Genomic data as the "hitchhiker's guide" to cattle adaptation: tracking the milestones of past selection in the bovine genome

Yuri T. Utsunomiya[1], Ana M. Pérez O'Brien[2], Tad S. Sonstegard[3], Johann Sölkner[2] and José F. Garcia[1,4]*

[1] Departamento de Medicina Veterinária Preventiva e Reprodução Animal, Faculdade de Ciências Agrárias e Veterinárias, Universidade Estadual Paulista (UNESP), Jaboticabal, São Paulo, Brazil
[2] Division of Livestock Sciences, Department of Sustainable Agricultural Systems, University of Natural Resources and Life Sciences (BOKU), Vienna, Austria
[3] Animal Genomics and Improvement Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MA, USA
[4] Laboratório de Bioquímica e Biologia Molecular Animal, Departamento de Apoio, Saúde e Produção Animal, Faculdade de Medicina Veterinária de Araçatuba, Universidade Estadual Paulista (UNESP), Araçatuba, São Paulo, Brazil

The bovine species have witnessed and played a major role in the drastic socio-economical changes that shaped our culture over the last 10,000 years. During this journey, cattle "hitchhiked" on human development and colonized the world, facing strong selective pressures such as dramatic environmental changes and disease challenge. Consequently, hundreds of specialized cattle breeds emerged and spread around the globe, making up a rich spectrum of genomic resources. Their DNA still carry the scars left from adapting to this wide range of conditions, and we are now empowered with data and analytical tools to track the milestones of past selection in their genomes. In this review paper, we provide a summary of the reconstructed demographic events that shaped cattle diversity, offer a critical synthesis of popular methodologies applied to the search for signatures of selection (SS) in genomic data, and give examples of recent SS studies in cattle. Then, we outline the potential and challenges of the application of SS analysis in cattle, and discuss the future directions in this field.

Keywords: positive selection, *Bos taurus*, *Bos indicus*, adaptation, single nucleotide polymorphism

INTRODUCTION

The ability of domestic cattle to convert low-quality forage into meat, milk, and draft power is of direct importance to the livelihood of the human species. This ability is tightly linked to the adaptation of indigenous and exotic cattle to the diverse environments found around the world, which may result from complex—mostly untold—stories of migration, expansion, exposure to diseases, admixture, climate changes, and selective pressures (Ajmone-Marsan, 2010). These past events have shaped the genetic diversity of domestic cattle throughout history, and their present genomes may shelter tractable signatures of these phenomena.

Footprints of selection, such as specific patterns of change in allele frequencies, diversity loss, and haplotype structure, are currently detectable from single nucleotide polymorphism (SNP) data by well-established methodologies (Sabeti et al., 2006; Oleksyk et al., 2010), and can unravel past responses of the cattle genome to natural and human-driven selection, as well as evidences of loci and variants underlying adaptive and economically important traits. Detecting these selection signatures (SS) may not only help to shed some light on the key adaptive events that have generated the enormous phenotypic variation observed

between cattle breeds today, but can also be of biotechnological relevance.

In recent years, SS studies are becoming increasingly popular because they offer a complementary strategy to genome-wide association (GWA) studies on mapping variants impacting traits of interest, helping to link phenotypes to gene function. In a typical GWA analysis, one starts from a phenotype and scan genotypes to identify underlying large and moderate effect variants (Bush and Moore, 2012). Generally, SS studies take the opposite direction: one starts with an evidence of selection in samples sharing geographical proximity, environmental factors or a common phenotype, and attempts to find selected mutations (Sabeti et al., 2006).

The major motivation for SS studies is that this type of approach can pinpoint chromosomal segments sheltering large effect mutations even if they no longer segregate in a population. In such cases, these variants cannot be detected by classical quantitative genetics methods unless linkage experiments are designed using crosses (Ramey et al., 2013). Another appealing feature of SS studies is that they typically require over 10 fold smaller sample sizes in comparison to GWA studies. Moreover, SS can reveal signals on genes controlling traits that are difficult, expensive or

even impossible to measure on a large population (i.e., disease resistance).

Although identifying SS is of paramount importance to uncover variants responsible for adaptive traits, its application to cattle data must be carefully interpreted as important demographic events such as severe population bottlenecks, genetic drift, and admixture, as well as confounding effects derived from the development of SNP panels, may give rise to false signals. Furthermore, SS have serious limitations in targeting specific traits, so assigning signals to phenotypes is a non-trivial problem.

Here, we provide a brief review of the demographic history of cattle as it is known, present a critical review of some of the main methodologies available for the detection of putative loci under selection, and provide examples of recently published results in cattle. Then, we outline the potential and the challenges of the application of these methods to cattle data, and discuss the future directions in this field.

CATTLE DEMOGRAPHIC HISTORY
DIFFERENTIATION AND DOMESTICATION
The humpless taurine (*Bos taurus*) and the humped indicine or zebu cattle (*Bos indicus*) descend from the wild ox or auroch (*Bos primigenius*), which has been extinct since 1627 (Mona et al., 2010). The two populations of wild aurochsen that formed the ancestral pool for these interfertile cattle species may have diverged some 280,000 years ago (Murray et al., 2010), and were subjected to many demographic events before being domesticated by our species, including severe bottlenecks, admixture and natural selection.

Although the domestication of taurine and zebu cattle is still an open question and an active field of investigation, evidences collected over 20 years of research on molecular genetic diversity of cattle (see Groeneveld et al., 2010, for further review), combined with historic and archaeological data, support that these species were independently domesticated in at least two episodes (Loftus et al., 1994; Bruford et al., 2003): some 10,000 years ago, taurine animals were captured from the wild in the Fertile Crescent (modern-day countries of Israel, Jordan, Lebanon, Cyprus and Syria, and parts from Egypt, Turkey, Iraq, Iran, and Kuwait); 1500 years later, zebu cattle were domesticated in the Indus valley (present Northeastern Afghanistan, Pakistan, and Northwestern India).

EXPANSION
Taurine cattle have an almost cosmopolitan distribution today. From the domestication centre in Southwestern Asia, they followed human migrations and slowly expanded over Asia and Europe. An independent domestication episode of taurine cattle in Northeastern Africa is disputed, but as molecular data are not conclusive, the divergence of African and Eurasian taurine populations from a common ancestor domesticated in the Fertile Crescent is well supported (Hanotte et al., 2002; Ajmone-Marsan, 2010; Decker et al., 2014). Other isolated micro-events of domestication are also not discarded, not even interbreeding of domestic animals with wild aurochsen (Bonfiglio et al., 2010). Taurine cattle reached the New World by European importations after the discovery of America in the late fifteenth century, and their descendants living today are broadly referred as Creole cattle.

Zebu cattle also spread around the world accompanying human migrations, but became more endemic in tropical and subtropical regions due to their adaptability to these environments. Zebu cattle were probably introduced in the African continent in the seventh century by imports of *B. indicus* sires during Arabian invasions (Bradley et al., 1996). Indian zebu cattle were only introduced in Central and South America in the early twentieth century, and started to be massively imported to the American continent after 1950.

FORMATION OF SPECIALIZED BREEDS AND FURTHER CATTLE ADAPTATION
After domestication, farmers started to control cattle mating according to their interest in traits such as size, behavior, and milk production. This "breed the best to the best and hope for the best" strategy exerted a high artificial selective pressure, triggering a severe decline in cattle effective population size. Estimates from a genome-wide linkage disequilibrium analysis using a medium density SNP panel suggested that domestication was responsible for a 50 to 70-fold decline in the effective population size in comparison to the wild ancestors of taurine and zebu cattle (The Bovine HapMap Consortium, 2009).

About 200 years ago, farmers invented the concept of breed and started to limit germplasm exchange in order to standardize cattle populations based on morphology and performance, giving rise to over 1200 cataloged cattle breeds today (Taberlet et al., 2008; see also: http://www.ansi.okstate.edu/breeds/cattle/). In spite of this apparent rich source of genetic resources, over 200 additional breeds are already extinct, and many others are at risk (FAO, 2007).

Due to their high productive performance for milk and meat traits, some cattle breeds were adopted worldwide, such as the Holstein-Friesian, which is present in 128 countries (FAO, 2007). However, most of these specialized groups of genetically distinct animals are local, and even though they are not under the spotlights, they exhibit high adaption to their environments. Therefore, small local breeds should be considered as primary targets of SS studies, as several of them are endangered and their genomes shelter footprints of adaptation.

IMPLICATIONS TO POSITIVE SELECTION MAPPING
In theory, neutrally evolving alleles form the majority of the genome variation within and between species, and events such as population contraction, expansion, migration, isolation and admixture are responsible for random drift of such alleles. While the demographic history of a population determines its neutral allele frequency spectrum, alleles that impact fitness, survival, reproduction, or traits of human interest are subjected to natural or artificial selection, such that their frequencies deviate from the distribution of neutral alleles. Therefore, mapping loci under selection implicates searching for outlier alleles that substantially differ from the genome background.

The idea of outlier loci is dependent upon knowledge about the frequency distribution of neutrally evolving alleles. As different populations have distinct demographic histories, their allele frequency spectrum also differ. Consequently, statistical methods designed to map loci under selection must be calibrated to either a demographic model for the population under study or

to the empirical distribution of scores across the genome, assuming most of the genome is evolving under neutrality. As most of the events that shaped cattle diversity at the species and breed level are still obscure, methods that are robust across different demographic scenarios that take advantage of genome-wide scores to detect outlier loci are more appropriate to analyze cattle data. Nevertheless, one can use the demographic model for cattle differentiation, domestication and expansion provided in this section and combine with specific models for specialized cattle breeds to simulate the distribution under neutrality and compare with empirical data.

METHODS FOR DETECTING POSITIVE SELECTION IN THE CATTLE GENOME

As proposed by Darwin and Wallace (1858), positive selection is the phenomenon where phenotypes that increase the likelihood of survival and reproduction (i.e., that increase an individual's fitness) become more prevalent in populations over time. In the context of the genome sequence, if a specific allele confers advantage, its carrier is more likely to thrive and leave more offspring than non-carriers, causing the haplotype containing that beneficial allele to spread quickly and increase in frequency in the population (Sabeti et al., 2002, 2006).

The majority of the genetic variation found within and between populations is deemed to have little or no effect on fitness, so that haplotype frequencies follow Hardy-Weinberg expectations (Kimura, 1968; Hellmann et al., 2003). Therefore, positive selection leaves distinctive tractable patterns of genetic variation that differ from the neutrally evolving background DNA sequence. These patterns are broadly referred as "signatures" or "footprints" of selection (Oleksyk et al., 2010).

In this section, we describe classes of signatures in terms of different signals. In each class, whenever convenient, we describe the basis of some popularly adopted methods to highlight their strengths and weaknesses for cattle studies. For a broader overview, we suggest the reader to consult other previously published reviews on the topic (Sabeti et al., 2006; Oleksyk et al., 2010; Vitti et al., 2013; Qanbari and Simianer, 2014). Additionally, we provide examples of studies reporting putative loci under selection in different cattle breeds.

LOCAL GENETIC DIVERSITY DEPRESSION

At each generation, recombination shuffles and breaks haplotypes down, producing linkage equilibrium. When selection increases the frequency of an advantageous allele, neighbor variants "hitchhike" and rise in frequency together so quickly that recombination does not prevent linkage disequilibrium, causing an entire chromosome segment to lose diversity. Therefore, positive selection can be probed by searching for chromosome regions where heterozygosity is much lower than expected under neutrality (Oleksyk et al., 2008).

Ramey et al. (2013) scanned Illumina® BovineSNP50 (50 k) genotypes of over 6000 animals from 13 taurine and one zebu breed using an *ad-hoc* method to identify selective sweeps reaching fixation. Briefly, they considered a candidate locus under selection if at least five contiguous SNPs presenting a minor allele frequency (MAF) > 0.01 spanned a chromosome segment

of 200 kb or more. Although the strategy was successful to find validated selective sweeps, such as the *POLL* locus that controls horn development (Georges et al., 1993), they showed that the SNP ascertainment bias of the 50 k assay incurred in false positives in breeds that are genetically distant from the SNP discovery breeds. This is not unexpected, as the influence of the 50 k bias on population genetics parameters, such as heterozygosity, genetic structure and differentiation, is well documented (Decker et al., 2009). Therefore, methods relying on heterozygosity are largely prone to type I errors generated from the SNP discovery process. Ideally, genotype data should be generated from less biased commercial arrays, such as the Illumina® BovineHD (HD), or from re-sequencing efforts.

Another approach to search for deficit of heterozygosity is the identification of islands of runs of homozygosity (ROH). As long stretches of consecutive homozygous genotypes indicate identical-by-descent haplotypes (Gibson et al., 2006; Lencz et al., 2007), ROH have been recently used to characterize genome-wide inbreeding in cattle (Purfield et al., 2012; Ferenčaković et al., 2013; Kim et al., 2013). However, as recombination is random, the distribution of ROH across samples is expected to be highly heterogeneous under neutrality, so genomic hotspots of ROH can be a signal of selective sweep (Curik et al., 2014). Interestingly, the length of the run is negatively correlated with the number of generations back to the selective pressure or inbreeding event (Howrigan et al., 2011), so ROH size can be used to date the age of the selective sweep.

Focusing on the identification of ROH islands produced by recent artificial selection in U.S. Holstein cattle, Kim et al. (2013) proposed a new statistic, namely F_{RL}, that compares local homozygosity between a population under artificial selection and a control population. Briefly, for each SNP, animals are scored as 1 if the locus is encompassed by a ROH or 0 otherwise. Then, F_L is computed as the proportion of animals with scores equal to 1. Finally, F_{RL} is obtained as:

$$F_{RL} = \ln\left(\frac{F_{L(selected)}}{F_{L(control)}}\right)$$

All scores are standardized (i.e., subtracting each value by the average score and dividing by the standard deviation) to produce a distribution with mean zero and variance one. Extreme positive values of F_{RL} represent changes in allele frequency in the artificially selected population in comparison to the control group, and therefore reflect homozygosity attributable to recent inbreeding, drift or selection at the analyzed locus. As for the approach adopted by Ramey et al. (2013), SNP density and ascertainment bias are important confounders that can produce false positive ROH (Ferenčaković et al., 2013), so commercial SNP assays must be used with caution here.

Methods that search for local absence of heterozygosity are most powerful to detect haplotypes that hitchhiked to fixation. However, allowing for some heterozygosity can be of help to detect partial sweeps, i.e., haplotypes under ongoing selection that have not reached fixation yet. For a given bi-allelic site, let n_{REF} be the number of observed reference alleles, n_{ALT} the number of observed alternative alleles, and p and q their respective

frequencies. The expected frequency of heterozygote genotypes under Hardy-Weinberg Equilibrium is given by:

$$H = 2pq = \frac{2n_{REF}n_{ALT}}{(n_{REF}+n_{ALT})^2}$$

This expectation can be easily extended for a chromosomal segment containing multiple bi-allelic sites as:

$$H = \frac{2\Sigma n_{REF}\Sigma n_{ALT}}{(\Sigma n_{REF} + \Sigma n_{ALT})^2}$$

Rubin et al. (2010) proposed running sliding windows across the genome, calculating H for these windows and then standardizing the obtained values. The method named as ZH_p produces standard deviation scores, and extremely negative values represent chromosome windows where the regional diversity is substantially lower than the average genome diversity. Although the method has been successfully used to map candidate variants under selection in chickens (Rubin et al., 2010), dogs (Axelsson et al., 2013) and pigs (Rubin et al., 2012), no studies applying ZH_p to cattle data have been published to date.

CHANGE IN THE ALLELE FREQUENCY SPECTRUM

After a complete selective sweep (i.e., the haplotype containing the selected variant reaches fixation), new mutations slowly restore local diversity in the course of many generations. As mutations are generally rare and may take a large number of generations to drift to high frequency under neutral evolution, the local heterozygosity depression signal is deemed to remain robust for several generations after the selective pressure has occurred.

Newly acquired mutations or derived alleles (i.e., variants that differ from the original or ancestral allele) occur in lower frequency in comparison to ancestral alleles under neutrality, but when they arise within a selective sweep they will hitchhike to high frequency quickly in the selected population. Therefore, another class of signals that emerge after the depression in local diversity is the enrichment for moderate to high frequency derived alleles. Several methods have been proposed in this category, including Tajima's (1983), Fay and Wu's (2000), and $\triangle DAF$ (Grossman et al., 2010).

The limitation of the use of these methods in cattle data is that inference of ancestral alleles should be preferably performed by comparison of domesticated cattle genomes with wild type genomes. As these are not available, ancestral allele information for cattle SNP assays were derived from genotypes of outgroup species, such as Gaur (*Bos gaurus*), Buffalo (*Bubalus bubalis*) and Yak (*Bos grunniens*), which are assumed to descend from a common founder *Bovinae* species (Matukumalli et al., 2009; Utsunomiya et al., 2013). However, as not all SNP probes hybridize against the genomes of these outgroup species, ancestral allele information for cattle is incomplete. Future availability of genome assemblies for these outgroup species may help to better infer ancestral status for common SNPs.

One way to bypass the limitation of ancestral allele information is by using samples from several genetically distinct populations to estimate average allele frequencies that could represent the spectrum in the common ancestral population. In this approach, instead of searching for enrichment of high frequency derived or rare alleles, one may look for a shift in the allele frequency spectrum in one population in comparison to the average across populations. Stella et al. (2010) successfully applied the negative composite log-likelihood (*CLL*) approach, an extension of the composite likelihood ratio (*CLR*) test (Kim and Stephan, 2002; Nielsen et al., 2005) to 13 taurine, 4 zebu and 2 synthetic breeds, and reported several candidate loci under selection. These included *KIT* (mast/stem cell growth factor receptor gene), responsible for the piebald and color sidedness phenotype (Durkin et al., 2012), and *MC1R* (melanocortin 1 receptor gene), incriminated in the black/red coat color in Holstein and Angus (Klungland et al., 1995). These loci were further confirmed using whole genome sequence data of German Fleckvieh (Qanbari et al., 2014).

Briefly, *CLL* is computed for each SNP window as:

$$CLL = -\sum_{i=1}^{k} \ln\left[1 - Pr\left(d < |p_i - \bar{p}_i| \mid \mu_i\right)\right]$$

Where, relative to SNP i, d is any random value from the theoretical distribution of allele frequencies with mean $\mu_i = \bar{p}_i$, $\bar{p}_i$ is the reference allele frequency averaged across populations, and p_i is the allele frequency in the population being investigated. The theoretical distribution of allele frequencies can be modeled as a binomial or a normal approximation to the binomial distribution. As calculations are very straightforward and only require a dataset with multiple breeds instead of ancestral allele information, it is general enough for cattle SNP data.

Another extension of the *CLR* statistic was proposed by Chen et al. (2010), namely *XPCLR* (cross-population composite likelihood ratio test), which attempts to detect a shift in allele frequency in a target population in respect to a reference population. Lee et al. (2014) applied *XPCLR* to Holstein and Hanwoo whole genome sequence data and reported that the chromosome segment encompassing the kappa-casein gene (*CSN3*) exhibited high *XPCLR* scores in Holstein cattle.

LONG-RANGE HAPLOTYPES

The concept of Extended Haplotype Homozygosity (*EHH*), first introduced by Sabeti et al. (2002), attempts to identify haplotypes that increased so rapidly in frequency that recombination could not substantially break them down, so linkage disequilibrium presents a long-range persistency. Briefly, consider N chromosomes in a sample, and G unique haplotypes extending from a core SNP site to an upstream or downstream position x, with each group g having n_g observations. The *EHH* score for the entire sample is calculated as:

$$EHH = \frac{\sum_{g=1}^{G} \binom{n_g}{2}}{\binom{N}{2}}$$

This score serves as a proxy for the probability of identity-by-descent of haplotypes within the chromosome segment being investigated. Generally, *EHH* is calculated at varying distances from the core SNP position, so that the decay of *EHH* as a

function of physical distance can be assessed to determine the extension of the haplotype homozygosity. From this seminal concept, a family of statistical methods was developed in order to scan entire genomes in the search for evidence of selection.

Voight et al. (2006) proposed to measure how rapidly EHH decays from a core SNP site by calculating the area under the EHH curve,

$$iHH = \int_a^b EHH(x)dx$$

where iHH represents the definite integral of EHH evaluated over the domain of the chromosome segment delimited by upstream position a and downstream position b where EHH decays to some arbitrary small value (originally 0.05). As the area under the curve is not tractable analytically, a trapezoid quadrature with non-uniform grid can be adopted as a deterministic approximation:

$$iHH \sim \sum_{k=1}^K (x_{k+1} - x_k) \frac{(EHH_{k+1} + EHH_k)}{2}$$

A within population score, namely Integrated Haplotype Score (iHS), for a given site i, was introduced by Voight et al. (2006) as the log-ratio between the integrated EHH for the haplotypes containing the ancestral allele (iHH_A) and the derived allele (iHH_D):

$$iHS_i = \ln\left(\frac{iHH_{A,i}}{iHH_{D,i}}\right)$$

These scores are then standardized to have mean zero and variance one.

Extremely negative standardized iHS values have been of particular interest in human genetics, as they represent a recently acquired mutation that increased very rapidly in frequency (i.e., there is a partial sweep due to ongoing selection) or a haplotype that hitchhiked to fixation and then became enriched for derived alleles (Voight et al., 2006). However, a sweep can also produce large positive iHS values at nearby SNPs if ancestral alleles hitchhike with the selected site, so the chromosome region surrounding the selected variant typically exhibits a cluster of extreme positive and negative iHS values. Furthermore, in the context of cattle data, artificial selection and domestication probably favored "beneficial" alleles in the sense of human interest, regardless if it is ancestral or derived. Therefore, both positive and negative values should be investigated in cattle data. This implicates that the absolute value of standardized iHS scores should be preferred over the signed values, or, equivalently, that a two tailed hypothesis test should be assumed. As only partial ancestral allele information is available for cattle SNP assays, and the search for footprints of selection by iHS in cattle should disregard the direction of the sweep, a more appropriate generalized version of iHS can be postulated as the log-ratio between the integrated EHH for an arbitrary reference allele (iHH_REF) and for the alternative allele (iHH_ALT).

One of the limitations of this method is that if a given marker presents a nearly or completely fixed allele in the population being analyzed, this allele will have no integral to be calculated or an integral close to zero, so the log-ratio will result in a positive or negative infinite value. In this scenario, the calculation of iHS must be conditioned by $iHH_{REF} > 0$ and $iHH_{ALT} > 0$, which indirectly leads to a minor allele frequency (MAF) constraint. This limitation renders iHS underpowered to detect very recent nearly fixed selective sweeps, which are of primary interest in the cattle community. However, as discussed earlier, a crucial point to be considered is that contiguous chromosomal segments containing SNPs with MAF = 0 can also result from SNP chip ascertainment bias, which may produce false positive signals.

Tang et al. (2007) and Sabeti et al. (2007) have independently developed equivalent methods, Rsb and XPEHH, respectively, which attempt to compare long-range haplotypes between populations in order to increase the power of selective sweep detection. The most crucial improvement is that, for each population being analyzed iHH is calculated for the entire sample, instead of being partitioned between derived and ancestral alleles. This eliminates the MAF constraint and recovers the power to detect sweeps reaching fixation. The comparison with a population where the selective sweep may not have occurred adds extra power to the method. Calculations are performed as follows:

$$XPEHH_i = Rsb_i = \ln\left(\frac{iHH_{pop1,i}}{iHH_{pop2,i}}\right)$$

Where, relative to SNP i, $iHH_{pop1,i}$ is the integrated EHH in the first population and $iHH_{pop2,i}$ is the integrated EHH in the second population. Scores are also standardized to produce a distribution of standard deviates. Positive values indicate selective sweeps in the population used in the numerator, while negative values indicate selection in the population used in the denominator. Here, it is easy to keep track of the signals by using one-tailed hypothesis tests.

Studies applying EHH-based methods to cattle data are numerous (for instance, Hayes et al., 2008; Gautier and Naves, 2011; Qanbari et al., 2011, 2014; Flori et al., 2012; Utsunomiya et al., 2013; Huson et al., 2014). The reported loci are deemed to be genome responses to a variety of different selective pressures, such as milk and meat production, coat color, heat stress, and reproductive performance. Among these, one particularly interesting selective sweep, most likely related to adaptation to heat stress, has been reported in Creole cattle, including Senepol, Carora, Romosinuano, and cross-bred lineages (Flori et al., 2012; Huson et al., 2014). These cattle breeds present the slick hair coat phenotype, a dominant trait associated to heat tolerance in tropically adapted cattle that descend from Spanish cattle introduced to the New World. The chromosome segment containing the selective sweep ranges from 37.5 to 39.6 Mb on chromosome 20, with a variable peaking position (39.5 or 37.7 Mb) depending on the SNP panel (BovineSNP50 or BovineHD) and dataset analyzed (Flori et al., 2012; Huson et al., 2014). The disputed positional candidate genes are the retinoic acid induced 14 (RAI14 or NORPEG), prolactin receptor (PRLR), and S-phase kinase-associated protein 2 (SKP2). A strong candidate mutation has been recently proposed for PRLR, a single base deletion in exon 10 (ss1067289408) predicted to cause a frameshift that introduces

a premature stop codon (p.Leu462*) and consequent loss of 120 C-terminal amino acids from the long isoform of the prolactin receptor (Littlejohn et al., 2014).

POPULATION DIFFERENTIATION

Following the same principle as in *Rsb/XPEHH*, *XPCLR* and F_{RL}, although positive selection may act across populations sharing geographical proximity, environmental factors or a common phenotype, outgroup populations may not share the same selective pressure. Therefore, changes in allele frequency promoted by selection in one group will not be detectable in the other, and large differences in allele frequencies between populations will be observed.

The fixation index F_{ST} (Wright, 1950; Weir and Cockerham, 1984) and its abundant estimators is the gold standard method for detecting highly differentiated loci between populations. In essence, given the average allele frequency $\bar{p}$ across subpopulations, F_{ST} is simply the ratio between the variance in the allele frequency in different subpopulations $\sigma_S^2 = \sum_{j=1}^{k} (p_j - \bar{p})^2$ and the variance in the total population $\sigma_T^2 = \bar{p}(1 - \bar{p})$. In pair-wise comparisons, calculations simplify to:

$$F_{ST} = \frac{(p_1 - p_2)^2}{(p_1 + p_2)(2 - p_1 - p_2)}$$

Scores can be averaged across SNP windows or smoothed against genomic positions using a local variable bandwidth kernel estimator (Porto-Neto et al., 2013).

Flori et al. (2009) applied F_{ST} to three French dairy cattle breeds (Holstein, Normande, and Montbéliarde), and found that some of the putative loci under selection in that study overlapped genes that strongly affect milk production traits, such as the growth hormone receptor (*GHR*), and coat color, for instance *MC1R*. Porto-Neto et al. (2013) generated a comprehensive map of divergent loci between taurine and zebu cattle using over 777,000 SNPs and 13 cattle breeds, and reported that the highest scoring locus in the F_{ST} analysis maps to chromosome 7:47.2–53.7 Mb, which shelters a cluster of immune-related genes and *SPOCK1*, a gene previously implicated in puberty (Fortes et al., 2010).

Extensions of F_{ST} led to the development of the *FLK* test (Lewontin and Krakauer, 1973; Bonhomme et al., 2010; Fariello et al., 2013), which not only account for effective population size and hierarchical structure among populations, but also have known distributional properties under neutrality. Briefly, let q_i be the vector of reference allele frequencies at marker i for the populations being compared, and q_0 the ancestral allele frequency for the same allele. The *FLK* method relies on the linear model:

$$q_i = 1_n q_0 + e$$

where the residual term e is assumed $\mathcal{N}(0, V_i)$, and V_i is the expected variance-covariance matrix for vector q_i. This matrix is modeled as:

$$V_i = \mathcal{F} q_0 (1 - q_0)$$

where $\mathcal{F}$ is a kinship matrix. The diagonal elements of $\mathcal{F}$ represent the expected inbreeding coefficients in each respective population, and off-diagonal elements represent the amount of drift accumulated on the different branches of the population tree. A weighted least squares estimate of q_0 is then obtained as:

$$\hat{q}_0 = (1_n' \mathcal{F}^{-1} 1_n)^{-1} 1_n' \mathcal{F}^{-1} q_i$$

The *FLK* score is a measure of goodness-of-fit of this model, and is calculated as the deviance (i.e., residual sum of squares):

$$FLK_i = (q_i - \hat{q}_0 1_n)' V_i^{-1} (q_i - \hat{q}_0 1_n)$$

Under the assumption of a star-like tree-pure drift model (i.e., the populations being compared evolved in parallel from a single ancestral population, with no mutations or admixture), matrix $\mathcal{F}$ can be simplified to $\mathcal{F} = I_n \overline{F_{ST}}$, where I_n is an identity matrix and $\overline{F_{ST}}$ is the average F_{ST} over all SNP loci. In this case, the average allele frequency across populations is an unbiased estimator of q_0, and the deviance is simplified to $F_{ST}(n-1)/\overline{F_{ST}}$, which gives a test statistic that is linearly correlated with F_{ST}. As discussed later, while F_{ST} has no known theoretical distribution under the neutral model, *FLK* can be modeled as a chi-squared distribution. Moreover, *FLK* outperformed F_{ST} in simulations, especially when scores were computed based on haplotypes (*hapFLK*) instead of single markers (Fariello et al., 2013). This method is a suitable alternative to F_{ST} for cattle data.

DISTRIBUTIONS UNDER THE NULL HYPOSTHESIS AND *p*-VALUES

Although different in formulation and assumptions, all methods presented so far attempt to address the same null and alternative hypothesis: H_0 = locus is neutral; H_1 = locus is not neutral. Under the hypothesis of neutrality, *EHH*-based methods are standard normal deviates (Voight et al., 2006; Sabeti et al., 2007; Tang et al., 2007). Hence, the probability that SNP i with *Rsb* or *XPEHH* score x_i is neutral in the population used as numerator can be approximated by an upper tail *p*-value derived from the normal cumulative density function (CDF) Φ:

$$Pr(Rsb, XPEHH > x_i \mid neutral) = 1 - \Phi(x_i)$$

Likewise, the probability that SNP i with *iHS* score x_i is neutral, given that both reference and alternative alleles are of interest, can be approximated by a two-tailed *p*-value:

$$Pr(|iHS| > |x_i| \mid neutral) = 1 - |\Phi(x_i) - \Phi(-x_i)|$$
$$= 1 - 2|\Phi(x_i) - 0.5|$$

Recently, concerns have emerged on the interpretation of *p*-values in signatures of selection analyses (Simianer et al., 2014). It is argued that, at least for the cases of *EHH*-based methods, scores standardized using genome-wide data are not test statistics in the classical sense but only deviates from an average (Voight et al., 2006), so *p*-values would only represent quantiles from an empirical distribution, rather than formal significance values. This implies that the probability of obtaining an arbitrary score given the locus is neutral cannot be exactly computed once the

underlying true probability distribution may vary according to different demographic scenarios. The major caveat here is, for any given test discussed, the proposed asymptotic distribution under the hypothesis of neutrality is still largely based on coalescent simulations with demographic models calibrated for human data. However, if the majority of the genome is under neutrality and the null distribution is robust to a wide range of demographic models, one may expect that genome-wide distribution of scores in cattle data should mimic the simulations performed for the human model, serving as a control (Gianola et al., 2010). This has been observed in SS studies in cattle (Gautier and Naves, 2011; Flori et al., 2012; Utsunomiya et al., 2013), and therefore these theoretical distributions under neutrality are suitable for cattle data.

In the case of F_{ST}, although approximate distributions are available (e.g., exponential or beta), such approximations are sub-optimal. One advantage of FLK over F_{ST} is that scores are asymptotically distributed as χ^2_ν under the null hypothesis, where degrees of freedom ν can be equivalently calculated as *number of subpopulations* $- 1$ or as the average FLK across all loci. Upper tail p-values can be computed as:

$$Pr\left(FLK > x_i \mid neutral\right) = 1 - F_\nu(x_i)$$

where F_ν is the χ^2 CDF with ν degrees of freedom.

Although F_{ST} and CLL scores have unknown distributions under neutrality, both rely on population comparisons. The use of datasets consisting of multiple populations allows for permutation tests via random sampling of individuals or random sorting of population labels in order to compute an empirical null distribution. Nevertheless, permutation tests are computationally intensive and may be impracticable in re-sequencing data.

The ZHp method suffers from the same problem, with the additional limitation of not allowing data permutations when a single population is surveyed. Although a truncated normal distribution could be suggested to approximate its null distribution, this has not been properly assessed in practice. Another challenge is deriving significance values when scores are averaged across SNP windows. Following the assumption that the majority of the loci (single markers or SNP windows for that matter) are under neutrality, an empirical CDF derived from genome-wide scores should converge to the underlying true CDF, so probability values could be empirically obtained from a step function:

$$Pr\left(Test > x_i \mid neutral\right) = 1 - ECDF(x_i)$$
$$= 1 - \frac{(number\ of\ scores\ \leq\ x_i)}{number\ of\ observed\ scores}$$

The probability that SNP i with score x_i in a given test is selected is not as trivial to approximate. While the distributions under the hypothesis of neutrality seems to be robust across a wide range of demographic models, the distribution under the hypothesis of selection may vary widely depending upon the demographic history that shaped the allele frequency spectrum. Therefore, there is no unique theoretical distribution to represent selected variants, and coalescent simulations using well calibrated demographic models are required in order to generate empirical distributions.

This has been successfully done for human data (Grossman et al., 2010), but is highly challenging in cattle as most of the demographic history of the bovine species is yet to be uncovered. Nevertheless, promising methods to infer demographic scenarios from the data, including estimates of population sizes and population separations over time, are now emerging (Schiffels and Durbin, 2014), which could be useful to elucidate cattle history and customize coalescent simulations based on empirical data.

AVAILABLE SOFTWARE AND ANALYSIS BEST PRACTICES
Table 1 summarizes all the software mentioned below. The first step to be considered before starting a SS analysis is filtering out poor quality data and markers and samples that are not informative or that may confound the analysis. Discussing the particularities of measures of quality control is beyond the scope of the present article, but some of the best practices in SS studies can emerge from common sense. First, metrics such as Hardy-Weinberg Equilibrium (HWE) and MAF, highly used in GWA studies, should be applied with caution. Elimination of markers with extreme deviations from HWE expectations may counteract the whole SS enterprise, as in this type of study outlier loci are being sought. Likewise, MAF controls may cause signals that are reaching fixation to be completely lost. A situation where HWE and MAF thresholds can be benign is when only markers with extreme excess of heterozygotes are eliminated, and MAF constraints are applied to the pooled allele frequencies across all populations in the study. We suggest *PLINK* v1.07 (Purcell et al., 2007) or *PLINK* v1.90 for this first data screening. Other usual filters such as GenCall and GenTrain scores and call rates should follow the same guidelines as in GWA studies.

Another important issue is cryptic relatedness. Eliminating samples of closely related animals is of paramount importance to reduce false positive signals. We have previously reported an algorithm to find the maximum independent set based on identity-by-descent, i.e., maximize the number of samples while eliminating first and second degree relationships using SNP data (Utsunomiya et al., 2013). Other software such as *PLINK* v1.90 and *GCTA* v1.24.2 (Yang et al., 2011) also provide means to find the optimal set of independent individuals.

The next key step is producing high quality phased data. There are several methods and implementations for phasing, such as *fastPHASE* v1.2 (Scheet and Stephens, 2006), *Beagle* v3.3 or later (Browning and Browning, 2008), and *SHAPEIT2* (O'Connell et al., 2014). Although *fastPHASE* exhibits high phasing accuracy, it is orders of magnitude more computationally intensive than *Beagle* or *SHAPEIT2*. It is important to notice that phasing is not a straightforward procedure, and is highly prone to errors. Consequently, results from haplotype-based methods should be assessed with caution. The effect of haplotype errors on SS results remains underexplored and at some extent neglected.

Regarding the SS analysis *per se*, *EHH*-based methods can be computed using *Sweep*, *selscan* (Szpiech and Hernandez, 2014) or the R package *rehh* (Gautier and Vitalis, 2012). F_{RL} has a dedicated software, as well as *CLR* (available at: and *XPCLR*. For ROH-based tests, runs can be computed using either *PLINK* or *SNP & Variation Suite v7.6.8* or later, and F_{RL} can be easily computed with home grown scripts. In the cases of F_{ST}, CLL, ZH_p and

Table 1 | Available software for signatures of selection analysis.

Software	Analysis	Availability	References
PLINK v1.07 and v1.90beta	Quality control, data management, cryptic relatedness check, and computation of runs of homozygosity and FST	http://pngu.mgh.harvard.edu/~purcell/plink/ https://www.cog-genomics.org/plink2	Purcell et al., 2007
SNP & Variation Suite v7.6.8	Quality control, data management, cryptic relatedness check, and computation of runs of homozygosity and FST	http://www.goldenhelix.com	Golden Helix, Bozeman, MT, USA
GCTA v1.24.2	Quality control, data management, and cryptic relatedness check	http://www.complextraitgenomics.com/software/gcta/	Yang et al., 2011
fastPHASE v1.2	Phasing and imputation of missing genotypes	http://stephenslab.uchicago.edu/software.html#fastphase)	Scheet and Stephens, 2006
Beagle v3.3 and v4.0	Phasing and imputation of missing genotypes	http://faculty.washington.edu/browning/beagle/beagle.html	Browning and Browning, 2008
SHAPEIT2	Phasing and imputation of missing genotypes	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html	O'Connell et al., 2014
Sweep	EHH-based methods	http://www.broadinstitute.org/mpg/sweep/	Sabeti et al., 2002
selscan	EHH-based methods	https://github.com/szpiech/selscan	Szpiech and Hernandez, 2014
rehh	EHH-based methods	http://cran.r-project.org/web/packages/rehh/index.html	Gautier and Vitalis, 2012
FLK	Computation of the FLK statistics	https://qgsp.jouy.inra.fr/index.php?option=com_content&view=article&id=50&Itemid=55	Bonhomme et al., 2010; Fariello et al., 2013
SweepFinder	Computation of the CLR statistics	http://people.binf.ku.dk/rasmus/webpage/sf.html	Nielsen et al., 2005
XP-CLR v1.0	Computation of the XPCLR statistics	http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html	Chen et al., 2010
cosi and cosi2	Coalescent simulations	http://www.broadinstitute.org/~sfs/cosi/ http://www.broadinstitute.org/~ilya/cosi2/	Schaffner et al., 2005; Shlyakhter et al., 2014
MSMS	Coalescent simulations	http://www.mabs.at/ewing/msms/index.shtml	Ewing and Hermisson, 2010

the method reported by Ramey et al. (2013), allele frequencies and genotype counts can be obtained from either *PLINK* or *SNP & Variation Suite*, and calculation of scores can be easily implemented in *R* (available at: http://www.r-project.org/) or other languages. Simulations of population genetics datasets under specific demographic models, including neutral and selected loci, can be performed using coalescent simulators such *cosi* (Schaffner et al., 2005), *cosi2* (Shlyakhter et al., 2014), or *MSMS* (Ewing and Hermisson, 2010).

COMBINING SELECTION SIGNALS

The available methodologies to detect positive selection differ substantially from each other in terms of the pattern of genetic variation encrypting a "signal." However, all of them have a shared objective: to identify loci that have undergone positive selection. Indeed, at least for recent selective pressures (up to a few thousand generations back), a selected variant is expected to be in a chromosome segment where there has been loss of diversity, enrichment for derived or rare alleles, population differentiation, and highly frequent long-range haplotypes. Therefore, collecting evidence across different methodologies targeting distinct classes of signals may help in identifying loci under positive selection. This section explores the statistical properties and limitations of some of the available methods designed to combine different methods for signatures of selection.

COMPOSITE OF MULTIPLE SIGNALS (CMS)

In the original implementation of the method (Grossman et al., 2010), CMS is a local test designed to narrow down signals detected from s distinct genome-wide scans, and is defined as the approximate joint posterior probability that a given variant is selected:

$$CMS_{local} = \Pr\left(selected \mid x_{ij}\right)$$

$$= \prod_{j=1}^{s} \frac{\Pr\left(x_{ij}|selected\right)\Pr\left(selected\right)}{\Pr\left(x_{ij} \mid selected\right)\Pr\left(selected\right) + \Pr(x_{ij}|neutral)\Pr(neutral)}$$

The genome-wide extension (Grossman et al., 2013) focuses on the product of the Bayes Factors for each one of the tests to be combined. For each test, BF is computed as the ratio between the posterior and prior odds:

$$CMS_{GW} = \prod_{j=1}^{s} BF_{ij} = \prod_{j=1}^{s} \frac{\Pr(x_{ij}|selected)\Pr(selected)}{\Pr(x_{ij}|neutral)\Pr(neutral)}$$

In the absence of prior information on the number of loci under selection across the genome, CMS scores simplify to:

$$\mathrm{CMS_{GW}} = \prod_{j=1}^{s} \frac{\mathrm{Pr}(x_{ij}|selected)}{\mathrm{Pr}(x_{ij}|neutral)}$$

A challenging aspect of the implementation of these methods is computing $\mathrm{Pr}(x_{ij}|selected)$, which requires simulations under clear demographic assumptions. In contrast, as discussed earlier, many methods designed to detect markers under positive selection allow for approximating $\mathrm{Pr}(x_{ij}|neutral)$ from asymptotic theoretical or empirical distributions. Therefore, composite tests considering only the distributions under neutrality are more appropriate for cattle data.

The original CMS score can be modified to take advantage of the assumed distributions under neutrality in order to relax demographic assumptions and avoid expensive simulations. The most essential modification involves reformulating the problem of detecting markers departing from neutrality. Instead of considering the assessment of whether a marker has been selected or not, one can look for support from the data against the null model, i.e., that the marker does not fit well to the neutral model. First, let the new CMS score be the approximate joint posterior probability of a given variant not being neutral:

$$\mathrm{CMS_{new}} = \mathrm{Pr}\left(not\ neutral|x_{ij}\right)$$

$$= \prod_{j=1}^{s} \frac{\mathrm{Pr}(x_{ij}|not\ neutral)\mathrm{Pr}(not\ neutral)}{\mathrm{Pr}\left(x_{ij}|neutral\right)\mathrm{Pr}(neutral)} + \mathrm{Pr}\left(x_{ij}|not\ neutral\right)\mathrm{Pr}(not\ neutral)$$

Here, $\mathrm{Pr}(x_{ij}|neutral)$ is computed directly from its theoretical distribution, and $\mathrm{Pr}\left(x_{ij}\mid not\ neutral\right)$ is computed as $1 - \mathrm{Pr}(x_{ij}|neutral)$. Also, it can be assumed that the prior $\mathrm{Pr}(not\ neutral) = 1 - \mathrm{Pr}(neutral)$. Therefore, the new CMS score can be re-written as:

$$\mathrm{CMS_{new}} = \prod_{j=1}^{s} \frac{\left[1 - \mathrm{Pr}\left(x_{ij}|neutral\right)\right]\left[1 - \mathrm{Pr}(neutral)\right]}{\mathrm{Pr}\left(x_{ij}|neutral\right)\mathrm{Pr}(neutral)} + \left[1 - \mathrm{Pr}\left(x_{ij}|neutral\right)\right]\left[1 - \mathrm{Pr}\left(neutral\right)\right]$$

Likewise, the genome-wide modified CMS scores can be reformulated as:

$$\mathrm{CMS_{GW-new}} = \prod_{j=1}^{s} \frac{\mathrm{Pr}(x_{ij}|not\ neutral)}{\mathrm{Pr}(x_{ij}|neutral)} = \prod_{j=1}^{s} \frac{1 - \mathrm{Pr}(x_{ij}|neutral)}{\mathrm{Pr}(x_{ij}|neutral)}$$

It is important to note that this modification does not allow for the same interpretation as the original CMS method: the composite likelihood does not indicate selection, but rather, that a marker does not fit well the neutral model.

META ANALYSIS OF SELECTION SIGNALS (META-SS)
Following the ideas expanded from the landmark publication of Grossman et al. (2010), given the probabilities under the null hypothesis for each test, our interest is to identify loci presenting consistent rejection of the neutral model across the different tests. For any given statistic, p-values are uniformly distributed in the interval between 0 and 1 under the null hypothesis. This property makes possible to use an inverse CDF, such as the Gaussian density, to produce scores for each test derived from a single theoretical distribution. Therefore, for each SNP i and test j, a new score can be computed as $Z_{ij} = \Phi^{-1}(1 - P_{ij})$, where P_{ij} is the p-value. These Z-transformed p-values can be then averaged and standardized to produce a composite score.

We have previously described $meta-SS$ (Utsunomiya et al., 2013), an abstraction of the Stouffer Z-transformation for combining different selection signals using the aforementioned framework. As the Stouffer method assumes the tests are uncorrelated under the shared null hypothesis and the use of pair-wise comparisons produce correlated scores, a weighted average was originally proposed to penalize dependent tests:

$$meta-SS = \frac{\sum_{j=1}^{s} \omega_j Z_j}{\sqrt{\sum_{j=1}^{k} \omega_j^2}}$$

where ω_j is the weight for test j. In this setting, a uniform penalization can be applied to control for the inflation of correlated tests. As this penalization does not incorporate the strength of correlations among tests, the $meta-SS$ test can be modified to explicitly account for the magnitude of correlations between scores. Considering all scores are equally weighted, the corrected composite score can be computed as:

$$meta-SS = \frac{\sum_{j=1}^{s} Z_j}{\sqrt{k + 2R}}$$

where R is the sum of all pairwise Pearson's product-moment correlations. Under the hypothesis of neutrality, these composite scores are distributed as $N(0, 1)$, so the higher is the Z-transformed value, the worse the marker fits the neutral model. Upper tail p-values can then be obtained from the standard normal CDF.

Obvious limitations from $meta-SS$ and CMS is the inability to incorporate statistics for which p-values cannot be derived. Randhawa et al. (2014) proposed Composite Selection Signals (CSS), a nonparametric interpretation of $meta-SS$, where fractional ranking is used instead of p-values to combine different tests. Briefly, the vector of test statistics for method j is first sorted and then ranked, taking values 1, ..., k. Next, the vector of ranks is re-scaled by dividing all elements by $k + 1$, thus producing a variable ranging from 0 to 1. These re-scaled ranks are treated as they were p-values for the test statistics, and then combined as in the $meta-SS$ approach. This strategy is equivalent to computing probabilities from an empirical CDF using a step function, as discussed earlier, which has an appealing feature: as fractional ranking can be generated for any particular test, signature blending is made feasible even if the theoretical distributions are unknown or if scores have been averaged in chromosome windows. However, a caveat is that the magnitude of the actual test statistics may be lost, so one may expect loss of power compared to the use of theoretical or simulated distributions.

PRINCIPAL COMPONENTS ANALYSIS

Simianer et al. (2014) proposed combining different tests by applying an eigendecomposition of the correlation matrix of the scores. The attractive feature of this method is the possibility of using standardized scores instead of approximate probabilities. However, as each principal component has heterogeneous loadings from each test, deriving a single synthetic score that summarizes all different tests remains a challenge in this framework.

CHALLENGES AND FUTURE DIRECTIONS

In theory, genome-wide genotypes are a vast source of information that can be explored in the search for large effect mutations that underwent selection. However, the existing data and methods still suffer from power issues and confounding effects that can give rise to false positive and false negative signals.

Although simulations suggest that only marginal gains in power are obtained when the sample size is increased from tens to hundreds of unrelated samples, marker density and allele frequency spectrum seems to impact power dramatically (Lappalainen et al., 2010; Simianer et al., 2014). Genotypes derived from commercial SNP arrays have two important limitations in this context: (1) incomplete genome coverage by markers; and (2) ascertainment bias. The search for SS must be preferentially performed using high density SNP panels, although optimal average intermarker distances to detect a sweep may vary depending on the effective population size, extent of linkage disequilibrium and the nature of the signal. Regarding ascertainment bias, commercial SNP arrays are suitable for cattle populations that are closely related to the breeds used in the SNP discovery process, but there is no guarantee they will be informative in genetically distant populations. Indeed, with a few exceptions, little congruence has been reported between candidate selected loci identified using whole genome sequence and different commercial genotyping platforms in African humans not included in the HapMap data (Lachance and Tishkoff, 2013). Altogether, these arguments suggest that re-sequencing data is the optimal choice in SS studies in cattle. At some extent, the HD assay is appropriate, as it has a high-density coverage of the genome with SNPs that are less biased than competing panels.

Another important source of confounding comes from the methods available to detect SS. First, all methods assume that individuals have no recent relationships in their pedigrees, a condition that is hardly true and generally ignored. It is essential to filter the data for cryptic relationships and assure to include only samples that are unrelated for at least two generations. Second, most of the methods rely on haplotypes and SNP coordinates, so further improvement of phasing strategies and of the bovine reference genome assembly is crucial to assure high quality results. Third, variants can depart from neutrality not only due to positive selection, but also as consequence of demographic events such as bottlenecks, genetic drift and admixture. Distinguishing loci under selection from neutrally evolving loci remains a major challenge in the field, and will require refinement of existing methods and development of new tests. Nevertheless, combining signals across different methods seems to be a promising approach to mitigate the individual methodological limitations.

Also, when available, the concomitant analysis of environmental data (e.g., temperature, humidity, precipitation, disease prevalence, etc.) may be of great help in distinguishing true positives and accelerating the link between signal and phenotype (Lv et al., 2014).

Well-planned study designs will be crucial to exploit the full potential of SS in the detection of large effect mutations favored by selection. The identification of common adaptive phenotypes, together with geographical information data, should be an important player in sampling and decisions of population comparisons. Cattle breeds that are not highly productive but that exhibit genetic local adaptation should be considered as priority targets, as their environmental fitness was probably forged by hundreds of years of natural and artificial selection. In the context of artificial selection for complex traits, as large cattle pedigree cohorts for genomic selection become available, it will be soon possible to actually assess rapid changes in allele frequency using historical data, rather than present date data only. First demonstrations of such ideas were presented by Decker et al. (2012), and are likely to be incorporated as routine monitoring tools of genomic resources in breeding programs in the future.

Recently, comparing candidate loci across GWA studies has been facilitated in cattle with the advent of the Animal QTLdb (Hu and Reecy, 2007). Similarly, results from a SS scan on the human 1000 Genomes data reported by Grossman et al. (2013) were made available through the CMS viewer tool (available at: http://www.broadinstitute.org/mpg/cmsviewer/). Pybus et al. (2014) have also launched a comprehensive database of SS in the 1000 Genomes data (available at: http://hsb.upf.edu/). The research community would highly benefit from the development of a SS database for livestock species, which would not only facilitate cross-referencing, but would also help researchers willing to dig deep into the functional meaning of the signals to select promising candidates emerging from multiple preexisting studies.

Finally, similarly to the recent developments in human SS (Kamberov et al., 2013), unraveling the functional relevance of the putative selected variants will demand interdisciplinary reasoning, compilation of a wide range of data types (e.g., transcriptomic, proteomic), and assemblage of an arsenal of *post-hoc* assays, such as genomic editing, culture, phenotyping and challenge of specific cell lines, production of knock-out models, and generation of cross-bred lines for confirmatory segregation analyses.

REFERENCES

Ajmone-Marsan, P. (2010). On the origin of cattle: how aurochs became cattle and colonized the world. *Evol. Anthropol. Issues News Rev.* 19, 148–157. doi: 10.1002/evan.20267

Axelsson, E., Ratnakumar, A., Arendt, M.-L., Maqbool, K., Webster, M. T., Perloski, M., et al. (2013). The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495, 360–364. doi: 10.1038/nature 11837

Bonfiglio, S., Achilli, A., Olivieri, A., Negrini, R., Colli, L., Liotta, L., et al. (2010). The enigmatic origin of bovine mtDNA haplogroup R: Sporadic interbreeding or an independent event of Bos primigenius domestication in Italy? *PLoS ONE* 5:e15760. doi: 10.1371/journal.pone.0015760

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et al. (2010). Detecting selection in population trees: the lewontin and krakauer test extended. *Genetics* 186, 241–262. doi: 10.1534/genetics.110.117275

Bradley, D. G., MacHugh, D. E., Cunningham, P., and Loftus, R. T. (1996). Mitochondrial diversity and the origins of African and European cattle. *Proc. Natl. Acad. Sci. U.S.A.* 93, 5131–5135. doi: 10.1073/pnas.93.10.5131

Browning, B. L., and Browning, S. R. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi: 10.1016/j.ajhg.2009. 01.005

Bruford, M. W., Bradley, D. G., and Luikart, G. (2003). DNA markers reveal the complexity of livestock domestication. *Nat. Rev. Genet.* 4, 900–910. doi: 10.1038/nrg1203

Bush, W. S., and Moore, J. H. (2012). Chapter 11: genome-wide association studies. *PLoS Comput. Biol.* 8:e1002822. doi: 10.1371/journal.pcbi.1002822

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res* 20, 393–402. doi: 10.1101/gr. 100545.109

Curik, I., Ferenčaković, M., and Sölkner, J. (2014). Inbreeding and runs of homozygosity: a possible solution to an old problem. *Livest. Sci.* 166, 26–34. doi: 10.1101/gr.100545.109

Darwin, C., and Wallace, A. (1858). On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *J. Proc. Linn. Soc. Lond. Zool.* 3, 45–62. doi: 10.1111/j.1096-3642.1858. tb02500.x

Decker, J. E., McKay, S. D., Rolf, M. M., Kim, J. W., Molina Alcalá, A., Sonstegard, T. S., et al. (2014). Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. *PLoS Genet.* 10:e1004254. doi: 10.1371/journal.pgen.1004254

Decker, J. E., Pires, J. C., Conant, G. C., McKay, S. D., Heaton, M. P., Chen, K., et al. (2009). Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proc. Natl. Acad. Sci. U.S.A.* 106, 18644–18649. doi: 10.1073/pnas.0904691106

Decker, J. E., Vasco, D. A., McKay, S. D., McClure, M. C., Rolf, M. M., Kim, J., et al. (2012). A novel analytical method, Birth Date Selection Mapping, detects response of the Angus (Bos taurus) genome to selection on complex traits. *BMC Genomics* 13:606. doi: 10.1186/1471-2164-13-606

Durkin, K., Coppieters, W., Drögemüller, C., Ahariz, N., Cambisano, N., Druet, T., et al. (2012). Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* 482, 81–84. doi: 10.1038/nature 10757

Ewing, G., and Hermisson, J. (2010). MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* 26, 2064–2065. doi: 10.1093/bioinformatics/btq322

FAO. (2007). *The State of the World's Animal Genetic Resources for Food and Agriculture.* Available online at: http://www.fao.org/docrep/010/a1250e/ a1250e00.htm.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929–941. doi: 10.1534/genetics.112.147231

Fay, J. C., and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413.

Ferenčaković, M., Sölkner, J., and Curik, I. (2013). Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genet. Sel. Evol.* 45:42. doi: 10.1186/1297-9686-45-42

Flori, L., Fritz, S., Jaffrézic, F., Boussaha, M., Gut, I., Heath, S., et al. (2009). The genome response to artificial selection: a case study in dairy cattle. *PLoS ONE* 4:e6595. doi: 10.1371/journal.pone.0006595

Flori, L., Gonzatti, M. I., Thevenon, S., Chantal, I., Pinto, J., Berthier, D., et al. (2012). A quasi-exclusive European ancestry in the Senepol tropical cattle breed

highlights the importance of the slick locus in tropical adaptation. *PLoS ONE* 7:e36133. doi: 10.1371/journal.pone.0036133

Fortes, M. R. S., Reverter, A., Zhang, Y., Collis, E., Nagaraj, S. H., Jonsson, N. N., et al. (2010). Association weight matrix for the genetic dissection of puberty in beef cattle. *Proc. Natl. Acad. Sci. U.S.A.* 107, 13642–13647. doi: 10.1073/pnas.1002044107

Gautier, M., and Naves, M. (2011). Footprints of selection in the ancestral admixture of a New World Creole cattle breed. *Mol. Ecol.* 20, 3128–3143. doi: 10.1111/j.1365-294X.2011.05163.x

Gautier, M., and Vitalis, R. (2012). Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. *Bioinformatics* 28, 1176–1177. doi: 10.1093/bioinformatics/bts115

Georges, M., Drinkwater, R., King, T., Mishra, A., Moore, S. S., Nielsen, D., et al. (1993). Microsatellite mapping of a gene affecting horn development in *Bos taurus. Nat. Genet.* 4, 206–210. doi: 10.1038/ng0693-206

Gianola, D., Simianer, H., and Qanbari, S. (2010). A two-step method for detecting selection signatures using genetic markers. *Genet. Res. (Camb.)* 92, 141–155. doi: 10.1017/S0016672310000121

Gibson, J., Morton, N. E., and Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* 15, 789–795. doi: 10.1093/hmg/ddi493

Groeneveld, L. F., Lenstra, J. A., Eding, H., Toro, M. A., Scherf, B., Pilling, D., et al. (2010). Genetic diversity in farm animals–a review. *Anim. Genet.* 41(Suppl. 1), 6–31. doi: 10.1111/j.1365-2052.2010.02038.x

Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., et al. (2013). Identifying recent adaptations in large-scale genomic data. *Cell* 152, 703–713. doi: 10.1016/j.cell.2013.01.035

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 327, 883–886. doi: 10.1126/science.1183863

Hanotte, O., Bradley, D. G., Ochieng, J. W., Verjee, Y., Hill, E. W., and Rege, J. E. O. (2002). African pastoralism: genetic imprints of origins and migrations. *Science* 296, 336–339. doi: 10.1126/science.1069878

Hayes, B. J., Lien, S., Nilsen, H., Olsen, H. G., Berg, P., Maceachern, S., et al. (2008). The origin of selection signatures on bovine chromosome 6. *Anim. Genet.* 39, 105–111. doi: 10.1111/j.1365-2052.2007.01683.x

Hellmann, I., Ebersberger, I., Ptak, S. E., Pääbo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.* 72, 1527–1535. doi: 10.1086/375657

Howrigan, D. P., Simonson, M. A., and Keller, M. C. (2011). Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics* 12:460. doi: 10.1186/1471-2164-12-460

Hu, Z. L., and Reecy, J. M. (2007). Animal QTLdb: beyond a repository - A public platform for QTL comparisons and integration with diverse types of structural genomic information. *Mamm. Genome* 18, 1–4. doi: 10.1007/s00335-006-0105-8

Huson, H. J., Kim, E. S., Godfrey, R. W., Olson, T. A., McClure, M. C., Chase, C. C., et al. (2014). Genome-wide association study and ancestral origins of the slick-hair coat in tropically adapted cattle. *Front. Genet.* 5:101. doi: 10.3389/fgene.2014.00101

Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., et al. (2013). Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152, 691–702. doi: 10.1016/j.cell.2013.01.016

Kim, E. S., Cole, J. B., Huson, H., Wiggans, G. R., Van Tassel, C. P., Crooker, B. A., et al. (2013). Effect of artificial selection on runs of homozygosity in U.S. *Holstein cattle. PLoS ONE* 8:e80813. doi: 10.1371/journal.pone.0080813

Kim, Y., and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765–777.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–626. doi: 10.1038/217624a0

Klungland, H., Vage, D. I., Gomez-Raya, L., Adalsteinsson, S., and Lien, S. (1995). The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. *Mamm. Genome* 6, 636–639. doi: 10.1007/BF003 52371

Lachance, J., and Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: Why it is important, and how to correct it. *Bioessays* 35, 780–786. doi: 10.1002/bies.201300014

Lappalainen, T., Salmela, E., Andersen, P. M., Dahlman-Wright, K., Sistonen, P., Savontaus, M., et al. (2010). Genomic landscape of positive natural selection

in Northern European populations.*Eur. J. Hum. Genet.* 18, 471–478. doi: 10.1038/ejhg.2009.184

Lee, H. J., Kim, J., Lee, T., Son, J. K., Yoon, H. B., Baek, K. S., et al. (2014). Deciphering the genetic blueprint behind Holstein milk proteins and production.*Genome Biol. Evol.* 6, 1366–1374. doi: 10.1093/gbe/evu102

Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M., et al. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci. U.S.A.* 104, 19942–19947. doi: 10.1073/pnas.0710021104

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.

Littlejohn, M. D., Henty, K. M., Tiplady, K., Johnson, T., Harland, C., Lopdell, T., et al. (2014). Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nat. Commun.* 5, 5861. doi: 10.1038/ncomms6861

Loftus, R. T., MacHugh, D. E., Bradley, D. G., Sharp, P. M., and Cunningham, P. (1994). Evidence for two independent domestications of cattle. *Proc. Natl. Acad. Sci. U.S.A.* 91, 2757–2761. doi: 10.1073/pnas.91.7.2757

Lv, F., Agha, S., Kantanen, J., Colli, L., Stucki, S., and Kijas, J. W. (2014). Adaptations to climate-mediated selective pressures in sheep. *Mol. Biol. Evol.* 31, 3324–3343. doi: 10.1093/molbev/msu264

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., et al. (2009). Development and characterization of a high density snp genotyping assay for cattle. *PLoS ONE* 4:e5350. doi: 10.1371/journal.pone.0005350

Mona, S., Catalano, G., Lari, M., Larson, G., Boscato, P., Casoli, A., et al. (2010). Population dynamic of the extinct European aurochs: genetic evidence of a north-south differentiation pattern and no evidence of post-glacial expansion. *BMC Evol. Biol.* 10:83. doi: 10.1186/1471-2148-10-83

Murray, C., Huerta-Sanchez, E., Casey, F., and Bradley, D. G. (2010). Cattle demographic history modelled from autosomal sequence variation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 365, 2531–2539. doi: 10.1098/rstb.2010.0103

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575. doi: 10.1101/gr.4252305

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* 10:e1004234. doi: 10.1371/journal.pgen.1004234

Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). Genome-wide scans for footprints of natural selection. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 365, 185–205. doi: 10.1098/rstb.2009.0219

Oleksyk, T. K., Zhao, K., De La Vega, F. M., Gilbert, D. A., O'Brien, S. J., and Smith, M. W. (2008). Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS One* 3:e1712. doi: 10.1371/journal.pone.0001712

Porto-Neto, L. R., Sonstegard, T. S., Liu, G. E., Bickhart, D. M., Da Silva, M. V. B., Machado, M., et al. (2013). Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. *BMC Genomics* 14:876. doi: 10.1186/1471-2164-14-876

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Purfield, D. C., Berry, D. P., McParland, S., and Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. *BMC Genet.* 13:70. doi: 10.1186/1471-2156-13-70

Pybus, M., Dall'Olio, G. M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., et al. (2014). 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res.* 42, D903–D909. doi: 10.1093/nar/gkt1188

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., et al. (2011). Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics* 12:318. doi: 10.1186/1471-2164-12-318

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T. M., Fries, R., et al. (2014). Classic selective sweeps revealed by massive sequencing in cattle. *PLoS Genet.* 10:e1004148. doi: 10.1371/journal.pgen.1004148

Qanbari, S., and Simianer, H. (2014). Mapping signatures of positive selection in the genome of livestock. *Livest. Sci.* 166, 133–143. doi: 10.1016/j.livsci.2014.05.003

Ramey, H. R., Decker, J. E., McKay, S. D., Rolf, M. M., Schnabel, R. D., and Taylor, J. F. (2013). Detection of selective sweeps in cattle using genome-wide SNP data. *BMC Genomics* 14:382. doi: 10.1186/1471-2164-14-382

Randhawa, I. A. S., Khatkar, M. S., Thomson, P. C., and Raadsma, H. W. (2014). Composite selection signals can localize the trait specific genomic regions in multi-breed populations of cattle and sheep. *BMC Genet.* 15:34. doi: 10.1186/1471-2156-15-34

Rubin, C. J., Megens, H. J., Martinez Barrio, A., Maqbool, K., Sayyab, S., Schwochow, D., et al. (2012). Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19529–19536. doi: 10.1073/pnas.1217149109

Rubin, C.-J., Zody, M. C., Eriksson, J., Meadows, J. R. S., Sherwood, E., Webster, M. T., et al. (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* 464, 587–591. doi: 10.1038/nature08832

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419, 832–837. doi: 10.1038/nature01140

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., et al. (2006). Positive natural selection in the human lineage. *Science* 312, 1614–1620. doi: 10.1126/science.1124309

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918. doi: 10.1038/nature06250

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15, 1576–1583. doi: 10.1101/gr.3709305

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. *Am. J. Hum. Genet.* 78, 629–644. doi: 10.1086/502802

Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* 46, 919–925 doi: 10.1038/ng.3015

Shlyakhter, I., Sabeti, P. C., and Schaffner, S. F. (2014). Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* 30, 3427–3429. doi: 10.1093/bioinformatics/btu562

Simianer, H., Ma, Y., and Qanbari, S. (2014). "Statistical problems in livestock population genomics," in *Proceedings, 10th World Congress of Genetics Applied to Livestock Production* (Vancouver, BC).

Stella, A., Ajmone-Marsan, P., Lazzari, B., and Boettcher, P. (2010). Identification of selection signatures in cattle breeds selected for dairy production.*Genetics* 185, 1451–1461. doi: 10.1534/genetics.110.116111

Szpiech, Z. A., and Hernandez, R. D. (2014). Selscan: an efficient multithreaded program to perform ehh-based scans for positive selection. *Mol. Biol. Evol.* 31, 2824–2827. doi: 10.1093/molbev/msu211

Taberlet, P., Valentini, A., Rezaei, H. R., Naderi, S., Pompanon, F., Negrini, R., et al. (2008). Are cattle, sheep, and goats endangered species? *Mol. Ecol.* 17, 275–284. doi: 10.1111/j.1365-294X.2007.03475.x

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460.

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 Biol.* 5, 1587–1602. doi: 10.1371/journal.pbio.0050171

The Bovine HapMap Consortium. (2009). Genome-wide survey of snp variation uncovers the genetic structure of cattle breeds. *Science* 324, 528 –532. doi: 10.1126/science.1167936

Utsunomiya, Y. T., Pérez O'Brien, A. M., Sonstegard, T. S., Van Tassell, C. P., do Carmo, A. S., Mészáros, G., et al. (2013). Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. *PLoS ONE* 8:e64280. doi: 10.1371/journal.pone.0064280

Vitti, J. J., Grossman, S. R., and Sabeti, P. C. (2013). Detecting natural selection in genomic data. *Annu. Rev. Genet.* 47, 97–120. doi: 10.1146/annurev-genet-111212-133526

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.* 4:72. doi: 10.1371/journal.pbio.0040072

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

Wright, S. (1950). Genetical structure of populations. *Nature* 166, 247–249. doi: 10.1038/166247a0

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.