ-values among loci are not normal (the logit and Gompit transforms would be expected to be more nearly so).

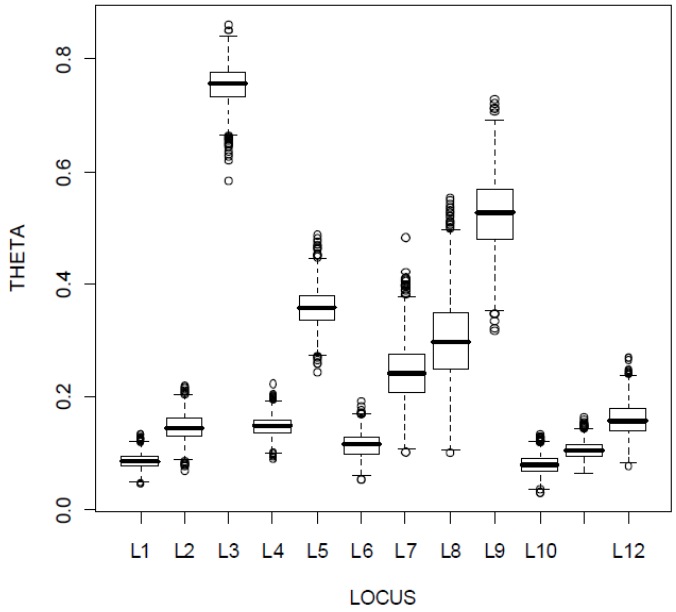

Figure 7. Boxplot of the posterior distributions of θ parameters in 12 isozyme loci of the argan tree in Morocco (data originally from Petit et al., 1998).

This should not be an issue because the mixture model was not used for testing hypotheses, its objective, rather, was to explore a clustered structure. Since there are only 12 posterior means, the mixture models must have less than 12 parameters, otherwise, a perfect fit would be obtained. The mixture model fitted to the posterior mean estimates $\bar{\theta}_l$ postulated that

$$\bar{\theta}_l \text{ or } \log\left(\frac{\bar{\theta}_l}{1-\bar{\theta}_l}\right) \text{ or } -log\,(-log\,(\bar{\theta}_l)) \sim \sum_{k=1}^{K} \pi_k N(\bar{\theta} \mid \mu_k, \sigma_k^2),$$

where K is the number of components of the mixture (clusters of posterior means of θ-values or transforms thereof), π_k is the probability that $\bar{\theta}_l$ belongs to cluster k (subject to $\sum_{k=1}^{K} \pi_k = 1$, and μ_k and σ_k^2 are the mean and variance, respectively, of component k. For example, if $k = 2$, there are 5 "free" parameters in the mixture, if $k = 4$, there are 11 such parameters, so it is not sensible to fit a model with more than 4 components. Mixture model parameters were estimated by maximum likelihood via the

expectation-maximization algorithm as implemented in the FlexMix package (Leisch, 2004) in the R project (R development core team, 2008). Upon convergence (assuming the stationary point was a global maximum), the conditional probability that $\overline{\theta}_l$ (or its transformation) belongs to cluster *k* is calculated as

$$\Pr(\text{locus } l \in \text{cluster } k | \text{ parameter estimates}) = \frac{\hat{\pi}_k N(\overline{\theta}_l \mid \hat{\mu}_k, \hat{\sigma}_k^2)}{\sum_{k=1}^{K} \hat{\pi}_k N(\overline{\theta}_l \mid \hat{\mu}_k, \hat{\sigma}_k^2)}$$

The locus was assigned to the cluster with the largest conditional probability. Models with different values of k = 1, 2, 3, 4 were compared using Akaike's information criterion (AIC), that is

$$AIC(K) = 2\left[pk - \sum_{l=1}^{12} \log\left(\sum_{k=1}^{K} \hat{\pi}_k N(\overline{\theta}_l \mid \hat{\mu}_k, \hat{\sigma}_k^2) \right) \right],$$

where p_K is the number of parameters for a model with K components (McLachlan and Peel, 2000). Models with the smallest *AIC* values are preferred. It is known that this criterion tends to overstate the number of components due to violation of regularity conditions in mixture models (Celeux and Soromenho, 1996).

Results of the mixture model analysis, by number of components fitted, are shown in Table 2. The AIC criterion favored a mixture with 2 clusters when the response was either θ or its Gompit transform, and a single component when the logit transformation was used. Clearly, with data from only 12 loci, the analyses did not have enough power to resolve heterogeneity in a finer manner. This would certainly not be the case with SNP data, where the number of marker loci typically oscillates between a few thousands in some animal species to close to a million in humans. Classification probabilities using K=2 and estimates of cluster mean and standard deviation are shown in Table 3. Irrespective of whether θ values were transformed or not, loci were clustered into two groups, one consisting of loci 3,5,7,8 and 9, possibly reflecting a selection signature, and the other one including the remaining loci, presumably representing neutral loci.

Table 2. Comparison of mixture models with 2, 3 or 4 components fitted to the 12 posterior means of θ-parameters and their logit or Gompit transforms in the argan tree data of Petit et al. (1998). AIC: Akaike's information criterion (models with smallest values are favored and indicated in boldface).

Variable	No. components (k)	Iterations to convergence	AIC
θ	k=1	2	-0.651
	k=2	16	**-6.299**
	k=3	36	-2.921
	k=4	39	3.079
$\log\frac{\theta}{1-\theta}$	k=1	2	**39.100**
	k=2	28	40.102
	k=3	77	44.392
	k=4	94	50.392
$-\log[-\log(\theta)]$	k=1	2	26.909
	k=2	36	**24.328**
	k=3	41	27.742
	k=4	48	33.742

Table 3. Conditional probabilities of membership to one of two clusters for mixture models fitted to the posterior means of _ for the 12 loci in the argan tree, and their logit, $\log\left(\frac{\theta}{1-\theta}\right)$, and Gompit, -log(-log (θ)); transformations (boldfaced probability indicates the cluster with largest probability of membership).

	θ means		logit(θ)		Gompit(θ)	
Locus	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	**0.93**	0.07	**0.91**	0.09	**0.91**	0.09
2	**0.92**	0.08	**0.83**	0.17	**0.89**	0.11
3	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
4	**0.92**	0.08	**0.82**	0.18	**0.88**	0.12
5	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
6	**0.95**	0.05	**0.91**	0.09	**0.93**	0.07
7	0.00	**1.00**	0.08	**0.92**	0.04	0.96
8	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
9	0.00	**1.00**	0.00	**1.00**	0.00	**1.00**
10	**0.92**	0.08	**0.89**	0.11	**0.89**	0.11
11	**0.95**	0.05	**0.92**	0.08	**0.93**	0.07
12	**0.87**	0.13	**0.76**	0.24	**0.83**	0.17
Cluster Mean	0.12	0.41	-2.03	-0.52	-0.11	0.76
Cluster standard deviation	0.03	0.21	0.32	1.02	0.67	0.13

The maximum likelihood estimates of the mean and variance of θ values in the cluster with loci 3,5,7,8 and 9 were 0.41±0.21, whereas the corresponding estimates in the other cluster were 0.12± 0.03. This assignment into clusters is consistent with the picture emerging from visual consideration of the box plots in Figure 7.

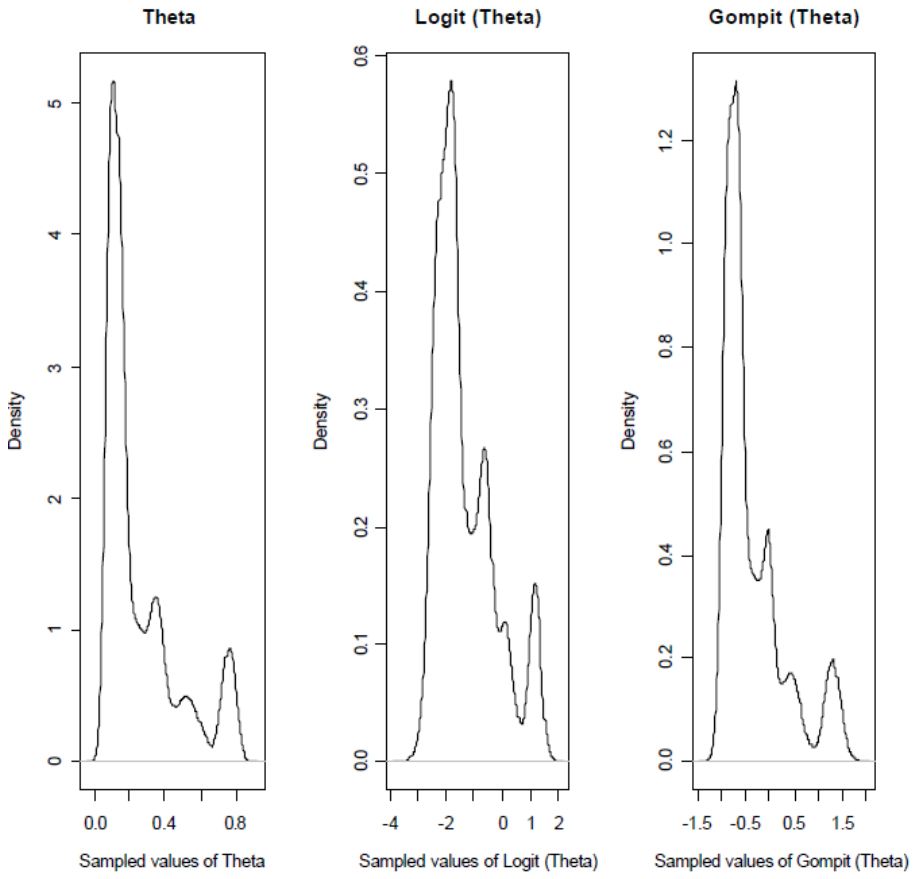

Figure 8. Non-parametric density estimates of θ-values (based on 2000 samples for each of 12 loci), logit (θ) and Gompit (θ): All samples treated as homogeneous, i.e., as generated from the same stochastic process.

In principle, a better approach is to feed the entire set of posterior samples to the clustering procedure, such that not only the location of the posterior distributions of the θs is considered, but their uncertainty as well. This is very appealing conceptually, but it creates havoc in the EM algorithm, which often fails to converge. For instance Qanbari et al. (personal communication) employed the procedure with posterior means (each calculated with 1 million samples from the corresponding posterior distribution) with about 35,000 SNPs in Hereford and Simmental cattle. When posterior means were used as data, the mixture model approach revealed the existence

of 4-5 clusters. However, when the 35 million samples were used as data points, the EM algorithm, as implemented in FlexMix, failed to converge.

DISCUSSION

The use of F-statistics for the study of genetic divergence between population dates back to Wright (1931). Holsinger and Weir (2009) have provided a justification for their usefulness, e.g., in association mapping and in detecting genomic regions affected by evolutionary processes, such as selection. These authors also reviewed different types of statistical methods for inferring F_{ST}, including Bayesian procedures. Method of moments estimation was prompted by the linear model formalism of Cockerham (1969, 1973), and a review is in Weir and Hill (2002). There has been an increased interest in Bayesian methods, and important contributions in this front have been made by Holsinger (1999, 2006), Beaumont and Balding (2004) and Guo et al. (2009).

In the Bayesian approaches that have been suggested, e.g., Holsinger (1999), the model poses a product binomial (or product multinomial in the case of multiple alleles) likelihood function for allelic frequencies, with conjugate prior distributions, such as beta or Dirichlet processes. Marginalizing over the allelic frequencies yields the beta binomial or Dirichlet-multinomial distributions used by Balding (2003) for likelihood-based inference. Holsinger (1999) matched the mean and variance of, e.g., the beta distribution, to the definition of θ, and obtained a joint posterior distribution which is a function of the unknown allelic frequencies, of θ (assumed exchangeable over all loci) and of the mean allelic frequencies in an undivided population. The implementation, as well as those of Beaumont and Balding (2004) and of Guo et al. (2009) requires Markov chain Monte Carlo sampling (MCMC). While the power and flexibility of hierarchical models coupled with MCMC are well known (Sorensen and Gianola, 2002), implementations are not trivial and monitoring of convergence to the equilibrium distribution is a delicate matter (Cowles and Carlin, 1996). The idea in these methods is that, under a neutral model, all θ (over loci) should be realizations of the same stochastic process. Outlying θ values may be suggestive of genomic regions

affected by selection. Typically, it is argued that loci are either neutral, subject to balancing selection or to directional selection favoring alleles in specific environments, e.g., Akey el al. (2002). However, the assignment of loci to specific types of processes is often arbitrary.

The present paper follows ideas of Holsinger (1999) but it differs in two important respects. The proposed method has two steps. First, allelic frequencies are assigned a non-informative prior, so that the mutual borrowing of information between loci is limited, leading to less shrinkage of frequencies towards a common value, in maximum likelihood there is no shrinkage at all, an issue criticized by Haldane (1948). Samples of allelic frequencies can be obtained directly (actually, their posterior distributions are tractable, analytically), and these draws are used to form draws from the posterior distribution of locus-specific θ parameters, using the parametric definition of F_{ST} as a function of allelic frequencies. The first step was illustrated with hypothetical data and with type II diabetes data in Myles et al. (2007). The step leads to estimates of the posterior distribution of the θs which can be used to explore underlying structure, presumably caused by different evolutionary forces. In the second step, the structure is explored by using features of the posterior distribution of the θs (posterior means or transformations thereof) as response variables in a mixture model. Data from Petit et al. (1998) on 12 isozyme loci in 12 populations of the argan tree in Morocco were used to illustrate the second step. Here, the posterior means of θ are treated as belonging to a mixture of normal distributions which is then resolved into data-supported components. Since the final objective is that of clustering loci according to their similarity in θ-values, departures from normality are arguably of little consequence. Here, logit and Gompit transformations were examined, and the clustering procedure produced exactly the same results. Using Akaike's information criterion as a gauge for model comparison, it was suggested that the 12 estimates of θ clustered into two groups, one representing putatively neutral loci (provided that this group reflects variation due to drift), and another one possibly corresponding to genomic regions affected by selection. With 12 loci only, it is unreasonable to expect a finer clustering structure. An ongoing study is applying the two-step procedure to

large scale SNP data in an animal population and this will be reported in a future communication.

The method proposed here extends naturally to multiple alleles. In this case the likelihood is product multinomial and the beta prior distribution is replaced by a Dirichlet distribution with minimum information content. The posterior distribution of the allelic frequencies is product Dirichlet, which is simple to sample from. Then, samples from the posterior distribution of θ_l would be drawn by evaluation of formulae similar to those in Nei (1973) where θ-parameters are averaged over alleles. For example, one could define

$$\theta_l = \sum_{m=1}^{M} \overline{p}_{l,m} \frac{\dfrac{\sum_{r=1}^{R} (p_{r,l,m} - \overline{p}_{r,l,m})^2}{R}}{\overline{p}_{r,l,m}(1 - \overline{p}_{r,l,m})} = \sum_{m=1}^{M} \frac{\dfrac{\sum_{r=1}^{R} (p_{r,l,m} - \overline{p}_{r,l,m})^2}{R}}{\overline{p}_{r,l,m}(1 - \overline{p}_{r,l,m})}$$

where $p_{r,l,m}$ is the frequency of allele m at locus l in population r and $p_{l,m}$ is the unweighted average over the R populations.

In common with the studies of Holsinger (1999), Beaumont and Balding (2004), Weir et al. (2005) and Guo et al. (2009) the procedure presented here assumes that allelic frequencies are in linkage equilibrium, so that the likelihood of all allelic frequencies is either product binomial or product multinomial. Accommodating linkage disequilibrium, especially with dense batteries of marker loci, represents a formidable task and it is a challenge for future research. For example, Akey et al. (2002) and Weir et al. (2005) reported that θ values of loci in regions of high linkage disequilibrium were similar. Guo et al. (2009) address correlations due to linkage, but not due to linkage disequilibrium, and do so by introducing a spatial structure for loci located in the same chromosome. Specifically, they proposed an autoregressive model in which logit transforms of θ-values are correlated according to physical distance. The model is quite involved and requires MCMC computations. However, loci may be in linkage disequilibrium even though not being physically linked (Crow and Kimura, 1970), and such disequilibrium is very common in animal populations

(Sandor et al., 2006, de Roos et al. 2008, Lipkin et al., 2009, Qanbari et al. 2010), where finite size and selection under epistasis are factors in building up linkage disequilibrium. The two-step approach considered here could be enhanced by exploring algorithms alternative to EM as well as by consideration of different types of mixtures, e.g., of beta distributions, which are more appropriate for random variables taking values in (0, 1).

Acknowledgements

Part of this work was carried out while the senior author was a Visiting Professor at Georg-August-Universität, Göttingen (Alexander von Humboldt Foundation Senior Researcher Award). Support by the Wisconsin Agriculture Experiment Station, and by grant NSF DMS-NSF DMS-044371 is acknowledged. This study is part of the project FUGATO-plus GenoTrack and was financially supported by the German Ministry of Education and Research, BMBF, the Föderverein Biotechnologieforschung e.V. (FBF), Bonn, and Lohmann Tierzucht GmbH, Cuxhaven. S. Qanbari thanks the H. Wilhelm Schaumann Stiftung Hamburg for .nancial support. Prof. W. G. Hill is thanked for useful comments.

References

Akey, J. M. .2009. Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Research 19: 711.722

Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12: 1805-1814.

Balding, D. J. 2003. Likelihood-based inference for genetic correlations. Theoretical Population Biology 63: 221-230.

Beaumont, M. A., and Balding, D. J. 2004. Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology 13: 969-980.

Bernardo, J. M., and Smith, A. F. M. .1994. Bayesian Theory. Chichester: Wiley.

Cavalli-Sforza, L. L. 1966. Population structure and human evolution. Proc. R. Soc. Lond. B. Biol. Sci. 164: 362-379.

Celeux, G., and Soromenho, G. 1996. An entropy criterion for assessing the number of clusters in a mixture model. Classification Journal 13: 195-212.

Cockerham, C. C. 1969. Variance of gene frequencies. Evolution 23: 72-84.

Cockerham, C. C. 1973. Analyses of gene frequencies. Genetics 74: 679-700.

Corander, J., Waldmann, P., and Sillanpää, M. 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163: 367-374.

Cowles, M. K., and Carlin, B. P. 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. Journal of the American Statistical Association 91: 883-904.

Crow, J. F., and Kimura, M. 1970. An Introduction to Population Genetics Theory. Caldwell: Blackburn Press.

de Roos, A. P.W., Hayes, B. J., Spelman, R. J. and Goddard, M. E. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179: 1503-1512.

Guo, F., Dey, D. K., and Holsinger, K. E. 2009. A Bayesian hierarchical model for analysis of single nucleotide polymorphisms, diversity in multilocus, multipopulation samples. Journal of the American Statistical Association 104: 142-154.

Haldane, J. B. S. 1948. The precision of observed values of small frequencies. Biometrika 35: 297-303.

Holsinger, K. E. 1999. Analysis of genetic diversity in geographically structured populations: a Bayesian perspective. Hereditas 130: 245-255.

Holsinger, K. E. 2006. Bayesian hierarchical models in geographical genetics. In: Clark JS, Gelfand AE, editors. Applications of Computational Statistics in the Environmental Sciences. Oxford University Press; New York. pp. 25-37.

Holsinger, K. E., and Weir, B. S. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST} : Nature Review Genetics 10: 639-650.

Leisch, F. 2004. FlexMix: a general framework for .nite mixture models and latent class regression in R. Journal of Statistical Software 11: 1-18.

Lewontin, R. C., and Krakauer, J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.

Lipkin, E., Straus, K., Stein, T., Bagnato, A., Schiavini, F., Fontanesi, L., Russo, V., Medugorac, M., Foerster, M., Sölkner, J., Dolezal, M., Medrano, M. F., Friedmann, A., and Soller, M. 2009. Extensive long-range and non-syntenic linkage disequilibrium in livestock populations. Genetics 181: 691-699.

Myles, S., Hradetzky, E., Engelken, J., Lao, O. Nürnberg, P., Trent, R. J., Wang, X., Kayser, M., and Stoneking, M. 2007. Identi.cation of a candidate genetic variant for the high prevalence of type II diabetes in Polynesians. European Journal of Human Genetics 15: 584-589.

McLachlan, G. and Peel, D. 2000. Finite Mixture Models. New York: Wiley.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National Academy of Sciences 70: 3321-3323.

Petit, R. J., El Mousadik, A. and Pons. O. 1998. Identifying populations for conservation on the basis of genetic markers. Conservation Biology 12: 844-855.

Qanbari, S., Pimentel, E. C. G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R. and Simianer, H. (2009) The pattern of linkage disequilibrium in German Holstein cattle. Animal Genetics (In press).

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robertson, A. 1975. Gene frequency distributions as a test of selective neutrality. Genetics 81: 775-785.

Sabeti, P. C., Scha¤ner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky,S., Palma,A., Mikkelsen, T. S., Altshuler, D. and Lander, E. S. 2006. Positive natural selection in the human lineage. Science 312: 1614-1620.

Sandor, C., Farnir, F., Hansoul, S., Coppieters,W., Meuwissen, T. and Georges, M. 2006. Linkage disequilibrium on the bovine X chromosome: characterization and use in quantitative trait locus mapping. Genetics 173: 1777-1786.

Sorensen, D. and Gianola, D. 2002. Likelihood, Bayesian and MCMC Methods in Quantitative Genetics. New York: Springer.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M. and Hill, W. G. 2005. Measures of human population structure shows heterogeneity among genomic regions. Genome Research 15: 1468-1476.

Weir, B. S. and Cockerham, C. C. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

Weir, B. S. and Hill, W. G. 2002. Estimating F-statistics. Annual Review of Genetics 36: 721-750.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16: 97-159.

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics 15: 323-354.

APPENDIX A: First derivatives of θ with respect to allelic frequencies

Let $\overline{p}_{.,l} = \sum_{r=1}^{R} p_{r.l}$. From (4), the derivative is

$$\frac{\partial}{\partial p_{r,l}}\theta$$

$$= \frac{1}{\left(\frac{R\sum_{r=1}^{R} p_{r,l} - \left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}\right)}\left(2p_{r,l} - \frac{2\sum_{r=1}^{R} p_{r,l}}{R}\right) - \frac{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}}{\left(\frac{R\sum_{r=1}^{R} p_{r,l} - \left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}\right)^2}\left(\frac{R - 2\sum_{r=1}^{R} p_{r,l}}{R}\right)$$

$$= \frac{2\theta_l}{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}}\left(p_{r,l} - \frac{\sum_{r=1}^{R} p_{r,l}}{R}\right) - \frac{\theta_l}{\left(\frac{R\sum_{r=1}^{R} p_{r,l} - \left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}\right)}\left(\frac{R - 2\sum_{r=1}^{R} p_{r,l}}{R}\right)$$

$$= \left[\frac{2\left(p_{r,l} - \frac{\sum_{r=1}^{R} p_{r,l}}{R}\right)}{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}} - \frac{\left(1 - \frac{2\sum_{r=1}^{R} p_{r,l}}{R}\right)}{\sum_{r=1}^{R} p_{r,l} - \frac{\left(\sum_{r=1}^{R} p_{r,l}\right)^2}{R}}\right]\theta_l$$

$$= \left[\frac{2\left(p_{r,l} - \overline{p}_{.,l}\right)}{\sum_{r=1}^{R} p_{r,l}^2 - \frac{\left(R\overline{p}_{.,l}\right)^2}{R}} - \frac{\left(1 - 2\overline{p}_{.,l}\right)}{\sum_{r=1}^{R} p_{r,l} - \frac{\left(R\overline{p}_{.,l}\right)^2}{R}}\right]\theta_l$$

$$= \left[\frac{2\left(p_{r,l} - \overline{p}_{.,l}\right)}{R\left(\overline{p_{.,l}^2} - \overline{p}_{.,l}^2\right)} - \frac{\left(1 - 2\overline{p}_{.,l}\right)}{R\left(\overline{p}_{.,l} - \overline{p}_{.,l}^2\right)}\right]\theta_l$$

$$= \left[\frac{2\left(p_{r,l} - \overline{p}_{.,l}\right)}{\overline{p_{.,l}^2} - \overline{p}_{.,l}^2} - \frac{\left(1 - 2\overline{p}_{.,l}\right)}{\overline{p}_{.,l}\left(1 - \overline{p}_{.,l}\right)}\right]\frac{\theta_l}{R} \qquad (13)$$

APPENDIX B: Approximate Bayesian analysis

An approximate Bayesian analysis without sampling from the posterior distribution is also possible. An approximation to the mean and variance of the posterior distribution of θ_l can be obtained using a Taylor series expansion about the modes $\tilde{p}_{r,l}$ of the allelic frequencies. Let now $\overline{\nabla} = \left(\frac{\partial}{\partial p_{r,l}} \theta_l \right)_{p_{r,l} = \tilde{p}_{r,l}}$ be an R×1 vector of first derivatives evaluated at the posterior mode estimates (8) of the allelic frequencies. Then, approximately

$$\theta_l \approx \theta_l + \overline{\nabla}'(p_l - \tilde{p}_l),$$

where $\tilde{p}_l$ is the vector of posterior mode estimates of allele frequencies in the R groups. Then, approximately

$$\begin{aligned} & E(\theta_l | DATA) \\ \approx\ & \tilde{\theta}_l + \sum_{r=1}^{R} \left\{ \frac{\partial}{\partial p_{r,l}} \theta_l \right\}_{p_{r,l}=\tilde{p}_{r,l}} \left(\frac{n_{r,A_l} + \frac{1}{2}}{2n_r + 1} - \frac{n_{r,A_l} - \frac{1}{2}}{2n_r - 1} \right) \\ =\ & \tilde{\theta}_l + \sum_{r=1}^{R} \left\{ \left[\frac{2(p_{r,l} - \overline{p}_{\cdot,l})}{\overline{p^2}_{\cdot,l} - \overline{p}^2_{\cdot,l}} - \frac{(1 - 2\overline{p}_{\cdot,l})}{\overline{p}_{\cdot,l}(1 - \overline{p}_{\cdot,l})} \right] \frac{\theta_l}{R} \right\}_{p_{r,l}=\tilde{p}_{r,l}} \left(\frac{n_{r,A_l} + \frac{1}{2}}{2n_r + 1} - \frac{n_{r,A_l} - \frac{1}{2}}{2n_r - 1} \right) \end{aligned} \quad (14)$$

Likewise,

$$Var(\theta_l \mid DATA) \approx \overline{\nabla}' Var(p_l \mid DATA) \overline{\nabla}.$$

Since allelic frequencies have mutually independent distributions, the R _ R variance-covariance matrix $Var(\theta_l \mid DATA)$ is diagonal with elements given by (9). Thus

$$Var(\theta_l \mid DATA) \approx \overline{\nabla}' Diag\left[\frac{(n_{r,A_l}+\frac{1}{2})(n_{r,a_l}+\frac{1}{2})}{(2n_r+1)^2(2n_r+2)}\right]\overline{\nabla}' \tag{15}$$

In short, each θ_l ($l = 1, 2, \ldots, L$) statistic will have a point estimate and an assessment of uncertainty, e.g., a credibility interval of size 95% given by the 2.5% and 97.5% percentiles of the corresponding posterior distribution estimated from samples, or from using a normal theory approximation, e.g.,

$$\widetilde{\theta}_l + 2\sum_{r=1}^{R}\left\{\left[\frac{2(p_{r,l}-\overline{p}_l)}{\overline{p_l^2}-\overline{p}_l^2}-\frac{(1-2\overline{p}_l)}{\overline{p}_l(1-\overline{p}_l)}\right]\frac{\theta_l}{R}\right\}_{p_{r,l}=\widehat{p}_{r,l}}\left(\frac{n_{r,A_l}+\frac{1}{2}}{2n_r+1}-\frac{n_{r,A_l}-\frac{1}{2}}{2n_r-1}\right)$$

$$\pm 1.96\sqrt{\sum_{r=1}^{R}\left\{\left[\frac{2(p_{r,l}-\overline{p}_l)}{\overline{p_l^2}-\overline{p}_l^2}-\frac{(1-2\overline{p}_l)}{\overline{p}_l(1-\overline{p}_l)}\right]\frac{\theta_l}{R}\right\}^2_{p_{r,l}=\widehat{p}_{r,l}}\frac{(n_{r,A_l}+\frac{1}{2})(n_{r,a_l}+\frac{1}{2})}{(2n_r+1)^2(2n_r+2)}}.$$

5th CHAPTER

Application of Site and Haplotype-Frequency Based Approaches for Detecting Selection Signatures in Cattle

S. Qanbari*, D. Gianola†, B. Hayes‡, F. Schenkel§, S. Miller§, G. Thaller, H. Simianer***

* Animal Breeding and Genetics Group, Department of Animal Sciences, Georg-August University, 37075 Göttingen, Germany

† Department of Animal Sciences and Department of Dairy Science, University of Wisconsin-Madison, Madison, Wisconsin 53706

‡Animal Genetics and Genomics, Primary Industries Research Victoria, 475 Mickleham Rd, Attwood, VIC 3049, Australia.

§ Centre for Genetic Improvement of Livestock, Animal and Poultry Science Department, University of Guelph, Guelph, Ontario, N1G 2W1 Canada

** Institute of Animal Breeding and Animal Husbandry, Christian-Albrechts-University, 24098 Kiel, Germany

(Submitted)

ABSTRACT

Background

'Selection signatures' delimit regions of the genome that are, or have been, functionally important and have therefore been under either natural or artificial selection. In this study, two different and complementary methods—integrated Haplotype Homozygosity Score (|iHS|) and population differentiation index (F_{ST})—were applied to identify traces of decades of intensive artificial selection for traits of economic importance in modern cattle.

Results

We scanned the genome of a diverse set of dairy and beef breeds from Germany, Canada and Australia genotyped with a 50K SNP panel. Across breeds, a total of 109 extreme |iHS| values exceeded the empirical threshold level of 5% with 19, 27, 9, 10 and 17 outliers in Holstein, Brown Swiss, Australian Angus, Hereford and Simmental, respectively. Annotating the regions harboring clustered |iHS| signals revealed significant enrichment for functional genes like SPATA17, MGAT1, PGRMC2 and ACTC1, COL23A1, MATN2, respectively, in the context of reproduction and muscle formation. In a further step, a new Bayesian F_{ST}-based approach was applied with a set of geographically separated populations including Holstein, Brown Swiss, Simmental, North American Angus and Piedmontese for detecting differentiated loci. In total, 127 regions exceeding the 2.5 per cent threshold of the empirical posterior distribution were identified as extremely differentiated. In a substantial number (56 out of 127 cases) the extreme F_{ST} values were found to be positioned in poor gene content regions which deviated significantly ($p < 0.05$) from the expectation assuming a random distribution. However, significant F_{ST} values were found in regions of some relevant genes such as SMCP and FGF1.

Conclusions

Overall, 236 regions putatively subject to recent positive selection in the cattle genome were detected. Both |iHS| and F_{ST} suggested selection in the vicinity of the Sialic acid binding Ig-like lectin 5 gene on BTA18. This region was recently reported

to be a major QTL with strong effects on productive life and fertility traits in Holstein cattle. We conclude that high-resolution genome scans of selection signatures can be used to identify genomic regions contributing to within- and inter-breed phenotypic variation.

BACKGROUND

The domestication of cattle (*Bos taurus* and *Bos taurus indicus*) 8,000–10,000 years ago [1] had a significant impact on human civilization. Since that time, a broad range of either natural as well as man made factors (e.g., geography, environment, culture and directional artificial selection) has led to diversity in cattle: Today we know more than 800 cattle breeds across the world. The cattle genome therefore represents a significant opportunity for identifying genetic variation that contributes to phenotypic diversity and for detecting genome response to strong directional selection from both domestication and subsequent artificial selection.

Recently a number of studies with different analytical concepts have been conducted to detect signals of recent positive selection on a genome-wide scale [2, 3, 4, 5, 6, and 7]. The methods used are based either on the allele frequency spectrum or on properties of haplotypes segregating in populations. For example, comparing F_{ST} values among loci provides an estimate of how much genetic variability exists between, rather than within, populations [8, 9]. This statistic assumes that geographically variable selective forces favor different variants in different regions. Hence, between-population allele frequency differences may be more extreme in genome regions harboring such variants. The method can be used to scan patterns of variation over many loci. Akey et al. (2002) [10] suggested using the loci in the tails of the empirical distribution as candidate targets of selection. Another approach to infer evidence of past selection is the “Extended Haplotype Homozygosity” (EHH) test [11] which identifies regions with an unusually long range of haplotype and a high population frequency. Voight et al. (2006) [12] developed the “integrated Haplotype Score” (|iHS|), an extension of EHH, based on the comparison of EHH

between derived and ancestor alleles within a population. In this concept, directional selection favoring a new mutation results in a rapid increase in the frequency of the selected allele along with the background haplotype in which the mutation arose. This phenomenon increases linkage disequilibrium (LD) on the chromosomes which harbor the derived (selected) allele, but not the unselected allele, which therefore acts as a "control". Thus, this measure is most sensitive to a rapid increase in the frequency of the derived allele at a selected site, but the derived allele must have existed only on a distinct background (haplotype) prior to selection and must not have reached fixation yet [12, 13]. After fixation, the |iHS| statistic may continue to identify regions of high LD surrounding the selected site, but may not detect selection at the selected region itself because fixation will eliminate variation at and near the selected site.

In this study we scan the genome of a diverse set of cattle breeds including dairy and beef breeds based on the 50K SNP panel. Besides identifying selection footprints common to all breeds, these analyses examine how divergent directions of positive selection may have affected the genomic pattern of those breeds. Our analyses focus primarily on two haplotype and site frequency based statistics: the |iHS| and F_{ST} statistics. These tests were chosen because previous power analyses suggest they are largely complementary—|iHS| has good power to detect selective sweeps at moderate frequency, while in contrast, F_{ST} is most powerful to detect selection on fixed variation [14]. Applying the |iHS| test with a new Bayesian method of F_{ST}, we report a panel of 236 regions putatively subject to recent positive selection confirming the higher differentiation index and longer haplotype consistency for a strong QTL recently detected in Holstein cattle.

METHODS

Animals

A diverse set of animals collected from Germany, Australia and Canada were used for this study. Table 1 summarizes information of 3876 animals included in our study. The main subset involves Holstein (HS), Simmental (SI) and Brown Swiss (BS)

breeds which are part of the total population of cattle genotyped for the genomic selection program in Germany. These breeds are highly selected, essentially for milk production (HF and BS) or for milk and beef (SI). The second subset consisted of 900 individuals collected from 6 beef breeds genotyped in Australia. Another subset of beef cattle included 103 North American Angus (CN) and 43 Piedemontese (PI) collected from Ontario, Canada. The first data set (data set I) consisted of the German breeds mentioned above together with the Australian beef breeds; it was used for LD based analysis in this study. In contrast, the second data set (data set II) included the German breeds together with the Canadian sample and was used for the site frequency approach.

Table 1. Description of samples

Breed	**Code**	**Data set**		**Sample (n)**	**Country**	**Purpose**
Holstein	HS	I	II	2091	Germany	Dairy
Brown Swiss	BS	I	II	277	Germany	Dairy
Simmental	SI	I	II	462	Germany	Dual-purpose
North American Angus	CA	-	II	103	Canada	Beef
Piedemontese	PI	-	II	43	Canada	Beef
Australian Angus	AA	I	-	232	Australia	Beef
Brahman	BR	I	-	80	Australia	Beef
Belmond Red	BE	I	-	166	Australia	Beef
Hereford	HR	I	-	158	Australia	Beef
Murray Gray	MG	I	-	57	Australia	Beef
Santa Gertrudis	SG	I	-	126	Australia	Beef
Shorthorns	SH	I	-	81	Australia	Beef

SNP genotypes and data preparation

Semen or blood samples were used as source of genomic DNA. All samples were genotyped using the Illumina Bovine SNP 50K BeadChip [15]. However, they were

genotyped on multiple platforms and at different times. To ensure the highest possible data quality a series of filters were employed to remove lower quality markers and insecure genotypes for individuals. We filtered out samples with ≥5% missing genotypes and SNP loci assigned to unpositioned contigs. Genotypes were also discarded if they had quality scores <95%.

We used only autosomal SNPs with minor allelic frequencies (MAF) ≥ 0.05 in the LD based analysis (data set I). Haplotypes were then reconstructed for each chromosome using default options in fastPHASE [16]. Reconstructed haplotypes were inserted into HAPLOVIEW v4.1 [17] to estimate LD statistics based on pair-wise r^2 and construct the blocking pattern in the candidate regions of interest for selection signature analysis. Both paternal and maternal haplotypes were utilized for selection signature analyses.

For the analysis of site frequency spectrum, all SNPs that passed quality control were used in the final analysis, so that loci with MAF <5% or fixed in some populations were included as well. After quality control and removal of individuals with high proportion of missing genotypes (≥ 5 %), data set II consisted of 40,595 common SNPs typed on 2976 animals in 5 breeds (Table 1). The number of heterozygous loci was determined and used to estimate the average heterozygosity for all individuals across the breeds. Allele frequencies and observed and expected heterozygosity for each SNP were also estimated.

Calculation of |iHS| values

We employed the *iHS* test to evaluate the evidence of positive selection based on haplotype frequencies as described by Voight et al. (2006) [12]. The *iHS* statistic measures the extent of local LD, partitioned into two classes: haplotypes centered upon a SNP that carry the ancestral *versus* the derived allele. For the purpose of this study we used the set of ancestral alleles identified and reported in Matukumalli et al. (2009) [15]. This statistic is applied to individual SNPs and begins by calculating the integrated EHH [11, 6], which is defined as the integral of the observed decay of EHH

(i.e. the area under the curve of EHH versus distance) away from a specified core allele until EHH reaches 0.05. This integrated EHH (iHH) (summed over both directions away from the core SNP) is denoted iHH_A or iHH_D, depending on whether it is computed for the ancestral or derived core allele. The unstandardized *iHS* is then calculated as follows:

$$\text{unstandardized } iHS = \ln\left(\frac{iHH_A}{iHH_D}\right)$$

This quantity is standardized such that it has a mean of 0 and variance of 1 irrespective of allele frequency at the core SNP (see Voight et al. 2006 [12] for details).

$$iHS = \frac{\ln\left(\frac{iHH_A}{iHH_D}\right) - E\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD\left[\ln\left(\frac{iHH_A}{iHH_D}\right)\right]}$$

Large positive or negative values of *iHS* indicate unusually long haplotypes carrying the ancestral or derived allele, respectively.

Population differentiation index

In this study we estimated $F_{ST} = \theta$ statistic [9] using a new Bayesian algorithm proposed by Gianola et al. (2010) [18]. The procedure has two steps. First, allelic frequencies are assigned a non-informative prior, leading to less shrinkage of frequencies towards a common value. In maximum likelihood there is no shrinkage at all, an issue criticized by Haldane (1948) [19]. Samples of allelic frequency can be obtained directly because their posterior distributions are tractable analytically and those draws are used to form draws from the posterior distributions of locus-specific θ–parameters, using the parametric definition of F_{ST} as a function of allelic frequency (see [18] for more details). This step leads to estimates of the posterior distribution of θ which can be used to explore any underlying structure, presumably caused by

different evolutionary forces. In the second step the structure is explored by using features of the posterior distribution of θ (posterior means or transformations thereof) as response variables in a mixed model.

RESULTS

Marker and LD statistics

Table 2 presents a descriptive summary of data characteristics across breeds for data set I. The average observed heterozygosity and mean MAF were similar in all dairy and dual purpose breeds, while the MAF was generally lower and more variable in beef breeds. The second data set consisted of 40,595 common SNPs typed in 5 breeds which covered 2544.1 Mbp of the genome (Btau 4.0 assembly) with 62.68 ± 58.3 Kbp average adjacent marker spacing. Analysis of the entire panel of across-breed SNPs revealed a non uniform distribution of allele frequencies by breed (results not shown).

Table 2. Genome wide summary of marker statistics for the breeds used in LD based analysis (data set I).

Breed	SNP (n)	MAF (%)	ObsHET (%)	Inter-marker distance (kb)	Max gap (kb)
Holstein	39474	28.2±13	37.2±12	64.45±62.5	2081.4
Brown Swiss	35226	27.7±13	36.6±13	72.26±72.8	2081.4
Simmental	37976	27.5±13	37.0±12	67.06±69.8	2145.7
Australian Angus	44938	24.3±15	32.3±16	56.70±52.4	2081.5
Brahman	45173	16.4±14	23.7±17	56.40±51.3	1677.8
Belmond Red	47416	24.1±15	32.3±16	53.74±47.9	1677.8
Hereford	45322	25.5±15	34.1±16	56.22±52.1	2081.5
Murray Gray	41369	24.4±15	33.3±17	61.52±59.0	2081.5
Santa Gertrudis	46809	23.6±15	31.7±17	54.44±48.9	1677.8
Shorthorns	42280	21.7±15	28.5±16	60.26±56.9	2081.5

We compared the extent of LD among breeds. In order to visualize the decay of LD we plotted r^2 as a function of inter-marker distance (Figure 1). As expected, the level of pair-wise LD as measured by r^2 decreases with marker distance within each breed. The decrease is more or less pronounced across the different breeds up to a rather high average value (0.05) at large distances (>3Mb).

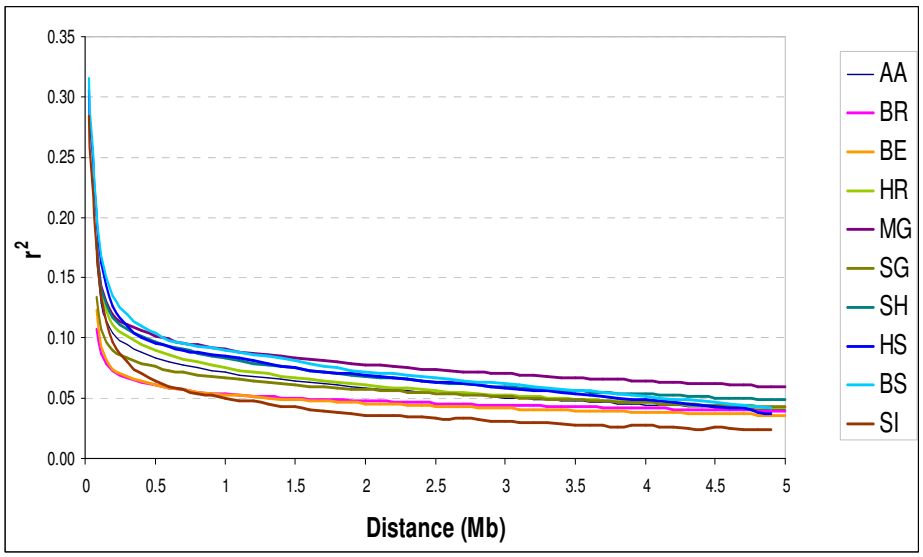

Figure 1. Decay of LD as a function of inter-marker distance in dairy and beef breeds

Signatures of positive selection revealed by |iHS|

To identify genomic regions that may have been targets of recent selection, we calculated |iHS| for each SNP across the genome of the breeds in the first data set. To facilitate comparisons of genomic regions either within dairy and beef groups or across breeds we split the genome into non-overlapping segments of 500 kb and averaged, in each segment, the |iHS| scores over the SNPs located in each window. 500 kb was chosen as the window size so as to have a sufficient number of SNPs in a window. Figure 2 presents the distribution of the average number of SNPs in windows sliding over the genome of breeds in data set I. We chose this length because of the

longer extent of LD in cattle compared to humans, in which the window length used is commonly around 200 Kb [11, 12].

We tested 5099 and 5055 sliding windows in beef and dairy groups respectively, involving a total of 49'559 |iHS| values. The mean |iHS| value was 0.74 and the highest estimated value was 3.41 for a region on chromosome 6 in BS. Across breeds, a total of 109 extreme windows exceeded the |iHS| value 1.96 with 19, 27, 9, 10 and 17 outliers in HS, BS, AA, HR and SI, respectively (Table S1).

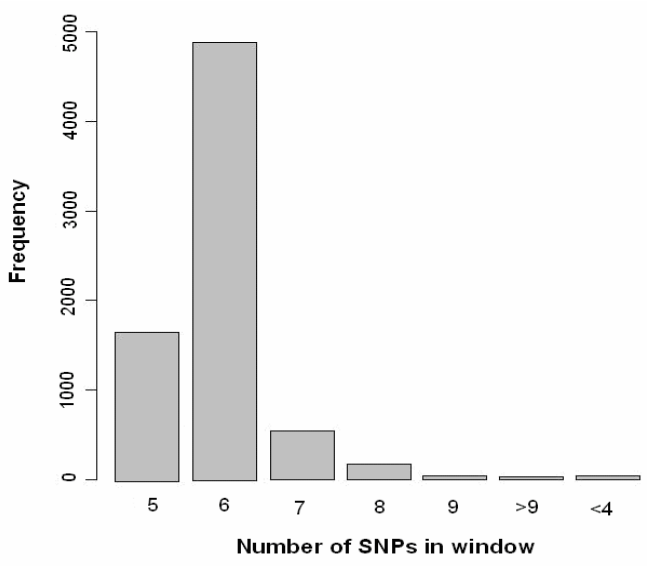

Figure 2. Distribution of the number of SNPs in 500 kb windows sliding over the genome of breeds in data set I.

In order to visualize the chromosomal distribution of outlier signals, we plotted the |iHS| statistic against the genomic position for all breeds (Figures 3 and S1). A panel of clustered signals representing strong evidence for selective sweeps appeared in a number of breeds. Interestingly, a substantial proportion of extreme |iHS| clusters was observed in the telomeric regions of chromosomes, probably due to the strong LD and particular structure of the genome in these regions (Figure 4). Apart from this we found evidence of selective sweeps in two regions in HS and two regions in BS. There were also five distinct clusters of |iHS| signals across the genome of AA and four clusters in HR. The clustered signals also overlapped among breeds in some cases

(Figure 3, S1 and Table 4). The regions with clustered signals reflect high values of LD and a slower decay of haplotype homozygosity for a long stretch around the mutation undergoing selection. It is evident that the signals are non-uniformly distributed across chromosomes and chromosome segments.

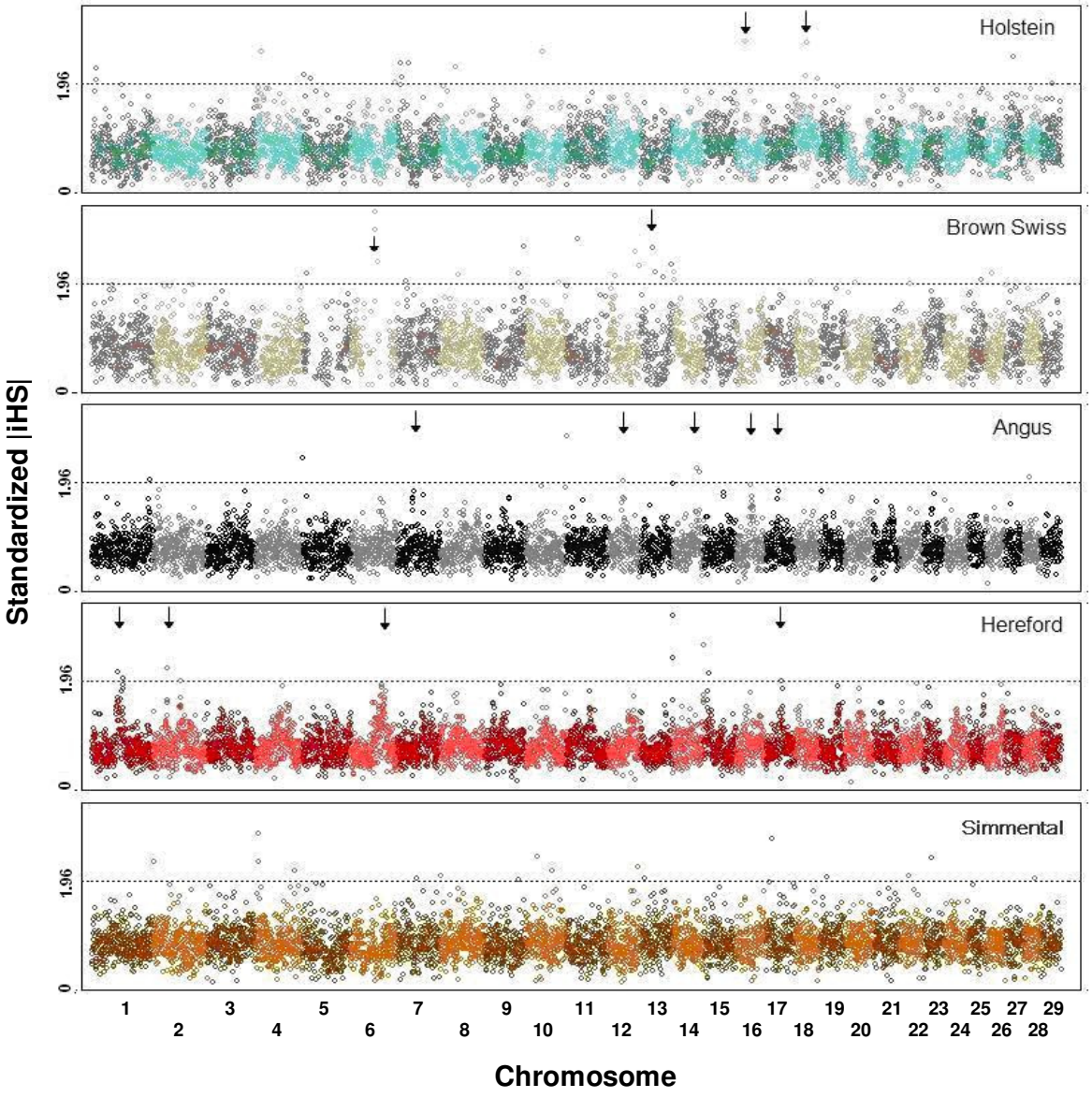

Figure 3. Genome wide distribution of |iHS| values for Holstein and Brown Swiss representing dairy vs. Australian Angus and Hereford representing beef breeds and Simmental being a dual purpose breed. Each dot represents a window of 500Kb and arrows display the clustered signals. Dashed lines are cutting the upper 0.05 of the |iHS| values.

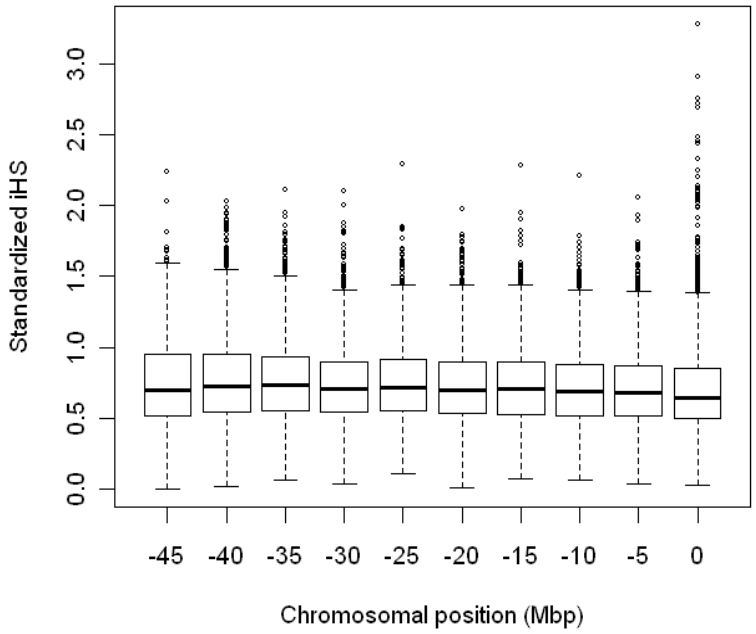

Figure 4. Box plot of standardised |iHS| values found in 5 Mbp bins from the telomere (Position = 0) accumulated over all chromosomes across breeds. More extreme |iHS| values are observed close to the telomeres.

To gain insight into the reliability of our analysis, we compared the |iHS| scores between Angus populations in Australia and Canada and the United States. To this purpose genotypes from 103 North American Angus were used. Because of the smaller sample size and subsequently a larger number of excluded loci (see Material and Methods) only 18'772 SNPs were left for further analyses. Of the total of 12'871 SNPs common between CA and AA, only 107 |iHS| scores overlapped in the 10% upper tail of the empirical distribution, thus basically indicating no major overlap of the regions detected to be under selection.

To assess the background of this result we conducted a cross-validation test [20] regarding the accuracy of |iHS| scores in the Holstein cattle. For this, the Holstein data set was split at random into two data sets, and |iHS| scores calculated from both data sets were found to be in very good agreement (Figure 5). The discordance observed in

the two Angus populations could be due to the sparser inter-marker intervals in the North American Angus which may lead to inefficient estimates of |iHS| scores. However, this difference can also be caused by a different genetic composition of the two populations as well as by different selection pressures in the two environments.

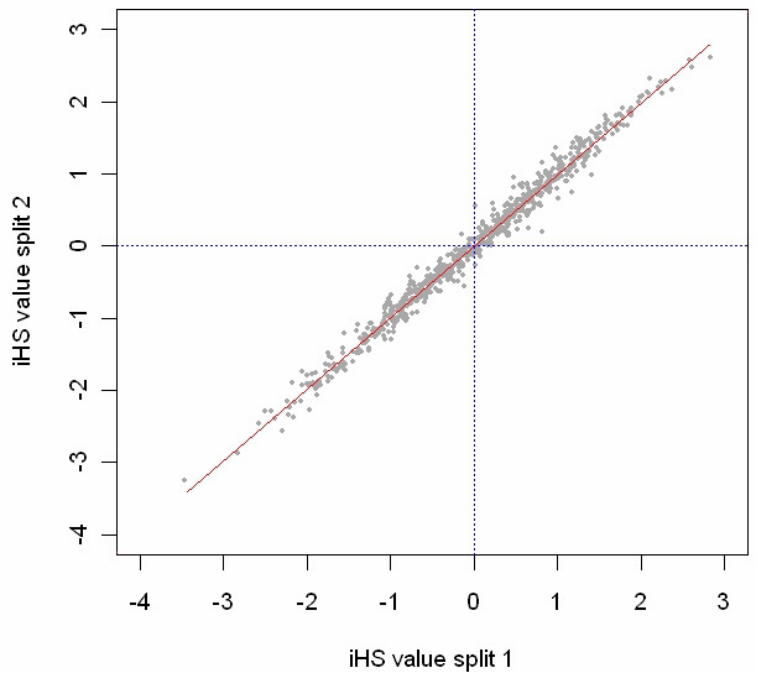

Figure 5. Cross-validation of iHS scores in Holstein data set. The iHS scores from a randomly chosen half data set animals (split 2) are plotted against the other half of the data (split 1).

Exploring the differentiated loci

We then investigated evidence for positive selection by assessing variation in allele frequency among populations, using the new Bayesian method proposed by Gianola et al. (2010) [18]. Data set II was used for this purpose. Several comparisons were made, varying the breeds and the sets of SNPs that were included. Summarized pairwise population comparisons of θ values are shown in Table 3. The θ values varied from 0 to 1, which at the extreme represent identity (F_{ST} = 0) or fixation of alleles in different populations (F_{ST} = 1). The mean posterior distribution of θ values between dairy breeds and between beef breeds respectively, was different from that

between dairy and beef breeds. F_{ST} between HS and CA was estimated as 0.27 ± 0.01 and between CA and PI as 0.02 ± 0.01. Fixation index estimated between two dairy breeds, HS and BS, was 0.05 ± 0.01.

Table 3. Summary statistics of the pair-wise estimates of F_{ST} and clustering information

	HS			BS			SI			CN		
	θ	K[1]	L[2]	θ	K	L	θ	K	L	θ	K	L
BS	0.05	5	4878									
SI	0.04	4	7796	0.04	5	7691						
AN	0.27	3	12106	0.29	4	5571	0.28	3	10882			
PI	0.27	3	19442	0.28	3	18637	0.27	3	8867	0.02	7	2247

[1] Number of clusters
[2] Number of SNPs with largest θ values representing the first cluster of loci

In a further step estimates of θ values (in this case, posterior means) per locus were clustered into groups. The expectation was that these clusters might be representative of different processes taking place in the populations such as balancing or directional selection, neutrality or any other specific process. The structure of clustering was explored by fitting a sequence of finite mixture models to the means of posterior distribution of θ values for each locus. Mixture model parameters were estimated by maximum likelihood via the expectation-maximization algorithm in the FlexMix package [21] in the R project. Results of mixture model analysis, by number of clusters favored by the average information criterion (AIC) and the number of loci representing the first cluster (a fraction of loci with largest θ values) in each comparison, are shown in Table 3. In a breed-by-breed comparison of θ, loci were classified into 3 to 7 clusters, possibly reflecting selection footprints left by different evolutionary forces.

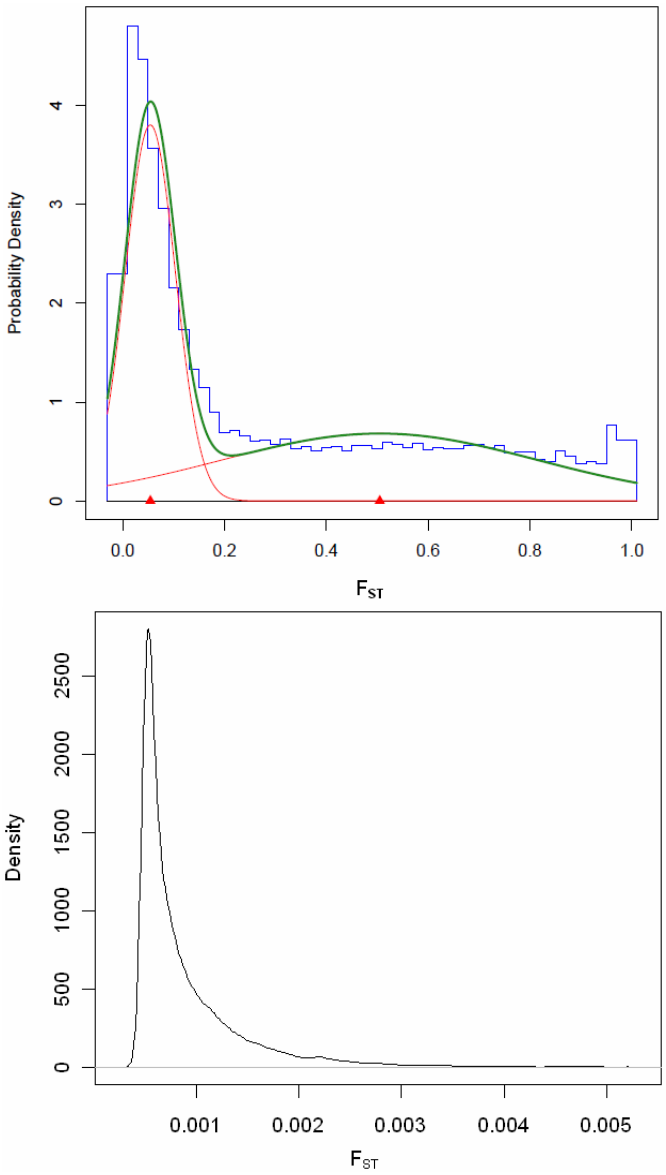

Figure 6. A. Histogram (blue) of the distribution of posterior means over loci of θ values between two dairy (HS and BS) and two beef breeds (CA and PI) and densities of the underlying mixture of two normals (green) and the respective components (red). B. Density plot of 39'474 F_{ST} values between two randomly derived Holstein sub-populations.

To determine if recent selection was responsible for the differences in allele frequencies between dairy and beef breeds, we examined θ among HS and BS versus CA and PI. In total, 4.3% of the posterior θ means among the 4 populations were <

0.01, 27.1% of the θ values were equal to or greater than 0.5, and the average θ was 0.3. Using Akaike's information criterion as a gauge for model comparison, genome-wide estimates of θ were clustered into two groups, one representing 19'471 putatively neutral loci, and another one included 21'124 loci possibly corresponding to genomic regions affected by selection (Figure 6A).

To address this in some further detail, we partitioned the Holstein population randomly into two sub-populations, then estimated F_{ST} and plotted the densities. As shown in Figure 6B, F_{ST} values between two sub-populations of no divergence derived from the same breed resulted only in a unimodal distribution indicating a uniform mode of selection over all evaluated loci.

Signatures of selection can be recognized when adjacent SNPs all show high F_{ST}, due to the hitch-hiking effect, implying divergent selection between breeds, or where adjacent SNPs all show low F_{ST}, implying balancing selection between breeds. Therefore, to facilitate comparisons of genomic regions within or across dairy and beef groups and to reduce locus-to-locus variation in the inference of selection we averaged the F_{ST} values into the non-overlapping windows of 500 kb across the genome. Evidence of the positive selection was assumed for windows in the extreme 2.5 % of the empirical distribution which resulted in 127 significant windows (Table S2).

To identify differentiated windows between dairy and beef genomic background pairwise F_{ST} comparisons denoted as HS–AN, HS–PI, BS–AN and BS–PI were examined and plotted across the genome (Fig. 7). All in all, 29% of the genomic windows with a differentiation index >0.3 overlapped in the four breed comparisons. Bovine chromosome (BTA) 9 with 80 windows covering 0.35 of the chromosome and BTA25 with 23 windows spanning on 0.26 of the chromosome presented the largest and smallest degree of differentiation in the genome. Figure 7 depicts the genome wide map of F_{ST} windows indicating the genomic position of the most diverse regions.

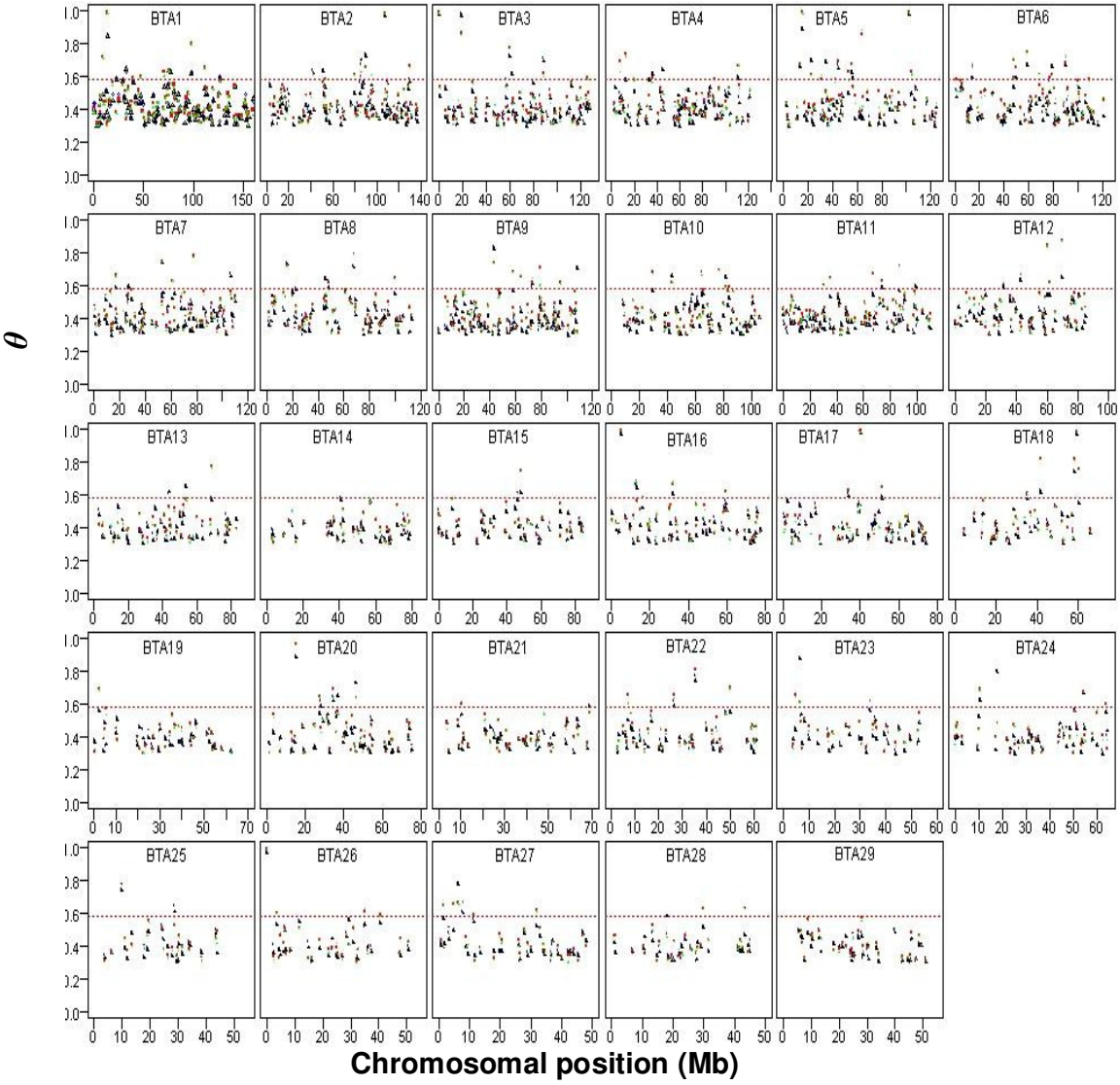

Figure 7. Windows with F_{ST} >0.3 in all pair-wise comparisons, indicating the genomic position of the most diverse regions between dairy and beef breeds. Blue, black, red, and green dots represent F_{ST} values for HS–AN, HS–PI, BS–AN, and BS–PI, respectively, in each window. Dashed lines display the threshold level of 2.5%.

Genomic annotation

We investigated the genomic regions containing extreme |iHS| and F_{ST} values using the fourth draft of bovine genome sequence assembly (Btau 4.0). A subset of genes and ESTs located in each region were identified. We screened this list for the biologically most interesting candidate genes in each region. Table 4 summarizes the

statistic estimated as well as the list of genes for 25 genomic regions presenting the most extreme peaks across breeds. Some regions overlapped with genes previously suggested being under selection. For example on chromosome 18 in the Holstein population, an outlier of |iHS| scores was in the interval 57.25–57.75 Mb. This interval contains Sialic acid binding Ig-like lectin 5 and Zinc finger protein 577 genes which recently were reported as candidates to have a strong effect on productive life and fertility traits in Holstein cattle [22].

Table 4. Summary statistics for windows representing extreme |iHS| and θ

Chr	Position (Mbp)	|iHS| or F_{ST}*	Breed	Gene/EST (n)	Candidate Gene	Function
18	57.25-57.75	2.2, *0.78*	HS	30	SIGLEC5,8,10	Sialic acid binding Ig-like lectin 5, 8, 10
16	19.75-20.25	2.6	HS	2	SPATA17	Spermatogenesis associated 17
6	61.75-62.75	3.41	BS	13	UGDH	UDP-glucose dehydrogenase
					APBB2	Amyloid beta (A4) precursor protein-binding, family B, member 2 (Fe65-like)
13	30.5-31.5	2.68	BS	8	TRDMT1	Cysteine and methionine metabolism
1	79-81.5	2.10	HR	6	SST	Somatostatin
2	34.5-36	2.26	HR	6	GCG	Glucagon
					FAP	Fibroblast activation protein, alpha
6	80-83		HR	9	SRD5A2L2	Lipid metabolism
7	39-41	1.9	AN	15	COL23A1	Collagen, type XXIII, alpha 1
					MGAT1	Fertilization and early development of the embryos
12	36-38	2.03	AN	19	ATP12A	ATPase activity
14	64-65	2.02	AN	6	MATN2	Developing cartilage rudiments
16	39-40	1.98	AN	14	NMNAT1	Methylenetetrahydrofolate reductase (NADPH) activity
17	31-32.5	2.05	AN/HR	15	PGRMC2	Progesterone receptor membrane component 2
2	70-73	2.06	MG/BE/ SH/BR	5	-	-
10	29-31	2.24	BE/SH	8	ACTC1	Actinin, Involved in the formation of filaments
1	12-13	0.92	-	0	-	
2	111.5-112	0.98	-	11	ABCB6	ATP-binding cassette, sub-family B (MDR/TAP), member 6
					GLB1L	Galactosidase, beta 1-like
3	119.2-119.7	*0.92*	-	11	SMCP	Sperm mitochondria-associated cysteine-rich protein
7	53.25-53.75	*0.74*	-	4	FGF1	A growth factor which stimulates growth or differentiation, key role in embryonic development
9	42-43	*0.78*	-	12	LACE1	Lactation elevated 1
					PPIL6	Peptidylprolyl isomerase (cyclophilin)-like 6
13	53.5-54	*0.98*	-	7	SIRPA	Signal-regulatory protein
16	4.75-5.25	*098*	-	5	-	-
17	39.5-40.5	*0.98*	-	4	-	-
18	58.25-58.75	*0.98*	-	15	-	-
20	15.25-15.75	*0.92*	-	8	ADAMTS6	-
22	35.25-35.75	*0.77*	-	3	-	-

* F_{ST} values are in italic

The window with the largest |iHS| value (3.41) was observed in BS spanning 61.75–62.75 Mb on chromosome 6. Of the 13 genes/ESTs in this region, UGDH (which acts in the carbohydrate metabolism pathways) may be a possible candidate to affect feed efficiency traits. Another strong |iHS| cluster which harbors the Somatostatin (SST) gene was observed on chromosome 1 in HR. Strong evidence of a sweep reflected by a set of windows was observed in the region 80-83 Mb of BTA6 in the vicinity of the SRD5A2 gene. The enzyme steroid 5-alpha-reductase converts testosterone into dihydrotestosterone and a polymorphism in this gene was shown to moderately increase the proportion of progressively motile spermatozoa in normozoospermic men [23]. We also found four clusters of outliers on BTA16 and BTA17, BTA2 and BTA10 which overlapped among some beef breeds.

DISCUSSION

The high level of observed phenotypic variation among domestic cattle is a result of both neutral demographic processes, weak but sustained natural selection and strong short-term artificial selection for divergent breeding goals. The task of separating these processes and identifying genes under the influence of artificial selection can be challenging. The efforts to identify genes affected by selection have so far been concentrated on species with well-characterized genomes, such as *Drosophila* and humans [11, 24]. The cattle genome offers an excellent opportunity to test the power of genome-wide analyses, as it has extensive LD [2, 25] caused by intensive selection, and it is expected that selection footprints would be correlated with genes affecting production traits or fitness.

In this study we presented an application of two complementary statistics of selection signatures in a diverse set of dairy and beef breeds. In the first step, regions of the genome that contained targets of putative positive selection revealed by long range LD were defined as windows in the extreme of the empirical distribution of the |iHS| statistic. This criterion resulted in 109 significant windows ($P \leq 0.05$). These signals generally differ from those reported by the Bovine HapMap consortium [2]. This is

probably due to the differences in sample size and marker densities between studies which both could limit accurate estimates of |iHS|. Mapping the corresponding genomic regions to the cattle genome sequence resulted in a large number of adjacent loci. The list of genes with signatures of positive selection was significantly enlarged by those involved in the biological processes such as anatomical structure development, muscle development, metabolism of carbohydrates and lipids, spermatogenesis and fertilization. We refined the complete list for the most important genes in the region of clustered signals that may have functional relevance for economic traits. A remarkable observation in this study is a strong selection signal confirmed by both |iHS| and F_{ST} analyses in the vicinity of Sialic acid binding Ig-like lectin 5 gene on BTA18. This QTL was recently reported to have large effects on calving ease, several conformation traits, longevity, and total merit in Holstein cattle [22]. We observed that other haplotypes present in this region display a shorter extent of homozygosity, indicating abundant historical recombination (Figure 8). Therefore, the long stretch of homozygosity observed in this region presumably is not simply due to a low local recombination rate but presumably reflects the combination of strong and recent selective pressure, pushing the beneficial mutation rapidly towards high frequency with a long conserved haplotype surrounding it. Although the low heritability of most of the aforementioned traits has not made them a primary breeding goal in selection programs, it could be hypothesized that applying sustained but weak negative selection against these traits has increased the frequency of favorable alleles and surrounding haplotypes in the Holstein population.

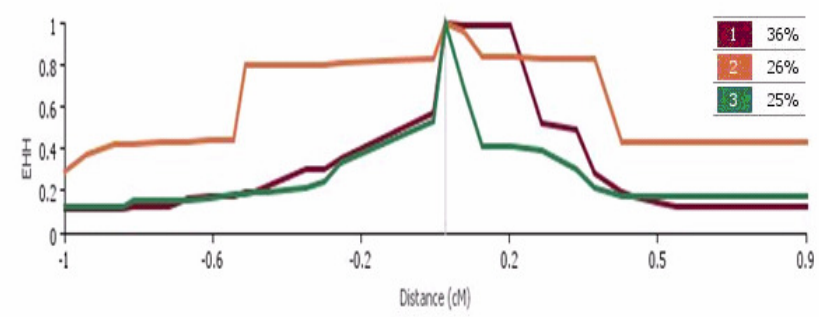

Figure 8. Frequencies of the haplotypes segregating in the region of extreme |iHS| in the interval 57.25–57.75 Mb on BTA18 in Holstein cattle. The extent of haplotype homozygosity was estimated by Sweep v.1.1 [11].

A cluster of signals reflecting strong evidence of selection was also observed in the vicinity of the Somatostatin gene on BTA1. We also found clusters of outliers which overlapped among some beef breeds (Table S1). These results show significant enrichment for genes such as SPATA17, MGAT1, PGRMC2 and SRD5A2 in the region of clustered signals which belong to a number of functional categories relevant to reproduction, including gamete generation, embryo development and spermatogenesis, and genes in these categories may be strong candidates for selection for fertility traits. These results generally are consistent with the observations of Flori et al. (2009) [3]. Another interesting observation was the strong evidence for selection in the region of genes related to muscle formation (e.g., ACTC1, COL23A1, MATN2, and FAP) in beef breeds. For example polymorphisms in the genes encoding Actinin are among the best characterized athletic-performance associated variants in human endurance athletes [26, 27]. Evidence for positive selection in the genomic region surrounding muscle related genes has also been reported in racing horses [28] and humans [29]. The presence of genes like Actinin, Collagen and Fibroblast activation protein as well as the gene responsible for developing cartilage rudiments in positively selected regions in beef cattle (Table 4) supports the supposition that selection for muscle related phenotypes has played a major role in the shaping the beef cattle. A better understanding of the role these genes play in the development, strength and integrity of muscles may contribute to improved knowledge of musculoskeletal traits and developing new marker systems for beef cattle breeding. Consistency of our observations with previous reports [3, 28] may suggest general themes about the types of genes that have been targets of positive selection in cattle.

We also optimized a new Bayesian approach for exploring the level of genetic differentiation to infer the selection signatures against the genome as a whole. This algorithm is able to deal with a large battery of marker information via probabilistic clustering of F_{ST} values. After examining F_{ST} among HS and BS versus CA and PI breeds using Akaike's information criterion it appears likely that genome-wide estimates of F_{ST} are clustered into two groups, one representing putatively neutral loci, and another one (possibly) corresponding to genomic regions affected by

selection. Annotation of the genes underlying the regions with extreme F_{ST} does not appear to reveal many strong candidates for positive selection with the possible exception of the SMCP and FGF1 genes (Table 4). A receptor of the latter gene (FGFR3) showed evidence of selection in a genome-wide sweep mapping study using F_{ST} among dog breeds [30]. This gene is responsible for achondroplasia (shortened limbs) in humans. As an explanation we suggest that selection may work on genes that were not considered the primary targets of selection so far. Some extreme peaks were observed in presumed gene deserts which may reflect selection acting on uncharacterized regulatory regions or simply fixation of non-coding DNA by genetic drift.

We found that 56 of the 127 significant F_{ST} values lie in poor gene content regions, defined by the frequency of coding sequences in the bracket of 1Mb surrounding the F_{ST} signal. To test whether this observation is a systematic deviation from the expected, we sampled 127 random positions with matching frequencies on the chromosomes, i.e. since 11 significant F_{ST} values were observed on BTA1, we also sampled 11 random positions on that chromosome. Of these 127 random positions, 35 were positioned in regions with poor gene content applying the same definition. The difference was tested with a χ^2 test revealing a significant difference on the 5 per cent error level.

This observation is consistent with the studies of Flori et al. (2009) [3], and Gu et al. (2009) [28] which reported F_{ST} signals in poor gene content regions in genome wide analyses of cattle and thoroughbred horse, respectively. Thus, these results in combination with the observations from Voight et al. (2006) [12], Carlson et al. (2005) [31] and Wang et al. (2006) [32] on human population data suggest that non-coding regions may have been important for adaptive evolution.

We examined the validity of F_{ST} analysis by testing some candidate major genes in our data set. The results revealed F_{ST} values larger than expected ($P < 10\%$) for regions harboring the Casein cluster, GHR, STS, LP, IGF-1 and MSTN genes which are supposed to be targets of artificial selection. The observation of selection evidence

in the region of the GHR gene on BTA20 is consistent with the reports of Flori et al. (2009) [3] and Hayes et al. (2009) [4], the latter based on a study comparing Angus and Holstein. The presence of the longer than expected haplotype homozygosity in this region was also observed in Holstein cattle [6]. 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 [2], also matched with the outlier F_{ST} windows in our study.

Overall, the average F_{ST} of dairy vs. beef breeds was equal to 0.3 which is substantially higher than the differentiation index reported previously between Holstein and Angus [2, 7]. The higher average of F_{ST} as well as the similar pair-wise F_{ST} within dairy and beef breeds might reflect the dominating influence of a substantial number of fixed SNPs in the pair-wise comparisons of breeds and groups.

The two metrics applied yielded a total of 236 regions putatively subject to positive selection. To investigate how frequently selective events were unique or shared between methods, we assessed the number of overlapping signals. A panel of 6 significant signals was overlapping (Table 5). Interestingly, most of these were found in Holstein cattle, which may reflect a comparatively higher pressure of selective breeding in this breed.

Table 5: Overlapping signals revealed by both |iHS| and F_{ST} metrics.

Chr	Position (Mbp)	Breed	F_{ST}	|iHS|
4	12.5	HS	0.67	2.62
8	40.5	HS	0.59	2.33
10	30	SI	0.64	2.48
10	43.5	HS	0.64	2.63
18	58	HS	0.78	2.12
22	26	BS	0.63	1.99

Overall, comparing our scan for selection with the results of previous genome-wide studies revealed a modest overlap with some notable exceptions. Different hypotheses can be proposed to explain these incongruities. From the methodological point of view, a possible reason could be due to the differences in the computational analyses between the studies. In other words, the statistical tests used in each study are recovering selective events from different time periods and/or for different stages of the selective sweep. Even for tests that should detect similar types of selective events (e.g., scans that identify unusually long haplotypes), low statistical power further decreases the probability of overlap [14]. In addition, most studies report only the most significant results (i.e. outliers in the 1% empirical distribution). Therefore, the results presented in this study are probably a conservative estimate of overlap between studies.

Population demographic history can also impart similar patterns on DNA sequence variation, making inferences on selection difficult. For example, population expansion can lead to an excess of low frequency alleles compared with the number expected under the standard neutral model. Likewise, recent positive selection for a putative mutation may have started from a higher initial frequency of beneficial alleles [33]. Such an allele might e.g. be imported into a breed through crosses with other breeds. In such a case beneficial alleles may be included in diverse haplotypes and LD based estimators would not be able to trace the selection signature. Crossbreeding can also generate false selection signatures, if e.g. a large conserved piece of a chromosome from another breed is mixed with many shorter segments from the original breed.

From the technical point of view, the density of the markers is also critical for the power of such studies and could be a source of discrepancy. It was shown earlier with LD based analyses that core regions are more likely to appear where the marker density is greater than the average [6]. This would imply that the availability of genotyping arrays with an increased genome-wide marker density (by a factor >10) will allow a more reliable and comprehensive screening of the genome for signatures of selection by LD based tests. Moreover, although sliding window analyses facilitate inferences of selection by reducing locus-to-locus variation, the size of the window is

often subjectively determined which can influence the final results and interpretations. One potential refinement would be to adjust window sizes to local levels of LD [34], although it remains unclear how to account for varying levels of LD between populations. Finally, the incongruities can also result from a lack of power given the sample size available for some of the breeds in this study, and complex genomic interactions.

CONCLUSIONS

In this study genomic scans based on site frequency and haplotype data led to the detection of 236 regions putatively subject to recent positive selection in the cattle genome. Our results confirmed the higher differentiation index as well as the longer than expected haplotype consistency in the vicinity of Sialic acid binding Ig-like lectin 5 gene on BTA18, which was recently reported as a strong QTL in the Holstein cattle [22]. However, the overlap between the identified regions via |iHS| with previous studies is modest. Analysis of population differentiation revealed signatures of selection occurring in regions of the genome thought to be nonfunctional, which may reflect selection acting on uncharacterized regulatory regions or simply fixation of non-coding DNA by genetic drift due to the absence of any selection. Clearly, many challenges remain, including the development of efficient methods of differentiating the effects of drift and selection, identifying the causal gene driving the signature of selection observed across large genomic regions, and functionally characterizing the suspected targets of selection. Independent confirmation studies with larger sample sizes and/or SNP densities are required. Our results may be of future interest for identifying signatures of recent positive artificial selection between the cattle breeds or as additional evidence for any polymorphisms that show associations with beef or milk traits.

Abbreviations

iHS: integrated Haplotype Homozygosity Score, EHH: Extended Haplotype Homozygosity, F_{ST}: Population fixation index, HS: Holstein, BS: Brown Swiss, SI: Simmental, CA: North American Angus, PI: Piedmontese, AA: Australian Angus,

HR: Hereford, SH: Shorthorns, BR: Brahman, BE: Belmond red, MG: Murray Gray, SG: Santa Gertrudis

Authors' contributions

SQ carried out the data analyses, drafted and prepared the manuscript for submission. HS supervised the study and contributed in revising and editing the manuscript. BH coordinated in data analysis, provision of data and writing support. DG and FS coordinated in the interpretation of data as well as critically revising the manuscript. GT, SSM and SPM participated in provision of study material and manuscript improvement and also provided administrative support. All authors read and approved the manuscript.

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. The authors thank the Cooperative Research Centre for Beef Genetic Technologies for supplying some of the data used here. The Canadian genotypes were made available through funding from the Ontario Ministry of Agriculture Food and Rural Affairs, the Ontario Cattlemen's Association and the Agriculture Adaptation Council.

References

Loftus RT, MacHugh DE, Bradley DG, Sharp PM, Cunningham P: Evidence for two independent domestications of cattle. Proc. Natl. Acad. Sci. USA 1994, 91:2757–2761.

The Bovine HapMap consortium: Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 2009, 324:528-532.

Flori L, Fritz S, Jaffrezic F, Boussaha M, Gut I, Heath S, Foulley JL, Gautier M: The genome response to artificial selection: a case study in dairy cattle. PLoS One 2009, 4: e6595.

Hayes BJ, Chamberlain AJ, Maceachern S, Savin K, McPartlan H, MacLeod I, Sethuraman L, Goddard ME: A genome map of divergent artificial selection between Bos Taurus dairy cattle and Bos taurus beef cattle. Animal Genetics 2009, 40:176-184.

Hayes BJ, Lien S, Nilsen H, Olsen HG, Berg P, Maceachern S, Potter S, Meuwissen TH: The origin of selection signatures on bovine chromosome 6. Animal Genetics 2008, 39:105-111.

Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H: A Genome-Wide Scan for Signatures of Recent Selection in Holstein Cattle. Animal Genetics 2010a, (in press).

MacEachern S**,** Hayes B**,** McEwan J, Goddard M: An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics 2009,10**:**181.

Wright S: The genetical structure of populations. Annals of Eugenics 1951, 15:323-54.

Cockerham CC: Variance of gene frequencies. Evolution 1969, 23:72–84.

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD: Interrogating a high-density SNP map for signatures of natural selection. Genome Research 2002, 12:1805-1814.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES: Detecting recent positive selection in the human genome from haplotype structure. Nature 2002. 419:832–837.

Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. PLoS Biology 2006, 4:e72.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E. *et al:* Genome-wide detection and characterization of positive selection in human populations. Nature 2007, 449:913-918.

Biwas S, Akey JM: Genomic insights into positive selection. Trends in Genetics 2006, 22:437-446.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS, Van Tassell CP: Development and Characterization of a High Density SNP Genotyping Assay for Cattle. 2009, PLoS ONE 4:e5350.

Scheet P, Stephens M: A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 2006, 78:629–644.

Barrett JC, Fry B, Maller J, Daly MJ: HaploView: Analysis and visualization of LD and haplotype maps. Bioinformatics 2005, 21: 263–265.

Gianola D, Simianer H, Qanbari S: A two-step method for detecting selection signatures using genetic markers. Genetics Research 2010, (in press).

Haldane JBS: The precision of observed values of small frequencies. Biometrika 1948, 35:297-303.

Whittaker J.C, Haley C, Thompson R: Weighting of information in marker-assisted selection. Genetical Research 1997, 69:137–44.

Leisch F: FlexMix: A general framework for finite mixture models and latent class regression in R. Journal of Statistical Software 2004, 11:1-18, URL http://www.jstatsoft.org/v11/i08/.

Cole JB, VanRaden PM, O'Connell JR, Van Tassell CP, Sonstegard TS, Schnabel RD, Taylor JF, Wiggans GR: Distribution and location of genetic effects for dairy traits. Journal of Dairy Sci. 2009, 92:2931–2946.

Peters M, Saare M, Kaart T, Haller-Kikkatalo K, Lend AK, Punab M, Metspalu A, Salumets A: Analysis of Polymorphisms in the SRD5A2 Gene and Semen Parameters in Estonian Men. Journal of Andrology 2009 (in press).

Wall JD, Andolfatto P, Przeworski M: Testing models of selection and demography in *Drosophila simulans.* Genetics 2002, 162:203–216.

Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H: The pattern of linkage disequilbriom in German Holstein cattle. Animal Genetics 2010b (in press).

Yang N, MacArthur DG, Gulbin JP, Hahn AG, Beggs AH, Easteal S, North K: ACTN3 genotype is associated with human elite athletic performance. American Journal of Human Genetics 2003, 73:627–631.

Chan S, Seto JT, MacArthur DG, Yang N, North KN, et al: A gene for speed: contractile properties of isolated whole EDL muscle from an alpha-actinin-3 knockout mouse. American Journal of Physiology- Cell Physiology 2008, 295:C897–904.

Gu J, Orr N, Park SD, Katz LM, Sulimova G, et al: A Genome Scan for Positive Selection in Thoroughbred Horses. PLoS ONE 2009, 4:e5767.

MacArthur DG, Seto JT, Raftery JM, Quinlan KG, Huttley GA, et al: Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. Nature Genetics 2007, 39:1261–1265.

Pollinger JP, Bustamante CD, Adi Fledel-Alon A, Schmutz SM, Gray MM, Wayne RK: Selective sweep mapping of genes with large phenotypic effects. Genome Research 2005, 15:1809–1819.

Carlson CS, Thomas DJ, Eberle MA, Swanson JA, Livingston RJ, Rieder MJ, Nickerson D: Genomic regions exhibiting positive selection identified from dense genotype data. Genome Research 2005, 15:1553–1565.

Wang ET, Kodama G, Baldi P, Moyzis RK: Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceeding of National Academy of Science of USA 2006, 103:135–140.

Innan H, Kim Y: Pattern of polymorphism after strong artificial selection in a domestication event. Proc Natl Acad Sci USA 2004, 101:10667–10672.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG: Measures of human population structure show heterogeneity among genomic regions. Genome Research 2005, 15:1468–1476.

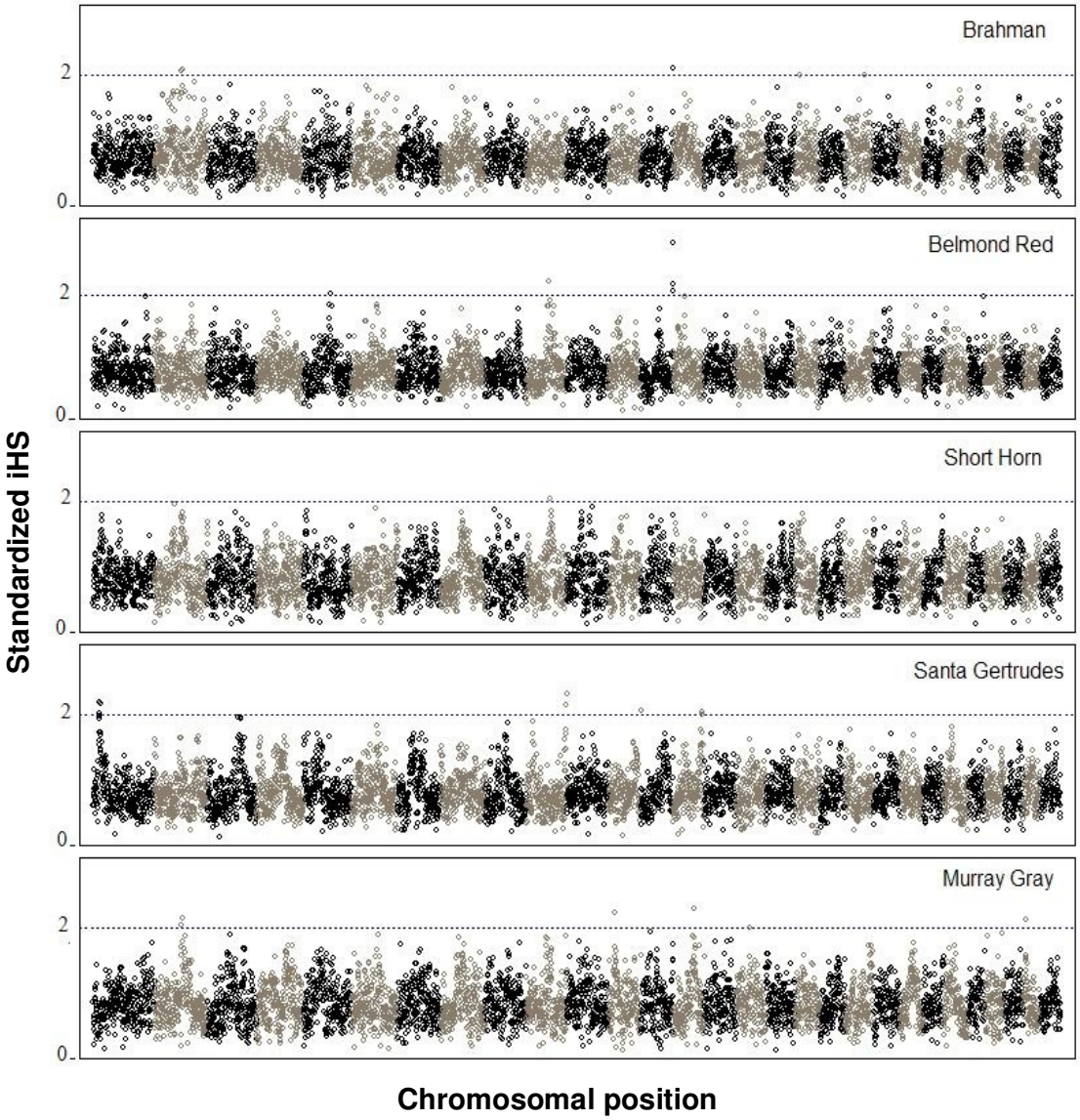

Figure S1. Genome wide distribution of |iHS| values across the genome of beef breeds. Dashed lines display the threshold level of 0.05.

Table S1. Genomic regions associated with extreme |iHS| values. |iHS| values averaged over non-overlapping windows of each 500kb.

Position (Mb)	Chr	AN	BR	BE	HE	MG	SG	SH	HS	SI	BS
105.5	10	2.91	1.09	1.24	1.01	1.30	2.33	0.52	NA	1.16	1.24
0.25	5	2.48	0.82	0.98	0.68	0.49	0.89	0.61	0.70	0.50	0.88
0.25	5	2.48	0.82	0.98	0.68	0.49	0.89	0.61	NA	NA	0.52
63.5	14	2.29	0.75	0.49	0.53	0.32	0.46	0.45	0.49	1.27	1.31
69	14	2.21	0.37	0.66	0.82	0.93	1.03	0.54	0.47	0.48	0.63
13.5	28	2.10	0.25	0.63	0.26	0.29	1.25	0.07	0.74	0.66	0.12
151	1	2.06	0.51	0.45	0.99	0.41	0.92	0.75	1.43	1.24	1.09
38.5	12	2.03	1.17	0.77	0.44	0.53	0.60	0.24	NA	NA	NA
83.5	13	0.67	2.10	2.08	2.46	1.73	1.59	1.77	0.96	0.67	0.91
74	2	0.53	2.07	1.15	0.95	0.83	0.86	1.01	0.75	0.52	0.73
71	2	0.99	2.05	0.73	1.50	1.67	1.60	1.73	1.01	1.04	0.37
84	13	1.98	1.39	2.91	3.28	0.88	1.30	0.63	0.87	0.73	0.57
59	10	0.76	1.06	2.24	1.01	0.68	0.58	0.59	0.31	0.50	0.78
83	13	0.55	0.90	2.20	1.25	1.45	0.97	1.28	0.65	1.19	1.26
69.5	5	1.05	1.12	2.03	0.90	0.99	0.58	0.78	0.78	1.35	NA
84	13	1.98	1.39	2.91	3.28	0.88	1.30	0.63	0.87	0.73	0.57
81.5	14	1.24	0.70	1.38	2.72	1.53	0.45	1.77	NA	NA	NA
83.5	13	0.67	2.10	2.08	2.46	1.73	1.59	1.77	0.96	0.67	0.91
35	2	0.67	0.85	0.59	2.26	0.98	0.53	0.60	0.65	0.70	0.61
68	1	0.93	0.50	1.22	2.19	0.74	0.62	1.25	0.70	0.88	0.58
11.5	15	0.74	0.58	0.96	2.16	0.84	0.51	0.93	0.81	0.73	0.59
81	1	0.35	0.60	1.07	2.06	0.50	1.32	0.76	0.67	1.19	0.20
81.5	1	0.65	0.99	1.08	2.06	0.46	1.13	0.57	0.35	1.41	0.92
71.5	2	0.98	0.60	0.97	2.02	1.55	0.81	1.48	0.46	0.70	0.69
41.5	17	0.46	0.47	0.75	2.00	0.86	0.98	1.40	1.09	0.85	0.03
54.5	14	0.60	0.31	0.84	0.57	2.29	0.71	0.47	1.19	0.89	NA
17	12	0.86	1.06	0.73	0.67	2.22	0.44	0.25	1.36	0.81	1.00
75	2	1.18	0.19	0.64	0.61	2.14	0.99	1.55	0.62	1.00	0.27
6	28	0.91	0.63	1.39	0.47	2.11	0.77	1.08	0.69	0.41	0.80
70.5	2	0.46	0.63	0.99	1.68	2.03	0.70	1.53	0.65	0.65	0.35
105.5	10	2.91	1.09	1.24	1.01	1.30	2.33	0.52	NA	1.16	1.24
17	1	0.77	1.29	0.77	1.36	0.76	2.18	1.05	0.58	1.02	1.31
19.5	1	1.17	0.48	0.36	0.27	0.70	2.16	1.38	0.80	0.16	1.41
105	10	1.91	1.27	0.71	1.77	1.86	2.14	1.08	0.47	0.54	0.68
85.5	12	0.89	NA	0.03	1.21	1.54	2.06	NA	NA	NA	NA
77	14	1.00	0.34	0.88	0.86	0.75	2.04	0.44	0.92	0.60	1.35
16.5	1	0.53	0.66	0.79	0.71	0.73	2.00	1.00	0.72	0.77	1.15
105.5	10	2.91	1.09	1.24	1.01	1.30	2.33	0.52	NA	1.16	1.24
17	1	0.77	1.29	0.77	1.36	0.76	2.18	1.05	0.58	1.02	1.31
19.5	1	1.17	0.48	0.36	0.27	0.70	2.16	1.38	0.80	0.16	1.41
105	10	1.91	1.27	0.71	1.77	1.86	2.14	1.08	0.47	0.54	0.68
85.5	12	0.89	NA	0.03	1.21	1.54	2.06	NA	NA	NA	NA
77	14	1.00	0.34	0.88	0.86	0.75	2.04	0.44	0.92	0.60	1.35
16.5	1	0.53	0.66	0.79	0.71	0.73	2.00	1.00	0.72	0.77	1.15
61	10	0.29	0.79	1.45	0.93	1.19	0.54	2.03	0.60	0.29	0.67
21	16	0.23	0.53	0.62	0.61	0.39	0.17	1.03	2.83	0.80	NA
30	18	0.75	0.32	0.27	1.03	0.55	1.04	0.61	2.81	0.30	NA
43.5	10	0.21	0.48	0.49	0.72	0.50	0.29	0.39	2.63	0.86	0.63
12.5	4	0.55	0.62	0.68	0.50	0.91	0.82	0.36	2.62	1.88	NA
21	27	0.41	0.15	0.60	1.41	0.16	0.34	0.47	2.54	0.93	0.70

Position (Mb)	Chr	AN	BR	BE	HE	MG	SG	SH	HS	SI	BS
29.5	7	0.68	0.69	0.82	1.00	0.75	0.90	0.86	2.41	1.60	0.44
9	7	0.48	1.06	0.36	0.34	0.86	0.98	0.39	2.40	0.37	0.06
40.5	8	0.26	0.55	0.41	0.37	1.18	0.44	0.38	2.33	0.82	NA
11	1	0.70	0.56	0.83	0.38	0.22	0.55	0.40	2.30	0.47	NA
3	5	0.42	0.42	0.39	0.78	0.32	1.56	1.01	2.17	1.25	1.28
25	18	0.19	1.13	0.43	0.28	0.36	0.65	0.19	2.16	NA	NA
12	7	0.43	0.30	0.35	0.55	0.12	0.87	0.52	2.13	0.56	0.76
58	18	0.62	0.82	1.01	0.84	0.87	0.63	0.56	2.12	0.05	NA
21	5	0.31	0.33	0.76	0.93	0.64	0.37	0.69	2.10	1.34	NA
10.5	1	0.60	0.85	0.55	0.64	0.66	0.82	0.81	2.08	0.46	0.43
116.5	6	0.48	0.56	0.50	0.66	0.39	0.12	0.99	2.04	1.54	1.04
29	29	0.77	0.23	0.49	0.93	0.87	0.52	0.60	2.01	0.77	NA
6	4	0.72	0.47	0.21	0.39	0.22	0.83	0.61	1.83	2.92	NA
14	17	1.19	0.33	0.62	0.54	0.73	0.73	0.47	1.05	2.83	NA
30	10	0.70	0.40	0.48	0.59	0.78	0.34	0.98	0.56	2.48	NA
21.5	23	0.73	0.64	0.40	0.43	0.61	0.56	0.65	0.60	2.45	1.84
6.5	4	0.54	0.34	0.20	0.46	0.12	1.03	0.48	1.94	2.38	0.94
2.5	2	0.33	0.25	0.97	0.29	0.37	0.85	0.03	0.14	2.37	0.17
78	12	0.45	0.85	0.87	NA	NA	0.42	1.07	1.48	2.27	1.36
101	4	0.83	0.38	1.09	0.40	0.68	0.18	1.32	1.26	2.21	NA
67	10	0.58	0.34	0.51	0.72	1.14	0.32	0.78	0.20	2.19	NA
23.5	22	0.84	0.19	0.47	0.46	0.43	0.51	0.42	0.27	2.11	0.10
2	8	0.61	0.09	0.36	0.48	0.41	0.63	0.29	1.57	2.10	1.16
16.5	19	0.85	0.49	0.50	0.84	0.44	0.54	0.65	0.80	2.08	1.53
10.5	13	1.02	0.34	0.44	0.54	0.79	0.57	0.37	1.62	2.05	NA
28	28	0.69	0.84	0.65	1.30	0.84	0.45	0.37	1.56	2.04	1.33
48.5	7	0.38	0.67	0.87	0.59	0.60	0.38	0.19	NA	2.04	NA
87	9	0.34	1.00	1.44	0.49	1.20	0.84	0.59	0.66	2.02	0.44
62	6	0.77	0.40	0.63	1.27	1.23	0.90	0.95	0.57	0.56	3.41
62.5	6	0.58	0.65	1.02	1.12	1.52	0.79	1.87	0.44	0.46	3.05
29.5	11	0.66	0.94	0.65	0.78	0.65	0.80	1.05	0.73	0.86	2.86
103	9	0.79	0.38	0.45	0.69	0.52	0.48	0.59	1.22	0.80	2.71
30.5	13	0.87	0.41	0.50	0.70	0.90	0.67	0.69	0.36	1.33	2.68
70.5	12	0.65	0.30	0.73	0.88	0.96	1.06	0.70	0.58	0.78	2.62
71.5	6	1.09	0.98	0.78	1.28	1.26	0.95	0.28	0.51	NA	2.43
79	13	0.62	1.04	0.97	0.80	0.31	0.46	0.45	1.17	1.17	2.36
79.5	12	0.72	0.65	0.63	0.85	0.51	0.55	1.04	1.05	1.27	2.33
0.5	14	0.39	0.76	0.75	0.21	0.26	0.26	0.54	1.10	0.75	2.22
36.5	13	0.55	0.78	0.61	0.52	0.67	0.42	1.02	1.31	1.58	2.22
10	5	1.72	0.45	0.26	0.70	0.44	0.27	0.67	0.59	1.83	2.20
86.5	9	0.46	0.59	1.08	0.48	0.96	0.53	0.59	0.49	0.56	2.20
14.5	26	0.77	0.73	0.92	0.71	0.46	0.81	0.39	0.58	1.04	2.18
45.5	8	0.55	0.70	1.10	0.98	1.19	0.98	0.63	0.79	1.84	2.17
16	12	0.47	0.73	0.63	1.00	1.23	0.85	0.74	1.26	1.31	2.13
57.5	13	0.53	0.47	1.13	0.62	0.73	1.30	0.23	0.74	0.73	2.12
30.5	25	0.97	0.47	0.60	0.77	1.15	1.13	0.71	1.14	0.87	2.07
2	14	0.59	0.88	0.58	0.67	0.64	0.40	0.57	0.68	1.57	2.06
39	18	0.91	0.66	0.49	0.80	0.39	0.97	0.58	1.55	0.99	2.05
23	7	0.35	0.58	0.50	0.31	0.38	0.68	0.59	0.65	0.98	2.03
24.5	20	0.82	0.90	1.13	0.91	0.83	0.40	0.62	0.70	1.05	2.02
85	10	0.31	0.87	0.37	0.49	0.44	0.41	1.08	0.85	0.93	2.01

Table S2. Genomic regions associated with extreme θ values (P < 2.5%). θs averaged over non-overlapping windows of each 500kb.

Chr	Position (Mb)	θ
1	13.5	0.92
1	98	0.72
1	47	0.63
1	75.5	0.63
1	19	0.62
1	9	0.62
1	111.5	0.6
1	32	0.6
1	77	0.6
1	70.5	0.59
1	39.5	0.58
2	107	0.98
2	90	0.68
2	85.5	0.65
2	42.5	0.64
2	130	0.63
2	51	0.63
3	19	0.92
3	60.5	0.76
3	88.5	0.66
3	6	0.61
3	62	0.59
4	12.5	0.67
4	111.5	0.64
4	36	0.63
4	8	0.59
4	45.5	0.57
5	102.5	0.99
5	16	0.94
5	53	0.68
5	44.5	0.68
5	22.5	0.67
5	14	0.67
5	64	0.66
5	56	0.65
5	34.5	0.61
5	104.5	0.58
6	90	0.74
6	58.5	0.72
6	79	0.64
6	14.5	0.64
6	47.5	0.63
6	50	0.63
6	68.5	0.6
6	5	0.58
7	53.5	0.74
7	17	0.65
7	106.5	0.62

Chr	Position (Mb)	θ
7	77.5	0.62
7	61.5	0.61
7	27	0.6
8	15.5	0.73
8	68	0.7
8	48.5	0.64
8	100	0.63
8	45.5	0.59
8	61	0.57
9	43	0.79
9	58	0.63
9	73.5	0.62
9	79	0.6
10	43.5	0.64
10	29.5	0.64
10	64.5	0.63
10	81.5	0.62
11	87	0.67
11	67.5	0.65
11	74	0.62
11	35	0.61
12	69.5	0.77
12	60	0.74
12	43	0.66
12	32	0.62
12	61.5	0.6
12	36	0.58
13	54	0.99
13	69	0.69
13	44	0.59
14	57	0.57
15	48.5	0.68
15	46.5	0.6
16	5	0.98
16	13	0.67
16	32	0.65
16	13.5	0.61
16	59.5	0.6
17	40.5	0.99
17	40	0.78
17	34	0.64
17	51.5	0.62
18	59.5	0.98
18	58.5	0.77
18	42	0.73
18	60.5	0.66
18	35.5	0.62
19	2.5	0.64
20	15.5	0.93
20	46.5	0.69
20	34.5	0.66
20	27.5	0.62
20	37	0.62
21	10.5	0.61

Chr	Position (Mb)	θ
21	68.5	0.58
22	35.5	0.78
22	50	0.68
22	26.5	0.63
22	7	0.59
23	34	0.61
24	54.5	0.68
24	10.5	0.66
24	18	0.66
24	64	0.59
24	50.5	0.57
25	10	0.76
25	29	0.61
26	1.5	0.62
26	35	0.6
26	40.5	0.58
26	3.5	0.57
27	6.5	0.72
27	8	0.64
27	11.5	0.59
27	32	0.58
27	5	0.58
28	30	0.59
28	43.5	0.58

6th Chapter

GENERAL DISCUSSION

Genome-wide pattern of linkage disequilibrium

Quantifying the level of linkage disequilibrium is an important step for fine-scale mapping of QTL (e.g., Meuwissen and Goddard, 2000), genomic selection (Meuwissen et al., 2001), and increasing the understanding of genomic architecture and the historical population structure (e.g., Hayes et al., 2003). The first chapter of this thesis was designed to measure the extent of LD in German Holstein cattle. For this purpose we used the Illumina BovineSNP50 BeadChip and presented a second generation of LD map statistics for the Holstein genome which has four times higher resolution compared to the maps available so far. At a physical distance of less than 100 kb, an average r^2=0.21 ± 0.26 was observed. We compared our study to that of Sargolzaei *et al*. (2008) and Kim & Kirkpatrik (2009), among some others, who utilized r^2 as measure of LD and found a lower level of LD than previously reported in the literature.

The levels of LD are expected to be highly variable across the genome, due to several factors, such as variation in recombination rate and selection. For reliable results, this variation needs to be taken into account when designing experiments to exploit LD. Variation in rate of recombination across the genome is a key factor that contributes to the variance observed in patterns of LD. A number of researchers have focused on the distance at which average r^2 is reduced to 0.25, as a reasonable point to conclude there is useful LD to detect associations with complex traits (e.g.,Kruglyak, 1999; Ardlie et al., 2002). The reasoning for this r^2 cut-off is as follows: in a complex trait a large QTL may only explain approximately 10% of the phenotypic variation. If a marker only explains 10% of the total QTL variation, then the marker will only explain 2.5 % of the phenotypic variation. Detection of locus effects that cause smaller than 2.5% phenotypic variation requires exponentially increasing population sizes therefore such small effects would be considered undetectable in a moderate-sized study population. Based on the investigations of this study and assuming the size of bovine genome as 3 Gb, to achieve this level of LD the SNP spacing should be ~35 Kb to perform whole genome association study in *Bos taurus*. This implies the

use of more than 75,000 SNPs per individual, assuming that all SNPs are informative (with a MAF ≥ 0.05). However considering the fact that some of the SNPs may have low minor allele frequency in certain breeds, our results suggest that a nearly 100,000 SNPs should be sufficient to perform whole genome association study. According to the results of this study, the same power can be achieved by implementing a panel of 50,000 SNPs with moderate frequencies (e.g., MAF ≥ 0.15) which simultaneously improves the accuracy and magnitude of estimated LD between pairs of SNPs.

Some properties of LD metric r^2, such as its dependence on allele frequency and inter-marker distance, were also explored in this study. We showed that the magnitude of r^2 is dependent on the allele frequency, as such the average r^2 values between SNPs unmatched for allele frequency are much less than matched SNPs. In practice, this observation has applications for single-marker association studies where markers that have similar frequencies to the causative SNP can have high correlations with the causative allele. Indirectly, this property of r^2 has been previously observed, because larger sample sizes are required for mapping when an SNP has a very different frequency to that of the causative polymorphism (Zondervan and Cardon, 2004). Our results also demonstrated that the dependence of LD on the MAF difference between SNP pairs is stronger for SNPs in short distances. These results reveal that the minimizing the allele frequency difference between SNPs, provides a more sensitive and useful metric for analyzing LD across the bovine genome. Although an entirely frequency-independent measure of LD is not possible (Lewinton 1988), frequency matching between SNP pairs removes one major source of statistical noise when assessing the LD structure.

Effective population size (Ne) was another aspect in our dataset which is of relevance for whole genome LD analysis. Because the extent of LD is affected by both recent and past *Ne*, estimating historical *Ne* is useful to shed light on the evolutionary pattern of LD. In this study the recombination rates required for the inferring Ne were estimated directly from haplotype data. Our results showed in German Holstein cattle, the historical *Ne*, going back 500 generations, was approximately 1,200 individuals, in

contrast to the estimated ~100 individuals in recent generations. Although the figures for Ne might not be highly accurate, they nevertheless provide useful information on the trend in effective population size. In general, our results showed a persistent decline in effective population size which is consistent with the results of other studies (Bovine HapMap Consortium, 2009). The rapid decrease in Ne from a very large ancestral population is explained by several bottlenecks, associated with domestication, selection and breed formation (Bovine HapMap Consortium, 2009).

Investigation of possible traces of positive selection in cattle genome

During the last century, the Holstein Friesian breed has been propagated throughout the world and intensively selected, particularly with the introduction of new reproductive technologies. Consequently, genomic regions controlling traits of economic importance are expected to exhibit signatures of selective breeding. With the availability of an ever-increasing number of genetic markers, we are able now to analyze cattle genome on a more comprehensive level to identify what genome changes are associated with the phenotypic changes. Pioneered by human geneticist some tools have been developed to find traces of selection on genomic data. The chapters from three to five of our study were aimed to explore these traces in cattle genome.

Application of extended haplotype homozygosity in Holstein

In the first of these experiments on German Holstein we employed Sabeti's EHH statistic, one of the most popular of selection signature approaches (Sabeti et al., 2002). This test was designed to work with haplotypes. Unfortunately, robust inferences of recent positive selection from genomic data are difficult because of the confounding effects of population demographic history. Another important question with this approach concerns the appropriate null distributions of REHH values. Ideally one should use a set of loci that can be considered to evolve under neutral conditions. However, there are no a priori criteria for choosing such loci with confidence. To validate the efficiency of this test we therefore took the opposite

approach, namely to choose loci that are candidates for positive selection and compared them to the overall genome distribution. We focused on ten genes or gene clusters which are well-known to be related to dairy qualities and therefore were assumed to be potentially under recent selection. The results revealed a longer than expected range of LD in core regions harboring the Casein cluster, DGAT1, GHR, STS and LPR genes which are supposed to affect milk yield and milk composition traits in Holstein cattle. Consistent with previous reports (Grisart et al., 2003; Marques et al., 2008), the second most frequent haplotype of DGAT1 gene (frequency = 30%) showed the highest REHH in the core region. We observed that other haplotypes present in this region display a shorter extent of homozygosity, indicating abundant historical recombination. Therefore, the long stretch of homozygosity observed in this region presumably is not simply due to a low local recombination rate but likely reflects the combination of strong and recent selective pressure, pushing beneficial mutation rapidly towards high frequency with long conserved haplotype surrounding it. In order to test this hypothesis we examined the distribution of this haplotype in 146 animals for which the DGAT1 genotype was available. This comparison revealed an almost perfect association of GGGG haplotype in the region with the Lysine variant at DGAT1 gene which is related to the elevated milk fat content (Winter et al., 2002; Thaller et al., 2003). Allele frequency estimated for the lysine variant was 30% in the sample, which results that, most likely all of them are segregating with GGGG haplotype. This observation confirms that this variant is surrounded by a long range of haplotype and has been underlying recent positive selection.

As a further step ahead, a genome screen was directed to identify selection signatures across the genome of Holstein cattle. Preliminary exploration identified a total of 3741 core regions covering 18.5 % of the mapped genome. After estimating haplotype consistency, a total of 161 genomic regions displayed outlying peaks on a threshold level of 0.01 which were non-uniformly distributed across chromosomes. Bovine chromosomes 6 and 14 which harbor known genes and QTL for several economically important traits (Stone et al., 1999; Mosig et al., 2001; MacNeil and Grosz, 2002; Casas et al., 2003; Li et al., 2004; Ashwell et al., 2005; Nkrumah et al., 2007) showed

8 and 2 outliers, respectively. The number of peaks rises to 41 and 14, respectively, when the threshold was set to P<0.05. Based on the observations of the validation test on candidate genes, we concluded that a substantial proportion of the regions detected in this study is likely under selection.

Regarding the fact that multiple testing may have led to false positive results, we performed a further validation by aligning the 12 regions of extreme REHH to the bovine genome (Btau 4.0) to verify any coincidence of the signals observed with important genomic regions. We found co-location of a panel of genes such as FABP3 (Bionaz and Loor, 2008), HTR2A (Reist et al., 2003), CPN3 (Barendse et al., 2008), PTGER2 (Arosh et al., 2003) and some others with putative regions which previously suggested being under selection in cattle populations. For example FABP3 plays a key role in the regulation of the channeling of fatty acids toward copious milk fat synthesis in bovine mammary (Bionaz and Loor, 2008). There are also reports of associations with subcutaneous fat thickness in beef cattle (Roy et al., 2003) as well as milk fat content in sheep (Calvo et al., 2004). One interesting observation of this study was the HTR2A 5 gene which acts in serotoninergic pathways which are involved in economically important bovine gastrointestinal (GI) motility disorders, such as displaced abomasum and cecal dilatation/dislocation (Reist et al., 2003). It was also suggested that variants of this gene are related with behavioral disorders in human (Khait et al., 2005) and aggressiveness in canine (Peremans et al., 2003). This point looks more interesting when we compare the temperament behavior of modern cattle breeds, which have been bred during the last decades, to native cattle breeds worldwide. However, still more reference data in terms of statistical and functional significance will be required to confirm our finding at this locus.

Comparison of the pattern of selective sweeps revealed by EHH test among populations

Given the presence of the large number of false positives among possible true selective sweeps, it is important to find additional criteria of how the true cases can be identified. Schlötterer (2002) has suggested that signatures that are found in at least

two populations and/or with more than one statistics might be considered to be more reliable. We therefore used populations with different demographic and selective histories.

In the first step we scanned the genomes of Brown Swiss (n=277) and Simmental (n=462) breeds using EHH statistic based on 50k genotypes. The extent of haplotype homozygosity at region of 10 candidate genes was estimated. As shown in Table 1, six and three regions exhibited a longer than expected extent of haplotype homozygosity, respectively in Brown Swiss and Simmental breeds. It is evident that Holstein and Brown Swiss show more similarity with respect to the number of gene regions underling positive selection. This observation corresponds roughly to expectations when the history of formation of the breeds and their breeding purposes are considered. Further genome-wide screen also revealed 140 and 137 genomic regions with haplotype consistency longer than expected ($P \leq 0.01$) across the genome of Brown Swiss and Simmental, respectively, in contrast to 161 regions in Holstein.

To confirm the chromosomal regions containing selection evidences in Holstein, we examined the co-location of selection signature at significance level of ($P \leq 0.05$) across the genome of the three breeds. Our analysis revealed 55 and 48 regions which coincided between Holstein vs. Brown Swiss and Holstein vs. Simmental, respectively. There are also 55 overlaps between Brown Swiss and Simmental. Overall, we found only 7 overlapping regions across the genome of three breeds (Table 2) which included the Prolactin receptor gene on BTA20. However, the resulting pattern from the tracing of the sweep signatures in three breeds was generally not consistent.

Table 1. Comparison of the significance of the haplotype homozygosity revealed by the EHH test among Simmental (SI), Brown Swiss (BS) and Holstein (HS).

Candidate Region	P-Value*		
	SI	BS	HS
DGAT1	- / 0.19	**- / 0.03**	**- / 0.06**
Casein Cluster	0.29 / 0.12	**0.08 / 0.04**	**0.01 / 0.01**
GH	0.21 / 0.17	**0.02 / 0.01**	0.86 / 0.90
GHR	**0.23 / 0.10**	0.45 / 0.34	**0.10 / 0.08**
SST	0.76 / 0.84	0.22 / 0.44	**0.03 / 0.07**
IGF-1	0.35 / 0.24	**0.12 / 0.11**	0.38 / 0.55
ABCG2	0.19 / 0.18	0.29 / 0.19	0.76 / 0.79
LEP	**0.003 / 0.03**	0.15 **/ 0.02**	0.45 / 0.42
LPR	**0.08 / 0.11**	0.35 / 0.40	**0.04 / 0.04**
PIT-1	0.22 / 0.24	0.06 / 0.17	0.67 / 0.69

* P-values for REHH statistic are presented for upstream and downstream sides from core region, respectively, for the most longest haplotype with frequency > 25%

Table.2. List of overlapping regions with extreme EHH in Simmental, Brown Swiss and Holstein and candidate genes located nearby

Symbol	Gene	Chr	Position (Mbp)
ABHD10	Esterase-lipase	1	56.75-57.25
TSGA14	Testis specific, 14	4	95.75-96.25
CRY1	Cryptochrome 1	5	75.75-76.25
MGP	Matrix Gla protein	5	101.75-102.25
-	-	12	62.75-63.25
HSPB3	Heat shock 27kDa protein 3	20	26.25-26.75
PRLR	Prolactin receptor	20	41.5-43

Application of F_{ST} statistic to find standing variation

The rationale of selective sweep mapping is that during breed formation natural or artificial selection should imparts a distinct signature on genomic regions harboring loci that influence the specific phenotype that is selected. In chapters four and five of this thesis we addressed this issue. In this study, we developed a new Bayesian approach for exploring differentiated loci and applied it to a set of geographically separated populations with identical or diverse breeding goals. We estimated F_{ST} for 40,595 SNPs either for pair-wise comparisons or across the dairy vs. beef breeds. This algorithm was able to deal with a large panel of marker information. Our results suggested almost similar level of differentiation in pair-wise comparisons within the dairy and beef breeds. Clustering the genome-wide estimates of θ values between Holstein and Brown Swiss versus Angus and Piedemontese breeds, using Akaike's criterion, resulted in two groups. One representing putatively neutral loci, and the other possibly corresponding to the genomic regions affected by selection. Overall, the average F_{ST}, comparing of dairy vs. beef breeds, was equal to 0.3 which is substantially higher than the differentiation index reported by MacEachern et al., (2009) between Holstein and Angus. The higher average of F_{ST} as well as the similar pair-wise F_{ST} within dairy and beef breeds might reflect the overweighting influence of a large number of fixed SNPs in the pair-wise comparisons of breeds and groups.

Selection of a favorable variant is expected to result in a higher level of differentiation for neighboring SNPs. In several instances, outlier SNPs tended to cluster to similar regions (e.g. BTA2 or BTA18). Hence, in order to identify footprints of selection at the regional level we adopted the strategy proposed by Weir et al., (2005) consisting in performing average of SNP F_{ST} over sliding windows. Linkage disequilibrium was shown to decay within 1–2 Mb in the analyzed breeds. However a strong selective effect could sweep loci that are located considerably further away. We chose 500kb because of the longer extent of LD in cattle compared to human, in which the considered length is usually less than 200 Kb (Sabeti et al., 2002. Voight et al., 2006). However, it remains difficult to define, *a priori*, an optimal window size since it

would depend on the strength and timing of selection which are expected to be highly variable.

As summarized in Table S1, 127 regions with extreme scores (P-value<0.025) were identified when considering F_{ST} across populations. Annotation of the genes underlying these regions with the extreme F_{ST} revealed some genes (e.g., SMCP and FGF1genes). A receptor of the latter gene (FGFR3) showed evidence of selection in a genome-wide sweep mapping study using F_{ST} among dog breeds (Pollinger et al., 2005). This gene is responsible for achondroplasia (shortened limbs) in Humans. However, F_{ST} results do not appear to report strong candidates in the region of extreme signals. As an explanation, we can theorize that selection may work on the genes that have not been considered the primary targets of selection so far. In addition for most extreme regions identified, we were not able to propose candidate genes on the basis of the gene content in the vicinity of the peak location. These results mostly revealed gene deserts in the location of extreme peaks, which may reflect selection acting on uncharacterized regulatory region or simply fixation of non-coding DNA by genetic drift in the absence of any selection. This observation is consistent with the reports of Flori et al. (2009), and Gu et al. (2009) which reported poor gene content regions in genome wide analyses of Cattle and Thoroughbred horse, respectively, using 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 have been an important substrate for adaptive evolution.

We also examined the validity of F_{ST} analysis by testing some candidate major genes in our data set. The results revealed F_{ST} values larger than expected (P < 10%) for regions harboring the Casein cluster, GHR, STS, LP and IGF-1 genes which are supposed to be targets for artificial selection. The observation of selection evidence in the region of GHR gene on BTA20 is consistent with the reports of Flori et al. (2009) and Hayes et al. (2009) which the latter reported it between Angus and Holstein breeds. The presence of the longer than expected of haplotype homozygosity in this

region was also observed in the validation of EHH test in current study. 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 (Bovine HapMap consortium 2009), also matched to the outlier F_{ST} windows in our study.

Tracing the on-going sweeps

The iHS test (Voight et al., 2006), a derivation of EHH, was also applied on a diverse set of cattle breeds in this study and results were presented in chapter 5. We defined regions of the genome that may contain targets of positive selection as windows in the extreme of empirical distribution. This criterion resulted in 94 significant windows ($P \leq 0.05$). Interrogating the corresponding genomic regions to the cattle genome sequence resulted in a large number of flanking loci. The list of genes with signatures of positive selection was significantly enriched with those involved in the biological processes such as anatomical structure development, muscle development, metabolism of carbohydrates and lipids, spermatogenesis and fertilization. We refined the complete list for the most important genes in the region of clustered signals that may have functional relevance to the economical phenotypes. A remarkable observation in this study was a perfect overlap between an extreme |iHS| window and a major QTL on BTA18 which was recently reported to have large effects on calving ease, several conformation traits, longevity, and total merit in Holstein cattle (Cole et al., 2009). A cluster of signals reflecting strong evidence of selection was also observed in the vicinity of SST gene. We found also clusters of outliers which overlapped among some beef breeds. These results show significant enrichments for genes such as SPATA17, MGAT1, PGRMC2 and SRD5A2 in the region of clustered signals which belong to a number of functional categories relevant to reproduction, including gamete generation, embryo development and spermatogenesis and genes in these categories may provide strong candidates for selection for fertility traits. Another interesting observation was the strong evidence for selection in the region of genes related to muscle formation (e.g., ACTC1, COL23A1, MATN2, and FAP). For example polymorphisms in the genes encoding Actinin are among the best

characterized athletic-performance associated variants in human endurance athletes (Yang et al., 2003; Chan et al., 2008). Evidence for positive selection in the genomic region surrounding muscle related genes has been also reported in racing horses (Gu et al., 2009) and humans (MacArthur et al., 2007). The presence of the genes like Actinin, Collagen and fibroblast activation protein as well as the gene responsible for developing cartilage rudiments in positively selected regions in beef cattle suggest that selection for muscle related phenotypes has played a major role in the shaping the beef cattle. A better understanding of the role that these genes play in the development, strength and integrity of muscle may contribute to improved knowledge of musculoskeletal traits and developing new marker systems for breeding beef breeds with better performance. These results generally are consistent with the observations of Flori et al. (2009) and begin to suggest general themes about the types of genes that have been targets of positive selection in cattle genome.

How can the discrepancies in the results be explained?

Overall, we found a modest overlap between the results of previous genome-wide studies and our scans for selection. How can the discrepancies in these results be explained? From the methodological point of view, first, given the varying statistical tests used to detect signatures, we should not expect complete agreement between studies. More specifically, different studies are probably detecting different selective events. For example, |iHS| statistic has the greatest power to detect incomplete and on-going selective sweeps. Conversely, tests based on the site frequency spectrum like F_{ST} has greater power to identify sweeps where the advantageous allele is approaching fixation or completed sweeps in which new mutations are occurring on selected haplotypes that over time will lead the patterns of genetic variation to equilibrium. In short, the statistical tests used in each study are recovering selective events from different time periods and for different stages of the selective sweep. Second, even for tests that should detect similar types of selective events, low statistical power further decreases the probability of overlap. Third, most studies report only the most significant results (i.e. outliers in the 1% empirical distribution). Therefore, the results

presented in this study are probably a conservative estimate of overlap between studies. Finally, the false positive rate in genome-wide scans for selection is likely to be high.

Population demographic history can also reveals similar patterns on DNA sequence variation, making inferences of selection difficult. For example, population expansions can lead to an excess of low frequency alleles compared to the number expected under the standard neutral model. Likewise, if the recent positive selection for a putative mutation has been started from a higher initial frequency of beneficial alleles; such an allele might, for instance, be imported into a breed through crosses with other breeds, in such a case beneficial alleles may be included in diverse haplotypes and LD based estimators would not be able to trace the selection signature.

From the technical point of view, the density of the markers is also critical for the power of such studies and could be a source of discrepancy. It was shown earlier (see chapter 3 for more details) with LD based analyses that core regions more likely appeared where the marker density is greater than the average (Qanbari et al., 2010). This would imply that the anticipated arrival of genotyping chips with increase genome-wide marker density (by a factor ~10) would allow a more reliable and comprehensive screening of the genome for signatures of selection by LD based tests. As discussed earlier, the size of the window in genome sliding strategy is also a source of discrepancy. One potential refinement would be to adjust window sizes to local levels of LD (Weir et al., 2005), although how to account for varying levels of LD between populations remains unclear. Finally, the incongruities can also result from the complex genomic interactions or lack of power, given the sample size available for some of the breeds in this study.

Conclusions and remaining challenges

Based on the results of this research we conclude that high-resolution genome scan using dense markers is capable to identify outlier regions that potentially contain

genes contributing to within and inter-breed phenotypic variation. Genomic scans based on site frequency spectrum and haplotype data led to the detection of a surprising high frequency of regions subjected to recent positive selection in cattle genome. Many of the regions showing extreme values for the statistics seem to play important roles in economically important traits in cattle and can now serve as starting points for formulating biological hypotheses. Results from either site frequency or haplotype based methods also showed evidence of positive selection for some candidate regions harboring major genes. However, in general, the overlap between the identified regions and those from previous studies was modest. Analysis of population differentiation revealed signatures of selection occurring in regions of the genome thought to be nonfunctional which may reflect selection acting on uncharacterized regulatory regions or reflect simply the fixation of non-coding DNA by genetic drift.

Clearly, many challenges remain to be investigated, including the development of efficient methods to differentiate the effects of drift and selection, identifying the causal gene driving the signature of selection observed across large genomic regions and the functional characterization of the suspected targets of selection. Our results may be of future interest for identifying signatures of recent positive artificial selection between cattle breeds and as an additional evidence for polymorphisms that show associations with beef or dairy traits.

References

Ardlie, K. G., Kruglyak, L., and Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3: 299–309.

Arosh, J. A., Banu, S. K., Chapdelaine, P., Emond, V., Kim, J. J., MacLaren, L. A., Fortier, M. A. 2003. Molecular Cloning and Characterization of Bovine Prostaglandin E2 Receptors EP2 and EP4: Expression and Regulation in Endometrium and Myometrium during the Estrous Cycle and Early Pregnancy. Endocrinology 144: 3076-3091.

Ashwell, M. S., Heyen, D. W., Weller, J. I., Ron, M., Sonstegard, T. S., Van Tassell, C. P., and Lewin H.A. 2005. Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle. Journal of Dairy Science 88: 4111–9.

Barendse, W., Harrison, B., Bunch, R. J., Thomas, M. B., and Turner, L. B. 2009. Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. BMC Genomics 10:178.

Bionaz, M., and Loor, J. 2008. Gene networks driving bovine milk fat synthesis during the lactation cycle. BMC Genomics 9: 366.

Biswas, S., and Akey, J. M. 2006. Genomic insights into positive selection. Trends in Genetics 22: 437-446.

Carlson, C. S., Thomas, D. J., Eberle, M. A., Swanson, J. A., Livingston, R. J., Rieder, M. J., Nickerson, D. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. Genome Research 15: 1553–1565.

Casas, E., Shackelford, S. D., Keele, J. W., Koohmaraie, M., Smith, T. P., and Stone, R. T. 2003. Detection of quantitative trait loci for growth and carcass composition in cattle. Journal of Animal Science 81: 2976–83.

Chan, S., Seto, J. T., MacArthur, D. G., Yang, N., North, K. N., et al. 2008. A gene for speed: contractile properties of isolated whole EDL muscle from an alpha-actinin-3 knockout mouse. American Journal of Physiology- Cell Physiology 295: C897–904.

Flori, L., Fritz, S., Jaffrezic, F., Boussaha, M., Gut, I., Heath, S., Foulley, J. L., Gautier, M. 2009. The genome response to artificial selection: a case study in dairy cattle. PLoS One. 4: e6595.

Gianola, D., Simianer, H., and Qanbari, S. 2010. A two-step method for detecting selection signatures using genetic markers. Genetics Research (in press).

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M., and Snell, R. 2002. Positional candidate cloning of a QTL in dairy cattle: Identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Research 12:222–231.

Gu, J., Orr, N., Park, S. D., Katz, L. M., Sulimova, G., et al. 2009 A Genome Scan for Positive Selection in Thoroughbred Horses. PLoS ONE 4: e5767.

Hayes, B. J., Lien, S., Nilsen, H., Olsen, H.G., Berg, P. et al. 2008. The origin of selection signatures on bovine chromosome 6. Animal Genetics 39: 105-111.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., and Goddard, M. E. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research 13: 635–43.

Hayes, B. J., Chamberlain, A. J., Maceachern, S., Savin, K., Mcpartlan, H. et al. 2009. A genome map of divergent artificial selection between Bos Taurus dairy cattle and Bos taurus beef cattle. Animal Genetics 40: 176-84

Innan, H., and Kim. Y. 2004. Pattern of polymorphism after strong artificial selection in a domestication event. Proceedings of the National Academy of Sciences of the United States of America 101: 10667-10672.

Khait, V. D., Huang, Y., Zalsman, G., Oquendo, M. A., Brent, D. A. Harkavy-Friedman, J. M, Mann, J. J. 2004. Association of Serotonin 5-HT2A Receptor Binding and the T102C Polymorphism in Depressed and Healthy Caucasian Subjects. *Neuropsychopharmacology* 30: 166-172.

Kim, E. S., and Kirkpatrick, B. W. 2009. Linkage disequilibrium in the North American Holstein population. Animal Genetics 40: 279-288.

Kruglyak, L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22:139-144.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49–67.

Li, C., Basarab, J., Snelling, W. M., Benkel, B., Kneeland, J., Murdoch, B., Hansen, C. and Moore, S. S. 2004. Identification and fine mapping of quantitative trait loci for backfat on bovine chromosomes 2, 5, 6, 19, 21 and 23 in a commercial line of Bos taurus. Journal of Animal Science 82: 967–972.

MacArthur, D. G., Seto, J. T., Raftery, J. M., Quinlan, K. G., Huttley, G. A., et al. 2007. Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. Nature Genetics 39: 1261–1265.

MacNeil, M. D., and Grosz, M. D. 2002. Genome-wide scans for QTL affecting carcass traits in Hereford × composite double backcross populations. Journal of Animal Science 80: 2316–2324.

Marques, E., Schnabel, R., Stothard, P., Kolbehdari, D, Wang, Z., Taylor, J. F., Moore, S. S. 2008. High density linkage disequilibrium maps of chromosome 14 in Holstein and Angus cattle. BMC Genetics 9: 45.

MacEachern, S.**,** Hayes, B.**,** McEwan, J., and **G**oddard, M. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics 10**:** 181.

Meuwissen, T. H. E., Goddard, M. E., 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genetic Selection Evolution 33:605–634.

Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157: 1819-1829.

Mosig, M. O., Lipkin, E., Khutoreskaya, G., Tchourzyna, E., Soller, M., and Friedmann A. 2001. A whole-genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics 157: 1683–1698.

Nkrumah, J. D., Sherman, E. L., Li, C., Marques, E., Crews, D. H., Bartusiak, R., Murdoch, B., Wang, Z., Basarab, J. A., and Moore, S. S. 2007. Primary genome scan to identify putative QTL for feedlot growth rate, feed intake and feed efficiency of beef cattle. Journal of Animal Science 85: 3170–3181.

Peremans, K., Audenaert, K., Coopman, F., Blanckaert, P., Jacobs, F. Otte, A, Verschooten, F., van Bree, H., van Heeringen, K., Mertens, J., Slegers, G., Dierckx, R. 2003. Estimates of regional cerebral blood flow and 5-HT2A receptor density in impulsive, aggressive dogs with 99mTc-ECD and 123I-5-I-R91150. European Journal of Nuclear Medicine and Molecular Imaging 30: 1538-1546.

Pollinger, J. P., Bustamante, C. D., Adi Fledel-Alon, A., Schmutz, S. M., Gray, M. M., and Wayne, R. K. 2005. Selective sweep mapping of genes with large phenotypic effects. Genome Research 15: 1809–1819.

Prasad, A., Schnabel, R. D., McKay, S. D., Murdoch, B., Stothard, P., Kolbehdari, D., Wang, Z., Taylor, J. F., and Moore, S. S. 2009. Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. Animal Genetics 39: 597-605.

Qanbari, S., Pimentel, E. C. G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A. R. & Simianer, H. (2009) The pattern of linkage disequilibrium in German Holstein cattle. Animal Genetics (In press).

Reist, M., Pfaffl, M.W., Morel, C., Meylan, M., Hirsbrunner, Blum, J. W., Steiner, A. 2003. Quantitative mRNA analysis of eight bovine 5-HT receptor subtypes in brain, abomasum, and intestine by real-time RT-PCR. J. Receptor Signal Transduction 23, 271–287.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J, Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E. S. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., and Schaeffer L. R. 2008. Extent of linkage disequlibrium in Holstein cattle in North America. Journal of Dairy Science 91: 2106–2117.

Schlötterer, C. 2002. A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics 160:753-763.

Stone, R. T., Keele, J. W., Shackelford, S. D., Kappes, S. M., and Koohmaraie, M. 1999. A primary screen of the bovine genome for quantitative trait loci affecting carcass and growth traits. Journal of Animal Science 77: 1379–84.

Thaller, G., Krämer, W., Winter, A., Kaupe, B., Erhardt, G., Fries, R., 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. Journal of Animal Science 81: 1911–1918.

The Bovine HapMap consortium, 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324: 528-532.

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

Wang, E. T., Kodama, G., Baldi, P., Moyzis, R. K. 2006 Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceeding of National Academy of Science of USA 103: 135–140.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., Hill, W. G. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Research 15: 1468–1476.

Winter, A., Krämer, W., Werner, F. A., Kollers, S., Kata, S., Durstewitz, G., Buitkamp, J., Womack, J.E., Thaller, G., and Fries, R.. 2002. Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-

CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proceeding of National Academy of Science of USA 99: 9300–9305.

Yang, N., MacArthur, D. G, Gulbin, J. P., Hahn, A. G., Beggs, A. H., Easteal, S., North, K. 2003. ACTN3 genotype is associated with human elite athletic performance. Ameican Journal of Human Genetics 73: 627–631.

Zondervan, K. T., and Cardon, L. R. 2004. The complex interplay among factors that influence allelic association. Nature Reviews Genetics 5: 89-100.