TESIS DE LA UNIVERSIDAD DE ZARAGOZA

2018

91

Elena Flavia Mouresan

Evaluation of the efficiency of the procedures of genomic selection in the autochthonous spanish beef cattle populations

Departamento

Producción Animal y Ciencia de los Alimentos

Director/es

Varona Aguado, Luis

ISSN 2254-7606

Prensas de la Universidad Universidad Zaragoza

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctora

EVALUATION OF THE EFFICIENCY OF THE PROCEDURES OF GENOMIC SELECTION IN THE AUTOCHTHONOUS SPANISH BEEF CATTLE POPULATIONS

Autor

Elena Flavia Mouresan

Director/es

Varona Aguado, Luis

UNIVERSIDAD DE ZARAGOZA

Producción Animal y Ciencia de los Alimentos

2016

FACULTAD DE VETERINARIA

Evaluation of the efficiency of the procedures of Genomic Selection in the Autochthonous Spanish Beef Cattle populations

Memoria presentada por:

Elena – Flavia Mouresan

Para optar al Grado de Doctor

Mayo 2016

FACULTAD DE VETERINARIA

D. LUIS VARONA AGUADO, Catedrático del Departamento de Anatomía, Embriología y Genética Animal de la Facultad de Veterinaria de la Universidad de Zaragoza,

CERTIFICA:

Que la tesis titulada: **"Evaluation of the efficiency of the procedures of Genomic Selection in the Autochthonous Spanish Beef Cattle populations"**, con proyecto de tesis aprobado el 5 de MAYO de 2014 por el programa de doctorado en PRODUCCIÓN ANIMAL, y de la que es autor ELENA - FLAVIA MOURESAN, ha sido realizada bajo mi dirección y cumple los requisitos necesarios para optar al grado de Doctor por la Universidad de Zaragoza.

Zaragoza, 09 de Mayo de 2016

Fdo: Prof. Dr. Luis Varona Aguado

INDEX

Chapter 1: Performance of genomic selection under a single-step approach in Autochthonous Spanish beef cattle populations

Chapter 2: Evaluation of the potential use of a meta-population for Genomic Selection in Spanish Beef Cattle Populations

FIGURE INDEX

Materials

Chapter 1: Performance of genomic selection under a single-step approach in autochthonous Spanish beef cattle populations.

Chapter 2: Evaluation of the potential use of a meta-population for Genomic Selection in Spanish Beef Cattle Populations

Chapter 4: On the haplotype diversity along the genome in Spanish Beef Cattle populations.

ANNEXE 3

TABLE INDEX

Introduction

Materials

Chapter 1: Performance of genomic selection under a single-step approach in autochthonous Spanish beef cattle populations.

Chapter 2: Evaluation of the potential use of a meta-population for Genomic Selection in Spanish Beef Cattle Populations

Chapter 3: Genetic architecture of the persistency of linkage disequilibrium across seven Spanish beef cattle populations.

ANNEXES

ANNEXE 1

ANEXO 4

ACKNOWLEDGEMENTS

First of all, I would like to sincerely thank my director, Luis Varona, for giving me this opportunity and for having the patience to teach me. I also feel grateful for my office colleagues, Juan Altarriba, Carlos Moreno, Aldemar Gonzalez-Rodriguez and Sebastian Munilla and for their valuable contribution in this thesis. Likewise, I appreciate the advices and suggestions for my thesis from the colleagues of the INIA institution, the University of UAB and of all the other institutions and universities that participated in this project.

Secondly, I would like to thank my family for the help and support they gave me, especially during my first steps here in Spain, and my friends in all corners of this planet for keeping me company through skype.

Finally, I would also like to acknowledge the Spanish cattle breeders associations of *Asturiana de los Valles, Aveleña-Negra Iberica, Bruna des Pirineus, Morucha, Pirenaca, Retinta* and *Rubia Gallega*, as well as the Spanish Federation of Ganado Selecto, for providing the necessary samples and data for this research project.

ABSTRACT

Genomic Selection (GS) has been a great success for the dairy cattle industry and its application in other species, like pigs, has been gradually growing. However, several factors impede its application in the beef cattle industry. Among those, the great number of breeds with limited census, the limited use of artificial insemination and the poor phenotyping strategies lead to low quality reference populations that produce less accurate predictions. The Autochthonous Spanish beef cattle populations have limited census but they play a crucial role in the maintenance of the economic activity of the human rural population and provide high quality products. Their breeding programs are based on BLUP genetic evaluations, while the use of DNA markers are restricted to major genes or paternity checks. The main objective of this study was to investigate the potential application of Genomic Selection in these populations under both single and multiple population approaches.

The biological material used for the development of this thesis was 171 triplets (sire/dam/offspring) from seven Autochthonous Spanish beef cattle populations (*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; *Rubia Gallega* -RG-, n=24) that were genotyped for 777,962 SNP markers with the *BovineHD BeadChip*. Additionally, the genealogical data and phenotypic data of weaning weights were available for two of the populations (*Pirenaica*, -Pi-, and *Rubia Gallega*, -RG-).

The first study analyzed the efficiency of the application of GS in two Autochthonous Spanish beef cattle populations (Pi and RG), under a single-step approach. Several genotyping strategies were tested, as well as, other factors like marker density,

effective population size, mutation rate and heritability of the trait. The results obtained showed gains in accuracy with respect to pedigree BLUP evaluation in all cases. The greatest benefit was obtained when the candidates to selection had their genotypes included in the evaluation. Moreover, genotypes from the individuals with the most accurate predictions maximized the gains but other suboptimal strategies also yielded satisfactory results. Further, the gains in accuracy increased with the marker density reaching a plateau around 50,000 markers. Likewise, the effective population size and the mutation rate showed to have an effect on the increase of accuracy, both increasing the accuracy with decreasing values. Finally, the results obtained from RG population showed greater gains with respect to Pi population for both traits because the wider implantation of artificial insemination.

The second study investigated the potential application of GS under a multi-breed model. Purebred and combined reference sets were used for the genomic evaluation and several scenarios of different genetic architecture of the trait were investigated. The single breed evaluations yielded the highest within breed accuracies. Across breed accuracies were found low but positive on average showing the genetic connectedness of these populations. The admixed populations resulted in lower accuracies compared to single-breed evaluation but showed a small advantage over small-sized purebred reference sets over the accuracies of subsequent generations. The accuracies obtained when combining all populations together resulted lower than those obtained from simple individual selection. The genetic architecture of the trait showed no significant effect of the accuracy with the exception of rare variants which yielded slightly lower results and higher loss of predictive ability over the generations.

The success of the GS from a multi-breed reference population is linked to the persistency of the linkage disequilibrium (LD) between populations. The third study attempted to analyze the genetic architecture of the persistency of LD across the seven Spanish populations. The methods used were VarLD and CorLD which showed different results. The VarLD method was able to detect differences in the LD pattern between populations, but it cannot detect the genomic regions of high LD persistency between populations. On the other hand, CorLD highlighted several genomic regions of high LD persistency among all populations. The genes located in these regions participated in metabolic pathways that included processes of cell adhesion, synapse assembly and organization and nervous system development, all associated with the *Protocadherin* gene family. The incorporation of the information of local LD persistency into the GS model yielded similar to slightly lower accuracies both for within and across breed predictions.

Finally, the haplotype diversity analysis along the genome of the seven Spanish populations showed genomic regions with substantially higher diversity, located near the telomeres and lower near the central part of the chromosome which are greatly conserved across populations. This strong concordance in the genomic regions of high haplotype diversity between populations suggest that they are mainly structural and caused probably by the higher mutation or recombination rate.

RESUMEN

La Selección Genómica (SG) ha constituido un indudable éxito en la mejora genética de vacuno de leche, y su aplicación en otras especies, como el porcino, está siendo introducida gradualmente. Sin embargo, existen varios factores que han impedido su desarrollo en vacuno de carne. Entre otros, el gran número de poblaciones de censo limitado, la reducida implantación de la inseminación artificial y la insuficiente cantidad de fenotipos que permitan generar poblaciones de referencia. Las poblaciones autóctonas de vacuno de carne en España tienen un censo muy limitado, pero juegan un papel vital en el mantenimiento de la actividad económica rural y en la producción de productos de alta calidad. Hasta ahora, sus esquemas de mejora están basados en las evaluaciones genéticas mediante BLUP, y la utilización de la información molecular se restringe a escasos genes mayores, como la *Miostatina*, y a la aplicación de test de paternidad. Por lo tanto, el principal objetivo de esta tesis doctoral es investigar la potencial aplicación de la Selección Genómica en estas poblaciones, tanto a partir de una aproximación especifica en cada población, como mediante una aproximación conjunta de varias poblaciones.

El material biológico que se ha utilizado en el desarrollo del trabajo ha consistido en 171 tripletes (padre/madre/descendiente) procedentes de siete poblaciones locales de vacuno de carne (*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; *Rubia Gallega* -RG-, n=24) que se genotiparon para 777,962 marcadores SNP mediante el *BovineHD BeadChip*. Además, se utilizó la información genealógica y fenotípica procedente de dos de las poblaciones (*Pirenaica* –Pi- y *Rubia Gallega* – RG-).

El primer trabajo analizó la eficiencia de la aplicación de la SG en dos poblaciones (Pi y RG), bajo la aproximación “single-step”. Se analizaron varias estrategias de genotipado, así como otros factores, como la densidad de marcadores, el tamaño efectivo de la población, la tasa de mutación y la heredabilidad del carácter. Los resultados mostraron que la SG siempre proporciono un incremento de la precisión sobre la evaluación genética mediante BLUP. Pese a todo, el mayor beneficio se obtuvo cuando los candidatos a la selección estaban genotipados. Además, se probó que la estrategia de genotipado que muestreaba a los individuos con menor error de predicción maximizaba la precisión, pero que, a pesar de ello, estrategias sub-óptimas también ofrecieron resultados satisfactorios. Por otra parte, se observó un incremento de la precisión a mayor densidad de genotipado, pero que alcanzó un “plateau” en torno a 50,000 marcadores. Del mismo modo, valores menores de tamaño efectivo y de tasa de mutación proporcionaron un incremento en la precisión. Finalmente, los resultados obtenidos a partir de la población RG fueron superiores a los obtenidos en la población Pi, debido a la mayor implantación de la inseminación artificial.

El segundo estudio abordo la potencial aplicación de la GS bajo un modelo multi-población. Para ello, se definieron poblaciones de evaluación compuestas a partir de individuos de una y de varias poblaciones y se analizaron bajo varios escenarios de arquitectura genética de los caracteres. Las evaluaciones en población única proporcionaron la mayor precisión dentro cada población. Las precisiones de la predicción entre poblaciones fueron muy bajas, aunque siempre positivas poniendo en evidencia la conexión genética entre poblaciones. Las poblaciones compuestas proporcionaron una precisión menor si se comparan con la evaluación en población

única, pero mostraron un ligero incremento sobre poblaciones de referencia de tamaño equivalente en una única población. Por otra parte, la arquitectura genética de los caracteres no mostro ningún efecto relevante, salvo el escenario de simulación que utilizó variantes raras para la generación de la variabilidad genética de los caracteres. En él, se observaron una menor precisión y una mayor pérdida de la misma a lo largo de las generaciones.

El éxito de la SG a partir de una población de referencia compuesta está relacionado con la persistencia del desequilibrio de ligamiento (DL) entre poblaciones. El tercer estudio pretendió analizar la arquitectura genética de la persistencia de DL entre las siete poblaciones analizadas. Se utilizaron dos métodos (VarLD y CorLD) que mostraron resultados diferentes. El método VarLD permitió detectar diferencias entre los patrones de DL entre poblaciones, pero no pudo identificar las regiones de mayor persistencia de DL. Por el contrario, el método CorLD sí que fue capaz de detectarlas. Los genes localizados en estas regiones de mayor persistencia entre poblaciones participan en rutas metabólicas que incluyen procesos de adhesión celular, sinapsis, organización y desarrollo del sistema nervioso, asociadas, en general, con la familia génica de las protocaderinas (*Protocadherin).* Pese a todo, la incorporación de la información acerca de la persistencia de fase de DL en los modelos de evaluación genómica no proporcionó ningún incremento de precisión tanto dentro como entre poblaciones.

Finalmente, se analizó la diversidad haplotípica a lo largo del genoma en las siete poblaciones y se mostró una notable heterogeneidad en la misma. En general, la diversidad fue mayor en la cercanía de los telómeros que en la parte central de los cromosomas. Es destacable que las regiones de alta diversidad haplotipica fueron

coincidentes entre poblaciones, sugiriendo que las causas de esta diversidad son estructurales, como pueden ser las tasas de mutación y recombinación locales.

INTRODUCTION

Spanish beef cattle breeds

The first signs of bovine domestication took place in the valleys of the rivers Euphrates and Tigris around 10,000 years ago (Helmer *et al.,* 2005; Hongo *et al.,* 2009). Later, the domestic livestock started appearing in western Anatolia and the south-east of Europe, south of Italy and Central Europe around 8,000 years ago, mainly due to the cattle's growing economic importance for the production of milk and meat (Vigne and Helmer, 2007; Vigne, 2008). The domesticated cattle reached the Iberian Peninsula through two distinct routes. The first went through Central Europe while the second reached the Mediterranean coast through the African continent, Egypt specifically. These two migration flows and the effect of the Iberian environment gave birth to 3 ethnic branches that classify traditionally the Spanish autochthonous cattle: *B. Taurus ibéricus, B Taurus cantábricus* and *B. Taurus turdetanus* (Sánchez-Belda, 1984).

In May 2015 Spain possessed a population of more than 6 million bovine animals. 1.5 million of them are registered in one of the 45 breeds officially recognised by the Ministry of Agriculture, Alimentation and Environment. There are 8 autochthonous breeds in development (*Autóctona de Fomento*), 31 autochthonous breeds in danger of extinction (*Autóctona en Peligro de Extinción*) and 6 foreign integrated (*Integrada*) breeds (Table 1). Up to 35% of the bovine population belongs to the autochthonous breeds in development (*Lidia:* 13.3%, *Asturiana de los Valles:* 6.4%, *Parda de Montaña:* 3.8%, *Avileña-Negra Ibérica:* 3.6%, *Pirenaica:* 2.7%, *Rubia Gallega:* 2.6%, *Retinta:* 1.9% and *Morucha:* 1.1%). Moreover, among the autochthonous breeds in danger of extinction *Asturiana de la Montaña:* 1.4%, *Bruna dels Pirineus:* 0.9% and *Tudanca:* 0.9%, stand out for their importance.

Table 1. Breed, Status, Evolutive Tendency and Number of Animals.

Breed	Classification	Evolutive tendency of the population	Total animals
Asturiana de los Valles	Autóctona de Fomento	Expansión	94,682
Avileña-Negra Ibérica	Autóctona de Fomento	Expansión	53,428
Morucha	Autóctona de Fomento	Expansión	16,378
Pirenaica	Autóctona de Fomento	Expansión	40,026
Retinta	Autóctona de Fomento	Expansión	29,394
Rubia Gallega	Autóctona de Fomento	Recesión	38,797
Parda de la Montaña	Autóctona de Fomento	Expansión	55,509
Lidia	Autóctona de Fomento	Recesión	195,967
Albera	Autóctona en vía de extinction	Expansión	763
Alistana-Sanabresa	Autóctona en vía de extinction	Recesión	3,351
Asturiana de la Montaña	Autóctona en vía de extinction	Expansión	21,460
Avileña-Negra Ibérica Variedad Bociblanca	Autóctona en vía de extinction	Expansión	813
Berrenda en Colorado	Autóctona en vía de extinction	Expansión	5,791
Berrenda en Negro	Autóctona en vía de extinction	Recesión	3,169
Betizu	Autóctona en vía de extinction	Expansión	884
Blanca Cacereña	Autóctona en vía de extinction	Recesión	1,049
Bruna dels Pirineus	Autóctona en vía de extinction	Expansión	13,542
Cachena	Autóctona en vía de extinction	Expansión	4,195
Caldelá	Autóctona en vía de extinction	Recesión	1,304
Canaria	Autóctona en vía de extinction	Expansión	1,314
Cárdena Andaluza	Autóctona en vía de extinction	Expansión	1,001
Frieiresa	Autóctona en vía de extinction	Expansión	673
Limiá	Autóctona en vía de extinction	Expansión	852
Mallorquina	Autóctona en vía de extinction	Expansión	499
Marismeña	Autóctona en vía de extinction	Recesión	2,204
Menorquina	Autóctona en vía de extinction	Expansión	1,514
Monchina	Autóctona en vía de extinction	Expansión	2,060
Morucha Variedad Negra	Autóctona en vía de extinction	Expansión	3,664
Murciana-Levantina	Autóctona en vía de extinction	Expansión	37
Negra Andaluza	Autóctona en vía de extinction	Expansión	2,417
Pajuna	Autóctona en vía de extinction	Recesión	727
Palmera	Autóctona en vía de extinction	Expansión	596
Pasiega	Autóctona en vía de extinction	Expansión	447
Sayaguesa	Autóctona en vía de extinction	Expansión	1,569
Serrana de Teruel	Autóctona en vía de extinción	Expansión	411
Serrana Negra	Autóctona en vía de extinción	Expansión	498
Terreña	Autóctona en vía de extinción	Expansión	2,474
Tudanca	Autóctona en vía de extinción	Expansión	13,075
Vianesa	Autóctona en vía de extinción	Expansión	2,401
Blonda de Aquitania	Integrada	Expansión	12,234
Charolesa	Integrada	Recesión	13,200
Fleckvieh	Integrada	Recesión	7,023
Frisona	Integrada	Recesión	760,554
Limusina	Integrada	Recesión	48,144
Parda	Integrada	Expansión	10,967

(Ministerio de Agricultura, Alimentación y Medio Ambiente, 2014)

Traditionally, a large part of these breeds were destined towards a triple production of meat, milk and workforce. Nonetheless, in present day, the autochthonous breeds are destined almost exclusively towards meat production except for the *Lidia* breed. The production systems of these populations are extensive or semi-extensive and are not homogeneous. In fact, at least 3 main production systems can be highlighted:

- Pasture system: It is located in areas with infertile lands of low agricultural aptitude. These lands are found at the west and southwest of the Iberian Peninsula. The breeds that are used in such systems are rustic breeds, fully adapted to the difficult climatic conditions like *Avileña-Negra Ibérica*, *Morucha* and *Retinta.*
- Mountain system: The mountainous areas, like the Pyrenees, were always associated with beef production. The breeds that are exploited under this type of system are *Bruna dels Pirineus*, *Pirenaica*, *Parda de Montaña* and *Asturiana de la Montaña.* During the winter the animals are kept in the valleys or near the villages and are feed with hay. In the spring and autumn the animals are taken to the mid-mountain pastures while in the summer the animals are taken higher to the mountains to benefit from the lower temperatures there and the local vegetation.
- Humid not mountainous regions: The regions of Spain that apply this system are located mainly in Galicia, Asturias and Cantabria. These regions have pastures of excellent quality and are also suitable for the production of artificial pastures. The main autochthonous breeds that are used under this system are the *Rubia Gallega* and the *Asturiana de los Valles* (Revilla, 1997).

The distribution pattern of the different production systems is related to the geographic distribution of these breeds in the country and the variability in the muscle

development, growth capacity (Piedrafita *et al.*, 2003) and in the carcass and meat quality (Gil *et al.*, 2001; Piedrafita *et al.*, 2003; Serradilla *et al.*, 2008). Therefore, *Asturiana de los Valles*, *Rubia Gallega* and *Pirenaica* show a higher muscle development, while *Retinta*, *Avileña-Negra Ibérica* and *Morucha* can be classified as rustic breeds. *Bruna dels Pirineus* and *Parda de Montaña* occupy an intermediate place. Equivalently, the populations with higher muscle development present better carcass conformation and greater percentage of lean meat, while the more rustic populations give carcasses with a higher degree of fatness and a poorer conformation. These populations present a higher degree of intramuscular fat infiltration that gives a better taste, aroma and tenderness of the meat.

In general, the Spanish autochthonous breeds are characterized by high genetic variability that grants them the ability to adapt to climatic changes, and along with the process of evolution and selection have promoted their rusticity. As a result, these breeds are highly resistant to the local diseases and have a better capacity to take advantage of the low quality pasture resources (Hoffmann, 2010). Additionally, they play an important role in the maintenance and the development of the rural population covering around 30% of the human alimentation needs and contributing to food safety (Molina, 2010). The economic importance of the local breeds is based on the lower production costs and the higher quality of the products. The particular characteristics of the autochthonous breeds permit their use in an open range exploitation system and therefore reduce the production cost by avoiding the high-cost maintenance of the intensified exploitation and moreover deal with the health implications and the environmental burden of such systems (Ibañez and Mas 1997). Nowadays, the consumers are demanding products of distinguished quality, produced under the concept of environmental respect and consideration of the effect this procedure has

on climate change. Thus, the quality is one of the principal characteristics that distinguish the autochthonous breeds from the specialized ones.

Most of the breeding programs of the Spanish Autochthonous beef cattle populations started in the late 80's and the 90's (Serradilla, 2008). Nowadays, their breeding schemes are based on the evaluation of the direct and maternal effects of growth related traits and morphology, and, in some cases, they also include carcass and meat quality traits (carcass conformation, degree of fatness or pH after sacrifice) and reproductive traits (calving ease, precocity or interval between parities). The breeding evaluation of these traits is performed using the mixed model BLUP (Henderson, 1984) and the use of molecular information is only restricted to single genes of special interest, like the MTSN (*Miostatin*) gene in the *Asturiana de los Valles* population or to check paternity using a standard microsatellite set. (Serradilla, 2008)

Genomic Selection

During the XXth Century, the advances of population and quantitative genetic theories provided tools for the prediction of breeding values for the candidates to selection. These techniques have allowed a remarkable increase of the genetic progress in all livestock populations. In particular, this “genetic” revolution started in the 40’s to 60’s with the application of selection indexes (Hazel, 1943) that allow to weight the information provided by related individuals and also to incorporate the phenotypic information that proceed from genetically correlated traits. Later on, the advent of new statistical developments allowed the use of mixed model procedures (Henderson, 1973), such as BLUP (Best Linear Unbiased Prediction). The BLUP method improves the prediction provided by the selection indexes thanks to the joint prediction of breeding values and the estimation of some known systematic effects, such as sex, age of parity or contemporary groups. The application of the BLUP procedure spread out after the development of simple rules to compute the required inverse of the numerator relationship matrix (**A**), that were developed by Henderson (1976) and Quaas (1976). The BLUP method takes into account the effects of drift, selection and assortative mating (Kennedy and Sorensen, 1988) and it is easy to generalize into a multiple trait context (Henderson and Quaas, 1976). Until very recently, BLUP has been the method of choice to obtain predictions of breeding values for most of the livestock populations.

Molecular Markers and Genomic Selection

Since the 90’s, molecular information was made available due to the advances of techniques of molecular biology. This new source of information gave the opportunity to enhance the response to selection by incorporating it to traditional breeding

programs, especially for traits that present difficulties in their improvement by traditional selection (Dekkers et *al.,* 2004). Such traits are those with low heritability (ex. Reproductive traits) and traits whose phenotypes are difficult to obtain (ex. Disease resistance or meat quality). The first attempts of direct selection at DNA level came along with a method called Marker Assisted Selection (MAS). This method consists of locating the genes or the Quantitative Trait Loci (QTL) underlying the trait of interest and then incorporating that information in the procedure for the prediction of breeding values (Kennedy *et al.,* 1992). Although, some major QTLs were detected for some traits in cattle (ex. DGAT1, fat content in milk -Grisart *et al.,* 2001- and CDH1, affects infectious pancreatic necrosis virus -Moen et al., 2015-) and pigs (RYR1 -Fujii *et al.,* 1991- causing the porcine stress syndrome, ESR- Rothschild *et al.,* 1996-, related with prolificacy or IGF2- Van Laere *et al.,* 2003- affecting fatness and growth), the majority of the traits of interest had very few QTLs located and less than 10% of the genetic variance explained. Therefore, the difficulty of detecting genes, the small portion of the genetic variance explained by the few genes detected and the fact that the traits of economic interest are controlled by many genes with small effect led to a low uptake of this method by the industry.

More recently, the identification of a large number of Single Nucleotide Polymorphisms (SNPs) along the genome, as a by-product of the sequencing efforts (Daetwyler *et al.,* 2014), and the development of SNP-chip genotyping technology (Gunderson, *et al.,* 2005), that made affordable the genotyping of thousands of these markers at low cost, led Meuwissen *et al.* (2001) to propose a new method of selection denoted as Genomic Selection (GS). These authors proposed the use of a dense marker map for the prediction of total breeding values. The basic idea underlying the GS approach is that it is expected that some of the SNP markers are located near the QTLs of traits

of interest and due to linkage disequilibrium (LD) between them, they should be inherited jointly. In this way, all of the QTLs affecting a trait may be in LD with one or more markers. As a consequence, if there are enough markers to cover the whole length of the genome, the additive effects of the QTLs can be captured by the markers without the necessity of locating them. Initially, the GS procedures are based on estimating the effects associated to the markers in a reference population where genotypes and phenotypes are available for all individuals. These estimates are then combined with the genotypes of the selection candidates to produce genomic estimated breeding values (GEBV) of the candidates to selection. These candidates are usually young animals that do not have reliable trait records or no records at all. Moreover, with this approach of the GS methodology the genealogical information is not strictly needed.

Genomic Selection Methods

The statistical model used for the simultaneous estimation of the SNP effects in a reference population is:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X\beta} + \boldsymbol{e} \text{ ,}$$

where $\boldsymbol{y}$ is the vector of phenotypes, $\mathbf{1}$ is a vector of ones, μ is the trait's mean, $\boldsymbol{\beta}$ is the vector of the effects associated to the markers, $\boldsymbol{X}$ is the genotype matrix and $\boldsymbol{e}$ is the vector of residual effects. The GEBVs are calculated as follows:

$$\boldsymbol{GEBVs} = \boldsymbol{Z\widehat{\beta}} \text{ ,}$$

where $\boldsymbol{Z}$ is the genotype matrix of the selection candidates and $\widehat{\boldsymbol{\beta}}$ is a vector containing the estimated effects of the markers. The number of effects to estimate is usually greater than the number of available phenotypes and therefore the traditional

statistical methods that treat the SNP effects as fixed are unable to perform this task. The standard solution to this problem is to introduce some kind of regularization of the marker effects (Gianola, 2013). There are several methods to implement this regularization that could be classified as:

Gaussian Regularization

The most standard regularization is the Gaussian, proposed by Meuwissen *et al.* (2001) that provides BLUP estimates of the SNP marker effects by assuming that they follow a Gaussian distribution with equal variance for all SNPs. The procedure involves a quadratic penalization numerically equivalent to a ridge regression (Hoerl and Kennard, 1970). This method is denoted SNP-BLUP and the estimates are a lineal combination of the phenotypes. If the genotypes $\boldsymbol{X}$ are standardized in such way that they have mean 0 and standard deviation 1 for every SNP, the SNP-BLUP method is equivalent to the GBLUP method (Habier *et al.,* 2007; VanRaden, 2008; Goddard, 2009), that is similar to the traditional BLUP described by Henderson (1973), but it uses a genomic relationship matrix (**G**) instead of the standard pedigree relationship matrix (**A**). The genomic relationship matrix (VanRaden, 2008) is built from molecular information in a way that, individuals that share identical by state genotypes for a larger number of markers are more similar and therefore, have larger values in the corresponding cell of the matrix. The main advantages of this method are that 1). It is computationally faster, 2) the existing BLUP software can be used just by replacing the pedigree relationship matrix with the genomic relationship matrix, 3) it provides breeding values automatically and, 4), it is very easy to generalize to multiple trait analysis.

Further, in real populations, it is frequently not feasible for entire populations to be genotyped because of its high cost or logistical constrains (i.e. slaughtered or foreign animals). Thus, in order to use the phenotypic data of non-genotyped animals a multiple-step GS approach has to be implemented. With the traditional approach (Meuwissen *et al.,* 2001), the first step is to create pseudo-phenotypes for the genotyped animals from the phenotypes of its ungenotyped relatives. An example of pseudo-phenotype is the average daughter production for a dairy bull. The second step comprises of a genomic prediction using the pseudo-data and their genotypes and finally, the third step combines the traditional EBV and the GEBV into a total EBV (e.g. VanRaden, 2008). The advantages of such system include no change to the regular evaluations and simple steps for predicting genomic values for young genotyped animals. However, the disadvantages include loss of information due to weights caused by different amount of information in the original data set and potential bias caused by selection. Moreover, in species like sheep, swine and beef cattle, or traits like maternal traits, pseudo-records are more difficult to compute or estimate. To cope with the loss of information of the multiple-step process, Legarra *et al.* (2009) suggested a simplification to this process by performing a joint evaluation using all phenotypic, pedigree and genomic information into a single step. To achieve this, these authors suggested to compute an **H** matrix that combines the numerator relationship matrix (**A**) for the non-genotyped animals with the genomic relationship matrix (**G**) for the genotyped animals. Additionally, Misztal *et al.* (2009) developed an efficient computing strategy to obtain solutions to mixed model equations in which the numerator relationship matrix is modified in order to account for genomic information. The idea behind this method is to include ungenotyped animals and take advantage of all sources of information available.

Non-Gaussian Regularization

The Gaussian regularization assumes that all SNPs come from a normal distribution with constant variance and results in all the SNPs having some effect onto the analysed trait. Biologically, it seems more reasonable to assume that some of the thousands of markers are in LD with a QTL, and therefore can capture their effect and some markers are not in LD with any gene and cannot capture any effect. To achieve this basic idea, several methods have been developed to incorporate different prior assumptions using different distributions or mixtures of distributions.

Bayes A

Originally proposed by Meuwissen *et al.* (2001), this method assumes that all SNPs have an effect drawn from a normal distribution with zero mean and a variance associated to each marker. The prior distribution of the locus-specific variance is a scaled inversed Chi-squared distribution. Consequently, the prior distribution of SNP effects are t shaped. In this way, some markers are allowed to have larger variances than others and therefore have a larger effect.

Bayes B

As before, this method assumes a normal distribution on the marker-effects and variance associated to each marker just as Bayes A. However, Bayes B differs from Bayes A as to the assumptions made for the distribution of the variance. A mixture of distributions on the variance is used, where the variance is zero with probability π and distributed as in Bayes A with probability 1-π (Meuwissen *et al.,* 2001). The election of π is arbitrary with no justification.

Bayes Cπ & Dπ

To address some drawbacks of Bayes A and Bayes B, such as the prior probability of π and the hyper-parameters of the prior distribution of the variance, Habier *et al.* (2011) described the Bayes Cπ and Bayes Dπ methods. In Bayes Cπ a common variance to all markers is assumed with probability 1-π and variance zero with probability π. Additionally, the proportion π of markers is treated as unknown and is estimated from the data. Bayes Dπ imposes a prior on the scale parameter of the inverse chi-square distribution, which is the prior distribution of the variance of marker-effects.

Bayes R

This method assumes a mixture of several normal distributions with different variances in order to allow for SNP effects from 0 to moderate or large. Thus, each SNP effect is assigned to one of the proposed distribution with a probability that is calculated from data (Erbe *et al.,* 2012).

Bayesian LASSO

This method (Park and Casella, 2008) was proposed for genomic selection by De los Campos *et al.* (2009) and Usai *et al.* (2009). Here, a double exponential prior distribution is assumed for the marker-effects with parameter λ. This procedure performs a larger shrinkage on the marker-effects than previous methods in a way that a large number of markers are estimated with a very small effect, and only a few markers are allowed to have larger effects. The degree of shrinkage is determined by the parameter λ, which has to be estimated previously to the analyses. Park and Casella, (2008) proposed the use of Empirical Bayes by Marginal Maximum Likelihood using an appropriate hyper-prior for the estimation of λ and Legarra *et al.* (2011) proposed a modification of this method (BL2Var) which considers two different

variances for the distribution of marker-effects and the residuals. Moreover, there is no need to pre-estimate the parameter λ as it is estimated from the data simultaneously with the marker effects. Up until now Bayesian LASSO has been widely applied for genomic evaluations as it provides accurate predictions for low density genotyping (Usai *et al.,* 2009) and for traits that are regulated by many genes with a small effect (Cleveland *et al.,* 2010).

Machine Learning and non-parametric methods

As an alternative to the above described regularization based on prior distributions, several machine learning or non-parametric procedures have been proposed. In this sense, Croiseau *et al.* (2011) suggested the implementation of the elastic net algorithm (Zou and Hastie, 2005). This is a combination of Ridge – BLUP- (Hoerl and Kennard, 1970) and Lasso regression (Tibshirani, 1994) weighted by a parameter α which takes values from 0 to 1. When α=0, a BLUP model is defined whereas α=1 a LASSO model is chosen. Additionally, a pre-selection of markers can be applied prior to the analyses. The purpose of this method is to provide a more flexible tool to apply shrinkage on the SNP effects.

With respect to non-parametric procedures, some methods that have been proposed are:

- Reproducing Kernel Hilbert Spaces Regression (RKHS) (Gianola *et al.,* 2006)
- Support vector machines (Moser *et al.,* 2009)
- Random Forest (Sun, 2010)
- Neural Networks (Gianola *et al.,* 2011).
- Boosting (Gonzalez-Recio *et al.,* 2010)

All these methods resulted in accuracies similar or even higher than the ones obtained by the Bayesian methods (Gonzalez-Recio *et al.,* 2010) and they give the possibility of capturing interactions between genes and between genes and environment (Sun, 2010) or to capture non-linear relations (Gianola *et al.,* 2011).

In fact, these algorithms are more attractive for application to the complex situations found in biological systems, as they are able to accommodate additive, dominant or even epistatic effects. It is believed that these methods can approach the genetic architecture of a trait more than the linear models. However, all of them need a case-specific tuning and sometimes they involve very strong computation efforts. Thus, their implementation in large populations is more difficult that methods based on a simple regularization of additive marker effects.

Comparison of Methods

In simulation studies the non-Gaussian methods outperform the Genomic BLUP methodology (Meuwissen *et al.,* 2001; Habier *et al.,* 2007; Clark *et al.,* 2011). Nevertheless, in real data this does not always occur (Moser *et al.,* 2009; Erbe *et al.,* 2012; Heslot *et al.,* 2012). The reasons behind this phenomenon (Daetwyler *et al.,* 2010) might be the genetic architecture of some traits that renders true the assumption that all markers have an effect, the extent of LD over large genomic distances and therefore the higher number of SNPs associated to a gene, and finally the low marker density and the need for more SNPs in order to capture the QTL effect.

Moreover, the Single Step approach (Aguilar *et al.,* 2010) is a simpler method to combine all information in a straightforward way, with the additional advantage of requiring little changes to existing software. Further, there have been several studies (Aguilar *et al.,* 2010; Chen *et al.,* 2011; Aguilar *et al.,* 2011; Baloche *et al.,* 2014) that

showed that its accuracy is usually as high as any other method and sometimes even higher. Apart from the accuracy, some additional benefits of this method include:

- the automatic accounting of all relatives of genotyped animals and their performance,
- the simultaneous fit of genomic information and estimates of other effects (e.g. contemporary groups),
- the extra accuracy in genotyped animals is transmitted to all their relatives,
- any model using relationship matrices can be fit using combined relationship matrices,
- It provides an analytical framework for alternative modelling of data (Legarra *et al.,* 2014a).

Genomic Selection in Livestock Populations

According to a recent review by Meuwissen *et al.* (2016), the rate of implementation of Genomic Selection in livestock industries is variable. It has been very quick and successful for dairy cattle, while for other species the uptake has been slower, with less satisfying results. The genomic evaluation in dairy cattle populations has become a standard since almost 10 years. In fact, the first unofficial USDA evaluations based on SNP genotypes were released in April 2008 and became official for Holsteins and Jerseys in January 2009 and for Brown Swiss in August 2009. Genotyping of thousands of animals has been financed by research grants and contributions from AI and breed organizations. A key factor in the construction of these large reference populations is the collaboration between countries by establishing consortiums (Eurogenomics, The North American Consortium, “rest of the world” consortium). Worldwide, approximately 2 million dairy cattle have been genotyped for the purpose

of GS. More than half come from the USA from 4 dairy breeds: Holstein (934,780 animals), Jersey (120,439 animals), Brown Swiss (19,588 animals) and Aryshire (4,767 animals) (Wiggans, https://www.cdcb.us/Genotype/cur_density.html). In fact, the majority of the genotyped animals in many countries are now heifer calves, because the cost of genotyping is low and allows to genotype heifer calves for the purpose of choosing which heifer to retain in the herd (Pryce and Hayes 2012; Weigel *et al.,* 2012). The accuracy of genomic prediction in dairy cattle exceeds 0.8 for production traits and 0.7 for fertility, longevity, somatic cell count and other traits (e.g., Wiggans *et al.,* 2011; Lund *et al.,* 2011). These accuracies were possible to obtain due to the large reference populations of each breed that contain many progeny-tested bulls with highly accurate phenotypes and to the fact that the GEBVs are often used to predict close relatives of the animals in the reference population. The achievement of these high accuracies provides an alternative to traditional progeny test, leading to a very important reduction of the generation interval (Hayes *et al.,* 2009a).

In pig breeding the most important breeding step is the selection of elite boars in the nucleus herd (Ibañez-Escriche *et al.,* 2014). In contrast with dairy cattle, the boar test recordings come generally before the selection and therefore extra gains due to a reduction of the generation interval are limited. The implementation of GS in pig breeding is therefore mainly directed at traits whose recording is invasive such as slaughter quality (Samore *et al.,* 2015) or maternal traits that cannot be recorded on the boars (Lillehammer *et al.,* 2011). However, the elite breeding nucleus animals are selected for purebred performance in a favourable environment but pork is produced by crossbred pigs in commercial environment. In this sense, Esfandyari *et al.* (2015) showed that by genotyping crossbred pigs and recording their performance can increase the response to selection and improve purebred nucleus animals for

crossbred performance. Further, pig breeding can obtain additional benefits by the implementation of alternative models that include dominance (Su *et al.,* 2012; Vitezica *et al.,* 2013) and epistatic effects (Muñoz *et al.,* 2014) with the aim of predicting cross-breeding performance (Vitezica *et al.,* 2016)

The poultry industry is investigating the use of GS and its implementation to the traditional breeding programs. In this sense, Wolc *et al.* (2015) conducted an experiment in layers to test the effectiveness of GS in genetic gain over traditional selection. The response to selection was greater for the GS line for most of the traits included in the index of selection, and in some cases even doubled. In broilers most traits can be recorded on both sexes at early age and therefore the application of GS is not as obvious as in layers. However, as in pig breeding, possible uses are for selection to improve crossbred performance in commercial environment and for traits that cannot be recorded in the nucleus such as disease challenge tests.

The implementation of GS in sheep and goat breeding is still on preliminary steps in most of the populations although some pilot studies have been developed in dairy (Duchemin *et al.,* 2012, Legarra *et al.,* 2014b) or meat quality traits (Daetwyler *et al.,* 2012) and some studies have also reported evaluations of the potential implementation of such schemes (Shumbusho *et al.,* 2015; Casellas and Piedrafita, 2015). The main conclusions of these studies is that GS can accelerate the selection response, with special interest in dairy production, mimicking the dairy cattle scheme, or in meat production for traits of difficult recording.

Finally, the implementation of genomic selection on other species (Ibañez-Escriche and González-Recio, 2011), such as rabbit or fish, is still under development, although there have been some studies that focused on the relevance of genomic selection on

disease resistance in fish (Villanueva *et al.,* 2011; Yañez *et al.,* 2014; Castillo-Juarez, 2015).

Genomic Selection in beef cattle

The beef cattle industry has been more reluctant on the uptake of this new technology. In a recent review, Berry *et al.* (2016) analysed the factors that have hindered the development and implementation of GS in beef cattle relative to dairy cattle:

- Multiple breeds and crossbreds: One of the most important factors is that the beef cattle industry is comprised by multiple breeds and crossbreds. Unlike dairy cattle, where the predominant breed is Holstein-Freisian, a plethora of British and Continental beef breeds are used in temperate climates, each with effective population sizes greater than Holstein-Freisian, and each with their own breed-specific attributes. Moreover, in tropical climates the *Bos indicus* (Nellore and Brahman) and *taurindicus* (Brangus and Bradford) breeds are preferred.
- Lack of artificial insemination: When compared to dairy cattle, a smaller proportion of beef calves are generated from artificial insemination (AI). This fact results in fewer bulls with highly accurate genetic evaluations and therefore, the need of larger reference populations, which are more difficult and expensive to assemble.
- Poor international genetic connectedness: The lack of AI in beef cattle contributes to poor connectedness among populations in different countries and as a result collaborations between countries are more difficult to establish.
- Low levels of phenotyping: Accurate genomic predictions are predicated on access to large quantities of phenotypic information (Daetwyler *et al.,* 2008). However, phenotyping strategies in beef production systems tend to be poorer than those of dairy, especially in commercial populations. Sire recording is also

generally poor in many beef production systems especially where multi-sire mating is practised.

- Lower-margin business model: The lower economic margins of beef production gives little motivation for investment that leads to poor adoption rates of genomic technologies to advance gain in beef. Reduced uptake, in turn, impedes the growth of the reference population necessary to improve the accuracy of predictions. The development of a genomic selection-based breeding program requires an initial investment in genotyping and phenotyping as well as in necessary infrastructure to deliver routine genomic evaluations. In contrast, the high accuracy of genomic prediction achieved in many dairy populations, coupled with it being a generally higher profit margin business, justifies the investment by the producers in genotyping to aid the selection of candidate female replacements (Weigel *et al.,* 2012).

However, in some beef breeds, genomic selection in now applied on a large scale. In the USA, more than 52,000 Angus animals have been genotyped for GEBV evaluation (Lourenco *et al.,* 2015), although the accuracies reported in are lower than those in dairy cattle ranging from 0.3 to 0.7 (Van Eenennaam *et al.,* 2014). This comes as a result of the lower quality of the reference population of beef cattle compared to dairy cattle. In fact, the accuracy of genomic predictions is influenced by effective population size, number of animals with genomic and phenotypic information (Daetwyler *et al.,* 2008), and the relatedness of the reference population to the candidate animal population (Pszczola *et al.,* 2012). Generally, the beef cattle reference populations contain fewer progeny tested animals within a breed and, in addition, the validation population may be less closely related to the reference population than in dairy cattle.

Following the Angus example, the American Hereford Association developed a training population and, in a similar way, other breed associations gradually followed (Saatchi *et al.,* 2011). In Europe, genomic evaluations in beef cattle are currently not official. However, Ireland will launch official genomic proofs in early 2016 for all beef breeds, based on a one-step multi-breed genomic evaluation, which includes more than 100,000 animals with genotypes and phenotypes. Moreover, genomic evaluations for UK Limousin cattle were planned to be available in December 2015 based on a reference population of 720 Limousin animals with high-density genotypes and an additional 1,700 animals with medium-density genotypes. In addition, in February 2015 unofficial French genomic evaluations were made available for Charolais, Limousin and Blonde d'Aquitaine based on a two-step approach blended with traditional genetic evaluation using a selection index approach. Finally, in Australia, genomic evaluations for Angus and Brahman populations are available, and in South America, some degree of implementation started on 2008 following the same approach used in North America (Barry *et al.,* 2016).

Across breed Genomic Selection

Until now, Genomic Selection has been implemented mainly in the dairy cattle industry, where the existence of a large enough reference population, permits to achieve highly accurate predictions (Hayes *et al.,* 2009). However, the beef cattle industry does not follow the structure of dairy cattle industry and, in some populations, the construction of a large enough reference population presents serious difficulties, due to the existence of many and small populations. Therefore, the within-breed evaluation gives poor results due to the small size of the training sets. Moreover, the estimations obtained from one breed cannot be applied to other breeds as they usually give very low accuracies (Harris *et al.,* 2008).

To avoid this inconvenience, De Roos *et al.* (2009) proposed pooling animals from different populations to obtain a large training set. His results showed that adding individuals from the second population to the training set (composed only by the first population), had some effect on the reliability of the genomic breeding values in the first population and it was most beneficial when the heritability of the trait was low. Furthermore, when the two populations had diverged for only few generations and the marker density was high, the information from the second populations was most valuable.

Since then, a number of studies have compared the predictive ability of genomic models trained in a joined reference population by combining populations of the same breed or populations of different breeds. In dairy cattle the predominant breed globally is Holstein-Freisian. Nonetheless, there are many smaller dairy breeds that are restricted by small reference populations of progeny tested bulls and therefore have low reliabilities of GEBV (Goddard *et al.,* 2009). Table 2 presents the results from

several studies on combining different dairy cattle populations. Three categories of results can be noted: The first category includes the combination of same-breed populations were it is clear that increases the accuracies of GEBVs especially where the exchange of genetic material between populations is large. Large improvements are realized when combining populations in North America (Schenkel *et al.,* 2009; VanRaden *et al.,* 2012) and in the EuroGenomics collaboration (Lund *et al.,* 2010). Similar results were obtained by Zhou *et al.* (2013) for genomic predictions for Chinese HF using a joint reference with Nordic HF.

The second category of results comes from joining more distinct breeds that use common bulls. An example of such combination are the Nordic red breeds (Danish Red, Swedish Red, Finnish Ayrshire and Norwegian Red) where a high exchange of genetic material is occurring (Brøndum *et al.,* 2011). The gain in the reliability of the GEBVs for these breeds is substantial, but smaller than combining populations of the same breed.

Finally, a third group of studies attempted to join populations of more distantly related breeds. Among them, Karoui *et al.* (2012) combined Holstein, Normande and Montbeliard and found a slight increase in reliabilities for production traits of the breed with the smallest population size. There are several studies (Hayes *et al.,* 2009b; Pryce *et al.,* 2011; Olson *et al.,* 2012; Erbe *et al.,* 2012) that report on the effect of combining Holstein-Freisian and Jersey, two breeds with weak genetic relationships and, generally, no improvements are observed in the accuracies of GEBV for HF when Jersey animals are added to the reference populations, and for Jersey animals results are similar or worse when using 54k data and GBLUP methods (Hayes *et al.,* 2009; Erbe *et al.,* 2012). However, when using denser SNP panels, functional subset of

markers or Bayesian methods, increases in accuracy for the Jersey have been observed when adding HF animals to the reference population (Erbe *et al.,* 2012). Also, Olson *et al.* (2012) studied the effect on reliabilities when combining Brown Swiss, Jersey and HF using single trait and multi-trait models showing that with the single trait model the GEBV reliabilities increased slightly for Brown Swiss but decreased for Jersey and HF. On the contrary, when the multi trait model was used the negative effects were not observed and a small positive effect was observed for Brown Swiss and HF.

Few studies are available involving beef cattle populations. In general, beef cattle has more breeds, but smaller populations than dairy cattle within a country. Weber *et al.* (2012) investigated the accuracy of genomic predictions for six growth and carcass traits for populations including many breeds (crossed animals). The study reported that genomic predictions using multi-breed reference populations were more accurate than those obtained using a single-breed reference population. On the other hand, Kachman *et al.* (2013) reported that, for breeds in the reference data, genomic predictions from multi-breed and single-breed reference populations had similar accuracies. Moreover, Chen *et al.* (2013a) studying residual feed intake in Canadian Angus and Charolais beef cattle populations, found that when there is weak relationship between reference and test animals the combined reference data increased accuracies 1-2% in Angus and 3-4% in Charolais. Finally, Bolormaa *et al.* (2013) assessed the accuracy of genomic predictions for 19 traits including feed efficiency, growth, carcass and meat quality traits in Australian beef cattle populations. Using a GBLUP model, the combined reference population performed better than a single-breed reference population with a 4% increase in the accuracy averaged over traits and breeds.

The main conclusion of these studies is that across-breed genomic evaluation can provide useful results, depending on the genetic divergence of the involved populations and the marker density. However, the usefulness of this approach for a particular set of population should be evaluated before its practical implementation.

Table 2. Increase in accuracy/reliability when using joint dairy reference compared to a single reference population for milk-, protein and fat yield, fertility and Somatic Cell Score (SCS). All studies are performed using 54 k genotype data. Ref1 is the breed and country of origin for the single reference population, and Ref2 is the breeds and countries of origin for the joint reference. Reference sizes are given as number of bulls (+number of cows). R or R^2 in column five states whether the original paper uses the correlation or squared correlation to measure the validation accuracy. Breed codes: HF=Holstein-Friesian, JE=Jersey, BS=Brown-Swiss, DR=Danish Red, SR=Swedish Red, FA=Finnish Ayrshire, NR=Norwegian Red, VR=Danish/Swedish/Finnish Red, MB=Montbéliarde, NM=Normande. Country Codes: US=United States, IT=Italy, CA=Canada, UK=United Kingdom, CH=Czech Republic, AT=Austria, DE=Germany, NL=Netherlands, FR=France, CI=China, NO=Nordic, AS=Australia. Trait codes: NRR=Non Return Rate, CR=Calving Rate, UHI=Udder Health Index, DPR=Daughter Pregnancy Rate, IFC=Interval between Calving and First insemination, FC=Fat Content.

Ref1	Ref2	Ref1 size	Ref2 size		Milk	Protein	Fat	Fertility	SCS	Method	Citation
HF (US)	HF (US+IT+CA+UK)	10,534+22,800	18,508+22,800	R^2	2.1	2.3	2.3	3.8^{DPR}	3.5	GBLUP	VanRaden et al. (2012)
BS (US)	BS (CH+DE+AT)	812+374	1682+374	R^2	5.3	2.7	1.1	-3^{DPR}	0.8	GBLUP	VanRaden et al. (2012)
HF (CA)	HF (US)	1,097	4,127	R^2	9	8	12	3	10	GBLUP	Schenkel et al. (2009)
HF (NO)	HF (NO+DE+FR+NL)	3,077	10,880	R^2		13		5^{NRR}	13	GBLUP	Lund et al. (2011)
HF (DE)	HF (NO+DE+FR+NL)	3,676	14,479	R^2		2		10^{NRR}	15	GBLUP	Lund et al. (2011)
HF (FR)	HF (NO+DE+FR+NL)	3,071	12,078	R^2		4		10^{CR}	8	QTL-BLUP	Lund et al. (2011)
HF (NL)	HF (NO+DE+FR+NL)	3,472	9,618	R^2		5		3^{IFC}	8	Bayesian 2-mixture	Lund et al. (2011)
HF (CI)	HF (CI+NO)	13+1,572	4,411+1,572	R^2	29	32	25			Multitrait GBLUP	Zhou et al. (2013)
HF (CI) cows	HF (CI+NO)	80+1,572	4,478+1,572	R^2	11	5	5			Multitrait GBLUP	Zhou et al. (2013)
DR	VR	929	3,735	R^2	2	4	1	-3^{NRR}	2^{UHI}	Bayesian	Brøndum et al. (2011)
SR	VR	1,551	3,735	R^2	9	18	7	9^{NRR}	6^{UHI}	Bayesian	Brøndum et al. (2011)
FA	VR	1,562	3,735	R^2	12	13	6	5^{NRR}	10^{UHI}	Bayesian	Brøndum et al. (2011)
VR	VR+NR	3,367	5,717	R	1	1	2	0^{NRR}	2^{UHI}	GBLUP	Zhou et al. (2014a)
NR	VR+NR	2,076	5,433	R	5	8	5	2^{NRR}		GBLUP	Zhou et al. (2014a)
VR	VR+HF (NO)	3,437	6,552	R	1.4	1.1	1.0	0.4^{NRR}	0.4	GBLUP	Zhou et al. (2014b)
DR	VR+HF (NO)	3,437	6,552	R	5	3	2	2^{NRR}	1	GBLUP	Zhou et al. (2014b)
SR	VR+HF (NO)	3,437	6,552	R	2	2	2	0^{NRR}	0	GBLUP	Zhou et al. (2014b)
FA	VR+HF (NO)	3,437	6,552	R	1	0	0	0^{NRR}	0	GBLUP	Zhou et al. (2014b)
HF (NO)	VR+HF (NO)	3,115	6,552	R	0.6	0	0.4	-0.4^{NRR}	0.4	GBLUP	Zhou et al. (2014b)
MB	MB+NM+HF (FR)	950	4,896	R^2	2		6^{FC}	0^{CR}		GBLUP	Karoui et al. (2012)
NM	MB+NM+HF (FR)	970	4,896	R^2	2		0^{FC}	0^{CR}		GBLUP	Karoui et al. (2012)
HF (FR)	MB+NM+HF (FR)	2,976	4,896	R^2	1		1^{FC}	0^{CR}		GBLUP	Karoui et al. (2012)
BS (US)	BS+JE+HF (US)	506	7,168	R^2	4	3	4	-1^{DPR}	-1	Non-linear GBLUP	Olson et al. (2012)
JE (US)	BS+JE+HF (US)	1,361	7,168	R^2	-3	-2	-4	0^{DPR}	0	Non-linear GBLUP	Olson et al. (2012)
HF (US)	BS+JE+HF (US)	5,331	7,168	R^2	-4	-3	-3	0^{DPR}	0	Non-linear GBLUP	Olson et al. (2012)
HF (AS)	HF+JE (AS)	1,897	2,351	R	-1	0	0			GBLUP	Erbe et al. (2012)
JE (AS)	HF+JE (AS)	454	2,351	R	-3	-2	-3			GBLUP	Erbe et al. (2012)

(Lund et al., 2014)

OBJECTIVES

The main objective of this thesis is to evaluate the efficiency of the potential implementation of Genomic Selection in seven Autochthonous Spanish beef cattle populations (*Asturiana de los Valles, Avileña-Negra Ibérica, Bruna dels Pirineus, Morucha, Pirenaica, Retinta and Rubia Gallega*).

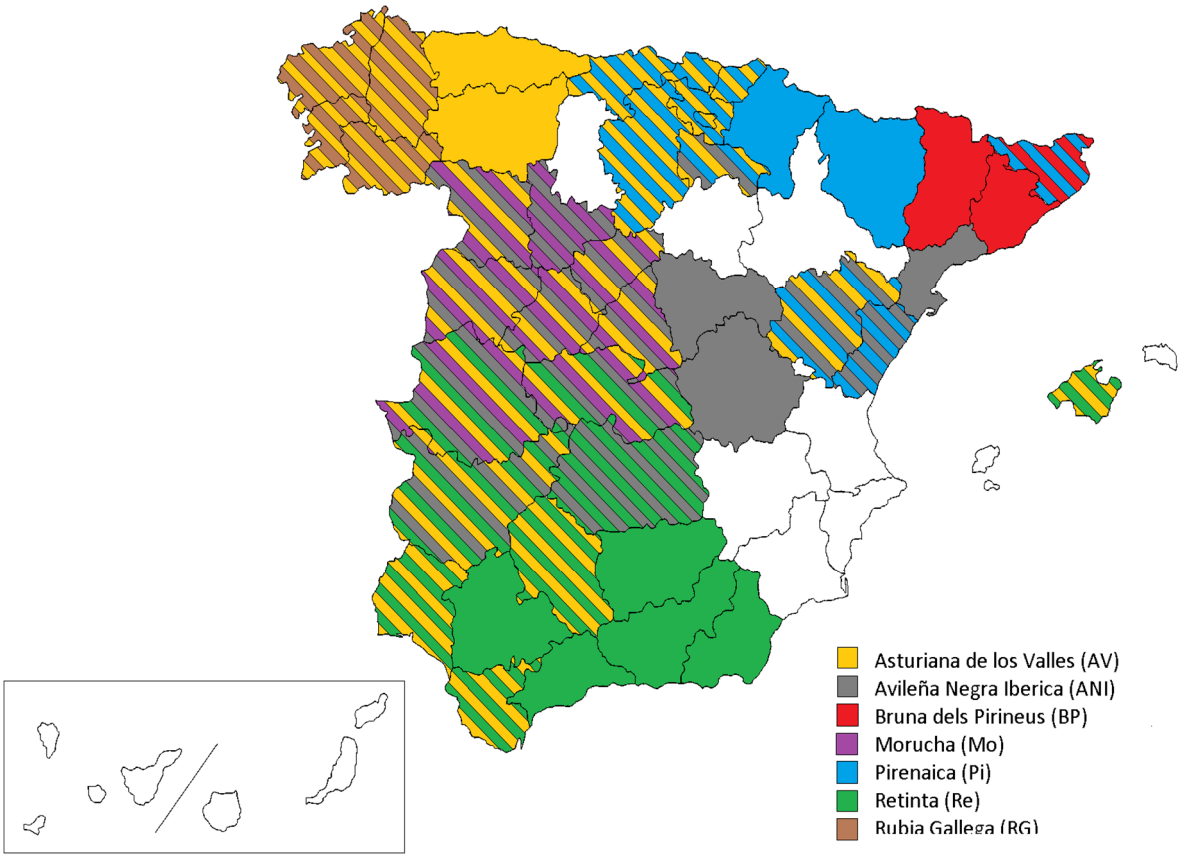
This main objective can be disentangled in the following secondary goals:

- Evaluate the potential improvement of the Single Step Genomic Selection within the genealogical and productive structure of the Spanish beef cattle populations.
- Evaluate the ability of prediction across populations within the Spanish beef cattle populations.
- Calculate the persistency of the haplotype phase across populations and evaluate its potential use for genomic selection across populations.
- Study the haplotype diversity along the genome for the Spanish beef cattle populations.

MATERIALS

The biological material used in this study was generated during the development of the AGL2010-15903 project. Biological samples of 171 triplets (sire/dam/offspring) were collected from 7 Spanish beef cattle populations (*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; *Rubia Gallega* -RG-, n=24), whose geographic distribution within the Iberian peninsula is presented in Figure 1.

Figure 1. *Geographic distribution map of the Spanish breeds included in this study.*



The animals were selected under the criteria of minimizing the genealogical relationship between them in order to capture as much of the variability as possible in each population. The individual samples were obtained through blood extraction from the caudal vein and were stored in tubes with EDTA anticoagulant according to the recommendations of Morton *et al.* (1993). Once the samples were collected they were processed according to the protocol described in *Prefiler™ Forensic DNA Kit of Applied Biosystems*, using *Mag-Max™ Express 96-Magnetic Particle Processor automated equipment.*

These samples were genotyped in a commercial laboratory (*Xenética Fontao, Lugo, España*) with the *Illumina BovineHD BeadChip* (Illumina, 2012) that contains 777,962 SNP markers. Further, the assembly of the genome was done using the *Bovine UMD3.1* database (Zimin *et al.*, 2009).

Additionally, the SNPs that were located on the autosomal chromosomes were filtered and those found in repetitive positions were excluded. During the filtering the following requirements were applied: 1) a Mendelian error inferior than 0.05 and 2) SNP and individual call rates higher than 95%. The quality control was performed using the PLINK software (Purcell *et al.*, 2007). At the end of the filtering process there were 703,707 SNP markers covering 2,510,350 kilobases (kb) of the autosomal chromosomes with a mean density of one marker per 3.57 kb. A more detailed description in presented in Table 3. The reconstruction of the parental haplotypes was conducted with the software *Beagle* (Browning and Browning, 2009) using the option of triplets (individual/sire/dam) analysis.

Table 3. *Distribution of the SNP markers along the autosomal chromosomes.*

BTA	Number of SNP	BTA	Number of SNP	BTA	Number of SNP
1	44,495	11	30,843	21	20,294
2	38,516	12	24,663	22	17,270
3	34,199	13	22,842	23	14,364
4	33,366	14	23,905	24	17,943
5	33,301	15	23,399	25	12,441
6	34,046	16	23,154	26	14,568
7	31,605	17	21,131	27	12,545
8	32,384	18	18,558	28	12,378
9	29,633	19	18,167	29	13,924
10	29,157	20	20,616		

Finally, along with the genotypic data, the genealogical and phenotypic data on birth weight were available for two populations, *Pirenaica* and *Rubia Gallega*. The data for *Rubia Gallega* comprised of 92,046 individuals in the genealogy and 64,030 birth weight data. The systematic effects considered for this trait were 1) sex with 2 levels, 2) age of mother with 16 levels and 3) herd-year-season (HYS) with 10,160 levels. Likewise, the data for *Pirenaica* included 55,203 individuals in the genealogy and 32,702 birth weight data. The systematic effects were the same with 2, 16 and 5,343 levels respectively.

CHAPTER 1

Performance of Genomic Selection under a single-step approach in Autochthonous Spanish beef cattle populations.

Introduction

The Genomic Selection (GS) methodology (Meuwissen *et al.,* 2001) has already been shown to be a promising development for animal breeding. In fact, the dairy cattle industry was quick to incorporate it in their selection schemes to produce highly accurate Genomic Breeding Values –GEBVs- (Hayes *et al.,* 2009a; Loberg and Durr, 2009; VanRaden *et al.,* 2009) and pig companies have started to used it regularly in elite populations (Forni *et al.*, 2011; Ostersen *et al.,* 2011).

However, the beef cattle industry has been more reluctant in the implementation of this technology due to several reasons (Berry *et al.,* 2016). On one hand, most of the beef cattle populations have a limited census, and on the other, the use of artificial insemination (AI) is lower than in dairy cattle. These phenomena restrict the presence of sires evaluated with high accuracy, and contribute to the poor connectedness among and within populations. As a consequence, the usual dairy cattle strategy to evaluate very young bulls, as an alternative of progeny testing (Hayes *et al.,* 2009a), cannot be automatically mimicked. Thus, the potential efficiency of the implementation of GS should be specifically tested in each population.

The first attempts to implement GS (Meuwissen *et al.,* 2001) suggested a two-step approach that involved training and testing populations. Later on, Habier *et al.* (2007) probe that the standard mixed model equations (Hendenson, 1984) can be easily adapted to incorporate genomic information through a genomic relationship matrix (**G**) and lead to predictions of GEBVs equivalent to the Gaussian Regularization proposed by Meuwissen *et al.* (2001). Further, Legarra *et al.* (2009)

and Aguilar *et al.* (2010) developed an extension of this model denoted as Single-Step GBLUP, which allows to predict at the same time the breeding values for genotyped and non-genotyped individual.

This approach could be useful for populations that cannot support a broad genotyping effort, such as the Autochthonous Spanish beef cattle populations. Thus, the objective of this study is to investigate the potential application of genomic selection under a single-step approach in two Spanish Autochthonous populations (*Pirenaica*, - Pi -, and *Rubia Gallega*, - RG -), as representatives of alternative genealogical structures due to the wide utilization of AI in RG, in contrast to Pi.

Materials and Methods

Simulation

An historical population of 100 individuals that evolved under random mating for 500 generations was simulated. The genome simulated comprised of 30 chromosomes with 2,000 markers each, from which 100 were randomly selected as causative mutations (QTLs). The mutation rate for both markers and causative mutations was fixed at 2.5×10^{-3}. These parameters were chosen in order to obtain genotypes of around 50,000 (50k) neutral markers mimicking the information provided by the *BovineSNP50 BeadChip* (Gunderson, *et al.,* 2005; Steemers *et al.,* 2006). The population of the last generation was used as the base population for the simulation over the available pedigree (see Material chapter) by gene-dropping providing simulated genotypes for markers and QTL for the pseudo-populations with the same genealogical structure as the real populations. Further, the QTL effects were drawn from a Gaussian distribution with mean zero and variance one.

After the simulation of the genotypes for all the individuals in the pedigree, pseudo-phenotypes were simulated for each individual that had a recorded phenotype on the real data set. Phenotypes were simulated by summing a general mean (1,000), the effects for the QTLs weighted by their specific genotype and a residual drawn from a Gaussian distribution with zero mean and a variance adequate to create two traits with heritability 0.1 and 0.4. Finally, the breeding values and the genotypes for all individuals in the pedigree were recorded.

Single step

The data provided by the simulation study were analysed by the standard BLUP analysis (Henderson, 1984) and by the single-step GBLUP –ssGBLUP- (Aguilar *et al.,* 2010). Both analyses were performed by the BLUPf90 program family (Misztal *et al.,* 2014).

The model used for all analyses was:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z_1 m} + \boldsymbol{Z_2 u} + \boldsymbol{e} ,$$

Where $\boldsymbol{y}$ is the vector of phenotypes, $\boldsymbol{b}$ is the vector of the systematic effects, sex with 2 levels and age of dam with 16 levels**,** $\boldsymbol{m}$ is the vector of the Herd-Year-Season random effect with 5,343 and 10,160 levels for Pi and RG respectively, $\boldsymbol{u}$ is the vector of additive genetic effects, and $\boldsymbol{e}$ is the vector of residuals. $\boldsymbol{X}$**,** $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ are the incidence matrices for $\boldsymbol{b}$**,** $\boldsymbol{m}$ and $\boldsymbol{u}$**,** respectively.

The only difference between ssGBLUP and BLUP is that in ssGBLUP the inverse of the numerator relationship matrix $\mathbf{A^{-1}}$ is replaced by matrix $\mathbf{H^{-1}}$ defined as:

$$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G^{-1}} - \boldsymbol{A_{22}^{-1}} \end{bmatrix},$$

Where $\boldsymbol{G}$ is the genomic relationship matrix and $\boldsymbol{A_{22}}$ is the numerator relationship matrix for the genotyped individuals.

The default parameter options of the BLUPF90 software such as minor allele frequency of 0.05, individual and SNP call rate of 0.90 and H matrix scaling parameters (α=0.05 and β=0.95) were used in all cases. In addition, variance components were assumed to be known.

Validation

The procedures were validated through the accuracy of the predictions that was calculated as the Pearson correlation between the estimated breeding values and the simulated breeding values for the individuals born in last available year (2014) that served as candidates to selection (579 for Pi and 1,738 for RG).

Simulation Scenarios

First, we developed a base scenario of simulation where we evaluated the accuracy of the procedure with respect to the following variables:

1) Heritability of the trait (h^2 = 0.1 and 0.4)
2) Number of historical individuals genotyped (4,000, 2,000, 1,000, 500 and 250)
3) Genotypes for sires and dams of candidates to selection (Yes or No)
4) Genotypes for the proper candidates to selection (Yes or No).
5) Phenotypic records for the candidates to selection (Yes or No).

Thus, the number of cases of simulation was 80, plus 4 cases of standard BLUP evaluation (two heritabilities with and without phenotypic records for the candidates to selection). In this study, the historical individuals were selected according to the estimated prediction error variance –PEV- achieved from a standard BLUP evaluation. Thus, the individuals were ranked according to their PEV, and the bottom 4,000, 2,000, 1,000, 500 and 250 historical individuals were selected to be genotyped, regarding the case of simulation.

In addition, we performed a sensitivity analysis by comparing the results of this base scenario with some other alternatives. These alternatives included

1) Replacing the Top Historical (TH) individuals with the individuals with lower PEV, but born exclusively from 2010 to 2013 (Top Recent – TR).-
2) Replacing the Top Historical individuals with a random sample between 2010 to 2013 (Random Recent – RR-)
3) Three combinations of the RR and TH strategies that included one quarter, one half or three quarters of TH individuals combined with RR individuals.
4) Five alternative marker densities including: 4,000 (4k), 10,000 (10k), 23,000 (23k), 100,000 (100k) and 200,000 (200k) neutral markers.
5) Two alternative effective population sizes (Ne) in the simulation of the historical population (50 and 200)
6) Two alternative mutation rates for markers and QTL (1×10^{-3} and 4×10^{-3}.)

Results and Discussion

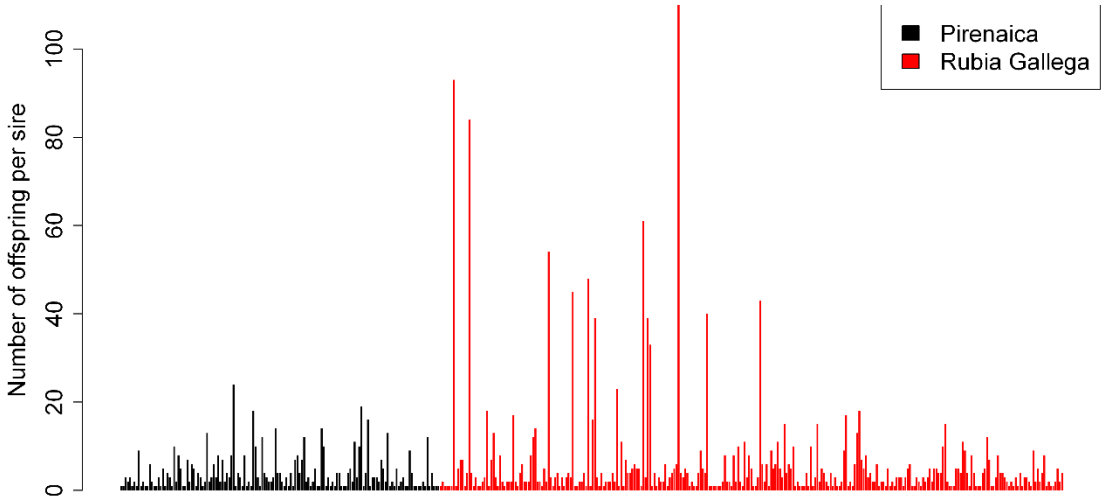
The datasets of the two populations used in this study differ significantly in their structure as it can be seen in Table 4. RG uses significantly more artificial insemination than Pi, as it is reflected on the pedigree structure. Thus, the number

of sires used for reproduction represents the 1.81% (1,669 animals) for RG while they represent the 5.45% (3,010) for the Pi. Moreover, the average number of offspring per sire is 47.82 (s.d. = 225) for RG and just 17.21 (s.d. = 39.64) for Pi. To reinforce this statement, Figure 2 shows the number of offspring born in the year 2014 per sire for both populations. In RG 283 sires have 1738 offspring (average 6.14 offspring per sire) while in Pi 145 sires have 579 (average 3.99 offspring per sire).

Table 4. *Comparison of the pedigree structures between the Rubia Gallega (RG) and Pirenaica (Pi) populations.*

	RG	**Pi**
№ of animals	92,046	55,203
№ of generations	16	25
Total № of sires	25,678	18,837
- With offspring	1,669	3,010
- Mean	47.82	17.21
(s.d.)	(225)	(39.64)
Total № of dams	66,368	36,366
- With offspring	35,156	23,373
- Mean	2.27	2.24
(s.d.)	(1.68)	(1.8)

Figure 2. *Number of offspring born in the year 2014 per sire.*



Standard BLUP evaluation

First, we evaluated the performance of the standard BLUP evaluation in each population and for each trait in order to define a reference point to compare the results of the alternative genotyping strategies. The results are presented in Table 5.

Table 5. *Accuracies (s.e) obtained from the BLUP evaluation.*

	Trait A h^2=0.4		**Trait B h^2=0.1**	
	Without 2014 data	With 2014 data	Without 2014 data	With 2014 data
Pi	0.554	0.724	0.446	0.515
	(0.011)	(0.004)	(0.013)	(0.011)
RG	0.550	0.727	0.479	0.549
	(0.010)	(0.004)	(0.012)	(0.007)

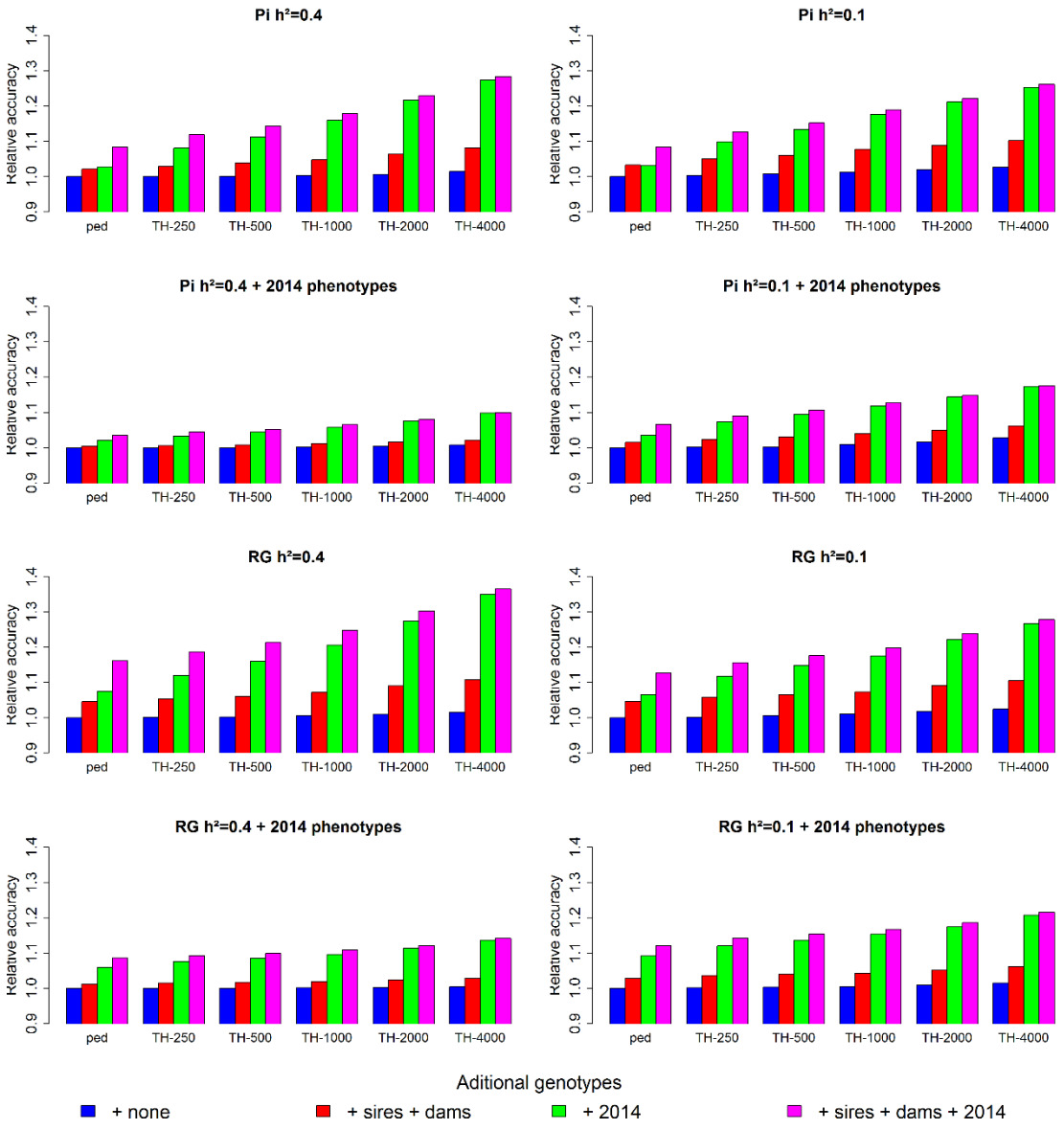
It can be observed that the accuracy of prediction for the individuals born in 2014 was very similar between populations when the heritability is higher (h^2=0.4), but there are remarkable differences between them for the low heritability (h^2=0.10).

The reason for this difference can be attributed to the genealogical structure of the RG population, due to the higher presence of IA that implies a higher accuracy of the prediction of breeding values of sires. As a consequence, this higher accuracy of the sires is reflected on their sons and daughters. This effect is more evident with lower heritability, because more progeny is needed to achieve a higher accuracy (Falconer and McKay, 1996). In addition, and as it was expected, the accuracy of the cases of simulation that included the phenotypes of the candidates to selection is higher. Finally, as it was also expected, the accuracy is higher for the scenarios with h^2=0.4 than with h^2=0.1.

Base Scenario

The detailed results of the accuracy of prediction for all cases of simulation are presented in Tables 1.1 to 1.4 of the ANNEXE 1. Moreover, in Figure 3 is presented a summary of the relative accuracy with respect to the standard BLUP procedure.

Figure 3. *Relative accuracy with respect to the standard BLUP procedure for the different alternatives of the base scenario.*



Relative accuracy with respect to standard BLUP evaluation. RG=Rubia Gallega, Pi=Pirenaica, ped=standard BLUP evaluation, TH-250=250 Top Historical genotypes, TH-500=500 Top Historical genotypes, TH-1000=1000 Top Historical genotypes, TH-2000=2000 Top Historical genotypes, TH-4000=4000 Top Historical genotypes, none=no additional genotypes, sires+dams=parents of the selection candidates, 2014=selection candidates.

One on hand, it can be observed that the accuracy of candidates to selection is always higher than the one provided by the standard BLUP procedure. This is a very important advantage of single-step (Aguilar et al., 2010) approach with respect to

the two-step approach of genomic selection (Meuwissen *et al.,* 2001) that requires a minimum number of genotyped and phenotyped individuals to compete with the pedigree-based approaches. Thus, the appropriateness of the single-step approach for populations that cannot afford huge genotyping efforts, like the Spanish autochthonous beef cattle breeds, is very clear.

As expected, the increase of accuracy is higher as the number of genotyped individuals increases. However, it should be noted that this gain could be only worthy when the candidates to selection are genotyped. In fact, the maximum gain obtained without genotyping the candidates to selection and their parents was just 2.7 ± 0.4 % (Pi, 4000 genotyped individuals, h^2=0.1) and this figure only increased up to 10.2 ± 1.0 % when genotypes of the sires and dams of the candidates to selection were added. On the contrary, and for the same scenario, the increase of accuracy goes up to 25.3 ± 2.2 % (4000 TH + candidates to selection genotyped) and 26.1 ± 2.2 % (4000 TH + sires and dams + candidates to selection). It should be noted that the sires and dams of the candidates to selection are frequently included within the TH individuals. So, only slight differences between both strategies were found. Genotyping of all candidate individuals could be an important effort for the breeders associations, although it is important to mention that the imputation technics work very efficiently (Khatkar *et al.,* 2012; Mulder *et al.,* 2012) even with low-density devices. Thus, a genotyping strategy that uses low density chips for candidate individuals can be appropriate.

Further, if we compare the performance of the method with respect to the heritability of the trait, it can be observed that the rate of increase of accuracy is higher for h^2=0.4 than for h^2=0.1 when the candidates to selection are not phenotyped. This

means, that the number of genotyped individuals required for traits with lower heritability is greater, because of the lower information provided by the phenotypes when heritability is low. The well-established strategies of genomic selection in dairy cattle (Hayes *et al.,* 2009a) involve the evaluation of genomic breeding values on sires with extremely high accuracies, overcoming the informativeness of each individual phenotype by averaging over a huge number of daughters. This strategy cannot be replicated with the population structure of smaller populations. However, this phenomenon is reverted when the candidates to selection are phenotyped, and the increase of accuracy is higher for the cases of simulation that involved a lower heritability. The cause of this difference can be attributed to scale effects due to the higher base accuracy for phenotyped individuals with a moderate or high heritability (h^2=0.4).

Moreover, and as it was expected, the increase of accuracy is much higher when the candidates to selection are not phenoyped. This specific scenario tries to represent traits that are measured late in life (e.g. maternal traits) or difficult and expensive to measure (e.g. Slaughter traits, Disease resistance). As an example, the maximum increase of accuracy for non-phenotyped candidates to selection was 36.5 ± 1.7 % (RG, 4000 TH, + sires and dams + candidates to selection, h^2=0.4) whereas in the same scenario but with own phenotype recorded, the percentage of increase was just 14.1 ± 0.4 %. This result confirms the appropriateness of the GS for traits that cannot be easily measured on the candidates to selection.

Finally, it is relevant to note that the increase of accuracy is higher for RG than for Pi, even when both populations start with the same base level of accuracy (standard BLUP, h^2=0.4). As before, the cause of this difference must be attributed to the

genealogical structure of the RG population due to the higher application of AI that provides a relevant number of individuals evaluated with very high accuracy. This highly accurate individuals contribute to the increase of the accuracy of the remaining individuals through the genomic relationship matrix, which can be considered as an improved estimate of the true genetic relationship between individuals, based on SNP markers instead of only on the pedigree information (Legarra *et al.,* 2014a) and by the detection of older relationships hidden in the pedigree information.

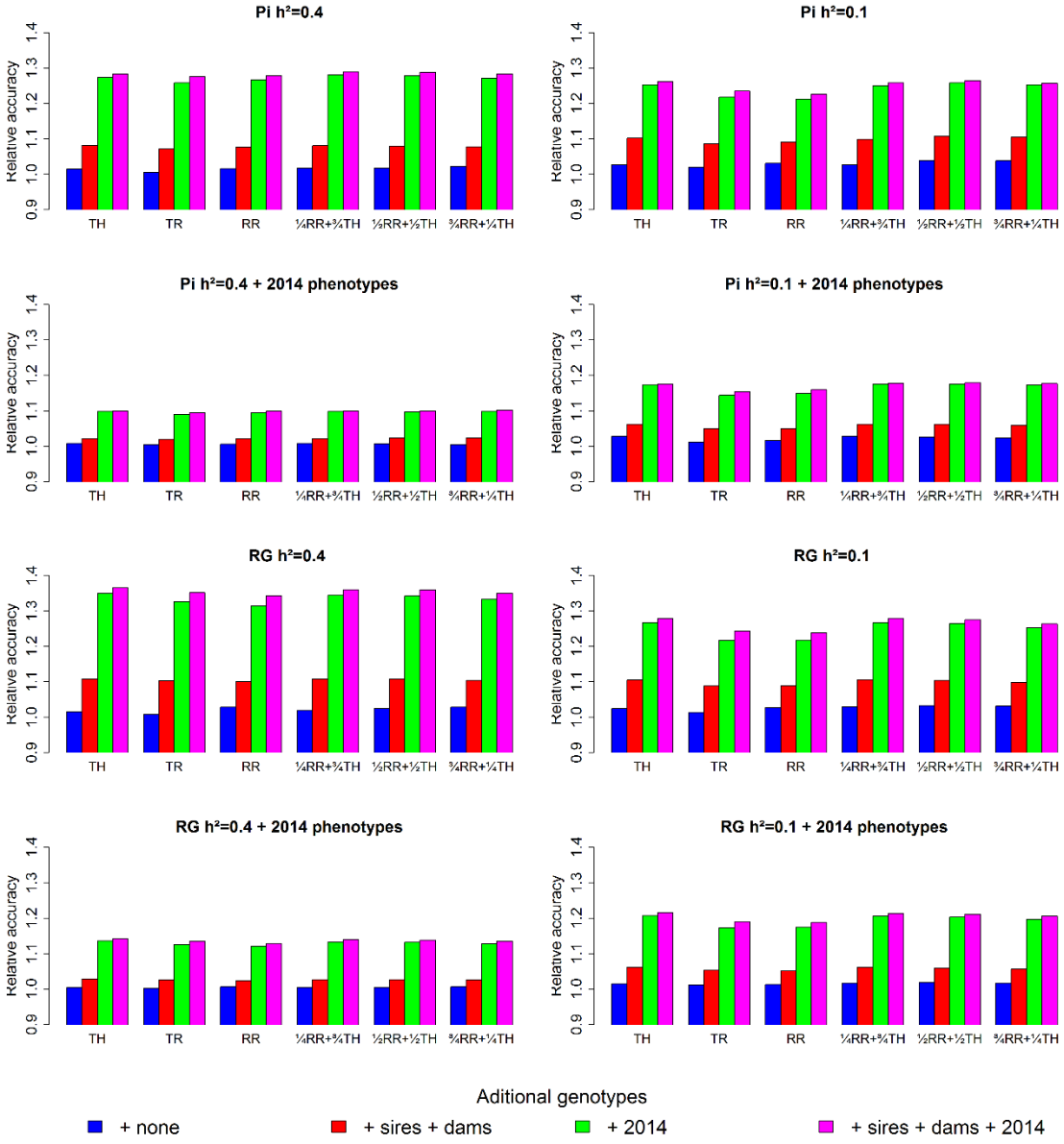
Sensitivity analysis

The results of the base scenario analysis covered a wide range of variables. However, it should be noted that they are conditioned to a set of predefined simulation parameters. Thus, and in order to extract more general conclusion we performed a sensitivity analysis with respect to the following variables: 1) the method of choice of genotyped individuals, 2) the marker density, 3) the effective population size along the evolutionary history of the population and 4) the mutation rates for QTL and SNP markers.

In first place, the results of the sensitivity analysis with respect to the method of choice of the genotyped individuals are presented in Figure 4. As it can be observed, there are no relevant differences in accuracy with respect to the method of choice of genotyped individuals when compared with the election of the TH individuals. It can be only noted a slight reduction of accuracy for lower heritabilities (h^2=0.1) when the TH individuals are replaced by RT and RR, that disappear when just one quarter of TH individuals were included in the genotyped subset. The consequences of this result implies that, although the most informative individuals (with lower PEV)

provide a better accuracy, the results are robust enough to suboptimal genotyping strategies forced by the availability of biological samples of older individuals.

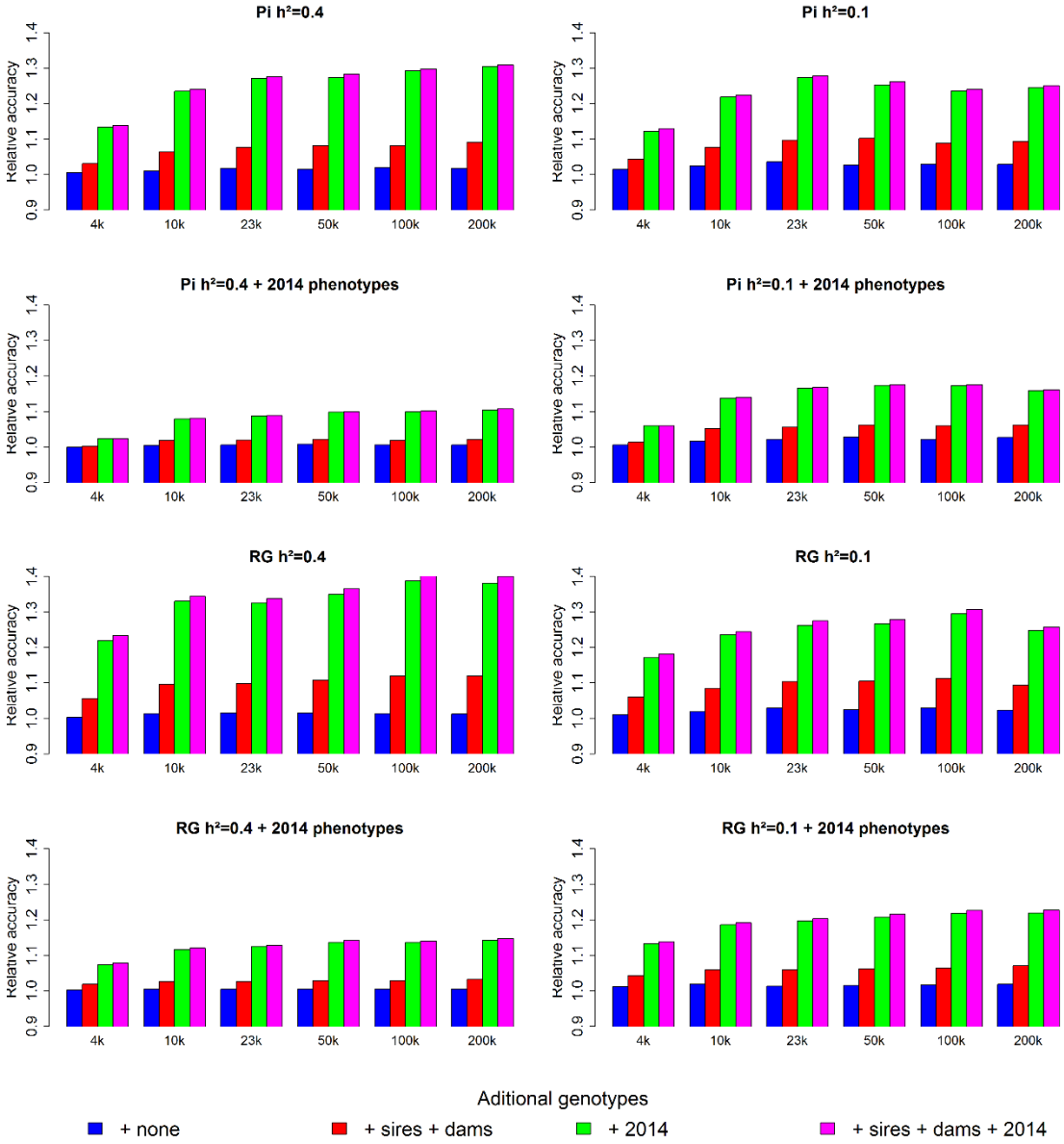
Figure 4. *Sensitivity analysis with respect to the genotyping strategies.*



Relative accuracy with respect to standard BLUP evaluation. RG=Rubia Gallega, Pi=Pirenaica, TH=Top Historical, TR=Top Recent, RR=Random Recent, size of reference sets=4,000, none=no additional genotypes, sires+dams=parents of the selection candidates, 2014=selection candidates.

The second sensitivity analysis was focused on the marker density, as the base simulation scenario tries to represent the density that can be obtained by the *BovineSNP50 BeadChip*. The results are presented in Figure 5.

Figure 5. *Sensitivity analysis with respect to the marker density.*

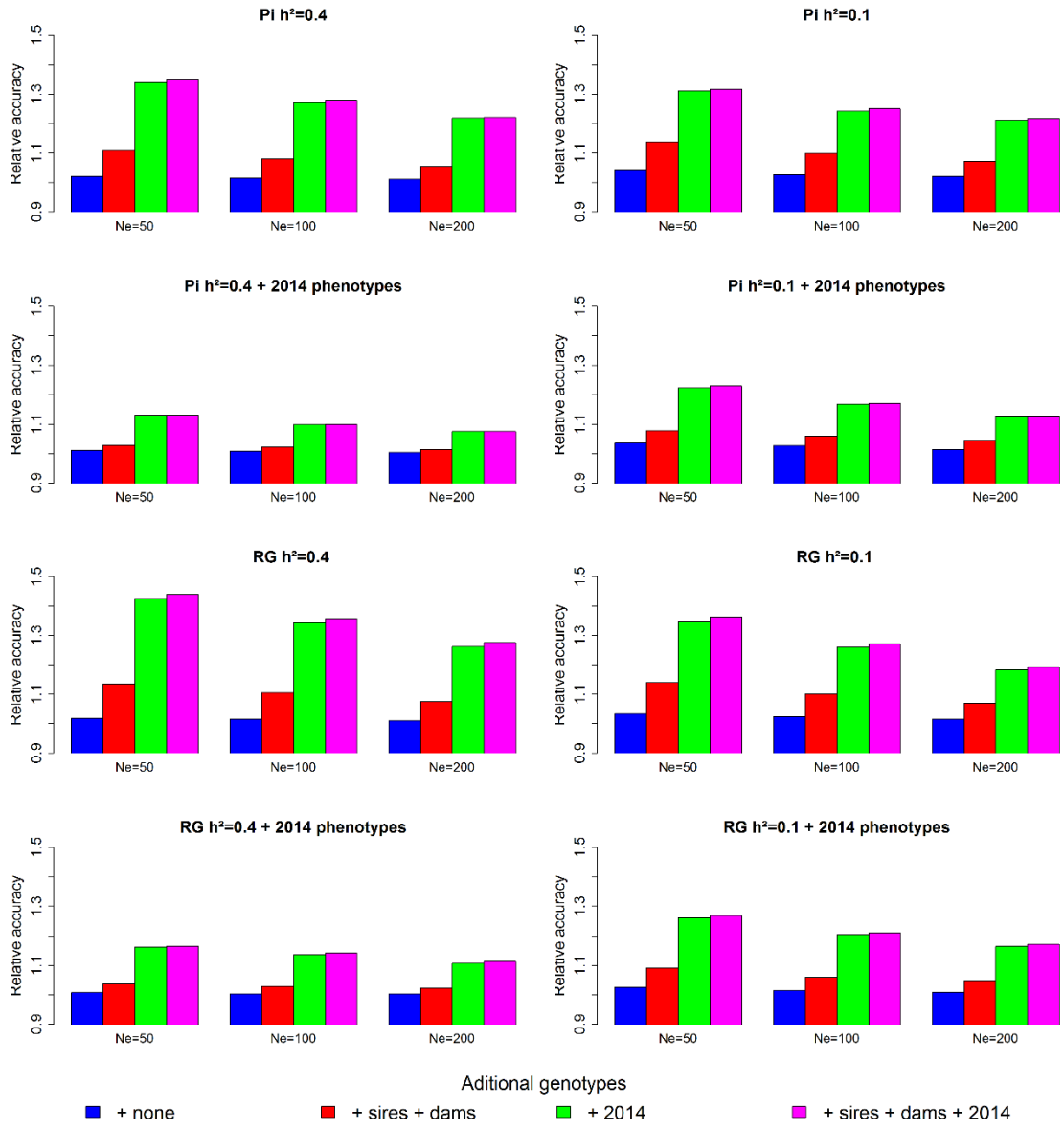


Relative accuracy with respect to standard BLUP evaluation. Genotype set used=TH 4,000 genotypes, RG=Rubia Gallega, Pi=Pirenaica, 4k=4,000 SNPs, 10k=10,000 SNPs, 23k=23,000 SNPs, 50k=50,000 SNPs, 100k=100,000 SNPs, 200k=200,000 SNPs, none=no additional genotypes, sires+dams=parents of the selection candidates, 2014=selection candidates.

The main conclusion of this analysis is that the accuracy of GS increases with marker density but it reaches a plateau around 50k, and further increases of accuracy are not obtained for higher densities. This result confirms the postulates of Cañas-Álvarez *et al.* (2016) that suggest that the Spanish autochthonous beef cattle populations need at least 38,000 segregating SNP markers. Thus, the potential increase that can be obtained from higher densities can be considered as negligible as also suggested by Solberg *et al.* (2008), even for unrelated individuals (Meuwissen, 2009).

Further, the results of the sensitivity analysis with respect to effective size of the evolutionary historical population are presented in Figure 6. As it can be observed, there is a reduction in accuracy as the Ne increases as predicted by Solberg *et al.* (2008). These authors proposed that equivalent accuracies can be obtained as a function of Ne x L (number of markers). Thus, doubling or halving the effective size implies that double or half of the markers are needed to achieve the same accuracy. However, the Ne of the Spanish autochthonous populations has been estimated around 50-60 (Cañas-Álvarez *et al.,* 2016) and, as a consequence, the results of the base simulation study presented earlier, can be considered as a conservative estimation of the potential increase of accuracy.

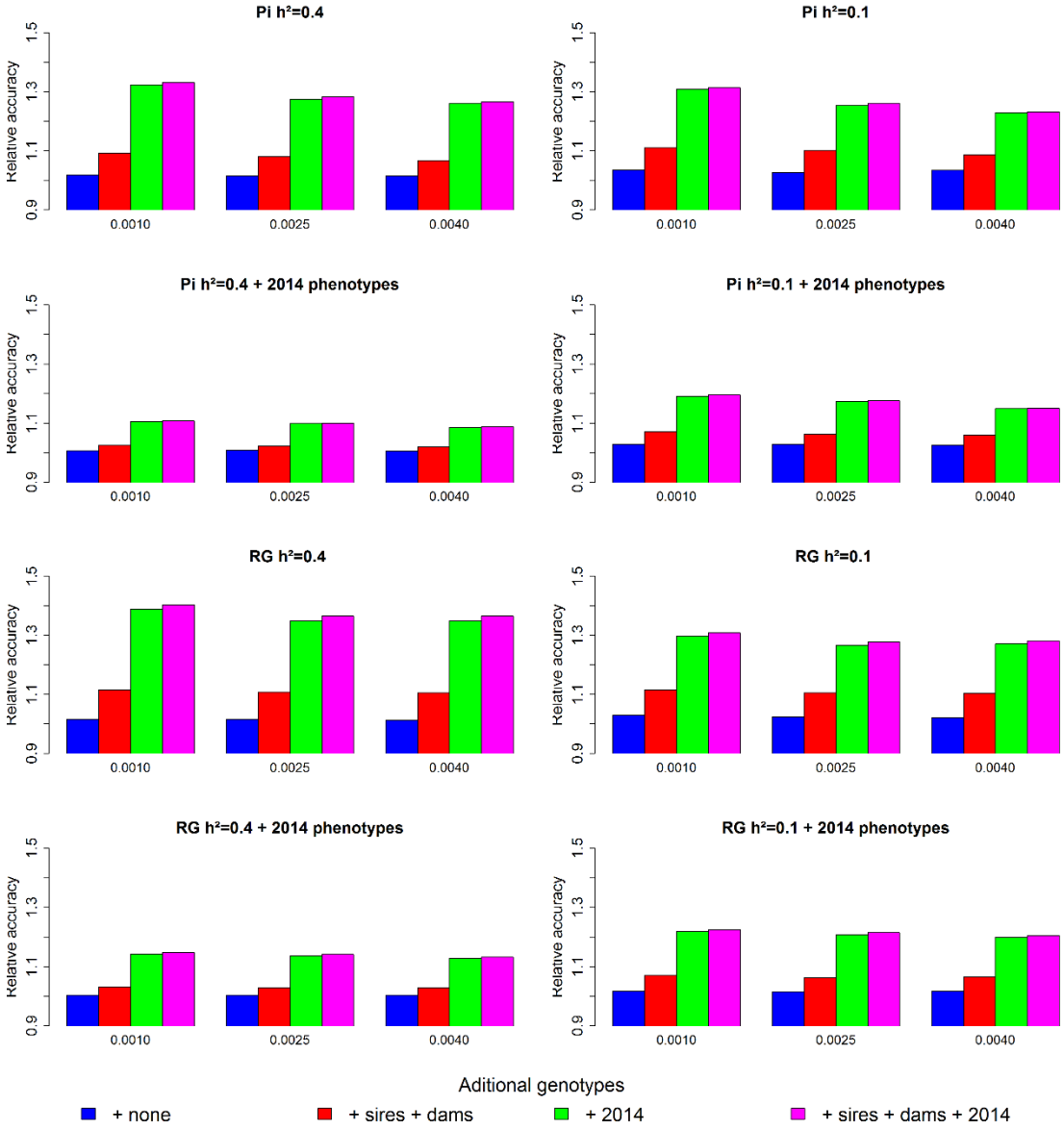
Figure 6. *Sensitivity analysis with respect to effective population size.*



Relative accuracy with respect to standard BLUP evaluation. Genotype set used=TH 4,000 genotypes, RG=Rubia Gallega, Pi=Pirenaica, Ne=50, Ne=100, Ne=200, none=no additional genotypes, sires+dams=parents of the selection candidates, 2014=selection candidates.

Finally, the last sensitivity analysis was devoted to mutation rate. The results are presented in Figure 7.

Figure 7. *Sensitivity analysis with respect to mutation rate.*



Relative accuracy with respect to standard BLUP evaluation. Genotype set used=TH 4,000 genotypes, RG=Rubia Gallega, Pi=Pirenaica, mutation rates tested = 1×10^{-3}, 2.5×10^{-3}, 4×10^{-3}, none=no additional genotypes, sires+dams=parents of the selection candidates, 2014=selection candidates.

The results showed only small differences in the accuracy when the mutation rate varied. However, there was a clear tendency to produce higher accuracies for lower

mutation rates. The reason for those differences can be attributed to the fact that higher mutation rates provide lower LD between SNP markers and QTL. However, the assumed mutation rates were extremely high with respect to estimations in the literature (Kumar and Subramanian, 2002; Hodgkinson and Eyre-Walker, 2011), and, as before, this result ensures that the output of our base simulation study consists of a conservative estimation of the potential increase of accuracy that can be achieved with GS in the Spanish autochthonous beef cattle populations.

Conclusions

The results of this study probe the appropriateness of the implementation of GS in the Spanish autochthonous populations, even though the genotyping efforts that can be achieved by the breeders association are intermediate or low. This advance can be obtained thanks to the single-step genomic selection approach (Legarra *et al.,* 2009; Aguilar *et al.,* 2010) that combines genomic and pedigree based relationships into the same relationship matrix. Thus, the pedigree based relationship matrix sets the lower bound of accuracy, and it is improved as more individuals with genotypes are incorporated into the genomic evaluation. As expected, the GS approach has been found more relevant for traits with low heritability or without own phenotypic information for the candidates to selection, and only when the candidates to selection are genotyped.

Finally, it is important to mention that the efficiency of GS is higher in RG than in Pi, because of the genealogical structure that is provided by the wider implantation of IA. So, a parallel increase of the rate of AI along with the genotyping efforts will lead to a greater success of GS in populations with a low percentage of AI.

CHAPTER 2

Evaluation of the potential use of a meta-population for Genomic Selection in the Autochthonous Spanish beef cattle populations.

Introduction

The advances in the area of molecular genetics have allowed the development of genotyping SNP chips that provide information throughout the genome (Gunderson, *et al.,* 2005). Along with the molecular advances, new statistical methods have been developed with the purpose of predicting the genomic breeding values of candidates to selection (Meuwissen *et al.,* 2001). The potential applications of these methods have been tested through simulation (Meuwissen *et al.,* 2001) and through cross-validation techniques in different species such as mice (Legarra *et al.,* 2008), dairy cattle (Luan *et al.,* 2009), aquaculture (Sonesson and Meuwissen, 2009) and poultry (González-Recio *et al.,* 2009).

Currently, genomic selection is a reality in dairy cattle (Hayes *et al.,* 2009a). Nevertheless, the implementation of such methods in the beef cattle industry is still questionable. The main limitations are the limited census of the beef populations, the great variability of the production systems, the narrower use of artificial insemination and the lower quality of phenotypic recording (Berry *et al.,* 2016).

To overcome these constraints, several authors (De Roos *et al.,* 2009; Toosi *et al.,* 2010, Kizilkaya *et al.,* 2010) have made efforts to increase the precision of the genomic predictions, in simulation studies, by using phenotypic and genomic information provided by several populations. Their results indicate that the use of a combined population is more helpful when the populations involved have diverged for a small number of generations, for populations of reduced size, and for traits of low heritability if high-density genotypes are available. However, with real data, some studies have obtained promising results (Weber *et al.,* 2012) whereas, some

others reported almost any advantage of multi-population genomic evaluation (Kachman *et al.,* 2013; Chen *et al.,* 2013a, Bolormaa *et al.,* 2013). Thus, the potential application of a meta-population approach should be studied in each specific case.

The Spanish autochthonous cattle breeds have a *Bos taurus* ancestral origin and it is estimated that they have a recent common origin (Beja-Pereira *et al.,* 2003; Cañas-Álvarez *et al.,* 2015) and with a quite important persistency of haplotype phase between them (Cañas-Álvarez *et al.,* 2016). These statements jointly with the small size of these populations and the limited economic resources available for genotyping suggest that these populations are good candidates for the use a meta-population for genomic selection. Thus, the objective of this study is to evaluate the efficiency of the potential implementation of multi-breed genomic selection in the Spanish beef cattle populations.

Material

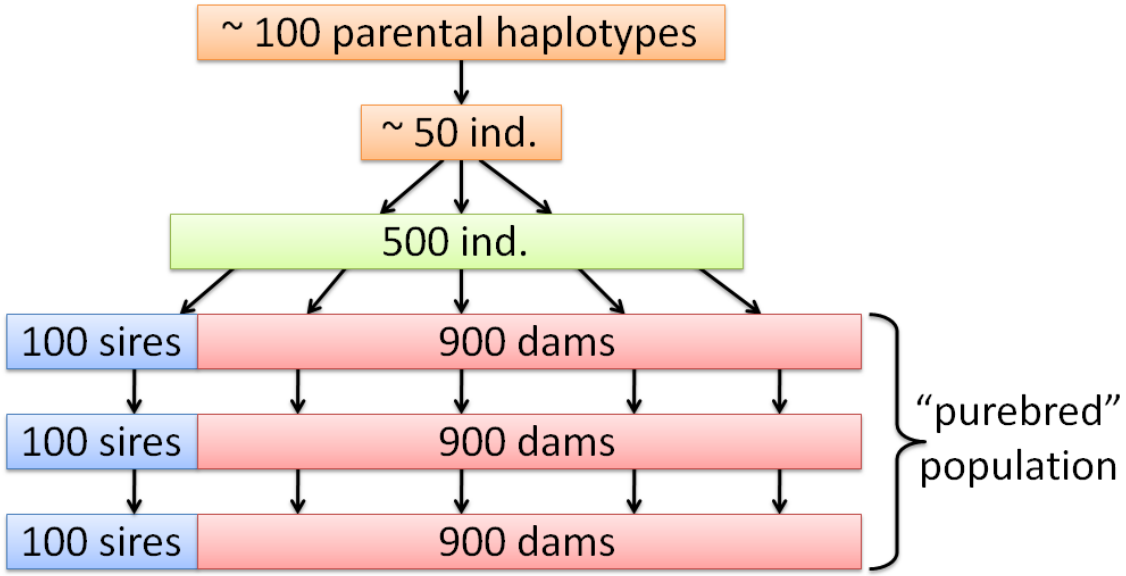
The data used in this study comprised of the genotypes with the *BovineHD Beadchip* for the 342 founder individuals of the triplets described in the MATERIAL chapter (Asturiana de los Valles – AV-, N=50, Avileña - Negra Ibérica – ANI-, N=48, Bruna dels Pirineus – BP-, N=50, Morucha –Mo-, N=50, Pirenaica –Pi-, N=48, Retinta – Re-, N=48 and Rubia Gallega –RG, N=48). Here, an additional quality control requirement applied was a minor allele frequency (MAF) of 0.01 resulting in 629,251 SNPs.

Methods

Simulation

The simulation structure tries to mimic the linkage disequilibrium structure of the analysed populations. Thus, for each breed, we defined a base population from the available paternal haplotypes. Thereinafter, for each population, the 629,251 SNP markers of the individuals of the first generation of 500 individuals were simulated by gene-dropping and assuming a map distance of 1 cM every Mb. The parents of this generation were selected randomly from the previous generation and ignoring their sex. Based on this generation, 3 more generations of 1,000 individuals (100 sires and 900 dams) were simulated with the same method, selecting the parents randomly but considering their sex this time. These three generations were used to establish the pseudo-populations for each initial population. A summary of this simulation structure is presented in Figure 8.

Figure 8. *Structure of the simulation strategy for the generation of pseudo-populations for each initial population.*



Further, in order to simulate the causative mutations of a trait, 3% of the SNP markers of each chromosome were randomly selected as QTLs, and they were attributed an additive effect sampled from a Gaussian distribution with zero mean and a standard deviation of one. Later on, for every individual, true genomic breeding values (TGBVs) were calculated as the sum of the effects of their genotype for the QTL polymorphisms. Moreover, phenotypes were simulated for all individuals summing to their TGBV, a trait mean (= 1,000) and a residual drawn from a Gaussian distribution with appropriate variance to generate two traits with heritability 0.4 and 0.1.

Additionally, and with the aim of defining a sensitivity analysis, 5 alternative scenarios of the genetic architecture of the quantitative traits were simulated beside the polygenic model followed above:

- 10G(20%): 10 randomly selected genes were added to the polygenic model which explained 20% of the total genetic variance.

- 4G(50%): 4 randomly selected genes were added to the polygenic model explaining 50% of the total genetic variance.

- Ex: the effects of the genes were drawn from an exponential distribution instead of a Gaussian distribution.

- LMAF: markers with extreme frequencies (rare variants) were chosen to simulate genes. (MAF≤0.05)

- 4MG: 4 QTLs were randomly selected to explain the 100% of the genetic variance with effects drawn from a normal distribution.

Genomic evaluation

The genomic evaluation was performed by means of *solveSNP* software (Legarra and Misztal, 2008) with the method RR BLUP (Meuwissen *et al.,* 2001) and under the following model.

$$y_i = \mu + \sum_{j=1}^{n} x_{ij} a_j + e_i$$

where y_i is the phenotype of the i^{th} individual, μ is the trait mean, n is the number of SNPs, x_{ij} is the genotype of the i^{th} individual for the j^{th} marker codified as 0,1 and 2, a_j is the substitution effect for the j^{th} marker and e_i is the residual effect of the i^{th} individual. Further, the prior distribution for the marker effects was the following multivariate Gaussian distribution:

$$\boldsymbol{a} \sim \boldsymbol{N}(0, \boldsymbol{I}\sigma_a^2)$$

where σ_a^2 is the marker variance whose prior distribution is assumed to be uniform within appropriate bounds.

The markers selected as causal mutations were excluded from the marker panel during the genomic evaluation. Later on, Genomic Breeding Values (GEBV) were calculated as:

$$GEBV_i = \sum_{j=1}^{n} x_{ij}\widehat{a_j}$$

Several scenarios of genomic evaluation were considered depending on the reference population used.

- Pure-bred: The reference population comprised of 3,000 individuals of one of the populations simulated. All seven populations were used as reference populations separately.

- Admixed ×2: The reference populations comprised of 3,000 (1,500 + 1,500) individuals from 2 pure-bred populations. All possible combinations were used as reference populations.

- Admixed ×7: One reference population comprised of 3,003 individuals with 429 individuals from each of the seven populations.

Additionally, reduced pure-bred populations of 1,500 and 429 individuals were used as reference populations with the goal of comparing them to the admixed scenarios under equal genotyping efforts within populations.

Validation

The populations used to validate the predictions included the 7 pure-bred populations and 3 additional generations of 1,000 individuals each for every population. The accuracy of the predictions was calculated as the Pearson correlation between the simulated breeding values and the predicted values. Each case of simulation was replicated 5 times and the results were averaged.

Results and Discussion

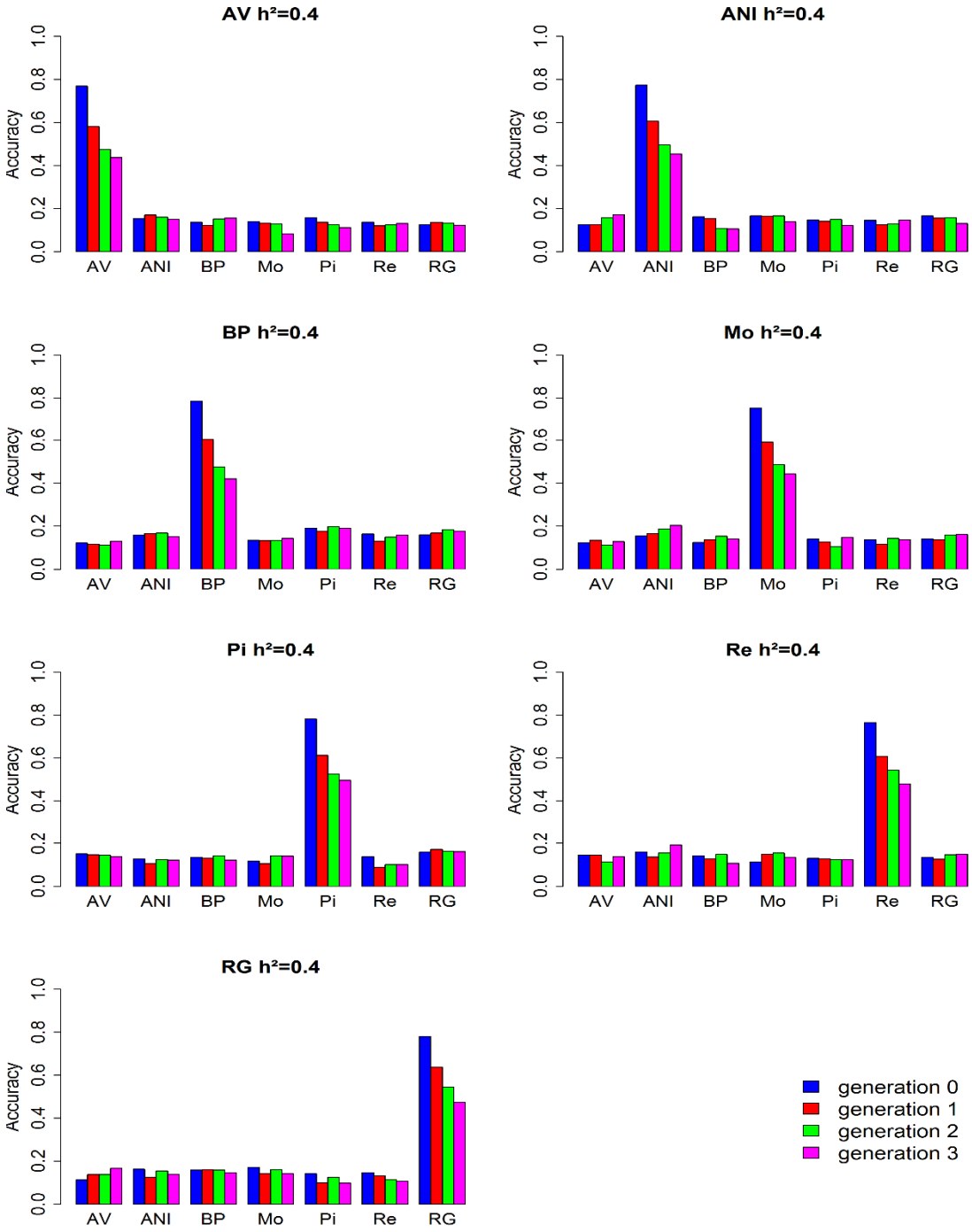
Single-breed evaluation

In the first scenario, the effects of the markers were estimated within each breed. Then, they were used to predict the GEBV within and across breeds. Figure 9 shows the results of the accuracies obtained for a trait with heritability 0.4 in all populations for generation 0 (where reference and validation populations are the same) and for 3 subsequent generations.

Within-breed accuracies at generation 0 were the highest, ranging from 0.785 (BP) to 0.754 (Mo). These results are slightly higher than those reported by Saatchi *et al.* (2011) and Van Eenennaam *et al.* (2014) from empirical field studies that ranged from 0.3 to 0.7 and from 0.22 to 0.69, respectively. The reason of the higher accuracies obtained in our study is probably the fact that the validation and the training sets have the maximum degree of relatedness. On the other hand, the across-breed accuracies were very low, with the highest value obtained when training in BP to predict over Pi (0.191) and the lowest when training in RG to predict over AV (0.112). These results confirm the postulate of Harris *et al.* (2008) that indicated that training in one population and validating in another is not effective.

However, it is remarkable that all the average estimates are positive and the results are coherent with the studies of persistence of LD phase by Cañas-Alvarez *et al.* (2016) in the same populations.

Figure 9. *Accuracy from single breed genomic evaluation (h^2=0.4)*

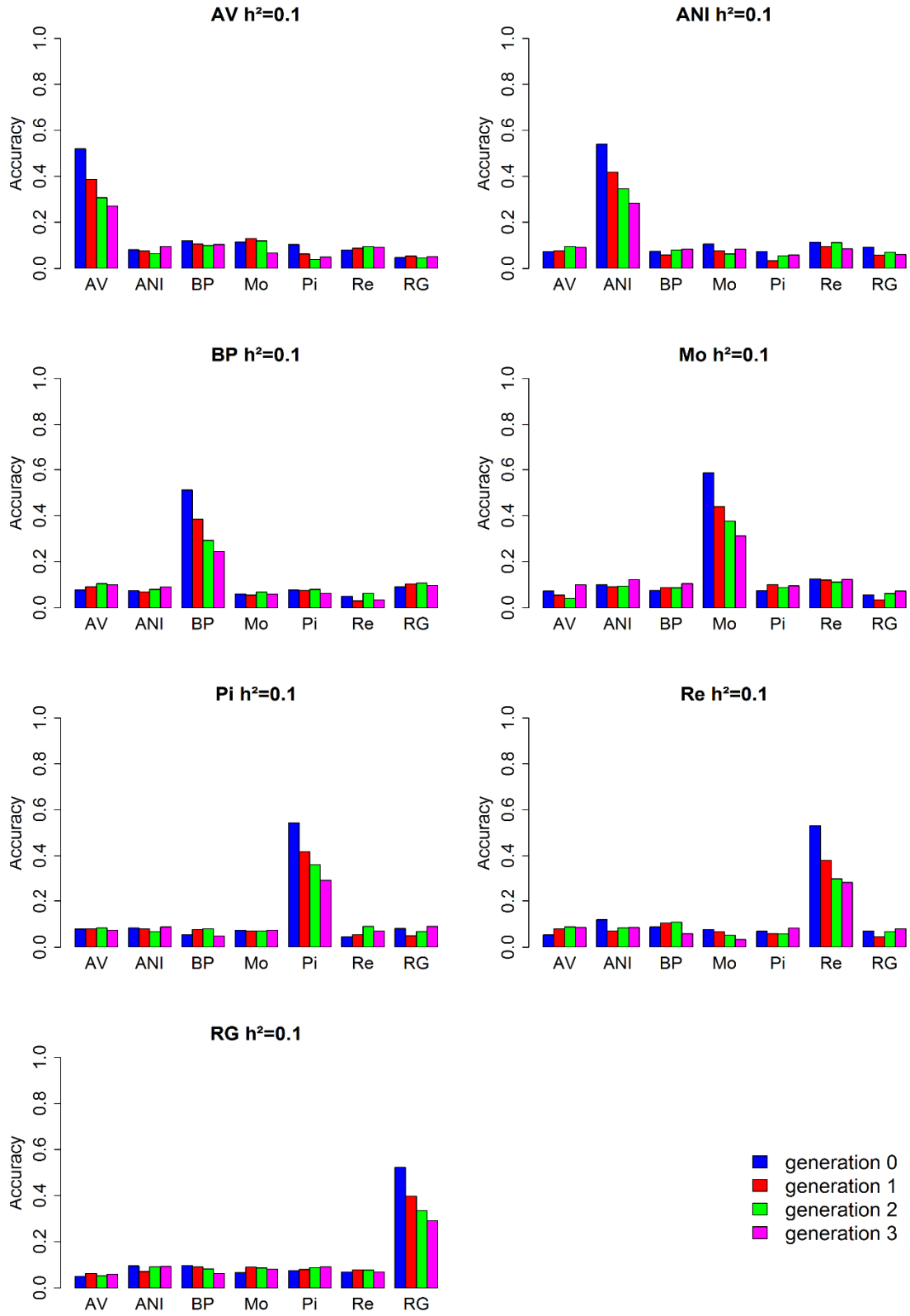


AV=Asturiana de los Valles, ANI=Avileña-Negra Iberica, BP=Bruna dels Pirineus, Mo=Morucha, Pi=Pirenaica, Re=Retinta, RG= Rubia Gallega, generations 0, 1, 2, 3 = distance in generations between the training and validation sets.

Additionally, 3 more generations of 1,000 individuals each were simulated for each population and used for validation. When predicting the subsequent generations the within-breed accuracies resulted on average lower by 21.6% in generation 1 with values between 0.637 (RG) and 0.580 (AV), 34.5% in generation 2 with values between 0.545 (RG) and 0.475 (AV) and 40.7% in generation 3 with values between 0.496 (Pi) and 0.420 (BP) regarding to generation 0. These results were expected and confirm the relevance of the relationship between the testing and training populations in the accuracy of genomic selection (Clark *et al.,* 2012). Further, the across-breed accuracies showed a small random fluctuation around the values in generation 0, because the relationship between testing and training populations is not modified in these cases.

The results of accuracy obtained when evaluating for a trait with heritability 0.1 resulted lower with values ranging between 0.587 (Mo) and 0.512 (BP) for within-breed predictions and between 0.125 (Mo over Re) and 0.045 (Pi over Re) for across breed predictions in generation 0. The loss of predictive ability over subsequent generations resulted to be higher than that of the previous case. The within-breed accuracies were 24.7% lower in generation 1 (0.440 (Mo) – 0.380 (Re)), 38.1% in generation 2 (0.377(Mo) – 0.292 (BP)) and 47.2% in generation 3 (0.312 (Mo) – 0.249 (BP)) (Figure 10). The higher loss of accuracy as the relationship between testing and training populations weakens is related to the crucial relevance of the information from relatives for low heritability traits (Falconer and McKay, 1996).

Figure 10. *Accuracy from single breed genomic evaluation (h^2=0.1)*



AV=Asturiana de los Valles, ANI=Avileña-Negra Iberica, BP=Bruna dels Pirineus, Mo=Morucha, Pi=Pirenaica, Re=Retinta, RG= Rubia Gallega, generations 0, 1, 2, 3 = distance in generations between the training and validation sets.

Evaluation in admixed ×2

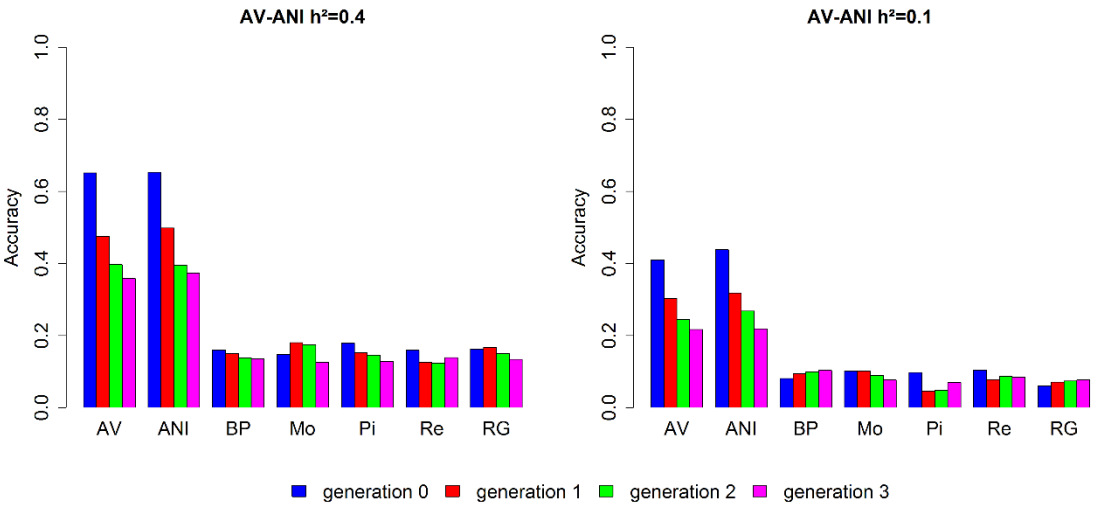
The training sets used in this second scenario were set up by mixing data from two purebred populations with equal proportion of each. All possible combinations were considered which resulted in 21 different admixed populations. Table 6 contains the results of the predictive ability of these populations over the purebred populations for generation 0 for a trait with heritability of 0.4. When the purebred validation population was included in the admixed training set the accuracies ranged from 0.680 (AV-BP over BP, BP-Pi over BP, BP-RG over RG, Re-RG over RG) and 0.639 (BP-MO, Mo-Pi, Mo-Re over Mo). However, when the purebred validation population was not included in the training set the accuracies resulted similar to the previous scenario of single breed evaluation and ranging between 0.202 (AV-BP over Pi) and 0.114 (Pi-RG over AV).

Table 6. *Accuracy from genomic evaluation in admixed ×2 populations in the generation 0 (h^2=0.4)*

Training sets	*Validation sets*						
	AV	ANI	BP	Mo	Pi	Re	RG
AV-ANI	**0.651**	**0.653**	0.160	0.148	0.179	0.159	0.163
	(0.018)	(0.003)	(0.020)	(0.017)	(0.023)	(0.015)	(0.010)
AV-BP	**0.652**	0.173	**0.680**	0.139	0.202	0.174	0.166
	(0.019)	(0.027)	(0.011)	(0.016)	(0.014)	(0.014)	(0.026)
AV-Mo	**0.651**	0.154	0.147	**0.640**	0.174	0.163	0.146
	(0.020)	(0.015)	(0.011)	(0.013)	(0.017)	(0.010)	(0.016)
AV-Pi	**0.650**	0.152	0.160	0.131	**0.669**	0.150	0.156
	(0.019)	(0.022)	(0.023)	(0.020)	(0.010)	(0.007)	(0.019)
AV-Re	**0.652**	0.179	0.161	0.119	0.159	**0.659**	0.162
	(0.018)	(0.018)	(0.009)	(0.018)	(0.012)	(0.012)	(0.021)
AV-RG	**0.650**	0.158	0.163	0.144	0.165	0.160	**0.678**
	(0.019)	(0.010)	(0.013)	(0.011)	(0.011)	(0.015)	(0.017)
ANI-BP	0.139	**0.654**	**0.678**	0.158	0.188	0.176	0.183
	(0.020)	(0.004)	(0.0151)	(0.011)	(0.010)	(0.017)	(0.017)
ANI-Mo	0.131	**0.655**	0.137	**0.640**	0.154	0.163	0.159
	(0.029)	(0.003)	(0.019)	(0.013)	(0.021)	(0.017)	(0.024)
ANI-Pi	0.127	**0.652**	0.154	0.152	**0.668**	0.155	0.172
	(0.025)	(0.003)	(0.024)	(0.017)	(0.010)	(0.016)	(0.019)
ANI-Re	0.146	**0.654**	0.146	0.144	0.147	**0.660**	0.172
	(0.007)	(0.002)	(0.018)	(0.018)	(0.012)	(0.012)	(0.009)
ANI-RG	0.124	**0.653**	0.155	0.163	0.148	0.154	**0.679**
	(0.028)	(0.004)	(0.015)	(0.007)	(0.016)	(0.013)	(0.017)
BP-Mo	0.127	0.172	**0.677**	**0.639**	0.179	0.175	0.156
	(0.029)	(0.009)	(0.012)	(0.012)	(0.018)	(0.013)	(0.019)
BP-Pi	0.126	0.171	**0.680**	0.144	**0.668**	0.170	0.174
	(0.006)	(0.038)	(0.011)	(0.015)	(0.010)	(0.017)	(0.018)
BP-Re	0.139	0.192	**0.679**	0.130	0.169	**0.661**	0.177
	(0.021)	(0.018)	(0.011)	(0.011)	(0.032)	(0.012)	(0.019)
BP-RG	0.126	0.177	**0.678**	0.155	0.172	0.178	**0.680**
	(0.025)	(0.025)	(0.012)	(0.009)	(0.019)	(0.010)	(0.0018)
Mo-Pi	0.120	0.156	0.138	**0.639**	**0.667**	0.156	0.152
	(0.012)	(0.010)	(0.024)	(0.013)	(0.011)	(0.015)	(0.036)
Mo-Re	0.133	0.175	0.133	**0.639**	0.134	**0.659**	0.154
	(0.029)	(0.020)	(0.023)	(0.012)	(0.018)	(0.012)	(0.018)
Mo-RG	0.115	0.155	0.139	**0.640**	0.139	0.160	**0.679**
	(0.034)	(0.014)	(0.021)	(0.012)	(0.017)	(0.013)	(0.018)
Pi-Re	0.128	0.175	0.152	0.125	**0.667**	**0.658**	0.172
	(0.011)	(0.015)	(0.026)	(0.024)	(0.011)	(0.013)	(0.020)
Pi-RG	0.114	0.154	0.159	0.148	**0.667**	0.154	**0.679**
	(0.011)	(0.016)	(0.014)	(0.009)	(0.011)	(0.004)	(0.017)
Re-RG	0.126	0.174	0.154	0.134	0.124	**0.659**	**0.680**
	(0.025)	(0.009)	(0.017)	(0.014)	(0.015)	(0.013)	(0.017)

As an example, Figure 11 shows the results obtained from training in the AV-ANI population for all generations. As expected, the predictive ability of these populations over the subsequent generations, when the validation sets were included in the training set, was lower than that of generation 0. Though, the loss in predictive ability resulted slightly higher than that of the previous case. On average the accuracies resulted 23%, 35% and 43.2% for generations 1, 2 and 3 respectively. The reason of this higher decrease of accuracy may be the limited number of individuals in the training population (1500 vs. 3000) with a direct relationship with the testing population.

Figure 11. *Accuracy from the admixed ×2 population (AV-ANI) genomic evaluation.*



AV=Asturiana de los Valles, ANI=Avileña-Negra Iberica, BP=Bruna dels Pirineus, Mo=Morucha, Pi=Pirenaica, Re=Retinta, RG= Rubia Gallega, generations 0, 1, 2, 3 = distance in generations between the training and validation sets.

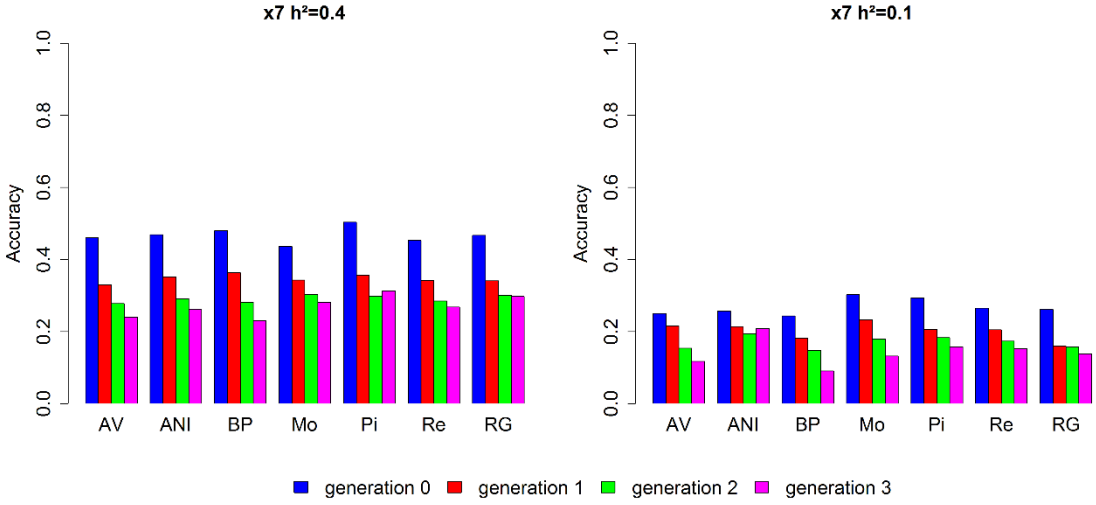
Moreover, the results with heritability 0.1 were similar, although the overall accuracies resulted lower than previously. Accuracies for the populations included in the training set ranged between 0.470 (AV-Mo and BP-Mo over Mo) and 0.401 (BP-Pi over BP). On the other hand, the results for the populations not included in

the admixture were also lower, from 0.127 (ANI-Mo over Re) to 0.031 (AV-Re over RG). The loss of accuracy in the subsequent generations was 26.1%, 39.2% and 48.4% for generations 1, 2 and 3 respectively.

Evaluation in admixed ×7

Finally, the last training set used for genomic evaluation was constructed by combining data from 429 individuals from each purebred population (total 3,003 individuals). Figure 12 shows the results obtained for all populations and all generations. The accuracies were between 0.503 (Pi) and 0.436 (Mo) for the trait with h^2=0.4 and between 0.302 (Mo) and 0.243 (BP) for the trait with h^2=0.1, while the loss of accuracy with the generations was 25.7%, 37.7% and 42.2% for the first trait and 24.3%, 36.3% and 46.8% for the second trait. The accuracies were lower than those in the previous scenarios and, as before, the loss of accuracy when training and testing populations are more distant is greater with lower heritability.

Figure 12. *Accuracy from admixed ×7 genomic evaluation.*

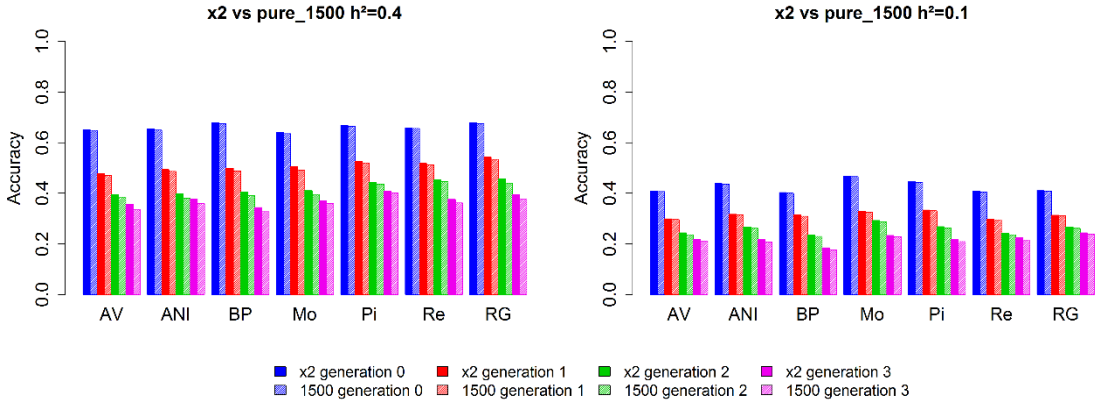


AV=Asturiana de los Valles, ANI=Avileña-Negra Iberica, BP=Bruna dels Pirineus, Mo=Morucha, Pi=Pirenaica, Re=Retinta, RG= Rubia Gallega, generations 0, 1, 2, 3 = distance in generations between the training and validation sets.

Admixed vs reduced purebred

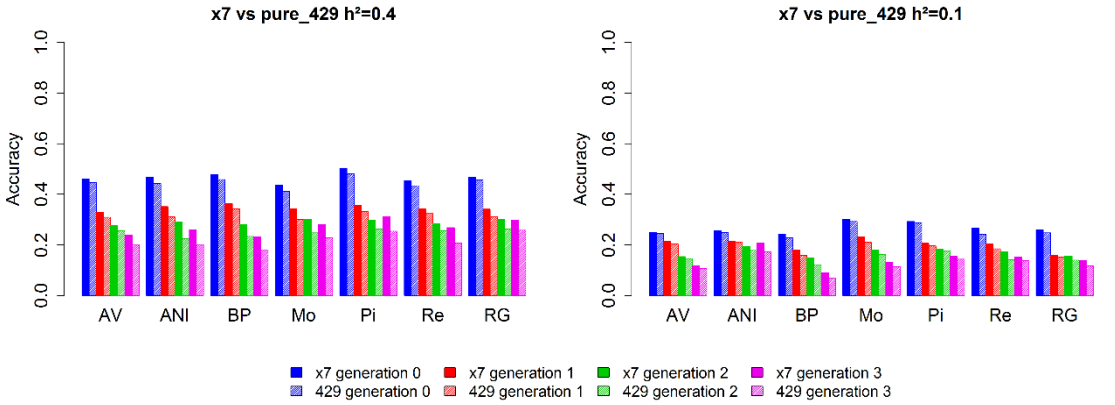
For each scenario, we also performed genomic evaluations using the groups of individuals selected (1,500 and 429 individuals) to make up the admixed populations separately for each purebred population in order to compare them with the admixed populations and evaluate the effect of adding individuals from other populations to increase the size of the training dataset. The figures 13 and 14 present the results of this comparison.

Figure 13. *Comparison between the results of the admixed ×2 and purebred genomic evaluation with 1,500 individuals per population.*



AV=Asturiana de los Valles, ANI=Avileña-Negra Iberica, BP=Bruna dels Pirineus, Mo=Morucha, Pi=Pirenaica, Re=Retinta, RG= Rubia Gallega, generations 0, 1, 2, 3 = distance in generations between the training and validation sets, ×2=admixed training set from 2 purebred populations (1,500+1,500 individuals), pure_1500=purebred training set with 1,500 individuals.

Figure 14. *Comparison between the results of the admixed ×7 and purebred genomic evaluation with 429 individuals per population.*



AV=Asturiana de los Valles, ANI=Avileña-Negra Iberica, BP=Bruna dels Pirineus, Mo=Morucha, Pi=Pirenaica, Re=Retinta, RG= Rubia Gallega, generations 0, 1, 2, 3 = distance in generations between the training and validation sets, ×7=admixed training set from all 7 purebred populations (7×429 individuals), pure_429=purebred training set with 429 individuals.

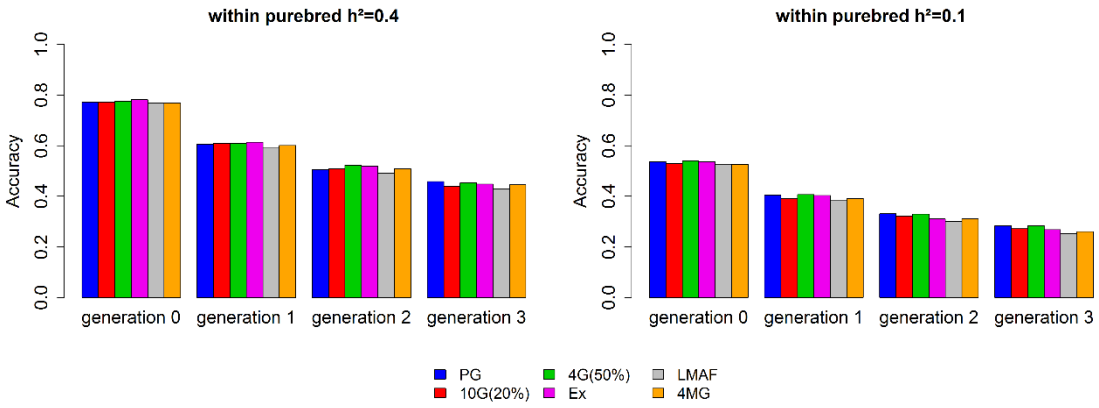
They show that adding information from another population to a small sized population is beneficial in all cases. The admixed ×2 populations performed slightly better than the reduced purebred populations with 1,500 individuals with 0.6%, 1.8%, 3.1% and 4.3% higher accuracies for generations 0, 1, 2, and 3 respectively for a trait with h^2=0.4 and 0.7%, 1.1%, 2.2% and 3.6% for a trait with h^2=0.1. This superiority of the admixed population was more obvious between the admixed ×7 and the reduced purebred populations of 429 individuals. Here, the gain in accuracy with the number of generations was 4.4%, 8.9%, 17% and 23.6% for the first trait (h^2=0.4) and 4.3%, 7.4%, 11.9% and 16.8% for the second trait (h^2=0.1). The most probable cause of this phenomenon is that as the relatedness between the training set and the validation set weakens the predictions are based more on the short range LD between the markers and the genes than on the pure genetic relationship between individuals. Thus, the admixed populations perform better because of the

higher number of data and the fact that mixing data breaks down the long distance LD created by relatedness and leaves the effects of the short range LD that persists through generations (Falconer and McKay, 1996).

Genetic architecture of the trait

Finally, we also compared the consequences of alternative genetic architecture of the traits. Thus, along with the polygenic traits (PG) simulated above, 5 more cases of genetic architecture were simulated as described earlier. In Figure 15, we present the average within breed accuracies for both traits (h^2=0.4 and h^2=0.1), obtained from training in purebred populations for all the populations simulated and for all generations. The values obtained resulted similar in all cases. Small differences can be observed only for the case that the traits are controlled by rare variants (LMAF) with MAF lower than 0.05, and where the loss was slightly greater with the number of generation.

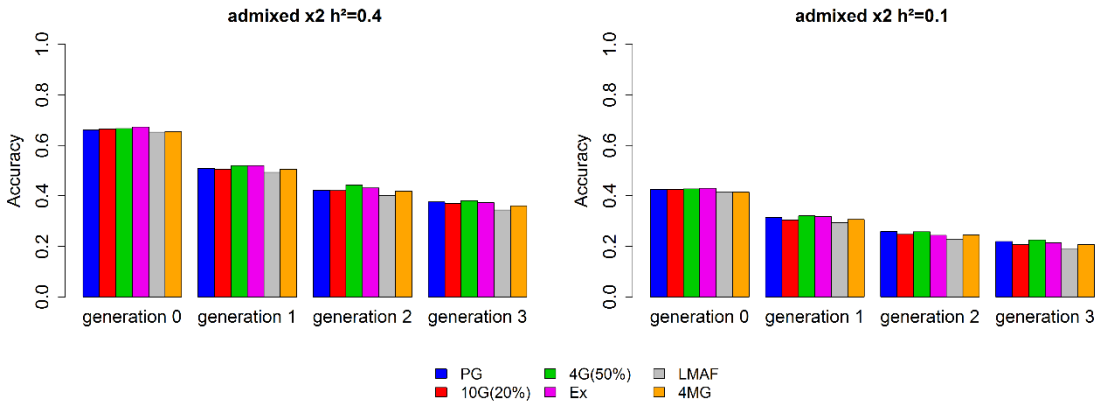
Figure 15. *Accuracy from single-breed genomic evaluation under different genetic architecture scenarios.*



PG=Polygenic effects, 10G(20%)=Polygenic effects + 10 genes explaining 20% of the genetic variance, 4G(50%)=Polygenic effects + 4 genes explaining 50% of the genetic variance, Ex=Polygenic effects drawn from an exponential distribution, LMAF=polygenic effects with low allelic frequencies (≤0.05), 4MG=4 major genes.

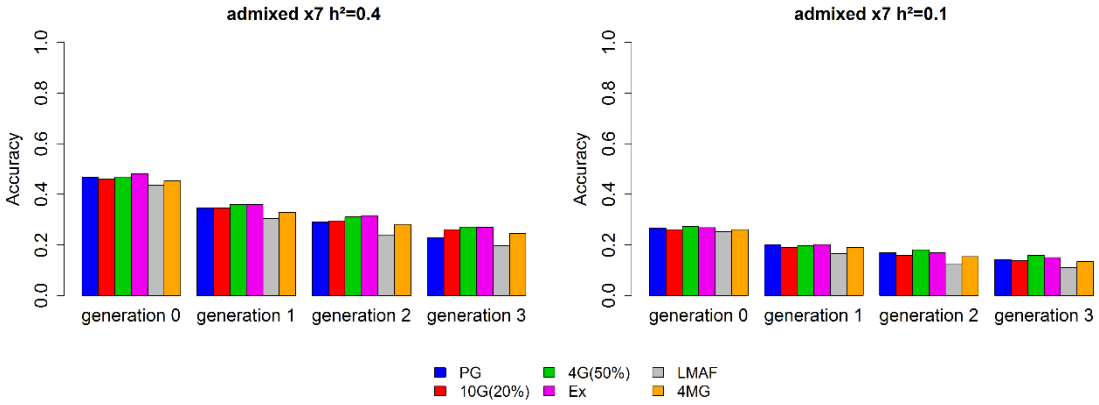
Similarly, the results from the admixed ×2 and admixed x7 training sets showed little differences among cases (Figures 16 and 17, respectively). As before, only the LMAF case gives slightly lower accuracies. This phenomenon is coherent with the results obtained by Wientjes *et al.,* (2015), that indicated that when the QTLs controlling the genetic variability of the traits have lower frequencies the ability of prediction of Genomic Selection in lower. However, although this has been suggested as the cause of the missing heritability (Gibson, 2012), the evidence for the percentage of genetic variation that rare variants produce is low and some authors have shown that these rare variants explain a small percentage of the missing heritability of complex traits in human (Gusev *et al.,* 2014) or cattle (Gonzalez-Recio *et al.,* 2015).

Figure 16. *Accuracy from admixed ×2 genomic evaluation under different genetic architecture scenarios.*



PG=Polygenic effects, 10G(20%)=Polygenic effects + 10 genes explaining 20% of the genetic variance, 4G(50%)=Polygenic effects + 4 genes explaining 50% of the genetic variance, Ex=Polygenic effects drawn from an exponential distribution, LMAF=polygenic effects with low allelic frequencies (≤0.05), 4MG=4 major genes.

Figure 17. *Accuracy from admixed ×7 genomic evaluation under different genetic architecture scenarios.*



PG=Polygenic effects, 10G(20%)=Polygenic effects + 10 genes explaining 20% of the genetic variance, 4G(50%)=Polygenic effects + 4 genes explaining 50% of the genetic variance, Ex=Polygenic effects drawn from an exponential distribution, LMAF=polygenic effects with low allelic frequencies (≤0.05), 4MG=4 major genes.

Moreover, when comparing the accuracies obtained from an admixed population and a reduced sized purebred population (Figures 18 and 19) the results follow those of the PG case with the only exception of the LMAF where the reduced purebred training sets yielded higher accuracies than those of the admixed training sets with the number of generations. In the LMAF case the markers selected to simulate the causal mutations where selected under the condition of having extreme frequencies (MAF≤0.05). As a consequence, the LD between the neutral markers and the genes is lower even at close distances and therefore, the reduced purebred training sets perform better than the admixed training sets because there is a larger proportion of family LD than sort range historical LD, although the family LD is decaying with the number of generations.

Conclusions

The results obtained in this study indicate that the use of a meta-population provides reasonable but not completely satisfactory accuracies. Though, the admixed populations seem to have a small advantage when predicting individuals more distant from the training set because the across breed and multi-breed genomic predictions are based more on the LD between markers and QTLs than family relationships.

CHAPTER 3

Genetic architecture of the persistency of linkage disequilibrium across seven Spanish beef cattle populations.

Introduction

The advances in the area of molecular genetics have allowed the development of dense SNP genotyping devices (Gunderson *et al.,* 2005) that have provided information throughout the genome of several livestock species. Along with these molecular advances, new statistical methods have been developed with the purpose of predicting the genomic breeding values of candidates to selection (Meuwissen *et al.,* 2001). Genomic selection is a reality in dairy cattle (Hayes *et al.,* 2009) and its implementation is consolidating in other livestock species, such as pig (Lillehammer *et al.,* 2011; Samore *et al.,* 2014; Esfandyari *et al.,* 2015) or small ruminants (Shumbusho *et al.,* 2015; Casellas and Piedrafita, 2015). Nevertheless, its introduction into the routine selection schemes of beef cattle has been considerably slower. Several causes can be argued for this delay (Berry *et al.,* 2016). Among them, the limited census of the beef populations and the smaller implantation of artificial insemination (AI) as compared to dairy populations play a crucial role.

One possible alternative that minimizes these problems is the potential use of information from multiple populations for the genomic evaluation. Nonetheless, the results from simulation studies, as presented in the previous chapter, as well as from experimental data (Kachman et al., 2013) show little progress. Theoretically, the success of the genomic evaluation from multiple populations is linked to the persistency of the linkage disequilibrium (LD) between the populations in such way that the LD between markers and QTLs is maintained in all populations. Several authors have studied the persistency between populations as one more measure of genetic diversity (de Roos *et al.,* 2008; Villa-Angulo *et al.,* 2009; Cañas-Álvarez *et al.,* 2016) but the genetic architecture of this persistency has been barely studied.

Therefore, the objective of this study is to analyse the pattern of the linkage disequilibrium persistency between seven Spanish beef cattle populations along the autosomal chromosomes. On one hand, the results will provide valuable information about the evolutionary history of these populations (Teo et al., 2009) and, on the other, they may be also used to improve the across population estimates of Genomic Breeding Values (GEBVs).

Material

The data used in this study comprised of the *BovineHD Beadchip* genotypes of the 342 founder individuals of the triplets described in the MATERIAL section (*Asturiana de los Valles* – AV-, N=50, *Avileña - Negra Ibérica* – ANI-, N=48, *Bruna dels Pirineus* – BP-, N=50, *Morucha* –Mo-, N=50, *Pirenaica* –Pi-, N=48, *Retinta* – Re-, N=48 and *Rubia Gallega* –RG, N=48). Here, an additional quality control requirement was a minor allele frequency (MAF) of 0.05 in pairs of populations, resulting in around 550,000 segregating markers for each pair of populations (see Table 7. for detailed results).

***Table 7**. Number of segregating SNP markers between all pairs of populations.*

Pairs of populations	Nº SNP markers	Pairs of populations	Nº SNP markers
AV-ANI	555,373	**BP-Mo**	543,305
AV-BP	557,588	**BP-Pi**	534,336
AV-Mo	555,769	**BP-Re**	535,997
AV-Pi	540,390	**BP-RG**	544,350
AV-Re	547,893	**Mo-Pi**	529,281
AV-RG	553,868	**Mo-Re**	541,225
ANI-BP	538,327	**Mo-RG**	542,682
ANI-Mo	545,324	**Pi-Re**	522,670
ANI-Pi	524,630	**Pi-RG**	529,577
ANI-Re	536,595	**Re-RG**	535,677
ANI-RG	537,882		

Methods

Persistency of Linkage Disequilibrium

Two measures of local persistency of linkage disequilibrium (LD) were used for sliding windows of 50, 100 and 200 SNP that cover an average size in physical distance of 226.85 kb (s.d. 110.2 kb), 457.85 kb (s.d. 175.3 kb) and 919.79 kb (s.d. 279.4 kb), respectively.

CorLD: The LD between all markers of each region was calculated as:

$$r = \frac{D}{\sqrt{p_A p_a p_B p_b}},$$

where $D = p_{AA}p_{BB} - p_{Ab}p_{aB}$ (Falconer and Mackay, 1996), and p_A, p_a, p_B and p_b are the allele frequencies p_{AA}, p_{BB} are the homozygous haplotype frequencies and p_{Ab}, p_{aB} are the heterozygous frequencies.

Finally, to compare the persistency of LD across populations the Pearson correlation was calculated between the values of LD (*r*) of all markers within each region for each pair of populations and for all sizes of windows.

VarLD (Teo et al., 2009): The LD between all markers of each region was calculated as the signed r^2:

$$r^2 = \frac{(p_{AB} - p_A p_B)^2}{p_A p_a p_B p_b} (-1)^{I(p_{AB} < p_A p_B)} ,$$

where $I(p_{AB} < p_A p_B)$ denotes an indicator function taking a value of one when $p_{AB} < p_A p_B$, and zero otherwise. Once the correlation matrices between the markers of each region are calculated, an eigen-decomposition is performed on each LD matrix in order to obtain a diagonal matrix with entries comprising of the sorted eigenvalues in descending order. The raw VarLD score is the absolute difference between the diagonal matrices of the two populations, and the magnitude of this score provides a measure for the extent of dissimilarity between the correlation matrices. This analysis was performed with the *VarLD* software (Ong *et al.,* 2010).

Identification of candidate genes and metabolic pathways

Once the correlations of CorLD and the raw scores of VarLD were obtained, the regions with values of CorLD higher than the top 0.1% and the regions with values of VarLD lower than the bottom 0.1% of the empirical distributions were selected for each pair of populations separately. Finally, we used the Biomart tool of Ensembl (www.ensembl.org; Flicek, 2013) to detect the genes located in the selected regions and the Enrichr tool (Chen *et al.,* 2013b) to identify the biological pathways in which these genes participate. The most significant pathways were selected under the

criteria of the adjusted p-value (< 0.05) as given by the Enrichr tool which uses the Benjamini-Hochberg method of correction for multiple hypotheses testing (Benjamini and Hochberg, 1995).

Integration into across-breed genomic evaluation

The results of the local patterns of linkage disequilibrium were used to define alternative models to perform across breed genomic evaluation. The base model was the GBLUP model that we consider in the previous chapter:

$$y_i = \mu + \sum_{j=1}^{n} x_{ij} a_j + e_i$$

where y_i is the phenotype of the i^{th} individual, μ is the trait mean, n is the number of SNPs, x_{ij} is the genotype of the i^{th} individual for the j^{th} marker codified as 0,1 and 2, a_j is the substitutions effect for the j^{th} marker and e_i is the residual effect of the i^{th} individual. Further, the prior distribution for the marker effects was the following multivariate Gaussian distribution:

$$\boldsymbol{a} \sim \boldsymbol{N}(0, \boldsymbol{I}\sigma_a^2)$$

where σ_a^2 is the marker variance whose prior distribution is assumed to be uniform within appropriate bounds.

Later on, we defined alternative models by modifying the prior distribution of marker effects with the consideration of the local estimate of the persistency of LD disequilibrium between populations from corLD. Thus the identity matrix (**I**) of the prior distributions of markers was replaced by a **T** matrix. This matrix is diagonal but

with values that correspond to the measure of the local persistence of LD. Several alternatives of prior distribution were proposed:

1. The local estimate of corLD that ranges between -1 and 1.
2. The square of corLD.
3. The cube of corLD.

The aim of these alternative prior distributions was to assign a different intensity of regularization depending on the persistency of LD between a specific pair of populations. The procedure was tested in one of the scenarios of simulation described in the previous chapter (**PG**) and considering only two chromosomes and in one pair of populations (Avileña Negra-Ibérica –AV- and Asturiana de los Valles –ANI-), that included 71,329 SNP markers and 2,205 QTLs that describe two traits with heritabilities 0.4 and 0.1.

Results and Discussion

Architecture of Linkage Disequilibrium

The analysis of the similarity of the LD patterns between populations was performed for sliding windows of 50, 100 and 200 SNP markers for both methods (CorLD and VarLD). After a visual inspection (see ANNEXE 2 for one example between ANI and AV), we decide to focus our study in the case of sliding windows of 100 SNPs, because the results from 50 SNPs were extremely noisy, whereas the results from windows of 200 SNPs were less variable as they include huge genomic regions of around 1 Mb.

In first place, we compared the results from both procedures (VarLD and CorLD) along every pair of populations and locations in the genome and we found that the raw correlation between them was positive, although very low (0.146 ± 0.050). Note that a positive correlation indicates that the signals of both methods tend to be in different sense. Thus, they indicate that the signals of persistency of phase were different. In fact, VarLD was designed (Teo *et al.,* 2009) for detecting differences and it is more sensible to a strong divergence between r^2 in a single, or few, pair of markers within a genomic region. Whereas, CorLD is more robust to few outlier correlations as it is calculated as a correlation of correlation estimates, that is less dependent on single estimates of LD. Thus, VarLD is probably more capable to detect genomic regions that diverge between populations, whereas CorLD is able to identify the ones where the persistency of the LD phase is maintained on average.

However, if we analyse the average results of CorLD and VarLD for each pair of populations, which are presented in Figures 18 and 19 (respectively), it can be observed that the average pattern of divergence is similar for both procedures.

***Figure 18**. Heatmap the average CorLD signals between pairs of populations along the genome.*

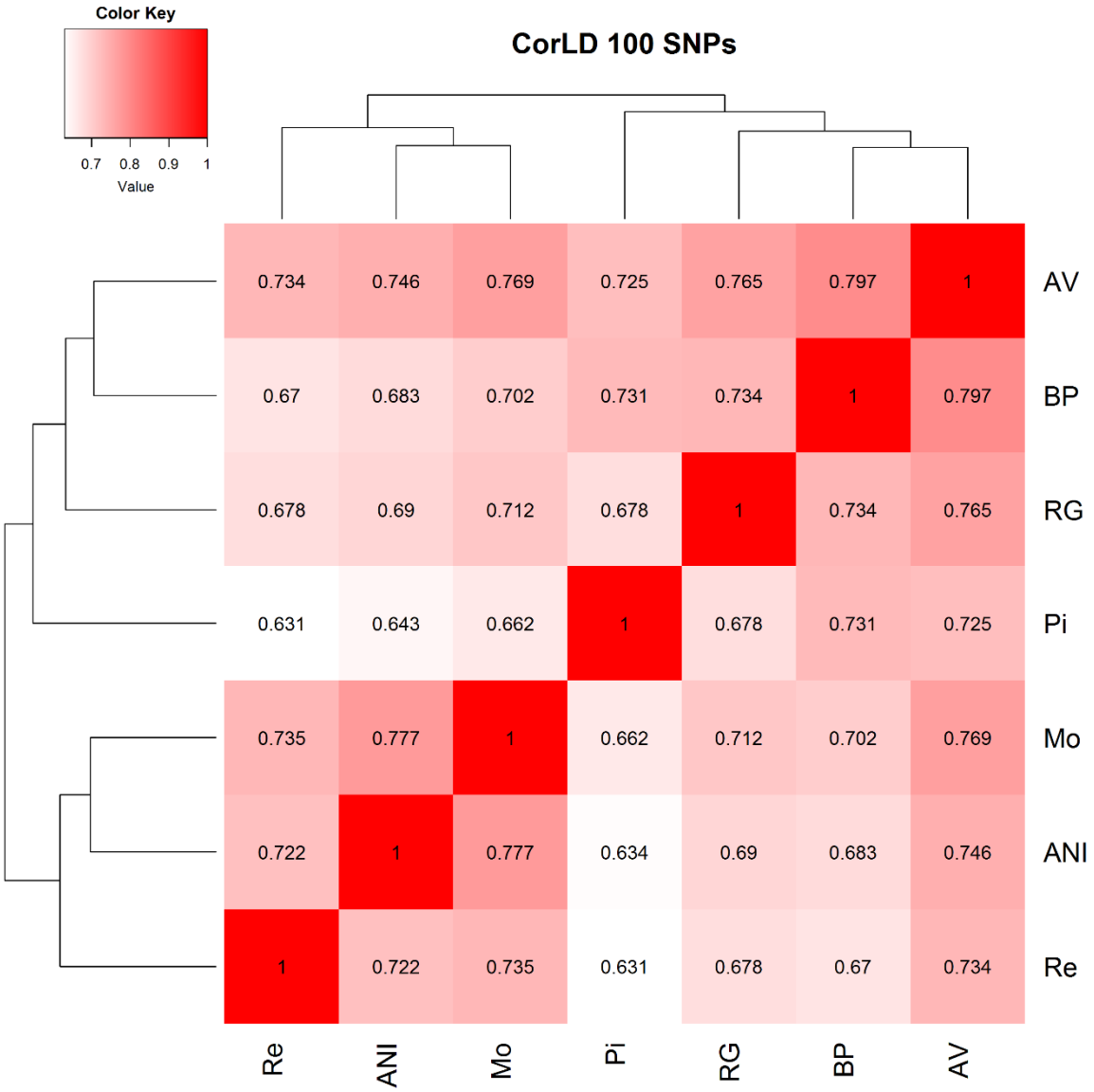
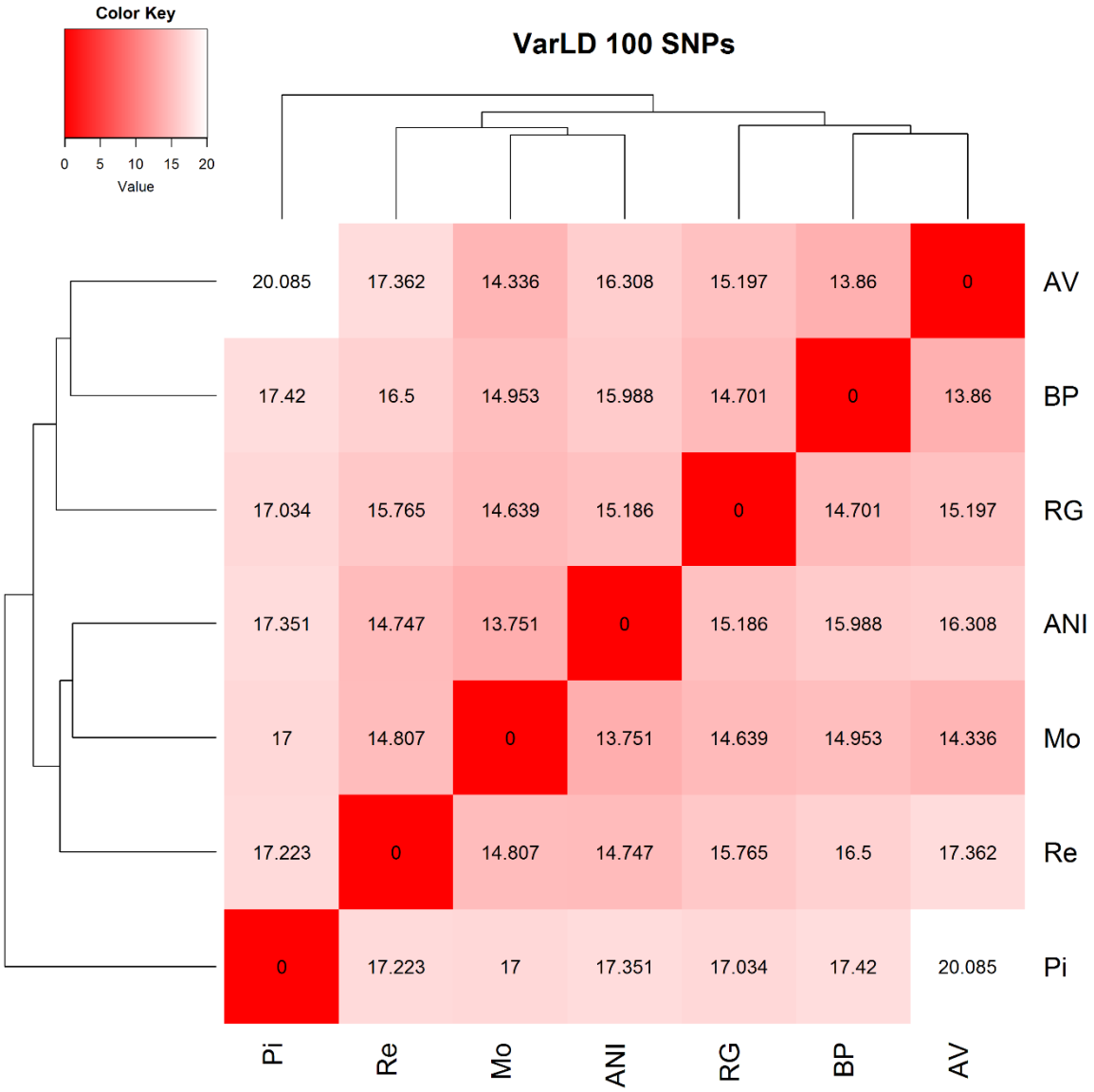


Figure 19. *. Heatmap the average VarLD signals between pairs of populations along the genome.*



Note that in the case of CorLD, the greater the value between two populations, the more similar patterns they have. On the contrary, for the VarLD analysis, the higher values indicate greater dissimilarity between the populations. Both methods showed different pairs of populations for the highest and lowest similarity. For CorLD the most similar were AV and BP (0.797) while for VarLD they were ANI and Mo (13.751). Likewise, for the most distant populations the results were Pi and Re

(0.631), and AV and Pi (20.085), respectively. However, if we look the dendrograms calculated using a neighbour-joining algorithm (Saitou and Nei, 1987), the pattern of classification between populations was similar. Both methods clustered the populations in two groups: ANI, Mo and Re on one hand and RG, AV and BP on the other. The only difference is the location of the Pi population, though it is consistently the most separated population. These results are in strong agreement with the results of divergence between these populations calculated using principal components (Cañas-Álvarez *et al.*, 2015) or phase persistency (Cañas-Álvarez *et al*., 2016). Moreover, the results are also in agreement with the traditional classification of the Spanish cattle populations (Sanchez-Belda, 1984) and with their geographical localization (Re, Mo and ANI are located in central and south of Spain while AV, RG and BP in the north). The divergence of Pi could be attributed to some degree of mixture with French populations given its localization close to the border between Spain and France.

The detailed plots of VarLD and CorLD for each pair of populations are presented in the ANNEXE 3 and here, as an example, we present exclusively the results (Figures 20 and 21) for all genomic regions along the autosomal chromosomes for one pair of populations (AV and BP).

Figure 20. *Manhattan plot of the CorLD estimates along the genome for the AV-BP comparison for regions of 100 SNPs.*

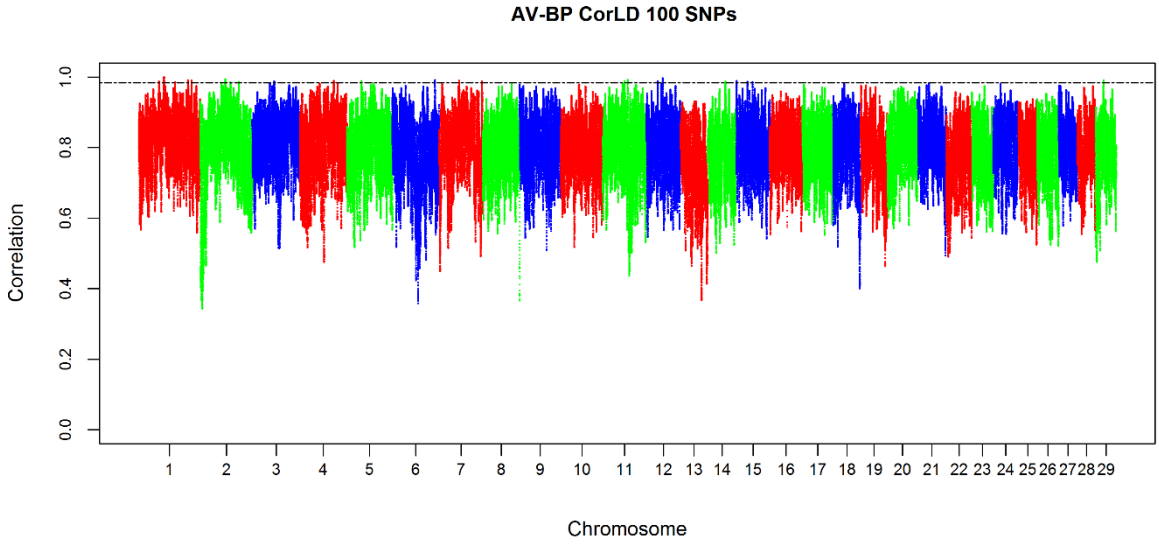
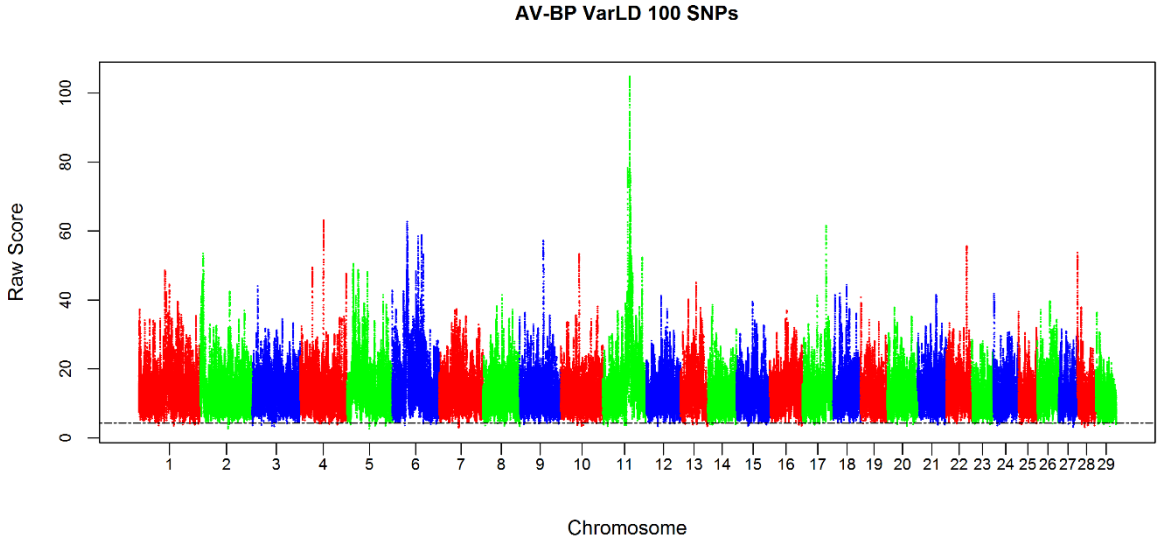


Figure 21*. Manhattan plot of the VarLD estimates along the genome for the AV-BP comparison for regions of 100 SNPs.*



As it can be observed, with the VarLD method is easier to distinguish genomic regions with more divergence, whereas with the CorLD method it is possible to detect both similitude and divergence between the genomic regions. Moreover, it

can also be observed that the concordance between both methods is very low, as pointed out by the correlation between them.

Identification of candidate genes and metabolic pathways

Once the correlations of CorLD and the raw scores of VarLD were obtained, the regions with values of CorLD higher than the top 0.1% and the regions with values of VarLD lower than the bottom 0.1% were selected for each pair of populations separately. Finally, we used the Biomart tool of Ensembl (www.ensembl.org; Flicek, 2013) to identify the genes located in the selected regions and the Enrichr tool (Chen *et al.,* 2013) to identify the biological pathways in which these genes participate.

The number of regions identified for each method differed significantly, with 120 regions for CorLD and 1,323 regions for VarLD. Later on, the regions that appeared in at least 10 pairs of populations were selected for further analysis, resulting in 14 regions for CorLD and just 6 for VarLD that are presented in Tables 8 and 9, respectively. As it can be seen, both methods revealed regions on different chromosomes, and without any region in common. Moreover, the size of the regions resulted significantly larger for VarLD than for CorLD.

***Table 8**. Genomic regions with values of CorLD higher than the 0.1% of the empirical distribution in at least 10 population pairs and the genes located there.*

Chromosome	Start position (pb)	End position (pb)	Distance (pb)	Nº of population pairs	Genes
1	66357406	66814280	456,874	21	STXBP5L, POLQ, ARGFX, FBXO40, HCLS1, GOLGB1
1	103439187	104057596	618,409	12	-
1	131212540	131860599	648,059	15	FOXL2, PIK3CB, FAIM, CEP70, ESYT3, MRAS, NME9
1	139666464	140197279	530,815	10	CPNE4, MRPL3, NUDT16, NEK11, U6
2	65221817	66070654	848,837	16	LYPD1, GPR39, SLC35F5, ACTR3
6	81225864	81933084	707,220	14	5S_rRNA, TECRL
7	53081367	54296580	1,215,213	17	CYSTM1, PFDN1, HBEGF, SLC4A9, U6, SRA1, APBB3, SLC35A4, CD14, TMCO6, IK, WDR55, DND1, HARS, HARS2, ZMAT, Vault, PCDHA3, PCDHA5, PCDHA11, PCDHA13, PCDHB1, PCDHB4, PCDHB6, PCDHB5, PCDHB7, PCDHB16, PCDHB9, PCDHB14, PCDHB15, TAF7, PCDHGA2, PCDHGB1, PCDHGB2, PCDHGA5, PCDHGA7, DIAPH1
8	45936344	46523114	586,770	12	APBA1, PTAR1, C9orf135
9	43672327	44358409	686,082	16	QRSL1, RTN4IP1, AIM1, ATG5, PRDM1
9	94763755	95310973	547,218	11	ARID1B, 5S_rRNA, bta-mir-2481
11	60241338	60772988	531,650	14	FAM161A, CCT4, COMMD1
12	41082553	42420488	1,337,935	21	-
15	43737	1399526	1,355,789	12	OR9G1, Olfactory receptors (x10)
21	28955131	29776774	821,643	11	TJP1, U4, TARSL2, TM2D3, PCSK6, SNRPA1

Table 9. *Genomic regions with values of VarLD lower than the 0.1% of the empirical distribution in at least 10 population pairs and the genes located there.*

Chromosome	Start position (pb)	End position (pb)	Distance (pb)	Nº of population pairs	Genes
7	106727846	108064926	1,337,080	13	-
13	80211353	83560269	3,348,916	14	ATP9A, SALL4, ZFP64, TSHZ2, snoU2_19, snoU2-30, ZNF217, BCAS1, CYP24A1, PFDN4, DOK5
28	9764836	13048195	3,283,359	11	RYR2, SNORA25, ZP4, 5S_rRNA, U6atac, U6, CHRM3, ZNF33B
28	15127565	17033972	1,906,407	10	FAM13C, SLC16A9, bta-mir-2403, CCDC6, ANK3, U6, CDK1, RHOBTB1
29	44972462	46682401	1,709,939	10	RAB1B, YIF1A, TMEM151A, CD248, RIN1, BRMS1, B4GAT1, U6, SLC29A2, NPAS4, MRPL11, PELI3, DPP3, ZDHHC24, ACTN3, CTSF, CCDC87, CCS, RBM14, RBM4, RBM4B, SPTBN2, C11orf80, RCE1, PC, LRFN4, bta-mir-2408, C11orf86, SYT12, RHOD, KDM2A, ADRBK1, ANKRD13D, SSH3, CLCF1, RAD9A, TBC1D10C, CARNS1, RPS6KB2, PTPRCAP, CORO1B, GPR152, CABP4, TMEM134, AIP, PITPNM1, CDK2AP2, CABP2, GSTP1, NDUFV1, DOC2G, NUDT8, TBX10, ALDH3B2, UNC93B1, ALDH3B1, NDUFS8, TCIRG1, CHKA, KMT5B, C11orf24, LRP5, 5S_rRNA, PPP6R3
29	46862596	50768717	3,906,121	14	MRPL21, IGHMBP2, MRGPRF, TPCN2, SNORD14, CCND1, ORAOV1, FGF19, FGF4, FGF3, 5S_rRNA, ANO1, FADD, CTTN, SHANK2, DHCR7, NADSYN1, MRGPRG, OSBPL5, U6, CARS, NAP1L4, PHLDA2, SLC22A18, CDKN1C, KCNQ1, TRPM5, TSSC4, CD81, TSPAN32, ASCL2, TH, INS IGF2, bta-mir-483, TNNT3, LSP1, TNNI2, SYT8, CRLF2, AP2A2, bta-mir-2409, IFITM10, 7SK, CHID1, TSPAN4, POLR2L, CD151, CRACR2B, PNPLA2

The number of genes harboured in these regions resulted in 92 for CorLD and 147 for VarLD. These gene sets were used for the enrichment analysis in order to reveal the most relevant biological pathways involved. The most significant pathways for the results obtained from CorLD are presented in Table 10. A total of 8 relevant pathways were revealed, that include processes of cell adhesion, synapse assembly and organization and nervous system development. The genes that participate in each of these pathways belong mainly to the *Protocadherin* gene family which are located on chromosome 7. This result is coincident with Su *et al.* (2014) that found that the haplotype diversity of this genomic region is reduced, and, as a consequence, the LD is increased consistently in all populations. On the other hand, the results obtained from VarLD yielded no significant pathways confirming that the procedure is more able to detect differences between populations than persistency of LD.

***Table 10**. Main biological pathways that the genes found in the CorLD regions participate.*

Biological pathway	**Genes**
Homophilic cell adhesion via plasma membrane adhesion molecules (GO:0007156)	PCDHGA7; PCDHGA5; PCDHGA2; PCDHGB2; PCDHB15; PCDHB14; PIK3CB; PCDHA13; PCDHA11; PCDHB1; PCDHB16; PCDHGB1; PCDHB6; PCDHA5; PCDHB5; PCDHB4; PCDHA3; PCDHB9; PCDHB7
Cell-cell adhesion (GO:0098609)	PCDHGA7; PCDHGA5; PCDHGA2; PCDHGB2; PCDHB15; PCDHB14; PIK3CB; PCDHA13; PCDHA11; PCDHB1; PCDHB16; PCDHGB1; PCDHB6; PCDHA5; PCDHB5; PCDHB4; PCDHA3; PCDHB9; PCDHB7
Cell-cell adhesion via plasma-membrane adhesion molecules (GO:0098742)	PCDHGA7; PCDHGA5; PCDHGA2; PCDHGB2; PCDHB15; PCDHB14; PIK3CB; PCDHA13; PCDHA11; PCDHB1; PCDHB16; PCDHGB1; PCDHB6; PCDHA5; PCDHB5; PCDHB4; PCDHA3; PCDHB9; PCDHB7
Calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules (GO:0016339)	PCDHB16; PCDHB6; PCDHB5; PCDHB14; PCDHB4; PCDHB9
Synapse assembly (GO:0007416)	PCDHB16; PCDHB6; PCDHB5; PCDHB14; PCDHB4; PCDHB9
Synapse organization (GO:0050808)	PCDHB16; PCDHB6; PCDHB5; PCDHB14; PCDHB4; PCDHB9
Nervous system development (GO:0007399)	PCDHB6; PCDHB15; PCDHA5; PCDHB4; APBA1; PCDHA3; PCDHA11; ARID1B
Amino acid activation (GO:0043038)	TARSL2; HARS; QRSL1; HARS2

Integration into across-breed genomic evaluation

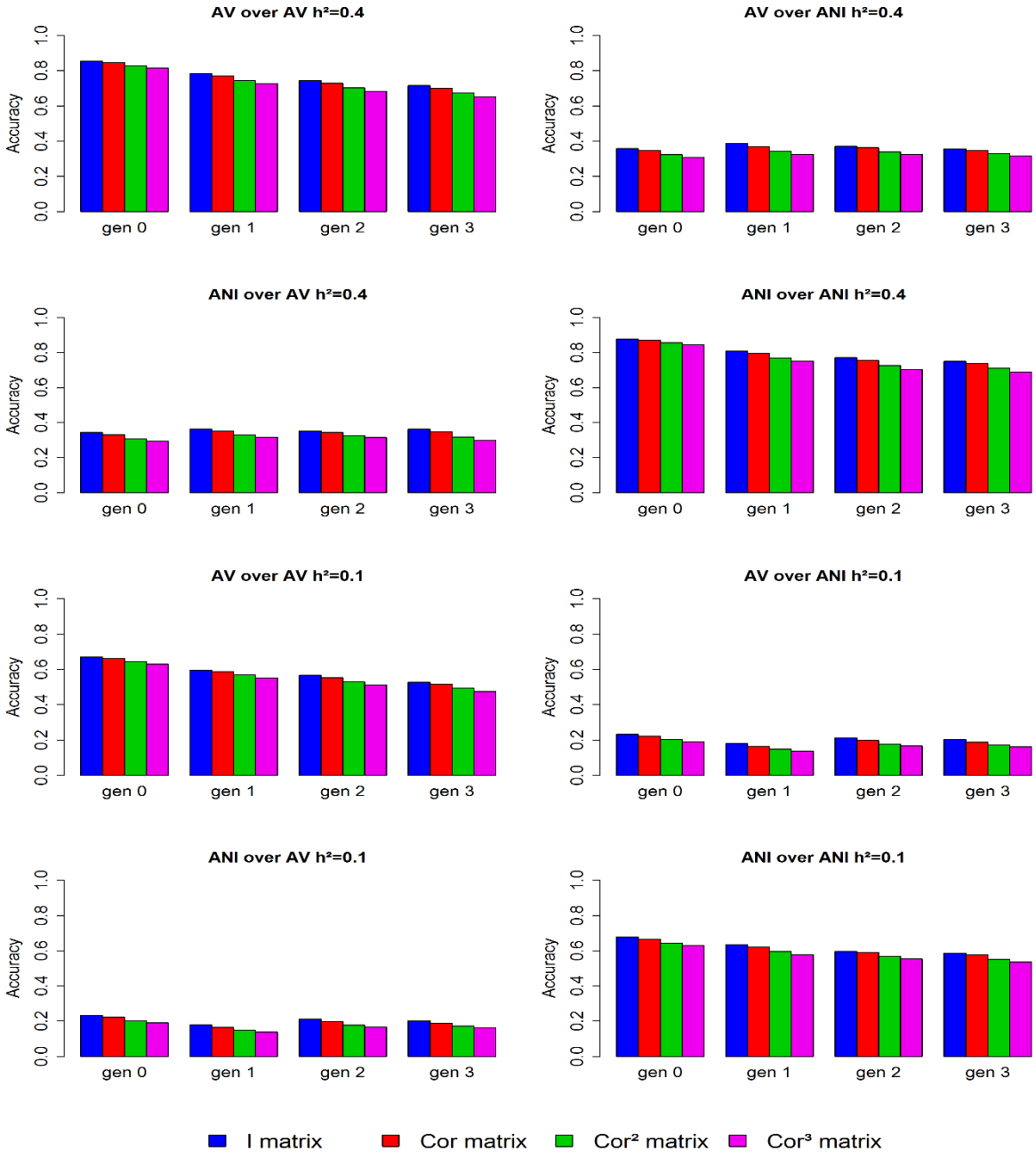
The results of the local patterns of linkage disequilibrium were used to define alternative models to perform across breed genomic evaluation by modifying the prior distribution of the marker effects with the consideration of the local estimate of the persistency of LD disequilibrium between populations from the results of CorLD. A reduced version of the base scenario (**PG**) from Chapter 2 was used considering two populations (AV and ANI) and just the two first chromosomes. The CorLD estimates were calculated for regions of 100 SNP markers and, as before, the results were averaged over 5 replicates of the analysis.

The average correlation among the genomic regions along the two chromosomes ranged between 0.041 (±0.023) and 0.819 (±0.011). The **T** matrix was constructed by replacing the diagonal of the identity matrix with the estimates of CorLD (Cor) for each marker adding a constant of 0.1 in order to avoid convergence problems. Moreover the square (Cor^2) and the cube (Cor^3) of these estimates were tested with the goal of maximizing the difference between the most similar and most dissimilar regions.

The results from both populations and for both traits (trait A with h^2=0.4 and trait B with h^2=0.1) are presented in Figure 22. As it can be noted, the model using the identity matrix performed slightly better than the Cor model, for both within and across breed predictions. Additionally, the Cor^2 and the Cor^3 models showed a further reduction of the accuracy compared to the previous model indicating that the parameterization of the prior distribution is not adequate. In fact, these results are in accordance with the results of Zhou *et al.* (2014) where they constructed a

weighted **G** matrix using the local persistency of LD across populations, calculated in the same way as CorLD but for genomic regions of just 5, 10 and 15 SNP markers.

Figure 22. *Accuracy of prediction within and between populations under genomic prediction models that use of local phase persistency.*



I matrix = Identity matrix, Cor, Cor^2 and Cor^3 matrix = diagonal matrices containing the CorLD estimates, the square and the cube of those estimates, respectively, gen 0, 1, 2 and 3 = distance of the validation set from the training set in generations. AV= Asturiana de los Valles, ANI= Avileña-Negra Iberica.

Conclusions

From this study it is clear that the two procedures of estimating the local persistency of LD between populations (CorLD and VarLD) yielded dissimilar results. That is because these methods are designed for different purposes. On one hand, the VarLD method is more efficient in detecting the differences between LD persistency and it is more sensible to a strong divergence between r^2 in a single, or few, pairs of markers within a genomic region, whereas, the CorLD method is less dependent on single estimates of LD and thus it is able to identify the ones where the persistency of the LD phase is maintained on average. Some genomic regions detected by the CorLD methodology were coincident between populations. The metabolic pathways identified for these regions were associated with the *Protocadherin* gene family on chromosome 7. The integration of the estimates of local LD persistency into the across-breed genomic evaluation showed no improvement in the accuracies indicating that the use of this information is not as straightforward.

CHAPTER 4

On the haplotype diversity along the genome in the Autochthonous Spanish beef cattle populations.

Introduction

The advent of massive genotyping technology has allowed the use of genomic information for genome-wide association studies –GWAS- (Bush and Moore, 2011) and genomic prediction of breeding values denoted as 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 plentiful evidence that the structure of linkage disequilibrium is not homogeneous along the genome (Ardlie *et al.,* 2002; Mckay *et al.,* 2007). In fact, the genome can be parsed into haplotype blocks of variable length, as described in human (Daly *et al.,* 2001, Gabriel *et al.,* 2002) and cattle (Mokry *et al.,* 2014), caused by the presence of 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), that are regulated by molecular mechanisms like the protein coded by the *PR domain-containing 9* (PRDM9) gene (Baudat *et al*., 2010). The analysis of the haplotype diversity is a classical measure of genetic diversity that is reduced in genomic regions that harbor genes under selection (Garud *et al*., 2015).

The objective of this study in to analyze the haplotype diversity along the genome of the Spanish Autochthonous beef cattle populations using the *BovineHD Beadchip* with the aim of identifying genome regions with higher haplotype diversity.

Material

The data used in this study comprised of the *BovineHD Beadchip* genotypes of the 342 founder individuals of the triplets described in the MATERIAL section (*Asturiana de los Valles* – AV-, N=50, *Avileña - Negra Ibérica* – ANI-, N=48, *Bruna dels Pirineus* – BP-, N=50, *Morucha* –Mo-, N=50, *Pirenaica* –Pi-, N=48, *Retinta* – Re-, N=48 and *Rubia Gallega* –RG, N=48).

Methods

The haplotypes of the parental chromosomes were established using two alternative software: BEAGLE (Browning and Browning, 2009) using the "TRIO" option, and SHAPEIT v2 (Delaneau *et al.,* 2013). Once the paternal and maternal haplotypes were defined for each genomic region, we calculated the number of haplotypes as the number of distinct haplotypes for a region, after phase reconstruction, defined by a given number of SNP or a map distance. In fact, the number of haplotypes was calculated for genomic regions defined by the number of SNPs (25, 50, 100 and 250) or by the genomic distance in kb (100, 250, 500 and 1,000).

Results and Discussion

The concordance between phases generated by BEAGLE and SHAPE-IT programs was very high (over 99.9% for all populations and chromosomes). Thus, we decided to present the analysis using exclusively the results provided by BEAGLE. This strong coincidence between software outputs confirms the robustness of haplotype phase reconstruction for trio families (Marchini *et al.,* 2006).

The results of the haplotype diversity along the genome for a meta-population composed by the seven Autochthonous beef cattle populations are presented in Figures 23 and 24 for regions of constant size (Figure 23) and fixed number of SNPs (Figure 24).

Figure 23 *Haplotype diversity along genomic regions of constant size for the meta-population of the seven autochthonous beef cattle populations.*

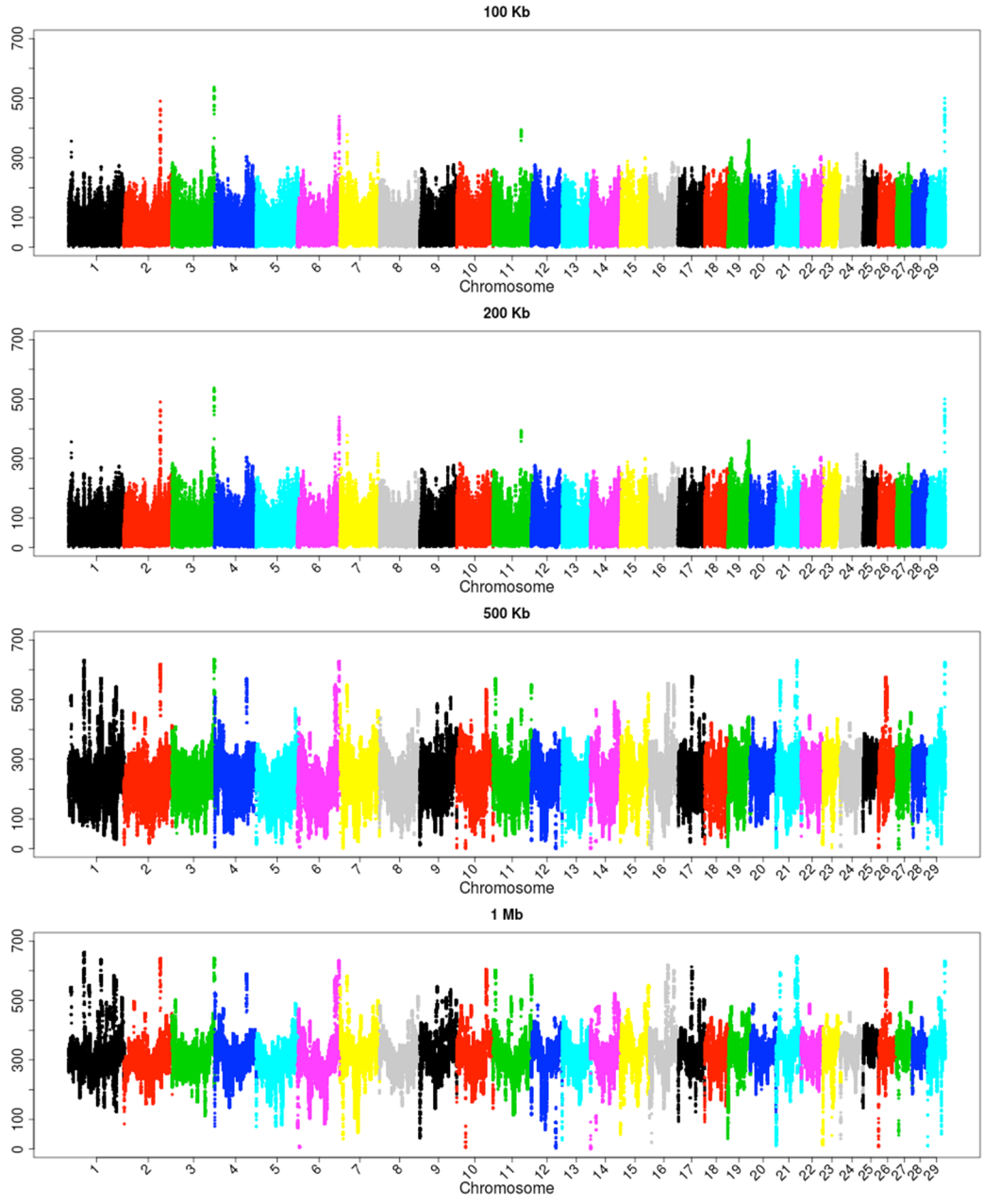
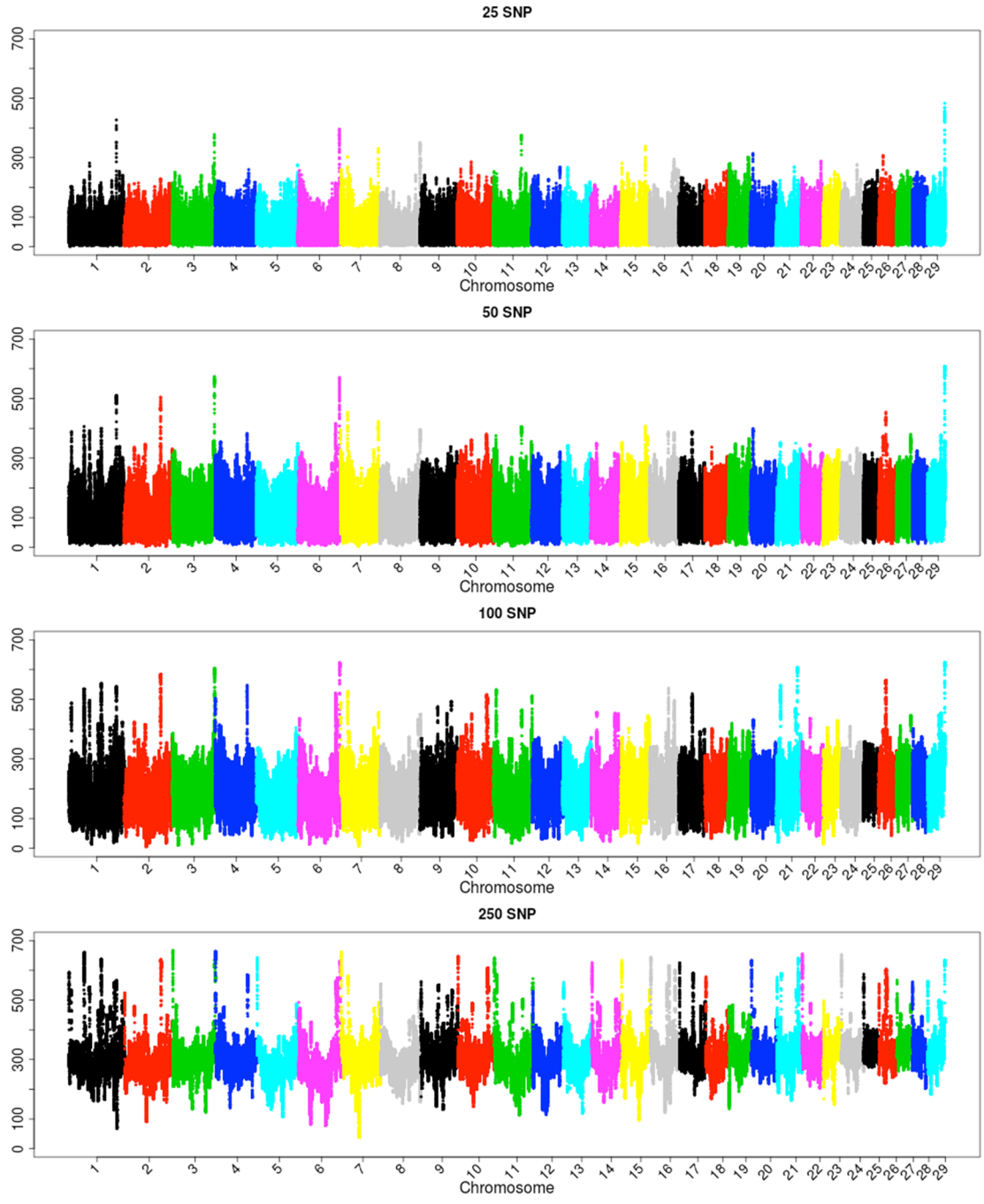


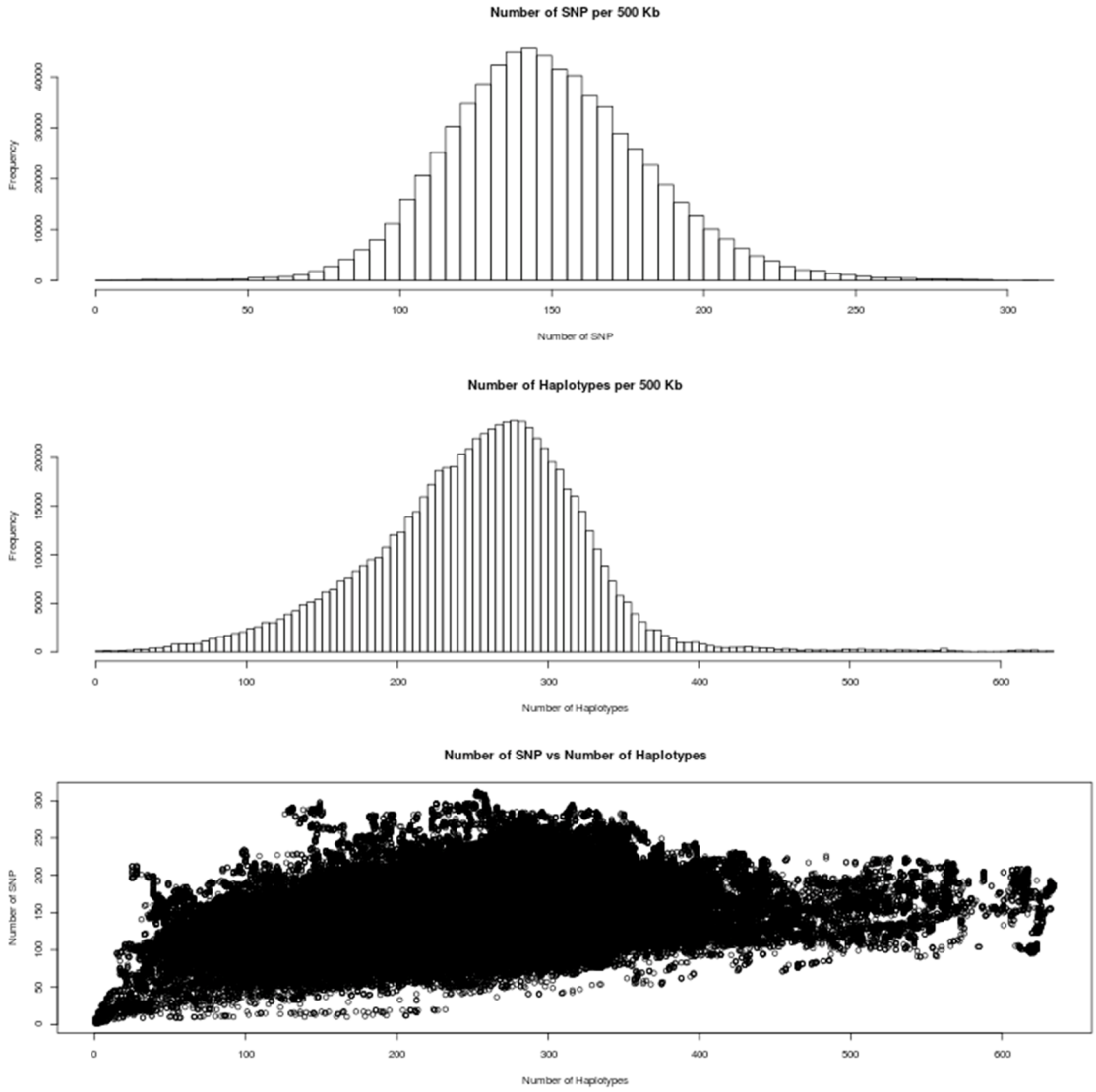
Figure 24. *Haplotype diversity along genomic regions of constant number of SNPs for the meta-population of the seven autochthonous beef cattle populations.*



As it can be observed, the results were similar. However, when the size (100 kb or 250 kb) or the number of SNPs (25 or 50 SNPs) were smaller, the results do not allow to identify clearly the variability in the number of haplotypes. On the other hand, the results from the analysis of wider genomic regions (1 Mb or 250 SNPs) provide a higher number of haplotypes with averages of 398.84 and 316.15 and standard deviations of 63.32 and 58.79, respectively. In fact, we think that intermediate windows (500 kb or 100 SNPs) provided a clearer picture of haplotype diversity, because the analysis of wider genomic regions can blur local signals. Both strategies have been used in the literature, although reconstruction of haplotypes using a constant number of SNPs is more frequent in simulation studies (Calus *et al.,* 2008). In fact, the results of both procedures were quite similar, with a correlation of 0.65 between the estimates of the number of haplotypes centered at each SNP marker. However, as the aim of this study is to analyze the haplotype diversity along the genome, we think that the use of a constant size of genome would provide a more accurate picture of haplotype diversity, caused by both modifications of mutation or recombination rate, and by the presence of selection processes. Thus, since now we will refer exclusively to the results generated by the analysis of a constant map segment (500 kb).

In first place, we analyzed the distribution of the number of SNPs present within these genomic regions of 500 kb (Figure 25, upper graph). We found that they follow an almost perfect Gaussian distribution with an average of 149.21 and standard deviation of 33.22, confirming the appropriateness of the SNP selection when the *BovinedHD Beadchip* was constructed.

Figure 25. *Distribution of the number of SNPs and Haplotypes and the relationship between them.*

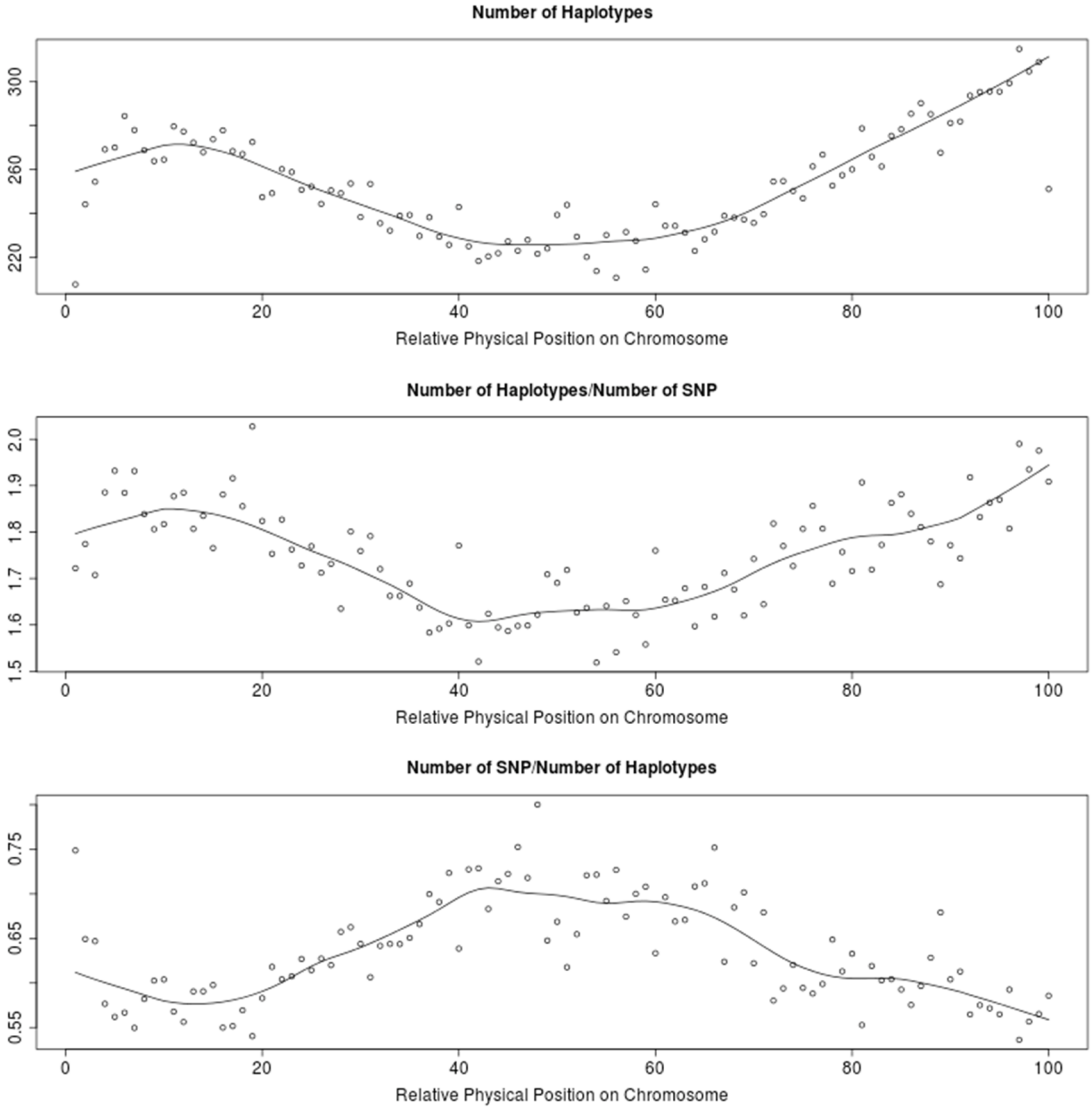


On the contrary, the distribution of the number of haplotypes within those genomic regions presents a clear asymmetry, as it has a long positive tail (Figure 25, middle graph). This fact indicates that the haplotype diversity is substantially higher in some regions of the genome. As expected, we found a positive relationship between the number of haplotypes and SNPs present in each specific region of the genome.

However, as it can be observed in Figure 25 (bottom graph), the genomic regions associated with a higher degree of haplotype diversity are not those with a higher number of SNPs, indicating that the presence of a large number of haplotypes is not a consequence of the overrepresentation of SNP markers.

Further, we analyzed the distribution of the haplotype diversity across the relative position within a chromosome (Figure 26, upper graph). As it can be observed, the results were as expected, and the haplotype diversity was higher in the genomic regions closer to the telomeres and lower in the central part of the chromosomes. The shape of the figure is almost equivalent to the one presented by Ma *et al.* (2015) for the male recombination rate. However, these authors found a decline of the female recombination rate at the distal part of the chromosomes.

Figure 26. *Haplotype diversity across the relative physical position within a chromosome*



Further, the results of the haplotype diversity within genomic regions of 500 kb for each of the seven analyzed population are presented in Figure 27, and the regions with a haplotype diversity within the top 1% are represented in Figure 28 and in Table 4.1 of ANNEXE 4.

Figure 27. *Haplotype diversity along the autosomal genome of seven Spanish autochthonous beef cattle populations for regions of 500 kb.*

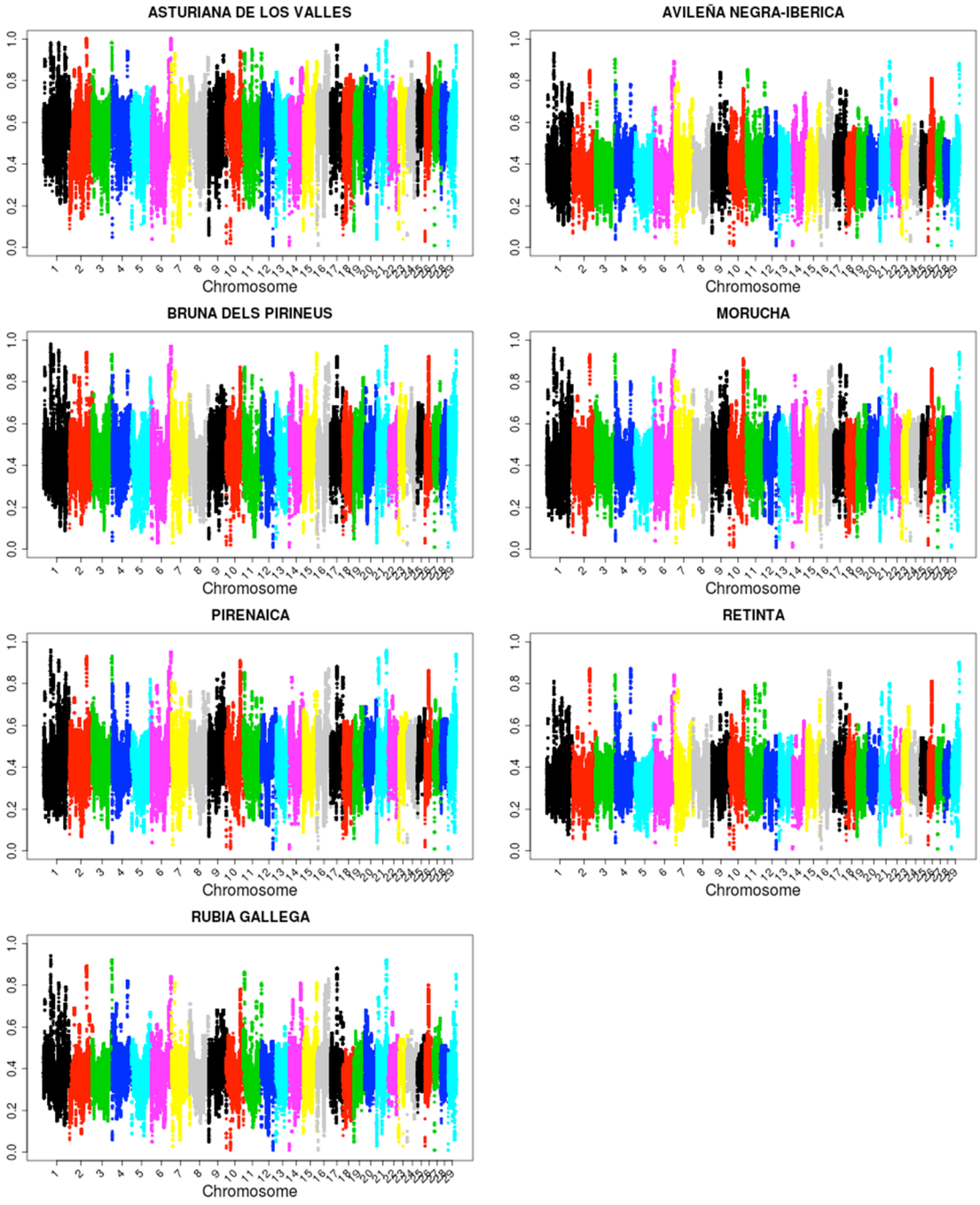
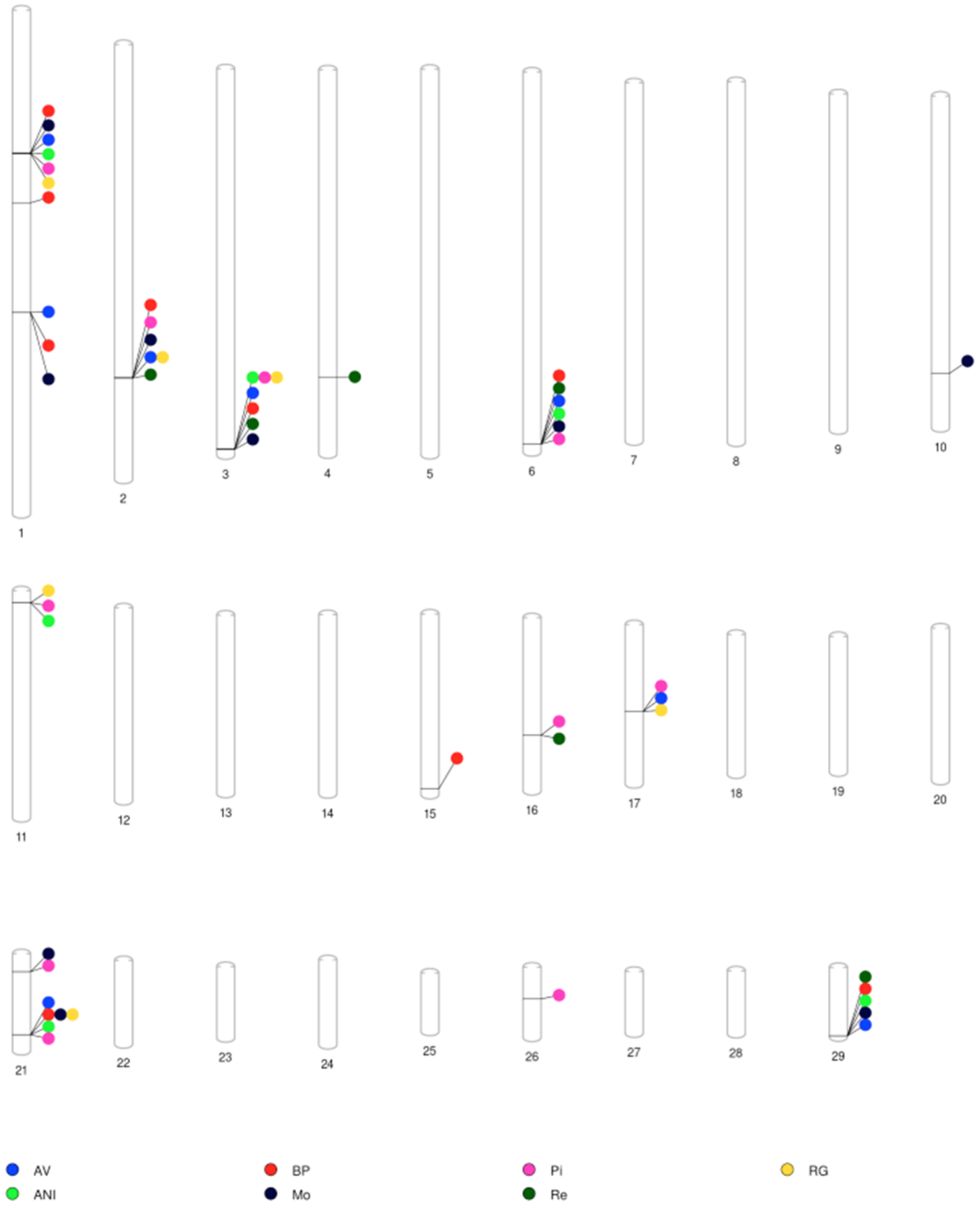
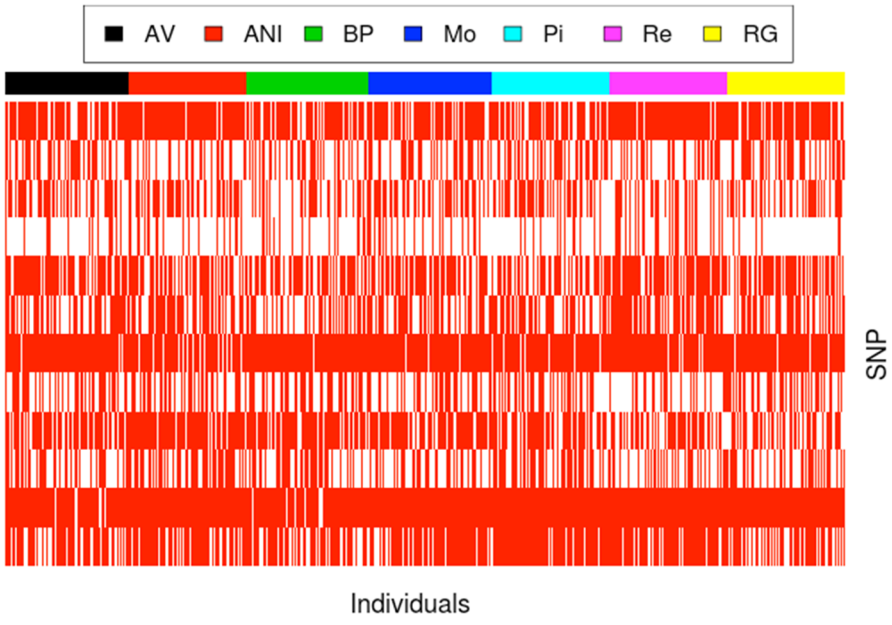


Figure 28. *Genomic regions identified with a number of haplotypes over an empirical top 1% for each of the populations.*



The location of these highly diverse regions is most frequent close to the telomeres, but some regions within the central part of the chromosomes can be also observed, such as in BTA1, BTA9, BTA11 and BTA29. As it can be observed, genomic regions with higher haplotype diversity are highly conserved across populations suggesting that the reason for that haplotype diversity is mainly structural, and probably associated with the recombination or mutation rate, because the higher the recombination or the mutation rate, the higher haplotype diversity is expected. As an example of this haplotype diversity, the specific haplotype configuration within a genomic region of the chromosome 3 is presented in Figure 29.

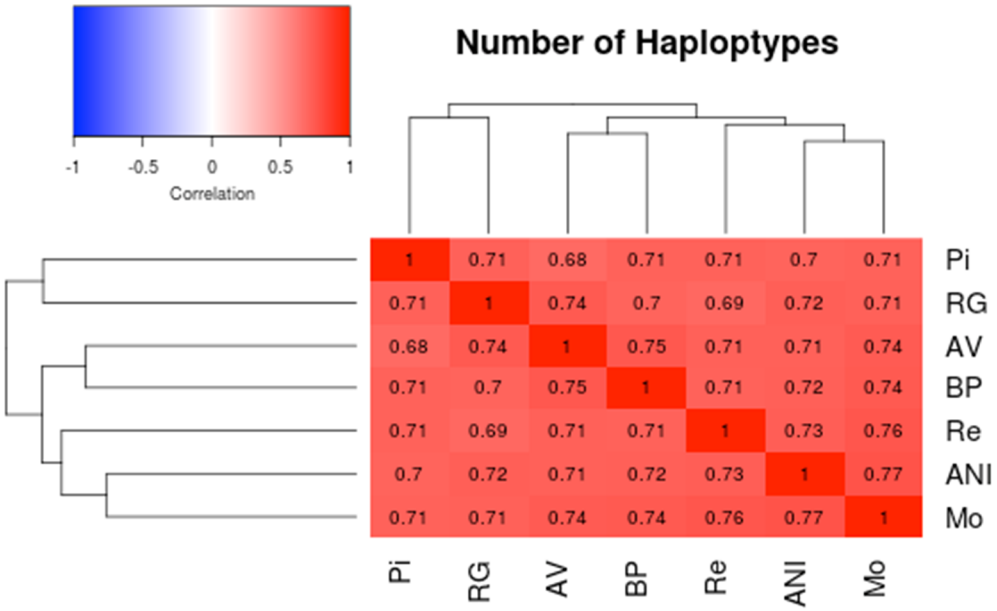
Figure 29. *Haplotype configuration of the 342 analysed phases in the genomic region of the chromosome 3 (119233792-119700584).*



Another probe of the concordance between populations can be observed in Figure 30, where the correlation between the number of haplotypes identified at each population is presented. These correlations ranged between 0.66 and 0.77.

However, the correlations are slightly higher between Re, Mo and AVI, that were previously identified as more genetically related by Cañas-Álvarez *et al.* (2015).

Figure 30. *Correlations between the number of haplotypes in the seven analyzed populations.*



The haplotype diversity is strongly related with the LD, and the results here presented confirm a strong heterogeneity of haplotype diversity along the genome. The applications that most frequently use genomic information, like GWAS (Bush and Moore, 2011) or GS (Meuwissen *et al.,* 2001), are based on the presence of linkage disequilibrium between SNP markers and QTLs. The aim of the procedures is to identify genomic regions associated with the variability of traits or to predict the breeding value of candidates of selection, respectively. However, the genomic regions with higher haplotype diversity are associated with a lower LD and thus,

genes of interest potentially located in those 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.

Conclusions

The results of this study confirm that the haplotype diversity is strongly variable along the cattle genome. Further, the comparison of the haplotype diversity among the seven analyzed populations suggest that the causes of this variability are mainly structural and probably associated with a higher recombination or mutation rate. The consequences of this variability in the application of genomic analysis, like genome-wide association studies or genomic selection should be studied in further research.

GENERAL DISCUSSION

Genomic Selection has been a great success for the dairy cattle industry (Hayes *et al.,* 2009) and its application in other species, like pigs (Brune, 2011; Forni *et al.,* 2011; Ostersen *et al.,* 2011), has been gradually growing during the last decade. However, its introduction in the beef cattle industry is not as straightforward as previously thought. Several factors exist that impede the uptake of this methodology by the breeders. Firstly, numerous breeds and crossbreds exist, each with limited census and with their own breed-specific attributes. Secondly, the restricted use of artificial insemination in these populations has as a result the poor connectedness among populations across countries, which, in turn, impedes the establishment of international collaborations. Thirdly, phenotyping strategies and sire recording tend to be poorer than those of dairy cattle, especially for commercial populations, and where multi-sire mating is practiced. And finally, the low-margin business model of the beef cattle industry gives little motivation for investment leading that way to poor adoption and low gains. In turn, this fact impedes the growth of the reference population necessary to improve the accuracy of the predictions and, thereinafter, to advance gain in beef (Berry *et al.,* 2016).

The autochthonous Spanish beef cattle populations have a limited census, and their effective sizes range between 50 and 60 (Cañas-Álvarez *et al.,* 2016). However, these populations play a crucial role in the maintenance of the economic activity of the rural population and, generally, provide high quality products of protected geographical indication or designation of origin. Nowadays, their breeding programs (Serradilla, 2008) are based on BLUP genetic evaluations of direct and maternal effects for growth traits and morphology, and, only in some cases, they also include carcass, meat quality or reproductive traits. Finally, the use of molecular information

is restricted to major genes, like MTSN in *Asturiana de los Valles* or for paternity checks. Thus, the main objective of this study was to investigate the potential application of Genomic Selection in these populations under both single and multiple population approaches.

In the first chapter we investigated the efficiency of the application of GS in two Autochthonous Spanish beef cattle populations (Pirenaica -Pi- and Rubia Gallega -RG-), as representatives of alternative genealogical structures, due to the wider implantation of artificial insemination in RG. The method of choice was the single-step approach (Aguilar *et al.,* 2010), because it's plausible that only a small percentage of the population could be genotyped in a short or medium range of time. Other kind of procedures, such as Bayes Cπ (Habier *et al*., 2011), Bayesian Lasso (De los Campos *et al*., 2009) and non-parametric procedures (González-Recio *et al*., 2014) require the establishment of a huge reference population that cannot be achieved by the local populations due to both limited census and economical limitations.

The first result of this study probes the appropriateness of the implementation of GS in the Spanish autochthonous populations, although the genotyping efforts that can be achieved by the breeders associations are intermediate or low. This is a very important advantage of the single-step approach as it combines genomic and pedigree based relationships into the same relationship matrix (Legarra *et al.,* 2009; Aguilar *et al.,* 2010), with respect to the two-step approach of genomic selection (Meuwissen *et al.,* 2001) that requires a minimum number of genotyped and phenotyped individuals to compete with the pedigree-based approaches.

Secondly, and with respect to the genotyping strategies, the accuracy improved with the number of individuals with genotypes introduced into the genomic evaluation, but the gains were worthy only in the case where the candidates to selection were genotyped. This fact implies that genotyping must be implemented routinely for the new candidates to selection, and that a single genotyping effort but not continuous in subsequent years cannot provide barely any increase in the accuracy of future individuals.

Moreover, although the most informative individuals (with lower PEV) provided better accuracies, the results showed that suboptimal genotyping strategies, due to limitations of the availability of biological samples or older individuals, are robust enough to achieve similar gains. Some of the Spanish local beef cattle populations (*Asturiana de los Valles*, *Avileña Negra-Ibérica* and *Retinta*) have been making a systematic collection of biological samples, while for other populations, the availability of biological samples of historical individuals is sparse and only restricted to sires used in AI. The main consequence of this result is that even when samples from older individuals are not available, substantial benefits from GS can be achieved similarly.

However, and although the genealogical information and the structure of phenotype information is considered in the simulation, the results of this study are constricted to the parameters of the simulation. It is not possible to consider in a simulation study all the possible variations or effects that may affect the output. Nevertheless, in order, to provide a broader view to the results of the study, a sensitivity analysis was performed. This analysis included the effects of the marker density, the effective size of the historical population and the mutation rate. As expected, the accuracy

increased with the marker density but reached a plateau around 50,000 SNPs confirming the postulates of Cañas-Álvarez *et al.*, (2016) that suggest that the Spanish autochthonous beef cattle populations need at least 38,000 segregating SNP markers. Thus, the potential increase that can be obtained from higher densities can be considered negligible as suggested also by Solberg *et al.*, (2008), even for unrelated individuals (Meuwissen, 2009).

The effective population size also showed to have an effect on the gains of accuracy as predicted by Solberg *et al.* (2008). The Ne of the Spanish autochthonous populations has been estimated around 50-60 (Cañas-Álvarez *et al.*, 2016). Therefore, and since our base scenario assumed a Ne of 100, the results obtained can be considered as a conservative estimate of the potential increase of accuracy. Likewise the mutation rate assumed in this study was extremely higher than estimations in the literature (Kumar and Subramanian, 2001; Hodgkinson and Eyre-Walker, 2011), which also leads to conservative estimates of gains.

Finally, it is important to mention that the efficiency of GS is higher in RG than in Pi, because of the genealogical structure that is provided by the broader implantation of AI in the RG breed. Consequently, a parallel increase of the rate of AI with the genotyping efforts will lead to a greater success of GS in populations with a low percentage of AI. Hence, in our view, the implementation of GS in the Spanish Autochthonous populations can be achieved by genotyping the candidates to selection of each generation and building gradually a reference population which will eventually include the most informative individuals. Finally, 50,000 SNP markers are sufficient to achieve high accuracies but the expansion of the application of AI is essential in order to maximize the gains.

The second study of this thesis intended to investigate the potential application of GS under a multi-breed model. The local Spanish beef cattle populations do not have the ability to construct a genotyped reference population similar to those obtained in international dairy or beef cattle populations and some authors (de Roos *et al.,* 2009) have proposed to pool animals from different populations in order to increase the size of that reference population and therefore improve the accuracy of the predictions. The cases of simulation studied here included single breed and multi breed genomic evaluations under several scenarios of alternative genetic architectures of the traits. The within breed predictions obtained from single breed evaluations resulted in the highest accuracies which are in concordance with the results of other studies (Saatchi *et al.*, 2011; Van Eenennaam *et al.*, 2014), The across breed predictions, though, resulted very low confirming the postulate of Harris *et al.* (2008) indicating that training in one population and validating in another is not effective. However, it is remarkable that all the average estimates are positive and the results are coherent with the studies of persistence of LD phase by Cañas-Alvarez *et al.* (2016) in the same populations. The accuracy dropped in the subsequent generations up to 40% (generation 3) for the within breed predictions as expected confirming the relevance of the relationship between the testing and training populations on the accuracy of GS (Clark *et al.,* 2012).

Moreover, the results obtained from the admixed reference populations yielded accuracies lower than those from single breed evaluation when the size of the training sets was equal (0.639 - 0-680 for the admixed x2 sets and 0.436 – 0.503 for the admixed x7 set). Note that the results from the admixed x7 training set in generation 0 were even lower than those that can be obtained using simple

individual selection (0.632 for h^2=0.4 and 0.316 for h^2=0.1), confirming that genomic selection requires a large reference population to be effective (Daetwyler *et al.,* 2008). Nonetheless, the admixed populations showed a small advantage in their predictive ability over the generations when compared to reduced pure-bred training sets. The most probable cause of this phenomenon is that as the relatedness between the training set and the validation set weakens, the predictions are based more on the short range LD between the markers and the genes than on the pure genetic relationship between individuals. Thus, the admixed populations perform better because of the higher number of data and the fact that mixing data breaks down the long distance LD created by relatedness and leaves the effects of the short range LD that persists through generations (Falconer and McKay, 1996).

In addition, when alternative genetic architectures of the traits were tested, the scenarios yielded similar results for all training sets with the exception of the case of rare variants. The rare variants controlling the trait of interest resulted in slightly lower accuracies in all cases and slightly higher loss of accuracy with the number of generations. This phenomenon is coherent with the results obtained by Wientjes *et al.* (2015), that indicated that when the QTLs controlling the genetic variability of the traits have lower frequencies the ability of prediction of Genomic Selection in lower. However, although this situation has been suggested to be the cause of the missing heritability (Gibson, 2012), the evidence of the percentage of genetic variation that rare variants produce is low and some authors have shown that these rare variants explain a small percentage of the missing heritability of complex traits in human (Gusev *et al.,* 2014) or cattle (Gonzalez-Recio *et al.,* 2016).

Theoretically, the success of the genomic evaluation from multiple populations is linked to the persistency of the linkage disequilibrium (LD) between populations in such way that the LD between markers and QTLs is maintained. Different populations may have different linkage phases for each specific genomic region, and therefore, it might be beneficial to include information about the persistency of the linkage phase between populations in the model of genomic evaluation. The third chapter of this thesis attempted to analyse the genetic architecture of the persistency of LD across the seven Spanish beef cattle populations. The two methods tested (CorLD and VarLD) yielded dissimilar results with a correlation between them of only 0.146 ± 0.050. This low correlation indicates that the signals of persistency of phase were clearly different. In fact, VarLD was designed (Teo *et al.*, 2009) for detecting differences and it is more sensible to a strong divergence between r^2 in a single, or few, pair of markers within a genomic region. Whereas, CorLD is more robust to few outlier correlations as it is calculated as a correlation of correlation estimates, that is less dependent on single estimates of LD. Thus, VarLD is probably more capable to detect genomic regions that diverge between populations, whereas CorLD is able to identify the ones where the persistency of the LD phase is maintained on average.

However, both procedures classified the populations in a similar way, and in agreement with the results of divergence between these populations calculated using principal components (Cañas-Álvarez *et al.*, 2015) or phase persistency (Cañas-Álvarez *et al.*, 2016). These results are also in concordance with the traditional classification of the Spanish cattle populations (Sanchez-Belda, 1984)

and with their geographical localization (Re, Mo and ANI in central and south of Spain and AV, RG and BP in the north).

Moreover, the main metabolic pathways associated to the genes located in regions with higher persistency for CorLD, included processes of cell adhesion, synapse assembly and organization and nervous system development, all associated with the *Protocadherin* gene family, which was also detected previously by Su *et al.*, (2014) in a haplotype diversity analysis. On the contrary, no significant pathways were found for VarLD, confirming that the procedure is more able to detect differences between populations than persistency of LD.

Finally, the information on the local persistency of LD obtained from CorLD was incorporated in the GS model in order to check its effects on the accuracy of the across breed predictions. All alternative models tested resulted in slightly lower accuracies than those obtained from the model using the identity matrix indicating that the parameterization of the prior distribution of the marker effects was not adequate. Similar results were obtained by Zhou *et al.* (2014) after introducing local estimates of LD persistency as weights to build an alternative **G** matrix. These results indicate that further research must be done in order to incorporate this kind of information in GS models.

Finally, as a side result, we also performed a haplotype diversity analysis along the genome of the Spanish populations. From the results of the study, we observed that the haplotype diversity is substantially higher in some regions of the genome and, as expected, is higher near the telomeres and lower in the central part of the chromosome, as previously reported by Ma *et al.* (2015) in cattle populations. Nevertheless, it is important to mention that the genomic regions with the highest

haplotype diversity are greatly conserved across populations. This probably suggests that the reason for that haplotype diversity is mainly structural, and associated with the recombination or mutation rate. The most used applications of genomic information, like GWAS (Bush and Moore, 2011) or GS (Meuwissen *et al.,* 2001), are based on the presence of linkage disequilibrium between SNP markers and QTLs. The aim of the procedures is to identify genomic regions associated with the variability of traits or to predict the breeding value of candidates to selection, respectively. However, the genomic regions with higher haplotype diversity are associated with lower LD and thus, genes of interest potentially located in those regions could be blurred under the standard procedures. Further research must be done to modify current procedures of GWAS or GS to incorporate structural information of the haplotype diversity in each specific region of the genome.

The main conclusion of this study is that GS is feasible in the Spanish local cattle populations. In fact, the application of single step procedures of GS makes its implementation beneficial for the accuracy of predictions from the first genotyping effort and even when the distribution of genotyped individuals is suboptimal. Moreover, the use of information from other populations can also increase the accuracy of prediction, especially for the individuals with loose genetic links to the phenotyped individuals. These results suggest that this procedure can be more beneficial for traits whose recording is made only in research experiments and not regularly in the standard development of the breeding programs. Finally, the last two studies indicate that there is a lot of additional information about the persistency of LD or the haplotype diversity that have the potential of being incorporated into the procedures of GS. Nevertheless, their integration in GS methods is not an easy task,

as we probed with the straightforward approximation that was tested in this thesis and further research is required.

CONCLUSIONS

1. The implementation of Genomic Selection under the single-step approach increases the accuracy of prediction over the standard BLUP even when few individuals were genotyped.
2. The increase of accuracy is worthy only when the candidates to selection are genotyped.
3. The results of accuracy are robust to variations in the genotyping strategy.
4. The efficiency of Genomic Selection is higher when the implantation of artificial insemination is broader.
5. The use of admixed populations for Genomic Selection provides a small advantage over a single population genomic evaluation when predicting individuals that are more genetically distant from the training set.
6. The VarLD method is more efficient in detecting the differences between LD, whereas the CorLD detect better the persistency of haplotype phase.
7. The metabolic pathways identified for genomic regions with high persistency were associated with the *Protocadherin* gene family
8. Haplotype diversity is strongly variable along the cattle genome, though the comparison among the seven analyzed populations suggests that the causes of this variability are mainly structural and probably associated with a higher recombination or mutation rate.

CONCLUSIONES

1. La implementación de la Selección Genómica bajo la aproximación “Single-Step” incrementa la precisión sobre la valoración BLUP incluso cuando se genotipa un conjunto pequeño de individuos.
2. El incremento de precisión es relevante exclusivamente para los candidatos a la selección genotipados.
3. Los resultados de precisión son robustos ante diseños alternativos de genotipado.
4. La eficiencia de la selección genómica es mayor a medida que se incrementa la implantación de la inseminación artificial.
5. La utilización de meta-poblaciones proporciona una ligera ventaja sobre el análisis en población única, que es más evidente cuando a medida que los candidatos a la selección están más alejados de los individuos utilizados en la valoración genómica.
6. El procedimiento VarLD es más eficiente para detectar diferencias de desequilibrio de ligamiento entre poblaciones, mientras que el método CorLD identifica mejor la persistencia de la fase haplotípica.
7. Las rutas metabólicas asociadas a las regiones genómicas con mayor persistencia de la fase haplotípica están asociadas con la familia génica de las Protocaderinas. (*Protocadherin*).
8. La diversidad haplotípica es muy variable a lo largo del genoma, aunque la comparación entre las siete poblaciones analizadas sugieren que las causas de esta variabilidad son en mayor medida estructurales, y asociadas probablemente a las tasas de mutación y recombinación.

REFERENCES

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J., 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, *93*(2), 743–52.

Aguilar, I., Misztal, I., Legarra, A., & Tsuruta, S., 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal Breeding and Genetics*, *128*(6), 422–428.

Ardlie, K.G., Kruglyak, L., & Seielstad, M., 2002. Patterns of Linkage Disequilibrium in the Human Genome. *Nature Reviews. Genetics* 3 (4):299–309.

Baloche, G., Legarra, A., Sallé, G., Larroque, H., Astruc, J.-M., Robert-Granié, C., & Barillet, F., 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. *Journal of Dairy Science*, *97*(2):1107–1116.

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., & de Massy, B., 2010. PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice" *Science (New York, N.Y.)* 327 (5967): 836–40.

Beja-Pereira, A., Alexandrino, P., Bessa, I., Carretero, Y., Dunner, S., Ferrand, N., Jordana, J., Laloe, D., Moazami-Goudarzi, K., Sanchez, A., Cañon, J., 2003. Genetic Characterization of Southwestern European Bovine Breeds: A Historical and Biogeographical Reassessment with a Set of 16 Microsatellites. *Journal of Heredity* 94 (3): 243–50.

Benjamini, Y., & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B* 57 (1): 289–300.

Berry, D. P., Garcia, J. F., & Garrick, D. J., 2016. Development and implementation of genomic predictions in beef cattle. *Animal Frontiers*, *6*(1), 32.

Bolormaa, S., Pryce, J. E., Kemper, K., Savin, K., Hayes, B. J., Barendse, W., Zhang, Y., Reich, C. M., Mason, B. A., Bunch, R. J., Harrison, B. E.,

Reverter, A., Herd, R. M., Tier, B., Graser, H. U., Goddard, M. E., 2013. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle. *Journal of Animal Science*, *91*(7), 3088–3104.

Brøndum, R. F., Rius-Vilarrasa, E., Strandén, I., Su, G., Guldbrandtsen, B., Fikse, W. F., & Lund, M. S., 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. *Journal of Dairy Science*, *94*(9), 4700–4707.

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 (5): 1084–97.

Bush, W.S., & Moore, J.H., 2012. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology* 8 (12): e1002822.

Calus, M P L, Meuwissen, T.H.E., De Roos, A.P.W., & Veerkamp, R.F., 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics* 178 (1): 553–561.

Cañas-Álvarez, J. J.; González-Rodríguez, A.; Munilla, S.; Varona, L.; Díaz, C.; Baro, J. A.; Altarriba, J.; Molina, A.; Piedrafita, J., 2015. Genetic diversity and divergence among Spanish beef cattle breeds assessed by a bovine high-density SNP chip. *Journal of Animal Science* 93(11):5164-5174

Cañas-Alvarez, J., Mouresan, E., Varona, L., Diaz, C., Molina, A., Baro, J.A., Altarriba, J., Carabano, M., Casellas, J., Piedrafita, J., 2016 Linkage disequilibrium, persistence of phase and effective population size in Spanish local beef cattle breeds assessed through a SNP high density chip. *Journal of Animal Science* (in press)

Casellas, J., & Piedrafita, J., 2015. Accuracy and expected genetic gain under genetic or genomic evaluation in sheep flocks with different amounts of pedigree, genomic and phenotypic data. *Livestock Science*, *182*, 58–63.

Castillo-Juárez, H., Campos-Montes, G. R., Caballero-Zamora, A., & Montaldo, H. H., 2015. Genetic improvement of Pacific white shrimp [Penaeus (Litopenaeus) vannamei]: perspectives for genomic selection. *Frontiers in Genetics*, *6*(March), 93.

Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A., & Muir, W. M., 2011. Effect of different genomic relationship matrices on accuracy and scale. *Journal of Animal Science*, *89*(9):2673–2679.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., Ma'ayan, A., 2013. Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool. *BMC Bioinformatics* 14: 128.

Chen, L., Schenkel, F., Vinsky, M., Crews, D. H., & Li, C., 2013a. Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. *Journal of Animal Science*, *91*(10):4669–4678.

Clark, S., Hickey, J., & van der Werf, J., 2011. Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution*, *43*(1), 18.

Clark, S.A., Hickey, J.M., Daetwyler, H.D., & van der Werf, J.H., 2012. The Importance of Information on Relatives for the Prediction of Genomic Breeding Values and the Implications for the Makeup of Reference Data Sets in Livestock Breeding Schemes. *Genetics Selection Evolution* 44 (1): 4.

Cleveland, M., Forni, S., Deeb, N., Maltecca, C., 2010 Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. *BMC Proceedings* 4, S6.

Coop, G., & Przeworski, M., 2007. An Evolutionary View of Human Recombination. *Nature Reviews. Genetics* 8 (1). Nature Publishing Group: 23–34.

Croiseau, P., Legarra, A., Guillaume, F., Fritz, S., Baur, A., Colombani, C., Robert-Granié, C., Boichard, D., Ducrocq, V., 2011. Fine tuning genomic evaluations

in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genetics Research*, *93*(06):409–417.

Daetwyler H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D., Chamberlain, A.J., Anderson, C., VanTassell, C.P., Hulsegge, I., Goddard M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D.A., Fries, R., Hayes B.J., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet*. 46(8):858-865.

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A., 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, *185*(3):1021–1031.

Daetwyler, H. D., Swan, a. a., van der Werf, J. H. J., & Hayes, B. J., 2012. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genetics, Selection, Evolution : GSE*, *44*(1), 33.

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A., 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE*, *3*(10), e3395.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., & Lander, E.S., 2001. High-Resolution Haplotype Structure in the Human Genome. *Nature Genetics* 29 (2): 229–32.

De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375–385.

De Roos, A.P.W., Hayes, B.J. & Goddard, M.E., 2009. Reliability of Genomic Predictions across Multiple Populations. *Genetics* 183 (4): 1545–53.

De Roos, A.P.W., Hayes, B.J., Spelman, R.J., & Goddard, M.E., 2008. Linkage Disequilibrium and Persistence of Phase in Holstein-Friesian, Jersey and Angus Cattle. *Genetics* 179 (3): 1503–12.

Dekkers, J. C., 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. *Journal of Animal Science.* E313-328

Delaneau, O., Zagury, J.F., & Marchini, J., 2013. Improved whole chromosome phasing for disease and population genetic studies. *Nat Methods.* 10(1):5-6.

Duchemin, S I., Colombani, C., Legarra, A., Baloche, G., Larroque, H., Astruc, J M., Barillet, F., Robert-Granie, C., Manfredi, E., 2012. Genomic selection in the French Lacaune dairy sheep breed. *Journal of Dairy Science*, *95*(5), 2723–2733.

Erbe, M., Hayes, B J., Matukumalli, L K., Goswami, S., Bowman, P J., Reich, C M., Mason, B A., Goddard, M E., 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, *95*(7), 4114–4129.

Esfandyari, H., Sørensen, A. C., & Bijma, P., 2015. A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genetics Selection Evolution*, *47*(1), 76.

Falconer, D.S. and Mackay, T.F.C., 1996. *Introduction to quantitative genetics* (4th ed.). Harlow: Addison Wesley Longman Limited.

Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A.K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W.M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H.S., Ritchie, G.R.S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S.P., Aken,

B.L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T.J.P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., Searle, S.M.J., 2013. Ensembl 2013. *Nucleic Acids Research* 41 (D1): D48–55.

Forni, S., Aguilar, I., & Misztal, I., 2011. Different Genomic Relationship Matrices for Single-Step Analysis Using Phenotypic, Pedigree and Genomic Information. *Genetics Selection Evolution* 43 (1)1–7.

Fujii J., Otsu, K., Zorzato, F, De león, S, Khanna V. K., Weiler J. E., O' Brien P. J., MacLennan, D. H., 1991. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* 25: 448-451.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D.. 2002. The Structure of Haplotype Blocks in the Human Genome. *Science* 296 (5576): 2225–29.

Garud, N.R., Messer, P.W., Buzbas, E.O., & Petrov, D.A., 2015. Recent Selective Sweeps in North American Drosophila Melanogaster Show Signatures of Soft Sweeps. Edited by Gregory P. Copenhaver. *PLoS Genetics* 11(2): 1–32.

Gianola, D., 2013. Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, *194*(3), 573–596.

Gianola, D., Fernando, R. L., & Stella, A., 2006. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. *Genetics*, *173*(3), 1761–1776.

Gianola, D., Okut, H., Weigel, K. A., & Rosa, G. J., 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics*, *12*(1), 87.

Gibson, G., 2012. Rare and Common Variants: Twenty Arguments. *Nature Reviews Genetics* 13 (2):135–45.

Gil, M., Serra, X., Gispert, M., Oliver, M. A., Sañudo, C., Panea, B., Olleta, J. L., Campo, M., Oliván, M., Osoro, K., García-Cachán, M. D., Cruz-Sagredo, R., Izquierdo, M., Espejo, M., Martín, M., and Piedrafita. J., 2001. The effect of breed-production systems on the myosin heavy chain 1, the biochemical characteristics and the colour variables of longissimus thoracis from seven Spanish beef cattle breeds. *Meat Sci*. 58:181–188.

Goddard, M., 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, *136*(2), 245–257.

Gonzalez-Recio, O., Daetwyler, H.D., MacLeod, I.M., Pryce, J.E., Bowman, P.J., Hayes, B.J., Goddard, M. E., 2015. Rare Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the Missing Heritability of Complex Traits in Cattle. *PloS One* 10 (12): e0143945.

González-Recio, O., Gianola, D., Rosa, G.J.M., Weigel, K.A., & Kranis, A., 2009. Genome-Assisted Prediction of a Quantitative Trait Measured in Parents and Progeny: Application to Food Conversion Rate in Chickens. *Genetics, Selection, Evolution : GSE* 41 (1): 3.

González-Recio, O., Weigel, K. a, Gianola, D., Naya, H., & Rosa, G. J. M., 2010. L2-Boosting algorithm applied to high-dimensional problems in genomic selection. *Genetics Research*, *92*(3), 227–37.

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., Cambisano, N., Mni, M., Reid, S., Simon, P., Spelman, R., Georges, M., Snell, R., 2001. 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 Res*. 12(2):222-231.

Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., Su, G., 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genetics Selection Evolution,* 43(1), 43.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS, 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Genet* 37: 549–554.

Gunderson, KL, Steemers, FJ, Lee, G, Mendoza, LG, Chee MS., 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet*. 37:549-554.

Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjalmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., Kahler, A.K., Hultman, C.M., Purcell, S.M., McCarroll, S.A., Daly, M., Pasaniuc, B., Sullivan, P.F., Neale, B.M., Wray, N.R., Raychaudhuri, S., Price, A.L., 2014. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *American Journal of Human Genetics* 95 (5): 535–52.

Habier, D, Fernando, R L., & Dekkers. J C M., 2007. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* 177 (4): 2389–97.

Habier, D., Fernando, R. L., & Dekkers, J. C. M., 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, *177*(4), 2389–2397.

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., Johnson, D.L., Spelman, R.J., 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. In Proc. 36th ICAR Biennial Session, Niagara Falls, USA. p325.

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., & Goddard, M. E., 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics, Selection, Evolution : GSE*, *41*, 51.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E., 2009a. Invited review: Genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science*, *92*(2), 433–43.

Hazel, L. N. 1943. The genetic basis for constructing selection indices. *Genetics* 28: 476-490.

Helmer, D., Gourichon, L., Monchot, H., Peters, J. y Saña Segui, M., 2005. Identifying early domestic cattle from Pre-Pottery Neolithic sites on the Middle Euphrates using sexual dimorphism. In J.D. Vigne, D. Helmer, y J. Peters (Eds.), *The First Steps of Animal Domestication: New Archaeozoological Approaches* (pp. 86–95). London: Oxbow Books.

Henderson, C. R and Quaas, R. L. 1976. Multiple trait evaluation using relative's records. *J. Anim. Sci*. 43:1188-1197.

Henderson, C. R. 1973. Sire evaluation and genetic trends. In Proceedings of the animal breeding and genetics symposium in honor of Dr. Lush, pp 10-41. American Society of Animal Science, Champaign, IL.

Henderson, C. R. 1976. A simple method for the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32:69-83.

Henderson, C.R., 1984. *Applications of linear models in animal breeding*. Guelph: University of Guelph Press

Heslot, N., Yang, H.-P., Sorrells, M. E., &Jannink, J.-L. 2012. Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Sci.*, *52*(1), 146–160.

Hodgkinson A., Eyre-Walker, A., 2011. Variation in the mutation rate across mammalian genomes. *Nature Review Genetics* 12:756-766.

Hoerl, A. E., Kennard, R. W. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Tecnometrics* 12:55-67.

Hoffmann, I., 2010. Climate change and the characterization, breeding and conservation of animal genetic resources. *Animal Genetics*, *41*(Suppl. 1), 32–46.

Hongo, H., Pearson, J., Öksüz, B. y Ilgezdi, G., 2009. The process of ungulate domestication at Cayönü, Southeastern Turkey: a multidisciplinary approach focusing on Bos sp. and Cervuselaphus. *Anthropozoologica*, *44*, 63–78.

Huber, D., Peter, G.F., 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics*, *198*(4), 1759–1768.

Ibáñez, M. and Mas, B., 1997. Razas bovinas autóctonas de interés. En C. Buxadé (Ed.), *Vacuno de carne: aspectos clave* (pp. 115–133). Madrid: Mundi-Prensa.

Ibañez-Escriche, N., & Gonzalez-Recio, O., 2011. Review. Promises, pitfalls and challenges of genomic selection in breeding programs. *Spanish Journal of Agricultural Research*, *9*(2), 404–413.

Ibañez-Escriche, N., Forni, S., Noguera, J. L., Varona, L., 2014. Genomic information in pig breeding: Science meets industry needs. *Lives. Sci*. 166:94-100.

Illumina Inc., 2012. BovineHD Genotyping BeadChip Datasheet. Recuperado de http://www.illumina.com/documents/products/datasheets/datasheet_bovineHD.pdf

Kachman, S.D., Spanger, M.L., Bennett, G.L., Hanford, K.J., Kuehn, L.a., Snelling, W.M., Thallman, R M., Saatchi, M., Garrick, D.J., Schnabel, R.D., Taylor, J.F., Pollak, E J., 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genetics, Selection, Evolution : GSE*, *45*(1), 30.

Karoui, S., Carabaño, M. J., Díaz, C., & Legarra Albizu, A., 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genetics Selection Evolution*, *44*(1), 10.

Kennedy B. W. and Sorensen D. A. 1988. Properties of mixed-model methods for prediction of genetic merit. In B. S. Weir, E. J. Eisen, M. M. Goodman and G. Namkoong (eds). Proceedings of the second international conference of quantitative genetics, pp 91-103. Sinauer Assoc., Sunderland, MA.

Kennedy, B. W., Quinton, M., van Arendonk, J. A. M. 1992. Estimation of effects of single genes on quantitative traits. *J. Anim. Sci*. 70:2000-2012.

Khatkar, M. S., Moser, G, Hayes, B. J., Raadsma, H. W., 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. *BMC Genomics* 13:538

Kizilkaya, K, Fernando, R.L. & Garrick, D.J., 2010. Genomic Prediction of Simulated Multibreed and Purebred Performance Using Observed Fifty Thousand Single Nucleotide Polymorphism Genotypes. *Journal of Animal Science* 88 (2): 544–51.

Kumar, S, Subramanian, S., 2002. Mutation rates in mammalian genomes. *PNAS* 99:803-808.

Legarra, A, & Misztal, I., 2008. Technical Note: Computing Strategies in Genome-Wide Selection. *Journal of Dairy Science* 91 (1): 360–66.

Legarra, A., Baloche, G., Barillet, F., Astruc, J M., Soulas, C., Aguerre, X., Arrese, F., Mintegi, L., Lasarte, M., Maeztu, F., Beltrán de Heredia, I., Ugarte, E., 2014b. Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. *Journal of Dairy Science*, *97*(5),

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I., 2014a. Single Step, a general approach for genomic selection. *Livestock Science*, *166*(1), 54–65.

Legarra, A., iAguilar, I., & Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, *92*(9), 4656–4663.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., & Fritz, S., 2011. Improved Lasso for genomic selection. *Genetics Research*, *93*(1), 77–87.

Legarra, A., Robert-Granié, C., Manfredi, E., & Elsen, J.M., 2008b. Performance of Genomic Selection in Mice. *Genetics* 180 (1): 611–18.

Lillehammer, M., Meuwissen, T. H. E., & Sonesson, A. K., 2011. Genomic selection for maternal traits in pigs. *Journal of Animal Science*, *89*(12), 3908–3916.

Loberg, A., and J. W. Dürr., 2009. Interbull survey on the use of genomic information. *Interbull Bull.* 39:3–14.

Lourenco, D. A. L., Tsuruta, S., Fragomeni, B. O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J. K., Amen, T. S., Wang, L., Moser, D. W., Misztal, I., 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *Journal of Animal Science*, *93*(6), 2653.

Luan, T., Woolliams, J A., Lien, S., Kent, M., Svendsen, M., & Meuwissen, T H E., 2009. The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics* 183 (3): 1119–26.

Lund, M S, de Roos, A P W, de Vries, A G, Druet, T, Ducrocq, V, Guillaume, F, Guldbrandtsen, B, Liu, Z, Reents, R, Schrooten, C, Seefried, M, Su, G, 2010. Improving Genomic Prediction by EuroGenomics Collaboration. In *9th World Congress on Genetics Applied to Livestock Production, 1-6 August 2010, Leipzig, Germany*, 150.

Lund, M. S., Su, G., Janss, L., Guldbrandtsen, B., & Brøndum, R. F., 2014. Invited review: Genomic evaluation of cattle in a multi-breed context. *Livestock Science*, *166*(1), 101–110.

Ma, L., O'Connell, J.R., VanRaden, P.M., Shen, B., Padhi, A., Sun, C., Bickhart, D.M., Cole, J.B., Null, D.J., Liu, G.E., Da, Y., Wiggans, G.R., 2015. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLoS Genetics* 11 (11): e1005387.

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., Donnelly, P., 2006. A Comparison of Phasing Algorithms for Trios and Unrelated Individuals. *American Journal of Human Genetics* 78 (3): 437–450.

McKay, S.D., Schnabel, R.D., Murdoch, B.M., Matukumalli, L.K., Aerts, J., Coppieters, W., Crews, D., Dias Neto, E., Gill, C.a., Gao, C., Mannen, H., Stothard, P., Wang, Z., Van Tassell, C.P., Williams, J.L., Taylor, J.F., Moore, S.S., 2007. Whole Genome Linkage Disequilibrium Maps in Cattle. *BMC Genetics* 8: 74.

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

Meuwissen, T., Hayes, B., & Goddard, M., 2016. Genomic selection: A paradigm shift in animal breeding. *Animal Frontiers*, *6*(1), 6.

Meuwissen, Theo H E., 2009. Accuracy of Breeding Values of 'Unrelated' Individuals Predicted by Dense SNP Genotyping. *Genetics, Selection, Evolution : GSE* 41: 35.

Misztal, I, Legarra, A., & Aguilar, I., 2014. Using Recursion to Compute the Inverse of the Genomic Relationship Matrix. *Journal of Dairy Science* 97 (6): 3943–52.

Misztal, I., Legarra, A., & Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science*, *92*(9), 4648–4655.

Moen, T., Torgersen, J., Santi, N., Davidson, W.S., Baranski, M., Ødegård, J., Kjøglum, S., Velle, B., Kent, M., Lubieniecki, K.P., Isdal, E., Lien, S., 2015. Epithelial cadherin determines resistance to infectious pancreatic necrosis virus in atlantic salmon. Genetics 200(4):1313-1326.

Mokry, F., Buzanskas, M., de Alvarenga Mudadu, M., do Amaral Grossi, D., Higa, R., Ventura, R., de Lima, A., Sargolzaei, M., Conceição Meirelles, S., Schenkel, F., da Silva, M., Méo Niciura, S., de Alencar, M., Munari, D., de Almeida Regitano, L., 2014. Linkage Disequilibrium and Haplotype Block Structure in a Composite Beef Cattle Breed. *BMC Genomics* 15 (Suppl 7): S6.

Molina, A., 2010. Biodiversidad y conservación de razas autóctonas de animales domésticos. *Ambienta*, *91*, 109–125.

Moser, G., Tier, B., Crump, R., Khatkar, M., & Raadsma, H., 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution*, *41*(1), 56.

Mulder H.A., Calus, M.P.L, Druet, T., Schrooten, C., 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *J. Dairy Sci*. 95:876-889.

Muñoz, P.R., Resende, M.F.R., Gezan, S.A., Resende, M.D.V., de los Campos, G., Kirst, M., Huber, D., & Peter, G.F., 2014. Unraveling Additive from Nonadditive Effects Using Genomic Relationship Matrices. *Genetics* 198 (4):1759–68.

Myers, S., Bottolo, L., Freeman, C., McVean, G., & Donnelly, P., 2005. A Fine-Scale Map of Recombination Rates and Hotspots across the Human Genome. *Science (New York, N.Y.)* 310 (5746). American Association for the Advancement of Science: 321–324.

Olson, K. M., VanRaden, P. M., & Tooker, M. E., 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *Journal of Dairy Science*, *95*(9):5379-5383.

Ong, R.T.H., & Teo. Y.Y., 2010. VarLD: A Program for Quantifying Variation in Linkage Disequilibrium Patterns between Populations. *Bioinformatics* 26 (9): 1269–1270.

Ostersen, T., Christensen, O.F., Henryon, M., Nielsen, B., Su, G., & Madsen, P., 2011. Deregressed EBV as the Response Variable Yield More Reliable Genomic Predictions than Traditional EBV in Pure-Bred Pigs. *Genetics Selection Evolution* 43 (1): 38.

Paigen, K., & Petkov, P., 2010. Mammalian Recombination Hot Spots: Properties, Control and Evolution. *Nature Reviews. Genetics* 11 (3): 221–233.

Park, T., & Casella, G., 2008. The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Piedrafita, J., R. Quintanilla, C. Sañudo, J. L. Olleta, M. M. Campo, B. Panea, G. Renand, F. Turin, S. Jabet, K. Osoro, M. C. Oliván, G. Noval, P. García, M. D. García-Cachán, M. A. Oliver, M. Gispert, X. Serra, M. Espejo, S. García, M. López, and M. Izquierdo., 2003. Carcass quality of 10 beef cattle breeds of the Southwest of Europe in their typical production systems. *Livest. Prod. Sci.* 82:1–13.

Pryce, J E., Gredler, B., Bolormaa, S., Bowman, P J., Egger-Danner, C., Fuerst, C., Emmerling, R., Sölkner, J., Goddard, M E, Hayes, B J, 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. *Journal of Dairy Science*, *94*(5), 2625–2630.

Pryce, J., & Hayes. B., 2012. A Review of How Dairy Farmers Can Use and Profit from Genomic Technologies. *Animal Production Science*. 52(2-3):180-184.

Pszczola, M., Strabel, T., Mulder, H. a, & Calus, M. P. L., 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science*, *95*(1), 389–400.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. y Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*, 559–575.

Quaas, C. R. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* 32: 949-953.

Revilla, R., 1997. Sistemas de explotación del ganado reproductor en zonas de montaña. En C. Buxadé (Ed.), *Vacuno de carne: aspectos clave* (pp. 229–249). Madrid: Mundi-Prensa.

Rothschild, M., Jacobson, C, Vaske, D, Tuggle, C., Wang, L., Short, T, Eckardt, G., Sasaki, S, Vincent, A., McLaren, D, Southwood, O, van der Steen, H., Mileham, A, Plastow, G., 1996. The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proc. Natl. Acad. USA*. 93:201-205.

Saatchi, M., McClure, M.C., McKay, S.D., Rolf, M.M., Kim, J., Decker, J.E., Taxis, T.M., Chapple, R.H., Ramey, H.R., Northcutt, S.L., Bauck, S., Woodward, B., Dekkers, J.C.M., Fernando, R.L., Schnabel, R.D., Garrick, D.J., Taylor, J.F., 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics, Selection, Evolution : GSE*, *43*(1), 40.

Saitou, N, & Nei, M., 1987. The Neighbour-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol Biol Evo* 4 (4): 406–25.

Samore, A. B., Buttazzoni, L., Gallo, M., Russo, V., & Fontanesi, L., 2015. Genomic selection in a pig population including information from slaughtered full sibs of boars within a sib-testing program. *Animal*, *9*(5), 750–759.

Sánchez-Belda, A., 198). *Razas bovinas españolas*. Madrid: Publicaciones de Extensión Agraria.

Schenkel, F S, M Sargolzaei, G Kistemaker, G.B. Jansen, P Sullivan, B.J. Van Doormaal, P. M. VanRadeN, and G.R. Wiggans., 2009. “Reliability of Genomic Evaluation of Holstein Cattle in Canada.” *Interbull Bulletin* 39 (39): 51–58.

Serradilla, J.M., 2008. Objetivos, organización y factores limitantes de los programas de selección de las razas autóctonas en España. En *XIV Reunión*

nacional de mejora genética animal, *Sevilla*. Recuperado de http://acteon.webs.upv.es/CONGRESOS/XIV%20Reunion%20MG%20SEVILLA%202008/Docs%20XIV/Serradilla_XIV.doc.

Shumbusho, F., Raoul, J., Astruc, J. M., Palhiere, I., Lemarié, S., Fugeray-Scarbel, A., & Elsen, J. M., 2015. Economic evaluation of genomic selection in small ruminants: a sheep meat breeding program. *Animal*, 1–9.

Solberg, T R, Sonesson, A K, Woolliams, J A, & Meuwissen, T H E, 2008. Genomic Selection Using Different Marker Types and Densities. *Journal of Animal Science* 86 (10): 2447–54.

Sonesson, A.K., & Meuwissen, T.H.E., 2009. Testing Strategies for Genomic Selection in Aquaculture Breeding Programs. *Genetics, Selection, Evolution : GSE* 41 (41): 37.

Steemers FJ, Chang W, Lee G, Barker DL,Shen R, et al., 2006. Whole-genome genotyping with the single-base extension assay. *Nat Methods* 3: 31–33.

Su, G., Christensen, O. F., Ostersen, T., Henryon, M., & Lund, M. S., 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. *PLoS ONE*, *7*(9), e45293.

Su, H., Koltes, J.E.; Saatchi, M., Lee, J., Fernando, R.L.; & Garrick, D.J., 2014 Characterizing Haplotype Diversity in Ten US Beef Cattle Breeds. *Animal Industry Report: AS 660, ASL R2846*.

Sun, Y. V., 2010. Multigenic Modeling of Complex Disease by Random Forests. *Advances in Genetics*, 72:73-99

Teo, Y.Y., Fry, A.E., Bhattacharya, K., Small, K.S., Kwiatkowski, D.P., Clark, T.G., 2009. Genome-Wide Comparisons of Variation in Linkage Disequilibrium. *Genome Research* 19 (10): 1849–60.

Tibshirani, R., 1994. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B*, *58*(1), 267–288.

Toosi, A., Fernando, R. L. & Dekkers, J. C. M., 2010. Genomic Selection in Admixed and Crossbred Populations. *Journal of Animal Science* 88 (1): 32–46.

Usai, M. G., Goddard, M. E., & Hayes, B. J., 2009. LASSO with cross-validation for genomic selection. *Genetics Research*, *91*(6), 427–436.

Van Eenennaam, A. L., Weigel, K. A., Young, A. E., Cleveland, M. A., & Dekkers, J. C. M., 2014. Applied Animal Genomics: Results from the Field. *Annual Review of Animal Biosciences*, *2*(1), 105–139.

Van Laere, A. S., Nguyen, M, Braunschweig, M, Nezer C., Collete, C., Moreau, L., Archibald, A. L., Haley, C. S., Buys, N., Tally, M., Andersson G., Georges, M., Andersson L., 2003. A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* 435: 832-836.

VanRaden, P M, C P Van Tassell, G R Wiggans, T S Sonstegard, R D Schnabel, J F Taylor, and F S Schenkel., 2009. “Invited Review: Reliability of Genomic Predictions for North American Holstein Bulls.” *Journal of Dairy Science* 92 (1):16–24.

VanRaden, P. M., 200). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, *91*(11), 4414–4423.

VanRaden, P.M., Olson, K.M., Null, D.J., Sargolzaei, M., Winters, M., & van Kaam, J.B.C.H.M., 2012. Reliability increases from combining 50,000- and 777,000-marker genotypes from four countries. *Interbull Bull*. 46:75–79.

Vigne, J.D. y Helmer, D., 2007. Was milk a “secondary product” in the Old World Neolithisation process? Its role in the domestication of cattle, sheep and goats. *Anthropozoologica*, *42*, 9–40.

Vigne, J.D., 2008. Zooarchaeological aspects of the Neolithic diet transition in the Near East and Europe, and their putative relationships with the Neolithic demographic transition. En J.P. Bocquet-Appel y O. Bar-Yosef (Eds.), *The*

Neolithic demographic transition and its consequences (pp. 179–205). New York: Springer.

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 (10): 19.

Villanueva, B., Fernandez, J., Garcia-Cortes, L. A., Varona, L., Daetwyler, H. D., & Toro, M. A., 2011. Accuracy of genome-wide evaluation for disease resistance in aquaculture breeding programs. *Journal of Animal Science*, *89*(11), 3433–3442.

Vitezica, Z. G., Varona, L., & Legarra, A., 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics*, *195*(4), 1223–1230.

Vitezica, Z. G., Varona, L., Elsen, J.-M., Misztal, I., Herring, W., & Legarra, A., 2016. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. *Genetics Selection Evolution*, *48*(1), 6.

Weber, K. L., Thallman, R. M., Keele, J. W., Snelling, W. M., Bennett, G. L., Smith, T. P. L., McDaneld, T. G., Allan, M. F., Van Eenennaam, A. L., Kuehn, L. A., 2012. Accuracy of genomic breeding values in multibreed beef cattle populations derived from deregressed breeding values and phenotypes. *Journal of Animal Science*, *90*(12), 4177–4190.

Weigel, K. A., Hoffman, P. C., Herring, W., & Lawlor, T. J., 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. *Journal of Dairy Science*, *95*(4), 2215–25.

Wientjes, Y.Cj., Calus, M.Pl., Goddard, M.E., & Hayes, B,J., 2015. Impact of QTL Properties on the Accuracy of Multi-Breed Genomic Prediction. *Genetics, Selection, Evolution : GSE* 47 (1). BioMed Central: 42.

Wiggans, G. R., Vanraden, P. M., & Cooper, T. A., 2011. The genomic evaluation system in the United States: past, present, future. *Journal of Dairy Science*, *94*(6), 3202–11.

Wolc, A., Zhao, H. H, Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., Preisinger, R., Stricker, C., Habier, D., Fernando, R. L., Garrick, D. J., Lamont, S. J., Dekkers, J. C. M., 2015. Response and inbreeding from a genomic selection experiment in layer chickens. *Genetics, Selection, Evolution : GSE*, *47*(1), 59.

Yáñez, J. M., Houston, R. D., & Newman, S., 2014. Genetics and genomics of disease resistance in salmonid species. *Frontiers in Genetics*. Frontiers, 5:415

Zhou, L, Lund, M.S., Wang, Y., & Su, G., 2014. Genomic Predictions across Nordic Holstein and Nordic Red Using the Genomic Best Linear Unbiased Prediction Model with Different Genomic Relationship Matrices. *Journal of Animal Breeding and Genetics* 131 (4): 249–57.

Zhou, L., Ding, X., Zhang, Q., Wang, Y., Lund, M. S., & Su, G., 2013. Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. *Genetics, Selection, Evolution : GSE*, *45*(1), 7.

Zhou, L., Heringstad, B., Su, G., Guldbrandtsen, B., Meuwissen, T H E, Svendsen, M., Grove, H., Nielsen, U S, Lund, M S, 2014a. Genomic predictions based on a joint reference population for the Nordic Red cattle breeds. *Journal of Dairy Science*, *97*(7), 4485–96.

Zhou, L., Lund, M. S., Wang, Y., & Su, G., 2014b. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *Journal of Animal Breeding and Genetics*, *131*(4), 249–257.

Zimin, A. V., A. L. Delcher, L. Florea, D. R.Kelley, M. C. Schatz, D. Puiu, R. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard , G Marçais, M.

Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg,. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10:R42.

Zou H, Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society. Series B (Methodological)* 67:301-320.

ANNEXES

ANEXXE 1

Table 1.1. Accuracies (s.e.) obtained in Pi for traits A (h^2=0.4) and B (h^2=0.1) without 2014 data

Pi	***Trait A (h^2=0.4)***				***Trait B (h^2=0.1)***			
	+none	*+sires +dams*	*+2014*	*+sires +dams +2014*	*+none*	*+sires +dams*	*+2014*	*+sires +dams +2014*
TH-250	0.555 (0.011)	0.571 (0.010)	0.598 (0.010)	0.619 (0.009)	0.447 (0.013)	0.467 (0.012)	0.488 (0.011)	0.500 (0.011)
TH-500	0.555 (0.011)	0.575 (0.011)	0.615 (0.010)	0.632 (0.009)	0.449 (0.013)	0.472 (0.013)	0.503 (0.011)	0.511 (0.011)
TH-1000	0.556 (0.011)	0.581 (0.011)	0.641 (0.009)	0.651 (0.008)	0.451 (0.013)	0.479 (0.012)	0.521 (0.010)	0.526 (0.010)
TH-2000	0.558 (0.011)	0.589 (0.011)	0.673 (0.009)	0.680 (0.008)	0.454 (0.013)	0.484 (0.013)	0.536 (0.011)	0.540 (0.011)
TH-4000	0.563 (0.011)	0.599 (0.011)	0.704 (0.008)	0.709 (0.007)	0.458 (0.013)	0.490 (0.013)	0.554 (0.010)	0.558 (0.010)

Table 1.2. Accuracies (s.e.) obtained in RG for traits A (h^2=0.4) and B (h^2=0.1) without 2014 data

RG	***Trait A (h^2=0.4)***				***Trait B (h^2=0.1)***			
	+none	*+sires +dams*	*+2014*	*+sires +dams +2014*	*+none*	*+sires +dams*	*+2014*	*+sires +dams +2014*
TH-250	0.550 (0.010)	0.578 (0.009)	0.615 (0.009)	0.651 (0.007)	0.480 (0.012)	0.506 (0.011)	0.535 (0.011)	0.552 (0.010)
TH-500	0.551 (0.010)	0.583 (0.009)	0.637 (0.009)	0.665 (0.007)	0.482 (0.012)	0.510 (0.011)	0.549 (0.011)	0.562 (0.010)
TH-1000	0.552 (0.010)	0.589 (0.009)	0.661 (0.007)	0.684 (0.006)	0.484 (0.012)	0.514 (0.011)	0.561 (0.010)	0.572 (0.009)
TH-2000	0.555 (0.010)	0.599 (0.009)	0.698 (0.007)	0.713 (0.006)	0.487 (0.012)	0.522 (0.010)	0.583 (0.009)	0.591 (0.008)
TH-4000	0.558 (0.010)	0.608 (0.008)	0.739 (0.006)	0.747 (0.005)	0.490 (0.012)	0.528 (0.010)	0.604 (0.009)	0.609 (0.008)

Table 1.3. Accuracies (s.e.) obtained in Pi for traits A (h^2=0.4) and B (h^2=0.1) with 2014 data

Pi	***Trait A (h^2=0.4)***				***Trait B (h^2=0.1)***			
	+none	*+sires +dams*	*+2014*	*+sires +dams +2014*	*+none*	*+sires +dams*	*+2014*	*+sires +dams +2014*
TH-250	0.724 (0.005)	0.728 (0.005)	0.748 (0.004)	0.756 (0.004)	0.515 (0.011)	0.526 (0.011)	0.551 (0.010)	0.526 (0.011)
TH-500	0.724 (0.005)	0.730 (0.005)	0.755 (0.004)	0.762 (0.004)	0.516 (0.011)	0.530 (0.011)	0.562 (0.010)	0.568 (0.010)
TH-1000	0.725 (0.005)	0.732 (0.005)	0.765 (0.004)	0.770 (0.004)	0.519 (0.011)	0.535 (0.011)	0.574 (0.009)	0.578 (0.009)
TH-2000	0.727 (0.005)	0.736 (0.005)	0.779 (0.004)	0.781 (0.004)	0.523 (0.011)	0.540 (0.011)	0.587 (0.009)	0.589 (0.009)
TH-4000	0.730 (0.005)	0.740 (0.005)	0.795 (0.004)	0.796 (0.004)	0.529 (0.011)	0.546 (0.010)	0.602 (0.009)	0.603 (0.009)

Table 1.4. Accuracies (s.e.) obtained in RG for traits A (h^2=0.4) and B (h^2=0.1) with 2014 data

RG	***Trait A (h^2=0.4)***				***Trait B (h^2=0.1)***			
	+none	*+sires +dams*	*+2014*	*+sires +dams +2014*	*+none*	*+sires +dams*	*+2014*	*+sires +dams +2014*
TH-250	0.727 (0.004)	0.738 (0.004)	0.782 (0.003)	0.795 (0.003)	0.549 (0.007)	0.568 (0.007)	0.614 (0.006)	0.626 (0.006)
TH-500	0.728 (0.004)	0.739 (0.004)	0.789 (0.003)	0.799 (0.003)	0.550 (0.007)	0.570 (0.007)	0.623 (0.006)	0.632 (0.006)
TH-1000	0.728 (0.004)	0.741 (0.004)	0.797 (0.003)	0.805 (0.003)	0.551 (0.007)	0.572 (0.007)	0.631 (0.006)	0.639 (0.005)
TH-2000	0.728 (0.004)	0.744 (0.004)	0.809 (0.003)	0.815 (0.003)	0.553 (0.007)	0.576 (0.006)	0.643 (0.005)	0.649 (0.005)
TH-4000	0.730 (0.004)	0.748 (0.004)	0.826 (0.002)	0.830 (0.002)	0.557 (0.007)	0.582 (0.006)	0.661 (0.005)	0.665 (0.005)

ANNEXE 2

Figure 2.1. Manhattan plot of the CorLD estimates along the genome for the AV-ANI for regions of 50 SNPs.

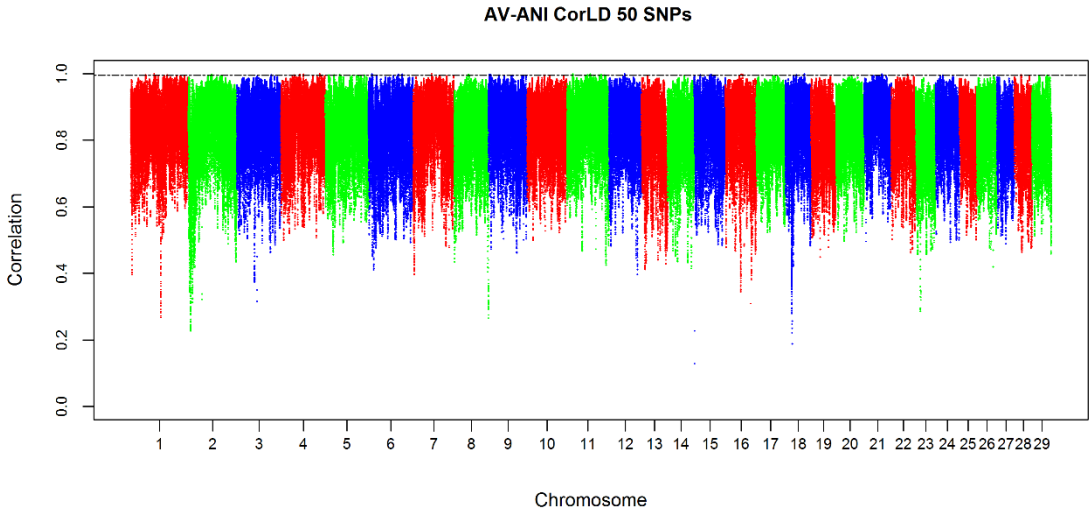


Figure 2.2. Manhattan plot of the VarLD estimates along the genome for the AV-ANI for regions of 50 SNPs.

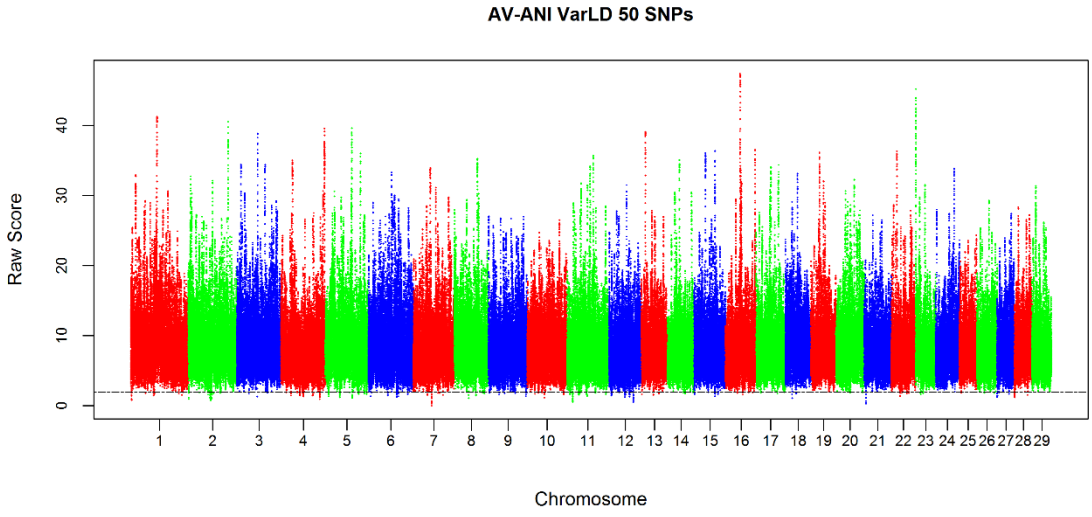


Figure 2.3. Manhattan plot of the CorLD estimates along the genome for the AV-ANI for regions of 100 SNPs.

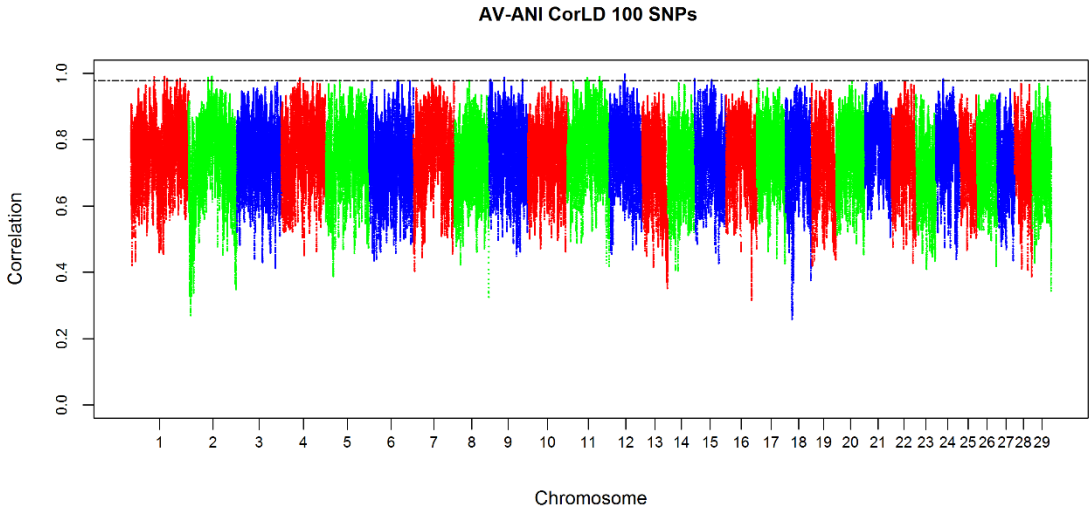


Figure 2.4. Manhattan plot of the VarLD estimates along the genome for the AV-ANI for regions of 100 SNPs.

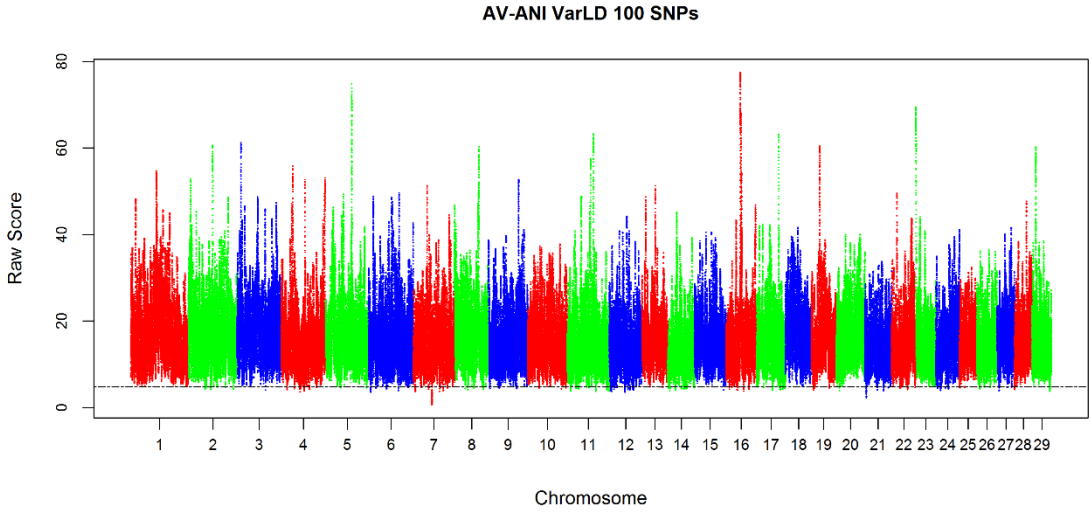


Figure 2.5. Manhattan plot of the CorLD estimates along the genome for the AV-ANI for regions of 200 SNPs.

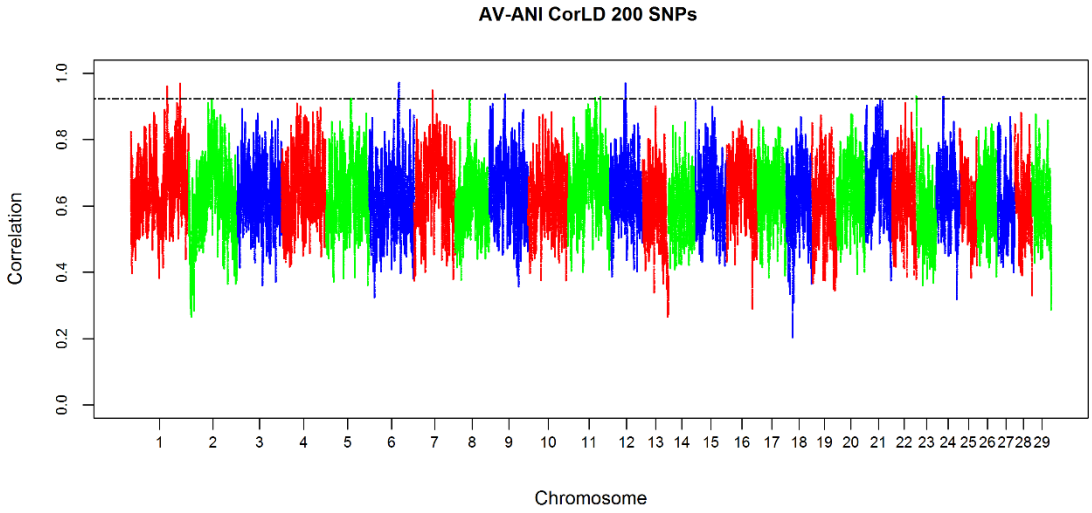
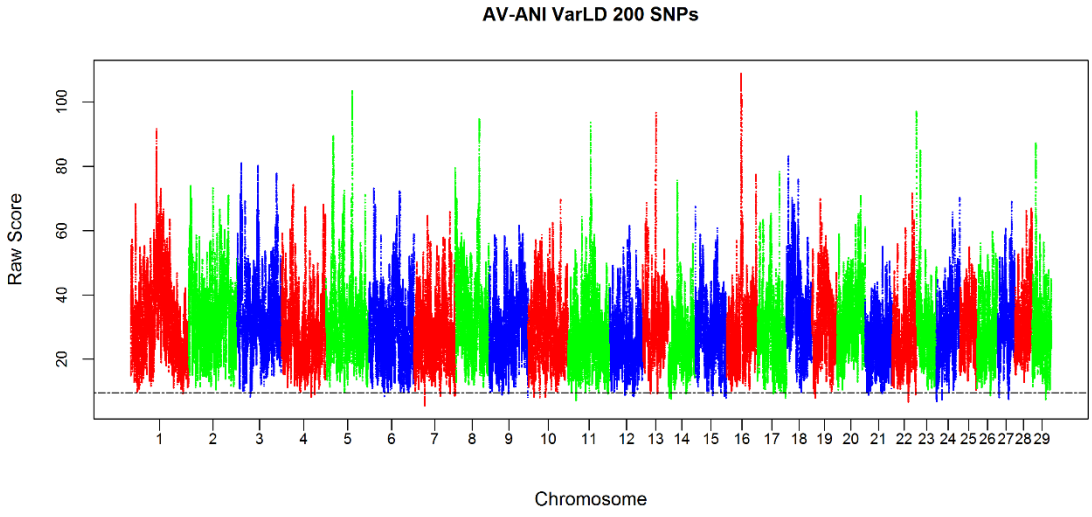


Figure 2.6. Manhattan plot of the VarLD estimates along the genome for the AV-ANI for regions of 200 SNPs.



Annexe 3

Figure 3.1. Manhattan plot of the CorLD estimates along the genome for the AV-ANI for regions of 100 SNPs.

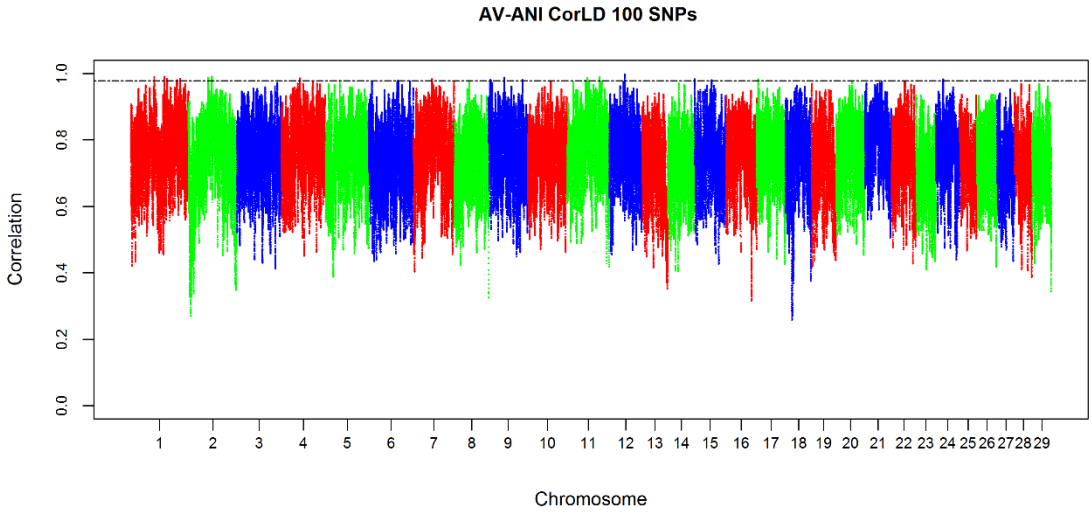


Figure 3.2. Manhattan plot of the VarLD estimates along the genome for the AV-ANI for regions of 100 SNPs.

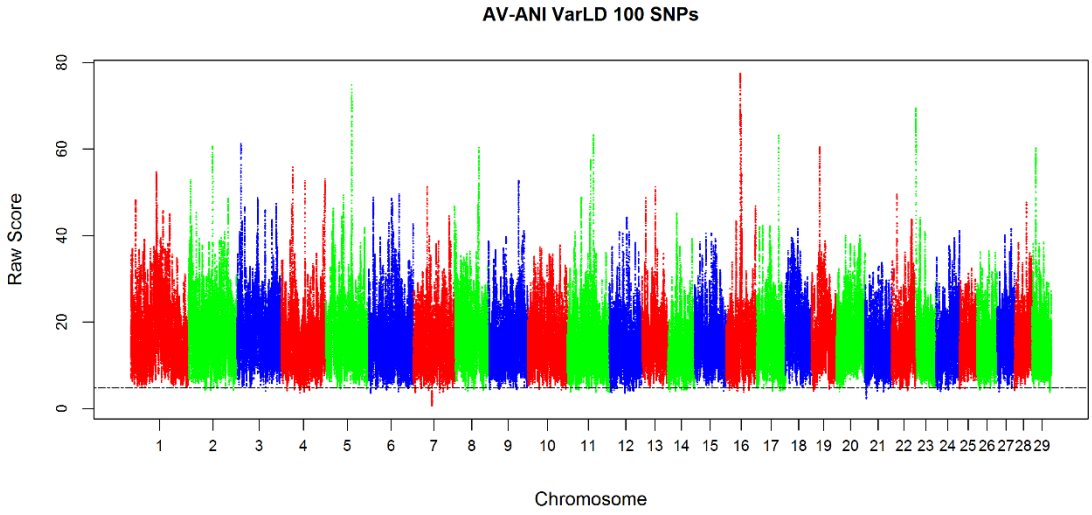


Figure 3.3. Manhattan plot of the CorLD estimates along the genome for the AV-Mo for regions of 100 SNPs.

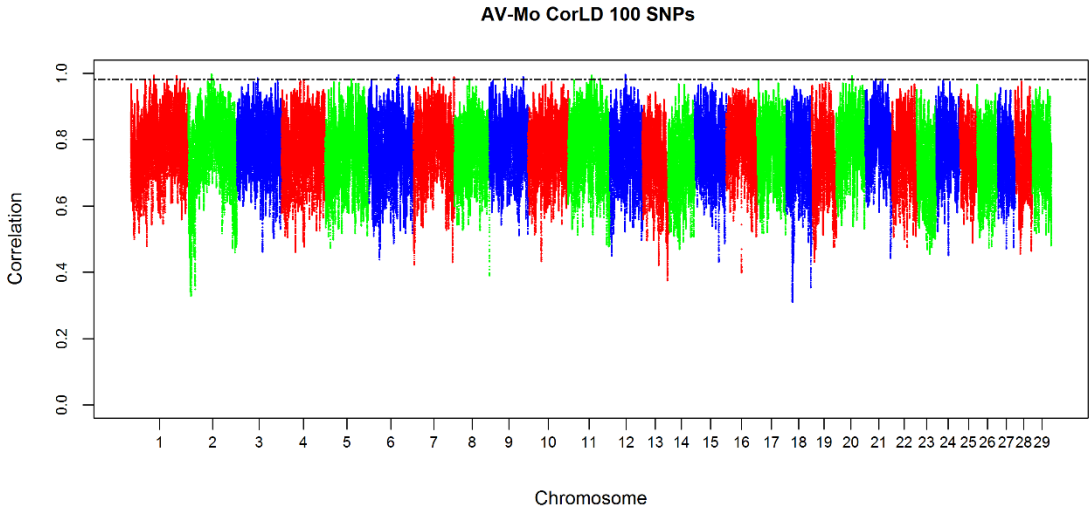


Figure 3.4. Manhattan plot of the VarLD estimates along the genome for the AV-Mo for regions of 100 SNPs.

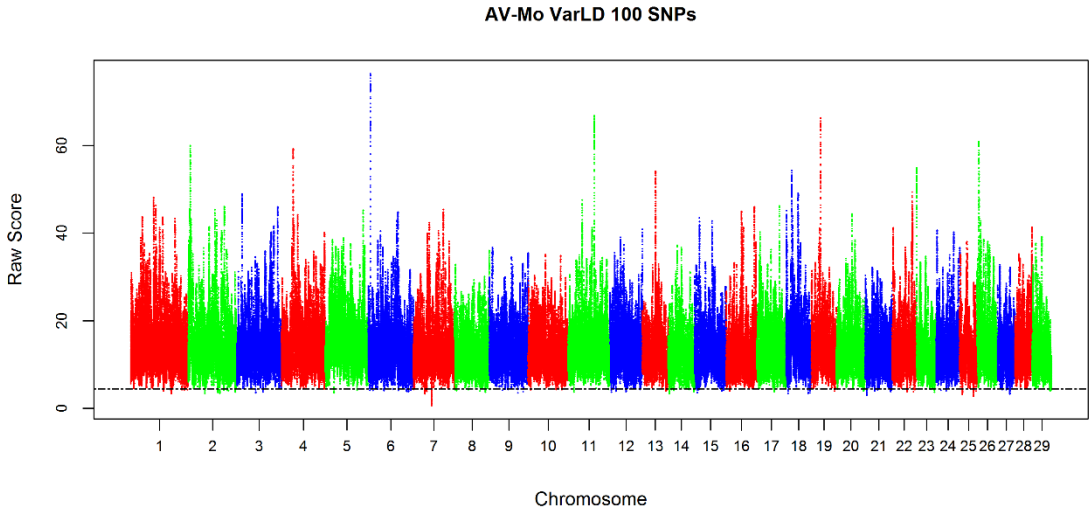


Figure 3.5. Manhattan plot of the CorLD estimates along the genome for the AV-Pi for regions of 100 SNPs.

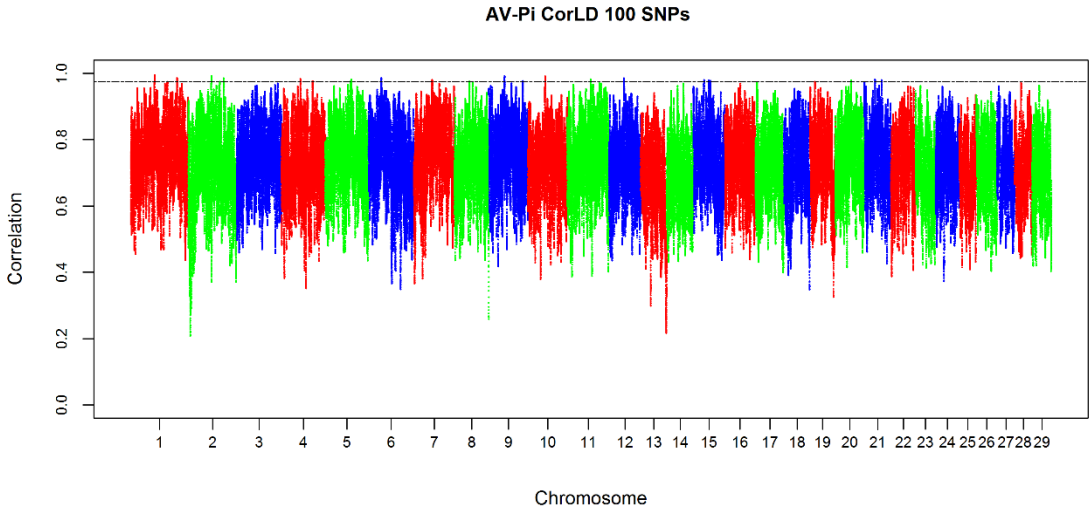


Figure 3.6. Manhattan plot of the VarLD estimates along the genome for the AV-Pi for regions of 100 SNPs.

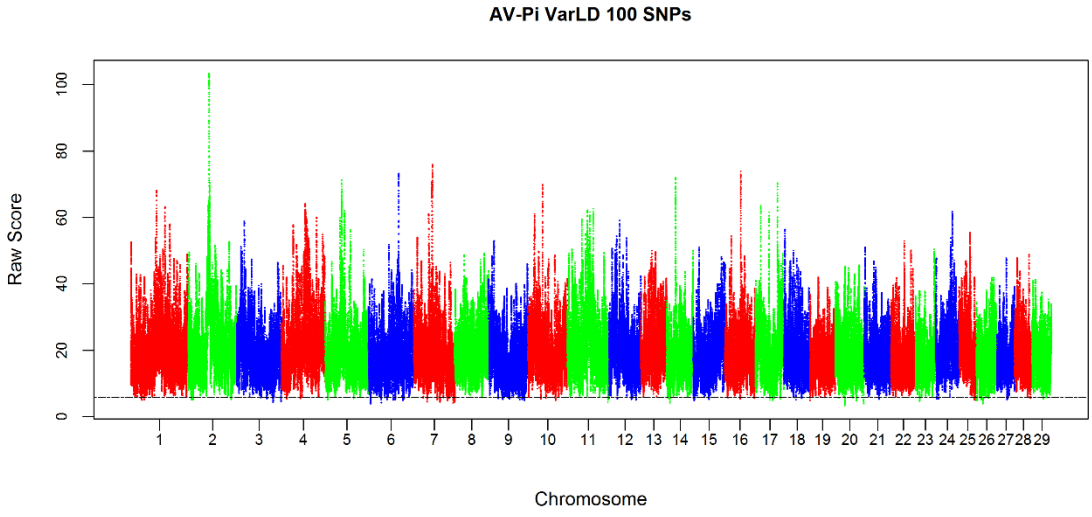


Figure 3.7. Manhattan plot of the CorLD estimates along the genome for the AV-Re for regions of 100 SNPs.

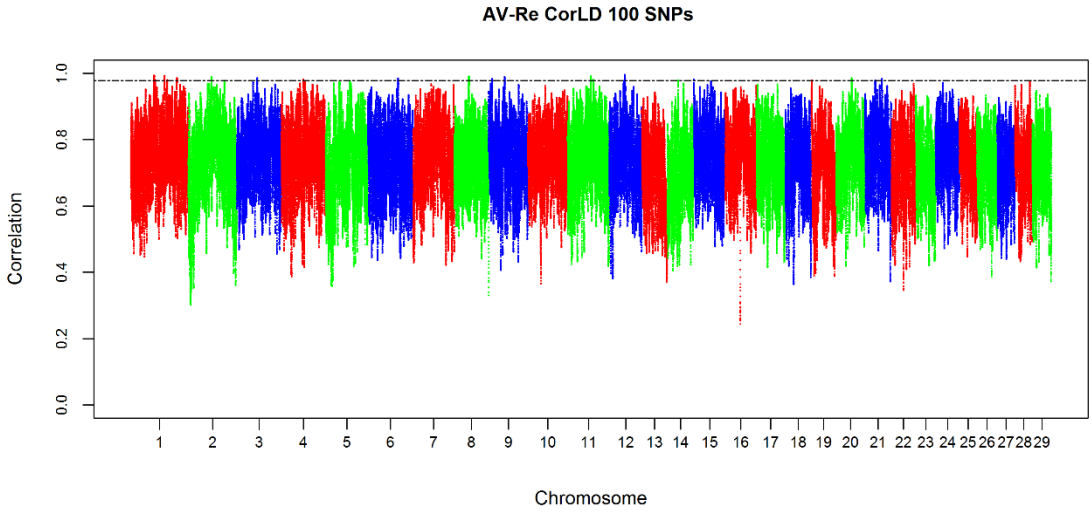


Figure 3.8. Manhattan plot of the VarLD estimates along the genome for the AV-Re for regions of 100 SNPs.

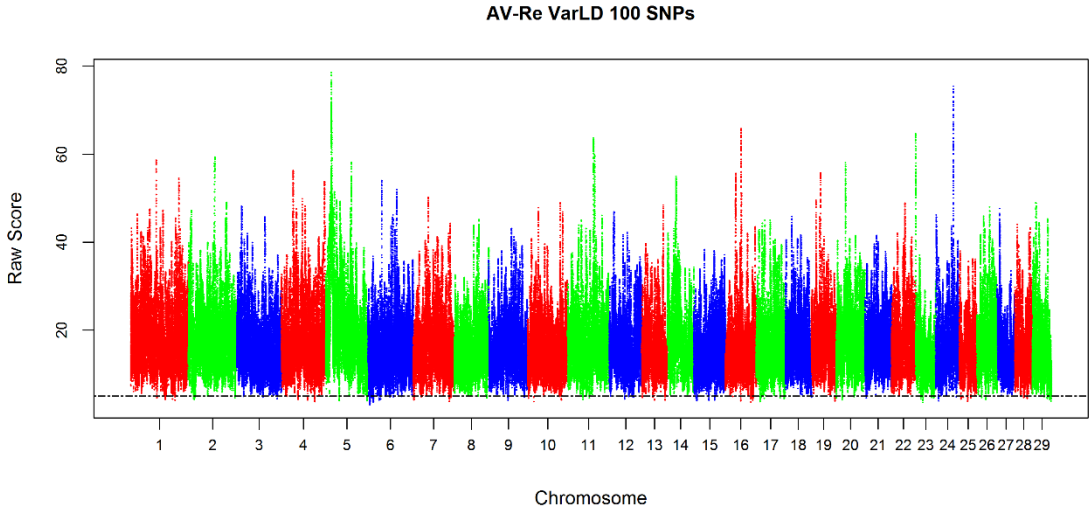


Figure 3.9. Manhattan plot of the CorLD estimates along the genome for the AV-RG for regions of 100 SNPs.

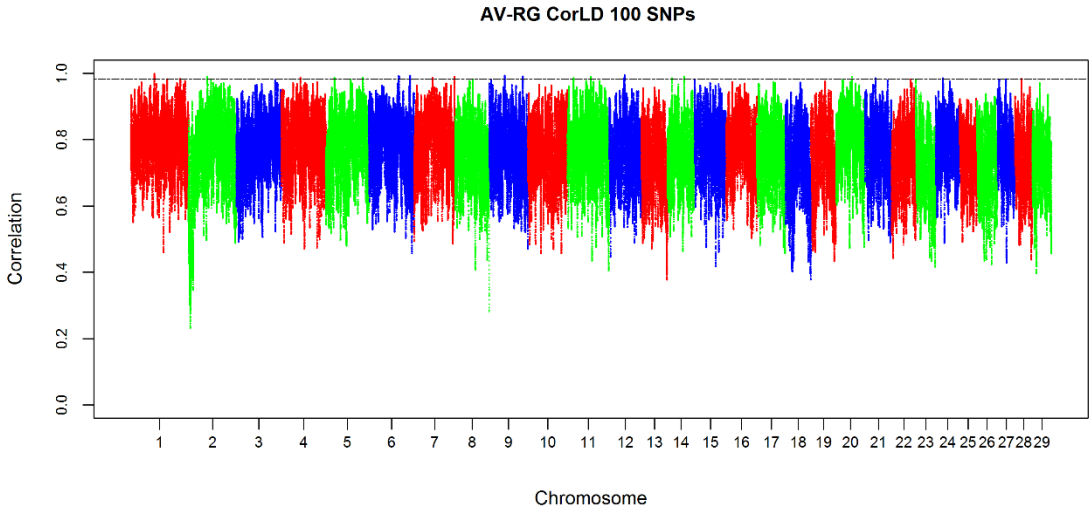


Figure 3.10. Manhattan plot of the VarLD estimates along the genome for the AV-RG for regions of 100 SNPs.

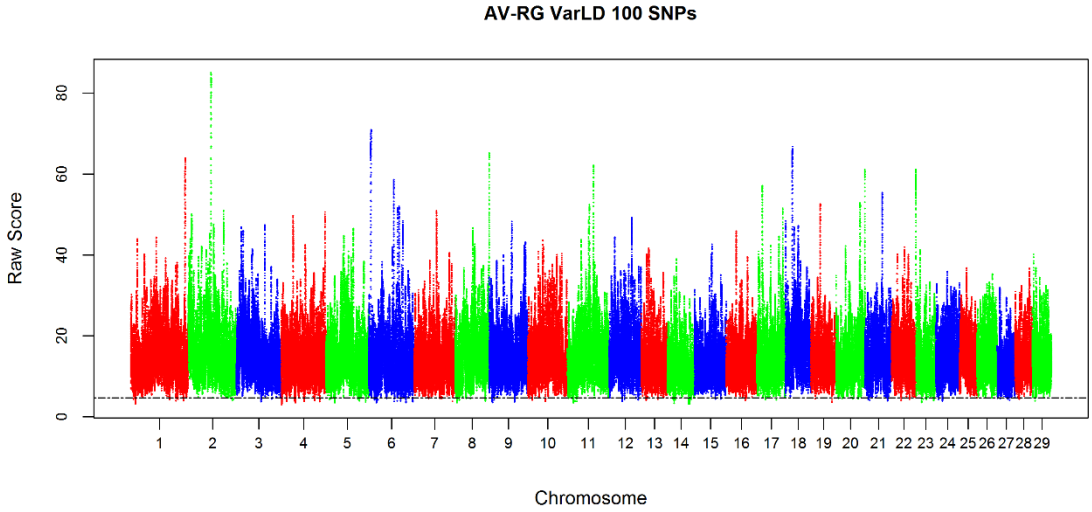


Figure 3.11. Manhattan plot of the CorLD estimates along the genome for the ANI-BP for regions of 100 SNPs.

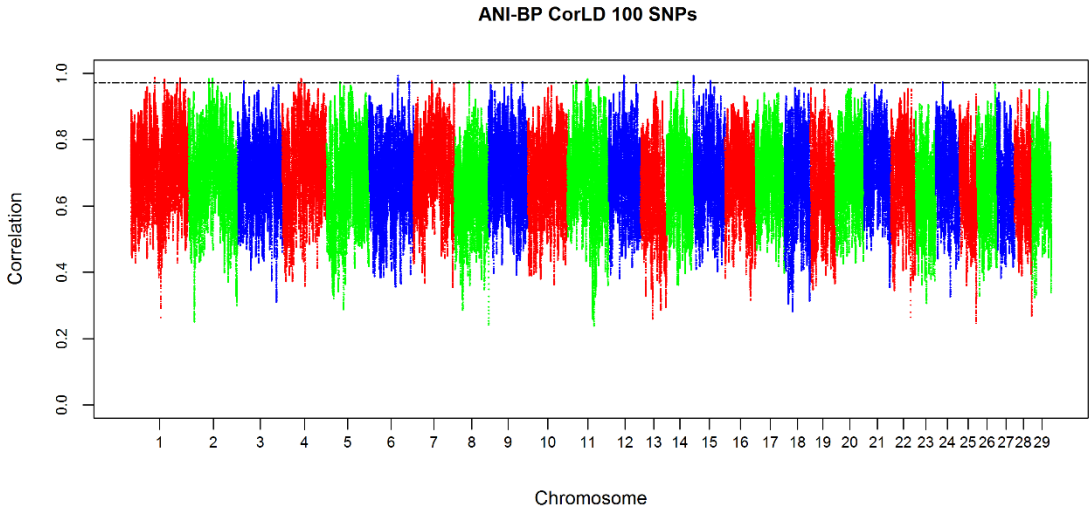


Figure 3.12. Manhattan plot of the VarLD estimates along the genome for the ANI-BP for regions of 100 SNPs.

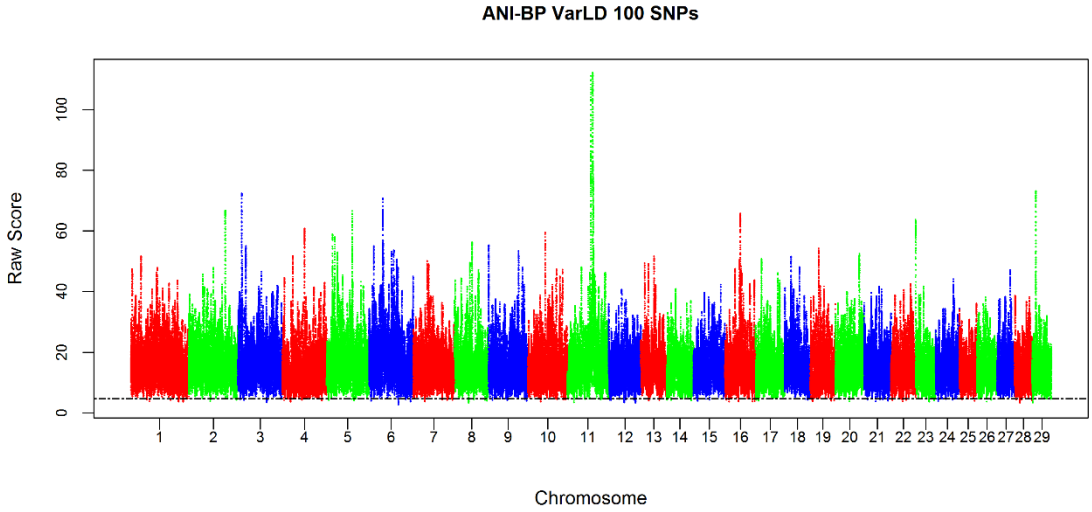


Figure 3.13. Manhattan plot of the CorLD estimates along the genome for the ANI-Mo for regions of 100 SNPs.

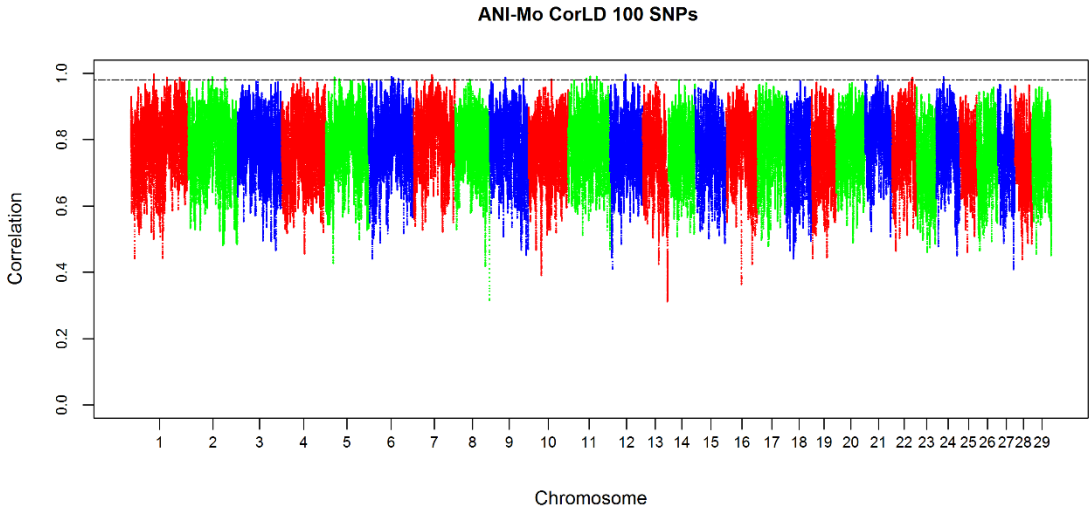


Figure 3.14. Manhattan plot of the VarLD estimates along the genome for the ANI-Mo for regions of 100 SNPs.

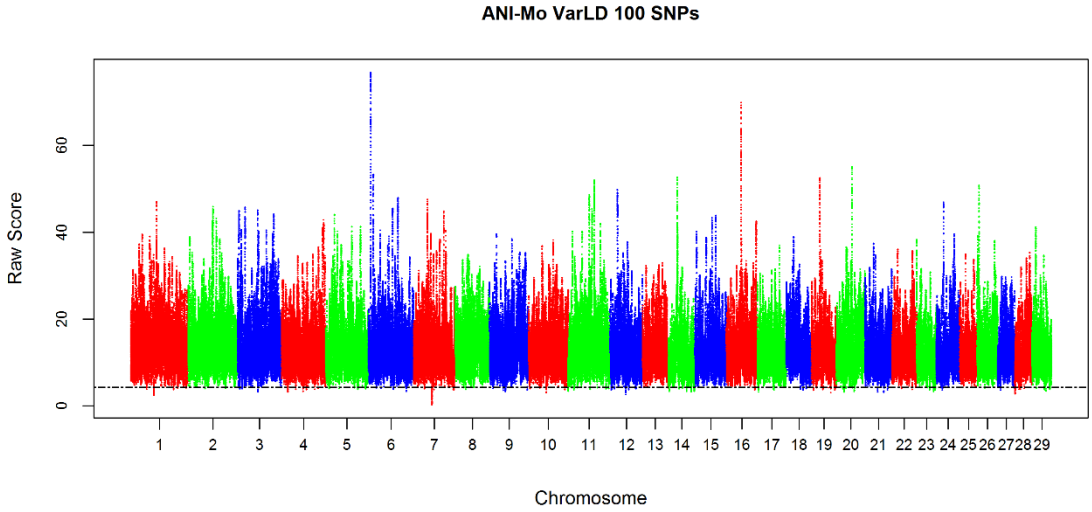


Figure 3.15. Manhattan plot of the CorLD estimates along the genome for the ANI-Pi for regions of 100 SNPs.

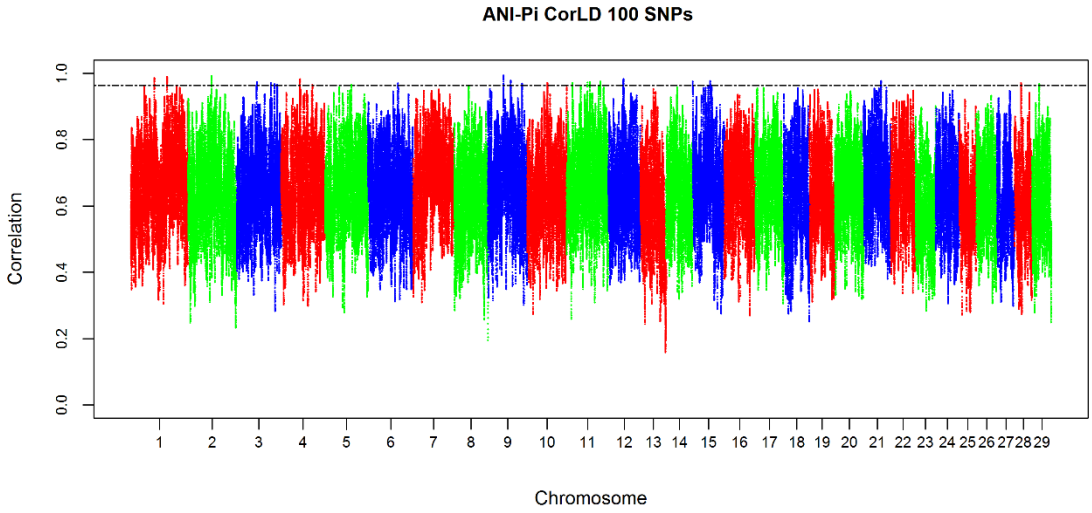


Figure 3.16. Manhattan plot of the VarLD estimates along the genome for the ANI-Pi for regions of 100 SNPs.

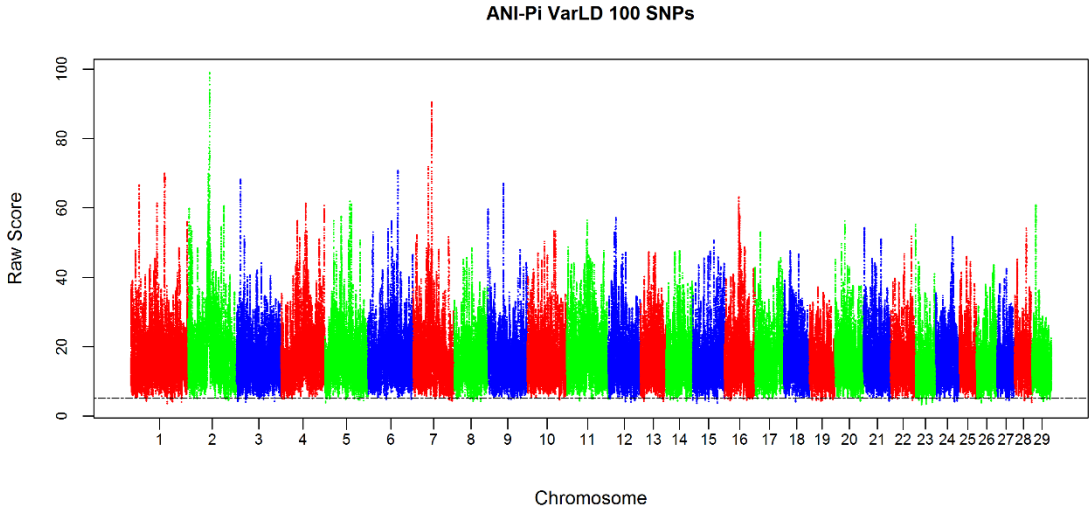


Figure 3.17. Manhattan plot of the CorLD estimates along the genome for the ANI-Re for regions of 100 SNPs.

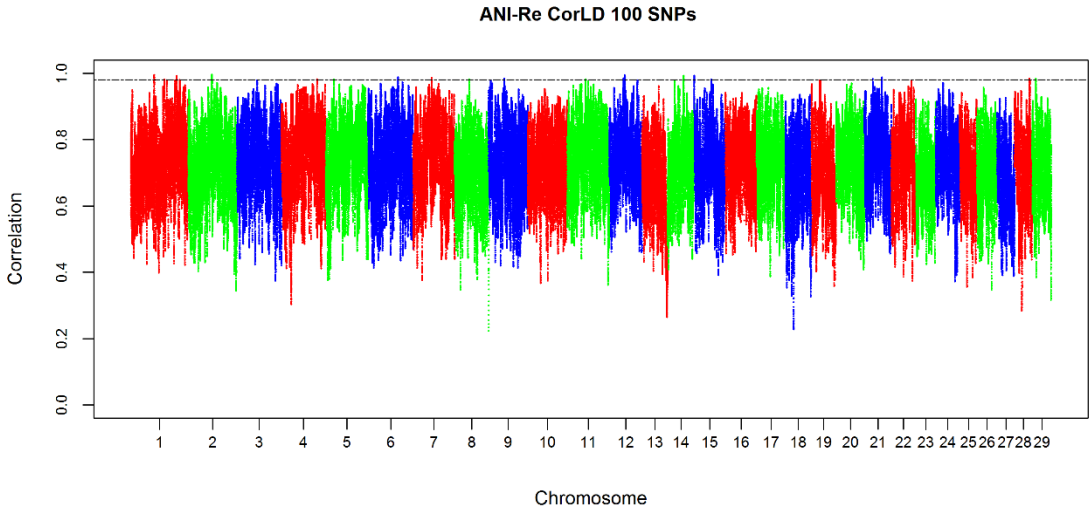


Figure 3.18. Manhattan plot of the VarLD estimates along the genome for the ANI-Re for regions of 100 SNPs.

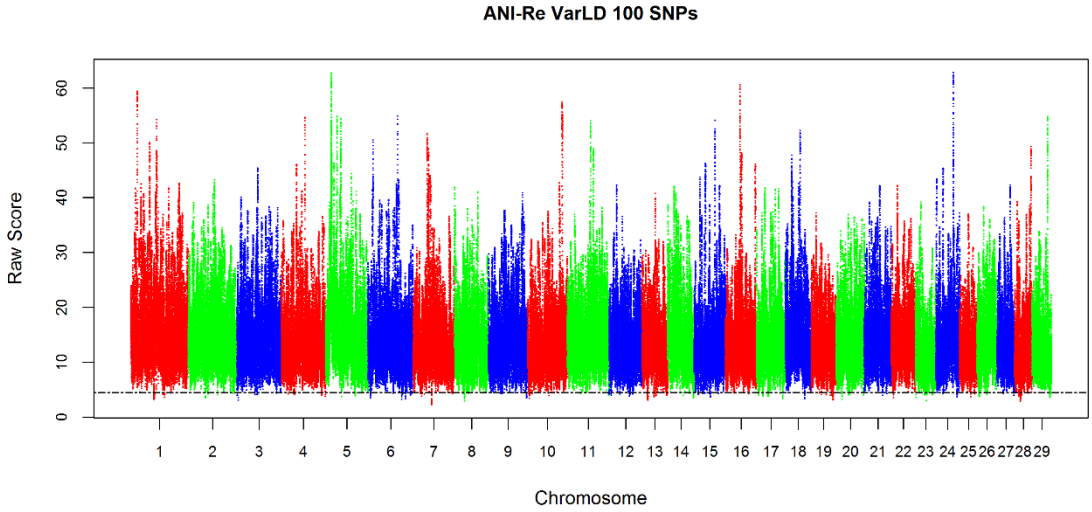


Figure 3.19. Manhattan plot of the CorLD estimates along the genome for the ANI-RG for regions of 100 SNPs.

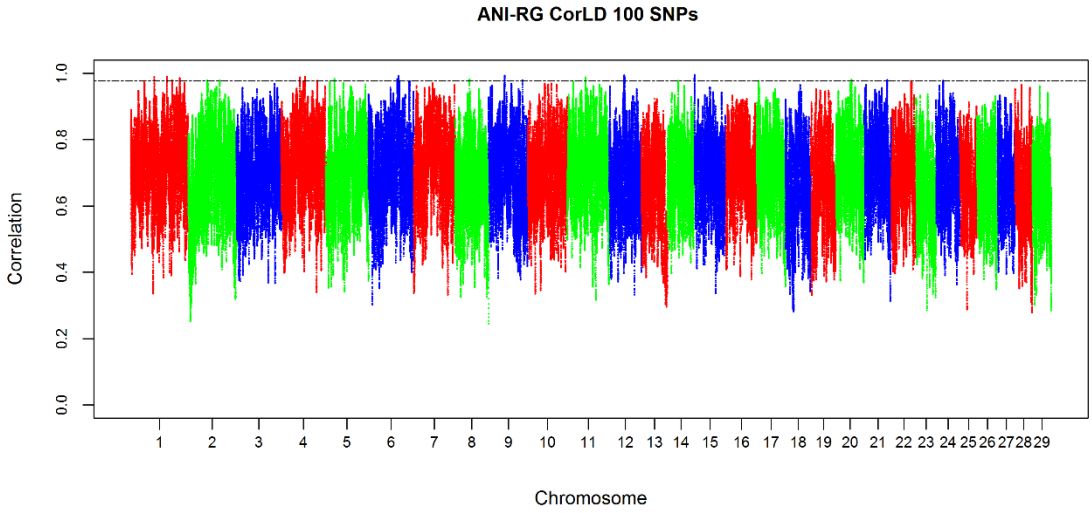


Figure 3.20. Manhattan plot of the VarLD estimates along the genome for the ANI-RG for regions of 100 SNPs.

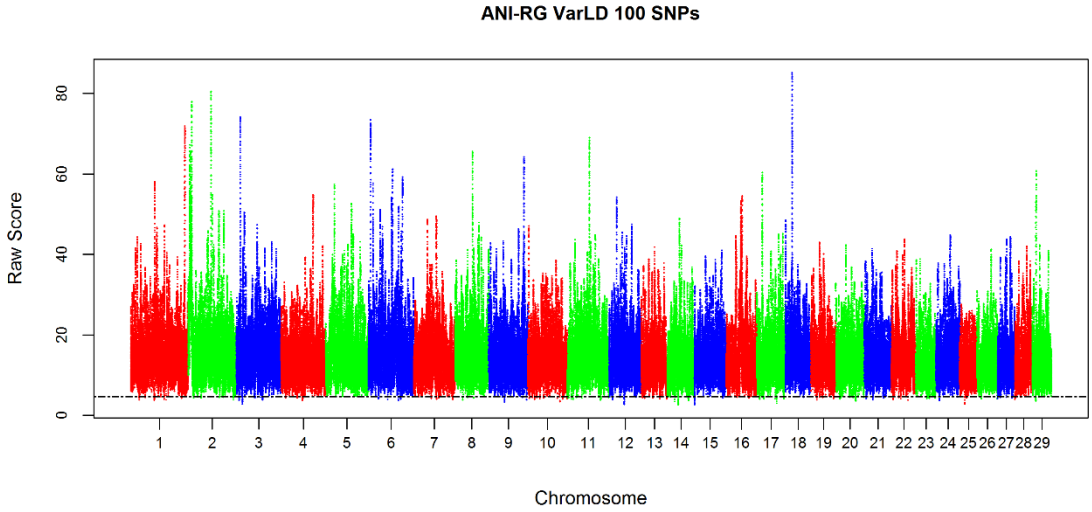


Figure 3.21. Manhattan plot of the CorLD estimates along the genome for the BP-Mo for regions of 100 SNPs.

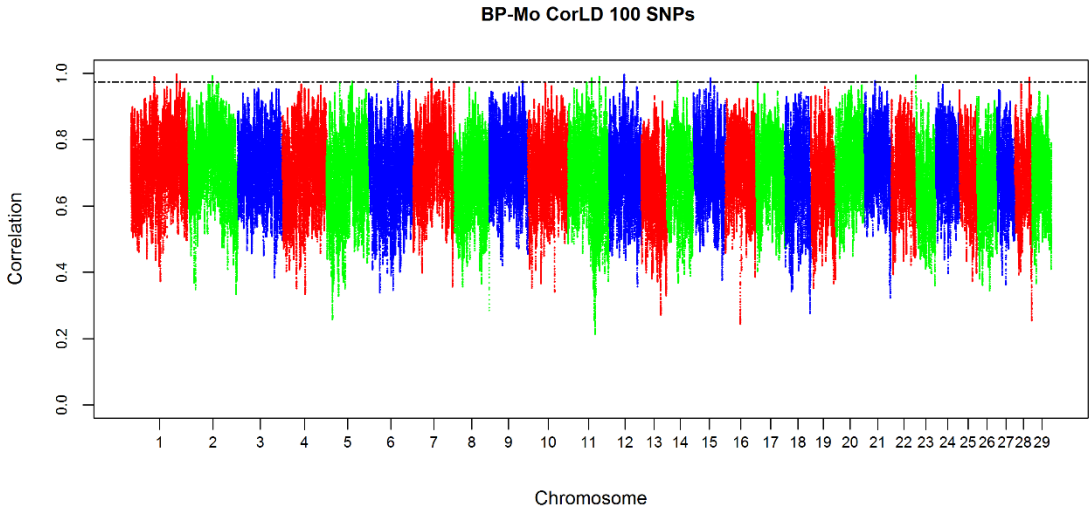


Figure 3.22. Manhattan plot of the VarLD estimates along the genome for the BP-Mo for regions of 100 SNPs.

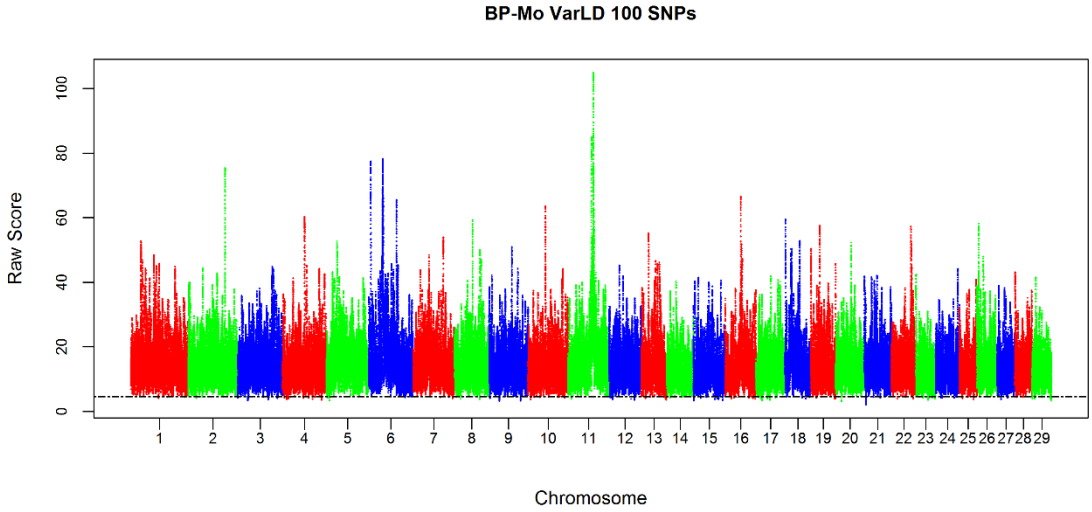


Figure 3.23. Manhattan plot of the CorLD estimates along the genome for the BP-Pi for regions of 100 SNPs.

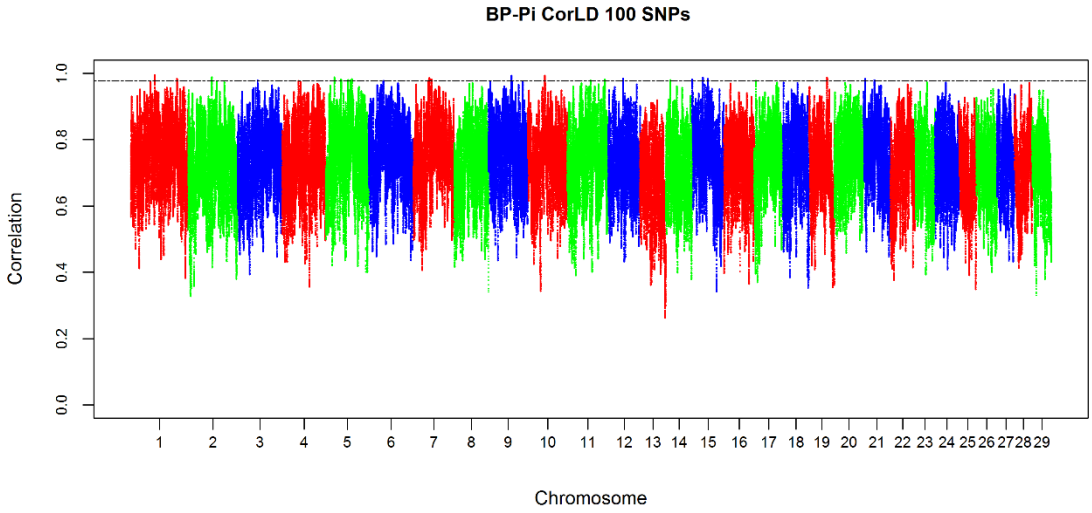


Figure 3.24. Manhattan plot of the VarLD estimates along the genome for the BP-Pi for regions of 100 SNPs.

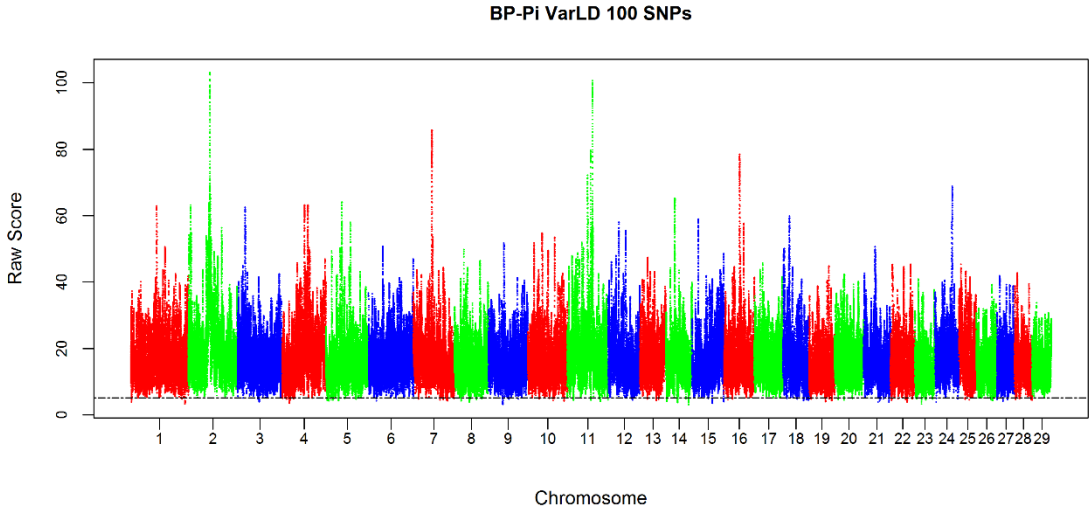


Figure 3.25. Manhattan plot of the CorLD estimates along the genome for the BP-Re for regions of 100 SNPs.

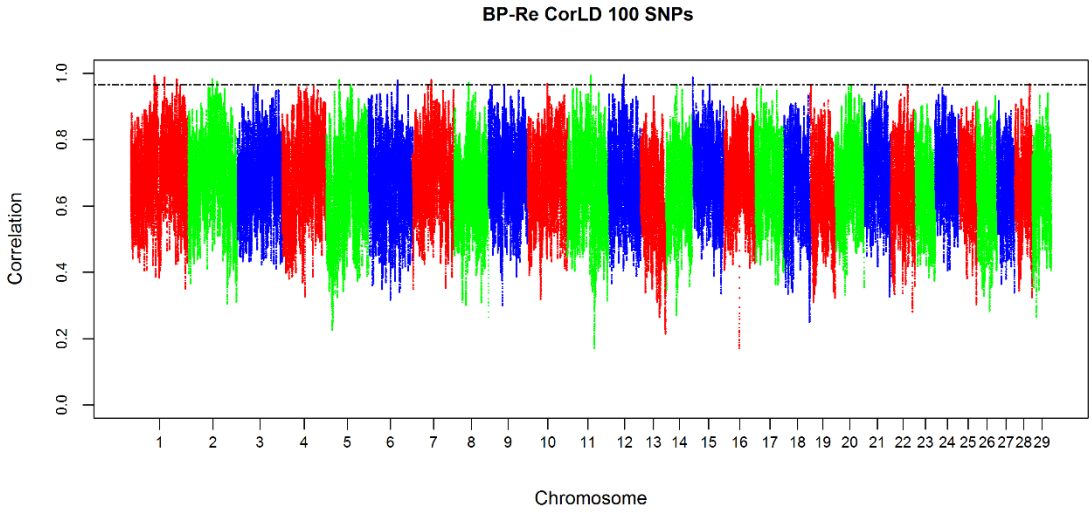


Figure 3.26. Manhattan plot of the VarLD estimates along the genome for the BP-Re for regions of 100 SNPs.

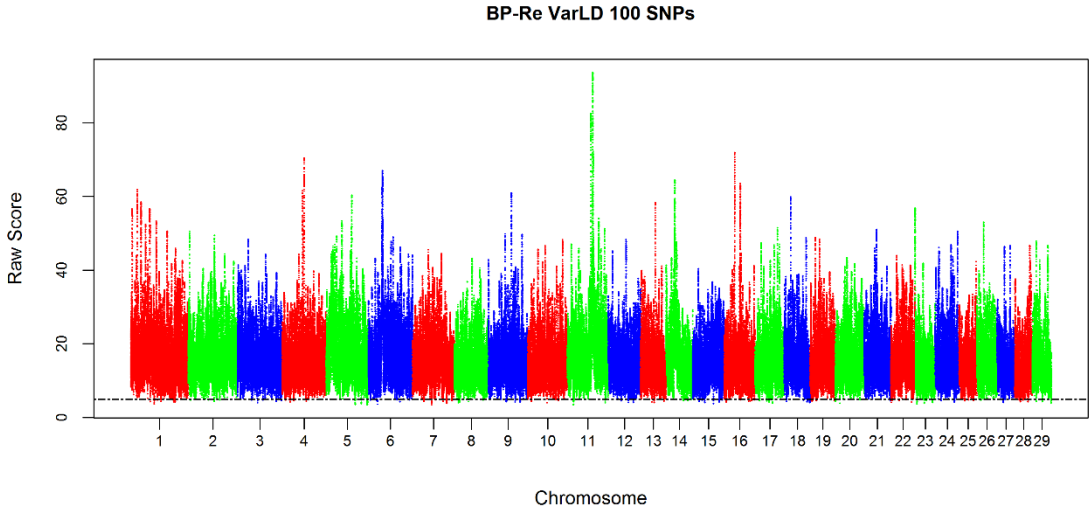


Figure 3.27. Manhattan plot of the CorLD estimates along the genome for the BP-RG for regions of 100 SNPs.

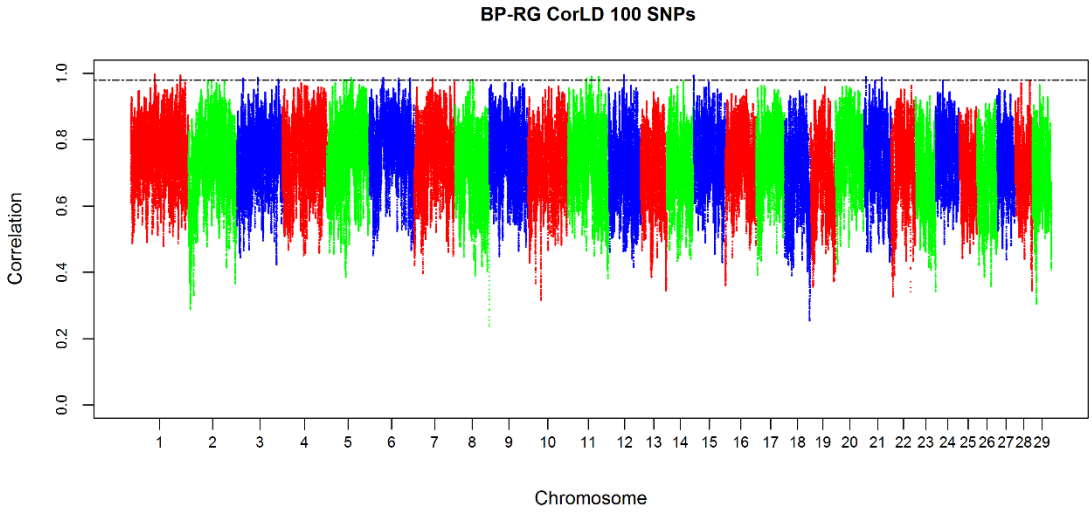


Figure 3.28. Manhattan plot of the VarLD estimates along the genome for the BP-RG for regions of 100 SNPs.

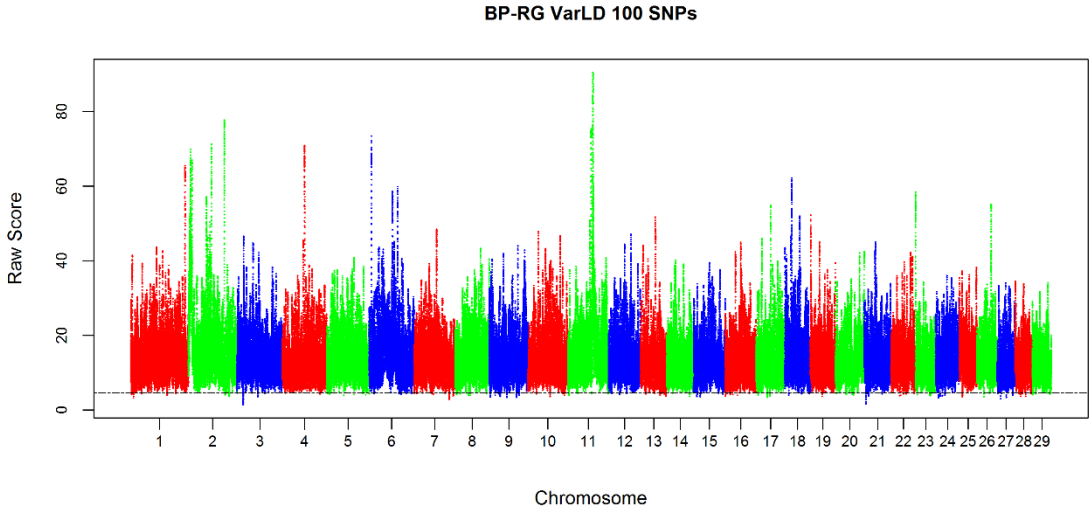


Figure 3.29. Manhattan plot of the CorLD estimates along the genome for the Mo-Pi for regions of 100 SNPs.

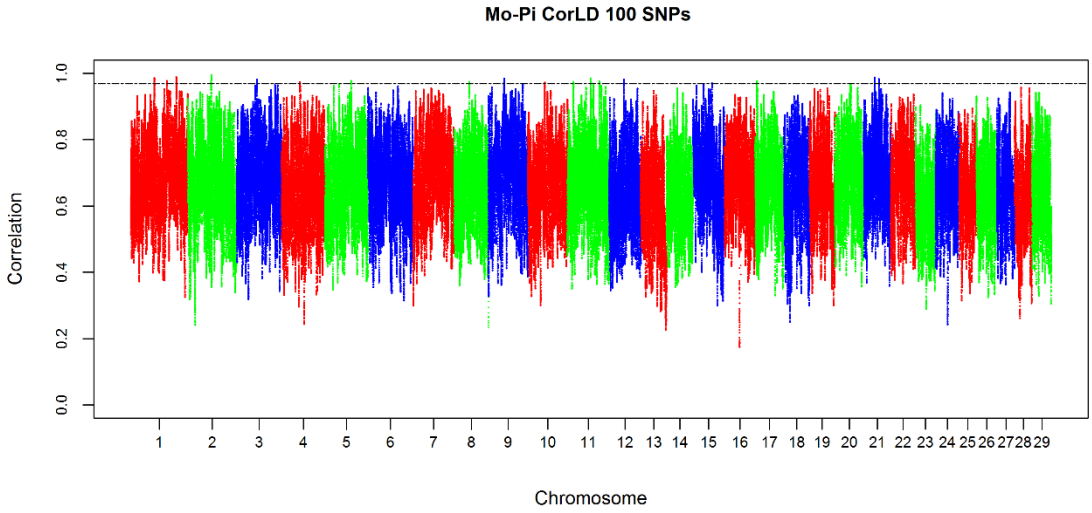


Figure 3.30. Manhattan plot of the VarLD estimates along the genome for the Mo-Pi for regions of 100 SNPs.

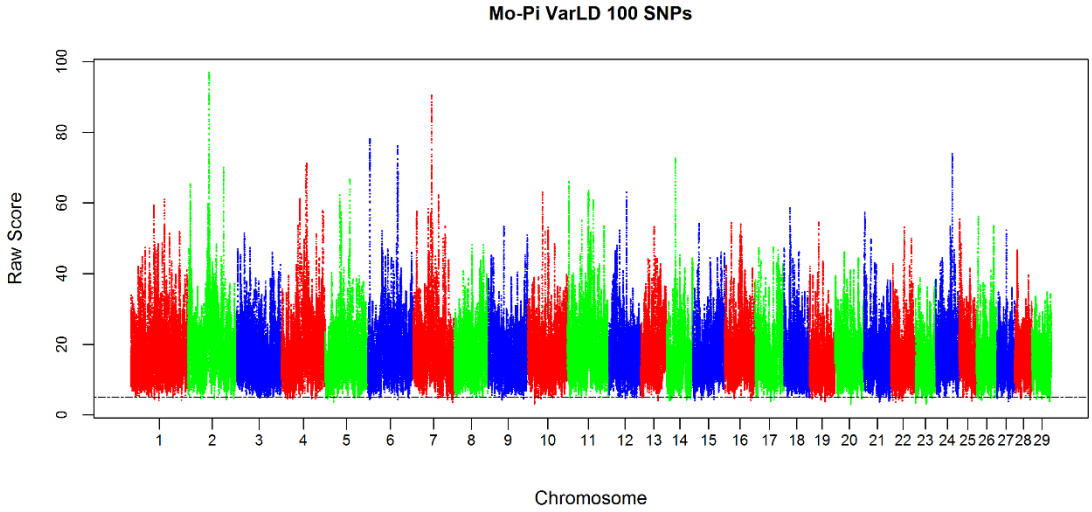


Figure 3.31. Manhattan plot of the CorLD estimates along the genome for the Mo-Re for regions of 100 SNPs.

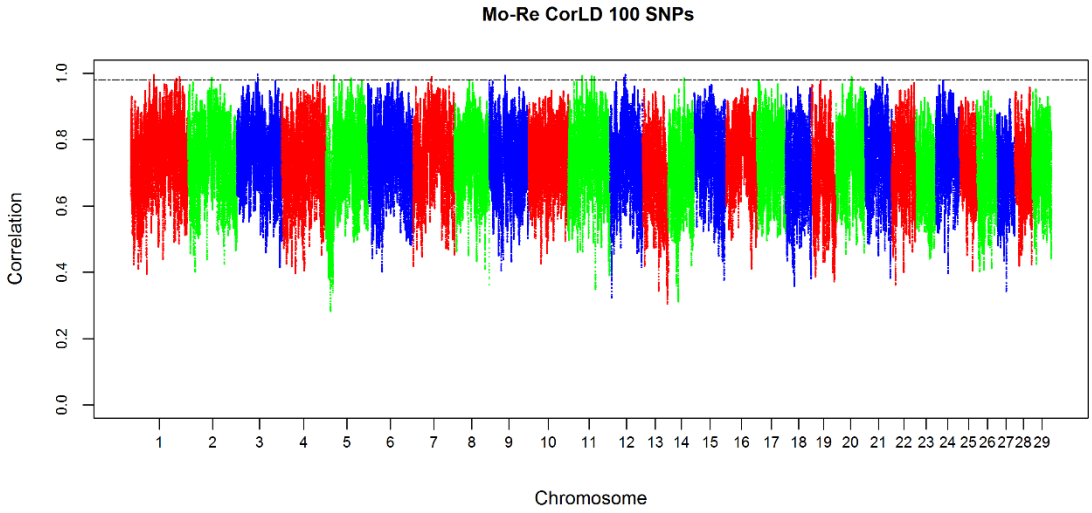


Figure 3.32. Manhattan plot of the VarLD estimates along the genome for the Mo-Re for regions of 100 SNPs.

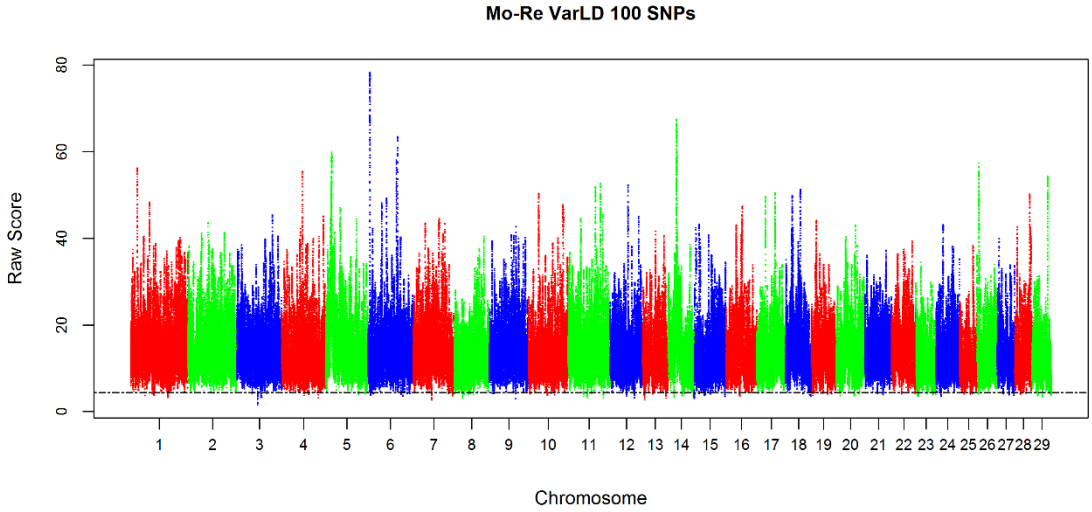


Figure 3.33. Manhattan plot of the CorLD estimates along the genome for the Mo-RG for regions of 100 SNPs.

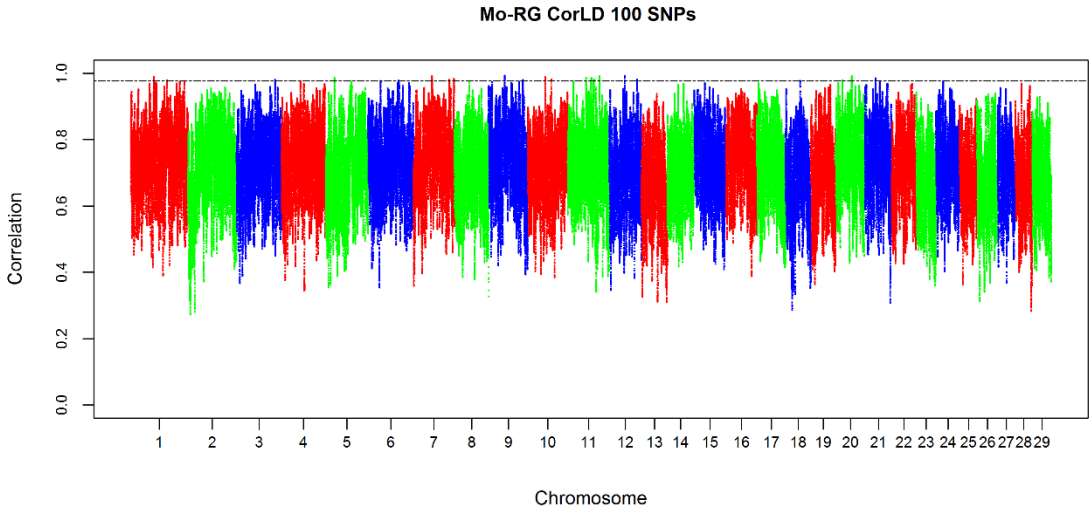


Figure 3.34. Manhattan plot of the VarLD estimates along the genome for the Mo-RG for regions of 100 SNPs.

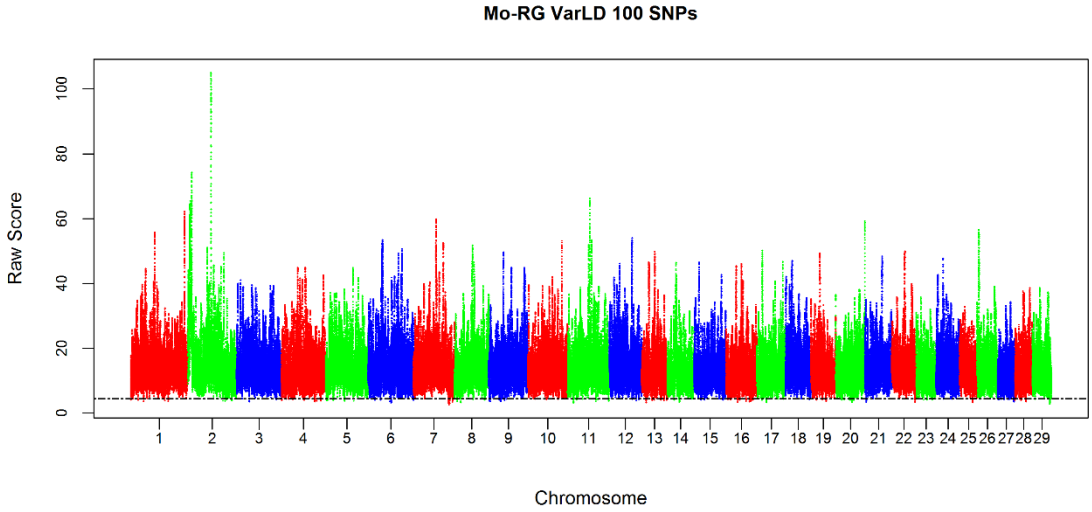


Figure 3.35. Manhattan plot of the CorLD estimates along the genome for the Pi-Re for regions of 100 SNPs.

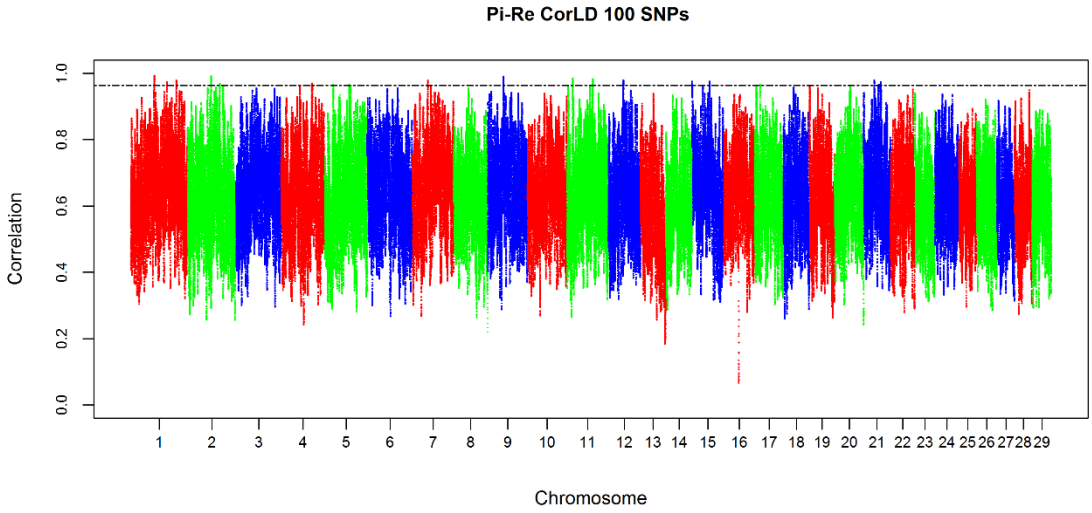


Figure 3.36. Manhattan plot of the VarLD estimates along the genome for the Pi-Re for regions of 100 SNPs.

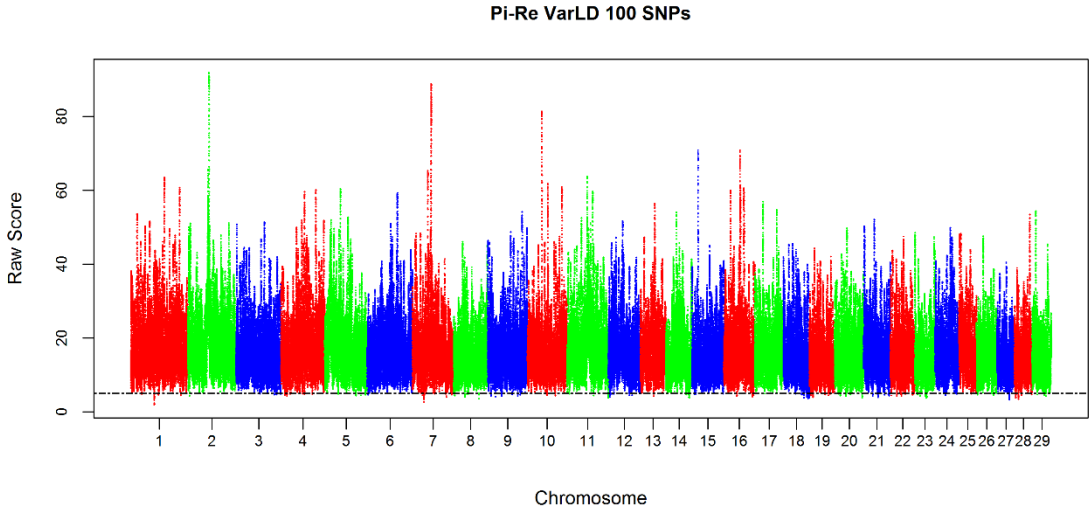


Figure 3.37. Manhattan plot of the CorLD estimates along the genome for the Pi-RG for regions of 100 SNPs.

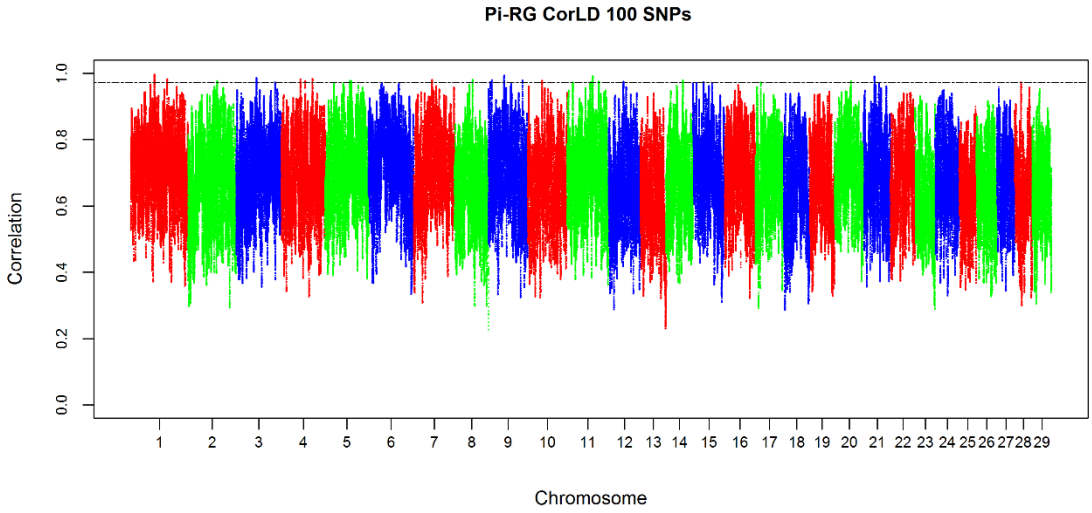


Figure 3.38. Manhattan plot of the VarLD estimates along the genome for the Pi-RG for regions of 100 SNPs.

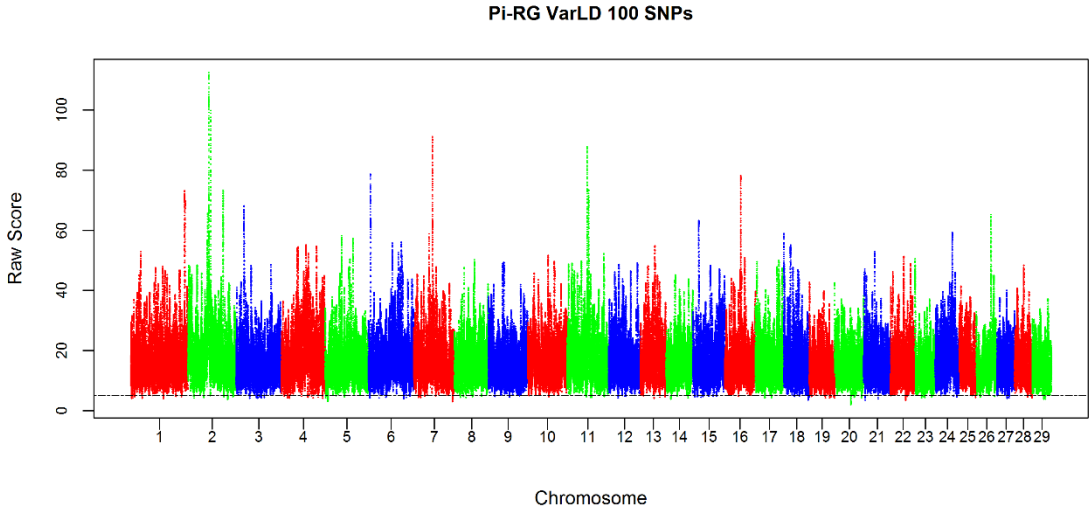


Figure 3.39. Manhattan plot of the CorLD estimates along the genome for the Re-RG for regions of 100 SNPs.

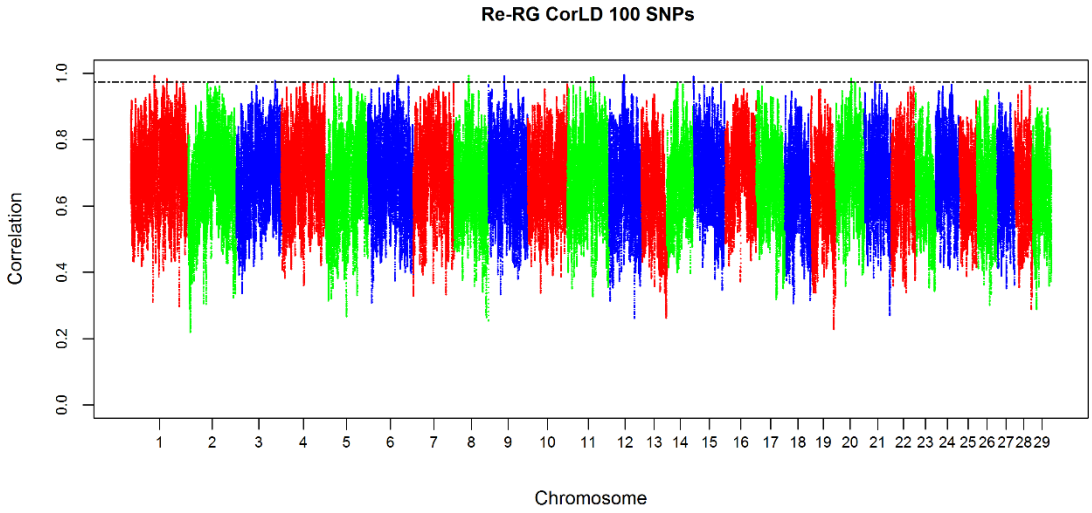
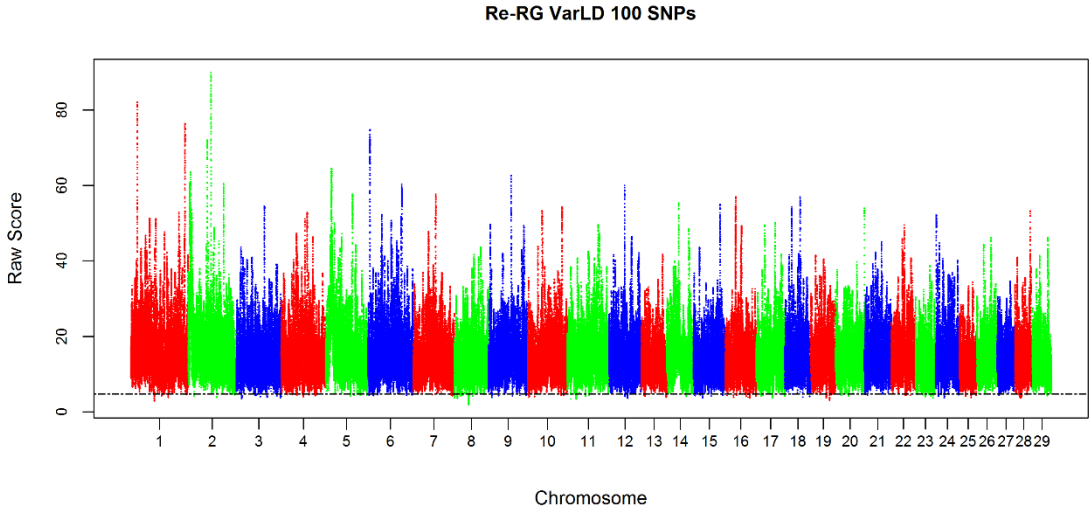


Figure 3.40. Manhattan plot of the VarLD estimates along the genome for the Re-RG for regions of 100 SNPs.



ANNEXE 4

Table 4.1. Genomic Regions with higher haplotype diversity.

BTA	Start	End	Populations	Genes
1	44485300	45386640	BP, Mo, AV, ANI, Pi, RG	EFHB, RAB5A, PP2D1, KAT2B, SGOL1
1	94927808	95300080	AV, BP, Mo	SPATA16, GPX5, ZSCAN23
2	104618864	105209704	BP, Pi, Mo, AV, RG, Re	TMEM196, PECR, XRCC5, MARCH4, SHOX, SMARCAL1, RPL37A
3	119233792	119700584	ANI, Pi, RG, AV, BP, Re, Mo	HDAC4, CSF2RA
6	116787544	117371128	BP, Re, AV, ANI, Mo, Pi	LDB2
11	5549297	5882842	RG, Pi, ANI	PDCL3
17	40887700	41123088	Pi, AV, RG	FNIP2
21	59640020	59962772	AV, BP, Mo, RG, ANI, Pi	SERPINA4, SERPINA5, SERPINA11, SERPINA12
29	50132888	50641688	Re, BP, ANI, Mo, AV	TNNT3, LSP1, TNNI2, SYT8, CRLF2, AP2A2, IFITM10, CHID1, TSPAN4