

## Selection signatures in autochthonous Spanish cattle breeds using site frequency spectrum statistics

S. Munilla<sup>1,2</sup>, A. González-Rodríguez<sup>2</sup>, E. F. Mouresan<sup>2</sup>, J. J. Cañas-Álvarez<sup>3</sup>, J. Altarriba<sup>2</sup>, C. J. Díaz<sup>4</sup>, A. Molina<sup>5</sup>, P. Martínez Cambor<sup>6</sup> and L. Varona<sup>2</sup>

<sup>1</sup> Universidad de Buenos Aires, Argentina, <sup>2</sup> Universidad de Zaragoza, Spain, <sup>3</sup> Universitat Autònoma de Barcelona, Spain, <sup>4</sup> INIA, Spain, <sup>5</sup> Universidad de Córdoba, Spain, <sup>6</sup> Universidad de Oviedo, Spain.

**ABSTRACT:** Autochthonous cattle breeds are a valuable reservoir of genetic diversity in Spain. Along their history, these breeds had been subjected to strong forces of differentiation and common selection. Based on genome-wide SNP data from five different autochthonous breeds, in this study we computed two classical selection signatures statistics in a sliding windows scan along the genome. The statistics chosen were Tajima's D and Fu and Li's D\*, both based on the folded site frequency spectrum. The one-thousandth percentile value of the empirical distribution of each statistic was interpreted as a selection signature. Both regions of common signals among breeds and regions that show breed-specific signals were detected. An example spanning the region that harbors the Myostatin gene confirms the ability of the methods to detect different patterns of DNA variation among breeds in specific areas of the genome.

**Keywords:** breed divergence; neutrality tests; folded spectrum

### Introduction

Autochthonous cattle breeds are considered an important reservoir of genetic and cultural variability in Spain. Although all of these breeds are known to have a *Bos taurus* ancestral origin (Beja-Pereira et al. (2003)), their differentiation started not very long ago when groups of individuals confined to very specific environments became reproductively isolated from each other. Subsequent runs of selection both to enhance adaptation and fit racial patterns produced the phenotypes that nowadays characterize them (see [www.feagas.com](http://www.feagas.com) for a detailed description of these breeds). Several studies have also shown that there are important differences in performance among them (e.g., Gil et al. (2001); Piedrafita et al. (2003)). On the other hand, it is known that most of these breeds have been under similar selection pressures, first towards labor, milk and meat production, and more recently exclusively towards meat production. This dynamic struggle between differentiation and common directional selective pressure have most probably shaped the genome in some very specific ways. The objective of this research was to search for such selection signatures in a sample of individuals from five of the most representative autochthonous Spanish beef cattle breeds using genome-wide SNP data. To do so, classical statistics based on the folded site frequency spectrum, namely Tajima's D (Tajima (1989)) and Fu and Li's D\* (Fu and Li (1993)) statistics, were computed in a sliding windows scan across the genome.

### Materials and Methods

**Data.** Twenty five unrelated families (sire-dam-offspring triplets) were sampled from five different autochthonous Spanish cattle breeds, and genotyped with the Illumina BovineHD 770K BeadChip (Illumina Inc., USA). The sampled breeds were: 1. Asturiana de los Valles (AV); 2. Avileña – Negra Ibérica (ANI); 3. Bruna dels Pirineus (BP); 4. Pirenaica (Pi); and 5. Retinta (Re). After standard procedures of SNP data edition, the Beagle software (Browning and Browning (2009)) was executed for parental haplotype reconstruction. Only parental data were used, since the site frequency spectrum statistics assume the DNA sequences are independent. Precisely, 100 DNA sequences (50 parents × 2 sequences per parent) were analyzed for AV and BP breeds, 96 sequences for ANI and Pi breeds (48 parents × 2 sequences per parent), and 86 for Re breed (43 parents × 2 sequences per parent). Neither sexual chromosomes nor mitochondrial DNA data was included in the analyses.

**Selection signatures tests.** Selection signatures are patterns of variation in DNA sequences attributed to the effect of natural or artificial selection. In this context, the so-called 'selection signatures methods' comprise a large and diverse collection of statistical tests on the hypothesis that the observed variation could be explained solely by mutation and drift processes under the Kimura's (1968) neutral evolution theory framework (see, for example, Oleksyk et al. (2010) for a classification attempt). Rejection of the hypothesis is usually interpreted as a selection footprint, although in certain circumstances demographic processes may also play an important role (e.g. Durrett (2008), chapter 2.3).

One important group of selection signatures methods are those based in the site frequency spectrum (in short, the 'spectrum'). In the context of the Kimura's (1969) infinite site model, under which it is assumed that no mutation can hit a specific site more than once, the site frequency spectrum is the distribution of the frequency of the derived (or mutated) allele in sample of size  $N$ . In other words, is the distribution of how many derived alleles appears once, twice, ..., up to  $N - 1$  times in the sample. This definition applies to the so-called 'unfolded' spectrum to distinguish the case when the ancestral allele is unknown. In such case, the spectrum is defined in terms of the least frequent allele in each site and it is termed the 'folded' spectrum (e.g. Achaz (2009)).

All the statistical tests based on the spectrum share a common structure: they are all computed as the standardized difference of two unbiased estimators of the scaled mutation rate ( $\theta$ ) under the neutral model (Ferretti et al. (2010)). The key point is that the spectrum probability

distribution is a function of this single unknown parameter and thus, a virtually infinite set of unbiased statistics for  $\theta$  could be constructed as a linear combination of the observed spectrum (Achaz (2009); Ferretti et al. (2010)). Specific selection and demography processes in a certain time span will have an effect on different features of the spectrum with respect to the neutral model. Thus, different tests may capture different signals in the data.

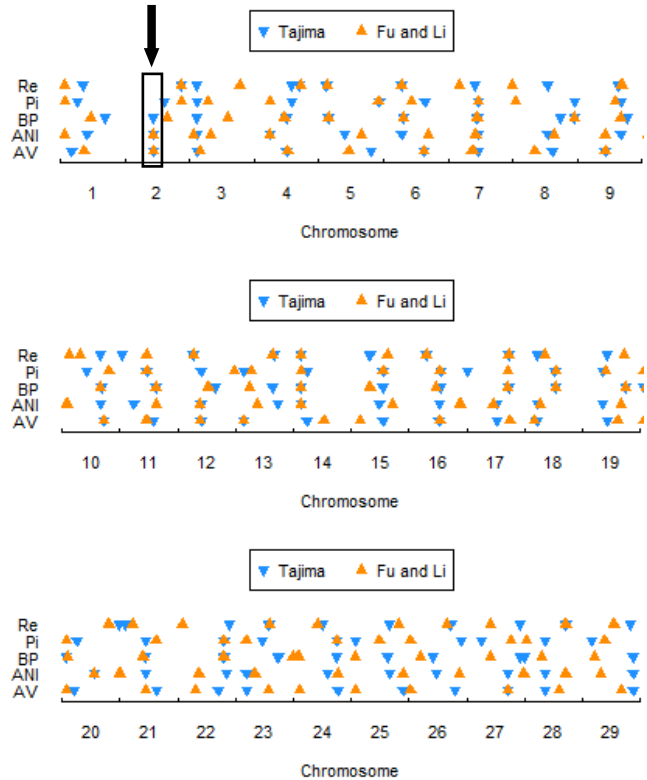
In this study we calculated two classical statistics based on the folded site frequency spectrum: Tajima's  $D$  statistic (Tajima (1989)) and Fu and Li's  $D^*$  statistic (Fu and Li (1993)). Tajima's  $D$  is computed as the difference between the estimate of  $\theta$  based on the average number of pairwise differences and the one based on the number of segregating sites. The expected value of both estimators under the neutral evolution model is zero. However, if low-frequency alleles are segregating in a certain region, either because a selective sweep is taking place or because it has recently finished, the statistic will take an extreme negative value (Tajima (1989)). In turn, Fu and Li's statistic is based on counting the number of singletons; i.e., the sites where a unique derived allele is observed. The rationale is that selection processes will extend time to coalescence so that a greater number of mutations may take place in external leaves of the tree and thus appear only once in the observed sample (Fu and Li (1993)).

**Implementation.** For each of the five breeds and the 29 autosomal chromosomes evaluated the two statistics were computed in a 100-SNP length sliding window scan across the genome. The length of the windows was chosen after an exploratory analysis focused on minimizing the number of SNPs included while retaining neat signals. To compute the statistics, specific Fortran codes were written based on the original formulation of Tajima (1989) and Fu and Li (1993). As SNP-chip data is subjected to strong ascertainment bias (Nielsen et al (2004)), the significance thresholds proposed by those articles are meaningless to reject the neutral hypothesis in our context. To overcome this problem, we used the percentiles of the empirical distribution of each statistic as a threshold for a selection signature. The advantages and disadvantages of this approach are later discussed.

## Results and Discussion

**Genome-wide scan.** Figure 1 summarizes the results of the scan across the entire bovine genome for each of the five breeds under study. The negative extreme 0.01% of the sampled values of the two statistics were plotted against its position in the genome. An overlook shows two distinct, salient features. Firstly, there are both regions of common signaling among breeds and regions that show breed-specific signals, as expected. Secondly, the values of the two statistics overlap at some regions but differ at others. This latter observation suggests that the two statistics are capturing different signals of the data. It is important to emphasize that the scale of representation does not allow a better assessment in specific regions. However, a deeper insight is achieved when these regions are spanned, as it is next shown by exploring the region in BTA2 where the Myostatin gene is located as an example.

**Myostatin gene signal.** The Myostatin gene, responsible for the bovine muscular hypertrophy syndrome or double muscle phenotype (McPherron and Lee (1997)), is located in BTA2. It is known that polymorphisms at this locus are segregating in the AV breed, whereas in the other breeds they are scarce or not present at all (Dunner et al. (2003)). Figure 2 spans the region where the Myostatin gene is located, and shows the values of the Tajima and Fu and Li's statistics within this region for the AV (where the gene is segregating) and Re (where it is not) breeds. Notice that a strong signal is observed in the specific region where the gene is located compared to the adjacent regions for AV but not for Re. In this particular case, the signal of Fu and Li's statistic is stronger.

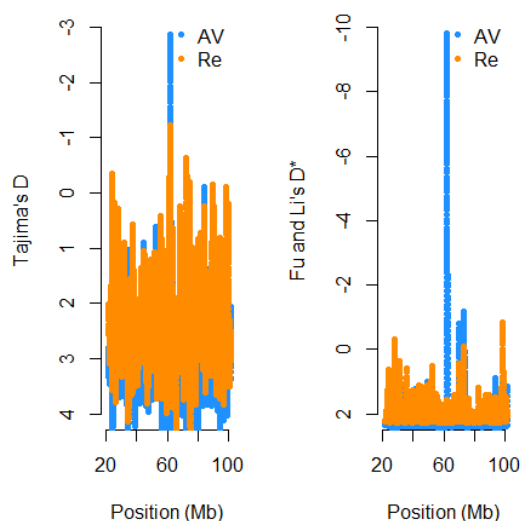


**Figure 1: Selection signatures along the genome for the five breeds analyzed based on site frequency spectrum classical statistics\*.**

\* Each symbol represents a one-thousandth percentile value of the empirical distribution of the corresponding statistic. Notice that an empirical distribution was built for each chromosome and breed based on the values computed for each window. The black box and the black arrow mark the region spanned in Figure 2.

**General discussion.** Two major issues should be raised on the interpretation of the results given by this analysis. First, it is important to understand that both Tajima and Fu and Li's statistics were originally designed to test neutrality by comparing sequences of DNA that have diverged in a time span greater than the one in which most of the *Bos taurus* cattle breeds have. As a consequence, the results should be taken with caution, although the signals are clearly indicating some active differentiation process in some specific DNA regions. Secondly, there is the sensitive issue of the ascertainment bias that afflicts our DNA data. To illustrate this, the average value for the Tajima and the

Fu and Li's statistics across breeds and chromosomes was 2.70 and 2.09, respectively, very far from a value close to 0 that would be expected on average under neutrality. This result is due to the distorted spectrum created by the ascertainment process of SNPs, where only intermediate frequency polymorphisms are retained (Nielsen et al (2004)). To overcome this problem, a possible solution is to construct significance thresholds by simulation, mimicking the scheme of SNP ascertainment in the data generation process (e.g. Voight et al. (2006)). Unfortunately, there is not a straightforward way of doing so. In this study we used instead the empirical distribution of the values of the statistics as a proxy to the appropriate thresholds. Of course, this approach will always "find" a signature within each unit of analysis (i.e., breed and chromosome), thus augmenting the possibility of having misleading signals. On the other hand, the availability of data from several breeds reinforce our approach, as it is extremely unlikely that random false positives occur simultaneously in several populations.



**Figure 2: Tajima's D and Fu and Li's D\* statistics computed for 100-SNP length windows spanned over the region that harbors the Myostatin gene in BTA2 chromosome.**

## Conclusion

Classical site frequency spectrum statistics showed ability to detect common and divergent patterns of DNA sequence variation among breeds in specific areas of the genome. However, how these statistics behave when the populations have recently diverged is still an open question. Additionally, the development of new and better ways to define significance thresholds for samples subject to strong ascertainment bias will enhance the potential to extract useful information from the data.

## Acknowledgments

This research was supported by the Spanish AGL2010-15903 and the UE FP7-289592-GENE2FARM grants. A. González-Rodríguez acknowledges the financial support given by the fellowship BES-2011-045434 of the Spanish government. J. J. Cañas-Álvarez acknowledges the financial support given by COLCIENCIAS through the Francisco José de Caldas fellowship 497/2009.

## Literature Cited

- Achaz, G. (2009). *Genetics* 183:249-258.
- Beja-Pereira, A., Alexandrino, P., Bessa, I. et al. (2003). *J. Hered.* 94:243-250.
- Browning, B. L. and Browning, S. R. (2009). *Am. J. Hum. Genet.* 84:210-223.
- Dunner, S., Miranda, M. E., Amigues, Y. (2003). *Genet. Sel. Evol.* 35:103-118.
- Durrett, R. (2008). *Probability models for DNA sequence evolution*. Springer, New York, USA.
- Ferretti, L., Perez-Enciso, M., and Ramos-Onsins, S. (2010). *Genetics* 186:353-365.
- Fu, Y. X. and Li, W. H. (1993). *Genetics* 133:693-709.
- Gil, M., Serra, X., Gispert, M. et al. (2001). *Meat Sci.* 58:181-188.
- Kimura, M. (1968). *Nature* 217:624-626.
- Kimura, M. (1969). *Genetics* 61:893-903.
- McPherron, A. C. and Lee, S. J. (1997). *Proc. Natl. Acad. Sci.* 94:12457-12461.
- Nielsen, R., Hubisz, M. J., and Clark, A. G. (2004). *Genetics* 168:2373-2382.
- Oleksyk, T. K., Smith, M. W., and O'Brien, S. J. (2010). *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 365:185-205.
- Piedrafitra, J., Quintanilla, R., Sanudo, C. et al. (2003). *Livest. Prod. Sci.* 82:1-13.
- Tajima, F. (1989). *Genetics* 123:585-595.
- Voight, B. F., Kudravalli, S., Wen, X. et al. (2006). *PLoS Biol.* 4(3): e72.