EVIDENCE OF SELECTIVE SWEEPS THROUGH HAPLOTYPE STRUCTURE OF PINZGAU CATTLE

Radovan KASARDA 1, Nina MORAVČÍKOVÁ 2, Veronika KUKUČKOVÁ 3, Ondrej KADLEČÍK 4, Anna TRAKOVICKÁ 5, Gábor MÉSZÁROS 6

ABSTRACT

The aim of this study was to identify the genomic regions potentially under positive selection through the determination of integrated Haplotype Homozygosity Score in dual-purpose Pinzgau cattle. The genomic data from in total 37 Slovak and 105 Austrian purebred bulls were obtained using Illumina high-density SNP panels. After applying the quality control the final dataset composed of overall 27,895 autosomal loci with SNP density at the level one locus per 89.31 kb. The single-site iHS values have been scored for each SNP and subsequently averaged into the non-overlapping 500 kb windows across the genome. Totally, 27,293 iHS values across 4968 sliding windows were tested for the evidence of selection signatures. The most significant iHS values were detected for 6 genomic regions on BTA 2, 4, 5, 7, 8 and 22. The iHS value was in average 0.03 and the highest score of iHS (2.09) has been found in region on BTA 5. Our results provide initial insight and basis for future studies of genome-wide associations and impact of artificial selection in Slovak Pinzgau cattle.

Key words: cattle, breeds, Pinzgau, genetics, selection signals, high-density SNP arrays, iHS score

1 INTRODUCTION

The Pinzgau cattle in Europe traditionally belongs to the dual-purpose cattle breeds with combined milk and meat production (Pavlík et al., 2013; Kadlecík et al., 2004). The nucleus of the breed is based upon two populations registered in Herd books of Austria and Slovakia. In Slovak republic the Pinzgau cattle is bred for more than 150 years. It was imported from the Austrian Alps and therefore the two populations are well connected (Kadlečík et al., 2011). The breeding management is specific in Slovakia however, where the genetic reserve is bred under intensive farming conditions in sub-mountainous regions. The nucleus of the breed is formed from almost 50 % (in Slovakia) and more than 65 % (in Austria) of animals in suckling-cow system (Kasarda et al., 2016). The breeders of Pinzgau cattle in Slovakia and also in Austria are interested in the future development of the breed, when integrating both pedigree and genomic data (Kasarda et al., 2009; Mészáros et al., 2010; Kasarda et al., 2015).

Generally, the formation of cattle breeds is the result of human selection, adaptation to different environments and demographic effects such as domestication, migration and selection, all contributing to the patterns of genetic diversity (Mancini et al., 2014). Artificial selection in cattle has resulted in divergent breeds that are specialized for either milk or meat production or raised...
as dual-purpose breeds. The selection strategies for traits of interest led to the increase of selection pressure on particular genomic regions that control traits of productive and economic importance (Zhao et al., 2015). When selection acts on a mutation, it also affects linked loci and leaves so-called signatures in the flanking chromosomal regions. If intensive selection pressure occurred only over a few generations, it is unlikely that recombination has an impact on haplotype structure, and thus it resulted in linkage disequilibrium patterns between the mutation and the neighbouring loci. The signal of such selection pressure on genes can be detected as a spectrum of allele frequencies among closely linked loci that is shifted towards extreme frequencies, an excess of homozygous genotypes, and a high frequency of long haplotypes (Saβeti et al., 2002; Makina et al., 2015; Zhao et al., 2015). The understanding of these signatures of selection is of interest from the perspective of evolutionary biology. Identification of candidate genes within these regions may ultimately help to further improve these breeds through informed management of breeding programmes (Qanbari and Simianer, 2014).

The availability of high density SNP genotyping arrays opens up the possibility of performing population genetic studies in cattle using thousands of markers distributed across the entire genome (Urbinati et al., 2016). The selection signature regions can be detected via several approaches. The studies applied in detection of selective sweeps are mainly based on estimation of differences in allele frequencies by contrasting pair of breeds through linkage disequilibrium patterns between the mutation and the neighbouring loci. Identification of candidate genes within these regions may ultimately help to further improve these breeds through informed management of breeding programmes (Qanbari and Simianer, 2014).

The understanding of these signatures of selection is of interest from the perspective of evolutionary biology. Identification of candidate genes within these regions may ultimately help to further improve these breeds through informed management of breeding programmes (Qanbari and Simianer, 2014).

The evidence of positive selection was determined based on evaluation of integrated Haplotype Homozygosity Score. The analysis of selective sweeps using iHS statistics is based on calculation of haplotype frequencies as specified by Voight et al. (2006). The haplotypes were reconstructed for each autosome using default options in fastPHASE software according to Scheet and Stephens (2006). Because the iHS approach requires the information of the ancestral and derived allele status for each locus, the set of ancestral alleles was established according to Rocha et al. (2014). In iHS statistics each locus is treated as core SNP and the test begins with calculation of extended Haplotype Homozygosity (EHH) for each core SNP. For a given focal SNP, the iHS is defined as the standardized log-ratio of the iHH (integrated EHH) for the ancestral (iHHa) and derived (iHHd) allele:

$$iHS^{(i)} = \frac{\log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right) - E_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)}{sd_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)}$$

The expectation $E_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)$ and standard deviation $sd_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)$ of the unstandardized score $\log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)$ are estimated from the empirical distribution at SNPs whose derived allele frequency $p$ matches the frequency at the core SNPs. Since iHS is standardized using the genome-wide empirical distributions, it provides a measure of how unusual the haplotypes around a given SNP are, relative to the whole genome (Voight et al., 2006). The iHS score has been computed using the R package “rehh” (Gautier and Vitalis, 2012). The single-site iHS values were averaged within non-overlapping windows of 500 kb across the genome. The observed major peaks defined regions with extreme iHS values based

## 2 MATERIAL AND METHODS

Genome wide single nucleotide polymorphism (SNP) data from 37 Slovak and 105 Austrian purebred Pinzgau bulls were used to detect selective sweeps. The genotyping was performed using Illumina high-density SNP panels (50K and 700K). Overall 42,444 SNPs common to both genotyping arrays were retained in reduced panel of loci. Markers assigned to unmapped regions or with unknown chromosomal position according to the latest bovine genome assembly (Btau 4.0) and SNPs positioned to sex chromosomes were removed. The quality control of genotyping data were performed according to Purcell et al. (2007) to exclude any autosomal locus with call rate lower than 90 %, minor allele frequency lower than 0.05 and HWE limit of 0.0001. The final dataset has been composed of 27,895 autosomal loci with SNP density at the level one locus per 89.31 kb.

The evidence of positive selection was determined based on evaluation of integrated Haplotype Homozygosity Score. The analysis of selective sweeps using iHS statistics is based on calculation of haplotype frequencies as specified by Voight et al. (2006). The haplotypes were reconstructed for each autosome using default options in fastPHASE software according to Scheet and Stephens (2006). Because the iHS approach requires the information of the ancestral and derived allele status for each locus, the set of ancestral alleles was established according to Rocha et al. (2014). In iHS statistics each locus is treated as core SNP and the test begins with calculation of extended Haplotype Homozygosity (EHH) for each core SNP. For a given focal SNP, the iHS is defined as the standardized log-ratio of the iHH (integrated EHH) for the ancestral (iHHa) and derived (iHHd) allele:

$$iHS^{(i)} = \frac{\log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right) - E_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)}{sd_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)}$$

The expectation $E_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)$ and standard deviation $sd_p \log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)$ of the unstandardized score $\log \left( \frac{iHH^{(i)}_a}{iHH^{(i)}_d} \right)$ are estimated from the empirical distribution at SNPs whose derived allele frequency $p$ matches the frequency at the core SNPs. Since iHS is standardized using the genome-wide empirical distributions, it provides a measure of how unusual the haplotypes around a given SNP are, relative to the whole genome (Voight et al., 2006). The iHS score has been computed using the R package “rehh” (Gautier and Vitalis, 2012). The single-site iHS values were averaged within non-overlapping windows of 500 kb across the genome. The observed major peaks defined regions with extreme iHS values based
on the outliers according to obtained boxplot distribution were evaluated as signals of the positive selection.

3 RESULTS AND DISCUSSION

The iHS approach introduced by Voight et al. (2006) was applied to determine the selective sweeps across Pinzgau populations. This approach is based on the ratio of the integrated EHH of two ancestral and derived core alleles. The ancestral alleles were defined based on genomic data of species considered to be a common founder of the Bovinae subfamily, in which the alleles were fixed at MAF equal to zero as demonstrated Utsunomiya et al. (2013) and Rocha et al. (2014). The single-site iHS values have been scored for each SNP and subsequently averaged into the non-overlapping 500 kb windows across the genome. The sliding windows size was adopted based on approach described by Qanbari et al. (2011). Overall 27,293 iHS values across 4968 sliding windows were tested for the evidence of selection pressure. Negative iHS values indicated long haplotypes carrying derived alleles and positive iHS values represented long haplotypes carrying ancestral alleles, and small or close to zero iHS values indicated similar rates of EHH decay for ancestral and derived SNP alleles (Urbinati et al., 2016). However, an extreme positive iHS score has been considered as a candidate because the ancestral allele itself may be the target of selection (Qanbari and Simianer, 2014). The signals of recent selection were defined as genomic regions with positive iHS values based on the outliers according to boxplot distribution (Fig. 1B). Thus, the genomic regions were considered as recently selected when the iHS score of multiple loci located within 0.5 Mb was greater than 1.18. Table 1 shows the summary of autosomal genomic regions displaying the strongest iHS signals. Total number of SNPs that can be regarded as loci under selection was lower in comparison to the study of Urbinati et al. (2016), Makina et al. (2015) or Zhao et al. (2015). The most significant iHS value exceeding the threshold 1.18 was detected across genome for 6 sliding windows (Table 1). Figure 1 shows the genome distribution of iHS values averaged within 500 kb against the genomic position. The highest score of iHS (2.09) has been found in the region on chromosome 5. Between the selective sweeps, no major overlap has been identified. The obtained genomic regions did not show strong signal of selection possibly due to close relatedness among both Pinzgau populations.

Table 1: Summary of genomic regions with the strongest iHS signals

<table>
<thead>
<tr>
<th>BTA</th>
<th>Start (Mb)</th>
<th>End (Mb)</th>
<th>Average iHS value</th>
<th>No. of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>49.86</td>
<td>60.68</td>
<td>1.32</td>
<td>46</td>
</tr>
<tr>
<td>4</td>
<td>59.28</td>
<td>60.25</td>
<td>1.55</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>17.19</td>
<td>17.75</td>
<td>1.29</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>101.25</td>
<td>105.27</td>
<td>1.47</td>
<td>23</td>
</tr>
<tr>
<td>8</td>
<td>87.64</td>
<td>93.78</td>
<td>1.39</td>
<td>51</td>
</tr>
<tr>
<td>22</td>
<td>31.23</td>
<td>44.27</td>
<td>1.42</td>
<td>84</td>
</tr>
</tbody>
</table>

Figure 1: Distribution of Integrated Haplotype Score (iHS) averaged within 500 kb across the genome (A) and the boxplot of the iHS values (B)
Generally, the massive use of artificial selection based on distinct breeding goals has shaped the phenotypic appearance of cattle breeds. In this context, selection signatures provide an useful information about the specific genomic regions that are conserved across generations (Mancini et al., 2014; Urbinati et al., 2016). The progressive breed differentiation was accelerated by introduction of breed standards, which accumulated specific phenotypes that visibly distinguished particular breeds (Rothammer et al., 2013). The analysis of evidence of recent positive selection in such breeds can provide the information on genomic regions under both artificial and natural selection and may help to identify the beneficial mutations underlying biological pathways for economically important traits (Zhao et al., 2015). In comparison to our results many studies reported strong signals of positive selection in genomic regions containing QTLs for economically important traits in milk and beef cattle based on multiple approaches, including IHS (Utsunomiya et al. 2013; Rothhammer et al., 2013; Makina et al., 2015; Urbinati et al., 2016).

4 CONCLUSION

The genome-wide analysis of the selective sweeps through haplotype structure of Pinzgau cattle identified several genomic regions that reflected the impact of breeding goals of Slovak and Austrian herds. The genomic regions potentially under selection showed only relatively weak signals, possibly due to the similar selection goals and close genetic connectedness between both Pinzgau populations. The results provide the first insight and basis for the future studies of genome-wide associations and impact of artificial selection on improvement of Pinzgau breeding programs.

5 ACKNOWLEDGEMENT

This study was supported by the Slovak Research and Development Agency (Contract No. APVV-14-0054 and uder contract SK-AT-2015-0016).

6 REFERENCES