Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8672
Doi: 10.4025/actascianimsci.v40i1.39015

ANIMAL BREEDING AND REPRODUCTION

Genome-wide association study using the Bayes C method for important traits in dairy yield in Colombian Holstein Cattle

Juan Carlos Rincón Flórez[1,2*], Albeiro López Herrera[2] and Jose Julián Echeverri Zuluaga[2]

*[1]Medicina Veterinaria y Zootecnia, Universidad Tecnológica de Pereira, Carrera 27 #10-02 Barrio Álamos, Pereira, Risaralda, Colombia. [2]Research Group: Biodiversity and molecular genetics, Universidad Nacional de Colombia, Facultad de Ciencias Agrarias, Medellín, Antioquia, Colombia. *Author for correspondence. E-mail: rincon.juan@utp.edu.co.*

ABSTRACT. The objective of this study was to determine the genome association between markers of bovine LD BeadChip with dairy important traits. Information of breeding program of the Universidad Nacional de Colombia was used. BLUP-EBVs were used for dairy yield (DY), fat percentage (FP), protein percentage (PP) and somatic cell score (SCS). 150 animals were selected for blood or semen DNA extraction and genotyping with BovineLD BeadChip (Illumina). Autosomal information was retained and the editing information was performed using Plink v1.07 software. The effects of SNPs were determined by Bayes C with GS3 software. The minor allele frequency for most of the markers on the bead chip was high, which increases the probability of finding important loci segregating in the population. Estimations of fraction markers with an effect were close to zero in almost all cases. The most important markers were mapped by trait using ENSEMBL. A total of 6,510 autosomal SNPs were retained, out of which only a proportion with effect was taken from the mixed function of Bayes C. Important genes as ANKS1B, CLCN1, NMBR and CTSD, were found for each trait for AL, FP, PP and SCS respectively. Finally, Bayes C estimation allowed to identify specific SNPs possibly associated with QTLs.

Keywords: dairy cattle; GWAS; QTLs, single nucleotide polymorphism.

Associação genômica usando o método Bayes C para características importantes na produção leiteira em Gado Holandês na Colômbia.

RESUMO. O Objetivo deste estudo foi determinar a associação genômica entre os marcadores do *chip* bovino LD (*LD BeadChip*) com características importantes na produção de leite. Foram utilizadas informações do programa de melhoramento genético da Universidade Nacional de Colômbia. BLUP-EBVs foram utilizados para a produção de leite (DY), porcentagem de gordura (FP), porcentagem de proteína (PP) e escore de células somáticas (SCS). 150 animais foram selecionados para extração de DNA do sangue ou sêmen e genotipados com o *chip* BovineLD (Illumina). A informação autossômica foi mantida e a edição da informação foi executada usando o programa Plink v1.07. Os efeitos dos SNPs foram determinados por Bayes C com o programa GS3. A frequência do alelo menor para a maioria dos marcadores no *chip* foi alta, o que aumenta a probabilidade de encontrar *locos* importantes segregando na população. As estimativas da fração de marcadores, com efeito, foram próximas de zero em quase todas as situações. Os marcadores mais importantes foram mapeados com ENSEMBL. Um total de 6510 SNPs autossômicos foram preservados, dos quais apenas uma proporção foi tomada com efeito a partir da função mista de Bayes C. Para cada característica foram encontrados genes importantes, como ANKS1B, CLCN1, NMBR e CTSD, para AL, FP, PP e SCS, respectivamente. Finalmente, a estimativa de Bayes-C permitiu a identificação de SNPs com possível associação com QTLs.

Palavras-chave: gado leiteiro; GWAS, QTLs; Polimorfismo de nucleotídeo único.

Introduction

The identification of millions of Single Nucleotide Polymorphisms (SNPs) in the bovine genome (Daetwyler et al.; 2014; Gibbs et al., 2009), along with the gradual reduction in genotyping and resequencing cost (Meuwissen & Goddard, 2010) have generated a real opportunity to use information from thousands of molecular markers for implementing the genomic selection. These advances have allowed association studies on a large scale, with the aim of strengthen breeding programs and improve understanding of the genetic variation of important traits in dairy yield (Daetwyler et al., 2014).

Genomic selection has been implemented mainly with high density bead chips in different dairy cattle breeds around the world (VanRaden et al., 2009). This uses information from a large number of DNA

markers to estimate individuals breeding values based on Linkage Disequilibrium (LD) between a specific marker and the Quantitative Trait Locus (QTL) (Meuwissen, Hayes, & Goddard, 2001).

Works based on Genome-Wide Association (GWAS), intend to identify markers, genomic regions, or causative mutations associated with productive traits, in order to improve the accuracy of estimated breeding values and the understanding of physiological processes and genetic architecture of dairy yield traits (Makowsky et al., 2011; Zhang et al., 2014).

GWAS studies have used estimation methodologies based on least squares or Restricted Maximum Likelihood (REML) repeatedly, with different settings for inferring the significance of the SNP effects and map specific QTLs, to reduce the problem of false positives rate and overestimation effect. To alleviate some of these problems, different approaches have been raised, including Bayesians that can reduce the problem (Peters et al., 2012) and can be used for Genome-Wide Association studies analysis as it is the case of the so-called Bayes C, and Bayes Cπ implemented for GWAS analysis (Legarra et al., 2015).

GWAS analyses typically select a small number of DNA markers, usually SNPs, which are closely linked with functional polymorphisms associated with quantitative traits of economic importance in the domestic species. The markers identified are subsequently subjected to post-GWAS tests with fine mapping techniques, in order to validate causal mutations with specific traits of interest (Yi, Breheny, Imam, Liu, & Hoeschele, 2015).

Holstein cattle has been selected for decades in many places around the world under different selection criteria and in accordance with production and market conditions of each country involved. Colombian Holstein cattle are the most used on specialized dairy farms and it is located in the high tropic under conditions different from those of other countries. Several GWAS in dairy yield have been reported in different countries (Zhang et al., 2014), but in tropical conditions there are few reports that allow to identify important regions for genetic improvement.

Given the above, the objective of this work was to contribute to the understanding of the genetic variance explained by multiple SNP of the BovineLD Bead Chip (Illumina, San Diego CA), using GWAS with Bayes Cπ approach, in order to identify also polymorphisms associated with Dairy Yield (DY, $h^2 = 0.16$), Fat Percentage (FP, $h^2 = 0.32$), Protein Percentage (PP,$h^2 = 0.30$), and Somatic Cell Score (SCS, $h^2 = 0.01$) which genetic parameters were previously estimated (Rincón, Zambrano, & Echeverri, 2015).

Material and methods

Study population

This work was performed with the information collected in dairy herds enrolled in the program of genetic evaluation and dairy control of the Universidad Nacional de Colombia at Medellín, and Colanta Cooperativa Ltda. The specific management conditions, food and health were variable in all herds, as well as its topography and geographical location.

Breeding values of 150 bovines, choosing as many bulls with daughters in different dairy herds in Antioquia, were taken. The animals were in areas of lower montane rainforest, with an average temperature of 14°C and at an altitude between 1800 and 2500 Meters Above Sea Level (MASL). The evaluated traits were Dairy Yield per lactation (DY), Fat Percentage (FP), Protein Percentage (PP), and Somatic Cell Score (SCS). Genetic values were used for each aspect of the 150 individuals (37 bulls and 113 cows), from which a blood sample in the case of cows or semen in sires were taken for DNA extraction and subsequent genotyping. Parents were taken given the lowest possible relationship between groups, but in some cases, there were a few couples and triplets (father-mother-daughter). The animal selection took into account the choice of bulls with as many daughters as possible in the population (9 national and 28 foreigners) and the sperm available in the market, because these can be very informative, possessing a number of major population haplotypes.

DNA extraction and genotyping

Blood samples were taken from the middle coccygeal vein using 5mL BD vacutainer tubes, with 18 needles and ethylene diamine tetra acetic acid (EDTA) as anticoagulant (BD Vacutainer TM). Once the samples were taken, they were stored at 4°C until processing. DNA extraction of some sires was made from semen straws of 250 and 400μL.

For blood and semen DNA extraction *DNeasy Blood & Tissue Kit®* and *QIAamp® DNA Mini Kit* were used respectively according to the manufacturer recommendations. DNA samples were analyzed in a NanoDrop (Qiagen, USA) to determine its concentration and to adjust them to 50 μg/μL. Purity was also determined by the absorbance ratio A260/A280 and finally, DNA integrity was

determined by electrophoresis in agarose gel at 0.8% (Amresco®). DNA samples were stored at 4°C until genotyping.

A total of 150 animals were genotyped with BovineLD Bead Chip (Illumina, San Diego CA) which covers a panel of 6,909 SNPs, in KosGenetic laboratory of the University of Milan (Italy). It was determined that the test of missing data was less than 0.1% and SNPs with Minor Allele Frequency (MAF) below 0.03 were discarded. Genotypes with Mendelian errors greater than 0.05 were also declared as missing data and only the SNPs present in autosomal chromosomes were used. R software (R Development Core Team, 2012) and plink v1.07 (Purcell, et al., 2007) were used for data editing. The genotypes were coded as 0, 1, or 2 according to the number of alternative alleles present. Missing data were imputed by Beagle software (Browning & Browning, 2009).

Statistical analysis

Traditional genetic values

Estimated Breeding Values (EBV) were taken from the predicted values using the Best Linear Unbiased Predictor (BLUP) in the breeding program of the Universidad Nacional de Colombia at Medellin, and Colanta Cooperativa Ltda.

SNPs estimated effects

The general statistical model used was:

$$y = \mu + u + \sum_{i=1}^{I} Z_i\, a_i + e$$

where y is the vector of genetic values for each evaluated trait, μ is the overall average, u is the vector of polygenic effects of individuals in the pedigree, i is SNPs number, Z_i corresponds to the vector of genotypes for the i – th SNP, a_i is the additive effect of each SNP, and e is the vector of residual effects.

In this study the Bayesian regression method called Bayes Cπ (Habier, Fernando, Kizilkaya, & Garrick, 2011) was used, where *a priori* constant for μ is assumed, and a distribution $u|A\sigma^2_u$ approximately N (0, $A\sigma^2_u$); where A is the matrix of relationships between individuals and σ_u^2 is the additive genetic variance not explained by SNPs. The distribution *a priori* for a_i was a mixed distribution dependent on the variance $\sigma^2_{a_i}$ and the probability π of having SNPs with effect. It is important to note that (Legarra et al., 2016) define π contrary to what (Meuwissen et al., 2001).

$$a_i|\pi,\sigma^2_{a_i} = \begin{cases} 0, con\ probalilidad\ (1-\pi) \\ \sim N\left(0,\sigma^2_{a_i}\right), con\ probabilidad\ \pi \end{cases}$$

In Bayes Cπ it is assumed that all the effects of SNPs have a common variance, distributed as an inverted chi-squared escalated with parameters ν_a and S_a^2 taken as in Bayes B (Meuwissen et al., 2001). For π a priori distribution U(0, 1) was assumed, considering the convergence difficulties when π is estimated simultaneously (Van den Berg, Fritz, & Boichard, 2013), an approach in which parameters were estimated first was prepared, and in cases where convergence was not achieved the proportion of SNPs with effect was fixed at 1% in order to estimate later the effect solutions of the SNPs included. In all cases the same approach was conducted, but in a way that all markers will be used ($\pi = 1$).

The determination of markers association was performed directly on Estimated Breeding Values (EBV) of the general population, because this is a direct additive genetic effect. To this end, the GS3 software that allows using the approach of Markov chain Monte Carlo (MCMC) to estimate the effect of each SNP among all was used. The procedure had 100000 iterations with a period of heating of 20000 and with corrections every 1000, according to the recommended minimum to achieve convergence (Legarra et al., 2016). The convergence diagnosis was verified visually by R software (R Development Core Team, 2012).

The effects of molecular markers were plotted by R Development Core Team, (2012) software according to their location in the genome. Additionally, the distribution was presented *a posteriori* for π and for the variance percentage explained by markers.

Defining localization map

The localization map for the most important markers was performed based on the assembly UMD 3.1 *(Bos taurus)* of NCBI and ENSEMBL, using Variant Effect Predictor (VEP) tool (McLaren et al., 2010). Clusters of genes were performed according to ENSEMBL based on construction UMD 3.1 of Variant effect predictor (VEP) (McLaren et al., 2010). Later, ontology was used (Gene Ontology Consortium) and previous reports of analyzed QTLs were searched, according to Animal QTLdb public database in section cattle QTLdb (Hu, Park, Wu, & Reecy, 2012), looking in windows of maximum 1.5Mb.

Results

Used breeding values show a significant variation between genetic values of sampled individuals. It is important to note that both positive and negative individuals were taken into account for genetic merit of the traits. Once filtered the information, it went from 6909 of SNPs to a total of 6716 SNPs that met set out criteria in editing. SNPs were found distributed in all chromosomes in the way presented in Table 1, out of which 6510 were autosomal and were used in subsequent analyzes.

Table 1. Description of SNPs number present in the bovineLD Bead Chip for each chromosome.

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of SNPs	390	341	310	301	305	306	282	292	268	269	275	224	209	18	219
Chromosome	16	17	18	19	20	21	22	23	24	25	26	27	28	29	X
Number of SNPs	205	190	176	180	205	184	163	150	176	134	142	138	124	134	206

Out of all evaluated SNPs, 6899 were processed by VEP, 6880 of which were recognized as existing variants and 19 as new or unreported. In total, 2251 SNPs of the BeadChip overlapping genes and 1303 with transcripts were found. Most of the mutations found in 6k Illumina BeadChip correspond to SNPs located in the intergenic spaces (57%), intronic regions (29%), or are synonymous mutations (1%). Only a small portion are in untranslated regions (UTRs) and missense mutations (1%) or upstream (5%) or downstream variants of a gene (5%), so that they can be biologically suggested as causative mutations. However, it has been shown that some intronic and/or non-coding variants often have some sort of relationship with specific QTLs, either because they have an unknown function or because they are associated with regions that affect expression (promoters, enhancers, among others) and messengers maturation, or because they are in LD with the causative mutation.

In determining the allele frequencies of different markers in the population, it was possible to observe a trend toward alleles of higher Minor Allele Frequency (MAF), which shows a clear trend towards intermediate and polymorphic markers selection in genotyping BeadChips, so that they segregate in different populations, generating a bias for assessment in population genetics. Only 11 markers out of all presented frequencies between 0.03 and 0.05; while 83 had MAF under 0.1. However, the markers with MAF higher values provide greater statistical power and in formativeness for association with phenotypic traits and may be a better alternative to use them in genomic selection programs, taking into account that there is a greater chance that they segregate in the population.

Considering that less frequently found variants (MAF < 0.05) correspond to rare variants in the population, they were selected to identify their location. Most of them were found in intergenic regions on different chromosomes and 3 of them within CNIH3, KCNIP1 and PPP1R13B genes, but in intronic regions.

Moreover, no evidence ($p > 0.05$) was found to claim that even one out of the 6510 markers remaining after editing was deviated from Hardy Weinberg equilibrium (HWE). However, some markers showed important differences between the observed and expected heterozygosity.

Subsequently, components of variance and *posterior* probability for π in each case were estimated using the MCMC algorithm of GS3 (Legarra et al., 2016). It is important to note that the π parameter has an opposite meaning to that defined by Meuwissen et al. (2001), π it is the fraction of SNPs that have an effect. The trait with the highest value of π was fat percentage, and the lowest was somatic cell score, however, by plotting the distribution *a posteriori*, it was not possible to differentiate in any case a strong peak in the estimation and generally a strong trend toward values near to zero was observed (Figure 1A).

According to the results, π values were set in 0.01 for all traits, since the results were very close to zero, but without a well – defined peak. Once π was set, the variance percentage explained by the markers was determined, the feature with higher value of the explained variance was SCS and the lowest was PP (Figure 1B).

In the case of fat percentage in milk, SNPs accounted for approximately 0.3% out of the genetic variance, 18% for DY, 1.8% for PP, and 97% for SCS (Figure 1B), which is a bit contrasting and evidences an effect of different genetic architectures among the evaluated traits.

The inclusion of all markers in the evaluation, allowed to estimate the variance effect again, showing that the use of all SNPs in the analysis causes a slight decrease in variance proportion explained by markers. Thus, FP had approximately 0.28% out of the variance explained by all markers, DY 0.39%, PP 1.4%, and SCS 78%.

It was possible to find SNPs with a greater effect for each trait when π was fixed in 0.01 (Figure 2B). However, some of the markers were close to zero and a much smaller proportion than π showed a significant peak in its magnitude. Table 2 presents a

description of the five most important markers by evaluated trait, including the gene to which it is related and its function. Among these the most important marker was rs110718748 because of its effect and its *posterior* probability for DY, it was found within an intron in ANKS1B gene, according to the version *Bos taurus* UMD 3.1 of NCBI. According to ontology, this gene is involved in various routes according to its isoform, with a significant function in overall protein synthesis (Table 2).

Furthermore, the most important marker for FP was rs109245784, this marker represents an intronic variation in the CLCN1 gene according to version NCBI UMD 3.1. This gene has a direct activity on chloride channels, according its ontology (Table 2). The most important marker for PP was rs29014693, which was found near the NMBR gene that plays an important role in various biological functions, including sensory activity, diet, gastric and pancreatic secretion, among others (Table 2). Finally, for SCS the marker rs109548201 had the biggest effect, it was found inside CTSD gene, and presents a clear immunological effect in animals (Table 2). Out of all markers tested, none was reported as a mutation with consequences on the loss of meaning or not synonymous.

The minor allele frequencies in the most important SNPs on different traits were between 0.19 and 0.49, evidencing greater importance of SNPs with intermediate frequencies (Table 2 in all cases, presumably because of their informativeness and importance on the genetic variance. In order to compare if there was a similar pattern in the solutions when a different value of π is used, the solutions graph was performed considering the inclusion of all markers on the BeadChip (π=1) (Figure 2B).

The figure shows that the higher values in solutions do not match for different traits; however, some individual solutions are important in both cases.

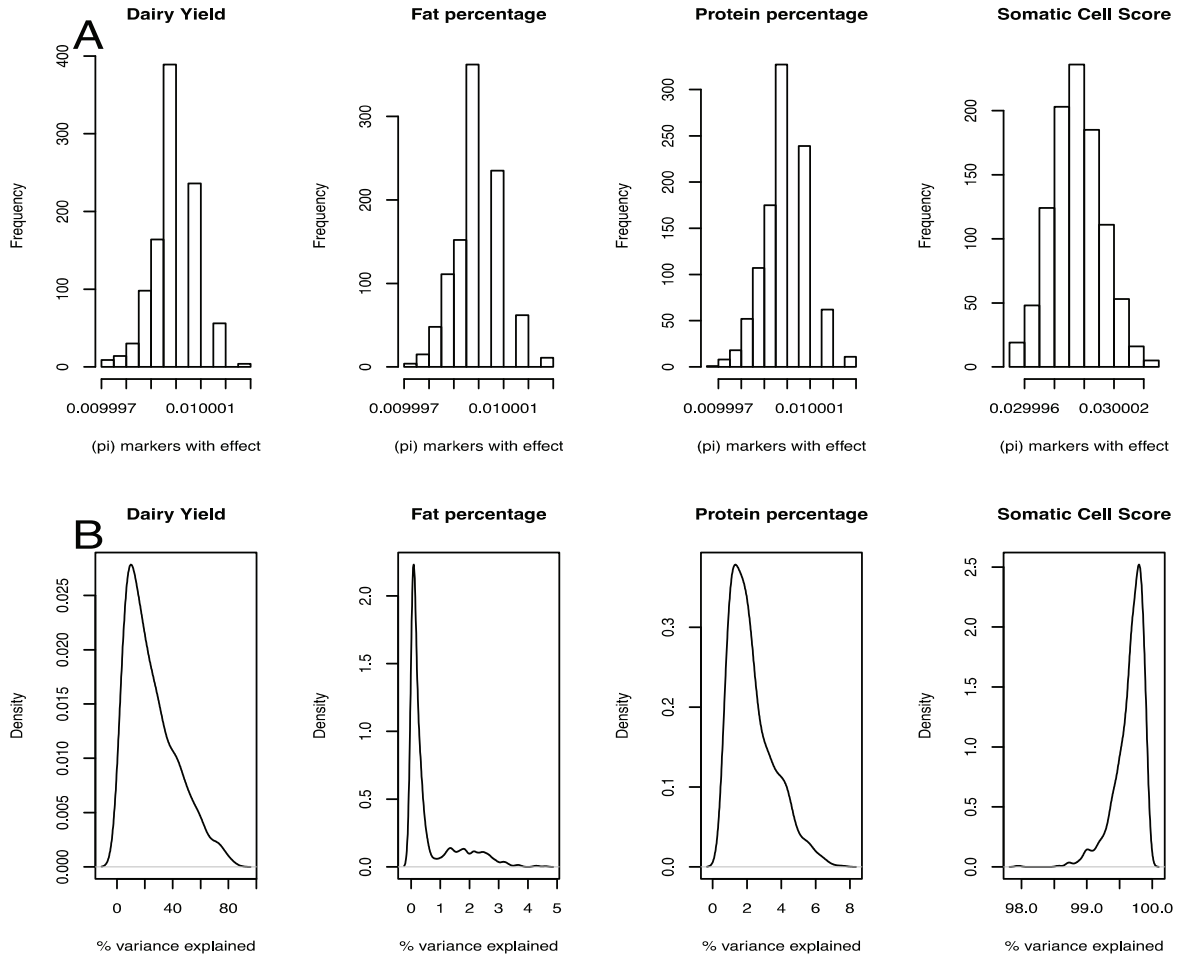

Figure 1. Distribution *a posteriori* with Bayes Cπ (A) Distribution *a posteriori* of markers with effect proportion (π) (B) Distribution *a posteriori* of variance percentage explained by molecular markers.

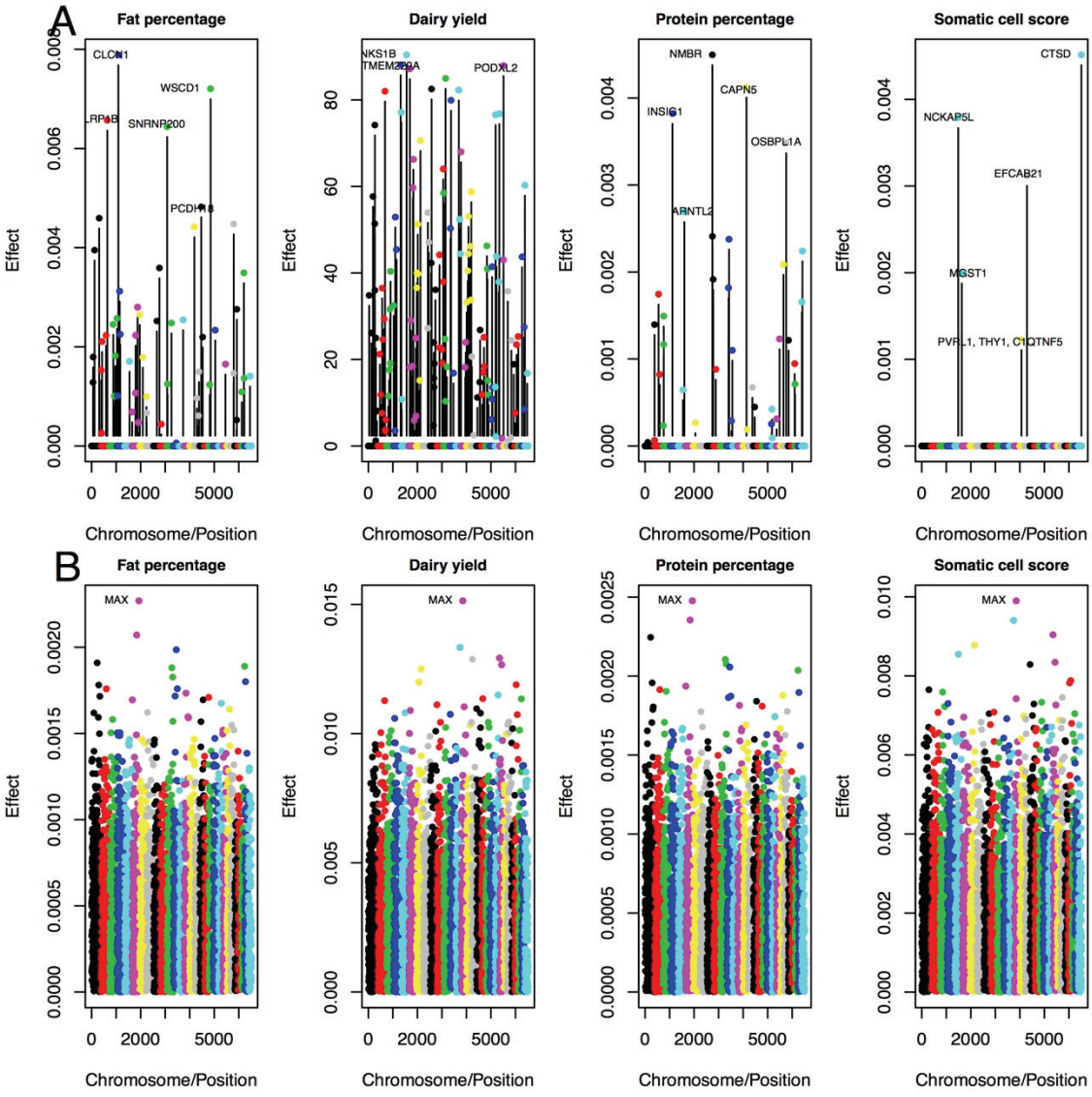

Figure 2. Effects of SNPs (A) with $\pi = 0.01$ and (B) with $\pi = 1$, on fat percentage, dairy yield, protein percentage, and somatic cell score in milk.

For example, in the BTA-6 for fat percentage (Figure 2B), it can be observed a sharp peak which is much lower when $\pi = 0.01$ (Figure 2A) but corresponding to the same marker at the top. Moreover, it is interesting to note that by including all markers, a peak on chromosome 14 is evidenced for fat percentage, which although is not the largest, is important because it corresponds to DGAT1 gene, which has been proposed repeatedly as a major gene for fat content in milk (Grisart et al., 2002; Wang et al., 2012). However, this peak is not observed when $\pi = 0.01$, possibly because there are some markers with greater effect. Finally, it should be noted that the estimation with a markers fraction allowed to explain a greater proportion of the variance in some cases and was therefore used in this work.

Discussion

The use of breeding values as response variable for estimating the effects of the markers, can include only additive genetic effects and isolate additional effects of more complex models, such that computational requirements can be reduced, looking to promote convergence and decrease computing time, especially in complex models that require joint estimation of several different parameters. Some works have directly used genetic merit to estimate additive effects of genetic variants (Calus, De Haas, & Veerkamp, 2013; Van Hulzen et al., 2012).

Table 2. Description of polymorphisms with greater effect on Dairy Yield per lactation (DY), milk Fat Percentage (FP), Protein Percentage (PP), and Somatic Cell Score (SCS).

Trait	rs code	Chromosome / position	MAF	Gene/consequence	Function
DY	Rs110718748	5/63899453	0.40	ANKS1B/intronic variant	There are different isoforms of the gene with neuronal regulation functions, regulation of global protein synthesis, APP regulated.
	Rs41607880	4/89380482	0.49	TMEM229A/close intergenic variant	Activity with binding transcription factors to specific DNA sequence
	Rs110425841	22/60508872	0.14	PODXL2/Intronic variant and upstream of ABTB1	PODXL2: transmembrane proteins binding to glycosaminoglycans ABTB1: Elongation factor activity in translation.
	Rs43483670	6/103079093	0.23	MAPK10/Intronic variant	JUN kinase activity and MAP kinase (Signaling)
	Rs41913085	11/19125116	0.45	VIT/Nearby intergenic variant	glycosaminoglycan binding
FP	Rs109245784	4/107540967	0.29	CLCN1/Intronic variant	Chloride channels activity
	Rs41571534	19/26182297	0.19	WSCD1/Nearby intergenic variant	sulfotransferase activity, milk fat
	Rs41670205	2/55831708	0.45	LRP1B/Intronic variant	Activity in cell surface proteins that bind and internalize ligands in the process of receptor - mediated endocytosis. Calcium and low - density lipoprotein binding.
	Rs43655765	11/2350093	0.49	SNRNP200 / Synonymous variant.	Small nuclear ribonucleoprotein (determining role in splicing)
	Rs110897514	17/20336068	0.43	PCDH18 / Nearby intergenic variant	Calcium dependent cell adhesion protein
PP	Rs29014693	9/80015418	0.45	NMBR / Nearby intergenic variant	Involved in many biological functions such as feeding, pituitary, gastric and pancreatic secretion, cell development and differentiation, among others.
	Rs41772701	15/57260972	0.47	CAPN5 / Intronic variant	Endopeptidase activity calcium dependent cysteine type.
	Rs29023352	4/118676822	0.38	INSIG1 / Nearby intergenic variant	Intermediary in cholesterol synthesis control. Plays a role in the growth and differentiation of tissue involved in metabolic control.
	Rs41576177	24/32677985	0.38	OSBPL1A / Intronic variant	It binds to phospholipids and cholesterol.
	Rs41624303	5/82878525	0.26	Arntl2 / Intronic variant	Activity in protein dimerization and activity of binding transcription factors to specific DNA sequence
SCS	Rs109548201	29:50361506	0.28	CTSD / Downstream variant (9PB)	Acid protease activity in intracellular protein breakdown. Involved in the pathogenesis of several diseases.
	Rs41589068	5:30275164	0.25	Among NCKAP5L and TMBIM6 / Intergenic	TMBIM6: Apoptosis modulator, calcium homeostasis. NCKAP5L; Activity associated to signaling mechanisms.
	Rs41797394	16:33121540	0.39	Locus uncharacterized near EFCAB2	EFCAB2 Binding activity to calcium ions
	Rs41602750	5:93871154	0.44	MGST1 / Nearby intergenic variant	Glutathione peroxidase activity and homodimerization
	Rs109119975	15:31153296	0.41	Intergenic variant near to immunological important gene cluster PVRL1, THY1, C1QTNF5	PVRL1: protein homodimerization activity. THY1: binding and activating protein kinase, cell-cell interaction and cell ligand. C1QTNF5: Plays a role in cell adhesion, related to tumor necrosis factor.

Generally, genetic values of sampled animals showed a significant variability, with values almost near zero, so it was possible to have animals positively and negatively assessed to estimate additive effects on evaluated markers. The polygenic effect was used in the model in order to reduce false positives showing up (Legarra et al., 2015), given that the used BeadChip includes only a small proportion of genetic variants that can be found, and therefore it is normal not to reach to explain a large proportion of the genetic variance in some traits.

Given that present markers on the BeadChip are often selected following a uniform distribution in the genome, we can see that there is a relationship between the size of the chromosome and the number of variants within it (Table 1), which can exert additional ascertainment bias on markers on the BeadChip. This type of markers distribution has been previously reported and seeks to place at least one marker in each haplotype block linked to a QTL and to promote the imputation process (VanRaden et al., 2013; Wiggans, Cooper, Van Tassell, Sonstegard, & Simpson, 2013).

Moreover, it is important to mention that in the search process for causative mutations and relevant biological information, it was possible to identify generally that most SNPs present on 6k genotyping BeadChip are located in regions in which an important biological consequence (synonymous or non-coding mutations) cannot be attributed.

However, a small number of markers could have consequences in the expression of specific genes, although it must be noted that most non-synonymous mutations have deleterious effects and therefore decrease segregation in population likelihood. It is also worth clarifying that it is logical to have few markers in coding regions if it is considered that a very small percentage of genetic material corresponds to genes (< 3%), for this reason most of reported variations are found in noncoding regions. However, markers with great effect in genome sections without apparent or noncoding function can be found, possibly because they have a role in regulating expression or messengers' maturation.

Furthermore, it is possible to find molecular markers in intergenic regions or non-coding regions with a significant effect, because they are linked with an important QTL for a special trait, which is very common and significant in species with a small effective number of population (Ne) without recent

expansion, such as Holstein cattle (Daetwyler et al., 2014; Villa-Angulo et al., 2009).

MAF distribution allowed to observe a trend toward markers with intermediate frequencies in the Holstein individuals sample, because genotyping BeadChip are biased towards common variants, so that they could be found segregating in different cattle populations worldwide. This bias has been known as *"Ascertainment bias"* and can be a problem for population genetics studies, especially those related to diversity and genetic differentiation (Lachance & Tishkoff, 2013).

On the other hand, this type of intermediate allele frequencies may be more important for genetic evaluation, because it improves testing power and different gene variants informativeness, due to the increase in the likelihood of identifying markers segregating in the population, so that a significant proportion of the genetic variance of important traits for dairy production can be explained. Importantly rare variants (MAF > 0.05) were few and were found in regions that apparently do not involve a direct genetic consequence, so they can be neutral variants that tend to fixation. However, it is necessary to note that rare variants are difficult to associate with a significant effect and require a high statistical power to one of these variants be significant on a specific feature, even in populations as Holstein which has a small effective number of the population. This can be one of the reasons why the increase from a density point is not advantageous for genomic selection genotyping (Lohmueller, 2014).

The π fraction assessment of SNPs with effects on phenotypic traits is defined in this work accordingly as presented by Legarra et al. (2016), which is the opposite to what Meuwissen et al. (2001) defined for Bayes B. In estimates achieved in this work, it was observed a trend towards π near to zero in all evaluated traits, indicating that only a small number of markers has effect and is important. Similar results have been reported in previous studies (Peters et al., 2012; Van den Berg et al., 2013). It should be clarified that convergence where π was deemed simultaneously was elusive and sometimes the distribution had trouble exhibiting defined peaks (Figure 1A); however, this has already been reported by using Bayes C in a genome-wide association study, it has even been reported that fixing π yields better results (Van den Berg et al., 2013).

On the other hand, Bayes C has been reported as a successful method for identifying great QTLs (Sun, Habier, Fernando, Garrick, & Dekkers, 2011) and has been used for different traits in beef cattle (Peters et al., 2012; Peters et al., 2013), pigs (Fan et al., 2011), and simulation studies with different genetic architectures (Van den Berg et al., 2013). Even recently Habier, Fernando, and Garrick (2013) discussing the linkage disequilibrium, concluded that BLUP is not able to define effects in some genome linked regions and recommended Bayesian methods with t distributions *a priori* that fit better in some cases where LD decays rapidly with distance. However, for large QTLs detection it was reported a significant effect of the genetic architecture of the evaluated traits, so that detection is more accurate using Bayes C for traits of medium to high heritability, with a moderate or low number of QTLs, and when it has a large number of records (Van den Berg et al., 2013), some of which are not met with the sample taken, so the mapping results should be approached with caution.

An interesting result emerges from the variance percentage assessment explained by genetic markers, because this percentage may be increased by fixing a π fraction of markers with effect, regarding the use of all markers. This means that it is possible to generate excess noise markers in the assessment, and therefore a much smaller proportion of variants can be used with the same advantages for the estimation, may even exceed them. However, to assess the actual effect it is necessary to propose scenarios that allow to properly define the true purpose of setting a π fraction for genome-wide association studies. Pérez-Enciso, Rincón and Legarra (2015), argued that using relevant biological information for genetic evaluation may be important, even recently it has been suggested a methodology that leverages the use of information from the GWAS studies for genomic evaluation programs, showing some interesting advantages especially in cases where small samples and low heritability traits are used, depending on traits genetic architecture thereof (Van den Berg, Boichard, Guldbrandtsen, & Lund, 2016; Zhang et al., 2014).

GWAS analyses have been successful in identifying new mutations associated with diseases and production traits. However, variants identified as statistically significant often explain a very low fraction of the genetic variance, even in features having high heritability (Clarke & Cooper, 2010). Many explanations have been proposed taking into account modeling issues, contrasting genetic architecture of the traits, including epistatic effects, problems of sample size, reference population and statistical technique applied (Makowsky et al., 2011).

In this paper the variance percentage explained by the markers was variable depending on the property, which in turn depends on the genetic architecture that

can be contrasting in the productive traits evaluated (Hayes, Pryce, Chamberlain, Bowman, & Goddard, 2010). Some studies indicate that the fraction of the additive variance explained by genes or regions near genes may be from 0.05 to 0.2, but more recent studies have reported up to 0.5, which is consistent with DY, FP and PP traits, but far from the explained variance in the case of Somatic Cell Score (SCS) (Misztal, 2011).

Some research has suggested that SNPs density increases the accuracy in the estimation of genomic breeding values of animals up to a tipping point where it begins to decrease the increase rate to form a plate, where increased density does not improve accuracy achieved (Harris, Creagh, Winkelman, & Johnson, 2011; VanRaden et al., 2013), although this depends on the genetic architecture of the evaluated traits (Gibbs et al., 2009). However, the importance of identifying specific regions where the SNPs has been highlighted, with the intention to reduce interference (noise) generated by the amount of information and to improve estimates for genetic values (Zhang et al., 2014). Although the BeadChip intensity used in this work is not the highest, it could be used for estimation and identification of significant effects and markers associated with important traits in milk production.

It is clear that the genetic architecture differs between the quantitative traits of importance for the dairy industry (Hayes et al., 2010). However, for some traits a large proportion of the genetic variance is associated with genomic regions with very small effect variants, and only a few traits present variants of great effect on a phenotypic characteristic and explain large proportion of the genetic variance (Dekkers, 2012). According to the genetic architecture, features which are governed by a greater number of QTLs, are less likely to identify false positives using the methodology Bayes C. In the same way the power of the test increases with heritability (Van den Berg et al., 2013).

The Bovine HapMap consortium (Gibbs et al., 2009), reported a low level of linkage disequilibrium (LD) above 1000 kb in different breeds of dairy cattle, which obviously has an influence on GWAS studies, as these exploit the LD between the marker and the QTL. A low density of QTLs may decrease the chance of finding a marker associated with QTL but also decreases the probability of having redundant markers, taking into account the homogeneous distribution of SNPs in commercial BeadChips (Wiggans et al., 2013). Related individuals can generate significant LD even in cases where there is no connection. However, it is possible to reduce false positives by the presence of related individuals with a model that uses the information in the pedigree (MacLeod, Hayes, & Goddard, 2009), as in the present work.

It is important to note that QTL mapping is generally accompanied by a large confidence interval in chromosomes (Manichaikul, Dupuis, Sen, & Broman, 2006), thus, identifying the causative mutation and important genes on a feature is often difficult to achieve even in specialized cattle as Holstein and where the effective number of the population is small and the haplotype blocks are larger (Villa-Angulo et al., 2009). By using linked markers and not directly causal mutation or a mutation closed to QTL, losses of accuracy for genomic selection after generations due to recombination processes are presented (Meuwissen & Goddard, 2010).

It is interesting to note that in graphics of effects, when $\pi = 0.01$ (Figure 2A) and $\pi = 1$ (Figure 2B) set scales were very different and the magnitudes of the estimated effects were much greater in the case of choosing a fraction of variants with effect.

Identifying markers when $\pi = 0.01$, allowed to present a summary of SNPs with greater effect, 2 out of which were previously reported as directly associated with a QTL, although not necessarily on the same assessed trait. Others were almost all in the region attributed to a QTL (without direct proof) or close to this (Table 3).

In this study no significant association with the DGAT1 gene was found for fat percentage, when Bayes C was used setting $\pi = 0.01$, which is strange if you consider that this gene has been considered a major gene for fat percentage in milk. However, when the analysis was performed including all markers, it was possible to observe a peak on chromosome 14 that corresponded with a marker into the gene (Grisart et al., 2002; Wang et al., 2012). According to the above, it is possible to fix a fraction of markers with effect on a feature; some important SNPs go unnoticed or are overshadowed by markers with a greater effect on the assessed trait.

The summary of assessed markers and the traits they were associated to, according to the report in the public database cattle QTLdb section Animal QTLdb (Hu, Park, Wu, & Reecy, 2012) is presented in Table 3.

It is worth noting that greater effect markers often were corroborated with QTLs, supporting the estimate presented in this paper when a fraction of π markers was fixed by Bayes C.

The results should be interpreted with caution because the number of animals may not be large enough, and for that reason only 5 markers with greater effect were taken, to make a subsequent possible fine mapping work in order to find QTN directly or better map identified QTLs, as they often have very wide confidence belts.

Table 3. Relationship of polymorphisms with greater effect on Dairy Yield per lactation (DY), milk Fat Percentage (FP), Protein Percentage (PP), and Somatic Cell Score (SCS) with previous reports of mapped QTLs.

Trait	rs code	Location	QTL type associated
DY	Rs110718748	In the interval attributed to a QTL. Direct association reported	Dairy yield, % fat, pregnancy and calving ease.
	Rs41607880	Close to a QTL (< 1Mb).	Milk production, fat and protein production, delivery first estrus interval, consumption.
	Rs110425841	In the interval attributed to a QTL.	Fatty acids in milk and somatic cell score, weight.
	Rs43483670	Close to a QTL (< 1.5 Mb)	Meat quality
	Rs41913085	In the interval attributed to a QTL. Direct association reported	Reproductive traits and content of Beta-lactoglobulin.
FP	Rs109245784	No QTL reported in the region.	
	Rs41571534	In the interval attributed to a QTL.	Myristoleic acid content, EBV for milk fat production, percentage of linoleic acid in milk,
	Rs41670205	Close to a QTL (< 1 Mb).	Protein production, grade of milk.
	Rs43655765	In the interval attributed to a QTL.	Logissimus lean muscle area, pentadecanoic acid content.
	Rs110897514	Close to a QTL (<0.5 Mb).	Production of milk fat, energy production in milk, fat percentage and protein production.
PP	Rs29014693	Close to a QTL (< 0.5 Mb).	Capric acid content.
	Rs41772701	Close to a QTL (< 0.5 Mb).	Relationship Omega 3 / Omega 6 content of docosahexaenoic acid
	Rs29023352	Close to a QTL (< 0.5 Mb).	Birth weight
	Rs41576177	Close to a QTL (< 1.5 Mb).	Fertility rates and feed conversion
	Rs41624303	Close to a QTL (< 1.5 Mb).	C22:1 fatty acid content
SCS	Rs109548201	Close to a QTL (< 0.5 Mb).	Lignoceric acid content, iron content.
	Rs41589068	In the interval attributed to a QTL.	Udder height, croup lenght, cell-mediated immune response, protein percentage in milk, fat production.
	Rs41797394	Close to a QTL (< 1.5 Mb).	Paratuberculosis susceptibility, early embryonic survival.
	Rs41602750	No QTL reported in the region.	
	Rs109119975	Close to a QTL (< 0.5 Mb)	Somatic cell score, udder composition index, subcutaneous fat, milk production.

Finally, some of the current approaches suggest to find strategies to identify causal variants in complex traits, with the aim of accelerating progress towards QTN (Quantitative Trait Nucleotide) based selection and more accurate prior knowledge-based technologies and genetic architecture of traits to be assessed (MacLeod, Hayes, & Goddard, 2014; Zhang et al., 2014). Future strategies may include selection markers with great effects for future high density BeadChips, development of technologies such as "Genome Editing", and selection with the intention of obtaining optimized individuals for specific environments, and genomic selection programs calibrated in particular conditions. It is important to note that the estimates and important markers can be variable among populations, and the more remote populations are the less likely the same regions have similar effects on a specific aspect.

Conclusion

The Bayes-C estimation allowed the identification of different regions possibly associated with QTLs for the evaluated traits. Markers in important genes as ANKS1B, CLCN1, NMBR and CTSD were found for DY, FP, PP and SCS, respectively. The reported genes corresponding with metabolic functions like protein synthesis regulation, sulfotransferase activity, cholesterol metabolism intermediary, among other functions. On the other hand, according to the genetic architecture, it was possible to predict more or less the genetic variance percentage of the traits, even the proportion of significant markers for traits were biased towards zero, which suggests that much of the molecular information generated noise in the estimates, so their debugging could have important effects on the estimates for genomic selection programs.

Acknowledgements

The lead author was sponsored by the 528 National Doctorates call of Colciencias (Colombia). This work was funded in part by the National Program of Projects to Strengthen Research, Development and Innovation in postgraduate studies at the Universidad Nacional de Colombia from 2013 to 2015.

References

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics, 84*(2), 210-223.

Calus, M. P. L., De Haas, Y., & Veerkamp, R. F. (2013). Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. *Journal of Dairy Science, 96*(10), 6703-6715.

Clarke, A. J., & Cooper, D. N. (2010). GWAS: heritability missing in action? *European Journal of Human Genetics, 18*(8), 859.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R. F., … Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics, 46*(8), 858.

Dekkers, J. C. M. (2012). Application of genomics tools to animal breeding. *Current Genomics, 13*(3), 207-212.

Fan, B., Onteru, S. K., Du, Z. Q., Garrick, D. J., Stalder, K. J., & Rothschild, M. F. (2011). Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. *PloS One, 6*(2), e14726.

Gibbs, R. A., Taylor, J. F., Van Tassell, C. P., Barendse, W., Eversole, K. A., Gill, C. A., & Dodds, K. G. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science, 324*, 528-532.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., … Snell, R. (2002). Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research, 12*(2), 222-231.

Habier, D., Fernando, R. L., & Garrick, D. J. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics, 194*(3), 597-607.

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics, 12*(1), 186.

Harris, B. L., Creagh, F. E., Winkelman, A. M., & Johnson, D. L. (2011). Experiences with the Illumina high density bovine beadchip. *Interbull Bulletin*(44).

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., & Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS Genetics, 6*(9), e1001139.

Hu, Z. L., Park, C. A., Wu, X. L., & Reecy, J. M. (2012). Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Research, 41*(D1), D871-D879.

Lachance, J., & Tishkoff, S. A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays, 35*(9), 780-786.

Legarra, A., Croiseau, P., Sanchez, M. P., Teyssèdre, S., Sallé, G., Allais, S., … Elsen, J. M. (2015). A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. *Genetics Selection Evolution, 47*(1), 6.

Legarra, A., Ricard, A., &. Filangi, O. (2016). *GS3: Genomic Selection - Gibbs Sampling - Gauss Seidel (and BayesCπ)* (ANR project Rules and Tools). Retrieved from http://snp.toulouse.inra.fr/~alegarra/

Lohmueller, K. E. (2014). The distribution of deleterious genetic variation in human populations. *Current Opinion in Genetics & Development, 29*, 139-146.

MacLeod, I. M., Hayes, B. J., & Goddard, M. E. (2009). A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. *Genetics Research, 91*(6), 413-426.

MacLeod, I. M., Hayes, B. J., & Goddard, M. E. (2014). The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. *Genetics, 198*(4), 1671-1684.

Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., & Los Campos, G. (2011). Beyond missing heritability: prediction of complex traits. *PLoS Genetics,* 7(4), e1002051.

Manichaikul, A., Dupuis, J., Sen, Ś., & Broman, K. W. (2006). Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. *Genetics, 174*(1), 481-489.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics, 26*(16), 2069-2070.

Meuwissen, T., & Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics, 185*(2), 623-631.

Meuwissen, T., Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157*(4), 1819-1829.

Misztal, I. (2011). FAQ for genomic selection. *Journal of Animal Breeding and Genetics, 128*(4), 245-246.

Pérez-Enciso, M., Rincón, J. C., & Legarra, A. (2015). Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. *Genetics Selection Evolution, 47*(1), 43.

Peters, S. O., Kizilkaya, K., Garrick, D. J., Fernando, R. L., Reecy, J. M., Weaber, R. L., … Thomas, M. G. (2012). Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers. *Journal of Animal Science, 90*(10), 3398-3409.

Peters, S. O., Kizilkaya, K., Garrick, D. J., Fernando, R. L., Reecy, J. M., Weaber, R. L., … Thomas, M. G. (2013). Heritability and Bayesian genome-wide association study of first service conception and pregnancy in Brangus heifers. *Journal of Aanimal Science, 91*(2), 605-612.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

R Development Core Team. (2012). *R Foundation for Statistical Computing.* Vienna, AU*:* R Development Core Team.

Rincón, F., Zambrano, A., & Echeverri, J. (2015). Estimation of genetic and phenotypic parameters for production traits in Holstein and Jersey from Colombia. *Revista MVZ Córdoba, 20*, 4962-4973.

Sun, X., Habier, D., Fernando, R. L., Garrick, D. J., & Dekkers, J. C. M. (2011). Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods. *BMC proceedings, 5*(Suppl 3), S13. doi: 10.1186/1753-6561-5-S3-S13.

Van den Berg, I., Boichard, D., Guldbrandtsen, B., & Lund, M. S. (2016). Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study. *G3: Genes, Genomes, Genetics, 6*(8), 2553-2561.

Van den Berg, I., Fritz, S., & Boichard, D. (2013). QTL fine mapping with Bayes C (π): a simulation study. *Genetics Selection Evolution, 45*(1), 19.

Van Hulzen, K. J. E., Schopen, G. C. B., van Arendonk, J. A. M., Nielen, M., Koets, A. P., Schrooten, C., & Heuven, H. C. M. (2012). Genome-wide association study to identify chromosomal regions associated with antibody response to Mycobacterium avium subspecies paratuberculosis in milk of Dutch Holstein-Friesians. *Journal of Dairy Science, 95*(5), 2740-2748.

VanRaden, P. M., Null, D. J., Sargolzaei, M., Wiggans, G. R., Tooker, M. E., Cole, J. B., … Doak, G. A. (2013). Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science, 96*(1), 668-678.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science, 92*(1), 16-24.

Villa-Angulo, R., Matukumalli, L. K., Gill, C. A., Choi, J., Van Tassell, C. P., & Grefenstette, J. J. (2009). High-resolution haplotype block structure in the cattle genome. *BMC Genetics, 10*(1), 19.

Wang, X., Wurmser, C., Pausch, H., Jung, S., Reinhardt, F., Tetens, J., … Fries, R. (2012). Identification and dissection of four major QTL affecting milk fat content in the German Holstein-Friesian population. *PloS One,* 7(7), e40711.

Wiggans, G. R., Cooper, T. A., Van Tassell, C. P., Sonstegard, T. S., & Simpson, E. B. (2013). Characteristics and use of the Illumina BovineLD and GeneSeek Genomic Profiler low-density bead chips for genomic evaluation1. *Journal of Dairy Science, 96*(2), 1258-1263.

Yi, H., Breheny, P., Imam, N., Liu, Y., & Hoeschele, I. (2015). Penalized multimarker vs. single-marker regression methods for genome-wide association studies of quantitative traits. *Genetics, 199*(1), 205-222.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., … Simianer, H. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PloS One, 9*(3), e93017.

Received on August 08, 2017.
Accepted on November 06, 2017.