rccp Revista Colombiana de Ciencias Pecuarias

Identifying signatures of recent selection in Holstein cattle in the tropic¤

Identificación de señales de selección reciente en ganado Holstein en el trópico

Identificação de sinais de seleção recente no gado Holandês no trópico

Juan C Rincón[1,2*], Zoot, MSc, PhD; Albeiro López[2], MV, Zoot, MSc, PhD; Julián Echeverri[2], Zoot, MSc, PhD.

[1]*Programa de Medicina Veterinaria y Zootecnia, Universidad Tecnológica de Pereira, Risaralda, Colombia.*

[2]*Grupo de Investigación Biodiversidad y Genética Molecular (BIOGEM), Facultad de Ciencias Agrarias, Departamento de Producción Animal, Universidad Nacional de Colombia, sede Medellín, Colombia.*

(Received: August 5, 2016; accepted: June 2, 2017)

doi: 10.17533/udea.rccp.v31n1a06

Abstract

Background: Holstein cattle have undergone strong selection processes in the world. These selection signatures can be recognized and utilized to identify regions of the genome that are important for milk yield. **Objective:** To identify recent selection signatures in Holstein from the Province of Antioquia (Colombia), using the integrated haplotype score (*iHS*) methodology. **Methods:** Blood or semen was extracted from 150 animals with a commercial kit. The animals were genotyped with the BovineLD chip (6909 SNPs). The editing process was carried out while preserving the loci whose minor allele frequency (MAF) was greater than 0.05. In addition, genotypes with Mendelian errors were discarded using R and PLINK v1.07 software programs. Furthermore, the extended haplotype homozygosity (EHH), *iHS* and the p-value were determined with the "rehh" package of R language. **Results:** The minor allele frequencies showed a tendency toward intermediate frequency alleles. In total, 144 focal markers were significant ($p<0.001$) for selection signatures. Some chromosomes showed a greater number of signatures than others. Many of the variants were found inside genes, although they were in intronic regions. Some important regions were associated with genes TRAPPC12, PANK3, ZNF16, OPLA and DPYSL4, which are related with cellular transport, excretion or metabolism. **Conclusion:** Identifying signatures of selection using the *iHS* method made it possible to determine some important regions for selection in Holstein cattle in the high tropics, some of which had been previously reported to be associated with quantitative traits loci (QTLs).

Keywords: *dairy cattle, genetic mapping, QTLs, selection pressure, single nucleotide polymorphism.*

¤ To cite this article: Rincón JC, López A, Echeverri J. Identifying signatures of recent selection in Holstein cattle in the tropic. Rev Colomb Cienc Pecu 2017; 31(1):45-58.

* Corresponding author: Juan C Rincón. Programa de Medicina Veterinaria y Zootecnia, Universidad Tecnológica de Pereira, Carrera 27 #10-02 Barrio Álamos, Pereira, Risaralda. AA: 97 - Código postal: 660003 Colombia. E-mail: rincon.juan@utp.edu.co

Resumen

Antecedentes: El ganado Holstein ha sido sometido a procesos fuertes de selección en el mundo. Estas señales de selección pueden ser reconocidas y utilizadas para identificar regiones del genoma importantes para la producción de leche. **Objetivo:** Identificar señales de selección recientes en ganado Holstein de la Provincia de Antioquia (Colombia), mediante la metodología de puntaje haplotípico integrado (*iHS*). **Métodos:** A 150 animales se les extrajo DNA de sangre o semen mediante un kit comercial y posteriormente se genotiparon los animales con el chip BovineLD (6909 SNPs). Se realizó edición conservando los loci con frecuencia del alelo menor (MAF) superior a 0,05. Además, se descartaron los genotipos con errores mendelianos, usando el software R y PLINK v1.07. La determinación de la homocigosidad haplotípica extendida (*EHH*), *iHS* y el valor p se realizó utilizando el paquete "rehh" de R. **Resultados:** Las frecuencias del alelo menor mostraron una tendencia hacia alelos de frecuencias intermedias. En total, 144 marcadores focales fueron significativos ($p<0,001$) para las señales de selección. Algunos cromosomas presentaron mayor número de señales de selección que otros. Muchas de las variantes focales se encontraron al interior de genes, aunque comúnmente en regiones intrónicas. Algunas de las regiones importantes estuvieron asociadas con genes como TRAPPC12, PANK3, ZNF16, OPLA y DPYSL4 que en general se encuentran asociados con funciones relacionadas con el transporte, excreción o metabolismo celular. **Conclusión:** La identificación de señales de selección usando el método *iHS* permitió determinar algunas regiones importantes para la selección en ganado Holstein del trópico alto, algunas de las cuales han sido previamente reportadas por su asociación a loci de características cuantitativas (QTLs).

Palabras clave: *ganado lechero, mapeo genético, polimorfismos de nucleótido simple, presión de selección, QTLs.*

Resumo

Antecedentes: O gado holandês tem sido objeto de processos de seleção fortes no mundo. Estes sinais de seleção podem ser reconhecidos e utilizados para identificar regiões do genoma importantes para a produção de leite. **Objetivo:** Identificar sinais de seleção recente em gado Holandês de la Província de Antioquia (Colômbia), através da metodologia de pontuação haplotípica integrada (*iHS*). **Métodos:** Foram usados 150 animais para a extração de DNA a partir de sangre ou sêmen usando kit comercial, os animais foram posteriormente genotipados com o chip BovineLD (6909 SNPs). A edição foi feita mantendo os loci com frequência do alelo menor (MAF) de 0,05; além disso, genótipos com erros mendelianos foram descartados usando o programa R e PLINK v1.07. A determinação da homozigosidade haplotípica estendida (*EHH*), *iHS* e valor p foi realizada utilizando o pacote estatístico R "reeh". **Resultados:** As frequências do alelo menor mostraram uma tendência inclinada a frequências intermédias. No total, 144 marcadores focais foram significativos ($p<0,001$) para os sinais de seleção. Alguns cromossomos apresentaram mais numero de sinais de seleção que outros. Muitas dos variantes focais foram encontradas dentro dos genes, embora comumente em regiões intrônicas. Algumas das regiões importantes foram associadas com genes como TRAPPC12, PANK3, ZNF16, OPLA e DPYSL4 que geralmente estão associadas a funções relacionadas com o transporte, a excreção ou metabolismo celular. **Conclusão:** A identificação de sinais de seleção usando o método *iHS* permitiu determinar algumas regiões importantes para a seleção no gado holandês do tropico alto, algumas destas regiões foram previamente relatados por sua associação com loci de características quantitativas (QTLs).

Palavras chave: *gado leiteiro, mapeamento genético, polimorfismo de nucleotídeo simples, pressão de seleção, QTLs.*

Introduction

Holstein is the most common breed in Colombian dairy farms. It was introduced into Colombia during the late nineteenth and early twentieth century from Holland and North America. From its arrival, Holstein cattle were located in the high tropics, where climate is colder. Subsequently, animals and semen were imported mostly from Europe and North America. In Colombia, breeding programs are new and selection of sires is often done without appropriate selection criteria (ACHF, 2009), so it is difficult to predict its consequences.

Recent advances in molecular technology have reduced sequencing and high density genotyping

costs in animals (Meuwissen and Goddard, 2010). This has made it possible to have large databases of single nucleotide polymorphisms (SNPs) of domestic animals, particularly in bovine cattle for which genomic selection is being implemented (Meuwissen *et al*., 2001).

Identifying regions with high local haplotype homozygosity in relation to the neutral scenario is a good strategy for identifying candidate genes subjected to natural or artificial selection (Qanbari *et al*., 2011). Taking into account that Holstein cattle has been selected for decades in many places around the world under different selection criteria and in accordance with the production and market conditions of each country involved, it is possible to use this approach to find regions associated with different traits of importance for milk yield. The search for these selection signatures has already been carried out in bovines (Druet *et al*., 2013; Ramey *et al*., 2013; Qanbari *et al*., 2014) regions associated with quantitative trait loci (QTL; Hayes et al., 2008).

To detect selection signatures, Sabeti *et al.* (2002) developed an estimator called extended haplotype homozygosity (*EHH*) to search for genetic footprints of positive selection in humans. The *EHH* is the probability that two randomly chosen chromosomes carry the same allele in a focal SNP and are identical by descent in the markers surrounding them. Voigth *et al.* (2006) developed an empirical test based on the integral of the observed decay of *EHH*, which was defined as *iHH.* After this, the test was complemented to define the *iHS test* as a log-ratio of *iHH* calculated in the derived and ancestral focal alleles. The approach suggested by Voight *et al*. (2006) involves comparing the linkage disequilibrium (LD) around a selected allele in relation to an unselected allele which acts as a control to the LD expected in the region. The proposed measure is known as standardized extended haplotype homozygosity (*iHS*).

Practically speaking, the integrated haplotype score (*iHS*), which is based on the extended haplotype homozygosity (*EHH*), makes it possible to detect selective pressure by identifying the regions of the genome where local LD increases. Observing the LD as a function of the distance may reflect the history of the population during distant generations or during more recent generations (with larger haplotypes). This, in turn, can be used to find selection signatures, since selected alleles increase in frequency in the population of a specific chromosome segment around it.

Therefore, the objective of this study was to identify signatures of recent selection in Holstein cattle from the province of Antioquia (Colombian high tropic) through the *iHS* methodology using the BovineLD genotyping chip. This can contribute to identify regions of the genome that were subjected to specific selection pressure according to the conditions of the population. Each signature can be associated to a QTL and this information can be used as relevant biological information to be included in the genomic selection program in Holstein cattle in Colombia.

Materials and methods

Ethical considerations

All experimental procedures were approved by the Institutional Committee for care and use of animals of Universidad Nacional de Colombia (Medellin city, Colombia; Act No 03, CEMED 015, 2012).

Genotype data and DNA extraction

This study was conducted with Holstein animals in the high tropics of the Province of Antioquia (Colombia). The animals are part of the genetic improvement program of Universidad Nacional de Colombia at Medellín city, and Colanta company (Cooperativa Lechera de Antioquia). Semen or blood samples were obtained from 150 animals (37 bulls and 113 cows with at least one calving). This was done following the protocols established by the genetic improvement program while avoiding practices considered as animal abuse. The specific animal husbandry, feeding and health conditions varied from herd to herd. This was also true for their topography and geographical location. The animals were selected from 35 dairy herds in 12 municipalities in the high tropics of Antioquia. The elevations at these locations ranged between 2,000 and 2,600 m.a.s.l, with average temperature between 13 and 17 ºC and precipitation between 2,000 and 4,000 mm, annually. All animals were kept on grazing and

supplemented with balanced commercial feed. Animal selection included as many bulls as possible that had daughters in the population (nine local and 28 foreigners) and sperm in the market, since these sires have many of the haplotypes found in the population, which renders them highly informative.

Blood samples were taken from the middle coccygeal vein using needles (18-gauge) with 5-mL BD Vacutainer® tubes containing EDTA as anticoagulant (BD Diagnostics, Franklin lakes, NJ, USA) and then stored at 4 °C until processed. Bull DNA was obtained from semen, which was extracted and stored in 250 and 400 µL straws.

The DNA was extracted according to the manufacturer's recommendations using two commercial kits, namely DNeasy Blood & Tissue Kit® for blood and QIAamp® DNA Mini Kit Protocol 1 for semen (Qiagen Inc, Valencia, CA, USA). Genotyping was performed using the BeadChip Bovine LD® (Illumina Inc., San Diego, CA, USA) chip, which has a total of 6,909 SNPs with an average distance of 383 kb (Illumina, 2013). Results were edited using R (R Development Core Team, 2012) and PLINK v1.07 programs (Purcell *et al*., 2007); only autosomal markers were kept. Lost data test were considered below 0.1% and the SNPs with a minor allele frequency (MAF) below 0.05 were discarded. Likewise, genotypes with Mendelian errors greater than 0.05 were also discarded. Each of the alleles was coded as 1 (ancestral allele) or 2 (derived allele), and a map file was built taking into account the reference kept in the bovineLD chip from illumine (Illumina Inc., San Diego, CA, USA). Only the haplotypes that could be mapped to the UMD 3.1 bovine genome assembly were taken into account.

Haplotype reconstruction and linkage analysis

The lowest possible kinship among groups was taken into account when selecting parents. However, there were a few couples and triplets (father-mother-daughter) in some cases. After the edition process, haplotypes were reconstructed for each chromosome individually and unrelated animals were taken into account. This was done using the Beagle 3.3 software (Browning and Browning, 2009). The reconstruction of the haplotypes was used for later analyses. Linkage analysis was performed in parallel with the program Haploview v4.1 (Barrett *et al*., 2005) to estimate linkage disequilibrium via the r^2 statistic.

EHH and iHS estimation

The evidence of positive or negative selection was obtained using the *iHS* test, based on haplotype frequencies, as described by Voigh *et al*. (2006). The *iHS* measures the expansion of the local LD, taking into account the haplotype centered on an SNP that carries the ancestral allele with respect to the haplotype that carries the derived allele. This statistic is applied to each SNP individually (as the center of the haplotype), thus it is necessary to calculate the integral of the decay of *EHH* for the ancestral (iHH_A) and derived (iHH_D) alleles. The *iHS* is standardized in such a way that its mean is 0 and its variance 1. This is achieved with the following equation:

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

Where:

iHS: Is the measure of the extension of the linkage.

$ln\left(\frac{iHH_A}{iHH_D}\right)$: Is the measure of the unstandardized *iHS* based on a specific SNP.

$E\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]$: Is *iHS* expectancy.

$SD\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]$: Is the standard deviation of the *iHS*.

The highest positive or negative values correspond to unusually large haplotypes carrying the ancestral or derived allele, respectively. The *iHS* scores of the SNPs were transformed in:

$$P_{iHS} = -log[1-2|\Phi(iHS)-0.5|]$$

Where:

$\Phi(iHS)$: Is the distribution of the Gaussian cumulative distribution. If we assume that the *iHS* data is normally distributed (under neutrality), may be interpreted as , where *P* is the p-value associated with two queues for the neutral hypothesis in which

there is no selection. The ratio of values for *iHS*>4 was plotted while taking into account the 15 markers around each focal SNP.

The estimation of the *EHH* and *iHS* was performed using R's "rehh" package (Gautier and Vitalis, 2012), which uses a function to detect selection signatures in dense marker data using the test based on extended haplotype homozygosity (*EHH*) that we discussed above. The software package makes it possible to estimate the statistics and create plots to visualize and interpret results.

Identification and annotation of candidate regions

In order to identify and annotate the candidate regions, the SNPs with the greatest effect, >4 ($p<0.00001$) were mapped. For nearby SNPs, the LD between contiguous markers was determined. Likewise, annotation was performed while making sure that the windows associated to the QTLs do not overlap. The most important SNPs were annotated based on the UMD 3.1 assembly (*Bos taurus*) from NCBI and ENSEMBL using the *Variant Effect Predictor tool* (VEP; McLaren *et al*., 2010) a web-based tool (the SNP Effect Predictor). Gene groupings and figures were made in accordance with the information from the ENSEMBL, based on the UMD 3.1 construction of VEP (McLaren *et al*., 2010) a web-based tool (the SNP Effect Predictor). An ontology from the *Gene Ontology Consortium* was used together with previous reports of QTLs from the *Animal QTLdb* public database, specifically those in the *cattleQTLdb* (Hu *et al*., 2013). Candidate region search and identification was performed with 1-Mb genome windows whose focal SNP was at the center (0.5 Mb for each side) of the window.

Results

After the editing phase, a total of 6,677 SNPs were available. These SNPs had a higher ratio of alleles with intermediate frequencies; this is typical of the genotyping chips that search for segregating markers in different breeds. Most SNPs corresponded to alleles in intergenic regions (57%), intronic regions (30%), upstream or downstream of the gene (9.4%). Only a very small percentage (~2%) was found in coding regions, although most instances of this corresponded to synonym variants. Only 15 markers were discarded because they could not be mapped to the UMD 3.1 bovine genome assembly. As a result of this, the last database had 6,662 markers available for use in the remaining assessments. SNPs were found evenly distributed across the entire genome, with the highest number in the large chromosomes and the lowest in the small chromosomes. This made it possible to adequately build a linkage map. After data edition, the SNPs frequencies were biased towards intermediate allele frequencies.

When the threshold for an absolute value of $p>2$ ($p<0.001$), a total of 144 SNPs were significant. However, when the threshold was set to 3 ($p<0.0001$), only 37 markers overcame it (Figure 1). For a threshold of 4 ($p<0.00001$), 10 markers were significant. Finally, only 5 markers were above a significance threshold of 5 ($p<0.000001$). This study shows the results obtained with markers beyond a threshold of 3 ($p<0.0001$); however, only markers with higher values were mapped and annotated.

Figure 2 shows the descriptive analysis of the SNPs with stronger signatures of selection. These SNPs were obtained with a threshold of 3, thus most of them correspond to intergenic regions. In spite of this, an interesting number of variants was found inside the genes or in regulating regions, although not specifically in coding regions (Figure 2A). When the threshold was set to 4, ten markers overcame this limit, half of them were found in intronic regions (50%). One case was reported in upstream variants (10%) and 4 variants were intergenic (40%). This shows a small change in favour of variants located inside genes or related to them. Nevertheless, no visible biological effects were observed.

It is worth mentioning that selection signatures did not appear randomly among the various chromosomes. On the contrary, some showed a more visible selection footprint, with a greater number of markers involved at the peaks. Three of the chromosomes stand out, namely chromosomes 8, 6, and 20, as they have 7, 5, and 5 of the 37 most important variants, respectively (Figure 2B).

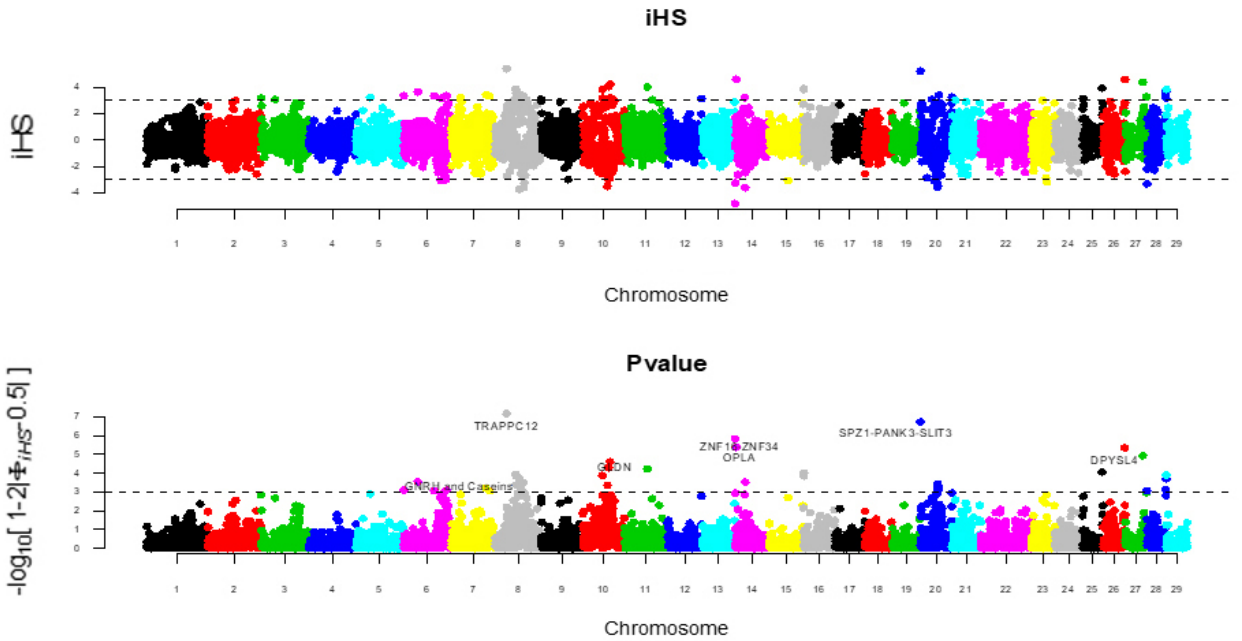

Figure 1. Plot of the *iHS* magnitude by chromosome and associated p-value. The dotted line represents the threshold of significance (p<0.0001).

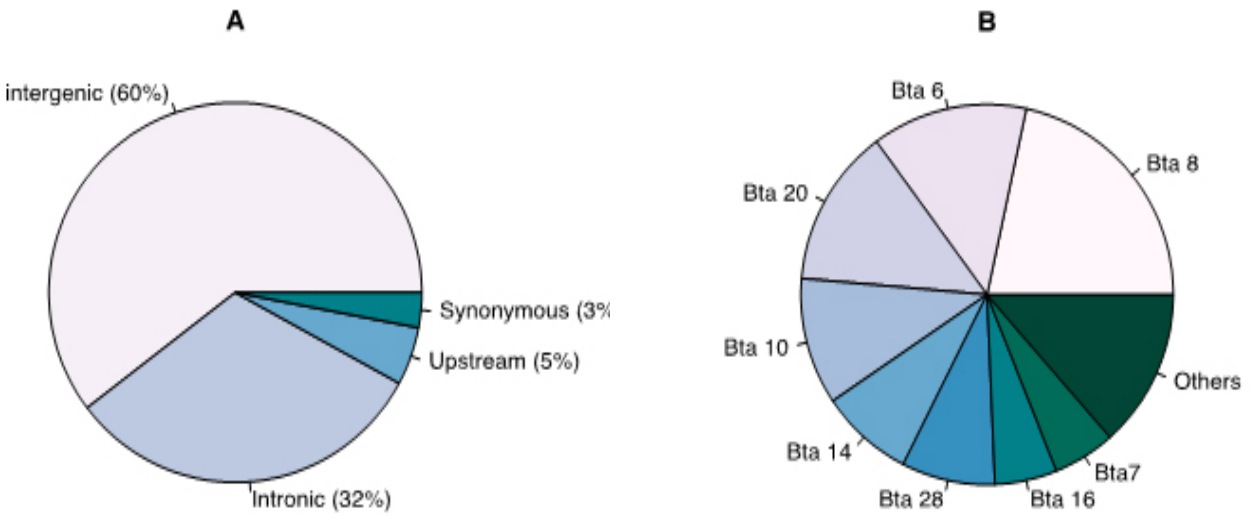

Figure 2. Types of variants (A) and chromosome (B) in which the significant variants (p<0.0001) for signatures of positive selection can be found. Bta: *Bos Taurus* Chromosome.

The most important chromosomes in terms of significant signatures are shown in Figure 3, where it is possible to see how far apart the various peaks are. For this reason, the LD between the various focal markers was low in most cases (LD<0.1), which hints at possible effects of different QTLs inside the same chromosome. In some cases, the SNPs found around the focal SNP had an important signal that could correspond to the same selected QTL. It is worth noting that some chromosomes had strong selection signatures (such as Bta 26 and 27). However, since only one signal was observed in the chromosome, they do not appear in an extended way, which does not mean that they are not important on different principal traits for dairy yield.

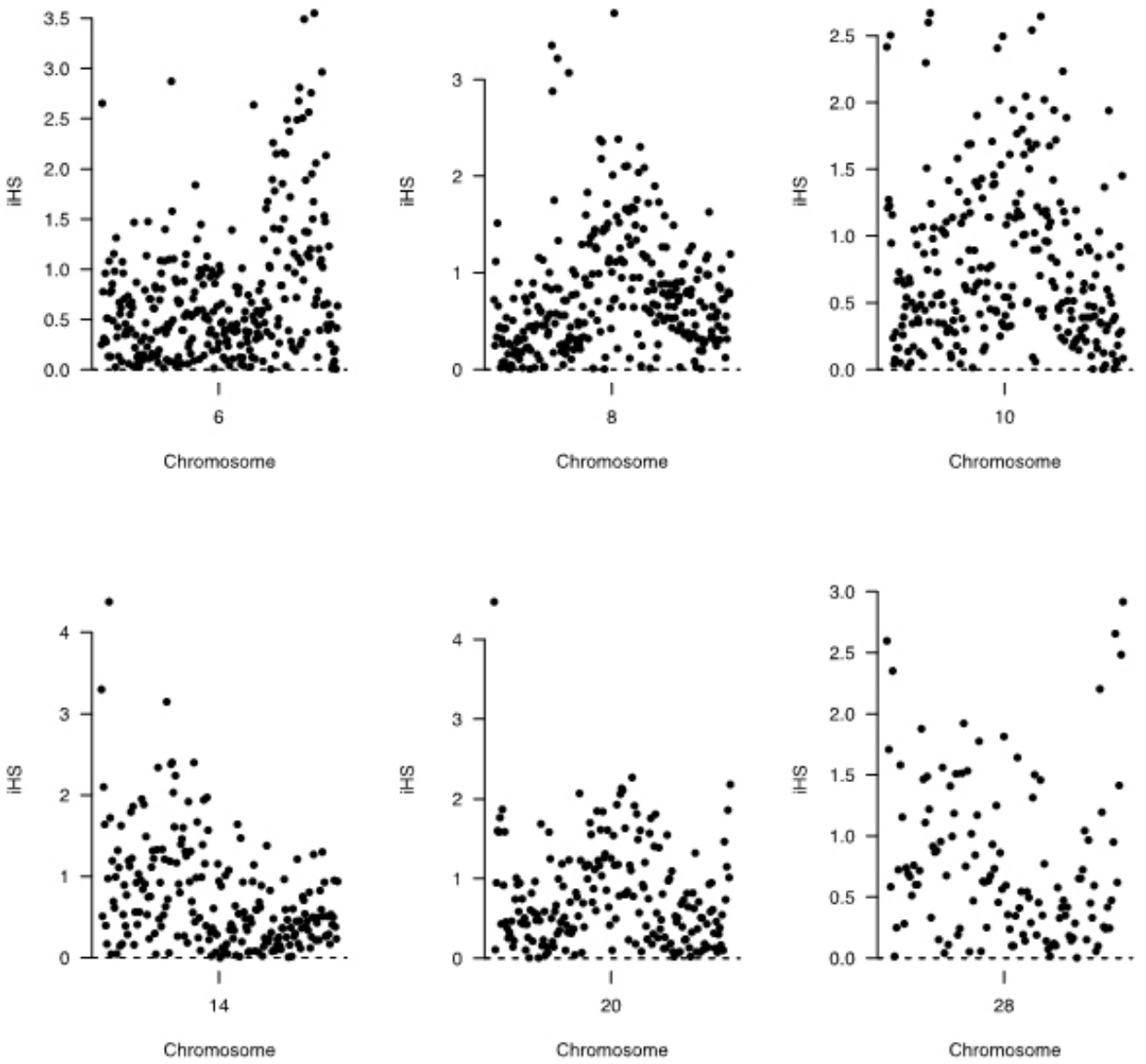

Figure 3. Magnitude plot of the *iHS* extended for chromosomes (Bta) 6, 8, 10, 14, 20, and 28.

Table 1 shows the description of the focal SNPs with the strongest selection signatures ($p \geq 4$). In most cases the SNPs were found inside genes or regulating regions, which suggest a possible effect of the gene on the traits associated with milk yield. Some SNPs were found in intergenic regions, but close to genes that are physiologically important for milk yield. The most important variants had minor allele frequencies ranging from 0.11 to 0.40. However, when including the significant variants with $p<0.0001$, the lowest MAF frequency was 0.087, and the highest was 0.497.

The marker with the strongest selection signature was found in an intron of the TRAPPC12 gene, which is close to some additional genes that are highly important for the immune response of mammals (Table 1). The second highest threshold corresponded to a marker located in an intergenic region composed of a SPZ1-PANK3-SLIT3 group.

Table 1. Description of the polymorphisms with stronger signatures of recent selection and their location in the UMD 3.1 bovine genome assembly.

rs code	Chromosome/ position	MAF	Gene/Consequence	Function
rs110170318	8:112966306	0.32	TRAPPC12/ (Intronic variant)	Plays a role in the vesicular transportation of proteins. Multiple transcripts per alternative splicing. Associated with multiple myeloma in humans
rs41565912	20:247319	0.21	Between PANK3 and SLIT3 gene (Intergenic variant)	PANK3: regulates coenzyme A biosynthesis SLIT3: Secreted protein with effects on cell migration
rs110090404	14:1463676	0.34	Between ZNF16 and ZNF34 gene (Intergenic variant)	ZNF16: Zinc finger proteins. Acts as a transcriptional activator. Can generate multiple transcripts ZNF34: Related to transcription regulation
rs110339989	14:1954317	0.19	OPLA (Upstream variant)	Catalyzes the cleavage of 5-oxo-L-proline to form L-glutamate coupled to the hydrolysis of ATP to ADP
rs109383357	26:51528962	0.11	DPYSL4 (Intron variant)	Hydrolase activity, important for signalling
rs109602238	27:39750933	0.21	Uncharacterized transcript close to NGLY1 gene	NGLY1: Glycosylates denatured forms of proteins linked to N in the cytoplasm and assists in proteosome-mediated degradation
rs110304273	10:59125626	0.36	GLDN (Intron variant)	Binding proteins involved in heterotypic cell-cell adhesion
rs43637158	10:55751369	0.40	UNC13C (Intron variant)	Plays a role in vesicular maturation during exocytosis
rs29019387	10:103908013	0.29	CTDSPL2 (Intron variant)	Phosphatase activity
rs41572782	11:42713681	0.32	Near BCL11A (Intergenic variant)	It works as a myeloid and B cell proto-oncogene. Lymphopoiesis essential factor
rs41664981	8:30772014	0.36	Close to MPDZ (Intergenic variant)	Causes clustering on the surface of cells
rs29026969	29:630087	0.32	PANX1 (Intronic variant)	Structural components of intercell junctions and hemichannels. Plays a role in calcium homeostasis (Ca^{2+} channel)

It is interesting to note that chromosome 6 had a significant peak (Figure 1; $p<0.0001$), which was not the largest but still corresponded to an SNP (rs110527224) that is very close to the casein genes (CSN1S1, CSN2, CSN1S2, CSN3) and to the GnRH gene. Moreover, the two most important peaks in chromosome 14 were in positions close to and surrounding the DGAT1 gene.

Figure 4 shows the *EHH* decay for the 6 focal markers related to the strongest signatures of selection. Significant differences between the ancestral and derived scenarios can be observed. It is worth noting that some of the focal markers are located at the start of the chromosome, therefore some of the *EHH* decay graphs are shortened.

In most of the cases shown in Figure 4, the haplotypes in the ancestral allele were larger than what was found in the derived allele. However, in one case (rs110339898), the haplotype of the derived allele was larger than its ancestral counterpart, which suggests stronger selection pressure in this allele.

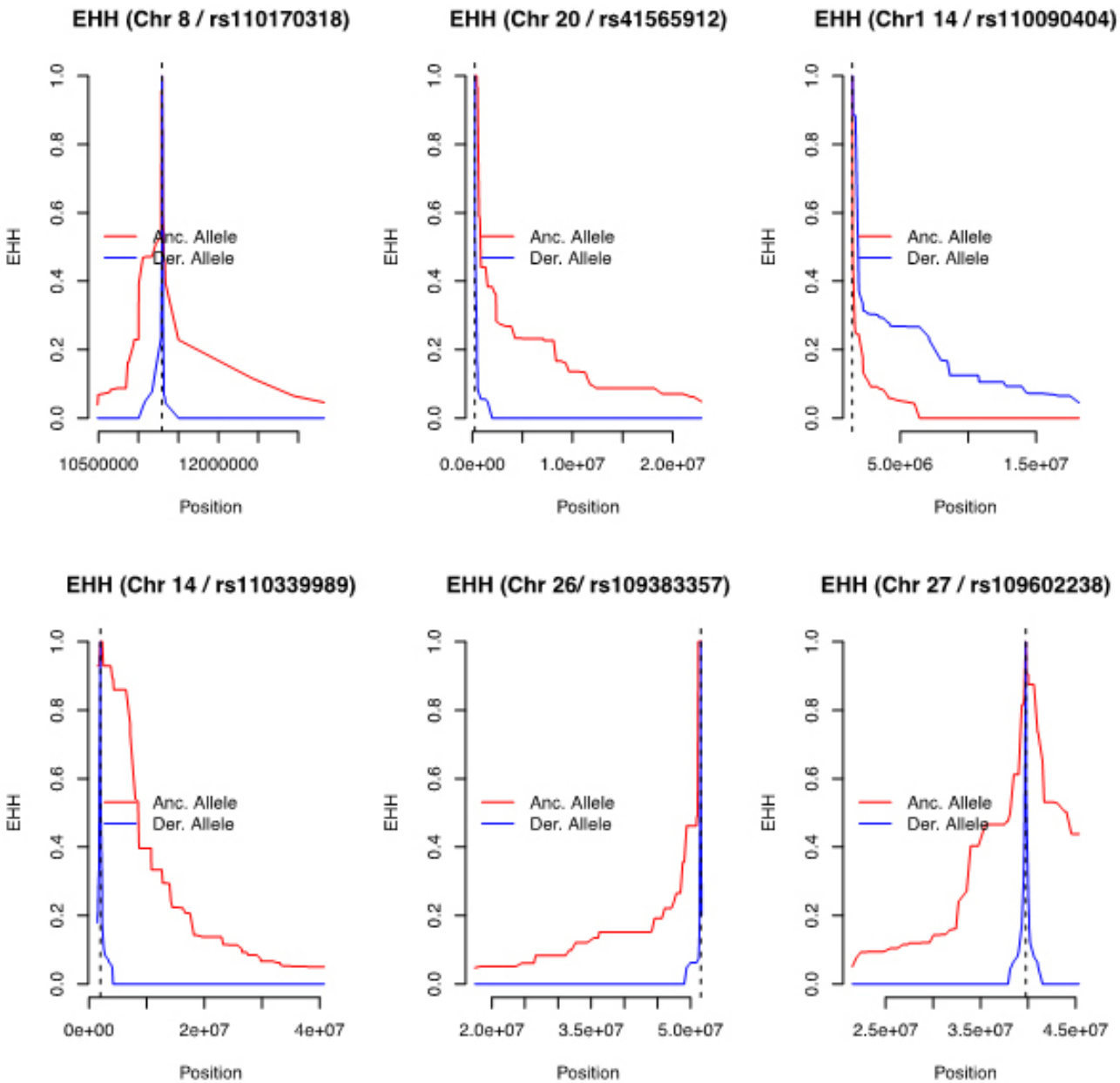

Figure 4. Extended haplotype homozygosity decay of the ancestral and derived alleles based on the 6 focal SNPs with stronger selection signatures in Holstein Cattle from Colombia.

Discussion

The greatest number of focal variants with strong selection signatures ($p<0.0001$) showed minor allele frequencies ranging from 0.087 to 0.497, with a mean near to 0.347. For this reason, it is clear that almost all focal markers, as well as the associated haplotype are segregating and have intermediate frequencies (Table 1). The marker with the lowest MAF values was SNP rs43497903, which is located inside an intron of gene AFAP1, which, according to the UMD 3.1 bovine genome assembly, is related to actin and phospholipid binding (McLaren *et al*., 2010).

It is worth noting that the BovineLD chip contains many SNPs that are segregating in the Holstein population. Even the focal markers subjected to selection pressure show medium allele frequencies, which makes it possible to exert some form of selection on the detected variants. We must clarify that the *iHS* is better at detecting selection when the alleles have intermediate frequencies. However, for the cases where allele frequency is very high, methodologies have been proposed which use more than one population. Such is the case of the methodologies using the *Rsb* statistic to increase detection power (Gautier and Vitalis, 2012). The power of the *iHS* also depends on other additional factors including SNP density, SNP selection bias, demographic history, availability of information in the haplotype phase, and selection intensity (Fay and Wu, 2000; Voight *et al*., 2006; Hayes *et al*., 2008). In addition, Hayes *et al*. (2008) reported a decrease in power under simulation scenarios when the selection pressure was too strong and there were many alleles of the haplotype near the fixation point. However, this was not the case in this study where most of the allele frequencies were intermediate.

As for genotyping density, it is necessary to take into account the effective number of the Holstein cattle population, which favours the presence of large haplotypes (Daetwyler *et al*., 2014). This in turn makes it possible to detect recent selection signatures in Holstein cattle even with the BovineLD chip, which has 6,909 markers with a spacing of 383 kb (Illumina, 2013). In this regard, Hayes *et al*. (2008) found significant linkage in chromosome 6 when using a relatively low marker density. According to Hayes *et al.* (2008), when the *iHS* values exceed the threshold of significance, the length of the population's haplotypes can be detected at up to 1 Mb of distance from the site where the selection on a QTL has increased the frequency of the favorable allele. In fact, even the homozygosity around the selected site can stretch beyond that value, reaching up to 40 Mb, as reported in a study conducted on canines (Pollinger *et al*., 2005). Considering this, and the fact that the effective size of the Holstein population is small (Pérez-Enciso, 2014) it is possible to build a linkage map even with a panel of markers that is not as dense as that of this study. Moreover, a study conducted by Hayes *et al*. (2008) used 403 SNPs inside chromosome 6, a density similar to some chromosomes in our study.

Linkage disequilibrium to find selection signatures is an interesting method that has been used successfully (Voight *et al*., 2006; Hayes *et al*., 2008). However, it is necessary to keep in mind that the LD in large intervals can be caused by a number of factors that are not necessarily related to the selection process (Hill, 1981). Nevertheless, tests utilizing *iHS* have been effectively used to detect recent selection signatures, both in humans (Voight *et al*., 2006) and in bovine cattle of different breeds (Hayes *et al*., 2008; Qanbari *et al*., 2014, 2011; Snelling *et al*., 2007). The identification of genome regions that have been under selection pressure may be used to find regions associated with QTLs which could later enrich genetic assessment programs through the use of *a priori* information (Pérez-Enciso *et al*., 2015).

It is worth mentioning that Holstein selection in Colombia has focused on the search for high milk yield with animals grazing on the mountains. Nevertheless, efforts have recently been directed to improve the amount of protein and fat in milk, thus disregarding other traits that have an economically important effect on milk yield. The strong selection signatures obtained in this study were associated with regions that have been reported (Table 2) to have QTLs particularly related to milk, fat and protein yield, which is consistent with the stated selection aims. Some QTLs were also associated to other traits, such as reproduction, growth and size, which usually have some form of inverse association with the most relevant milk yield traits. Even height has been reported as a trait strongly influenced by domestication (Karim *et al*., 2011) and has undergone strong selection pressure (Fortes *et al*., 2013).

Table 2. Polymorphisms with the strongest signature of recent selection and the previous reports of mapped QTLs from the *cattle QTLdb* (Hu *et al*., 2013).

Chromosome /rs code	Placement	Associated QTL type
8/rs110170318	In the interval attributed to a QTL	Associated with meat tenderness in Charolais, Simmental and Angus cattle (Rolf *et al*., 2011)
20/rs41565912	In the interval attributed to a QTL	Associated with milk protein yield in Holstein (Lund *et al*., 2008), body weight (Cole *et al*., 2011) and others
14/rs110090404	In the interval attributed to a QTL	Associated with milk, fat, protein yield in Nordic Red cattle (Iso-Touru *et al*., 2016)
14/rs110339989	In the interval attributed to a QTL	Associated with milk yield, the amount of fat and protein in milk, somatic cell scores in Holstein cattle (Iso-Touru *et al*., 2016)
26/rs109383357	Close to a QTL (<0.25 Mb of a QTL)	Associated with body weight at weaning and weight at birth in beef cattle (Lu *et al.,* 2013)
27/rs109602238	In the interval attributed to a QTL	Associated with milk fat acid percentage in Dutch dairy cattle (Bouwman *et al*., 2011)
10/rs110304273	In the interval attributed to a QTL	Associated with ease of calving, fertility and milk protein percentage in Holstein cattle (Höglund *et al*., 2009)
10/rs43637158	In the interval attributed to a QTL	Associated with the content of docosahexaenoic acid in Holstein x Charoláis cattle (Gutierrez-Gil *et al*., 2010)
10/rs29019387	In the interval attributed to a QTL	Associated with milk yield, protein and fat yield in Holstein (Milanesi *et al.,* 2008)
11/rs41572782	In the interval attributed to a QTL	Associated with milk kappa-casein percentage (Buitenhuis *et al*., 2016)
8/rs41664981	In the interval attributed to a QTL	Associated with parturition traits in Holstein cattle (Kühn *et al*., 2003)
29/rs29026969	Close to a QTL (<0.25 Mb of a QTL)	Associated with milk production, health and reproductive traits in Holstein cattle (Ashwell *et al.,* 2004)

Table 2 shows previous reports of QTLs located in the confidence interval to the focal SNP identified in this study. The description of the focal markers with the strongest signatures of selection as well as the identification of the QTLs associated with that interval were conducted in accordance with the report stored in the *Animal QTLdb* public database, section *cattleQTLdb* (Hu *et al*., 2013). It is worth noting that almost all regions reported as candidates were associated with the confidence interval for a previously reported QTL. Most of these QTLs were related to reproductive traits such as milk yield, fat and protein content, although the confidence interval was often large. Interestingly, two regions were also recently reported by Iso-Touru *et al*. (2016) in Holstein cattle (Table 2). The markers present in the same chromosome were often associated with different QTLs, suggesting that those QTLs do not overlap. Additionally, the linkage disequilibrium between focal markers was low in those cases (r^2<0.1).

The selection signatures presented in this paper were not evenly distributed throughout all the autosomes of the bovine genome. On the contrary, some chromosomes had higher selection pressure (Figure 2), which means that an important group of QTLs could be inside the same chromosome and not randomly distributed across the bovine genome. Some studies on dairy cattle have found similar distributions (Qanbari *et al*., 2011; Ramey *et al*., 2013). They have even shown variations between different breeds as a result of differential selection (Qanbari *et al*., 2014).

Interestingly, some studies have reported selection signatures inside chromosome 6 in dairy bovines (Hayes *et al*., 2008; Qanbari *et al*., 2014). In this study, some significant peaks were found in chromosome 6. There was even an interesting signature corresponding to an SNP (rs110527224) that was very close to the casein group (CSN1S1, CSN2, CSN1S2, CSN3). The signature was also near the GnRH gene, which could have been subjected to strong selection pressure. However, the most important selection regions did not occur in this chromosome, and those which appeared did not necessarily match those reported by the previously mentioned studies (Hayes *et al*., 2008; Qanbari *et al*., 2014) possibly because the

causal mutation has not yet been found, which raises the need for fine mapping studies in the QTL region.

It is interesting to clarify that the DGAT1 gene in bovine cattle has been repeatedly proposed as a major gene for milk fat (Grisart *et al*., 2002; Winter *et al*., 2002; Kühn *et al*., 2004; Wang *et al*., 2012), yet this study did not find any selection peak in the region occupied by the gene. Furthermore, according to the NCBI's bovine genome assembly —UMD 3.1—, the gene is located at chromosome 14 between positions 1,795,425-1,804,838b, at the beginning of the chromosome, in the centromeric region (Brown *et al*., 2015), near the area where we found the largest selection peak for chromosome 14 (Figure 3). It is important to note that, in spite of the fact that the chip has a variant inside the gene, no direct selection signature was found. However, the two most important variants were found at positions 1,463,676b and 1,954,317b, which surround the location of the gene and may be associated with regulating sequences, promoters or enhancers. In spite of this, the absence of a selection signature directly on the gene is not evidence of the absence of a QTL in the region; on the contrary, this study reaffirms the notion that this is a region of great importance where there could be QTLs related to milk fat. Thus, detailed, fine mapping studies should be conducted to attempt reducing the confidence interval for the QTL in order to find possible causal mutations or map the QTLs of the region in a more precise manner.

It is worth noting that the regions of the genome associated with traits of importance for milk yield identified by recent selection signatures are typical of the assessed population or of very close populations which have a very similar demographic history as well as genetic improvement and selection systems with similar criteria and objectives. For this reason, it is possible that studies on positive or negative selection made in different regions of the world could yield different results depending on the criteria, aims, time, demographic history and selection intensity of the genetic improvement programs for the assessed breed (Stella *et al*., 2010; Utsunomiya *et al*., 2013).

Finally, it is worth mentioning that, despite having various mapping approaches for over a decade, it is still difficult to find historical selection events on complex traits (Kemper *et al*., 2014) and different results may be obtained depending on the methodology applied (Qanbari *et al*., 2011). However, the combination of different tests might improve the ability to localize the candidate regions under selection (Utsunomiya *et al*., 2013), which is very important considering the relevance of QTN detection and knowledge of the genetic architecture of the traits of importance for milk yield to be used in genetic improvement and genomic selection programs (Pérez-Enciso *et al*., 2015). The objective of all this is the possibility of using methods that use causal variants such as "genome editing" in genomic selection in the near future. To achieve this, it would be advisable to use denser platforms or sequence data, including slightly more powerful approaches.

Conflict of interest

The authors declare they have no conflicts of interest with regard to the work presented in this report.

References

ACHF. Asociación Holstein de Colombia: 67 años de historia al servicio de la ganadería en Colombia. Holstein Colombiana 2009; 177:6-13.

Ashwell MS, Heyen DW, Sonstegard TS, Van Tassell CP, Da Y, VanRaden PM, Ron M, Weller JI, Lewin H. Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. J Dairy Sci 2004; 87:468-475.

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

Bouwman AC, Bovenhuis H, Visker MHPW, van Arendonk J.M. Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet 2011; 12:43.

Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T, Pruitt KD, Maglott DR, Murphy TD. Gene: A gene-centered information resource at NCBI. Nucleic Acids Res 2105; 43:D36-D42.

Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 2009; 84:210-223.

Buitenhuis B, Poulsen NA, Gebreyesus G, Larsen LB. Estimation of genetic parameters and detection of chromosomal regions affecting the major milk proteins and their post translational modifications in Danish Holstein and Danish Jersey cattle. BMC Genet 2016; 17:114.

Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker B, Van Tassell CP, Yang J, Wang S, Matukumalli LK, Da Y. Genome-wide association analysis of thirty-one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics 2016; 12:408.

Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, Esquerré D, Bouchez O, Rossignol M-N, Klopp C, Rocha D, Fritz S, Eggen A, Bowman PJ, Coote D, Chamberlain AJ, Anderson C, VanTassell CP, Hulsegge I, Goddard ME, Guldbrandtsen B, Lund MS, Veerkamp RF, Boichard D, Fries R, Hayes BJ. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 2014.

Druet T, Pérez-Pardal L, Charlier C, Gautier M. Identification of large selective sweeps associated with major genes in cattle. Anim Genet 2013; 44:758-762.

Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. Genetics 2000; 155:1405-1413.

Fortes MRS, Kemper K, Sasazaki S, Reverter A, Pryce JE, Barendse W, Bunch R, McCulloch R, Harrison B, Bolormaa S, Zhang YD, Hawken RJ, Goddard ME, Lehnert SA. Evidence for pleiotropism and recent selection in the PLAG1 region in Australian Beef cattle. Anim Genet 2013; 44:636-647.

Gautier M, Vitalis R. Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 2012; 28:1176-1177.

Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelman R, Georges M, Snell R. 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 Res 2002; 12:222-231.

Gutiérrez-gil B, Wiener P, Richardson RI, Wood JD, Williams JL. Identification of QTL with effects on fatty acid composition of meat in a Charolais × Holstein cross population. Meat Sci 2010; 85:721-729.

Hayes BJ, Lien S, Nilsen H, Olsen HG, Berg P, Maceachernb S, Potter S, Meuwissen THE. The origin of selection signatures on bovine chromosome 6. Anim Genet 2008; 39:105-111.

Hill WG. Estimation of effective population size from data on linkage disequilibrium. Genet Res 1981; 38:209-216.

Höglund JK, Buitenhuis AJ, Guldbrandtsen B, Su G, Thomsen B, Lund MS. Overlapping chromosomal regions for fertility traits and production traits in the Danish Holstein population. J Dairy Sci 2009; 92:5712-5719.

Hu ZL, Park CA, Wu XL, Reecy JM. Animal QTLdb: An improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res 2013; 41.

Illumina. BovineLD Genotyping BeadChip 2013; 3-6; [Access date: July 20, 2016]

Iso-Touru T, Sahana G, Guldbrandtsen B, Lund MS, Vilkki J. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genet 2016; 17:55.

Karim L, Takeda H, Lin L, Druet T, Arias JAC, Baurain D, Cambisano N, Davis SR, Farnir F,Grisart B, Harris BL, Keehan MD, Littlejohn MD, Spelman RJ, Georges M, Coppieters W. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nat Genet 2011; 43:405-413.

Kemper KE, Saxton SJ, Bolormaa S, Hayes BJ, Goddard ME. Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics 2014; 15:246.

Kuhn C, Bennewitz J, Reinsch N, Xu N, Thomsen H, Looft C, Brockmann GA, Schwerin M, Weimann C, Hiendleder S, Erhardt G, Medjugorac I, Forster M, Brenig B, Reinhardt F, Reents R, Russ I, Averdunk G, Blumel J, Kalm E. Quantitative trait loci mapping of functional traits in the German Holstein cattle population. J Dairy Sci 2003; 86:360-368.

Kühn C, Thaller G, Winter A, Bininda-Emonds ORP, Kaupe B, Erhardt G, Bennewitz J, Schwerin M, Fries R. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. Genetics 2004; 167:1873-1881.

Lu D, Miller S, Sargolzaei M, Kelly M, Voort GV, Caldwell T, Wang Z, Plastow G, Moore S. Genome-wide association analyses for growth and feed efficiency traits in beef cattle. J anim Sci 2013; 91:3612-3633.

Lund M, Sorensen P, Madsen P, Jaffrézic F. Detection and modelling of time-dependent QTL in animal populations. Genet Sel Evol 2008; 40:177-194.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 2010; 26:2069-2070.

Meuwissen T, Goddard M. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 2010; 185:623-631.

Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics 2001; 157:1819-1829.

Milanesi, E., Negrini, R., Schiavini, F., Nicoloso, L., Mazza, R., Canavesi, F., Miglior, F., Valentini A, Bagnato A, Ajmone-Marsan P. Detection of QTL for milk protein percentage in Italian Friesian cattle by AFLP markers and selective genotyping. J Dairy Res 2008; 75:430-438.

Pérez-Enciso M. Genomic relationships computed from either next-generation sequence or array SNP data. J Anim Breed Genet 2014; 131:85-96.

Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol 2015; 47:43.

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

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81:559-575.

Qanbari S, Gianola D, Hayes B, Schenkel F, Miller S, Moore S, Thaller G, Simianer H. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics 2011; 12:318.

Qanbari S, Pausch H, Jansen S, Somel M, Strom TM, Fries R, Nielsen R, Simianer H. Classic Selective Sweeps Revealed by Massive Sequencing in Cattle. PLoS Genet 2014; 10.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna Austria 2012.

Ramey HR, Decker JE, McKay SD, Rolf MM, Schnabel RD, Taylor JF. Detection of selective sweeps in cattle using genome-wide SNP data. BMC Genomics 2013; 14:382.

Rolf MM, Taylor JF, Schnabel RD, Mckay SD, Mcclure MC, Northcutt SL, Kerley MS, Weaber RL. Genome-wide association analysis for feed efficiency in Angus cattle. Anim Genet 2011; 43:367-374.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, 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.

Snelling WM, Chiu R, Schein JE, Hobbs M, Abbey C, Adelson DL, Aerts J, Bennett GL, Bosdet IE, Boussaha M, Brauning R, Caetano AR, Costa MM, Crawford AM, Dalrymple BP, Eggen A, Everts-van der Wind A, Floriot S, Gautie M, Gill C, Green RD, Holt R, Jann O, Jones SJ, Kappes SM, Keele JW, de Jong PJ, Larkin DM, Lewin H, McEwan JC, McKay S, Marra M, Mathewson C, Matukumalli LK, Moore SS, Murdoch, B, Nicholas FW, Osoegawa K, Roy A, Salih H, Schibler L, Schnabel RD, Silveri L, Skow LC, Smith TP, Sonstegard TS, Taylor JF, Tellam R, Van Tassell CP, Williams JL, Womack JE, Wye NH, Yang G, Zhao S. A physical map of the bovine genome. Genome Biol 2007; 8:R165.

Stella A, Ajmone-Marsan P, Lazzari B, Boettcher P. Identification of selection signatures in cattle breeds selected for dairy production. Genetics 2010; 185:1451-1461.

Utsunomiya YT, Pérez O'Brien AM, Sonstegard TS, Van Tassell CP, do Carmo AS, Mészáros G, Sölkner J, Garcia JF. Detecting Loci under Recent Positive Selection in Dairy and Beef Cattle by Combining Different Genome-Wide Scan Methods. PLoS One 2013; 8.

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

Wang X, Wurmser C, Pausch H, Jung S, Reinhardt F, Tetens J, Thaller G, Fries R. Identification and dissection of four major QTL affecting milk fat content in the German Holstein-Friesian population. PLoS One 2012; 7:e40711.

Winter A, Krämer W, Werner FO, Kollers S, Kata S, Durstewitz G, Buitkamp J, Womack JE, Thaller G, Fries R. 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. Proc Natl Acad Sci USA 2002; 99:9300-9305.