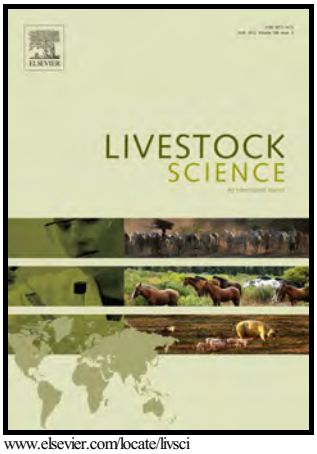
On the haplotype diversity along the genome in Spanish Beef Cattle populations

E.F. Mouresan, A. González-Rodríguez, J.J. Cañas-Álvarez, C. Díaz, J. Altarriba, J.A. Baro, J. Piedrafita, A. Molina, M.A. Toro, L. Varona



www.elsevier.com/locate/livsci

PII: S1871-1413(17)30130-0
DOI: http://dx.doi.org/10.1016/j.livsci.2017.04.015
Reference: LIVSCI3208

To appear in: *Livestock Science*

Received date: 11 November 2016
Revised date: 21 April 2017
Accepted date: 29 April 2017

Cite this article as: E.F. Mouresan, A. González-Rodríguez, J.J. Cañas-Álvarez C. Díaz, J. Altarriba, J.A. Baro, J. Piedrafita, A. Molina, M.A. Toro and L Varona, On the haplotype diversity along the genome in Spanish Beef Cattl populations, *Livestock Science,* http://dx.doi.org/10.1016/j.livsci.2017.04.015

On the haplotype diversity along the genome in Spanish Beef Cattle populations

E. F. Mouresan[1], A. González-Rodríguez[1], J.J. Cañas-Álvarez[2], C. Díaz[3], J. Altarriba[1,4], J. A. Baro[5], J. Piedrafita[2], A. Molina[6], M. A. Toro[7], L. Varona[1,4*]

[1]Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza, 50013-Zaragoza, Spain

[2]Grup de Recerca en Remugants, Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

[3]Departamento de Mejora Genética Animal, INIA, 28040. Madrid, Spain

[4]Instituto Agroalimentario de Aragón (IA2). 50013. Zaragoza, Spain

[5]Departamento de Ciencias Agroforestales, Universidad de Valladolid, 34004.Palencia, Spain

[6]MERAGEM, Universidad de Córdoba, 14071 Córdoba, Spain

[7]Departamento de Producción Animal. Universidad Politécnica de Madrid. 28040. Madrid, Spain

[*]Corresponding author. Tel.: +34 876 554209; fax: +34 976 761612. lvarona@unizar.es

Summary

This study analysed the haplotype diversity along the genome of seven Spanish Beef Cattle populations within regions of 500 kb using the information provided by the *BovineHD Beadchip.* The results of the analysis pointed out a strong variability of the haplotype diversity across the genome, which is greatly conserved across populations. This strong concordance between populations

suggests that the reasons behind it are intrinsic to the structure of the bovine genome and caused probably by the mutation or recombination rate. Nevertheless, some of the genomic regions with very large haplotype diversity are also due of genome assembly errors.

Keywords

Haplotype Diversity; Beef Cattle; Recombination; Mutation; Genome assembly errors

Introduction

The advent of massive genotyping technology has allowed the use of genomic information for genome-wide association studies –GWAS- (Bush and Moore, 2012) and for prediction of breeding values in Genomic Selection –GS- (Meuwissen et al., 2001). Both procedures make use of the linkage disequilibrium (LD) between causative mutations and neutral SNP markers. However, there is evidence that the structure of linkage disequilibrium is not homogeneous along the genome (Ardlie et al., 2002). In fact, the genome can be parsed into haplotype blocks of variable length, as described in human (Gabriel et al., 2002) and cattle (Mokry et al., 2014), caused by variability in the recombination rate across the genome (Myers et al., 2005). In general, the recombination rate is higher in the telomere regions of the chromosomes and lower near the centromere (Coop and Przeworski, 2007), but there is strong evidence of the presence of well-defined regions with a higher rate of recombination, denoted as recombination hotspots (Paigen and Petkov, 2010).

Material and Methods

The data used in this study comprised the *BovineHD Beadchip* genotypes of 171 trios (sire-dam-offspring) of seven beef cattle breeds (*Asturiana de los Valles* – AV-, N=25, *Avileña - Negra Ibérica* – ANI-, N=24, *Bruna dels Pirineus* – BP- N=25, *Morucha* –Mo-, N=25, *Pirenaica* –Pi-, N=24, *Retinta* – Re-, N=24 and *Rubia Gallega* –RG-, N=24). After filtering for mendelian error lower than 0.05 and individual and SNP call rate over 0.95, 707,307 SNP markers were considered.

First, we established the haplotype phases using two alternative software: BEAGLE (Browning and Browning, 2007) using the "TRIO" option, and SHAPEIT v2 (Delaneau et al., 2013). The concordance between phases generated by BEAGLE and SHAPEIT was very high (over 99.9%). Thus, we present exclusively the results provided by BEAGLE. Once the paternal and maternal haplotypes were established, we calculated haplotype diversity as the number of distinct haplotypes (or phases) present within a given genomic region of predefined size, and centered at each SNP. That is, for one SNP located at base pair 3,000,000 of a certain chromosome, the haplotype diversity for a genomic region of 500 kb is calculated by counting how many distinct haplotypes there are in the population for the SNPs located within the genomic region between 2,750,000 and 3,250,000 bp. In fact, we have calculated the haplotype diversity for sliding genomic regions of 100, 250, 500 and 1000 kb centered at each SNP.

When the size of the genomic regions was small (100 kb or 250 kb), the results did not allow to identify clearly the variability in the haplotype diversity, and on the other hand, the results from the analysis of wider genomic regions (1 Mb) provided a very large number of haplotypes (398.84 on average). Thus, we focused the analysis in windows of intermediate size (500 kb).

Results and Discussion

In first place, we analyzed the distribution of the number of SNPs present within genomic regions of 500 kb (Figure 1a). We found that they followed an almost perfect Gaussian distribution with an average of 149.21 (±33.22) SNPs. This indirectly confirms the adequacy of SNP selection when the *Bovine HD Beadchip* was constructed. Moreover, the distribution of the number of haplotypes within these genomic regions had a mean of 253.94 (±69.34) and presented a positive skewness (0.0596) (Figure 1b). This indicates that the haplotype diversity is relatively higher in some regions of the genome. As expected, we found a positive relationship between the number of haplotypes and SNPs present in each specific region of the genome, with a correlation between them of 0.36. However, as it is shown in Figure 1c, the genomic regions with the highest degree of haplotype diversity (larger number of haplotypes) were not those with a largest number of SNPs. It indicates that the presence of a large number of haplotypes is not only a consequence of the overrepresentation of SNP markers.

Further, the results of the haplotype diversity (number of distinct haplotypes) within genomic regions of 500 kb are presented in Figure 2a. The heterogeneity of the haplotype diversity was very large along the bovine chromosomes, but

regions with higher number of haplotypes were more frequent close to the telomeres, as it can be observed in Figure 2b, that presents the average haplotype diversity with respect to the relative physical position within the chromosome. This result was consistent with the pattern of the recombination rate provided by the study of Ma et al. (2015).

Further, we calculated the haplotype diversity for each of the seven analyzed populations separately (Supplementary Figure). We found that the correlations between the number of haplotypes identified within each population were very high and ranged from 0.68 (Pi and AV) to 0.77 (ANI and Mo) (Figure 3), being slightly higher between Re, Mo and AVI, previously found to have a relative higher degree of genetic relatedness (Cañas-Álvarez et al., 2015). Besides, in order to avoid the influence of the SNP density within the genomic regions, we also calculated the haplotype diversity after the correction by the number of SNPs within the genomic region using a linear regression. The correlations between populations were similar and ranged from 0.66 (Pi and AV) to 0.75 (ANI and Mo). These results suggest that the causes for the haplotype diversity are intrinsic to the structure of the bovine genome and related probably with the mutation or recombination rate. Nevertheless, some other causes can be also argued, such as the presence of some kind of natural or artificial selection or the incidence of structural variants in the genome. Both of them have been previously analyzed in these populations (Da Silva et al., 2014; González-Rodríguez et al., 2016) and these studies could not find any consistent pattern between populations, reinforcing the hypothesis that mutation and recombination rates are linked to the observed haplotype diversity. In addition,

the results were coherent with the strong heterogeneity of the recombination rate along the cattle genome found by Ma et al. (2015).

Moreover, some genomic regions presented an extremely high haplotype diversity (Table 1), and they were strongly conserved across populations (Table 1 and Supplementary Figure). The potential causes for this phenomenon can be the presence of genome assembly errors or the existence of very large recombination or mutation rates. In fact, when we compared the 20 genomic regions with highest haplotype diversity with the ones reported as genome assembly errors by Utsunomiya et al. (2016), we found that 13 of them overlapped (Table 1). Thus, they can be attributed to genome assembly errors, whereas the cause of the huge haplotype diversity of the remaining has to be deeply studied.

Despite the cause that generates it, the results presented here confirm a strong heterogeneity of haplotype diversity along the genome. The applications that most frequently use genomic information, like GWAS (Bush and Moore, 2012) or GS (Meuwissen et al., 2001), are based on the existence of linkage disequilibrium between SNP markers and QTLs. The aim is to identify genomic regions associated with the variability of traits or to predict the breeding value of candidates of selection, respectively. Besides, genomic regions with higher haplotype diversity are usually associated with a lower LD. In fact, in the analyzed dataset, there was a correlation of -0.64 between the average LD (r^2) between all markers and the number of distinct haplotypes present within each genomic region. Thus, genes of interest potentially located in these regions could be blurred by the standard procedures. Further research must be done to modify current procedures of GWAS or GS to incorporate the structural

information of the haplotype diversity in each specific region of the genome, such as the use of run of homozygosity (ROH) for genomic prediction (Luan et al., 2014) or GWAS (Biscarini et al., 2014).

Acknowledgements

The authors want to thank the AGL 2010-15903 grant from the Spanish government; the collaboration of Breed societies in collecting samples and the support of FEAGAS is also acknowledged. J. J. Cañas-Álvarez acknowledges the COLCIENCIAS support by the Francisco José de Caldas fellowship 497/2009 and A. González-Rodríguez acknowledges the financial support by the BES-2011-045434 fellowship.

References

Ardlie, K.G., Kruglyak, L., Seielstad, M., 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3, 299–309.

Biscarini, F., Biffani, S., Nicolazzi, E. L., *et al.*, 2014. Applying runs of homozygosity to the detection of associations between genotype and

phenotype in farm animals. In: Proceedings 10th World Congress of Genetics Applied to Livestock Production.

Browning, S.R., Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-Data inference for whole-Genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81, 1084–97.

Bush, W.S., Moore, J.H., 2012. Chapter 11: Genome-wide association studies. PLoS Computational Biology 8, e1002822.

Cañas-Álvarez, J.J., González-Rodríguez, A., Munilla, S., *et al.,* 2015. Genetic diversity and divergence among Spanish beef cattle breeds assessed by a bovine high-density SNP chip. Journal of Animal Science 93, 5164-5174.

Coop, G., Przeworski, M., 2007. An evolutionary view of human recombination. Nature Reviews Genetics 8, 23–34.

Da Silva, T. B. R, González-Rodríguez, A., Avilés, C., *et al.*, 2014. Analysis of copy number variants in Spanish autochthonous beef cattle breeds. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production.

Delaneau, O., Zagury, J.F., Marchini, J., 2013. Improved whole chromosome phasing for disease and population genetic studies. Nature Methods 10, 5-6.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., *et al.,* 2002. The structure of haplotype blocks in the human genome. Science 296, 2225–2229.

González-Rodríguez, A., Munilla, S., Mouresan, E. F., *et al.*, 2016. On the performance of tests for the detection of signatures of selection: a case study with the Spanish autochthonous beef cattle populations. Genetics Selection Evolution 48, 81.

Luan, T., Yu, X., Dolezal, M., *et al.*, 2014. Genomic prediction based on runs of homozygosity. Genetics Selection Evolution 46,64.

Ma, L., O' Connell, J. R., VanRaden, P. M., *et al.*, 2015. Cattle sex-specific recombination and genetic control from a large pedigree analysis. Plos Genetics 11, e1005387.

Meuwissen, T. H. E., Hayes, B. J., Goddard, M. E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Mokry, F., Buzanskas, M., de Alvarenga Mudadu, M., *et al.*, 2014. Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. BMC Genomics 15, S6.

Myers, S., Bottolo, L., Freeman, C., *et al.* 2005. A fine-Scale map of recombination rates and hotspots across the human genome. Science 310, 321–324.

Paigen, K., Petkov, P. 2010, Mammalian recombination hot spots: properties, control and evolution. Nature Reviews Genetics 11, 221–233.

Utsunomiya, A. T. H., Santos, D. J. A., Boison, S. A., *et al.*, 2016. Revealing misassembled segments in the bovine reference genome by high resolution linkage disequilibrium scan. BMC Genomics 17:705.

Figure captions:

Figure 1. Distribution of the number of SNPs (a) and haplotypes (b) for genomic regions of 500 Kb and the relationship between them (c).

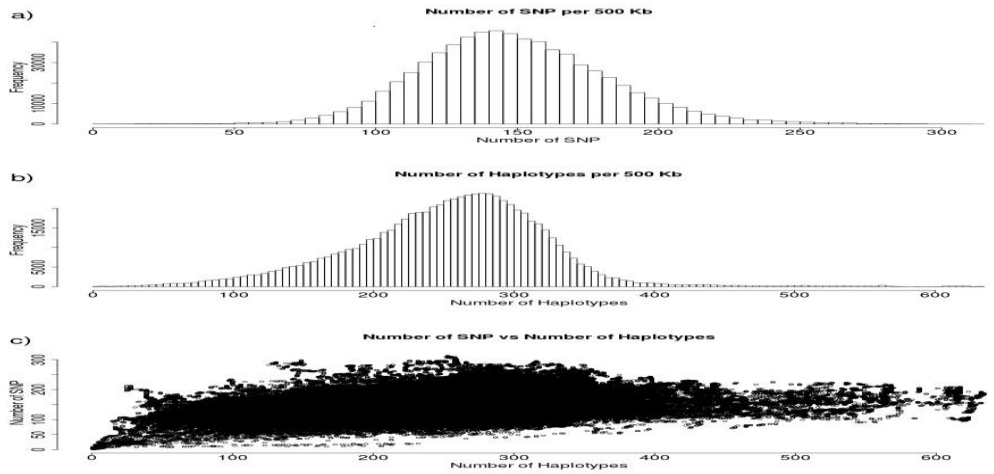


Figure 2. Haplotype diversity (number of distinct haplotypes) along the autosomal genome of seven Spanish autochthonous beef cattle populations for regions of 500 kb (a) and (b).

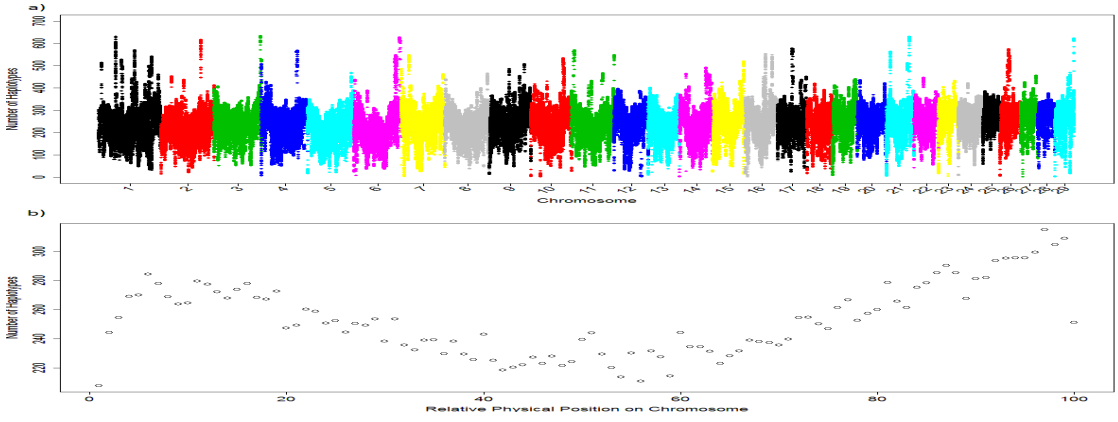


Figure 3. Correlations between the numbers of haplotypes for genomic regions of 500 kb in the seven analyzed populations (AV: Asturiana de los Valles, ANI: Avileña Negra Ibérica, BP: Bruna dels Pirineus, Mo: Morucha, Pi: Pirenaica, Re: Retinta, RG: Rubia Gallega).

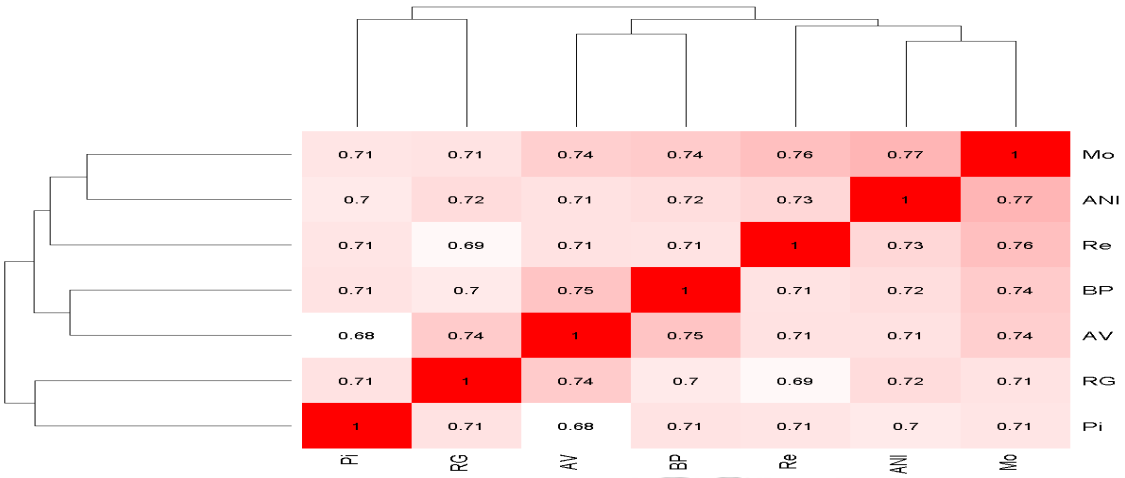


Table 1. Maximum number of haplotypes in the 20 genomic regions with the highest haplotype diversity, number of haplotypes per breed and overlapping with genome assembly errors.

BTA	Start	End	NH	AV	ANI	BP	Mo	Pi	Rt	RG	AE
1	44302055	45126351	632	97	93	97	96	85	78	94	YES
1	94927808	95300082	571	98	79	94	90	74	71	81	YES
1	138718684	138942001	542	91	78	85	84	69	69	78	YES
2	104611028	105209707	617	99	83	94	91	78	86	89	NO
3	119213805	119727574	634	98	90	93	92	85	83	92	NO
4	96579212	96976581	569	94	78	85	79	76	87	82	YES
6	106339203	107452318	549	90	79	88	84	74	73	69	YES

6	110012125	110061838	541	89	79	84	83	76	67	71	YES
6	116772416	117368074	628	100	89	97	95	79	84	84	NO
7	23596068	23768331	549	93	78	83	78	75	77	81	NO
10	87136104	87158386	533	95	76	81	89	70	76	73	YES
11	5540849	5998792	570	93	85	86	84	81	71	86	NO
11	106914516	107024981	549	93	77	77	73	78	80	80	NO
16	55480917	55715596	554	88	78	87	84	80	84	78	YES
16	70788898	70908905	549	90	76	88	85	78	78	83	YES
17	40631529	41138782	577	97	75	91	87	79	80	88	YES
21	13075227	13473716	564	93	80	85	92	80	72	74	YES
21	59626112	59962773	631	98	89	97	94	85	80	92	YES
26	23283996	25665978	575	93	81	84	86	79	80	80	YES
29	50141388	50641690	619	97	85	95	93	77	80	84	NO

BTA: Bos Taurus Chromosome, Start: Start position in bp, End: End position in bp, NH: Maximum number of Haplotypes, AV: Number of haplotypes in Asturiana de los Valles, ANI: Avileña Negra Ibérica, BP: Bruna dels Pirineus, Mo: Morucha, Pi: Pirenaica, Re: Retinta, RG: Rubia Gallega, AE: Identified (YET) or not (NO) as genome assembly errors by Utsunomiya et al, 2016).

Highlights

- Haplotype diversity is very variable along the autosomal genome.
- Haplotype diversity is very well conserved across populations.
- This concordance suggests that the reasons are intrinsic to the genome structure.