Original Paper

Assessing footprints of natural selection through PCA analysis in cattle

Nina Moravčíková[1]*, Veronika Kukučková[1], Gábor Mészáros[2],
Johann Sölkner[2], Ondrej Kadlečík[1], Radovan Kasarda[1]
[1]Slovak University of Agriculture in Nitra, Slovakia
[2]University of Natural Sciences and Life Sciences, Division of Livestock Sciences, Vienna, Austria

Article Details: Received: 2016-05-24 | Accepted: 2016-07-28 | Available online: 2017-06-20

http://dx.doi.org/10.15414/afz.2017.20.01.23–27

The aim of this study was to determine the population structure and to perform genome-wide scan of footprints of natural selection in cattle using principal component analysis. The applied statistics to identify the SNPs associated with selection pressure focused mainly on the extreme values of F_{ST} index. In our study the alternative individual-based approach adopted in the *PCAdapt* R package has been used. This approach is based on the assumption that markers extremely related to the population structure are also candidates for local adaptation of the population. The genotype data of 350 animals originating from four historically or geographically connected populations (Austrian Pinzgau, Slovak Pinzgau, Brown Swiss, Tyrol Grey) have been used to test this approach in cattle. As expected based on breed's origin the principal component analysis showed the division of animals in to the 3 separate clusters and the eigenvalues suggested to use of K = 3 as optimal number. The analysis of genomic regions harbouring signals revealed the candidate genes previously associated with muscle formation and immunity system. Detecting signals of adaptation that were also the targets of historical selection will allow in the future a better understanding of cattle origin.

Keywords: local adaptation, selection, cattle, SNP50 BeadChip, *PCAdapt*, population subdivision

1 Introduction

The development of sequencing and genotyping technologies will make possible to answer many important biological questions in conservation genetics that have been intractable until now. Today more than 800 cattle breeds were described across the world. The cattle genome therefore represents a valuable source for identifying genetic variation that contributes to evaluation of phenotypic diversity. From a genetic perspective, the evolution can be defined as changes in allele frequencies over time due to mutation, genetic drift, migration, and natural selection (Allendorf et al., 2010; Qanbari et al., 2011).

A challenge of genome-wide analysis is to determine patterns of nucleotide variation that can be explained by random drift versus selection pressure. Aspects of selection signatures depend on time, age and strength of selection events. Natural selection, the process by which organisms that are best adapted to their environment have an increased contribution of genetic variants to future generations, acts in at least three ways: positive, purifying and balancing selection (Oleksyk et al., 2010; Martins et al., 2016). Positive natural selection or local adaptation is the driving force behind the adaptation of individuals to their environment. In order to provide a list of variants that are potentially involved in natural selection, genome scans measure the genetic differentiation between populations considering that extreme values correspond to candidate regions (Duforet-Frebourg et al., 2015). Although high levels of differentiation can have various causes, adaptation of individuals to their local environment is a prominent explanation to such patterns of differentiation for adaptive loci exceeding neutral expectations (Duforet-Frebourg et al., 2014).

To detect the regions that have been targets of natural selection, various statistical approaches have been developed. One of the most frequently applied approach is based on the extreme values of the Wright's F_{ST} index that provides an estimate of genetic variability between populations (Nielsen et al., 2005, Weir et al., 2005). For selectively neutral loci the F_{ST} is determined by genetic drift that affects all SNPs across the genome in similar way. In contrast, natural selection has locus-specific effects that can cause deviations in F_{ST} values at selected and linked loci (Akey et al., 2002). However, there are

***Corresponding Author:** Nina Moravčíková, Slovak University of Agriculture in Nitra, Faculty of Agrobiology and Food resources, Tr. Andreja Hlinku 2, 949 76 Nitra, Slovakia, e-mail: nina.moravcikova@uniag.sk

important caveats with approaches related to F_{ST} because they require grouping individuals into populations which can be sometimes subjective and can result in the loss of important selection signals (Duforet-Frebourg et al., 2014). Bierne et al. (2013) found that the genome scan based on F_{ST} can produce many false positives of selection signature signals due to the various biological and statistical reasons. The computation of F_{ST} index becomes challenging mainly when the population is genetically homogenous, when defining subpopulations is difficult, and in the presence of admixture between individuals across analysed subpopulations (Waples and Gaggiotti, 2006; Novembre et al., 2008). Duforet-Frebourg et al. (2014) proposed alternative approach to determine candidate markers for natural selection based on principal component analysis (PCA) that use multivariate evaluation to the identify the population structure. The obtained correlations between genetic variants and each principal component provide a conceptual framework to identify the variants involved in local adaptation without a priory information of population structure (Duforet-Frebourg et al., 2015). The PCA based statistic implemented in *PCAdapt* R package provides three main advantages compared to the classical F_{ST} approach: works on individual basis, the computation time is reduced in comparison to methods that use the MCMC algorithms and candidate loci can be related to the different evolutionary events which correspond to the different principal components (Duforet-Frebourg et al., 2015; Luu et al., 2016).

The aim of this study was to perform genome-wide scan in order to determine the population structure and identify the regions that have been the targets of natural selection based on PCA method adopted in *PCAdapt* R package.

2 Material and methods

2.1 Genotyping data and quality control

To detect the footprints of selection the genotyping data from total of 350 animals originating from four historically and geographically connected populations (Austrian Pinzgau – AP, Slovak Pinzgau – SP, Tyrol Grey – TG and Brown Swiss – BS) was analysed. The final dataset was created by mergeing of new data from 37 SP breeding bulls that were genotyped by the Illumina BovineSNP50 v2 BeadChip and previously published information of 105 AP, 105 TG and 103 BS bulls obtained using the BovineSNP v1 BeadChip (Illumina Inc., San Diego, CA) as described Ferenčakovič et al. (2013). The 47065 SNPs common to both analysed datasets were retained in reduced panel of SNPs. From this in total of 35801 SNPs passed subsequent quality control that was conducted to eliminate any SNPs with call rate lower than 90%, minor allele frequency lower than 0.05 and Hardy-Weinberg equilibrium limit of 0.00001.

2.2 Analysis of selection footprints

The genome scan for selection was performed based on the approach adopted in *PCAdapt* R (Duforet-Frebourg et al., 2015) package according to Luu et al. (2016). The detection of outliers, the SNPs that are associated with selection, is based on the vector of *z*-scores obtained when regressing SNPs with the K principal components. The test statistic is the Mahalanobis distance, a multivariate method that measures the distance of the point from the mean. Denoting by $z^j = (z_1^j, ..., z_K^j)$ the vector of K *z*-scores between the *j*-th SNP and the first K PCs, the sqaured Mahalanobis distance is defined according to the Duforet-Frebourg et al. (2015) as:

$$D_j^2 = (z^j - z)\ \Sigma^{-1}\ (z^j - \bar{z})$$

where:
$\bar{z}$ and Σ are estimates of the z-score mean and covariance matrix of *z*-scores, respectively

The Mahalanobis distance should be transformed into the *p*-values to perform the multiple hypothesis testing. To determine the threshold of *p*-values Luu et al. (2016) recommend to use of false discovery rate (FDR) approach that provide a list of candidate markers with as expected proportion of false discoveries lower than specified value. The controlling of FDR is based on the *q*-value procedure that is adopted in the *q*-value *R* package (Storey, 2002) which transform the *p*-values into the *q*-values and allow the control of specified value α of FDR and detection of candidate SNPs with *q*-values lower than specified α (Luu et al. 2016; Duforet-Frebourg et al., 2015).

3 Results and discussion

An alternative approach based on PCA analysis (Duforet-Frebourg et al., 2014; Duforet-Frebourg et al., 2015) has been used to determine the population structure without a priori information about population subdivision and to perform genome scan to identify SNPs associated to local adaptation in cattle. As expected based on populations origin the first and the second principal components separated the population structure to the tree genetic clusters (Figure 1A). The Slovak and Austrian Pinzgau populations have been linked into the one group mainly due to the high genetic similarity between them that can be attributed to the common ancestors. The decay of eigenvalues confirmed to use of $K = 3$ as optimal because the eigenvalues decreased between $K = 3$ and $K = 5$. Figure 1B displays in decreasing order the percentage of variance explained by each PCs.

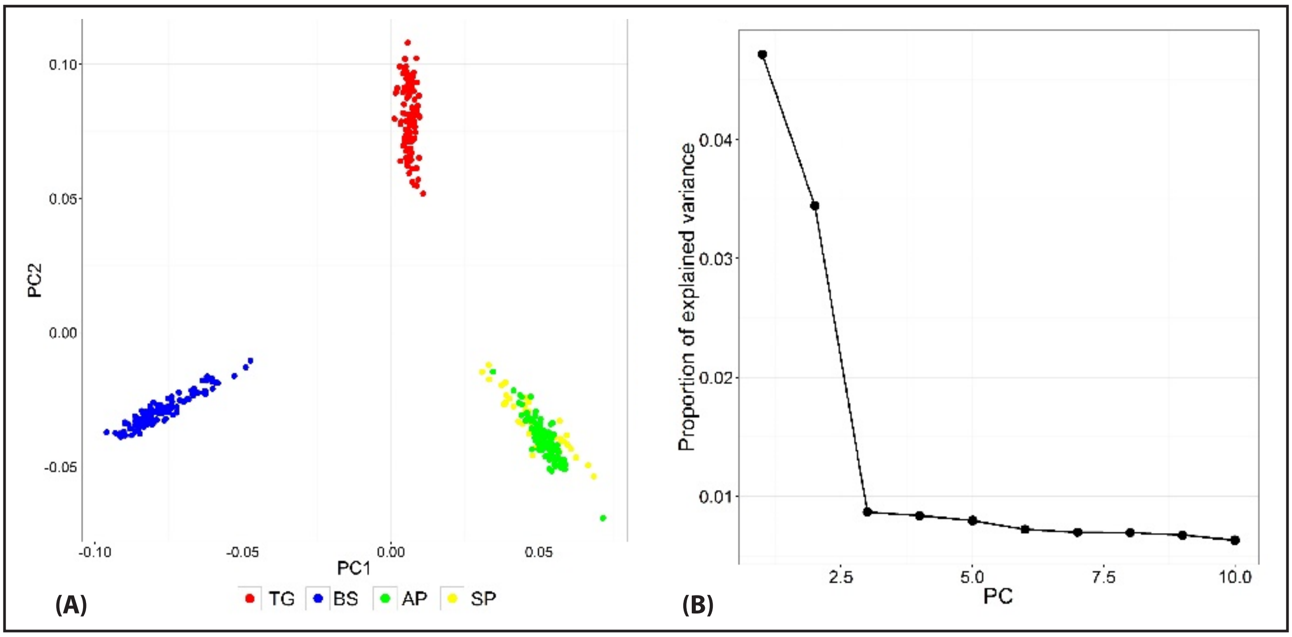

Figure 1 Score plot of population structure (A) and the proportion of variance explained by 10 PCs (B)

A histogram of *p*-values confirmed that most of them followed the uniform distribution. Figure 2A shows that the *p*-values were well calibrated since there was a mixture of uniform distribution and of a peaky distribution around 0, which corresponded to outlier loci (Storey, 2002; Luu et al., 2016). The distribution of the *p*-values was check also by using a Q-Q plot that confirmed the expected uniform distribution of the most of *p*-values (Figure 2B). The presence of outlier loci indicated the lowest *p*-values that were smaller than expectations.

Figure 3 shows a Manhattan plot indicating the main outlier SNPs that have been detected using the genome scan for footprints of natural selection. Based on the expected FDR equal to 10% we were able to determine 22 outliers with the approach implemented in *PCAdapt* package. Some of these were located near the genomic regions containing the candidate genes like *GHR*, *CAPN2*, *CAPN3*, *IL21* and *IL2* that were previously significantly associated mainly with muscle formation, immunity system as well as economically important production traits in cattle (McClure et al., 2012; Giusti et al., 2013; Gowane et al., 2014).

Compilation of the results from many studies in cattle provides an ideal opportunity to investigate how selection has influenced the variability and architecture of the bovine genome. Selection is likely to have eroded the levels of genetic variation that existed in the original domesticated population. At the same time, selection on

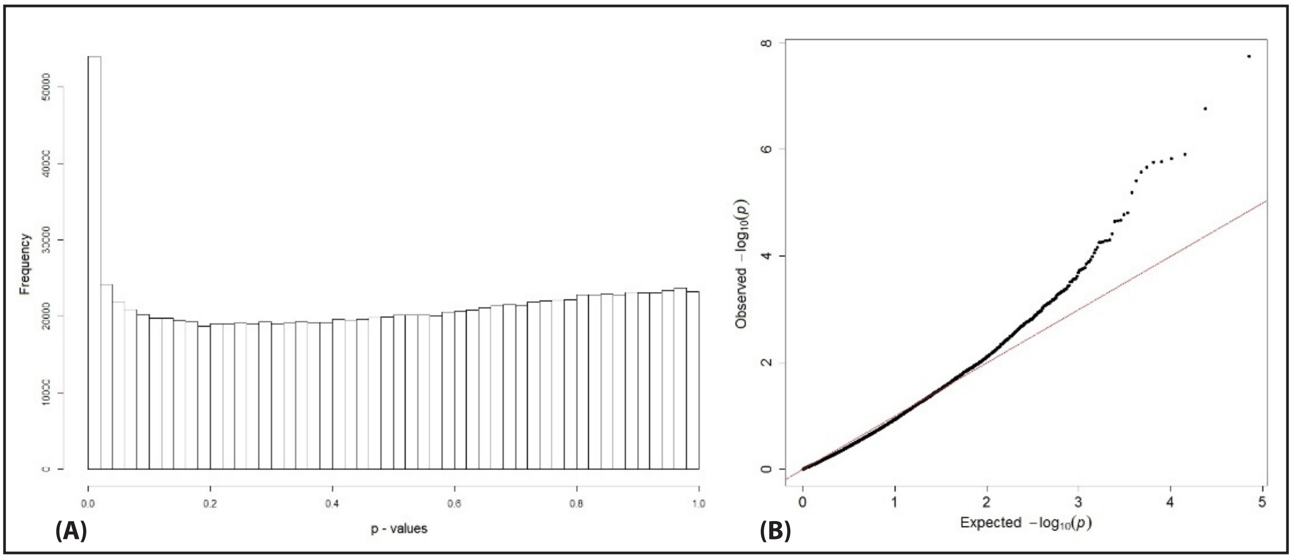

Figure 2 Histogram (A) and Q-Q plot (B) of p-values

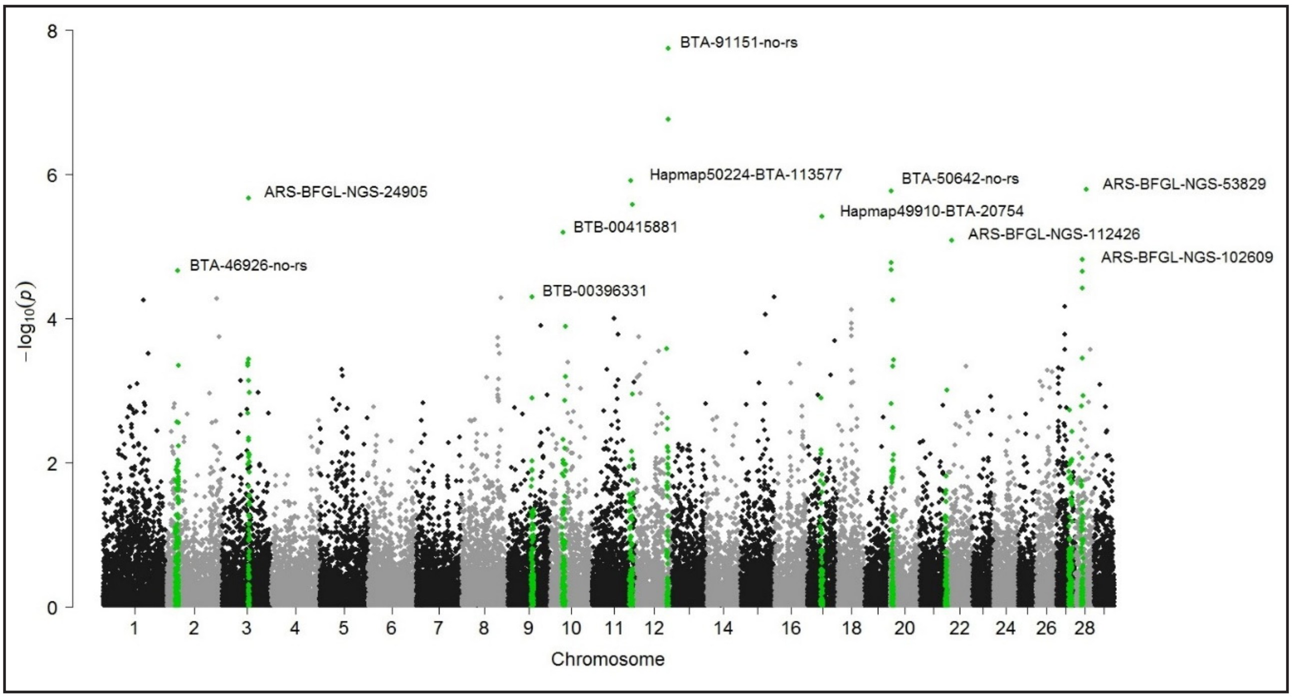

Figure 2 Manhattan plot of -log10 (*p*-values). The candidate genomic regions containing the SNPs associated with selection are coloured in green

a livestock breed has tended to fix specific variants that have become distinctive genetic signals of that breed compared with others (Gutiérrez-Gil et al., 2015). The presence of strong selective footprints across the bovine genome were tested in multiple populations using various approaches mainly based on site frequency spectrum, population differentiation represented by F_{ST} statistic and haplotype length (extend of linkage disequilibrium) (Qanbari et al., 2011; Druet et al., 2013; Mancini et al., 2014; Zhao et al. 2015). The number of identified selective sweeps varied across different studies, depending on methodological approach and analysed populations. A high number of selective sweeps was presented by Stella et al. (2010) for five specialized dairy cattle breeds (215 regions) and also by Druet et al. (2013) for 12 breeds of different production type (147 regions). The much lower proportion of selective footprints (16 regions) was found by Flori et al. (2009) for French dairy cattle breeds and Mancini et al. (2014) for Italian breeds. Despite the relatively low number of identified candidate loci in our study we showed that the alternative approach proposed by Duforet-Frebourg et al. (2015) can be a perspective alternative for the identification of selection sweeps in cattle.

4 Conclusions

The analysis of genomic regions associated with natural selection is necessary to understand the biological significance of molecular variations involved in adaptation mechanisms in cattle. Alongside commonly used methods for determination of selection footprints the individual-based approach using PCA analysis provided comparable results with previously published studies in this area. The genome scan for footprints of natural selection across Pinzgau, Braun Swiss and Tyrol Grey populations indicated 22 outlier loci that were the most strongly correlated to the observed population structure (FDR equal to 10%). Detected signals were mainly in genomic regions containing genes involved in muscle formation, body growth and immunity system, suggesting a connection to the natural selection events during breed development. Cattle breeds involved in this study belong to the mountainous group of cattle, traits of muscle formation, body growth and immunity are in agreement with the overall description of those breeds. Also in present valuable for farmers due to the resistance to harsh conditions, durability and longevity. The results indicate that those regions are important not artificially but naturally to survive local (mountain) conditions.

Acknowledgments

This work has been supported by the Slovak Research and Development Agency (IDs No. APVV-14-0054 and SK-AT-2015-0016). Part of the work was done during the stay of the first and second author at BOKU Vienna, supported by the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH, project SK 07/2016).

References

AKEY, J. M. et al. (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research,* vol. 12, pp. 1805–1814. doi:http://dx.doi.org/10.1101/gr.631202

ALLENDORF, F. W., HOHENLOHE, P. A. and LUIKART, G. (2010) Genomics and the future of conservation genetics. *Nature Reviews Genetics,* vol. 11, no. 10, pp. 697–709. doi:http://dx.doi.org/10.1038/nrg2844

BIERNE, N., ROZE, D. and WELCH, J. J. (2013) Pervasive selection or is it ...? Why are F_{ST} outliers sometimes so frequent?. *Molecular Ecology,* vol. 22, pp. 2061–2064. doi:http://dx.doi.org/10.1111/mec.12241

DRUET, T. et al. (2013). Identification of large selective sweeps associated with major genes in cattle. *Animal Genetics,* vol. 44, pp. 758–762. doi:http://dx.doi.org/10.1111/age.12073

DUFORET-FREBOURG, N. et al. (2015) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Molecular biology and evolution,* vol. 33, pp. 1082–1093. doi:http://dx.doi.org/10.1093/molbev/msv334

DUFORET-FREBOURG, N., BAZIN, E. and BLUM, M. G. B. (2014) Genome scans for detecting footprints of local adaptation using a bayesian factor model. *Molecular biology and evolution,* vol. 31, pp. 2483–2495. doi:http://dx.doi.org/10.1093/molbev/msu182

FERENČAKOVIĆ, M., SOLKNER, J. and CURIK, I. (2013) Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. *Genetic Selection Evolution,* vol. 45, no. 1, pp. 42. doi:http://dx.doi.org/10.1186/1297-9686-45-42

FLORI, L. et al. (2009) The genome response to artificial selection: a case study in dairy cattle. *PLoS One,* vol. 4, e6595. doi:http://dx.doi.org/10.1371/journal.pone.0006595

GIUSTI, J. et al. (2013) Expression of genes related to quality of *Longissimus dorsi* muscle meat in Nellore (*Bos indicus*) and Canchim (5/8 *Bos taurus* × 3/8 *Bos indicus*) cattle. *Meat Science,* vol. 94, no. 2, pp. 247–252. doi:http://dx.doi.org/10.1016/j.meatsci.2013.02.006

GOWANE, G. R. et al. (2014) The Expression of IL6 and 21 in Crossbred Calves Upregulated by Inactivated Trivalent FMD Vaccine. *Animal Biotechnology,* vol. 25, no. 2, pp. 108–118. doi:http://dx.doi.org/10.1080/10495398.2013.834826

GUTIÉRREZ-GIL, B., ARRANZ, J. J. and WIENER, P. (2015) An interpretive review of selective sweep studies in *Bos taurus* cattle populations: identification of unique and shared selection signals across breeds. *Frontiers genetics,* vol. 6, pp. 167. doi:http://dx.doi.org/10.3389/fgene.2015.00167

LUU, K., BAZIN, E. and BLUM, M. G. B. (2016) pcadapt: An R package for performing genome scans for selection based on principal component analysis. *bioRxiv.* doi:http://dx.doi.org/10.1101/056135

MANCINI, G. et al. (2014) Signatures of selection in five Italian cattle breeds detected by a 54K SNP panel. *Molecular Biology Reports,* vol. 41, pp. 957–965. doi:http://dx.doi.org/10.1007/s11033-013-2940-5

MARTINS, H. et al. (2016) Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *bioRxiv,* p. 054585. doi:http://dx.doi.org/10.1101/054585

McCLURE, M. C. et al. (2012) Genome-wide association analysis for quantitative trait loci influencing Warner-Bratzler shear force in five taurine cattle breeds. *Animal Genetics,* vol. 43, no. 6, pp. 662–673. doi:http://dx.doi.org/10.1111/j.1365-2052.2012.02323.x

NIELSEN, R. (2005) Molecular signatures of natural selection. *Annual Review of Genetics,* vol. 39, pp. 197–218. doi:http://dx.doi.org/10.1146/annurev.genet.39.073003.112420

NOVEMBRE, J. et al. (2008) Genes mirror geography within Europe. *Nature,* vol. 456, pp. 98–101. doi:http://dx.doi.org/10.1038/nature07566

OLEKSYK, T. K., SMITH, M. W. and O'BRIEN, S. J. (2010) Genome-wide scan for footprints of natural selection. Philosophical Transactions of the Royal Society B: *Biological Sciences,* vol. 365, pp. 185–205. doi:http://dx.doi.org/10.1098/rstb.2009.0219

QANBARI, S. et al. (2011) Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics,* vol. 12, pp. 318. doi:http://dx.doi.org/10.1186/1471-2164-12-318

STELLA, A. et al. (2010) Identification of selection signatures in cattle breeds selected for dairy production. *Genetics,* vol. 185, pp. 1451–1461. doi:http://dx.doi.org/10.1534/genetics.110.116111

STOREY, J. D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society,* Series B, vol. 64, pp. 479–498. doi:http://dx.doi.org/10.1111/1467-9868.00346

WAPLES, R. S. and GAGGIOTTI, O. (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology,* vol. 15, pp. 1419–1439. doi:http://dx.doi.org/10.1111/j.1365-294x.2006.02890.x

WEIR, B. S. et al. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Research,* vol. 15, no. 11, pp. 1468–1476. doi:http://dx.doi.org/10.1101/gr.4398405

ZHAO, F. et al. (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information. *Genetic Selection Evolution,* vol. 47, pp. 49. doi:http://dx.doi.org/10.1186/s12711-015-0127-3