DISCOVERY OF SIGNATURES OF SELECTION IN BEEF AND DAIRY CATTLE USING ULTRA HIGH-DENSITY SNP GENOTYPES

I.A.S. Randhawa[1,3], M.S. Khatkar[1], P.C. Thomson[2], R.D. Schnabel[4], J.F. Taylor[4] and H.W. Raadsma[1]

[1]Sydney School of Veterinary Science, University of Sydney, Camden, NSW, Australia
[2]School of Life and Environmental Sciences, University of Sydney, Camden, NSW, Australia
[3]School of Veterinary Sciences, University of Queensland, Gatton, QLD, Australia
[4]Division of Animal Sciences, University of Missouri, Columbia, MO, 65211 USA

SUMMARY

Molecular data can provide insights into historical natural or artificial selection events and the genetic architecture underlying breed-specific traits. We examined the impact of phase of ancestral and derived alleles, marker density and composition of reference population panels to validate known and discovered novel genomic regions under selection in Angus and Holstein cattle. Using a composite selection signals method which combines multiple tests into a single score, 57 regions in Angus and 55 regions in Holstein cattle were detected by using ultra-high-density genotypes (2.5M SNPs) compared to four regions in Angus and five regions in Holstein Friesian using a low-density 50K SNP genotyping strategy. The detected regions include many regions known to harbour variants associated with selected beef and dairy traits, as well as several novel putative selection signatures. We conclude that both marker density and composition of the reference panel affects the power to detect selection signatures in domestic cattle.

INTRODUCTION

Discovery of genetic variants and genes has been intensified during the past two decades to understand the biological control of agricultural and health traits in livestock species (Kemper and Goddard 2012). Selective breeding has improved trait performance by increasing the frequency of beneficial alleles throughout the genome. Genomic regions under selection are likely to be of functional importance and recurrent selection has left distinct imprints throughout the genome, called signatures of selection, by producing deviations in allele frequencies, reduced local nucleotide variability and increased linkage disequilibrium (LD) within long haplotypes (Randhawa *et al.* 2014). Angus and Holstein cattle, in particular, have been subjected to long-term selection for beef and dairy production, respectively, and have previously been studied to detect signatures of selection to identify the genomic regions that harbour functional variants influencing beef and dairy traits (Randhawa *et al.* 2016). Genomic investigations are frequently resource-intensive, however, detection of selection signatures can provide insights into the genetic architecture underlying breed-specific traits (Gibbs *et al.* 2009) in a relatively cost-effective manner.

Detection of selection signatures is strongly influenced by significance thresholds and the power of the selection tests, genotypic marker density, sample size, minor allele frequency and origin of ancestral alleles, structure of candidate populations, and the composition of the reference population (Randhawa *et al.* 2016). With the advent of ultra-high density genotyping platforms and whole-genome resequencing, the information per screened individual has risen exponentially, however often at a higher cost per tested sample, resulting in relatively few individuals per population being screened. As such, studies of selection signatures can be confounded by SNP density, sample size of candidate and reference populations. This study examined the impact of phase of ancestral and derived alleles, marker density and composition of reference population panels to validate known and to discover novel genomic regions under selection in Angus and Holstein cattle.

MATERIALS AND METHODS

A composite selection signals method (CSS), which here combined three tests into a single score (Randhawa *et al.* 2014), was used for the analysis of selection signatures. Firstly, the CSS method was evaluated by substituting one of the constituent tests, viz., ΔDAF (change in derived allele frequency), which requires ancestral and derived allele phase to be known, with the ΔSAF (change in selected allele frequency) test, which does not require known allele phases (Randhawa *et al.* 2014) by using phase-known low-density SNP data produced using the Illumina BovineSNP50 BeadChip assay ("50K SNPs") to assay 1630 samples representing 57 breeds of European cattle (Table 1) (Randhawa *et al.* 2014). Next, the CSS method was evaluated for the confounding effects of using different reference panel compositions in selection tests across breeds, by using ultra-high-density SNPs genotyped with a pre-screening assay comprising almost 3 million validated SNPs in collaboration with Affymetrix (Santa Clara, CA) to design the Axiom Genome-Wide BOS 1 Array Plate ("2.5M SNPs") comprising 105 samples (Table 1) from seven breeds: Angus, Holstein, Brahman, Hanwoo, Murray Grey, Simmental and Wagyu (Rothammer *et al.* 2013). For both datasets, imputation of missing genotypes and haplotype phasing was performed with BEAGLE 3.3 (Browning and Browning 2009). CSS scores were smoothed by averaging $-\log_{10}(p)$ values for SNPs within 1 Mb or 50 kb overlapping windows. The top 0.1% of smoothed CSS scores were used for significance thresholds and the boundaries of selection regions were defined by the contiguous SNPs within the top 1% and 0.5%, respectively for 50K and 2.5M datasets. Genome-wide significant regions under selection were detected in Angus and Holstein by using the 2.5M SNPs and these were compared with regions detected using the 50K dataset and the meta-assembly reported by Randhawa *et al.* (2016). It is noteworthy that the individuals from each of the breeds genotyped with the two panels were different. Therefore, this study provides a comparison of independent breed samples genotyped using two SNP densities.

Table 1. DNA samples, their geographic origin, country of sampling and genotypes data

Breeds	Geographic origin	Country of sampling	Two datasets	
			50K	2.5M
Angus	Scotland	Australia, New Zealand, USA	128	29
Holstein	Netherlands	Australia, France, New Zealand, USA	160	40
Other breeds' DNA	Worldwide	Worldwide	1342	36
Total DNA samples	-	-	1630	105
Reference samples*	-	-	165	45
SNPs (N) genotyped	-	-	54,609	2,575,339
SNPs (N) after QC	-	-	35,284	1,583,288

* Reference population comprised randomly selected equal numbers of samples from each breed with 50K (n = 165) or 2.5M (n = 45) genotypes.

RESULTS AND DISCUSSION

After quality control (call rate ≥ 95%), a total of 35,284 and 1,583,288 autosomal SNPs from the 50K and 2.5M datasets, with a mean inter-marker interval of 70.845 kb and 1.584 kb, respectively, and with MAF > 0.01 were used in the analyses (Table 1). Analyses of the 50K SNP dataset suggested that using either of the allele frequency-based tests of selection, ΔSAF and ΔDAF, detected the same regions under selection in both breeds within the top 0.5% of CSS scores. Correlations of scores between the two approaches were 0.92 in Angus and 0.91 in Holstein (Figure 1). Notably, using ΔSAF for computing CSS resulted in the detection of additional regions relative to using ΔDAF, because

the latter is limited to derived alleles, while the former can detect selection affecting both ancestral and derived alleles. Overall, 4 regions (BTAs: 4, 13, 18, 21) and 5 regions (BTAs: 8, 10, 13, 20, 26) containing selection signatures were detected in Angus and Holstein, respectively, using 50K SNPs. Of these 9 regions, all but one Holstein region (BTA 13) were detected in the analysis of the 2.5M SNPs, suggesting the robustness of the CSS approach to SNP density using independent samples.

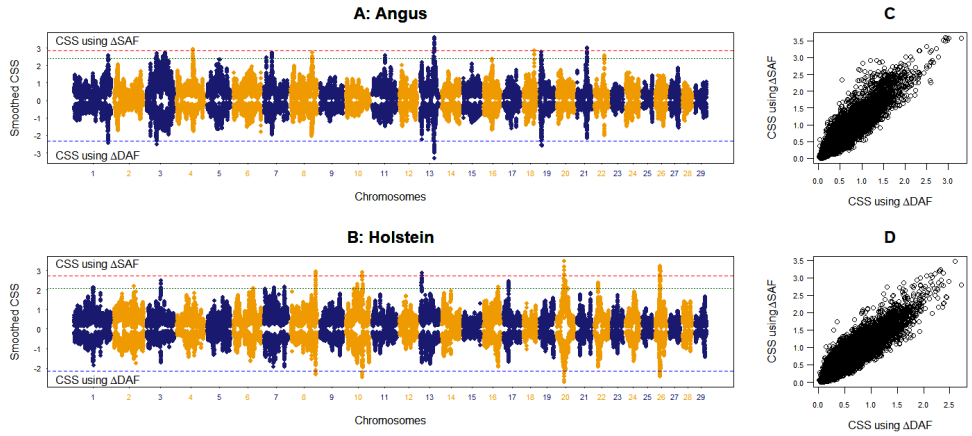

Figure 1. Comparison between the genome-wide smoothed CSS scores computed using ΔSAF and ΔDAF based on the 50K SNP dataset in Angus (A) and Holstein (B). Dashed (red and blue) and dotted (green) lines indicate the top 0.1% and 0.5% thresholds. Scatter plots of smoothed CSS scores computed using ΔSAF and ΔDAF in Angus (C) and Holstein (D)

Evaluation of three different reference panel compositions (1: single-breed, by Holstein vs Angus, 2: multi-breed, using all samples per breed, 3: multi-breed, with equal numbers of samples per breed) using three control regions on BTA6 (*ABCG2*, *KIT* and the casein cluster) in Holstein suggested that option 3 was optimal (results not shown). Hence, the final CSS analyses were performed using the optimal reference population strategy for Angus and Holstein and the ultra-high-density 2.5M SNPs and identified a total of 57 and 55 genomic regions, respectively (Figure 2).

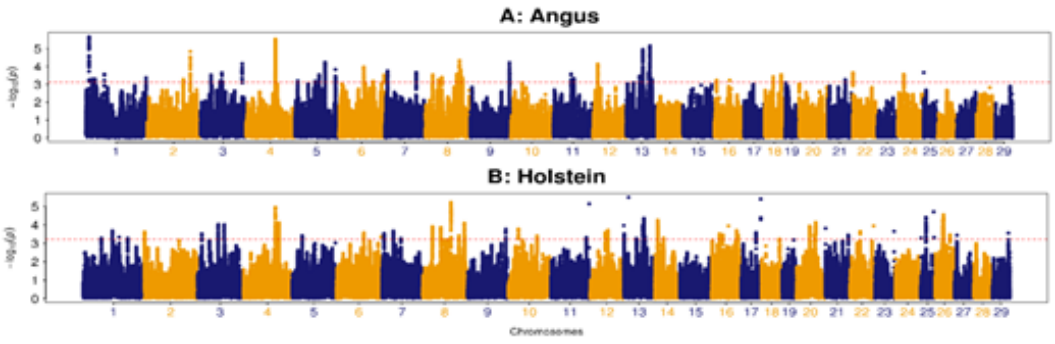

Figure 2. Manhattan plots showing smoothed CSS ($-\log_{10}(p)$) for Angus (A) and Holstein (B) using the 2.5M dataset. Horizontal dashed (red) lines indicate top 0.1% SNP thresholds

The significant CSS regions found in Angus and Holstein were compared against meta selection score (MSS) data, previously computed using a collection of published signatures of selection from 14 and 22 studies, respectively (Randhawa *et al.* 2016). Most of the significant CSS regions in both breeds coincided with regions with MSS ≤ 3. However, the distribution of MSS > 3, which indicates that a signature of selection has been validated in more than one study, was more frequently co-located with significant CSS regions in Holstein than in Angus. We also investigated the localization of CSS peaks for the presence of genes, because the bovine genome has variable gene density with an average of approximately nine genes per Mb. Gene locations in the UMD3.1 assembly indicate that 7.5% of 1 Mb genomic regions contain no genes (gene-sparse), whereas, 5% of 1 Mb regions contain 30-78 genes (gene-dense). Selection signatures in Angus and Holstein were mostly present in regions with variable gene-density and only ~7% (4 out of 57) and ~ 2% (1 out of 55), respectively, of significant CSS regions did not contain annotated genes. Given that the breeds are specialized for beef and dairy production, only five regions were found in common (BTAs: 1, 13, 19, 21 and 22).

This study detected several novel regions in both breeds; however, gene-dense regions can make it difficult to predict the underlying functional mutation. The identified signatures of selection show that selective forces have operated on the genetic architecture controlling growth and body size of beef cattle, and the physiological and anatomical structure of mammary glands, and quantity and quality of various milk components in dairy cattle. For example, *ABCG2* has been found to be involved in milk yield and composition and is a lactation regulator. *ABCG2*, along with the *NCAPG-LCORL* genes, has been found to be associated with stature in Holstein (Randhawa *et al.* 2016). *GHR* (Rothammer *et al.* 2013) is also a strong candidate gene with a major effect on milk yield and composition and is linked to many QTLs and is located in a region in which strong selection signatures have been identified in multiple cattle breeds (Khatkar *et al.* 2014). Additional genes underlying selective sweeps detected by CSS in this study, such as *SAR1B, AGTRAP* and *KIF1B* (Flori *et al.* 2009) are involved in the functioning of mammary glands, milk production and disease resistance in high producing dairy cows. The non-dairy related genes include *KIT* for white-spotting coat colour. Moreover, *PDGFRA* and *KDR* (Flori *et al.* 2009; Gautier and Naves 2011; Randhawa *et al.* 2014), *MGAT1*, *SPOCK1* (Gibbs *et al.* 2009) and *SIGLEC* genes (Khatkar *et al.* 2014) have been implicated with reproduction traits, due to their roles in fertilization, embryonic development and growth.

REFERENCES

Browning B.L. and Browning S.R. (2009) *The American Journal of Human Genetics* **84**: 210.
Flori L., Fritz S., Jaffrézic F., Boussaha M., Gut I., Heath S., Foulley J.-L. and Gautier M. (2009) *PLoS One* **4**: e6595.
Gautier M. and Naves M. (2011) *Mol. Ecol.* **20**: 3128.
Gibbs R.A., Taylor J.F., Van Tassell C.P., Barendse W., Eversole K.A., Gill C.A., Green R.D., Hamernik D.L., Kappes S.M., *et al.* (2009) *Science* **324**: 528.
Kemper K.E. and Goddard M.E. (2012) *Hum. Mol. Genet.* **21**: R45.
Khatkar M.S., Randhawa I.A.S. and Raadsma H.W. (2014) *Livestock Science* **166**: 144.
Randhawa I.A.S., Khatkar M.S., Thomson P.C. and Raadsma H.W. (2014) *BMC Genet.* **15**: 34.
Randhawa I.A.S., Khatkar M.S., Thomson P.C. and Raadsma H.W. (2016) *PLoS One* **11**: e0153013.
Rothammer S., Seichter D., Forster M. and Medugorac I. (2013) *BMC Genomics* **14**: 908.