Università degli Studi di Sassari

SCUOLA DI DOTTORATO DI RICERCA

Scienze dei Sistemi Agrari e Forestali

e delle Produzioni Alimentari

Indirizzo Scienze e Tecnologie Zootecniche

Ciclo XXIII

Use of genomic information in the genetic evaluation of livestock

dr. Giustino Gaspa

Direttore della Scuola	prof. Giuseppe Pulina
Referente di Indirizzo	prof. Nicolò P. P. Macciotta
Docente Guida	prof. Nicolò P. P. Macciotta

Anno accademico 2009- 2010

Alla Mia Famiglia

All’amato me stesso
(Vladimir Majakovskij)
…&

RINGRAZIAMENTI

Vorrei esprimere la mia gratitudine a tutti quelli che, in qualsiasi misura, hanno contribuito alla realizzazione di questo lavoro di tesi: ai professori, ricercatori e colleghi del corso di dottorato del Dipartimento di Scienze Zootecniche.

Vorrei ringraziare l'ente finanziatore, MIPAF (grant SELMOL) che ha consentito la realizzazione di questo lavoro.

Infine, un particolare ringraziamento va al Prof. Aldo Cappio-Borlino, e al Dr. Ezequiel Luis Nicolazzi.

ACKNOWLEDGEMENTS

I would like to acknowledge all people who contributed to any extent to this PhD thesis: Professors, researchers and postgraduate fellows of the Department of Animal Science.

I would like to thank the MIPAF (grant SELMOL) for funding my scholarship.

Finally, a special thank goes to Prof. Aldo Cappio Borlino and Dr. Luis Ezequiel Nicolazzi.

TABLE OF CONTENTS

INDEX OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

Chapter 4

CHAPTER 5

INDEX OF TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 1
GENERAL INTRODUCTION

USE OF MOLECULAR INFORMATION IN ANIMAL BREEDING

1.1 INFINITESIMAL MODEL AND SELECTION UNDER A FINITE LOCUS MODEL

1.1.1 The quantitative genetic approach

Most of the traits of economic interest in livestock have a complex quantitative expression coded by a large number of genes and affected by environmental factors. Statistical analysis of phenotypes and pedigree information allows to estimate the genetic merit (*breeding values*) of the animals candidate to selection following the Fischer's infinitesimal model, according to which observed phenotypes are determined by an infinite number of loci, each with an infinitesimal additive effect. Under this hypothesis, mean of a quantitative trait in a population can be modified choosing the best genotypes based on the breeding values estimated using Best Linear Unbiased Predictors (BLUP) methodology. In the best situation all sources of information on phenotypes and additive relationships among animals are included in a BLUP model to estimate a breeding value for all the animals in the population. Thus, the genetic gain (ΔG) per year, for a particular trait, could be obtained according the RENDEL and ROBERTSON (1950) formula:

——

where σ is the genetic standard deviation of the trait, ρ is the accuracy with which the breeding value of the selection candidate is estimate, *i* is the intensity of selection and T the generation interval or the average age of parents when their offspring are born.

Despite a generally considered limited theoretical foundation, the infinitesimal model (generally defined as *black box approach*) allowed to reach high rates of genetic improvement in many livestock species in the last decades. (DEKKERS and HOSPITAL 2002). A relevant constraint to the genetic progress is represented by the inverse relationship between accuracy of breeding values and generation interval, kept constant the other variable in the equation of genetic gain. Hence, the more reliable breeding value we want to estimate, the more time we need to wait. The generation interval is particularly large for sex-limited traits in the context of progeny testing in dairy cattle (about 60 months are needed to get the first estimated breeding value for a progeny tested bull). Several strategies have been proposed to increase the response to selection. For instance, the use of multiple ovulation embryo transfer (MOET) and in-vitro fertilization (IVF) (KRUIP *et al.* 1991; ROWSON 1971) were aimed at increasing the intensity of selection on the female line. Furthermore, in the past decades, thanks to the advances in the molecular techniques, a large number of genetic markers have been discovered. Possible strategies to use and integrate these new sources of information with the aim of enhancing the accuracy of selection has been extensively reviewed and proposed from different authors (FERNANDO and GROSSMAN 1989; LANDE and THOMPSON 1990; MEUWISSEN *et al.* 2001; DEKKERS 2004; DEKKERS and HOSPITAL 2002)

1.1.2 The finite locus model and the use of Quantitative Trait Loci (QTL) to enhance the response to selection

In the last 20 years, due to the application of advanced techniques in molecular genetics and statistics, several chromosomal regions that influence quantitative traits have been discovered. Moreover, the finite amount of DNA in the mammalian genome suggests that must be a finite number of loci that control the expression of quantitative traits (between 20,000 and 35,000 genes) (EWING and GREEN 2000), in contrast with the infinitesimal gene model. HAYES and GODDARD (2001) investigated the distribution of the QTL effects in dairy cattle and swine, enforcing the evidence that there are few genes with large effect and many of small effect.

How this relevant amount of knowledge is going to change the selection of farm animals is still an open issue. Combinations of molecular and classical quantitative information in a composite selection index have been proposed to increase the accuracy of selection (LANDE and THOMPSON 1990). Several approaches have been indicated to integrate molecular information in current breeding programs. The base principle is that genetic markers are available early in life, so that the accuracy of breeding values estimated for young animals can be increased and the generation interval reduced.

DEKKERS (2004) defined three types of genetic markers that can be used in practical implementation of molecular information into breeding programs:

1) Direct markers
2) LD markers
3) LE markers

The direct markers are those that code for a functional mutation; the LD markers are loci in population–wide linkage disequilibrium (LD) with the functional mutation; LE makers are in population-wide linkage equilibrium with the functional mutation, but in LD within family.

In particular, here we refer to marker assisted selection (MAS) and marker assisted introgression (MAI) to indicate the use of genetic markers in linkage disequilibrium with a QTL in breeding program (i.e. LD and LE markers). Whilst, in the gene assisted selection (GAS), the causative mutation (direct marker) of a gene that affect the expression of a quantitative trait is used for the calculation of molecular score of animals.

Use of molecular data represents an opportunity to enhance the response to selection especially for low-heritability traits, or whose phenotype is difficult or expensive to measure or expressed later in age. Sex-limited traits, such as milk production in dairy cattle, can be objectives of selection based on molecular data, in order to reduce the generation interval. For such traits the molecular based breeding value can be available early in life and for both gender (DEKKERS and HOSPITAL 2002).

Although advances in molecular genetics have been able to explain part of the genetic variances due to QTLs, the possibility of implement this information in a MAS program has been limited by several reasons. Firstly, only a limited number of genes have been identified. Secondly, in most cases the marker map used in the past were sparse, so that the QTL have been be mapped within very large confidence interval. Further, even using LD marker, the selection is not directly on the QTL or gene, but on the marker in LD with the QTL. Because of LD decreases across generations due to recombination, marker effect needs to be re-estimated frequently. Furthermore, the estimates of the QTL effects are generally biased (WELLER *et al.* 2005) and in particular are overestimated. All these issues, together with the relatively high cost of genotyping, have reduced the commercial application of marker information collected in more than one decade of researches on QTL mapping in livestock. So far, a successful example of MAS have been reported, both for simulated and real data, for the French MAS program (GUILLAUME *et al.* 2008a). The authors showed that marker assisted breeding values of Holstein bulls were on average 4% more accurate than the pedigree based breeding values (GUILLAUME *et al.* 2008b). Another example of application of MAS, for pre-selection of bulls before entering in progeny testing, has been proposed by BENNEWITZ *et al.* (2004b). Both of examples are based on the application of FERNANDO and GROSSMAN (1989) BLUP model.

1.1.3 Genome-wide approach to estimate breeding values: challenges and prospectives

More recently, the availability of high-throughput sequencing techniques allowed to discover thousands of single nucleotide polymorphism (SNP) spread across the whole genome in several livestock species. Currently, chips for genotyping animals at more than 50,000 marker loci are commercially available (VAN TASSELL *et al.* 2008) Such a map density is enough to find LD between marker and QTL, thus to looking for associations between traits and markers without specific knowledge of population structure.

These new techniques give rise to new opportunities for genetic evaluation of farm animals with a so called genome-wide approach (MEUWISSEN *et al*., 2001). On one hand, this new advance allows to explore the genome looking for QTLs and associations between SNP and phenotypes. On the other hand, it allows to use directly the marker information to estimate genomic breeding value (GEBV). In the former case we talk about genome-wide association (GWA) studies, while in the latter, the term genomic selection (GS) is generally adopted.

MEUWISSEN (2007) defined GS as Marker Assisted Selection on a genome-wide scale. Briefly, the GS rely on the segmentation of the genome using a dense marker map in thousands of bits, each contributing to the explanation of part of the genetic variance of a quantitative trait. The effect of each segment is estimated in a reference population (animal with phenotypes and genotypes). Then effects are used to predict the breeding values of another set of genotyped animals (prediction population) without phenotypes.

MEUWISSEN *et al.* (2001) first proposed to use dense marker information to predict the breeding values of animals. Afterwards, lots of models and approaches – mainly on simulated dataset – have been proposed to deal with the statistic key issue of practical implementation of GS: the great asymmetry of data matrix i.e., the number of effects (single marker or haplotypes) to estimate is highly greater than number of phenotypic records available.

In brief, potential advantages of using high density markers in genomic evaluation are the following: i) each QTL is expected to be in LD with at least one marker; ii) all the genetic variance is taken into account in the estimation of breeding values; iii) the animal can be genotyped early in life, and this may guarantee a reduction of generation interval; iv) furthermore, a better estimation of mendelian sampling term may give rise at lower inbreeding rates (DAETWYLER *et al.* 2007).

On the other hand, open challenges of GS are: i) the computational issues and the choice of a suitable statistical framework; ii) the size of the reference population, that should be large enough to ensure reliable estimates of DGV; iii) the practical implementation of GS in current breeding program, or adaptation of breeding program to genomic evaluation; iv) frequency of re-estimation of SNP effects; v) comparison of genomic predictions across countries.

1.2 OVERVIEW OF QTL MAPPING EXPERIMENT IN LIVESTOCK

Two approaches have generally been used to detect QTLs: the candidate gene approach and the anonymous marker approach. The former seeks causative mutations in all possible genes involved in the known biology of the considered trait, analyzing if variations of particular regions of DNA are significantly associated with variations on phenotypic expression of the trait. The latter, instead, assumes that the genes underlying a quantitative trait are unknown, and it makes use of neutral markers to scan the genome testing statistically the associations between markers and phenotypes.

1.2.1 Evolution of QTL mapping techniques in dairy cattle

Experimental design

Several statistical techniques to map QTL in animal populations have been proposed in literature (DOERGE 2002;ANDERSSON and GEORGES 2004; RON and WELLER 2007). The segregation analysis allows to follow the inheritance of marker alleles from parents to offspring. Most common experimental designs used for QTL mapping exploit the pedigree structure of current livestock populations to search for linkage disequilibrium between putative QTL and genetic markers. Higher power to map QTL is generally reached in back cross design or F2 design. However, these designs can be not realistically used in dairy cattle populations, due to the fact that most breeding programs are based on within breed selection (Ron and Weller 2007). However, paternal half sib families are very common in dairy cattle due to the extensive use of artificial insemination (AI). Experimental designs that exploit such family structure have been extensively used to seek QTLs. The daughter design (Figure 1) and the granddaughter design were mostly used for QTL detection (Weller *et al.* 1990)

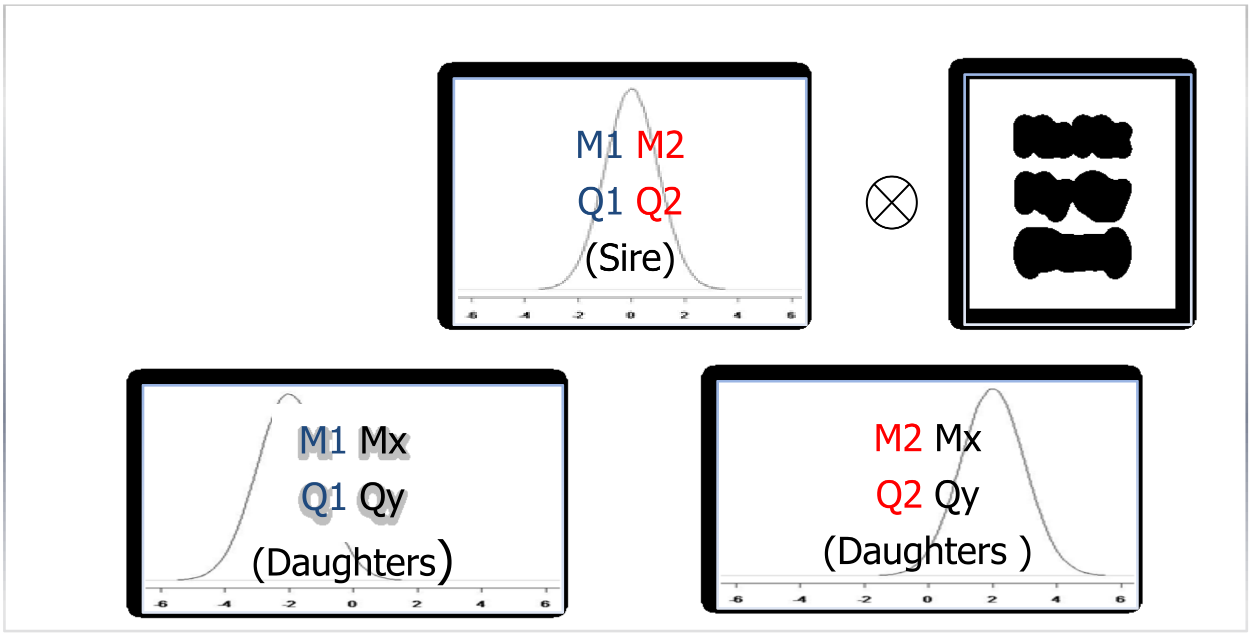


Figure 1 Scheme of daughter design

In the daughter design, groups of offspring of a sire, heterozygous for the marker, are sorted on the basis of the allele that they have inherited. Any significant phenotypic difference between the groups of offspring suggests that such marker is involved in the expression of a quantitative trait. In the granddaughter design genotypes are collected from AI bulls (grandsires) extensively used and groups of their sons (sires). The phenotypes are collected from a large number of daughters of the AI sires. This design uses marker information over two generation and allows to reduce markedly the number of genotype required to get the same power obtainable using daughter design (Weller *et al.* 1990). Variants of these designs have been proposed to reduce the number of genotyping costs, which in the past have represented the main economic limit to achieve sufficient power to detect QTLs. In particular, selective genotyping (Darvasi and Soller 1992), selective DNA pooling (Darvasi and Soller 1994) fractioned DNA pooling (Korol *et al.* 2007) have been proposed to reduce the cost of QTL mapping experiments. The rationale of the selective genotyping design is to determinate the linkage between marker loci and QTL by genotyping only individuals from the high and low tails of the distribution of phenotypes. Pooling DNA from extreme individuals in the tail and testing for linkage on marker allele frequencies derived from pooled samples represents one step beyond the selective genotyping and allows to reduces further the costs of genotype determination.

Using the former designs, only the effect of paternal QTL could be modeled and the contribution of maternal line are ignored. Such limitation is expressed by the fact that is not possible to fully exploit the informativeness of the markers. Some solutions have been proposed in literature to address this issue by using haplotypes. For example, the use of pedigrees with sufficient genetic links and identical by descent (IBD) mapping techniques, or combined linkage and linkage disequilibrium analysis (LDLA) (MEUWISSEN *et al.* 2002) have been proposed. The LDLA analysis is almost the same of the general pedigree linkage analysis but also included IBD estimates between founders that are not related through pedigree but are assumed to be related through a common, unknown, ancestor in whom the mutation was supposed to arise (DE KONING 2006). Through fine mapping and haplotype analysis it is possible to infer the phase of QTL and to detect haplotype, or single mutation for which at the QTL locus the genotype in the parents and offspring are consistent.

Once identified a block of conserved haplotypes in LD with causal mutation, it is possible to test the effect of the haplotype or single mutation on phenotype in the population without need to know the structure of the population performing LD mapping.

Statistical framework

Different statistical methods have been proposed to map QTLs in livestock population. The simplest approach is to perform a single-marker test to find which markers are associated with the phenotypic value of the quantitative trait analyzed. The null hypothesis usually tested is that the mean of the phenotypic value is not associated to the genotype at a particular locus. Some issues of this approach concerned: i) inability to provide an estimate of QTL location or recombination frequency between marker and QTL; ii) effect of the size of the sample to obtain sufficient power to detect QTL; iii) multiple testing and the choice of an appropriate significance threshold (DOERGE 2002).

Some issues of the single-marker analysis were overcome when the availability of genetic markers increased considerably due to advances in sequencing techniques. The use of additional genetic information on location and order of marker included in a genetic linkage map, not considered in the single-marker approach, allowed to calculate the frequency of recombination and the position of each marker in a linkage map. The technique of interval mapping (LANDER and BOTSTEIN 1994) provided a powerful tool for exploiting genetic information in order to map QTL more accurately. This method uses a pair of closed markers bracketing a portion of genome harboring a putative QTL. The interval mapping statistically tests for a single QTL at each position along the genome. The results are reported as logarithms of the odds (LOD) scores calculated by comparing the value of the likelihood function under the null hypothesis of no QTL segregating with the alternative hypothesis of presence of QTL. A value of LOD score of 3 means that there are 1 chance out of 1,000 to reject the null hypothesis of no presence of a QTL segregating for such position. All these approaches led to the identification of numerous QTL regions in several genome scan carried out across many countries in dairy cattle (GEORGES *et al.* 1995; HEYEN *et al.* 1999; MOSIG *et al.* 2001; OLSEN *et al.* 2002; RON *et al.* 2004ASHWELL *et al.* 2004; BAGNATO *et al.* 2008). Nonetheless, these QTLs have been mapped with moderate to large confidence intervals (QTL region spanning tens of centimorgans (cM) may harbor hundreds of genes) and for this reason they have had a limited use in MAS programs. Furthermore, the identification of the causative mutation or quantitative trait nucleotide (QTN) that underlying a mapped QTL region is even more tricky owing to the lack of direct relationship between phenotype and genotype. This is due to the fact that a single QTL explain only a proportion of the phenotypic variation, the rest being caused by other QTLs or environmental factors (ANDERSSON and GEORGES 2004). Multi-step strategies are necessary to detect and validate genes involved in the expression of complex quantitative traits (figure 2) (RON and WELLER 2007), and even if experimental evidences lead to a suitable candidate gene, is not ever possible to identify a QTN unambiguously (DE KONING 2006)

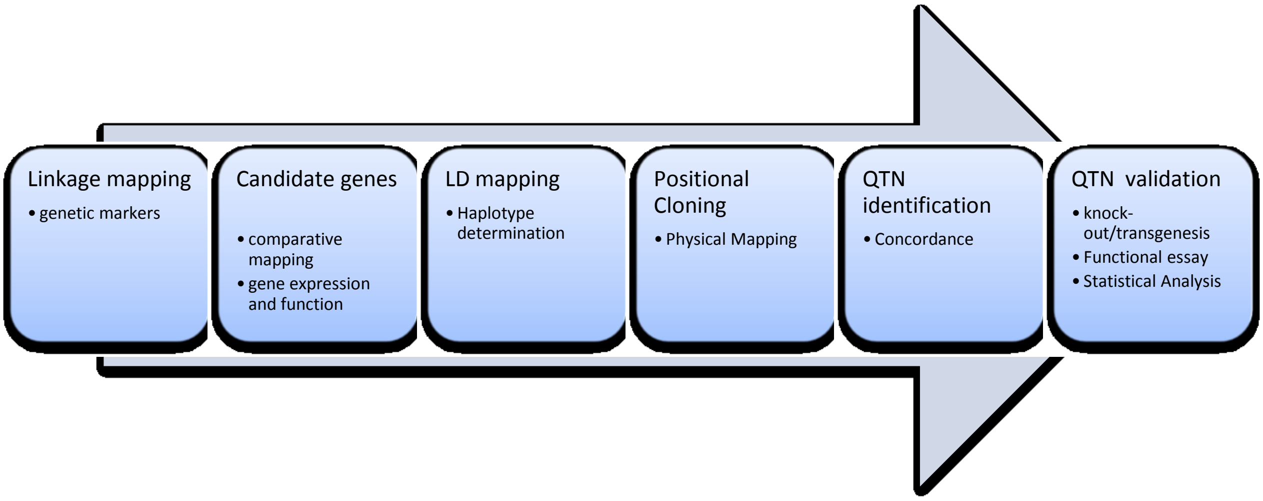


Figure 2 Multi-step strategy to detect and validate a QTN - modified from (RON and WELLER 2007)

1.2.2 Key Experiments and Polymorphisms that affect quantitative traits in livestock species

The existence of mendelian genetic factors affecting the expression of a quantitative trait has been demonstrated in a brilliant experiment by SAX (1923) who associated the weight of the seed of *Phaseoulus vulgaris* with the genotype at a locus that controlled the expression of a mendelian trait like the color of the seeds. Later, THODAY (1961) put the basis of definition of QTL or major genes concepts. However, the idea of increasing the rate of genetic improvement using molecular markers date to 1960 when NEIMANN-SORENSEN and ROBERTSON (1961) proposed to use the blood variant group (biochemical marker) to select for quantitative traits.

A work of SMITH (1967) put the basis of marker assisted selection and its usefulness specially: "*When normal selection is effective, further information on known loci can add only a little to the rate of improvement. But if normal selection is not very effective, as for characters of low heritability, or if indirect selection on relatives must be used (as for sex-limited or carcass traits) then known loci may add significantly to the rate of improvement possible*".

However, the extensive use of DNA variations started when DNA polymorphism discovery – restriction fragment length polymorphisms (RFLPs), minisatellite and microsatellite markers, single nucleotide polymorphisms (SNPs) – allowed to build genetic maps. These maps were initially very sparse but as soon as new polymorphisms were discovered, denser and denser genetic maps were created. The first extensive genome scan using a map covering almost the entire genome (60% of coverage) was carried out by GEORGES *et al.* (1995). Thereafter, several genome scans have been performed in livestock species across several countries and different species. These studies mainly focused on mapping QTLs, tracing the inheritance of microsatellite markers in group of progeny of sires that had different phenotypic expression according different experimental designs.

Before including a locus involved in the determinism of a quantitative trait in a program of gene or marker assisted selection it is necessary to establish the influence of such gene on the phenotype. Currently only few polymorphisms in gene sequences have been unambiguously linked to variation in quantitative traits (RON and WELLER 2007). Conversely, numerous genomic regions have been tested for the association with productive and functional traits, and in literature several polymorphisms have been reported to be associated to phenotypic trait values. However, many polymorphisms investigated in dairy cattle are suitable candidate genes but clear and concordant evidences of QTN have been obtained only in few cases. The K232A substitution in DGAT1 gene located on BTA14 and its association with increased milk fat content and protein percentage, and decreased milk yield (GRISART *et al.* 2002; SPELMAN *et al.* 2002; GRISART *et al.* 2004) and GDF8 (affecting double-muscling) in cattle represents one of the most popular example. Other genes involved in lipid metabolism FASN and ACC-α (ROY *et al.* 2006; MORRIS *et al.* 2007), both located on chromosome 19, have been reported to affect the milk fatty acid composition. Furthermore a polymorphism at the SCD gene (Valine to Alanine substitution at position 239) in chromosome 26 has been associated to change in milk fatty acid composition and milk production in Cattle (MELE *et al.* 2007; MACCIOTTA *et al.* 2008). Many other QTL mapping studies have found strong signals in chromosomal region and a more extensive review will be provided in chapter 1.

1.2.3 Linkage Disequilibrium in livestock and QTL mapping

Linkage disequilibrium is the nonrandom association of alleles in haplotypes at different loci within a population. The general principle of identification of QTL is based on the presence of LD between QTL alleles and marker loci. Let *A* and *B* being two markers in the same chromosome carrying two alleles A_1 A_2, and B_1 and B_2 respectively. Four different haplotypes (A_1B_1, A_1B_2, A_2B_1, A_2B_2) are possible. If the frequencies of A_1, A_2, B_1 and B_2 alleles is 0.5 in the population (random association), then the frequency of 0.25 for each haplotype is expected. In these situation the population is in linkage equilibrium. Any deviation from 0.25 in the haplotype frequency means presence of LD.

LD has been exploited in fine-scale mapping studies of human disease loci (CARDON and BELL 2001; RISCH and MERIKANGAS 1996) since the increasing availability of haplotype data represents the basis for historical or evolutionary inference. In livestock the use of haplotype data is mainly focused on identification of DNA regions affecting the expression of quantitative traits. The availability of dense map information allowed to use LD information for livestock QTL mapping and genomic breeding values estimation.

Both linkage analysis (LA) and LD mapping are techniques which exploit the LD existing in animal population to map QTL in a different way. Linkage analysis (LA) measures the association exploiting pedigrees and considers the LD that exist within families. In LA mapping the association between markers and QTL is broken down by recombination after few generations whereas LD mapping refers to associations between markers within populations of unrelated individuals. In the latter case the association persists for a considerable number of generations (i.e. makers and QTL in LD must be closely linked). Summarizing, pedigree studies analyze recombination events that involve exchanging *megabase* fragments of chromosomes whilst LD studies deal with fragment measured in *kilobase*. Hence, the allelic states of closely linked loci will be correlated, whereas those of distantly linked loci will be more-or-less independent (NORDBORG and TAVARE 2002). The causes of LD in natural conditions are different: i) genetic drift associated with reduction of population size; ii) mutations; iii) natural or artificial selection that may favor certain allelic combinations; iv) migration of a population that is mixed with another gene frequencies (FALCONER, 1996)

To analyze relationships between markers and QTLs is necessary to introduce the concept of distance map that defines the distance separating two genes or markers located on the same chromosome. The unit of distance map (*d*) (centimorgan (cM)) is not a fixed measure and does not correspond to a fixed number of base pairs, but depends on the number of recombinations which occur between two genes. Different functions which associate the frequency of recombination (θ) with the distance map have been proposed. The simplest function map is $d = \theta$ (MORGAN, 1909) or Morgan mapping function. Using the Morgan function the θ tends to underestimate d between two loci because it takes into account of recombination events occurred only in odd numbers. Other mapping function are Haldane (1919) based on Poisson distribution, where and Kosambi function (1944) .

Measure of LD

Different formulas have been proposed in literature to measure the extent of LD. Among these *D*, *D'*, r^2 and $\chi^{2'}$are presented below. The measure D was proposed by HILL (1981)

Where *freq*(A_1B_1) is the frequency of this haplotype in the population and likewise for the others haplotypes. D may be also expressed as function of the rate of recombination θ according to:

D = 0 indicates a state of linkage equilibrium (θ = 0.5), positive values of D indicates presence of linkage disequilibrium thus θ will be less than 0.5.

LD tends to decline in populations because the recombinant gametes (A_1B_2, A_2B_1) continue to occur from parental (A_1B_1, A_2B_2) and vice versa. The process is much slower when the rate of recombination (θ) is smaller. This process is described by where t indicates generations passed from generation 0. Figure 3 shows the decrease of LD through generation for different recombination rate.

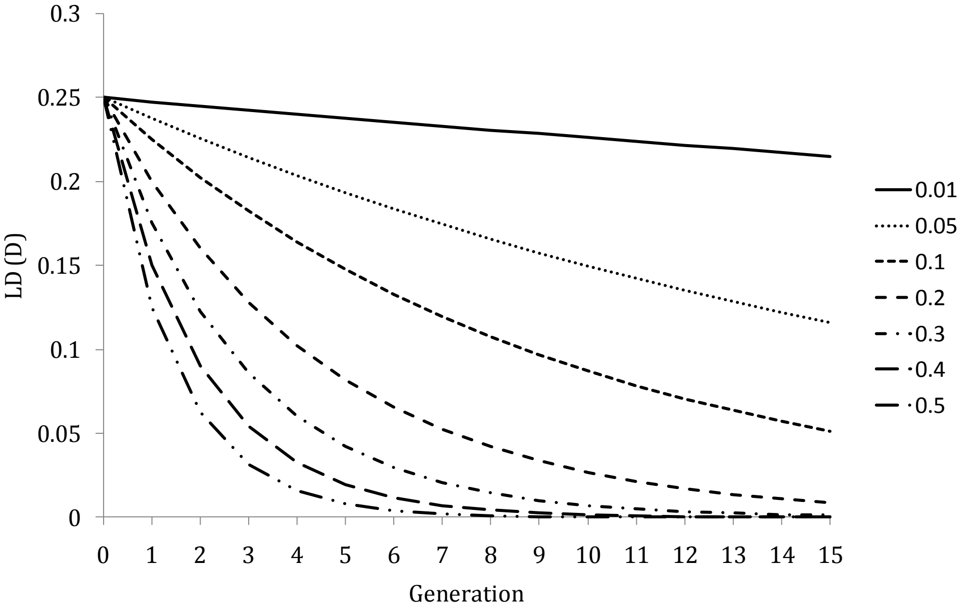


Figure 3. Decrease of LD as function of number recombination rate through generation

The measure D is strongly dependent on the frequency of individual alleles and it is not particularly useful for comparing the degree of LD between many pairs of loci. HILL AND ROBERTSON (1968) proposed a statistics, r^2, less dependent from allele frequency:

where *freq* (A_i) and *freq* (B_i) are the frequencies of *i*-th allele of *A* and B in the population respectively. The r^2 value varies from 0 for a pair of loci with no LD between them to1 for a pair of loci in complete LD (NORDBORG and TAVARE 2002). The usefulness of r^2 rely on the fact that this statistics measure of proportion of variance of QTL explained by a marker in LD with QTL.

Another commonly used measure of LD is D ':

where

The statistical value r^2 is preferred to D' as a measure of the extent of LD for two reasons: i) D' is strongly dependent on the frequency of individual alleles and ii) D' tends to overestimate LD with small samples or for low allele frequencies.

Previous measures of LD are suitable only for bi-allelic markers. To measure the LD between multi-allelic markers ZHAO *et al.* (2005) proposed the statistics χ^{2}' calculated as follow

Where: where *freq* (A_i) is the frequency in the *i*-th marker allele A *freq*(B_j) is the frequency of the j-th marker allele in B , *l* is the minimum number of alleles of marker A and B. In the case of bi-allelic markers is valid, the following identity $\chi^{2\prime}=r^2$

The use of DNA markers and the development of technologies for their analysis have allowed to explore the genomes of animals and to construct very dense genetic maps for all major species of livestock (figure 4). These high dense maps may guarantee that markers are tightly linked to QTL, enabling the LD mapping techniques to find genome-wide association between markers and phenotypes and genomic selection procedure.

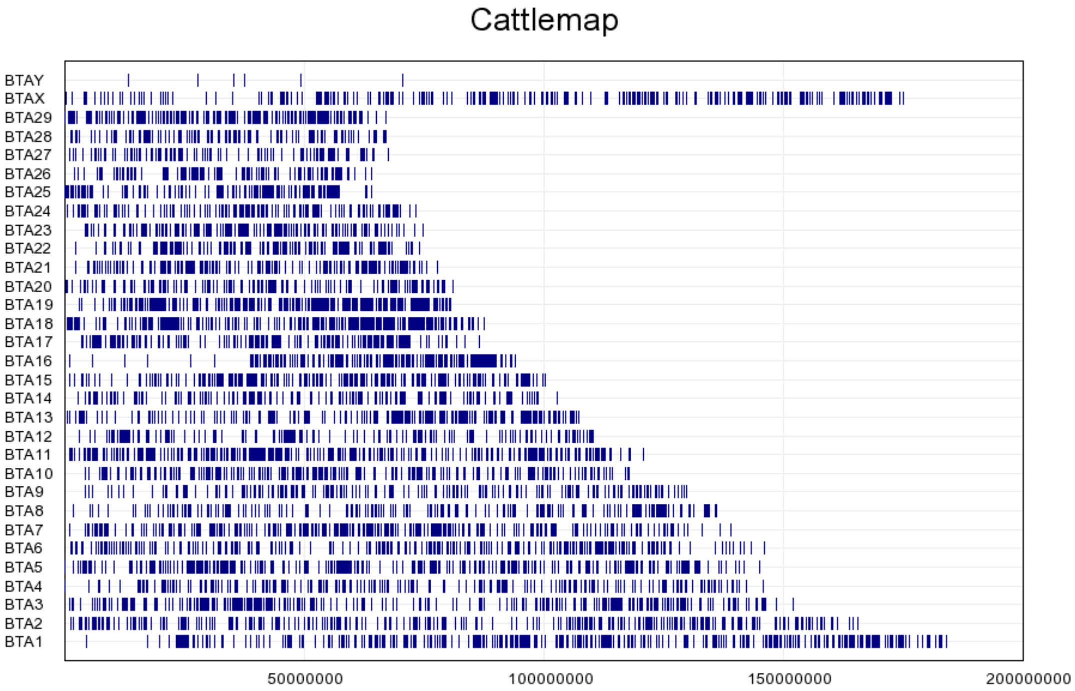


Figure 4. SNP cattle map *http://www.livestockgenomics.csiro.au/cow/cattlemap.html*

Linkage disequilibrium (LD) mapping of QTL exploits population level associations between markers and QTL. These associations are more likely due to the small distance between markers in the chromosome. These chromosome fragments, tracing back to the same common ancestor, will carry identical marker alleles or marker haplotypes. If a QTL is located somewhere within the chromosome segment, they will also carry identical QTL alleles. There are many QTL mapping techniques which exploit LD, the simplest one is the genome wide association test using single marker regression. Some statistical packages are also available to map QTL and conduct Genome-wide association studies (AULCHENKO *et al.* 2007b)

The main factors influencing the power of the association test to detect a QTL are: i) the level of linkage disequilibrium (r^2) between the marker and QTL; ii) the proportion of phenotypic variance explained by the QTL; iii) the sample size*; iv)* the allele frequency of the rare allele of the SNP or marker (*p)*. The power is particularly sensitive to low level of frequency (p<0.1); v) the significance level a set by the experimenter and the multiple testing issues.

1.3 GENOMIC SELECTION

The implementation of information about thousands of genetic markers into the current breeding programs has become feasible due to the availability of dense markers maps and to the quick development of SNP chip technology, now affordable also for some livestock species. Genomic selection (GS) is an new and important tool for the genetic improvement of farm animals which allows to estimate direct genomic values (DGV) of candidate to selection using dense marker maps without need to record the phenotypic performances of the animals (or of its relatives).

1.3.1 Genomic Selection: principle and applications

GS relies on the segmentation of the genome in thousands of intervals bracketed by contiguous SNPs and on the estimation of SNP (or haplotype) effects across the whole genome. Currently up to 54 K SNP chips are commercially available for cattle and 800K will be produced in the very next future (Illumina Inc. www.illumina.com). With such a density the chance of recombination between markers and QTLs is very low. In other words, each QTL is expected to be in LD with at least one marker (MEUWISSEN *et al.* 2001; CALUS 2010). In the GS framework, each SNP gives its contribution to the explanation of the total genetic variance for a quantitative trait, hence potentially the whole genetic variance may be explained by the markers (GODDARD and HAYES 2007).

Different statistical methods have been developed to capture the variance due to the genetic markers. Basically, the GS procedure relies on the estimation of the effect of each DNA segment in a reference population (animal with phenotypes and genotypes). These estimates are later used to predict the breeding values in the prediction population (genotyped animals without phenotypes). How to build the reference and prediction populations and how this choice affects the accuracy of DGV are still open issues. HABIER *et al.* (2010) pointed out how the relatedness among animals of the reference and prediction populations affects the accuracy of the DGV estimates. However, in the general GS framework, the estimation of the SNP effects is carried out in a set of proven bulls (for which reliable EBV are available). These estimates are used to predict the DGV of young bulls candidate to selection, and hardly these group of bulls are not somehow related in a dairy cattle population.

Several the statistical methods have been proposed for solving the main statistical issue of genomic selection: the great asymmetry in the data matrix (also called "*p>>n problem"*), due to the very large number of marker effects that need to be estimated (up to tens of thousands) in comparison to the limited number of phenotypic records generally available (around thousands). Several methodologies have been suggested to estimate marker effects with the basic aim to reduce the number of predictors. such methodologies can be classified into two main categories: *i)* methods that select a subset of original markers on the basis of their relevance to the considered trait, and *ii)* methods that summarize the information of original SNP with a smaller number of derivate variables using multivariate or non-parametric statistical technique. A further classification may distinguish between the methods that considers equal contribution of each SNP to the genetic trait variance, or techniques that assume different variance for each SNP thus taking into account the distribution of QTL effects, with many loci with small (or close to zero) effect and very few loci with large effect (HAYES and GODDARD 2001).

Advantages & open issues of GS

Use of GS may allow to achieve an extra genetic gain compared to the classical polygenic EBV estimation, due to the higher accuracy of estimated EBVs are, especially for low heritability, sex-limited and expensive or difficult to measure traits. Furthermore, the ability to reduce the generation interval due to an earlier estimation of the genetic merit is another advantage of using GS, considering that potentially each animal could receive an EBV at birth. Although QTL mapping is not the main goal of genomic selection, some statistical models (Bayesian methods, in particular) may be of help for identifying genome regions that affect a number of economic traits (COLE *et al.* 2009; CALUS 2010).

GS may radically modify the structure of livestock breeding programs, especially for dairy cattle. The potential usefulness of GS could be examined at different level: for instance, the genomic evaluations could be used to pre-select young bulls entering progeny test, or to select sire son or sire of dam. In the former cases the progeny testing scheme will disappear (GODDARD and HAYES 2007). A dramatic reduction of the cost of the genomic breeding program was predicted in a simulation study by SCHAEFFER (2006). Assuming an average accuracy of DGV of 0.75 at the time of birth of the animal, Schaeffer estimated a reduction of cost for proving bulls around at 92% with an increase in genetic gain twice as the current breeding schemes. The extra genetic progress is due to the reduction of generation interval (from 6.5 y to 1.75 y and from 6 y to 1.75 y for sire of sires and sire of dams, respectively). Further reduction of the generation interval was predicted for the dam of sire where genotyped (from 5 y to 2 y), no predicted changes concerned the dam of dams line of selection instead. Moreover an increase of the accuracy in the female side of the pedigree is expected. These results were also confirmed by KONIG *et al.* (2009) in a deterministic simulation carried out to compare the progeny test scheme against the genomic breeding program. The authors simulated a population of 100,000 cows and measured the discounted profit for a breeding goal of production and functionality. According to their findings, if the accuracy of genomic predictions were greater than 0.7, an economic advantage of genomic selection programs was up to a factor of 2.59. In both examples the increase of genetic gain and reduction of cost were due the reduction of interval generation and increased accuracy of DGV of young bulls.

Although these results seem to suggest that marker enhanced breeding values can replace the traditional genetic evaluation, as pointed out by GODDARD and HAYES (2007) a more realistic solution may found in integrating all the sources of information – phenotypes, pedigree and genomic – into an improved EBV (GEBV). It is now widely assessed that the use of molecular data may not replace phenotypic data recording. Moreover, the genotypes may not be determined for all the animals in the population, and alternative solutions could be sought. GODDARD and HAYES (2007), proposed the use of selection index theory to combine the DGV and traditional EBV into a genomic breeding value (GEBV). Another option provides to estimate genomic predictions in the whole population inferring genotypes for un-genotyped animals. Smaller SNP chips have also been proposed as solution to increase the quota of genotyped animals (HABIER *et al.* 2009; WEIGEL *et al.* 2009), and different solutions have been proposed to select the a subset of SNPS (MACCIOTTA *et al.* 2009; WEIGEL *et al.* 2009).

Open issues for practical implementation of GS in dairy cattle populations concern the size and composition of the reference population (CALUS 2010). Although the number of phenotypic records is not the only factor that affects the accuracy of DGV, its role is very important because of the influence on the cost of GS breeding programs. Simulation studies indicated two thousand phenotypic records as the minimum threshold for achieve reliable estimates of DGV (HAYES AND GODDARD, 2009). The accuracy of DGV also depend on the statistical methods used to estimate marker effects and on the heritability of the traits (MEUWISSEN *et al.* 2001). According to DAETWYLER *et al.* (2008) the accuracy of marker effects depend on the heritability of the trait and number of phenotypes. Moreover, the optimal composition of the reference population should include a wide range of phenotypes and genotypes representative of the entire population. Even if best accuracies of DGV have been achieved when juvenile animals are more related to the reference population (HABIER *et al.* 2007; CALUS 2010), these accuracies rapidly decrease as the distance in generations between reference and prediction population increases (HABIER *et al.* 2010).

Furthermore, re-estimation of markers effects is necessary because the level of LD tend to decrease through the generation due to the recombination. Hence, the association between markers and QTL are broken down over time. This fact imply a reduction of the accuracy of the DGV when chromosome segment effects were predicted from a reference population that is genetically far from the juvenile animals. As consequence, the marker effects need to be re-estimated every two or three generation (MEUWISSEN *et al.* 2001; DE ROOS *et al.* 2007).

The use of different prediction equations and different methodologies to estimates the SNP effects in different countries makes the DGV barely comparable. Some new methodologies are needed to standardize the procedure of calculation of international DGV.

1.3.2 Models for Genomic Selection and choice of the statistical framework

Since the large amount of data, the choice of an appropriate statistical model and the realization of an effective algorithm to solve the model represent two main critical point in GS.

Data editing of SNP genotypes is generally the first step in genomic selection and genome-wide association studies. The editing of marker data is needed to clean data from scanning error of machinery used to read DNA sequence and to remove uninformative data as well. There is no a defined protocol. However, the most frequent edits are: elimination of animals with missing genotypes over a low arbitrary threshold (maximum 5% generally), or animals for which is demonstrated the inconsistency between pedigree and markers data. Routinely, SNP with minor allele frequency (MAF) under a certain threshold (2-5%) are dropped as well as uninformative monomorphic markers. Moreover, markers that significantly deviate from Hardy Weinberg (HW) equilibrium (p-value 0.01 or 0.05 are the threshold used) are deleted from the analysis.

The second step for the implementation of GS is the estimation of marker effects. The base model to estimate the SNP could be described as:

Where y_i is the phenotypic record for the animal i; μ is the general mean; z_{ij} is and indicator variable for the genotyp*e* – coded as 0,1 or 2 for homozygous at first allele, heterozygous and homozygous at second allele, respectively – at locus *j* for the animal *i; g_i* is the SNP effect at locus *j; n* is the number of marker loci considered and e_i is the random residual.

However, different parameterizations of this model may be adopted as well as a polygenic effect could be fitted. The effect of the markers could be treated as fixed or as random. In the latter case the variance component associated to the SNP effect may be estimated in different ways. Basically, the easiest way is to consider each SNP contributing to the explanation of an equal amount of genetic variance. First, the additive genetic variance is estimated (with REML algorithm for instance), then it is divided by the number of effects that need to estimate simultaneously. Moreover, haplotype effects may be fitted instead of single marker genotypes. The association phase between SNP and QTL do not need to be known if SNP effects are fitted instead of haplotypes. This fact may simplify the calculation. In fact, sorting SNP markers into haplotypes is generally carried out using probabilistic algorithms based on the knowledge on the relationship among animals, or based only on the marker information (SCHEET and STEPHENS 2006). Although several softwares are currently available to determine the more likely haplotype phases, some of them are quite time consuming. Furthermore, the use of the haplotypes increases considerably the number of effects to estimate. Provided that in the case of GS the advantage of using haplotypes is relevant only for lower marker density (CALUS *et al.* 2008), it is often convenient to model just the marker genotype effect.

Once estimated the SNP effects, the next step is the calculation of DGV. DGV computation is straightforward according to the formula:

Where DGV_i is the direct genomic value of the animal *i* provided the estimation of SNP effects $\hat{g}_j$ at locus *j*. The estimation step may be carried out using many different procedures, nonetheless the total direct genomic value of the animal is the summation across the whole genome of the effects of the SNP genotypes.

The model described above is a general model, but different statistical implementations may be used. Selection of markers was proposed as a strategy to address to the main statistical issue *i.e.* number of effects that need to be estimate is much larger than number of phenotypic records available. As previously said, the different approaches can be grouped into two main categories: selection of subset of SNP and use of limited number of derivate variables.

Selection of SNP subsets

A simple approach for selecting SNP subset could be carried out by using single marker regression of SNP genotypes on phenotypes (or ANOVA) in order to evaluate for each SNP a possible significant association to the phenotype analyzed (MACCIOTTA *et al.* 2009; MEUWISSEN *et al.* 2001). Significance threshold may be adjusted using Bonferroni correction or permutations of data to take into account of multiple testing issue. Methods of pre-selection of subset of SNP have been also applied to develop low density SNP chips. SNP selection have been carried out both selecting SNPs based on their relevance to the phenotype considered and on their location (evenly spaced) along the genome. This techniques allowed to reduce up to a factor 100 the number of markers (MACCIOTTA *et al.* 2010; WEIGEL *et al.* 2009) needed with limited losses in DGV accuracy. Exploiting the machine learning theory and variables and features selection techniques – developed in the field of information technology (GUYON *et al.* 2003; KOHAVI and JOHN 1997) – multi-step procedure of selection have been proposed to deal with the issues of reduction of dimension of genomic data of mortality rate in broilers (LONG *et al.* 2007).

Approaches which incorporate the selection step have been proposed and they make use of Bayesian statistics. MEUWISSEN *et al.* (2001) first proposed the use of so called Bayes A and Bayes B methods. Briefly, in the Bayes A is the marker data are modeled at two level: at the data level, and at the level of variances of SNP effects. If we allow the variance of the effects across SNP to vary, Bayes A estimates both SNP effect and their variance simultaneously using a Gibbs sampling algorithm. MEUWISSEN *et al.* (2001) indicated that the distribution of genetic variance across SNP is characterized by many loci which no harboring QTL, and few loci which do contain QTL. This fact led to a modification of the algorithm. In fact, in the Bayes B method a further step (Metropolis-Hasting) is implemented to determine for each locus whether it has an effect on the phenotype or not, in the former case the effect of that locus is shrunk to zero. These methods are heavily affected by the prior information used to infer the SNP effect as pointed out by some authors (CALUS 2010; GIANOLA *et al.* 2009)

Other alternatives include non parametric methods like kernel reproducing Hilbert space regression (KRHS) (GIANOLA *et al.* 2006; GIANOLA and VAN KAAM 2008) for prediction of total genetic value for quantitative traits, using phenotypic and genomic data simultaneously allowing also to model interaction among SNPs. Different applications of KRHS have been performed on field data in literature on chicken and dairy cattle (GONZALEZ-RECIO *et al.* 2008; GONZALEZ-RECIO *et al.* 2009; MOSER *et al.* 2009; DE LOS CAMPOS *et al.* 2010) comparing such methodology with other models. RHKS regression resulted as much accurate as Bayes A approach considering whole genome data and more accurate when smaller subsets of informative SNP were used (GONZALEZ-RECIO *et al.* 2009)

Use of Derivate variables calculated from SNP data

The second class of techniques includes methods that summarize the marker information with a smaller number of derivate variables. Multivariate techniques like principal component (PC) analysis and partial least square regression (PLSR)(WOOLASTON *et al*, 2007; SOLBERG *et al.* 2009) have been used to reduce the dimensionality of data matrix. In PC analysis a number of uncorrelated PC explaining a large part of the SNP variance are used as predictors instead of the original variable. The accuracy of DGV using PC scores as predictor variables in a BLUP were similar to the accuracy obtained using the whole set of markers, but required much less computational time (MACCIOTTA *et al.* 2010). When the predictors are many and highly collinear, PLSR may be also used to constructing predictive models. Similarly to PC analysis, the main principle of PLSR is to extract orthogonal (*i.e.* uncorrelated) components from the original predictors matrix (latent components) and use them as predictors. Differently from PC analysis, which perform the extraction of PC using as criterion the maximization of the variance of the predictors matrix, PLSR extracts the latent components with the constrain to maximize the covariance between latent components and response variable. PLSR resulted a suitable technique to calculate DGV and its performances are similar to other techniques more time consuming (MOSER *et al.* 2009; SOLBERG *et al.* 2009). The key difference between the two class of methods (selection of *subset of SNP* and use of *derivate variable*) is that the methods that preselect a subset of SNP filter data on criteria that involve the association with the phenotype, whilst the multivariate techniques like PCA condense all the marker information into few derivate variables that are independent from the phenotypes used. Thus, PCA could be considered as trait independent technique. The PLSR method use simultaneously the information both on markers and phenotype to extract latent variable and may not be considered trait independent.

Effect of prior information of estimation of QTL effects

The estimation of the SNP effects could be carried out following several approaches, but a general feature is the assumption about the proportion variance explained by each chromosomal segment. To provide the ideal estimation of the SNP effects, such assumption should take into account the number of QTL underlying the trait and the prior distribution of QTL effects. However the number, the size of the QTL and the distribution of the effects are trait dependent and the number of detected QTL is function of the power of QTL mapping experiment (WELLER *et al.* 1990). Thus, not all the QTL have been discovered, but just those ones of biggest effect. HAYES and GODDARD (2001) estimated a number between 50 and 100 QTL affecting a generic quantitative trait using a meta-analysis approach. The predicted distributions of QTL were consistent with the hypothesis of many genes of small effect and few of large effect. They figures agree with the results of CHAMBERLAIN *et al.* (2007), who analyzed the results of a single experiment of QTL mapping and found that at least 30 QTL were likely to be segregating in the Holstein population examined for all core production traits. The distribution of QTL effects in some cases reflects the aforementioned pattern and clear examples have been reported in literature. In particular two polymorphism, *K232A* in DGAT1 gene and *F279Y* in GHR identified on bovine chromosome 14 and chromosome 20 respectively, explained about 50%

and 10% of trait variance of milk fat content and protein content respectively (GRISART *et al.* 2002; BLOTT *et al.* 2003). These two polymorphisms had a large effect also on the other production and quality trait. In particular, DGAT1 which explained such large amount of the trait variance seem to confirm the theoretical distribution of many genes of small effect and few genes of large effect. However, there are some traits for which this distribution might be not suitable and normally distributed QTL effects might be hypothesized (CALUS 2010)

If reliable prior knowledge of distribution of QTL effect is considered in modeling SNP effects this would guarantee a better estimation of SNP effect and a reduced bias. However, different options may be chosen when modeling the SNP data. If it is not assumed any shrinkage factor to the variance and the SNP genotype is treated as fixed effect, the ordinary least square fixed regression (LS-FR) approach is applied. Equal variance contribution of each chromosomal segments to the genetic variance may be assumed using for instance a Best Linear Unbiased Prediction (R-BLUP) approach, treating the SNP as random factors. Different variance shrinkage factors lambda have been proposed (λ). The alternative is to allow variance of each chromosomal segment to vary and to estimate SNP variances and effects simultaneously using a bayesian approach (BAYES) implemented through a Gibbs sampler algorithm drawing sample from known density distribution.

Ordinary Least Square fixed regression (LS-FR) for estimation of SNP effects

The simplest approach for estimating SNP effects makes no assumption regarding their distribution. The GS using LS-FR is a two step procedure. In the first step a subset of SNP are selected on the basis of their significant association with the phenotype according to the model

Where is the general mean, $\mathbf{1_n}$ is a vector of one whose dimension is the number *n* of records, $\boldsymbol{Q_i}$ is a incidence matrix that allocate the *i*-th SNP genotype to the phenotypic records, $\boldsymbol{g_i}$ is the vector of effects for the *i*-th SNP and $\boldsymbol{e}$ is the random residual. With such model each SNP is tested at once and a threshold is established to assess whether a SNP is significantly associated to the phenotype. In the second step the phenotypes are regressed on the selected SNP genotypes using a multiple linear regression. The SNP effects are estimate simultaneously only for the *m* selected SNP in the previous step, following the model:

In this case all the other SNP are not considered and set to zero. This fact leads to an overestimation of SNP effects Its magnitude is a function of the number of SNP retained and thus depends on the choice of appropriate significance threshold in the first step. Different options could be adopted like Bonferroni correction, permutation or false discovery rate (FDR) to deal with the issue of multiple testing, that arise when many null hypotheses are tested at the same time. In this case the theoretical alpha usually adopted to control the type error I is not appropriate as one hypothesis was tested instead. Some significant results are likely to be expected by chance even all the hypotheses are false (BALDING 2006). The threshold should be quite stringent otherwise there would be a degree of freedom problem as the number of predictor would be highly larger than phenotypic records and LS-FR could not be applicable. An additional drawback of such an approach is that only the larger effects are captured and consequently not all the genetic variance is explained by the marker (GODDARD and HAYES 2007). All these features make LS-FR estimation an unsuitable tool to estimate SNP effects as confirmed by application of LS-FR methods both in simulated and real datasets. Low accuracies of DGV predictions were obtained compared to other methods (accuracy of DGV ranged from 0.31 to 0.36 in MEUWISSEN *et al.* (2001) and from 0.26 to 0.55 in MOSER *et al.* (2009) where in both case BLUP and Bayesian method performed better).

R-BLUP approach and simultaneous estimation of whole SNP effects

A better alternative to LS-FR was proposed to overcome the issue of overestimation and bias of SNP effects. A feasible solution is to fit a model that assumes an equal contribution of each SNP to the genetic variance of the trait. If the QTL effects are drawn from a normal distribution with constant variance across the chromosomal segment, the estimates are BLUP and SNP effects may be estimated simultaneously (GODDARD and HAYES 2007). The model is:

in this case the all the *m* SNP are treated as random effects. **X** is the incidence matrix of a set of fixed effects **b**, ***Z*** (*n* individuals x *m* markers) is the incidence matrix that allocates SNP genotype to phenotypic record, ***g*** is the vector of random SNP effects and ***e*** is the vector of random residuals. The solution of the mixed model equation is

Covariance matrices of random effects (**G**) and residuals (**R**) may be modeled in different ways. In the simplest case no interaction is considered between loci, *i.e.* **G** and **R** are diagonal matrix, and equal contribution to the genetic variance as diagonal, where λ may assume different values.
If $\lambda=\sigma_e^2/\sigma_g^2$ as suggested by MEUWISSEN *et al.* (2001) σ_g^2 is estimated from the total additive genetic variance divided by the number of SNP fitted *i.e.* ——. In R-BLUP all the random effects have a common variance and the SNP with largest variance tend to be overestimated reducing the accuracy of prediction. Despite the overestimation of SNP effects, this methods lead to a better prediction than LS-FR approach. When the assumption of equal variance of each segment sampled from normal

distribution is valid the R-BLUP performances are similar or better than other methods that allow variance to vary between SNP (Calus 2010).

If the marker effects are normally distributed with constant variance, BLUP is useful tool to estimate effects of the markers and hence the DGV. Goddard (2009) demonstrated the R-BLUP model is equivalent to a conventional animal model in which the additive relationship matrix calculated from pedigree is replaced by the genomic relationship matrix (**G**) calculated from marker data, termed as Genomic BLUP (G-BLUP). The additive relationship matrix measure the expected fraction of alleles shared by the individuals of a population based on pedigree, whilst the genomic relationship matrix measure the actual fraction of alleles shared. If the classical animal model is considered, where ***Z*** is the incidence matrix that allocate the animal to the phenotypic records, and ***u*** is the vector o polygenic effects for all the animal in the population, the solution of such model is:

where $\mathbf{A}^{-1}$ is the inverse of additive relationship matrix, and $\lambda=\sigma_e^2/\sigma_u^2$.

In the G-BLUP model **G** replaces the **A** matrix. According to is the genomic relationship matrix can be calculated as where , **P** contains the allelic frequencies of the marker expressed as $2(p_i - 0.5)$., **M** denotes the matrix that specifies which marker alleles each individual inherited. If in **M** -1 (for the homozygote), 0 (for the heterozygote), 1 (for the other homozygote) parameterization is adopted, then the diagonals elements of **MM'** matrix count how many homozygous loci for each individual, and off-diagonals the number of alleles shared by relatives. The division by scales **G** to be comparable to the numerator relationship matrix **A** (With this formulation the mixed model equations become:

Where ***u*** in this case is the DGV and is equivalent to the DGV calculated estimating the marker effects using G-BLUP and summing up the effect for all chromosome segment. Furthermore, Hayes and Goddard (2008) demonstrated that the hereditability of a quantitative trait could be accurately estimated using a large number of markers (9,000 markers) to build the genomic relationship matrix instead of pedigree based relationship matrix in a simulation study. The additive variance component estimated with 5 generation of pedigree was not as much accurate as that one estimated using the whole set of markers (table 1)

Table 1. True and estimated heritability from marker based and from pedigree (Hayes and Goddard 2008)

Method	Values	SE
True heritability	0.33	-
1 generation of pedigree	0.22	0.04
5 generation of pedigree	0.26	0.03
Genomic relationship matrix (1000 markers)	0.21	0.03
Genomic relationship matrix (5000 markers)	0.30	0.03
Genomic relationship matrix (9000 markers)	0.32	0.03

If the number of markers used to estimate DGV is much larger than phenotypic records, the approach of using Genomic relationship matrix in BLUP animal model is more convenient than estimate the single marker effects, since much less number of parameters need to be estimated (Hayes *et al.* 2009b). A problem could be represented by the inversion of the **G** matrix, which in some cases may be singular. Efficient algorithms for solving the model have been proposed when large dataset are used (Legarra and Misztal 2008; VanRaden 2008).

Factor affecting the accuracy of genomic prediction

The assessment of goodness of genomic predictions is generally carried out by measuring the correlation between the DGV and the true breeding value (TBV) . Since the TBV is available only for simulated data, its expectation may be used when the evaluation is carried out on real data. The EBV (weighted or not to its reliability) is assumed as golden standard, and Pearson correlation coefficient are calculated ( ). Alternative evaluation may be done on the squared correlation ( )(VanRaden *et al.* 2009). An additional criterion to evaluate the genomic prediction is generally the bias of prediction measured by the regression coefficient $b_{EBV,DGV}$ between phenotype and DGV.

The characteristics of the reference population affect heavily the accuracy of genomic prediction. Number of animals in the reference population (Meuwissen *et al.* 2001; Muir 2007; Hayes *et al.* 2009b), number of markers and the level of LD (Calus *et al.* 2008; Solberg *et al.* 2008), heritability of the trait considered (Meuwissen *et al.* 2001; Kolbehdari *et al.* 2007) are the main factors affecting the accuracy of DGV. Moreover, additive genetic relationships in the reference population captured by the SNP affects the accuracy of genomic predictions both in simulated and real data (Habier *et al.* 2007; Habier *et al.* 2010).The choice of the statistical model and its parameterization (single markers or haplotypes) affect the accuracy of prediction as well (Calus *et al.* 2008).

Different theoretical formulation have been proposed for calculating the accuracy of DGV in animal without phenotypic record. DAETWYLER *et al.* (2008) proposed a formula to predict the accuracy of genomic prediction for animal without phenotype:

Where h^2 is the observed heritability, λ is the number of phenotypes per number of QTL loci. This equation allows to summarize the relationships between some factors affecting the accuracy as shown in figure 5, where the heritability was fixed at 0.3, and when the accuracy was evaluated in function of the heritability (figure 6)

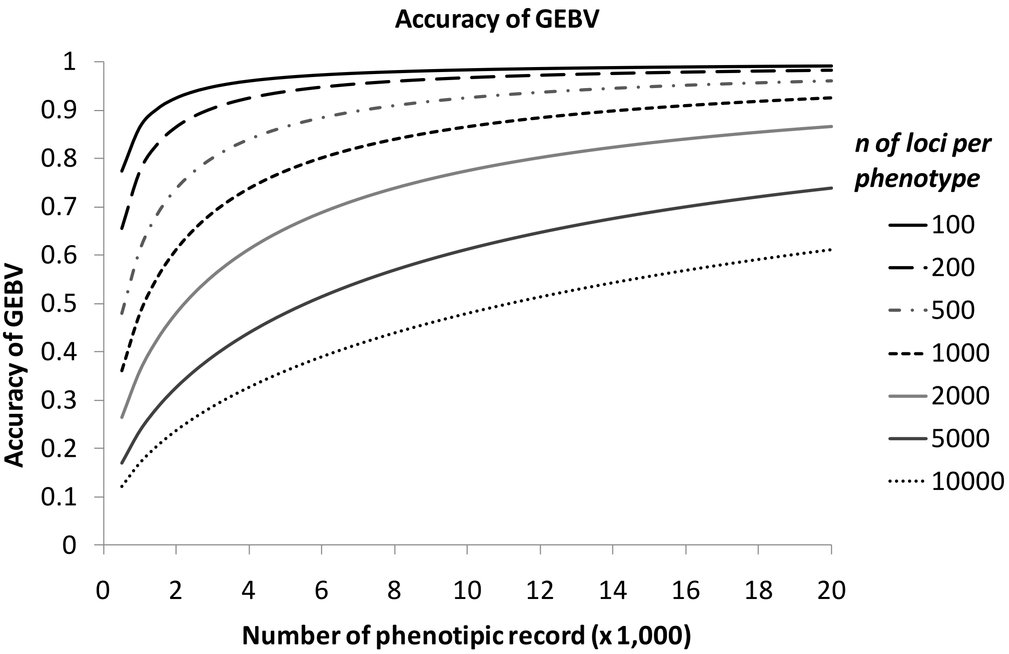


Figure 5. Predicted accuracy of DGV in function of number of phenotype per number of marker loci (heritability 0.3 and λ=10,5,2,1,0.5,0.2,0.1) calculated according to the formula of (DAETWYLER *et al.* 2008)

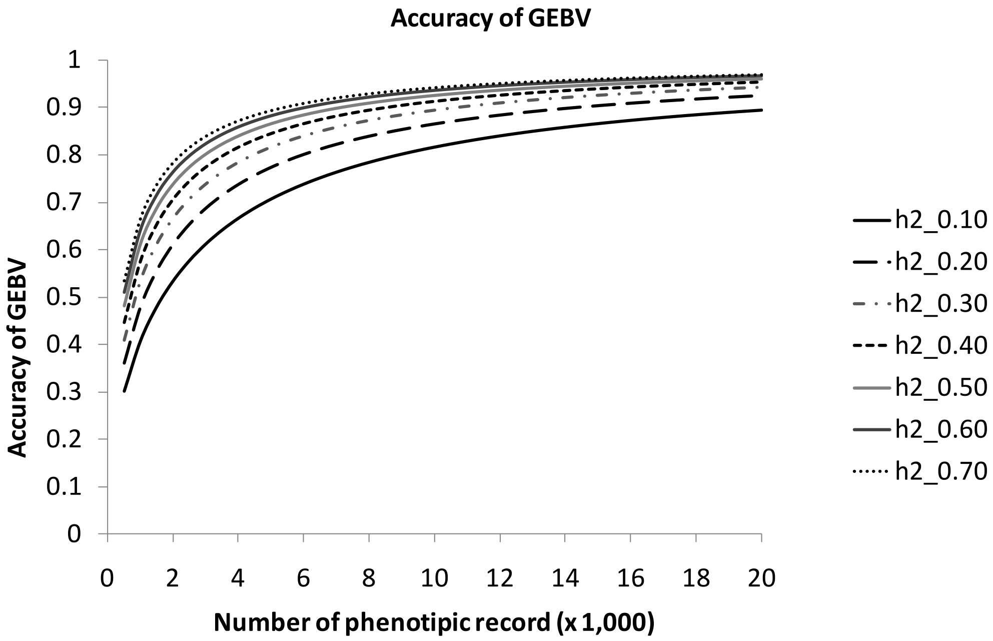


Figure 6 Predicted accuracy of DGV in function of number of phenotype and heritability (1000 loci x phenotype).

HAYES *et al.* (2009b) showed also similar relationships between accuracy of genomic predictions and number of phenotypic records in the reference population. In particular, for a quantitative trait of heritability of 0.3 and effective population size (Ne) of 100, about 12,500 individuals in the reference are needed to predict DGVs of un-phenotyped individuals with an accuracy of 0.7. If only 5,000 individuals are available a drop in the accuracy (0.5) with the same heritability is observed. A similar conclusion may be drawn using a different analytical approach, with the formula of DAETWYLER *et al.* (2008) for instance. By calculating the number of phenotypic records required to achieve the same values of accuracy showed by HAYES *et al.* (2009b) it is possible to observe similar patterns (figure 7).

Summarizing, the accuracy of DGV increases with of heritability and number of markers. Furthermore, the higher the number of animals with both genotypes and phenotypes the higher DGV accuracy in the prediction population.

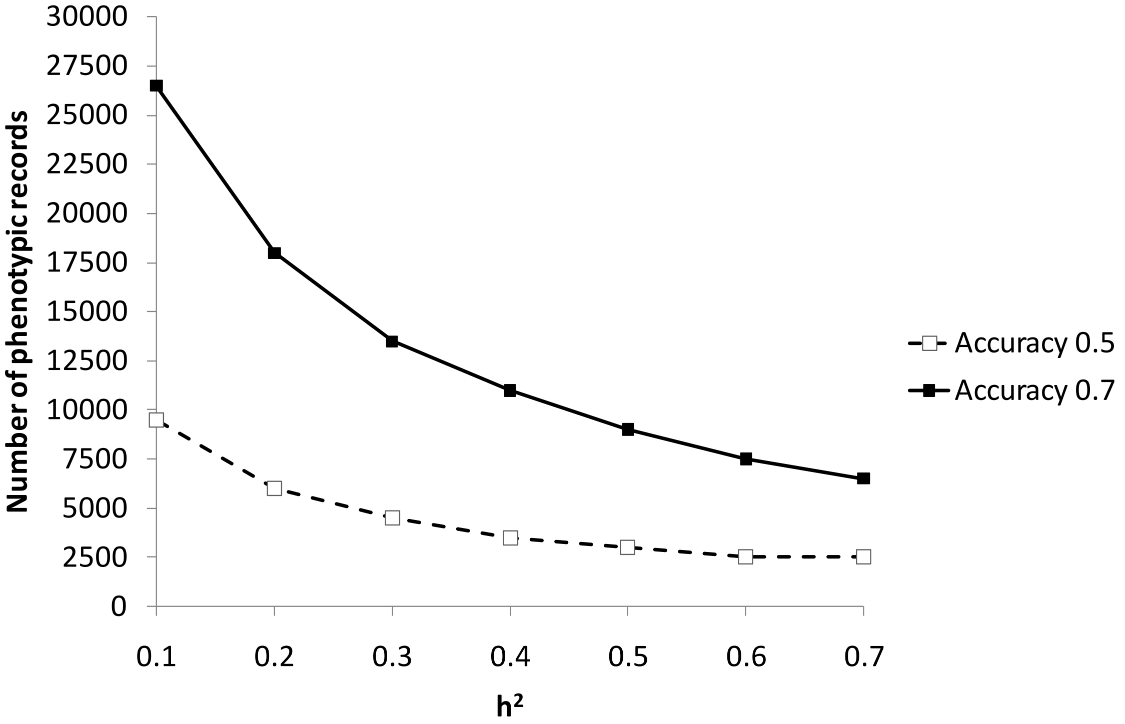


Figure 7. Relationship between number of phenotypic record in the reference population required to get two level of accuracy (0.5, 0.7) in function of the heritability calculated according to (DAETWYLER *et al.* 2008) setting 5,000 loci per phenotypic records.

1.3.3 Accuracy of Genomic Prediction: results from simulated dataset

Simulations of genomic selection programmes have been largely used to assess the potential of this tool for genetic evaluation of farm animals, before real data were available. Since 2001 (MEUWISSEN *et al.* 2001) many simulations have been carried out in literature both to propose suitable models for genomic evaluation and also to seek the main factors affecting the accuracy of genomic prediction using dense markers. Although strong prior assumptions affect simulation results, they have been useful to develop statistical model and propose solution to main issues of GS.

Table 2 reports values of DGV accuracies in juvenile animals (without phenotypic records) are reported across different studies. Not all the predictions are comparable because of different assumptions in simulation, as different number of markers and individuals in the simulated population. Main differences concern: the underlying genetic model with effect of QTL purely additive (CALUS *et al.* 2008) or simulation of dominance or epistasis (GIANOLA *et al.* 2006; GIANOLA and VAN KAAM 2008); the distribution of QTL effects and marker frequency; type of marker used; the method used to generate the LD (MEUWISSEN *et al.* 2001; MUIR 2007; KOLBEHDARI *et al.* 2007); and number of generations of random mating (eventually the number of generation of selection performed). The results reviewed are grouped on the basis of the methods used to estimate markers effects.

In general, results from simulated datasets (table 2) are quite in agreement as far as the accuracy of DGV in function of the statistical method is concerned. Bayesian methods ($r_{\mathrm{DGV,TBV}}$ 0.380-0.848, with several values above 0.70) perform better than BLUP ($r_{\mathrm{DGV,TBV}}$ ranging from 0.410 to 0.749) or LS-FR methods (0.124-0.610). Semi-parametric methods like RHKS gave results similar or better than Bayesian approaches. Multivariate techniques of SNP reduction perform similarly to BLUP approach (0.604-0.730).

Basically, the increase of density of SNP markers results in a better accuracy of prediction (MEUWISSEN *et al.* 2001; MUIR 2007; SOLBERG *et al.* 2009). Simulated results showed how the heritability of the trait affect positively the estimation accuracy (CALUS and VEERKAMP 2007; KOLBEHDARI *et al.* 2007; MUIR 2007) as confirmed also by theoretical expectations (DAETWYLER *et al.* 2008; HAYES *et al.* 2009b). Furthermore, as shown by CALUS and VEERKAMP (2007), the inclusion of polygenic effect in the estimation has just marginal positive effects on DGV accuracy both for high and low heritability traits. The inclusion of polygenic effect at low level of LD (r^2 <0.10) gave positive effect only for high heritability traits, confirming the effect of marker density on accuracy of DGV. The effect of the inclusion of an increasing number of individuals in the reference population results in a better genomic prediction as shown by several authors (MEUWISSEN *et al.* 2001; MUIR 2007; SOLBERG *et al.* 2008). These results found their theoretical justification in the reduction of the statistical asymmetry of data matrix due to the increased sample size. A further simulation carried

Table 2. Range of accuracies of genomic prediction across different method in simulated dataset.

Method	Accuracy	Authors
LS-FR	0.124-0.318 (0.318-0.363) [b]	(MEUWISSEN *et al.* 2001)
	0.49-0.61 [a]	(HABIER *et al.* 2007)
BLUP	0.579-0.732 (0.732-0.668)[b]	(MEUWISSEN *et al.* 2001)
	0.64-0.42[a]	(HABIER *et al.* 2007)
		(MUIR 2007)
	0.62-0.79 (0.60-0.75)[c]	(KOLBEHDARI *et al.* 2007)
	0.499-0.611(0.588-0.630)[d]	(PIMENTEL *et al.* 2009)
	0.76(0.41)[e]	(MACCIOTTA *et al.* 2010)
	[k]	(GUO *et al.* 2010)
BAYES A	0.798	(MEUWISSEN *et al.* 2001)
COMMON PRIOR	0.428-0.616(0.719-0.808)[k]	(GUO *et al.* 2010)
	0.424-0.615(0.711-0.805)[l]	(GUO *et al.* 2010)
BAYES B	0.708-0.848(0.848-0.737)[b]	(MEUWISSEN *et al.* 2001)
	0.690-0.860(0.626-0.827) [f]	(SOLBERG *et al.* 2008)
	0.802-0.821 (0.764-0.798) [g]	(SOLBERG *et al.* 2008)
	0.55-0.69[a]	(HABIER *et al.* 2007)
	0.38-0.55(0.36-0.55) [h]	(CALUS *et al.* 2008)
	0.73-0.79(0.74-0.80) [i]	(CALUS and VEERKAMP 2007)
MIXTURE PRIOR	0.474-0.657(0.745-0.829)[k]	(GUO *et al.* 2010)
	0.454-0.645(0.733-0.826)[l]	(GUO *et al.* 2010)
PCA-BLUP	0.604-0.665[g]	(SOLBERG *et al.* 2009)
	0.70-0.55 (0.73-0.56) [l]	(MACCIOTTA *et al.* 2010)
PLSR-BLUP	0.611-0.681[g]	(SOLBERG *et al.* 2009)
MLR (RKHS)	0.59(0.95) [j]	(GIANOLA *et al.* 2006)

[a] range of DGV accuracy of prediction population when for different time point far from reference population
[b] range of DGV accuracy for increasing phenotype records from 500 to 2200 and Ne=100 or (decreasing marker density: spaced from 1 up to 4 cM, Ne=100)
[c] range of DGV accuracy for h^2=0.5 scenario unequal QTL size and even or random distribution on genome (for $h^2$0.05 scenario unequal QTL size and even or random distribution on genome)
[d]DGV accuracy obtained applying to different BLUP method (two different Ridge Regression methods)
[e]DGV accuracy when considering the phenotype as response variable (or polygenic EBV)
[f] range of DGV accuracy at increasing density of marker loci using SNP genotype or (microsatellite)
[g]range of accuracy at increasing density from haplotype of SNP or (microsatellite)
[h]range of accuracy for DGV when h^2=0.05 or (h2=0.5)
[i]range of accuracy when the contribution of polygenic effect is not considered or (consiedered) at decreasing values of h^2
[j]accuracy of DGV for Multiple Linear Regression (MLR)Mixed Model and (RKHS regression)
[l]accuracy of DGV using raw phenotype as response variable and equal variance of each PC eigenvalues as prior variance in the mixed model equation (polygenic EBV as response variable)
[k]accuracy (squared correlation R^2)of DGV using EBV as response variable: from 30 to 100 daughter x bulls h^2=0.05 (0.30)
[l]accuracy (squared correlation R^2)of DGV using DYD as response variable: from 30 to 100 daughter x bulls h^2=0.05 (0.30)

out by GUO *et al.* (2010) showed how the influence of response variable (DYD or polygenic EBV) on the accuracy of DGV is method-dependent and of moderate effect (0.3 to 3.6% of difference in accuracy). The methods that reduce the number of SNP on the basis or their contribution to the variance (or summarize the information of the whole set of marker with few variable) gave the best results of DGV accuracy and less biased estimates.

1.3.4 Accuracy of Genomic Prediction: Results from field data

The main limitations of application of genomic selection on real data are the costs for genotyping the animals. In few years the cost of genotyping has drop dramatically and therefore has become feasible to genotype a high number of animal. The number of genotyped animal in US (cows included) in 2010 was around 33,434 (VANRADEN AND TOOKER, 2010). Table 3 reports the level of DGV accuracy achieved using real field data of dairy cattle across different methods and traits. The genotyped animals reported in the table 3 are all bulls and ranged from 479 to 5,535. Different methods have generally been adopted to build the reference and prediction set. In most of cases, older animals are used as training to predict the younger candidate to selection (VANRADEN *et al.* 2009). When the number of animals is too small *n*-fold cross validation are carried out to construct the reference set, leaving out for n time a certain percentage of animal and using the rest as training set to predict the genomic breeding values of former animals. Better accuracies are obtained when animals are chosen randomly (LUAN *et al.* 2009) to make the reference set, even if this not suitable for practical implementation in a genomic breeding program. Another way to build the reference set was carried out by HABIER *et al.* (2010) and put some constraints of relatedness among animal of training and prediction to evaluate the effects of additive relationship on DGV accuracy.

The average number of SNPs used were around 35,000 and about 19,000 for 54 k illumina bead chip (http://ww.illumina.net) and *Affymetrix* panel (http://www.affymetrix.com/) respectively, depending on data quality control, editing of SNP data and missing genotypes. The response variables used to estimate the SNP effects were both national EBV, de-regressed proof or DYD.

Differences in accuracy of DGV seem to rely on the number of the animals and relatedness among animals of reference and prediction, more than statistical methods. The range of DGV accuracies reported across different trait did not show how the estimation of SNP is somehow trait by method dependent. The accuracies obtained with LS-FR are the lowest (0.43-0.53), if the high number of phenotypic records in this study are considered. Furthermore the LS-FR works better when the number of predictors are smaller like in the case of MOSER *et al.* (2009). The prediction that used G-BLUP ranged from 0.153 up to 0.74 ($R^2 = 0.55$ in VANRADEN *et al.* (2009)). Bayesian methods gave values of accuracy from 0.128 to 0.790 ($R^2 = 0.63$ for fat percentage in VANRADEN *et al.* (2009)). Finally, the methods like PCA-BLUP or PLSR and RKSH perform similar to G-BLUP. The lowest figures for PCA-BLUP are likely to be a consequence of limited sample size. Considering that PLSR and PCA-BLUP use a limited number of derivate variables (reducing the original variable up to a factor 20) and allows to save much more computational time (MACCIOTTA *et al.* 2010; SOLBERG *et al.* 2009)

Table 3. Accuracies of genomic prediction across different methods and traits in real data of dairy cattle

Method	n (reference)1)	n SNP2)	Accuracy3)	Authors
LS-FR	1,945(1,239)	7,237	0.430-0.530[a]	(MOSER *et al.* 2009)
G-BLUP	5335 (3,576)	38,416	0.21-0.55[b]	(VANRADEN *et al.* 2009)
	500(400)	18,991	0.153-0.617(0.195-0.609)[c]	(LUAN *et al.* 2009)
	1181(781)	39,048	0.490-0.620	(HAYES *et al.* 2009a)
	1545(1068)		(0.450-0.620)[d]	
	1,945(1,239)	7,372	0.560-0.710[a]	(MOSER *et al.* 2009)
	5,212 (~)	42,302	0.481-0.572(0.510-0.603)[e]	(HARRIS and JOHNSON 2010)
	3863(~2096)	40,588	0.440-0.680(0.170-0.380)[f]	(HABIER *et al.* 2010)
BAYES				
MIXTURE	500 (400 cv)	18,991	0.128-0.601(0.192-0.612)[c]	(LUAN *et al.* 2009)
BAYES A	781(1,068)	39,048	0.470-0.710(0.470-0.690)[d]	(HAYES *et al.* 2009a)
	1,945(1,239)	7,372	0.560-0.710[a]	(MOSER *et al.* 2009)
BAYES SSVS	781(1,068)	39,048	0.470-0.700(0.405-0.700) [d]	(HAYES *et al.* 2009a)
BAYES LASSO	4,703(3,305)	32,518	0.612(0.428-0.567†;	(WEIGEL *et al.* 2009)
			0253-0.539‡)[h]	
NONLINEAR	3,576 (1,759)	38,416	0.190-0.630[b]	(VANRADEN *et al.* 2009)
BAYES B	500(400 cv)	18,991	0.130-0.607(0.189-0.601)[c]	(LUAN *et al.* 2009)
	3863(~2096)	40,588	0.500-0.680(0.290-0.470)[f]	(HABIER *et al.* 2010)
PCA-BLUP	863 (749)[479]	40,658	0.210-0.61[g]	MACCIOTTA et al, 2010
		(37,254)	(0.180-0.540)	
		[40,179]	[0.280-0.460]	
PLSR-BLUP	1,945(1,239)	7372	0.550-0.700 [a]	(MOSER *et al.* 2009)
SVR	1,945(1,239)	7372	0.580-0.720 [a]	(MOSER *et al.* 2009)

n=number of animal in the whole dataset (and in the reference population only)
1) number of SNP after editing procedure (3 chip set 54 k 25 k 9 k were used)
2) minimum and maximum DGV accuracies across productive and functional traits and different studies and methods
[a] range of DGV accuracy of prediction population for Australian economic index
[b] accuracy were expressed as $R^2$and the range is across production and functional trait
[c] range of accuracy for milk production trait estimated using 5 fold cross validation for cohort of animal whose phenotypes were masked on the basis of year of progeny test or (5 fold cross validation of random animal) to design the reference and prediction population
[d]range of DGV accuracy in Australian Holstein (Holstein +Jersey) population in the reference set with a multi-breed.
[e]range of DGV accuracy in NZ Hostein Holstein and NZ Jersey both not blending the DGV with Parent Average information and (using a blending approach)
[f] minimum and maximum of DGV accuracy for different constrain of additive relationship when building the reference set (DGV due to LD) for milk yield fat yield, protein yield and SCS in German Holstein
[g] range of DGV accuracy for Italian Holstein(Italian Brown Swiss) and [Italian Simenthal] building the reference set sorting the bulls by year of birth and using 2,564, 2,257, and 2,476 PC respectively.
[h] values of DGV accuracy using whole set of SNPs or (range of accuracy when selecting smaller subsets of SNPs of largest effect†, or evenly spaced in the genome ‡)

As far as the statistical method adopted is concerned, the Bayesian approach generally outperforms the other methods on simulated dataset. Especially Bayes B, which heavily relies on the prior assumption of distribution of QTL effect and gave similar results to methods implemented calculating the G-matrix in an animal model.

The NONLINEAR model (VANRADEN *et al.* 2009) (equivalent of Bayes B of MEUWISSEN *et al.* (2001)) performed quite good, even if the differences between G-BLUP and NONLINEAR approaches are not so large.

Some of the Bayesian methods showed an example of trait-model interaction. Interestingly, G-BLUP perform better than Bayesian approaches for some traits. Accuracy of DGV for milk yield in HAYES *et al.* (2009a) was better using G-BLUP approach, rather than Bayesian approach. This result is probably due to the genetic determinism of the trait where the number and distribution of gene underlying these traits approaches the normal distribution. Conversely, the genomic prediction for fat percentage (where DGAT1 explained 50% of the genetic variance) is more accurate using Bayesian approach, and in particular Bayes B (where the prior distribution reflect the real distribution of QTL effect for fat percentage).

Additional factors that affect the genomic prediction in real data are the number of SNP used in the prediction and the way to chose them (WEIGEL *et al.* 2009)

The large amount of QTL detected in several breeds of cattle should have led through the use of QTL information to marker assisted selection of farm animals. The aim to dissect the genetic architecture of quantitative traits through QTL mapping has not been fully exploited for several reasons. The imprecise location of detected QTL has permitted to use only part of this information in marker assisted selection schemes. Moreover designs to detect QTL, especially in dairy cattle, are tightly linked to the family structure of the population, and with such imprecise location of QTL only within family selection is a possible application. The refinement of the position of QTL detected might allow to use this large amount of data yielded in the past decade.

However thanks to the development of dense genomic map and the SNP chip technology, much more data than in the past have been produced, in few years. The use of SNP data should allow to overcome the problem of within family selection (the marker is supposed to be in LD with QTL) and the cost of genotyping is going down steeply. However new problems and challenges came up. In particular how to fully exploit thousands of markers in QTL detection and genomic selection.

There are several factors that affect the accuracy of genomic predictions, and the model used to breeding values estimation is one of the most relevant (at least in simulated studies). In real data, the small population size (excluding the US and Canada situation) that characterizes the European situation does not allow to reach the minimum number of animal needed to get an accurate genomic predictions in comparison to simulation studies. To overcome the problem of large number of predictor vs. number of animals, different variable selection techniques have been proposed. In the present thesis both the factor affecting the accuracy of genomic prediction and the reduction of number of predictor have been developed and compared to recent literature.

Objective of the Thesis

The overall objective of the present thesis was to investigate on the use of genetic markers in the marker assisted selection of farm animal. In particular, it has been investigate the way to better exploit the large amount of QTL data yielded in the recent literature through meta-analysis (chapter 2). Furthermore, the use of genomic marker to breeding values estimation, the study of the factor affecting the accuracy of prediction (chapter 3) and the use of multivariate techniques to reduce the number of predictors in genomic selection (chapter 4) have been studied.

REFERENCES

ANDERSSON, L., and M. GEORGES, 2004 Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet **5:** 202-212.

ASHWELL, M. S., D. W. HEYEN, T. S. SONSTEGARD, C. P. VAN TASSELL, Y. DA et al., 2004 Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. Journal of Dairy Science **87:** 468-475.

AULCHENKO, Y. S., S. RIPKE, A. ISAACS and C. M. VAN DUIJN, 2007 GenABEL: an R library for genome-wide association analysis. Bioinformatics **23:** 1294-1296.

BAGNATO, A., F. SCHIAVINI, A. ROSSONI, C. MALTECCA, M. DOLEZAL et al., 2008 Quantitative trait loci affecting milk yield and protein percentage in a three-country Brown Swiss population. Journal of Dairy Science **91:** 767-783.

BALDING, D. J., 2006 A tutorial on statistical methods for population association studies. Nat Rev Genet **7:** 781-791.

BENNEWITZ, J., N. REINSCH, F. REINHARDT, Z. LIU and E. KALM, 2004 Top down preselection using marker assisted estimates of breeding values in dairy cattle. Journal of Animal Breeding and Genetics **121:** 307-318.

BLOTT, S., J. J. KIM, S. MOISIO, A. SCHMIDT-KUNTZEL, A. CORNET et al., 2003 Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics **163:** 253-266.

CALUS, M. P. L., 2010 Genomic breeding value prediction: methods and procedures. Animal **4:** 157-164.

CALUS, M. P. L., T. H. E. MEUWISSEN, A. P. W. DE ROOS and R. F. VEERKAMP, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics **178:** 553-561.

CALUS, M. P. L., and R. F. VEERKAMP, 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics **124:** 362-368.

CARDON, L. R., and J. I. BELL, 2001 Association study designs for complex diseases. Nat Rev Genet **2:** 91-99.

CHAMBERLAIN, A. J., H. C. MCPARTLAN and M. E. GODDARD, 2007 The number of loci that affect milk production traits in dairy cattle. Genetics **177:** 1117-1123.

COLE, J. B., P. M. VANRADEN, J. R. O'CONNELL, C. P. VAN TASSELL, T. S. SONSTEGARD et al., 2009 Distribution and Location of Genetic effects for Dairy traits (vol 92, pg 2931, 2009). Journal of Dairy Science **92:** 3542-3542.

DAETWYLER, H. D., B. VILLANUEVA, P. BIJMA and J. A. WOOLLIAMS, 2007 Inbreeding in genome-wide selection. Journal of Animal Breeding and Genetics **124:** 369-376.

DAETWYLER, H. D., B. VILLANUEVA and J. A. WOOLLIAMS, 2008 Accuracy of predicting the genetic risk of disease using a genome-wide approach. Plos One **3:** e3395.

DARVASI, A., and M. SOLLER, 1992 Selective Genotyping for Determination of Linkage between a Marker Locus and a Quantitative Trait Locus. Theoretical and Applied Genetics **85:** 353-359.

DARVASI, A., and M. SOLLER, 1994 Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics **138:** 1365-1373.

DE KONING, D. J., 2006 Conflicting candidates for cattle QTLs. Trends Genet **22:** 301-305.

DE ROOS, A. P. W., C. SCHROOTEN, E. MULLAART, M. P. L. CALUS and R. F. VEERKAMP, 2007 Breeding value estimation for fat percentage using dense markers on Bos taurus autosome 14. Journal of Dairy Science **90:** 4821-4829.

DEKKERS, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim Sci. **82:** E313-328.

DEKKERS, J. C. M., and F. HOSPITAL, 2002 The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics **3:** 22-32.

DOERGE, R. W., 2002 Mapping and analysis of quantitative trait loci in experimental populations. Nature Reviews Genetics **3:** 43-52.

EWING, B., and P. GREEN, 2000 Analysis of expressed sequence tags indicates 35,000 human genes. Nature Genetics **25:** 232-234.

FALCONER, MACKAY, T. 1996, Introduction to Quantititative Genetics. Longman, New York, 1996

FERNANDO, R., and M. GROSSMAN, 1989 Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution **21:** 467 - 477.

GEORGES, M., D. NIELSEN, M. MACKINNON, A. MISHRA, R. OKIMOTO et al., 1995 Mapping Quantitative Trait Loci Controlling Milk-Production in Dairy-Cattle by Exploiting Progeny Testing. Genetics **139:** 907-920.

GIANOLA, D., G. DE LOS CAMPOS, W. G. HILL, E. MANFREDI and R. FERNANDO, 2009 Additive Genetic Variability and the Bayesian Alphabet. Genetics **183:** 347-363.

GIANOLA, D., R. L. FERNANDO and A. STELLA, 2006 Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. Genetics **173:** 1761-1776.

GIANOLA, D., and J. B. C. H. M. VAN KAAM, 2008 Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics **178:** 2289-2303.

GODDARD, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136:** 245-257.

GODDARD, M. E., and B. J. HAYES, 2007 Genomic selection. Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie **124:** 323-330.

GONZALEZ-RECIO, O., D. GIANOLA, N. LONG, K. A. WEIGEL, G. J. M. ROSA et al., 2008 Nonparametric Methods for Incorporating Genomic Information Into Genetic Evaluations: An Application to Mortality in Broilers. Genetics **178:** 2305-2313.

GONZALEZ-RECIO, O., D. GIANOLA, G. J. M. ROSA, K. A. WEIGEL and A. KRANIS, 2009 Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Genetics Selection Evolution **41:** -.

GRISART, B., W. COPPIETERS, F. FARNIR, L. KARIM, C. FORD et al., 2002 Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res **12:** 222-231.

GRISART, B., F. FARNIR, L. KARIM, N. CAMBISANO, J. J. KIM et al., 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc Natl Acad Sci U S A **101:** 2398-2403.

GUILLAUME, F., S. FRITZ, D. BOICHARD and T. DRUET, 2008a Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. Genet Sel Evol **40:** 91-102.

GUILLAUME, F., S. FRITZ, D. BOICHARD and T. DRUET, 2008b Short Communication: Correlations of Marker-Assisted Breeding Values with Progeny-Test Breeding Values for Eight Hundred Ninety-Nine French Holstein Bulls. J. Dairy Sci. **91:** 2520-2522.

GUO, G., M. LUND, Y. ZHANG and G. SU, 2010 Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. Journal of Animal Breeding and Genetics**:** no-no.

GUYON, I., ANDR\, \#233 and ELISSEEFF, 2003 An introduction to variable and feature selection. J. Mach. Learn. Res. **3:** 1157-1182.

HALDANE, J. B. S. 1919.The combination of linkage values, and calculation of distance between the loci of linked factor. J Genet. 299-309

HABIER, D., R. L. FERNANDO and J. C. M. DEKKERS, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics **177:** 2389-2397.

HABIER, D., R. L. FERNANDO and J. C. M. DEKKERS, 2009 Genomic Selection Using Low-Density Marker Panels. Genetics **182:** 343-353.

HABIER, D., J. TETENS, F. R. SEEFRIED, P. LICHTNER and G. THALLER, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution **42**.

HARRIS, B. L., and D. L. JOHNSON, 2010 Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. Journal of Dairy Science **93:** 1243-1252.

HAYES, B., and M. E. GODDARD, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution **33:** 209-229.

HAYES, B. J., P. J. BOWMAN, A. C. CHAMBERLAIN, K. VERBYLA and M. E. GODDARD, 2009a Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution **41**.

HAYES, B. J., and M. E. GODDARD, 2008 Technical note: Prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science **86:** 2089-2092.

HAYES, B. J., P. M. VISSCHER and M. E. GODDARD, 2009b Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). Genetics Research **91:** 143-143.

HEYEN, D. W., J. I. WELLER, M. RON, M. BAND, J. E. BEEVER et al., 1999 A genome scan for QTL influencing milk production and health traits in dairy cattle. Physiological Genomics **1:** 165-175.

HILL, W. G., 1981 Estimation of effective population size from data on linkage disequilibrium. Genetics Research **38:** 209-216.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics **38:** 226-231.

KOHAVI, R., and G. H. JOHN, 1997 Wrappers for feature subset selection. Artificial Intelligence **97:** 273-324.

KOLBEHDARI, D., L. R. SCHAEFFER and J. A. B. ROBINSON, 2007 Estimation of genome-wide haplotype effects in half-sib designs. Journal of Animal Breeding and Genetics **124:** 356-361.

KONIG, S., H. SIMIANER and A. WILLAM, 2009 Economic evaluation of genomic breeding programs. Journal of Dairy Science **92:** 382-391.

KOROL, A., Z. FRENKEL, L. COHEN, E. LIPKIN and M. SOLLER, 2007 Fractioned DNA pooling: a new cost-effective strategy for fine mapping of quantitative trait loci. Genetics **176:** 2611-2623.

KOSAMBI, D.D.1944. The estimation of map distances from recombination values. Ann Eugen.

KRUIP, T. A., M. C. PIETERSE, T. H. VAN BENEDEN, P. L. VOS, Y. A. WURTH et al., 1991 A new method for bovine embryo production: a potential alternative to superovulation. Vet Rec **128:** 208-210.

LANDE, R., and R. THOMPSON, 1990 Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics **124:** 743-756.

LANDER, E. S., and D. BOTSTEIN, 1994 Mapping Mendelian Factors Underlying Quantitative Traits Using Rflp Linkage Maps (Vol 121, Pg 185, 1989). Genetics **136:** 705-705.

LEGARRA, A., and I. MISZTAL, 2008 Technical note: Computing strategies in genome-wide selection. Journal of Dairy Science **91:** 360-366.

LONG, N., D. GIANOLA, G. J. M. ROSA, K. A. WEIGEL and S. AVENDAÑO, 2007 Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of Animal Breeding and Genetics **124:** 377-389.

LUAN, T., J. A. WOOLLIAMS, S. LIEN, M. KENT, M. SVENDSEN et al., 2009 The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics **183:** 1119-1126.

MACCIOTTA, N. P., G. GASPA, R. STERI, C. PIERAMATI, P. CARNIER et al., 2009 Pre-selection of most significant SNPS for the estimation of genomic breeding values. BMC Proc **3 Suppl 1:** S14.

MACCIOTTA, N. P. P., G. GASPA, R. STERI, E. L. NICOLAZZI, C. DIMAURO et al., 2010 Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. Journal of Dairy Science **93:** 2765-2774.

MACCIOTTA, N. P. P., M. MELE, G. CONTE, A. SERRA, M. CASSANDRO et al., 2008 Association Between a Polymorphism at the Stearoyl CoA Desaturase Locus and Milk Production Traits in Italian Holsteins. Journal of Dairy Science **91:** 3184-3189.

MACCIOTTA N. P. P., M. A. PINTUS, R. STERI, C. PIERAMATI, E. L. NICOLAZZI, E.SANTUS, D. VICARIO, J. T. VAN AAM, A. NARDONE, A. VALENTINI, AND P. AJMONE-MARSAN. Accuracies of direct genomic breeding values estimated in dairy cattle with a principal component approach. J. Anim. Sci. **88:** 532-533, E-Suppl. 2/J. Dairy Sci **93:**532-533. Proceeding ADSA-ASAS annual joint meeting July 11-15. Denver Colorado

MELE, M., G. CONTE, B. CASTIGLIONI, S. CHESSA, N. P. P. MACCIOTTA et al., 2007 Stearoyl-Coenzyme A Desaturase Gene Polymorphism and Milk Fatty Acid Composition in Italian Holsteins. Journal of Dairy Science **90:** 4458-4465.

MEUWISSEN, T., 2007 Genomic selection : marker assisted selection on a genome wide scale. Journal of Animal Breeding and Genetics **124:** 321-322.

MEUWISSEN, T. H. E., B. J. HAYES and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157:** 1819-1829.

MEUWISSEN, T. H. E., A. KARLSEN, S. LIEN, I. OLSAKER and M. E. GODDARD, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics **161:** 373-379.

MORRIS, C. A., N. G. CULLEN, B. C. GLASS, D. L. HYNDMAN, T. R. MANLEY et al., 2007 Fatty acid synthase effects on bovine adipose fat and milk fat. Mammalian Genome **18:** 64-74.

MOSER, G., B. TIER, R. E. CRUMP, M. S. KHATKAR and H. W. RAADSMA, 2009 A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution **41**.

MOSIG, M. O., E. LIPKIN, G. KHUTORESKAYA, E. TCHOURZYNA, M. SOLLER et al., 2001 A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics **157:** 1683-1698.

MUIR, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics **124:** 342-355.

NORDBORG, M., and S. TAVARE, 2002 Linkage disequilibrium: what history has to tell us. Trends in Genetics **18:** 83-90.

OLSEN, H. G., L. GOMEZ-RAYA, D. I. VAGE, I. OLSAKER, H. KLUNGLAND et al., 2002 A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. Journal of Dairy Science **85:** 3124-3130.

PIMENTEL, E. C., S. KONIG, F. S. SCHENKEL and H. SIMIANER, 2009 Comparison of statistical procedures for estimating polygenic effects using dense genome-wide marker data. BMC Proc **3 Suppl 1:** S12.

RENDEL, J., and A. ROBERTSON, 1950 Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. Journal of Genetics **50:** 1-8.

RISCH, N., and K. MERIKANGAS, 1996 The Future of Genetic Studies of Complex Human Diseases. Science **273:** 1516-1517.

RON, M., E. FELDMESSER, M. GOLIK, I. TAGER-COHEN, D. KLIGER et al., 2004 A complete genome scan of the Israeli Holstein population for quantitative trait loci by a daughter design. J Dairy Sci **87:** 476-490.

RON, M., and J. I. WELLER, 2007 From QTL to QTN identification in livestock - winning by points rather than knock-out: a review. Animal Genetics **38:** 429-439.

ROWSON, L. E., 1971 Egg transfer in domestic animals. Nature **233:** 379-381.

ROY, R., L. ORDOVAS, P. ZARAGOZA, A. ROMERO, C. MORENO et al., 2006 Association of polymorphisms in the bovine FASN gene with milk-fat content. Animal Genetics **37:** 215-218.

SCHAEFFER, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics **123:** 218-223.

SCHEET, P., and M. STEPHENS, 2006 A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics **78:** 629-644.

SMITH, C., 1967 Improvement of metric traits through specific genetic loci. Animal Science **9:** 349-358.

SOLBERG, T. R., A. K. SONESSON, J. A. WOOLLIAMS and T. H. E. MEUWISSEN, 2008 Genomic selection using different marker types and densities. Journal of Animal Science **86:** 2447-2454.

SOLBERG, T. R., A. K. SONESSON, J. A. WOOLLIAMS and T. H. E. MEUWISSEN, 2009 Reducing dimensionality for prediction of genome-wide breeding values. Genetics Selection Evolution **41:** -.

SPELMAN, R. J., C. A. FORD, P. MCELHINNEY, G. C. GREGORY and R. G. SNELL, 2002 Characterization of the DGAT1 gene in the New Zealand dairy population. J Dairy Sci **85:** 3514-3517.

VAN TASSELL, C. P., T. P. L. SMITH, L. K. MATUKUMALLI, J. F. TAYLOR, R. D. SCHNABEL et al., 2008 SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth **5:** 247-252.

VANRADEN, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science **91:** 4414-4423.

VANRADEN, P. M., C. P. VAN TASSELL, G. R. WIGGANS, T. S. SONSTEGARD, R. D. SCHNABEL et al., 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science **92:** 16-24.

WEIGEL, K. A., G. DE LOS CAMPOS, O. GONZALEZ-RECIO, H. NAYA, X. L. WU et al., 2009 Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science **92:** 5248-5257.

WELLER, J. I., Y. KASHI and M. SOLLER, 1990 Power of Daughter and Granddaughter Designs for Determining Linkage between Marker Loci and Quantitative Trait Loci in Dairy-Cattle. Journal of Dairy Science **73:** 2525-2537.

WELLER, J. I., M. SHLEZINGER and M. RON, 2005 Correcting for bias in estimation of quantitative trait loci effects. Genetics Selection Evolution **37:** 501-522.

WOOLASTON, A. F., B. TIER, AND R. D. MURISON, 2007 Principal components regression of SNP
data to predict genetic merit. Papers and abstracts from the workshop on QTL and marker assisted selelction, 22-23 March 2007, Toulouse, France, edited by A. Legarra

ZHAO, H., D. NETTLETON, M. SOLLER and J. C. DEKKERS, 2005 Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genetics Research **86:** 77-87.

CHAPTER 2

REVIEW OF QTL DETECTED ON DAIRY CATTLE AND META-ANALYSIS

ABSTRACT

A large number of quantitative trait loci (QTLs) affecting milk production and quality traits in dairy cattle have been reported in literature. A total of 150 papers by 90 authors were found on 29 scientific Journals for the period January 1995-February 2008. QTL meta-analyses have been carried out to estimate the distribution of QTL effects in livestock and to find consensus on QTL position. In this study seven selected variables were analyzed both with the Factor Analysis (FA) and Principal Component Analysis (PCA). Furthermore, five theoretical distributions (lognormal, weibull, normal, exponential and gamma distribution) were used to model QTL effect distributions of milk yield (MY), protein yield (PY), protein percent (PP), fat yield (FY), fat percent (FP), type traits (TT) and all milk production traits scaled by genetic standard deviation (AT). FA was able to explain 68% of the original variability with 3 latent factors: the first factor extracted is highly associated (0.98) to marker location along the chromosome and could be considered as a marker map index; the second factor shows loadings of 0.74 and 0.84 related to the number of animals involved and to the year of the experiment, respectively, and it can be regarded as an indicator of the dimension of the study; the third factor is correlated positively to the significance level of the statistical test (0.78), to the number of families (0.63) and, negatively, to the marker density (-0.43) and can be interpreted as an index of power of the experiment. Same patterns can be observed in the eigenvectors of PCA. Four PCs were able to explain about 80% of the original variance. The first two PCs basically underline the same structure found with the first two factors, whereas PC3 and PC4 summarize the structure of F3. The score that each QTL gets on each factor or PC could be useful tool classify the original QTL studies and make them more comparable once that the redundancy of information has been removed. The investigation on QTL effect distributions indicates the gamma function as the most suitable to fit data for all traits but MY and PP. The lognormal distribution fitted well FY, FP, PY and AT data, whereas the Weibull distribution showed a good fit only for FY, FP and PP.

Key words: QTLs, meta-analysis, dairy cattle; marker assisted selection (MAS)

Pubblished as Proceeding of conference

GASPA, G., M. A. PINTUS, R. STERI, S. SORBOLINI, AND N. P. P. MACCIOTTA 2009. *Analysis of distributions of estimated QTL effects for dairy cattle*. J. Anim. Sci. Vol. 92, E-Suppl. 1 pag. 397. Abstract

GASPA G., PERINI D., MARINO R., MAZZA R., SORBOLINI S., STERI R. (2009). *Multivariate meta-analysis of QTL mapping studies.* Italian Journal of Animal Science, Vol. 8 (Suppl. 2), p. 208. Abstract

2.1. INTRODUCTION

2.1.1. Genetic marker in animal breeding: a meta-analytic approach

The infinitesimal model used to explain the inheritance of quantitative traits has been the paradigm of selection so far. Its implementation in breeding programmes has yielded considerable increases in the genetic level of farm animals, especially for dairy cattle (DEKKERS and HOSPITAL 2002). However, being the expression of the economic traits controlled by a finite number of genes (CHAMBERLAIN *et al.* 2007 ; EWING and GREEN 2000), the finite locus model could be more appropriate to describe the inheritance pattern. The hypothesis a finite number of loci may led to a model with many genes of small effect and few ones with large effect (HAYES and GODDARD 2001). Moreover, the current availability of dense marker panel (chips with 54K SNP) allows to use the marker information in the prediction of breeding values of bulls and the knowledge of prior distribution of QTL effects may help on estimate with high accuracy the markers effect (GODDARD and HAYES 2007). Nevertheless the distribution of size of QTL effects is not well established yet for most traits.

Several genome scan studies carried out on livestock species reported a large number of quantitative trait loci affecting economic traits. The main aim of these studies was to integrate this information into marker assisted selection programs. However, commercial applications of MAS have been rather limited so far (BENNEWITZ *et al.* 2004b; DEKKERS 2004; GUILLAUME *et al.* 2008b). Currently, there is a general lack of consensus on QTL effects estimation and on chromosomal locations, with an average confidence interval for QTL position of more than 20 cM (KHATKAR *et al.* 2004).

In any case, the large amount of data available in literature may be exploited by meta-analysis to draw more general conclusions from results obtained in different experimental conditions, population investigated and statistical methodologies. Meta-analytic techniques have been initially proposed in social and medical sciences. The use of statistical methods to combine the results of independent research studies dates long time ago. Different objectives may be pursued using meta-analytic approaches and, conversely from classical descriptive review of a general topic, the meta-analysis may lead to new results. In an earlier application PEARSON (1904) collected correlation coefficients from several studies to determine the extent to which inoculation against smallpox disease was related to survival. Meta-analyses have often been carried out to analyze a series of studies on the same subject in medical science, allowing a quantitative summary of results. The meta-analysis is useful when the results of individual studies are conflicting or carried out only with limited sample size. In this case the findings are not reliable enough because of low statistical power and meta-analytic approach may help to enforce the evidence.

Five steps are generally required for conducting a meta-analysis:

1) clear and complete definition of meta-analytic objective(s);
2) search for most relevant paper on the topics;
3) definition of the criteria for inclusion of data in the analysis;
4) application of an appropriate mathematical-statistical method to fit the data;
5) critical evaluation of results.

Two approaches could be generally applied when carrying out a meta-analysis. In the first raw data from different experiments on the same topic are pooled and analyzed again with the aim to increase the sample size and the power of the experiment. As stated by LANDER and KRUGLYAK (1995), this approach would allow more robust results even though in many situation it is not feasible, especially when the published literature is too wide as in the case of linkage analysis studies to detect QLTs. The second, most common, approach is based on the collection of results from published papers and in the statistical correction for the effect of the study in order to draw more general conclusions.

Different methods have been proposed to meta-analyze the data retrieved from different studies. According to LI and RAO (1996) multiple replication studies tend to produce different results. They analyzed genetic effects from many independent quantitative sib-pair linkage studies using a random effect model . Each study used by LI and RAO (1996) evaluated the same markers using the same methodology. In humans, ALLISON and HEO (1998) proposed a meta-analysis technique “under the worst-case condition”. Briefly, they analyzed the P-value of five linkage studies that reported several markers, tested with different statistical techniques (multiple testing hypothesis and multiple marker tested) and with missing data. They pooled the *m* independent P-values into a single test of significance under the null hypothesis of no association in humans between OB genomic region and body mass index. They found a strong evidence (P-value =$1.5 \cdot 10^{-5}$) of an association for a marker in the Human OB gene.

GOFFINET and GERBER (2000) proposed a mathematical-statistical method for combining results from several independent studies which they have tested on simulated data. Later, this technique was used by KHATKAR *et al.* (2004) who analyzed 55 publications on QTL studies conducted on dairy cattle performing a meta-analysis looking for consensus on the position of QTLs influencing different milk production traits. Finally, HAYES and GODDARD (2001) studied the distribution of the effects of QTL in two populations (dairy cattle and pigs). The authors indicate the gamma distribution best suit to describe the distribution of effects: 17% QTLs explaining about 90% of genetic variance. HAYES and GODDARD (2001) estimated between 50 and 100 genes with a distribution where few QTLs have a major effect and many QTLs have small effects.

However, QTL meta-analysis techniques present peculiar problems. One is represented by the reduced availability of data in published papers: QTL effects are often missed, and marker positions are referred to different linkage maps. A common practice in QTL meta-analysis studies to cope with missing data is to reconstruct them whenever possible (GOFFINET and GERBER 2000; LANDER and KRUGLYAK 1995), or to rescale them in case of different marker maps. A further issue is whether QTLs reported in very close positions but with different effects in several studies should be considered the same or not. The extraction of one or of more synthetic variables from information reported in different QTLs studies could be a way to characterize the researches and to give an index of QTL reliability.

The objectives of the present study were: *i)* to build an updated data base of researches on QTLs in dairy cattle in order to carry out a descriptive statistical analysis of QTL results and to conduct a meta-analysis; *ii)* to seek latent variables able to characterize the research using multivariate dimension reduction techniques to analyze a data base of published QTLs; *iii)* to test some theoretical distributions to model QTL effect distributions all milk production traits.

2.2. MATERIALS AND METHODS

2.2.1 A data base of Quantitative Trait Loci studies for Dairy cattle

The relevant literature on dairy cattle QTL mapping was investigated. A total of 150 articles published on 29 scientific journals from January 1995 to February 2008 were retrieved (Table 1, Figure 1). Moreover, information reported on the following three specific online QTL data bases were also used and compared:

- http://www.vetsci.usyd.edu.au/reprogen/QTL_Map
- http://www.animalgenome.org/QTLdb,
- http://bovineqtlv2.tamu.edu.

More than thirty parameters were picked up from the articles (Table 2). A descriptive statistical analysis has been carried out with the aim of summarizing the principal features of the database. In particular the total number of QTL found and the trend of QTL detected per year; furthermore the breeds and the experimental designs used have been descripted. Traits analyzed and number of QTL detected per chromosome have been further reported. Several statistical models have been used to map QTL in outbreed populations, from less complex (Anova) to a higher level of complexity (Bayesian MCMC methods). Here we report a brief classification of the methods adopted and the phenotypes used as response variable. The significance value (P-value) is reported for the whole dataset. Furthermore, the effect of well-known (or novel detected) polymorphisms on phenotypic trait values have been used to compare the results of QTL detected in different studies at descriptive level (reported as *Bos Taurus* release 4.1). A graphic description of significance levels of QTL affecting milk core traits detected on chromosome 5 for milk production trait is reported . Trait analyzed were milk yield (MY), fat yield (FY), fat percentage (FP), protein yield (PY), protein percentage (PP). Finally, the effect of QTL affecting milk production traits is described.

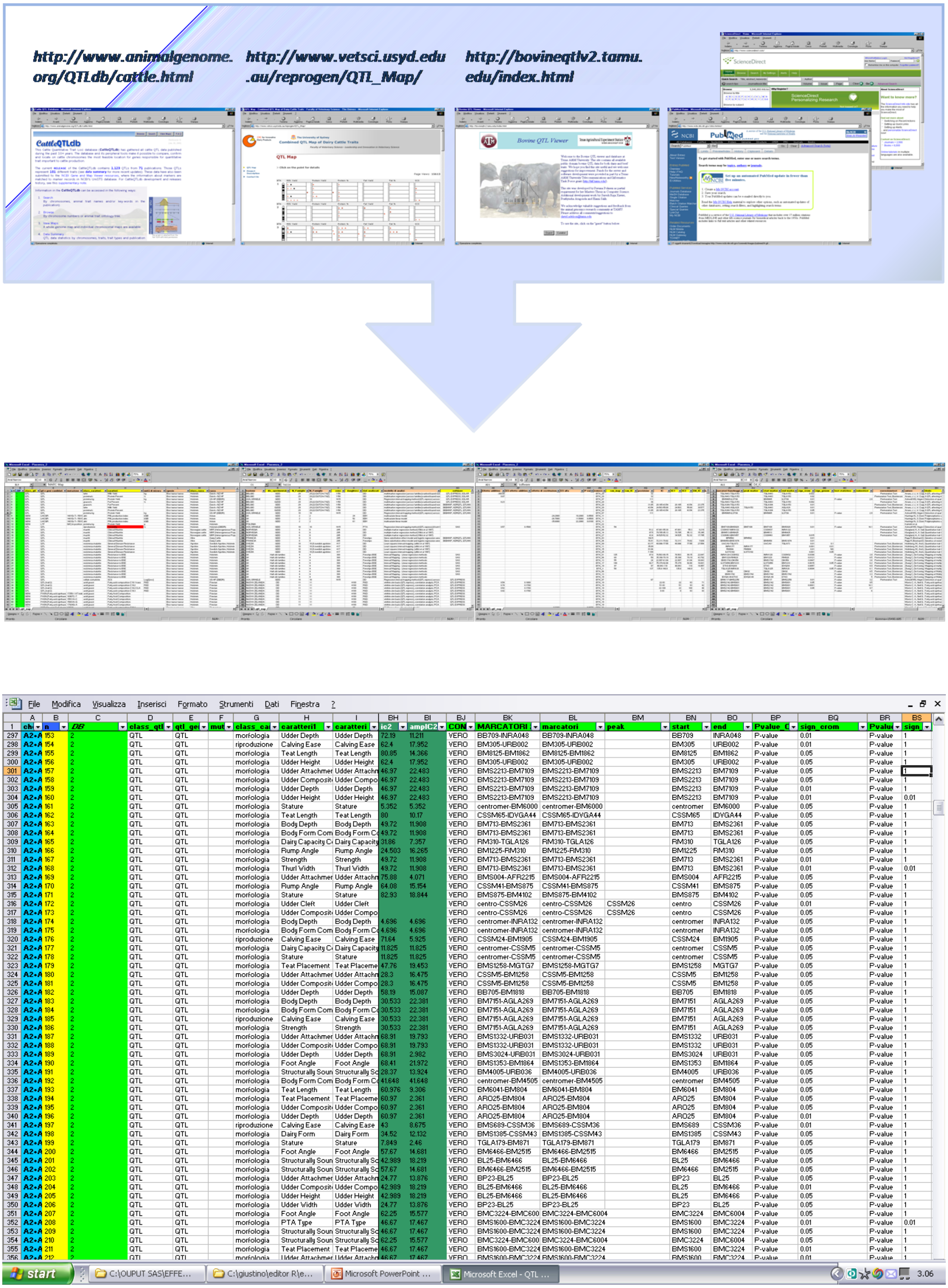
Table 1. Journals considered in the analysis and number of articles per Journal

Journal	n
1. Journal of dairy Science	48
2. Proceedings	17
3. Animal genetics	16
4. Genetics	13
5. Mammalian genome	8
6. Journal of animal breeding and genetics	6
7. Animal biotechnology	4
8. Journal of animal science	4
9. Journal of applied genetics	4
10. BMC Genetics	3
11. Journal of heredity	3
12. Physiological genomics	3
13. Genetics selection evolution	2
14. Genome research	2
15. Genomics	2
16. Italian journal of animal science	2
17. Animal science journal	1
18. Asian-Australian journal of animal science	1
19. Australian journal of agricultural research	1
20. BMC Genomics	1
21. BMC Veterinary research	1
22. Genetics Research Cambridge	1
23. Genetika	1
24. Journal of dairy research	1
25. Pigment cell research	1
26. Research in veterinary science	1
27. Veterinary Research	1
28. Veterinary Medicina	1

Table 2. Variables included in the database of QTL for dairy cattle.

Observed Variable	n‡
1. QTL or Candidate gene	2651
2. Mutation	846
3. Trait	2650
4. Measurement unit	1600
5. Breed	2641
6. Nation of experiment	2534
7. Experimental Design	2311
8. Number of family	2108
9. Number of sons	1682
10. Number of Daughters	928
11. Analyzed Phenotypes	2302
12. Single-multi QTL model	232
13. Software used	2404
14. Analytic model	762
15. Additive effect	550
16. SE additive effect	285
17. Dominance effect	142
18. SE dominance effect	116
19. Allelic substitution effect	104
20. SE allelic substitution effect	549
21. Absolute effect	1065
22. Genetic variance explained	353
23. Chromosome	2556
24. QTL location	1428
25. Location confidence interval	566
26. Flanking markers	1069
27. Significance level (chrom-wise)	1250
28. Significance level (genome-wise)	359
29. Statistic tests used	1028
30. Test value	770
31. Multiple test Correction	1491
32. Marker map used	1121
33. References	2648

† number entries for which were available the variable in the correspond field of the dataset.


http://www.animalgenome.org/QTLdb/cattle.html
http://www.vetsci.usyd.edu.au/reprogen/QTL_Map/
http://bovineqtlv2.tamu.edu/index.html
File
Modifica
Visualizza
Inserisci
Formato
Strumenti
Dati
Finestra
?
start
C:\OUPUT SAS\EFFE...
C:\giustino\editor R\e...
Microsoft PowerPoint ...
Microsoft Excel - QTL ...
3.06

	A	B	C	D	E	F	G	H	I	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS
1	ch	n	DB	class_qtl	qtl_ge	mut	class_ca	caratteri1	caratteri	ic2	ampIC2	CON	MARCATORI	marcatori	peak	start	end	Pvalue_C	sign_crom	Pvalu	sign_
297	A2+A	153	2	QTL	QTL		morfologia	Udder Depth	Udder Depth	72.19	11.211	VERO	BB709-INRA048	BB709-INRA048		BB709	INRA048	P-value	0.01	P-value	1
298	A2+A	154	2	QTL	QTL		riproduzione	Calving Ease	Calving Ease	62.4	17.952	VERO	BM305-URB002	BM305-URB002		BM305	URB002	P-value	0.01	P-value	1
299	A2+A	155	2	QTL	QTL		morfologia	Teat Length	Teat Length	80.85	14.366	VERO	BM8125-BM1862	BM8125-BM1862		BM8125	BM1862	P-value	0.05	P-value	1
300	A2+A	156	2	QTL	QTL		morfologia	Udder Height	Udder Height	62.4	17.952	VERO	BM305-URB002	BM305-URB002		BM305	URB002	P-value	0.05	P-value	1
301	A2+A	157	2	QTL	QTL		morfologia	Udder Attachmer	Udder Attachn	46.97	22.483	VERO	BMS2213-BM7109	BMS2213-BM7109		BMS2213	BM7109	P-value	0.05	P-value	1
302	A2+A	158	2	QTL	QTL		morfologia	Udder Composit	Udder Compo	46.97	22.483	VERO	BMS2213-BM7109	BMS2213-BM7109		BMS2213	BM7109	P-value	0.05	P-value	1
303	A2+A	159	2	QTL	QTL		morfologia	Udder Depth	Udder Depth	46.97	22.483	VERO	BMS2213-BM7109	BMS2213-BM7109		BMS2213	BM7109	P-value	0.01	P-value	1
304	A2+A	160	2	QTL	QTL		morfologia	Udder Height	Udder Height	46.97	22.483	VERO	BMS2213-BM7109	BMS2213-BM7109		BMS2213	BM7109	P-value	0.01	P-value	0.01
305	A2+A	161	2	QTL	QTL		morfologia	Stature	Stature	5.352	5.352	VERO	centromer-BM6000	centromer-BM6000		centromer	BM6000	P-value	0.05	P-value	1
306	A2+A	162	2	QTL	QTL		morfologia	Teat Length	Teat Length	80	10.17	VERO	CSSM65-IDVGA44	CSSM65-IDVGA44		CSSM65	IDVGA44	P-value	0.05	P-value	1
307	A2+A	163	2	QTL	QTL		morfologia	Body Depth	Body Depth	49.72	11.908	VERO	BM713-BMS2361	BM713-BMS2361		BM713	BMS2361	P-value	0.05	P-value	1
308	A2+A	164	2	QTL	QTL		morfologia	Body Form Com	Body Form Co	49.72	11.908	VERO	BM713-BMS2361	BM713-BMS2361		BM713	BMS2361	P-value	0.05	P-value	1
309	A2+A	165	2	QTL	QTL		morfologia	Dairy Capacity Ca	Dairy Capacity	31.86	7.357	VERO	RM310-TGLA126	RM310-TGLA126		RM310	TGLA126	P-value	0.05	P-value	1
310	A2+A	166	2	QTL	QTL		morfologia	Rump Angle	Rump Angle	24.503	16.265	VERO	BM1225-RM310	BM1225-RM310		BM1225	RM310	P-value	0.05	P-value	1
311	A2+A	167	2	QTL	QTL		morfologia	Strength	Strength	49.72	11.908	VERO	BM713-BMS2361	BM713-BMS2361		BM713	BMS2361	P-value	0.01	P-value	1
312	A2+A	168	2	QTL	QTL		morfologia	Thurl Width	Thurl Width	49.72	11.908	VERO	BM713-BMS2361	BM713-BMS2361		BM713	BMS2361	P-value	0.01	P-value	0.01
313	A2+A	169	2	QTL	QTL		morfologia	Udder Attachmer	Udder Attachn	75.88	4.071	VERO	BMS004-AFR2215	BMS004-AFR2215		BMS004	AFR2215	P-value	0.05	P-value	1
314	A2+A	170	2	QTL	QTL		morfologia	Rump Angle	Rump Angle	64.08	15.154	VERO	CSSM41-BMS875	CSSM41-BMS875		CSSM41	BMS875	P-value	0.05	P-value	1
315	A2+A	171	2	QTL	QTL		morfologia	Stature	Stature	82.93	18.844	VERO	BMS875-BM4102	BMS875-BM4102		BMS875	BM4102	P-value	0.05	P-value	1
316	A2+A	172	2	QTL	QTL		morfologia	Udder Cleft	Udder Cleft			VERO	centro-CSSM26	centro-CSSM26	CSSM26	centro	CSSM26	P-value	0.01	P-value	1
317	A2+A	173	2	QTL	QTL		morfologia	Udder Composit	Udder Compo			VERO	centro-CSSM26	centro-CSSM26	CSSM26	centro	CSSM26	P-value	0.05	P-value	1
318	A2+A	174	2	QTL	QTL		morfologia	Body Depth	Body Depth	4.696	4.696	VERO	centromer-INRA132	centromer-INRA132		centromer	INRA132	P-value	0.05	P-value	1
319	A2+A	175	2	QTL	QTL		morfologia	Body Form Com	Body Form Co	4.696	4.696	VERO	centromer-INRA132	centromer-INRA132		centromer	INRA132	P-value	0.05	P-value	1
320	A2+A	176	2	QTL	QTL		riproduzione	Calving Ease	Calving Ease	71.64	5.925	VERO	CSSM24-BM1905	CSSM24-BM1905		CSSM24	BM1905	P-value	0.05	P-value	1
321	A2+A	177	2	QTL	QTL		morfologia	Dairy Capacity Ca	Dairy Capacity	11.825	11.825	VERO	centromer-CSSM5	centromer-CSSM5		centromer	CSSM5	P-value	0.05	P-value	1
322	A2+A	178	2	QTL	QTL		morfologia	Stature	Stature	11.825	11.825	VERO	centromer-CSSM5	centromer-CSSM5		centromer	CSSM5	P-value	0.05	P-value	1
323	A2+A	179	2	QTL	QTL		morfologia	Teat Placement	Teat Placeme	47.76	19.453	VERO	BMS1258-MGTG7	BMS1258-MGTG7		BMS1258	MGTG7	P-value	0.05	P-value	1
324	A2+A	180	2	QTL	QTL		morfologia	Udder Attachmer	Udder Attachn	28.3	16.475	VERO	CSSM5-BM1258	CSSM5-BM1258		CSSM5	BM1258	P-value	0.05	P-value	1
325	A2+A	181	2	QTL	QTL		morfologia	Udder Composit	Udder Compo	28.3	16.475	VERO	CSSM5-BM1258	CSSM5-BM1258		CSSM5	BM1258	P-value	0.05	P-value	1
326	A2+A	182	2	QTL	QTL		morfologia	Udder Depth	Udder Depth	58.19	15.087	VERO	BB705-BM1818	BB705-BM1818		BB705	BM1818	P-value	0.05	P-value	1
327	A2+A	183	2	QTL	QTL		morfologia	Body Depth	Body Depth	30.533	22.381	VERO	BM7151-AGLA269	BM7151-AGLA269		BM7151	AGLA269	P-value	0.05	P-value	1
328	A2+A	184	2	QTL	QTL		morfologia	Body Form Com	Body Form Co	30.533	22.381	VERO	BM7151-AGLA269	BM7151-AGLA269		BM7151	AGLA269	P-value	0.05	P-value	1
329	A2+A	185	2	QTL	QTL		riproduzione	Calving Ease	Calving Ease	30.533	22.381	VERO	BM7151-AGLA269	BM7151-AGLA269		BM7151	AGLA269	P-value	0.05	P-value	1
330	A2+A	186	2	QTL	QTL		morfologia	Strength	Strength	30.533	22.381	VERO	BM7151-AGLA269	BM7151-AGLA269		BM7151	AGLA269	P-value	0.05	P-value	1
331	A2+A	187	2	QTL	QTL		morfologia	Udder Attachmer	Udder Attachn	68.91	19.793	VERO	BMS1332-URB031	BMS1332-URB031		BMS1332	URB031	P-value	0.05	P-value	1
332	A2+A	188	2	QTL	QTL		morfologia	Udder Composit	Udder Compo	68.91	19.793	VERO	BMS1332-URB031	BMS1332-URB031		BMS1332	URB031	P-value	0.05	P-value	1
333	A2+A	189	2	QTL	QTL		morfologia	Udder Depth	Udder Depth	68.91	2.982	VERO	BMS3024-URB031	BMS3024-URB031		BMS3024	URB031	P-value	0.05	P-value	1
334	A2+A	190	2	QTL	QTL		morfologia	Foot Angle	Foot Angle	68.41	21.972	VERO	BMS1353-BM1864	BMS1353-BM1864		BMS1353	BM1864	P-value	0.05	P-value	1
335	A2+A	191	2	QTL	QTL		morfologia	Structurally Soun	Structurally Sc	28.37	13.924	VERO	BM4005-URB036	BM4005-URB036		BM4005	URB036	P-value	0.05	P-value	1
336	A2+A	192	2	QTL	QTL		morfologia	Body Form Com	Body Form Co	41.648	41.648	VERO	centromer-BM4505	centromer-BM4505		centromer	BM4505	P-value	0.05	P-value	1
337	A2+A	193	2	QTL	QTL		morfologia	Teat Length	Teat Length	60.976	9.306	VERO	BM6041-BM804	BM6041-BM804		BM6041	BM804	P-value	0.05	P-value	1
338	A2+A	194	2	QTL	QTL		morfologia	Teat Placement	Teat Placeme	60.97	2.361	VERO	ARO25-BM804	ARO25-BM804		ARO25	BM804	P-value	0.05	P-value	1
339	A2+A	195	2	QTL	QTL		morfologia	Udder Composit	Udder Compo	60.97	2.361	VERO	ARO25-BM804	ARO25-BM804		ARO25	BM804	P-value	0.05	P-value	1
340	A2+A	196	2	QTL	QTL		morfologia	Udder Depth	Udder Depth	60.97	2.361	VERO	ARO25-BM804	ARO25-BM804		ARO25	BM804	P-value	0.01	P-value	1
341	A2+A	197	2	QTL	QTL		riproduzione	Calving Ease	Calving Ease	43	8.675	VERO	BMS689-CSSM36	BMS689-CSSM36		BMS689	CSSM36	P-value	0.01	P-value	1
342	A2+A	198	2	QTL	QTL		morfologia	Dairy Form	Dairy Form	34.52	12.132	VERO	BMS1385-CSSM43	BMS1385-CSSM43		BMS1385	CSSM43	P-value	0.05	P-value	1
343	A2+A	199	2	QTL	QTL		morfologia	Stature	Stature	7.849	2.46	VERO	TGLA179-BM871	TGLA179-BM871		TGLA179	BM871	P-value	0.05	P-value	1
344	A2+A	200	2	QTL	QTL		morfologia	Foot Angle	Foot Angle	57.67	14.681	VERO	BM6466-BM2515	BM6466-BM2515		BM6466	BM2515	P-value	0.05	P-value	1
345	A2+A	201	2	QTL	QTL		morfologia	Structurally Soun	Structurally Sc	42.989	18.219	VERO	BL25-BM6466	BL25-BM6466		BL25	BM6466	P-value	0.05	P-value	1
346	A2+A	202	2	QTL	QTL		morfologia	Structurally Soun	Structurally Sc	57.67	14.681	VERO	BM6466-BM2515	BM6466-BM2515		BM6466	BM2515	P-value	0.05	P-value	1
347	A2+A	203	2	QTL	QTL		morfologia	Udder Attachmer	Udder Attachn	24.77	13.876	VERO	BP23-BL25	BP23-BL25		BP23	BL25	P-value	0.05	P-value	1
348	A2+A	204	2	QTL	QTL		morfologia	Udder Composit	Udder Compo	42.989	18.219	VERO	BL25-BM6466	BL25-BM6466		BL25	BM6466	P-value	0.01	P-value	1
349	A2+A	205	2	QTL	QTL		morfologia	Udder Height	Udder Height	42.989	18.219	VERO	BL25-BM6466	BL25-BM6466		BL25	BM6466	P-value	0.05	P-value	1
350	A2+A	206	2	QTL	QTL		morfologia	Udder Width	Udder Width	24.77	13.876	VERO	BP23-BL25	BP23-BL25		BP23	BL25	P-value	0.05	P-value	1
351	A2+A	207	2	QTL	QTL		morfologia	Foot Angle	Foot Angle	62.25	15.577	VERO	BMC3224-BMC600	BMC3224-BMC6004		BMC3224	BMC6004	P-value	0.05	P-value	1
352	A2+A	208	2	QTL	QTL		morfologia	PTA Type	PTA Type	46.67	17.467	VERO	BMS1600-BMC3224	BMS1600-BMC3224		BMS1600	BMC3224	P-value	0.01	P-value	0.01
353	A2+A	209	2	QTL	QTL		morfologia	Structurally Soun	Structurally Sc	46.67	17.467	VERO	BMS1600-BMC3224	BMS1600-BMC3224		BMS1600	BMC3224	P-value	0.05	P-value	1
354	A2+A	210	2	QTL	QTL		morfologia	Structurally Soun	Structurally Sc	62.25	15.577	VERO	BMC3224-BMC600	BMC3224-BMC6004		BMC3224	BMC6004	P-value	0.05	P-value	1
355	A2+A	211	2	QTL	QTL		morfologia	Teat Placement	Teat Placeme	46.67	17.467	VERO	BMS1600-BMC3224	BMS1600-BMC3224		BMS1600	BMC3224	P-value	0.01	P-value	1
356	A2+A	212	2	QTL	QTL		morfologia	Udder Attachmer	Udder Attachn	46.67	17.467	VERO	BMS1600-BMC3224	BMS1600-BMC3224		BMS1600	BMC3224	P-value	0.01	P-value	1

Figure 1 QTL data base construction.

2.2.2 Multivariate meta-analysis of QTL mapping studies

Dataset

The raw dataset was edited to give more uniformity to the data mined from different studies and to organize it for the statistical analyses. To overcome the problem of different map locations, the flanking markers were mapped on release 4.1 of the *Bos taurus* genome sequence (www.ensembl.org). Their positions were retrieved from public databases or, when not available, was calculated in silico by blasting (http://blast.wustl.edu/) the markers' nucleotide sequence against the genomic sequence. Relationships between position of markers in cM (from published paper) and position of markers according to the physical map is shown in figure 2. Records were discarded if flanking markers or P-values were not available. Additional variables have been calculated from original raw data, and character variables have been transformed into discrete numeric variables. After these edits, the final archive consisted of 1,162 records.

To select the most relevant variable a preliminary exploratory data analysis was carried out examining on the whole data set Pearson and partial correlation matrices, and Kaiser's measure of sample adequacy (MSA) (CERNY and KAISER 1977) were calculated. High values of MSA (ranging from 0 to 1, at least greater of 0.60) indicate a latent structure underlying the data and suitability of the archive to multivariate factor (FA) analysis. After that, a variable selection step was carried out using a preliminary (FA) (see statistical analysis for full description of the factorial model) on the whole set of variables. Only those highly correlated with the common factors were retained. Also redundant variables were removed.

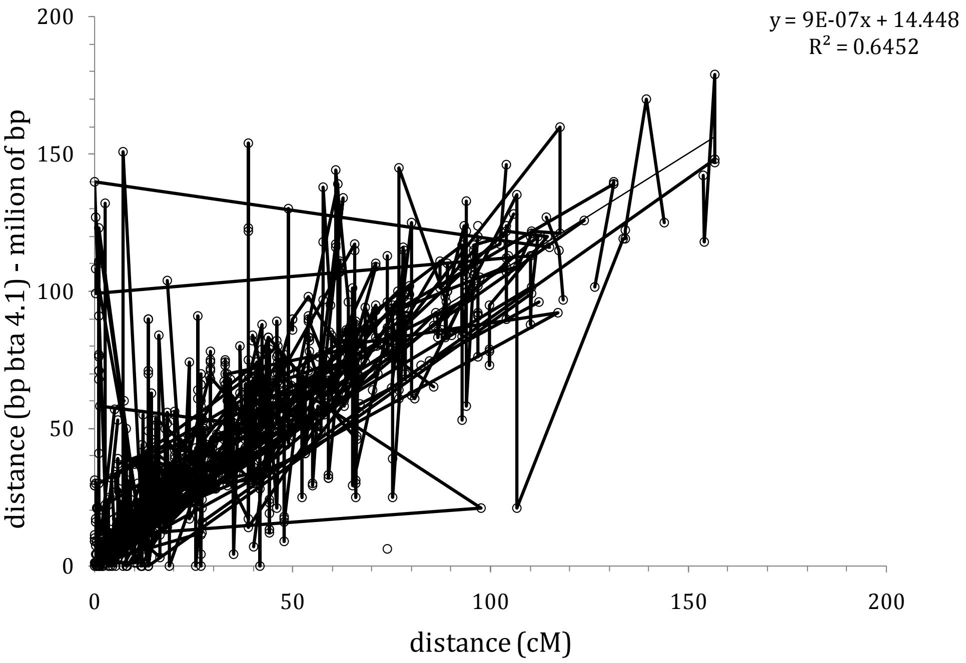


Figure 2. Relationship between position in cM and position in Mb estimated by linear regression

Statistical analysis

Selected variables were analyzed using the factor analysis and principal component analysis (PCA). In the factorial model the value of the variable X_i for the *i-th* observation can be decomposed as follow:

$$X_i = \sum_j b_{ij} \cdot F_j + e_i \qquad \text{(for } j=1,m\text{)}$$

where F_j is the *j*-th common factor (or latent variable), b_{ij} is called factor loading and weighs the *i*-th original variable in the composition of the *j*-th factor, *m* is the number of extracted factor, e_i is the uniqueness of the *i*-th variable (KRZANOWSKY, 2000).

Kaiser MSA was used to evaluate the suitability of dataset to FA. The proportion of variance explained by the common factors (~70% of the variance of the original variances) as well as the min-eigen criterion were used to choose the appropriate number of factor to retain. The principal factor method implemented in the PROC FACTOR of SAS (SAS INSTITUTE, 1996) was the method used to extract the common factors. Factor loading matrix (**B**) was rotated using the VARIMAX procedure to enhance the interpretation of extracted factors.

In the PCA the values of the principal component Y_i for the *i-th* observation is a linear combination of the original variables X_j.

$$Y_i = \sum_j a_{ij} \cdot X_j \qquad \text{(for } j=1,p\text{)}$$

where a_{ij} are the component coefficients – eigenvectors corresponding to the *m* largest eigenvalues of the correlation (covariance) matrix of the *p* original variables. In the PCA the number of the components extracted is equal to the number of original variables. The number of PC retained is generally function of the proportion of variance explained by the first *m* PC *(m<p).*

The main difference between PC and FA is that PCA is a mere data transformation. PCA explores the maximum variability direction in the space of the variable, hence no distributional assumptions are required. On the other hand, the model underlying the FA required some distributional assumptions. The variance of the original variables is divided into variance explained by the presence of latent factor (common variance or communality) and unique variance associated to each specific original variable (KRZANOWSKY, 2000). Hence, the PCA explore the variance of the multivariate system and the FA the covariance, and both factors and PCs are orthogonal (uncorrelated).

Factors scores (F) were then treated as new variables and analysed with the following mixed model using the PROC MIXED of SAS (SAS INSTITUTE, 1996):

$$F_{ijklm} = DESIGN_i + MODEL_j + TRAIT_k + STUDY_l + e_{ijklm}$$

DESIGN*i* = fixed effect of class experimental design (3 levels)

$MODEL_j$ = fixed effect of class model (2 levels)

$TRAIT_k$ = fixed effect of trait analysed (6 level)

$STUDY_l$ = random effect of study (75 level) associated to covariance matrix $\mathbf{G}\sim(0, \mathbf{I}\cdot\sigma_g^2)$

e_{ijklm} = random residual associated covariance matrix $\mathbf{R}\sim(0, \mathbf{I}\cdot\sigma_e^2)$

Table 3. Code of the factor used in the mixed model.

Analytic Model	code	Traits	Code
Anova	1	MY	1
Comparison-wise linkage test	1	FY	2
Mixed model	1	FP	3
Single Marker Regression	1	PY	4
Monte Carlo Markov Chain	2	PP	5
Composite Interval Mapping	2	CT	6
L+LD mapping	2		
ML approach for QTL mapping	2	Experimental design	Code
Multi-marker Regression	2	DD	1
Rank-based non-parametric approach	2	GDD	2
Variance component QTL mapping	2	DD-POOL	3

MY=milk yield, FY=fat yield, FP=fat percentage; PY=protein yield, PP=protein percentage; CT=conformation trait; DD=daughter design, GGD=granddaughter design, DD-POOL= daughter design with DNA pooling

2.2.3 Analysis of distribution of estimated QTL effects for dairy cattle

Three theoretical distributions (Gamma, Lognormal and Weibull distribution). were used to model QTL effect distributions of milk yield (MY), protein yield (PY), protein percent (PP), fat yield (FY), fat percent (FP), conformation trait (CT) and all milk production traits scaled by genetic standard deviation (AT). All data were retrieved from published QTL mapping experiment and included in the analysis on the basis of significance level (p-value<0.05). All the records that reported the QTL effects were used. The goodness of fit of three distributions was assessed using Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests using PROC UNIVARIATE of SAS (SAS Institute, 1996). The null hypothesis tested (H_0) is that the distribution follows the gamma (lognormal or weibull) distribution. If H_0 is rejected the distribution used is not suitable to fit the data. Standardization of the QTL effects was carried out by dividing the estimated effects values by the genetic standard deviation, whenever available from publication. These information were retrieved from literature when not reported. In most cases the conformation traits were already expressed in unit of genetic standard deviation and no further standardizations were applied.

The density functions fitted to the experimental data are shown in table 4. Two parameters are common to three function: the threshold parameter (θ) and the width of histogram interval (h). The threshold parameter θ must be less than the minimum data value and could be set to 0. In the present study maximum likelihood estimate of θ was computed from the data.

Table 4. Density function fitted to the QTL estimated effects

		parameter		
Function	Density function fitted	Scale	Shape	Threshold
Gamma		σ	α	θ
Lognormal		σ	ζ	θ
Weibull		σ	c	Θ

The gamma distribution is a two-parameter family of continuous probability distributions. It has a scale parameter (σ) and a shape parameter (α). If α is an integer then the distribution represents the sum of α independent exponentially distributed random variables, each of which has a mean of α.

A log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed. If Y is a random variable with a normal distribution, then X = exp(Y) has a log-normal distribution; likewise, if X is log-normally distributed, then Y = log(X) is normally distributed. It is occasionally referred to as the Galton's distribution and its analytical description is given in the table 4.

The Weibull distribution is a continuous probability distribution. The probability density function of a Weibull random variable X is described in table 4. The shape σ parameter and c >0 is the scale parameter of the distribution. For values of c = 1 become an exponential distribution.

Goodness-of-Fit tests

The procedure computes test statistics for the null hypothesis that the values of the analysed variable are a random sample from the specified theoretical distribution. For a specified distribution, the procedure attempts to calculate three goodness-of-fit tests that are based on the empirical distribution function (EDF): the Kolmogorov-Smirnov D statistic, the Anderson-Darling statistic, and the Cramer-von Mises statistic. When the p-value is less than the predetermined critical value, the null hypothesis is rejected and conclude that the data did not come from the assumed theoretical distribution.

The computational formulas for the EDF statistics use the probability integral transformation . If is the distribution function of , the random variable is uniformly distributed between 0 and 1. Given *n* observations computes the values by applying the aforementioned integral transformation. These tests are based on various measures of the discrepancy between the empirical distribution function and the proposed cumulative distribution function

The Kolmogorov-Smirnov statistic (D) is defined as . This class of statistics is based on the largest vertical difference between and. . The Anderson-Darling statistic (A^2) and the Cramer-von Mises statistic (W^2) belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference . Anderson-Darling is calculated as – , whereas the Cramer-von Mises statistic is computed as follow —— —— (SAS INSTITUTE, 1996)

2.3. RESULTS AND DISCUSSION

2.3.1 QTLs Detected for economic traits in dairy cattle

The 150 articles collected were published between January 1995 and February 2008 and contained 2,651 records which were included in the data base. The number of publications and, consequently, of reported QTLs has increased through the years (Figure 3). The number of records experienced a drop in 2008 both because articles were collected until February 2008. A further reason has been the decreasing number of QTL mapping studies carried out using microsatellite markers caused by new advances in SNP chip technology that allowed to use bi-allelic SNP, mainly with the aim to estimate genomic breeding values of the animal. GEBV estimation may be done without specific knowledge of QTL size or position.

Research studies have been carried out mainly on Holstein cattle (HF) (about 77% of the QTL records) followed by Brown Swiss (BR), Ayrshire (AYR), Norwegian Red cattle (NRC) and other minor breeds including Jersey (JER), Fleckvieh (FLE) and Swedish red and white (SRW) (Figure 4). This fact allows for possible comparisons of QTL effects among different breeds or populations.

Daughter and granddaughter designs (WELLER *et al.* 1990) were basically the experimental designs used, the former being the most frequent probably due to the greater power that could be achieved in comparison to the cost of the experiment (Figure 5). Selective DNA pooling (5.7%) and selective genotyping have been also used (0.6%). Although the last two designs allow to have a further reduction of the cost of the experiment, they were not widely used. They have been used especially in experiments carried out in Italy and Israel (COBANOGLU *et al.* 2005; LIPKIN *et al.* 1998; MOSIG *et al.* 2001; AJMONE-MARSAN *et al.* 2007; BAGNATO *et al.* 2008).

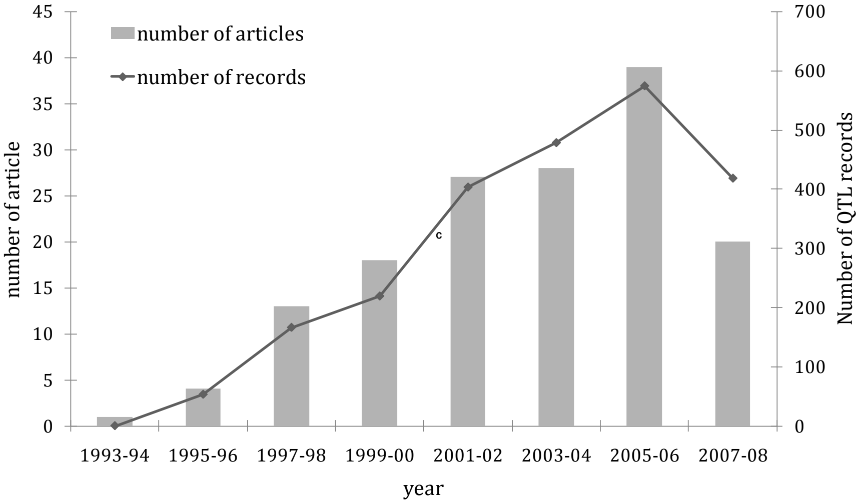


Figure 3. Number of article of QTL mapping studies and QTL retrieved from 1993 to 2008.

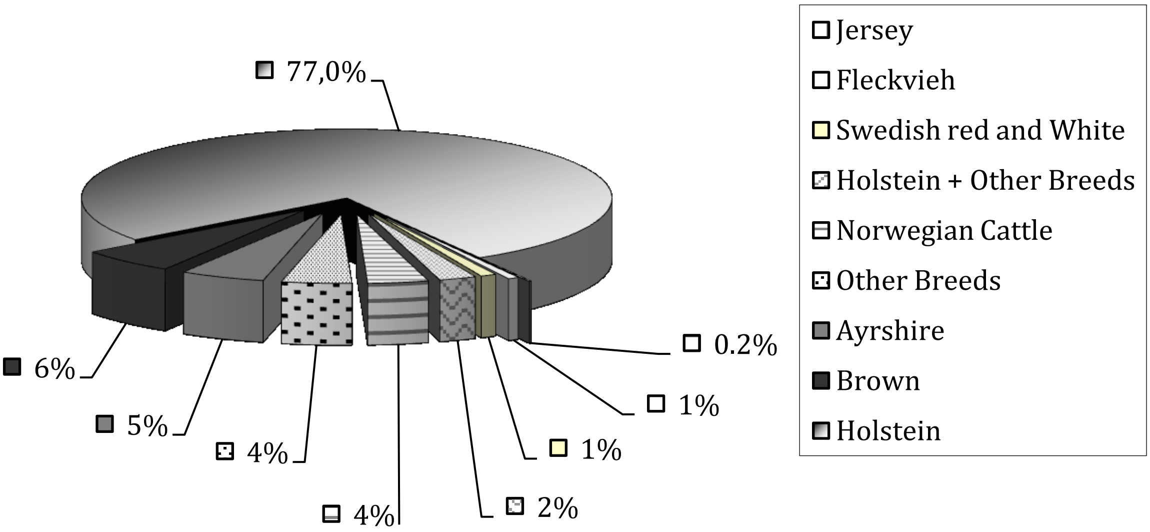


Figure 4. Proportion of dairy cattle breeds used in experiment of QTL-mapping.

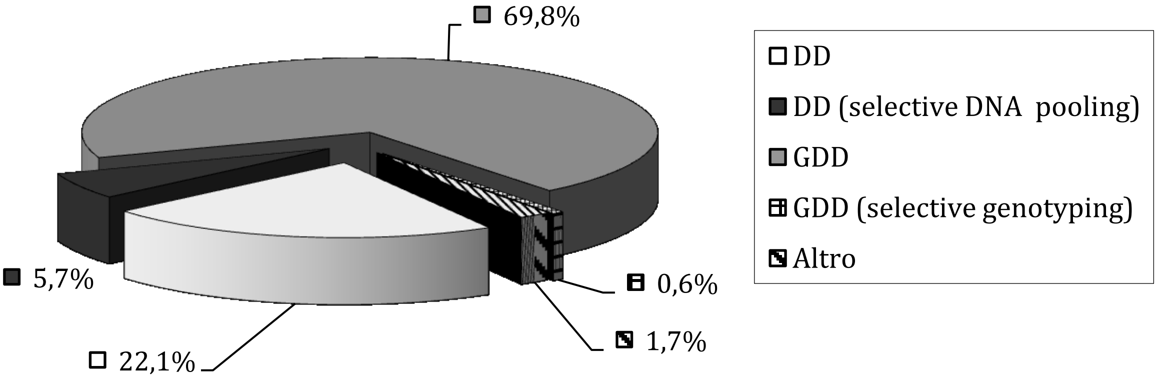


Figure 5. Proportion of different experimental design on QTL mapping studies

In the specific case of outbreed populations, the statistical models used to map QTLs are quite different and they depend also on the structure of the resource population. The analytic models used and their frequency are summarized in table 5. Multiple marker regression (KNOTT *et al.* 1996) for interval mapping was the most common technique (nearly 60% of the records). Comparison-wise linkage tests carried out at individual sire-by-marker level across families were used in about 10% of the cases. Single marker regressions and ANOVA summed up to 11%. The remaining analytical methods used were about 10% of the researches. The most frequently used response variable were Daughter Yield Deviations (DYD) followed by EBVs, Predicted transmitting ability (PTA) and de-regressed proofs (DRPF) (figure 6).

Table 5. Analytical model used in the experiment of QTL mapping

Analitic Model	% records
Multi-Marker Regression (Interval mapping)	57.1
Comparison-wise linkage test	9.3
Single Marker Regression	5.7
Anova	5.6
ML approach for QTL mapping	5.1
Composite Interval Mapping	2.9
Variance component QTL mapping	2.8
Mixed Model	2.4
L+LD mapping	2.4
Monte Carlo Markov Chain	2.2
Rank-based non parametric approach	1.1
Other	3.2

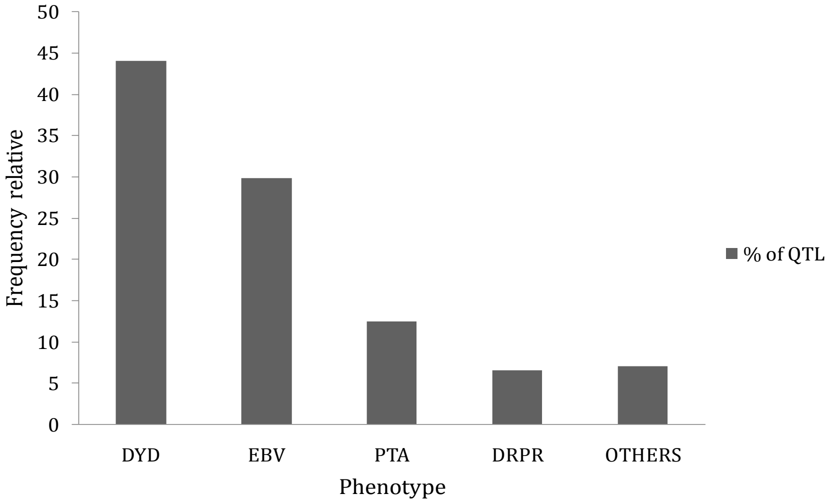


Figure 6. Response variables used in QTL mapping studies (DYD=daughter yield deviation; EBV=estimated breeding value; PTA=predicted transmitting ability; DRPR=de-regressed proof)

The average number of genotyped animals per study was 1,181, ranging from 50 to 5,470 depending on the experimental design used. Most of studies dealt with production traits (70% of the records) (Figure 7), even though a certain occurrence of studies on traits that are becoming of importance as breeding goals such as milk fatty acid (MFA) composition, disease resistance, longevity and lactation persistency has been highlighted. The highest number of QTLs was detected for protein percentage, followed by milk yield, protein yield and milk fat content. This fact underlie that the breeders put emphasis on these traits, still being the key breeding goals. Nonetheless, a major interest for functional trait is still growing and although few study dealt with conformation traits, the number of QTL detected was quite high.

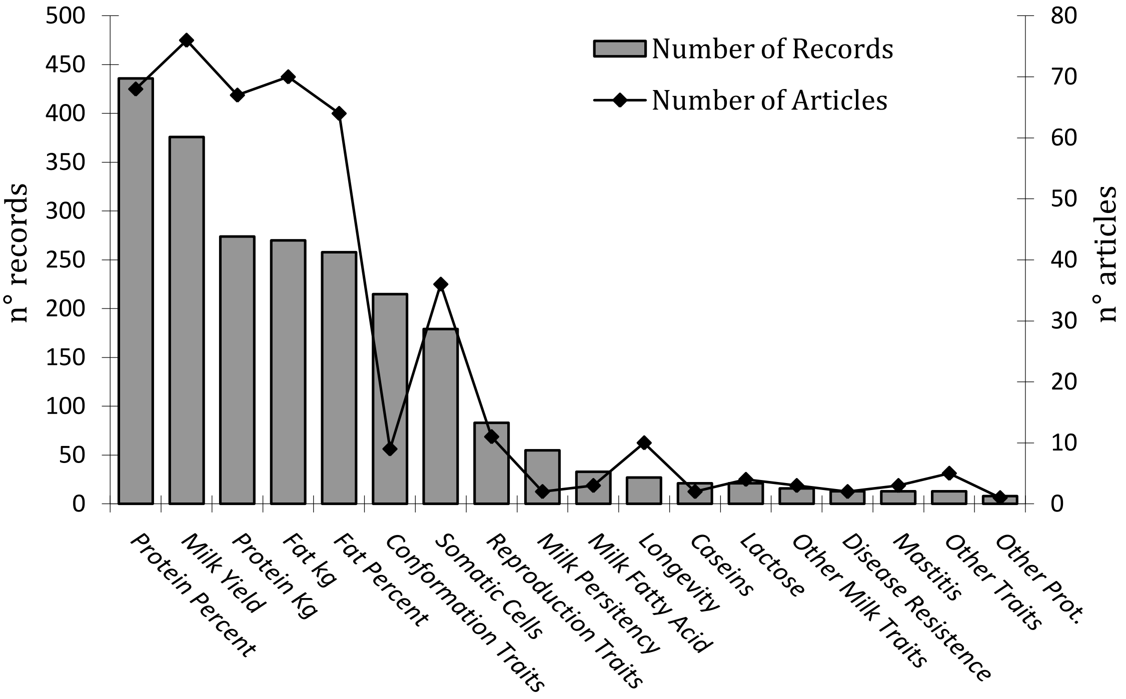


Figure 7. Number of publication and Record numbers grouped by traits analyzed.

Figure 8 shows the distribution of the QTL detected for milk production and quality traits by chromosome (*Bos Taurus autosome* (BTA)). It must be pointed out how just 6 chromosome (BTA3-6-7-14-20-26) cover about 60% of the total number of QTL records. In particular, BTA6 and BTA14 harbor the major number of QTLs for milk yield and composition: BTA6 for milk yield and protein content, and BTA14 for fat percentage, milk yield and protein content. In fact, the role of these genomic regions on the determinism of the milk protein content and fat content is well established. Casein cluster on BTA6 and DGAT1 on BTA14 respectively (COPPIETERS *et al.* 1998b; GRISART *et al.* 2002). Effect on milk yield have been also found on BTA20. Actually, GH receptor (BLOTT *et al.* 2003) and PRL receptor (VIITALA *et al.* 2006) have been mapped on this chromosome. The distribution of QTL affecting conformation traits is quite regular across the genome (figure 9) as well as for the functional traits (data not shown).

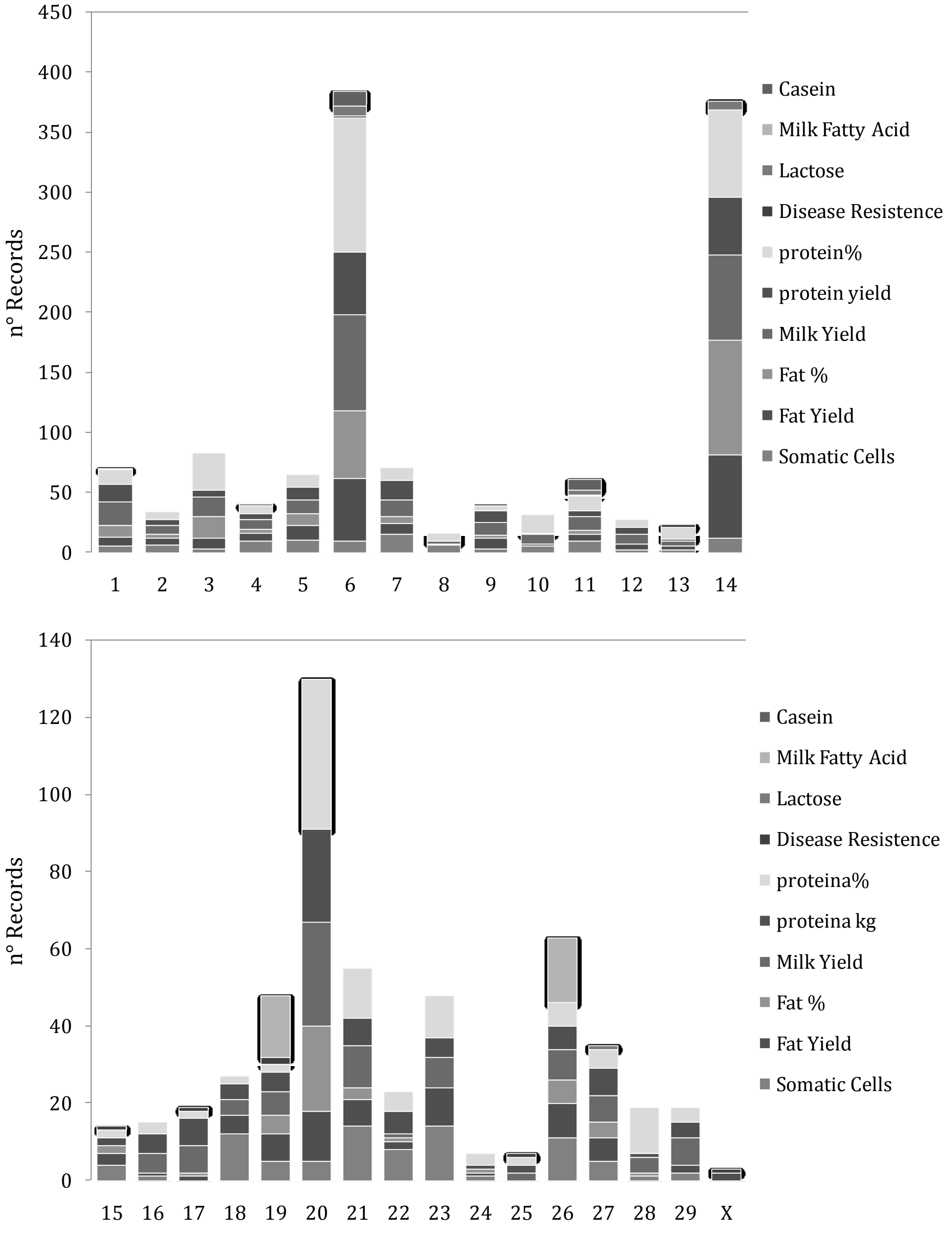


Figure 8. Distribution of number of QTL divided by trait records per chromosome

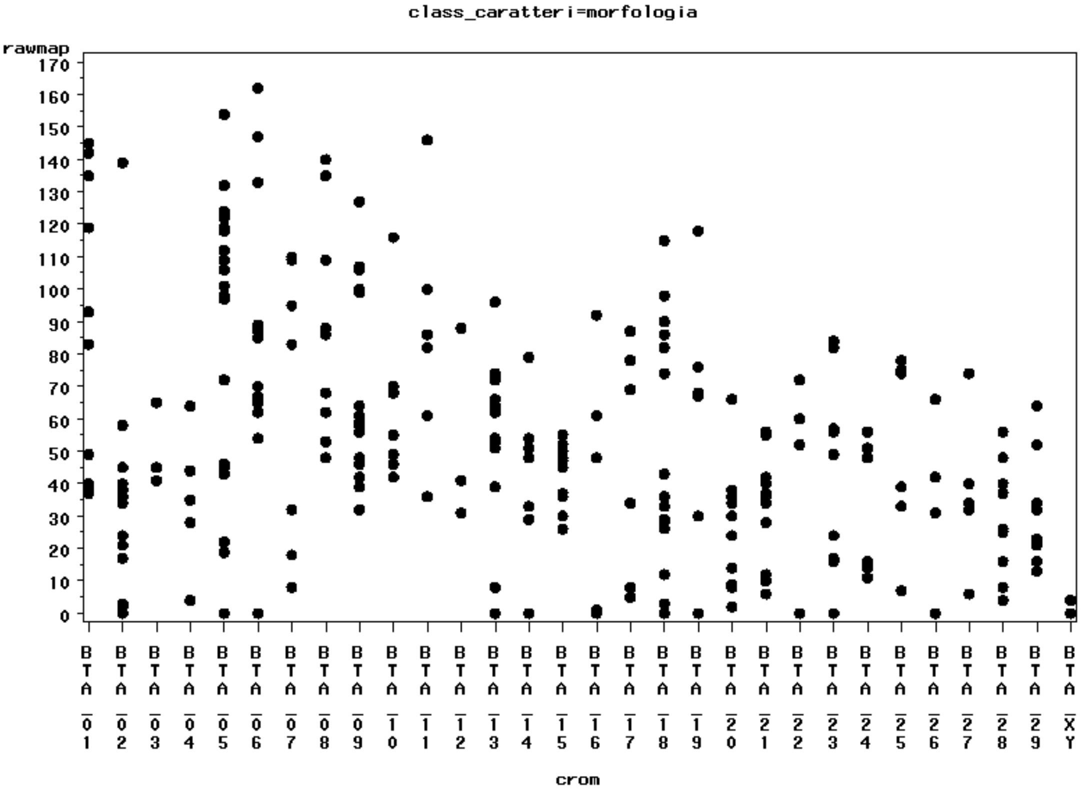


Figure 9. Position (cM) of QTL for conformation traits found across different studies along the genome

Table 6 reports the number of QTLs detected across traits according to the significance level. It is worth to notice how the most frequent class is that one which groups QTL with P-values less than 0.01, and about 76% of QTL have been detected with a significance level <0.05. However, a suggestive linkage has been reported for 15% of the database. Moreover, a certain number of QTL are not significantly associated with a phenotypic trait.

Table 6. Distribution of QTL significance level on 4 class of P-value

Class	P-value	nQTL
1	<0.01	923
2	0.01-0.05	663
3	0.05-0.1	322
4	> 0.1	169

Figure 10 reports for each position the significance level of QTL detected across studies and traits. It can be observed that QTLs found in different studies tend to be closer on the chromosomes, often to overlap each other. This result suggests that different study may report the same QTL, even though with slightly different positions. However, it must be pointed out that confidence intervals of QTL location, when reported in the paper, are quite large (24.2 cM, ranging from 0.2 to 150 cM). Actually, values of confidence intervals of QTLs position higher than 20 cM are extremely large for any efficient application of MAS (KHATKAR *et al.* 2004).

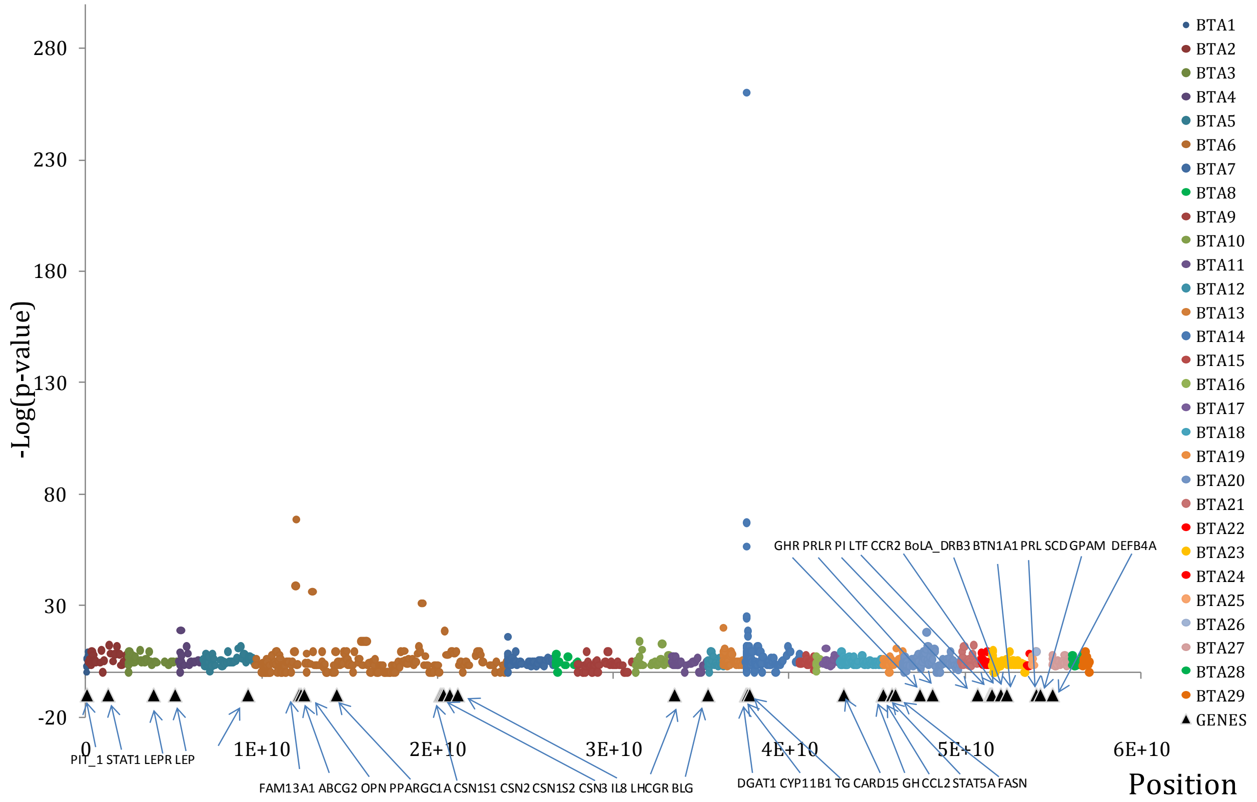


Figure 10. P-values of QTL detected across different studies and traits against position and most significant gene and polymorphism affecting dairy traits.

It is remarkable to notice (figure 10) the role of DGAT1 (BTA14) with the lowest p-value followed by FAM13A1, ABCG2 and OPN genes and casein cluster genes on BTA6. LEP on BTA3 showed a low p-values as well as the GHR and PLRL on BTA20. More details about these genes will be provided later, but this picture show how the meta-analytical approach, even though merely descriptive, may give a picture of the regions that affect phenotypic traits, using gene whose effect is known as "bookmark" (similar plots, are quite common using genome-wide approach to detect QTLs with dense maps).

Finally, the frequency distribution of QTL effects is reported in figure 11. For sake of simplicity the distributions for fat percentage, protein yield, milk yield and fat yield were reported. The distributions of QTL effects estimated in the collected studies look different according to the trait analyzed. The distributions of fat percentage and protein yield (figures 11a and 11b respectively) look like as there would be many QTLs of small effect and just few QTL with large ones. Figures 11c and 11d show the distribution of milk and fat yield respectively. These distributions might approximate the normal distribution.

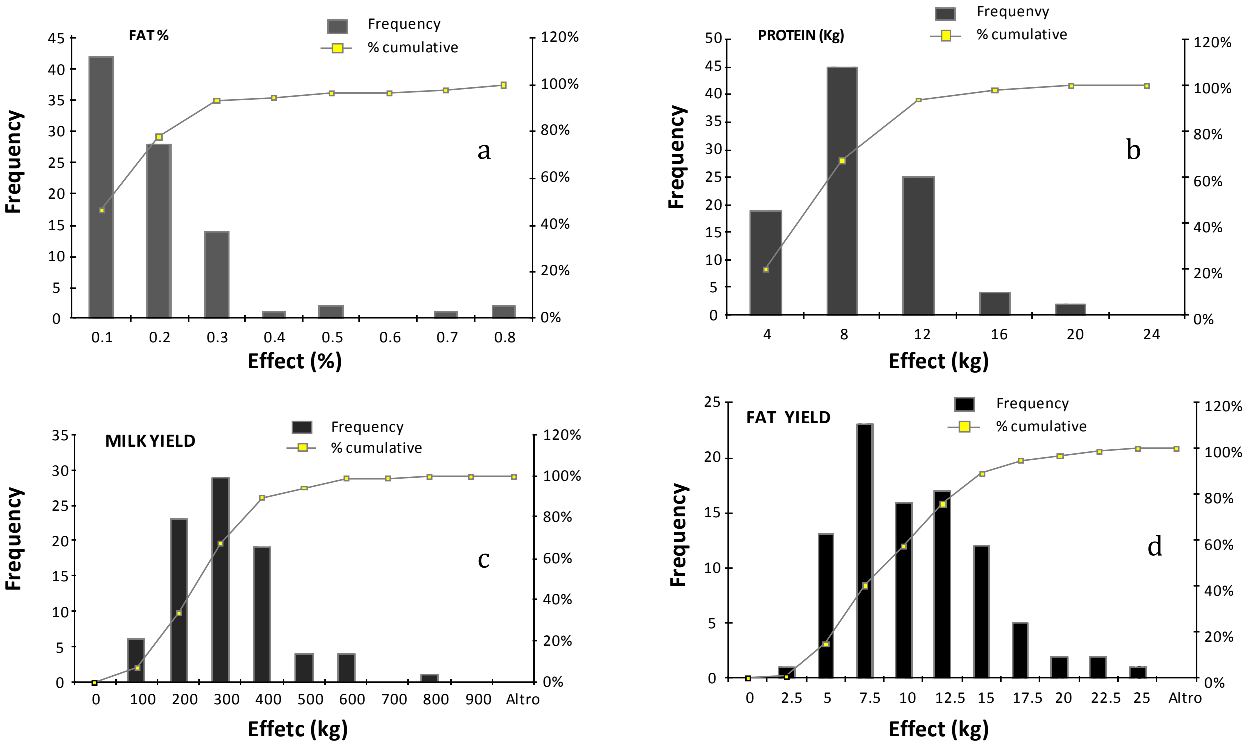


Figure 11. Frequency histogram of absolute values of the effects of QTL retrieved from published paper for fat percentage (a), protein yield (b), milk yield (c) and fat yield (d).

Comparison of QTL detected across different studies

In figure 11 are reported the QTLs detected for most significant chromosomes for five milk production and quality traist (MY, FY, FP, PY, PP) and known gene polymorphisms affecting productive and functional traits in cattle. In appendix an exhaustive review of all QTLs for dairy traits is provided, including more detail about position and significance level. The positions are reported in million of bp (Mb). In chromosome 1 (figure 12a) 34 QTLs have been found from 12 authors across three breeds, being Holstein Friesian most represented. No significant QTLs were found for FP.

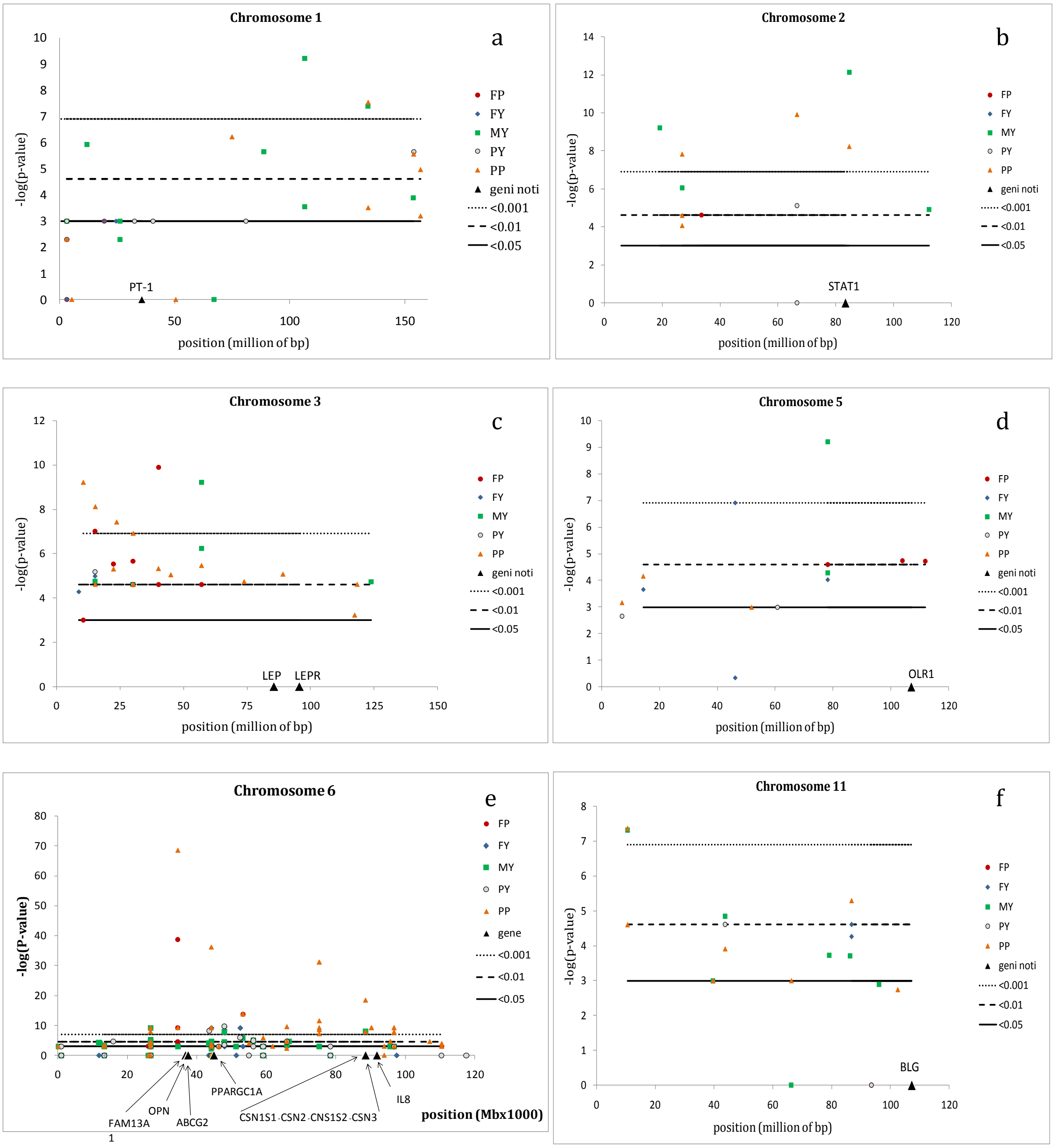


Figure 12. Significant QTL for 5 milk production traits [FP=fat percentage (red circle), FY=fat yield (blue diamond), MY=milk yield (green square), PY=protein Yield (open circle), PP=protein percentage (yellow triangle)] retrieved from published paper for most significant chromosome across the genome. The solid, dotted and dashed lines represent the significance threshold for p-value of 0.05 0.01 and 0.001 respectively. Black triangle on the *x* axis represent the known polymorphism affecting the trait.

Three QTLs out of 23 are highly significant (p-values<0.001) for milk yield (106 Mb, 134 Mb) and protein percentage (134 Mb) detected from (VIITALA *et al.* 2003) and (BAGNATO *et al.* 2008), in Finnish AYR and BR respectively. In the genomic region >100 Mb have been found 10 QTLs affecting MY, PY and PP under the p-values threshold of 0.05: 5 QTL on PP (MOSIG *et al.* 2001; BAGNATO *et al.* 2008) on BR and HF, 4 QTLs affecting MY on BR and AYR (BAGNATO *et al.* 2008; VIITALA *et al.* 2003) and 1 QTL on PY on HF (HEYEN *et al.* 1999). Around the PT-1 gene (a transcription factor acting in the pituitary gland controlling the transcription of growth hormone (GH), prolactin (PRL) genes and Thyroid stimulating Hormone) are present 6 different QTLs affecting mainly MY and PY. A polymorphism in the 3 exon of this gene have been associated to MY and productive live in Holstein (HUANG *et al.* 2008) and growth trait in Qinchuan cattle (ZHANG* *et al.* 2009). In chromosome 1 it possible to observe to QTL region that affect milk traits across different breed.

In the Chromosome 2, twelve QTLs have been reported from 6 studies (figure 12b). Two main chromosomal regions affecting MY and PP have been observed at around 27 and 85 Mb. Five QTLs are under the 0.05 significance threshold in the 27 Mb region: QTL for MY (VIITALA *et al.* 2003; BAGNATO *et al.* 2008) on BR and AYR; 3 QTLs on PP (HEYEN *et al.* 1999; ASHWELL *et al.* 2004; BAGNATO *et al.* 2008) in HF and BR; 1 QTL affecting FP in HF (ASHWELL *et al.* 2004). For positions greater than 70 Mb, five QTLs have been found in three different breed (HF,BR and NRC) in three different research studies. Interestingly the QTL affecting MY and PP at position 84 Mb (BAGNATO *et al.* 2008) and PP (MOSIG *et al.* 2001), are quite close to STAT1 gene (gene that regulates the transcription of some other genes involved in milk protein metabolism) which were associated with significant increases in milk, fat, and protein yields (COBANOGLU *et al.* 2006).

Figure 12c show QTLs detected on chromosome 3 for milk production trait. Thirty QTLs from 21 authors were reported on the graph. Seven QTLs affecting FP (HF, AYR), 2 FY (HF, NRC), 5 MY (HF, BR, AYR), 14 PY (HF, BR, AYR) and 2 PP (HF) were found. Seven QTLs overcome the 0.001 significance threshold: 2 for MY (VIITALA *et al.* 2003) 4 for PP (HEYEN *et al.* 1999; ASHWELL *et al.* 2001; VIITALA *et al.* 2003) and 1 for FP (HEYEN *et al.* 1999). It is interesting to notice that the high concentration of QTLs for PP in the range 10-25 Mb 5 QTL significant were found in 5 different studies (HEYEN *et al.* 1999; PLANTE *et al.* 2001; ASHWELL *et al.* 2001; BOICHARD *et al.* 2003; VIITALA *et al.* 2003). Furthermore a possible presence of one or more pleiotropic QTLs affecting MY PP and FP can be hypothesized in the region spanning between 15 Mb and 57 Mb. QTLs affecting three milk production traits have been detected by the same author in the same resource population around 15 Mb. Moreover, three QTL affecting MY PP and FP have also been identified in three different studies at around 57 Mb. Far from this region map the leptin (LEP) and leptin receptor (LEPR) genes (about

80 and 90 Mb respectively) which have been associated to milk production and feed intake (BANOS *et al.* 2008), even though the most common effect of these genes were found for growth trait in beef cattle (GUO *et al.* 2008; KULIG AND KMIEĆ, 2008 Three QTLs out of 23 are highly significant (p-values<0.001) for milk yield (106 Mb, 134 Mb) and protein percentage (134 Mb) detected from (VIITALA *et al.* 2003) and (BAGNATO *et al.* 2008) in Finnish AYR and BR, respectively. In the genomic region spanning beyond100 Mb, 10 QTLs affecting MY, PY and PP under the p-values threshold of 0.05 have been found: 5 QTL on PP (MOSIG *et al.* 2001; BAGNATO *et al.* 2008) on BR and HF, 4 QTLs affecting MY on BR and AYR (BAGNATO *et al.* 2008; VIITALA *et al.* 2003) and 1 QTL on PY on HF (HEYEN *et al.* 1999). Around the PT-1 gene (a transcription factor acting in the pituitary gland controlling the transcription of growth hormone (GH), prolactin (PRL) genes and Thyroid stimulating Hormone) are present 6 different QTLs affecting mainly MY and PY. A polymorphism in the thid exon of this gene has been associated to MY and productive live in Holstein (HUANG *et al.* 2008) and growth trait in Qinchuan cattle (ZHANG* *et al.* 2009). In chromosome 1 it possible to observe to a QTL region that affect milk traits across different breed.

In the Chromosome 2, twelve QTLs have been reported from 6 studies (figure 12b). Two main chromosomal regions affecting MY and PP have been highlighted at around 27 and 85 Mb respectively. Five QTLs are under the 0.05 significance threshold in the 27 Mb region: QTL for MY (VIITALA *et al.* 2003; BAGNATO *et al.* 2008) on BR and AYR; 3 QTLs on PP (HEYEN *et al.* 1999; ASHWELL *et al.* 2004; BAGNATO *et al.* 2008) in HF and BR; 1 QTL affecting FP in HF (ASHWELL *et al.* 2004). For positions greater than 70 Mb, five QTLs have been found in three different breed (HF,BR and NRC) in three different research studies. Interestingly, the QTL affecting MY and PP at position 84 Mb (BAGNATO *et al.* 2008) and PP (MOSIG *et al.* 2001) are quite close to STAT1 gene (gene that regulates the transcription of some other genes involved in milk protein metabolism) that has been found to be associated with significant increases in milk, fat, and protein yields (COBANOGLU *et al.* 2006).

Figure 12c shows QTLs detected on chromosome 3 for milk production trait. Thirty QTLs from 21 authors were reported on the graph. Seven QTLs affecting FP (HF, AYR), 2 FY (HF, NRC), 5 MY (HF, BR, AYR), 14 PY (HF, BR, AYR) and 2 PP (HF) were found. Seven QTLs overcome the 0.001 significance threshold: 2 for MY (VIITALA *et al.* 2003) 4 for PP (HEYEN *et al.* 1999; ASHWELL *et al.* 2001; VIITALA *et al.* 2003) and 1 for FP (HEYEN *et al.* 1999). It is interesting to notice that the high concentration of QTLs for PP can be found in the range 10-25 Mb. Five QTL were found in 5 different studies (HEYEN *et al.* 1999; PLANTE *et al.* 2001; ASHWELL *et al.* 2001; BOICHARD *et al.* 2003; VIITALA *et al.* 2003). Furthermore a possible presence of one or more pleiotropic QTLs affecting MY PP and FP can be hypothesized in the region spanning between 15 Mb and 57 Mb. QTLs affecting three milk production traits have been detected by the same author in the same resource population at around 15 Mb. Moreover, three QTL affecting MY PP and FP have also been identified in three different studies at around 57 Mb. Far from this region map the leptin (LEP) and leptin receptor (LEPR) genes (about 80 and 90 Mb respectively) which have been associated to milk production and feed intake (BANOS *et al.* 2008), even though the most common effect of these genes were found for growth trait in beef cattle (GUO *et al.* 2008; KULIG AND KMIEĆ, 2008).

In the Figure 12e, QTLs detected on chromosome 6 are reported. A total of 192 QTL were retrieved from 21 different authors. This chromosome is one of the most investigated, followed by BTA14 and BTA20. QTLs for all five milk production traits have been found. In particular: 25, 36, 45, 54 and 32 QTLs were identified for FP, FY, MY, PP and PY respectively. Thirty-eight QTLs were under the significance threshold of 0.001. In particular: 5 QTLs on FP detected by OLSEN *et al.* (2004) in NRC; 4 QTLs for FY both in HF and NRC (KUHN *et al.* 1999; OLSEN *et al.* 2004; SZYDA *et al.* 2005); 3 QTLs for MY in BR, HF and AYR identified by (VIITALA *et al.* 2003; SZYDA *et al.* 2005; BAGNATO *et al.* 2008) respectively. However, PP shows the highest number of QTLs detected (18 QTLs) on HF, BR, NRC and AYR (SPELMAN *et al.* 1996; ASHWELL and VAN TASSELL 1999; ASHWELL *et al.* 2001; MOSIG *et al.* 2001; OLSEN *et al.* 2002; VIITALA *et al.* 2003; OLSEN *et al.* 2004; BAGNATO *et al.* 2008;). Finally, two QTLs were detected for PY by (OLSEN *et al.* 2004; SZYDA *et al.* 2005). The lowest P-values were found for a QTL region in LD with four genes (FAM13A,OPN, ABCG2 PPARGC1A) located at around 40 Mb and for the casein cluster at position 88 Mb. All of these genes showed polymorphisms associated with milk protein production. In particular FAM13A1 (COHEN *et al.* 2004), a bovine gene close to a cluster of genes coding for proteins of the extracellular matrix, is revealed to be in LD with some QTLs as it affected the milk protein production. ABCG2 and OPN are genes very close each other that have been found to be associated with milk protein yield (SCHNABEL *et al.* 2005; RON *et al.* 2006; SHEEHY *et al.* 2009). OPN have been also found to be related to mastitis resistance (ALAIN *et al.* 2009). These two genes have been also indicated as conflicting QTN for milk protein content by DE KONING (2006). The whole casein cluster – αs1-casein (CSN1S1), αs2-casein (CSN1S2), β-casein (CSN2) and κ-casein (CSN3) – is a well known region that influence the quantity of milk protein. Several protein variants have been characterized in dairy cattle. Novel polymorphisms have been recently associated to difference in milk protein for CSN1S1 in German cattle (KUSS *et al.* 2005). Polymorphisms in CSN1S2 gene have been associated to difference in milk yield traits in German fleckvieh (BRAUNSCHWEIG, 2008). Furthermore, novel polymorphisms (CSN2) affecting milk production traits in NRC (NILSEN, 2009), and concentration of milk protein variants (HALLÉN *et al.* 2008) have been found. Polymorphism on CSN3 gene have also been associated to concentration of milk protein variants (HALLÉN *et al.* 2008) .

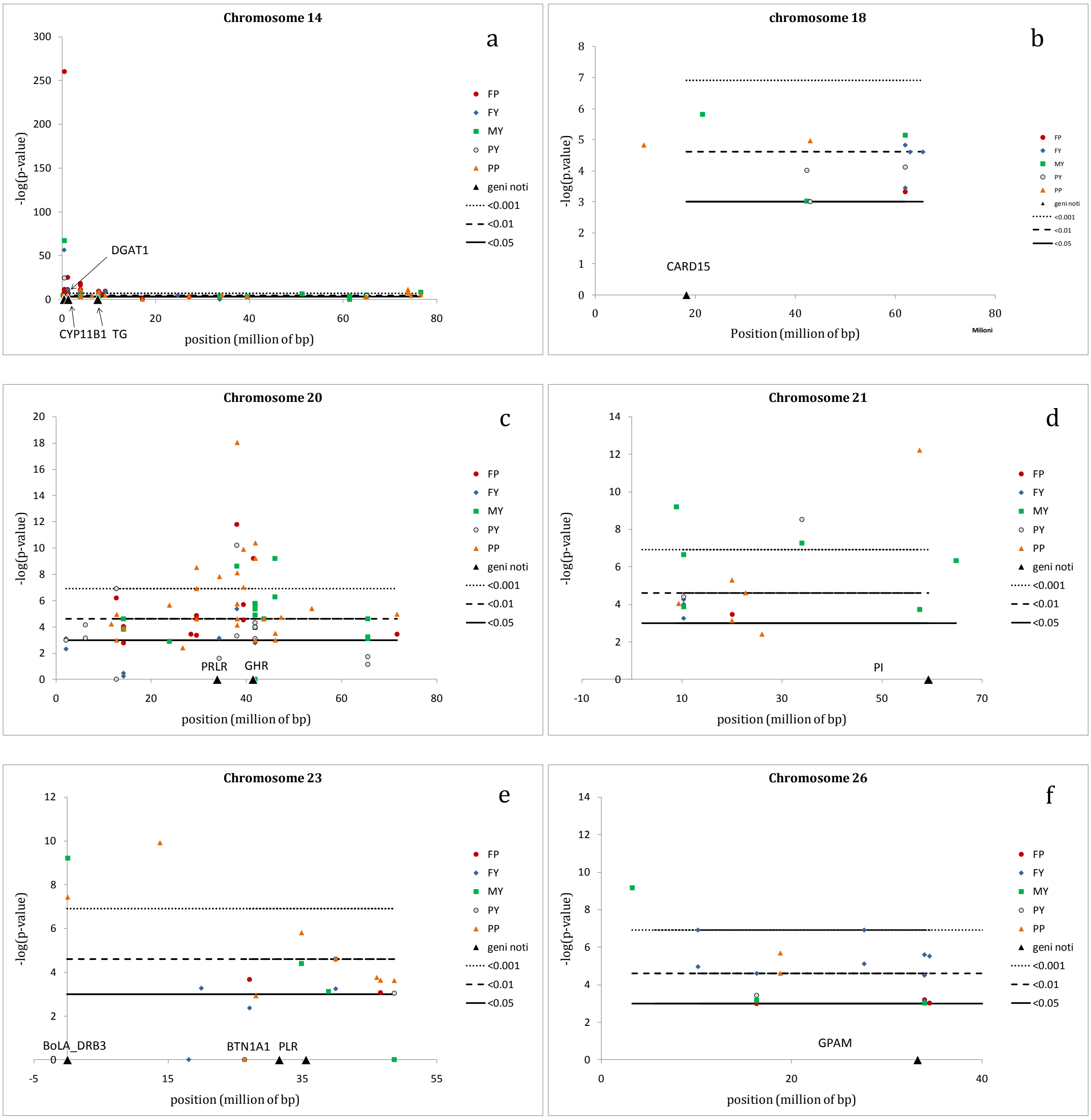


Figure 13. Significant QTL for 5 milk production traits (FP=fat percentage, FY=fat yield, MY=milk yield, PY=protein yield, PP=protein percentage) retrieved from published paper for most significant chromosome across the genome. The solid, dotted and dashed lines represent the significance threshold for p-value of 0.05 0.01 and 0.001 respectively.

Both chromosomes 14 and chromosome 20 showed a high number of QTLs detected (figure 13a, figure 13c respectively).

Chromosome 14 (figure 13a) harbors 148 QTLs detected in 21 studies: 69 QTLs for FP, 22 QTLs for FY, 22 QTLs for MY, 27 QTLs affecting PP and 8 QTLs influencing PY. Mostly, the QTLs on this chromosome influence the milk fat content, but also the protein content and milk yield. Association with other production and functional traits is provided in appendix I.

The majority of significant QTLs were located in the centromeric region spanning around 10 cM. At least 15 studies reported QTL for fat percentage in a region of 10 cM from centromere in HF, FLE and NOR (GEORGES *et al.* 1995; COPPIETERS *et al.* 1998a; COPPIETERS *et al.* 1998b; RON *et al*., 1998; ASHWELL *et al.* 2001; HEYEN *et al.* 1999; RIQUET *et al.* 1999; KIM and GEORGES 2002; WINTER *et al.* 2002; BOICHARD *et al.* 2003; THALLER *et al.* 2003; VIITALA *et al.* 2003; ASHWELL *et al.* 2004; KUHN *et al.* 2004; FONTANESI *et al.* 2005). Moreover, 8 significant QTLs influencing FY have been detected in the same genomic region (HEYEN *et al.* 1999; ASHWELL *et al.* 2001; LOOFT *et al.* 2001; KIM and GEORGES 2002; VIITALA *et al.* 2003; ASHWELL *et al.* 2004; BENNEWITZ *et al.* 2004a). Six QTLs for MY and 8 QTLs for PP were also detected in a10 cM region. This genomic region harbors one gene that heavily affects the milk fat content and milk traits in general. GRISART *et al.* (2002) refined the position of this QTL to a 3 cM chromosome interval bracketed by two microsatellite markers BULGE13 and BULGE09.They identified a strong candidate gene, Diacyl Glycerol Acyl Transferase (DGAT1) and a non-conservative lysine to alanine (K232A) substitution which showed an effect on milk fat content and other milk traits. Moreover, they report that DGAT1 explained about 50% of the phenotypic variance for fat percentage. The same authors gave further evidence of DGAT1 as QTN. GRISART *et al.* (2004) have expressed both DGAT1 alleles in Sf9 cells line by using in vitro assay to evaluate level of expression of K allele of DGAT1. They have shown that the K allele is characterized by a higher Vmax of the enzyme in producing triglycerides than the A allele. Moreover, SCHENNINK *et al.* (2007) and SCHENNINK *et al.* (2008) found that DGAT1 K232A polymorphism has a clear influence on milk-fat composition. K Allele is associated with more saturated fatty acid, a larger fraction of C16:0; and a smaller fractions of C14:0, unsaturated C18 and CLA.

Although the aforementioned results demonstrate how the genetic determinism of fat synthesis (milk fat content and composition) is largely explained by one gene, some authors suggested (BENNEWITZ *et al.* 2003) additional sources of genetic variance on this chromosome for milk fat content. Basically, the hypotheses explored were *i)* the presence of one or more DGAT1 additional alleles segregating in cattle population that were not previously identified *ii)* a second quantitative trait locus affecting these traits *iii)* or both hypothesis. KUHN *et al.* (2004) showed that alleles of the DGAT1 promoter in the 5' non-coding region derived from the variable number of tandem repeats

(VNTR) polymorphisms influence milk fat content in animals homozygous for the allele 232A at DGAT1. This promoter VNTR polymorphism influences the number of potential Sp1 binding sites (site that harbor DNA sequence to bind transcription factor) and therefore might regulate DGAT1 expression and also milk fat content (FURBASS *et al.* 2006). In addition, KAUPE *et al.* (2007) detected a dominant mode of effects for the DGAT1 K232A and promoter VNTR alleles.

Hypotheses that not involve VNTR polymorphism have been proposed by GAUTIER *et al.* (2007) that gave a further evidence of the potential presence of other genes underlying the milk fat traits by studying Normande, French Holstein and Montebeliarde breeds. In their researches, VNTR polymorphisms explained only a small fraction of the variance of the QTL for fat percentage in the 3 breeds simultaneously after correction for the effect of the K232A polymorphism. Therefore, their results suggest the existence of at least one other causative polymorphism not yet described. KAUPE *et al.* (2007) carried out a joint analysis of DGAT1 and another neighbor gene CYP11B1. KAUPE *et al.* (2007) found that CYP11B1 and DGAT1 together explained more of the variation in milk production traits than DGAT1 alone. Further analyses of segregation and characterization of DGAT1 in cattle population across different breeds in different countries were reported by BANOS *et al.* (2008), CONTE *et al.* (2010), LACORTE *et al.* (2006), PAREEK *et al.* (2005) and SANDERS *et al.* (2006).

Figure 13c shows 79 QTLs for milk production traits detected on BTA20 in 13 studies (GEORGES *et al.* 1995; ARRANZ *et al.* 1998; ASHWELL *et al.* 2001; MOSIG *et al.* 2001; PLANTE *et al.* 2001; OLSEN *et al.* 2002; BLOTT *et al.* 2003; ASHWELL *et al.* 2004; VIITALA *et al.* 2006; BAGNATO *et al.* 2008). The detected QTLs were 14, 8, 17, 27 and 13 for FP, FY, MY, PP and PP respectively. Thirteen QTLs were found to be highly significant (P-value <0.001) and spanning from 29 Mb to 46 Mb region and affect mostly PP (in HF and AYR) and secondly MY, FP and PY.

One or more QTLs with pleiotropic effects seems to be recognizable in the genomic region described above. Indeed, in that region have been identified two polymorphism in the GHs Receptor (F279Y substitution)(BLOTT *et al.* 2003) and PL receptor (VIITALA *et al.* 2006) genes that are heavily involved in the milk synthesis process. Moreover, 18, 24 and 23 QTLs affecting milk production traits where also detected on chromosomes 21, 23 and 26 respectively. In particular, the PRL gene that has been mapped on chromosome 26 is a suitable candidate to explain part of genetic variance of this chromosome (BRYM *et al.* 2005; SCHENNINK *et al.* 2009; LÜ *et al.* 2010). Furthermore, the A239V substitution in SCD gene has been associated to a greater content of *cis*-9 C18:1 (AA genotype) and total monounsaturated fatty acids and a higher C14:1/C14 ratio in comparison to VV genotype (MELE *et al.* 2007). Furthermore the same polymorphism in SCD has been associated to a higher level of milk and protein yield (MACCIOTTA *et al.* 2008).

QTLs found across different studies and breeds may used to enforce the evidence of presence of a region that affect a quantitative trait. However, the key question is: “QTLs detected in different studies are or not the same?”. Using a visual inspection of dataset is not possible to answer to this question. Although analytic tools need to be used to compare statistically QTLs found in different studies, a graphic may be still useful to evaluate in a qualitative way the amount of QTL and the genomic region most explored and drawn preliminary conclusion. The result of a multivariate meta-analytic techniques have been reporter below.

2.3.2 A multivariate Meta-analysis approach

The results of a preliminary exploratory data analysis are reported in figure 14 and table 7. The table 7 shows the Pearson and partial correlation matrices among all 18 variables retrieved from published articles and public databases. Some of the variables collected are highly correlated, but the majority shows low to moderate Pearson correlations. Looking at the partial correlation, these are in most of the cases lower than person correlations but not systematically. The idea is that there may be some latent structures of the data , even if not so clearly identified. Indeed, the MSA is not very high (0.53). Figure 14 shown the result of preliminary eigenvalue extraction on the whole dataset for the selection of a subset of variable which run the analysis with. The optimal number of factors retained where carried out on the basis of proportion of variance explained and min-eingen criterion (retain only the eigenvalues greater than one) indicate that 6 factor satisfied both criterion (at least 70% of variance explained).

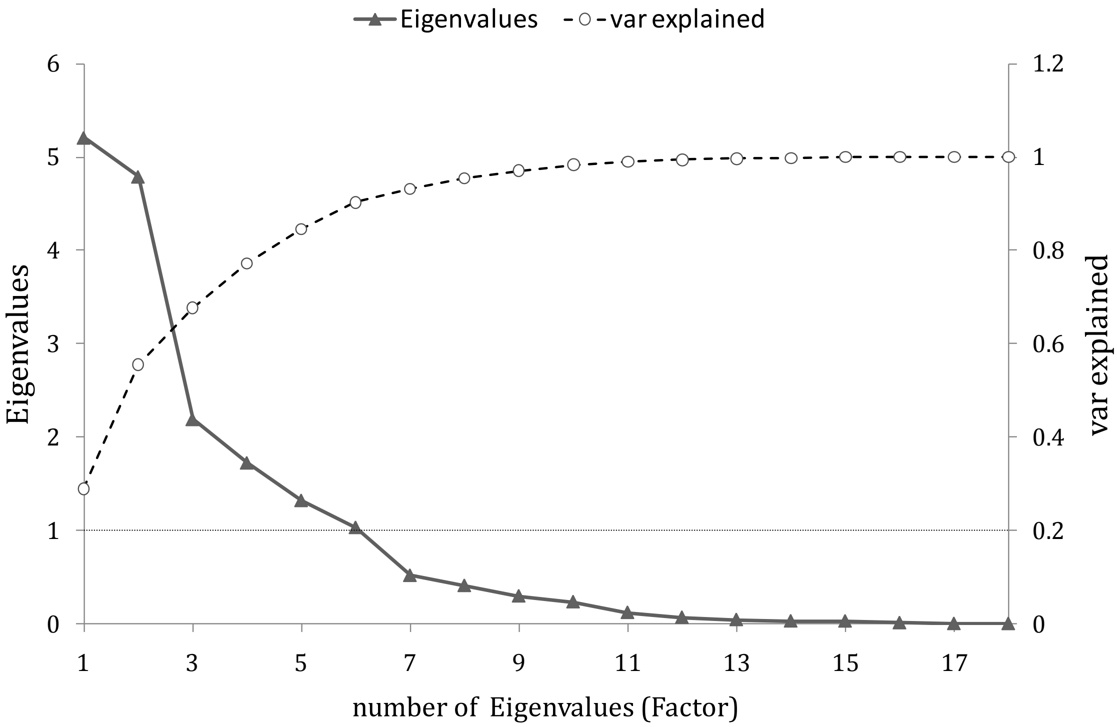


Figure 14. Scree plot of eigenvalue of FA carried out to preselect the variable (6 factor were retained)

Figure 15 and figure 16 report the factor loadings, i.e. the correlations among common factors and original variables. The criterion used to choose a subset of variables was to retain those that saturated all the six factor (factor loading at least greater than 0.70) excluding the redundant variables.

Table 7. Pearson correlation(above the diagonal) and Partial correlation (below the diagonal) among the variable used in FA.(in bold are highlighted the Pearson correlation correlations greater than 0.30, in red the Partial correlation whenever they were lower than Pearson Correlation for the same couple of variables)

	Nfam	Fsize	Ds	Dipv	Mod	SigC	Dist	Year	Mden	Peakp	Posf	Posm	Fsize_	Anim	Anim_	Mden	Pos_	Mar_
											Mf1	Mf2	class		class	class	class	dist
Nfam	*	-0.25	**-0.37**	**-0.43**	**-0.55**	**-0.38**	0.15	-0.29	0.13	**-0.36**	-0.01	**-0.41**	**0.93**	0.01	0.00	0.24	**-0.36**	**-0.51**
Fsize	-0.49	*	**0.92**	**0.33**	0.10	-0.13	-0.03	**0.48**	**-0.33**	-0.19	-0.22	-0.09	-0.09	**0.95**	**0.94**	**-0.53**	-0.19	0.06
Ds	**-0.33**	**0.09**	*	**0.43**	**0.40**	-0.14	0.01	**0.58**	**-0.34**	-0.16	-0.22	-0.04	-0.18	**0.87**	**0.83**	**-0.51**	-0.16	0.12
Dipv	**-0.14**	**0.07**	**0.19**	*	**0.36**	-0.11	0.17	**0.55**	**-0.47**	0.22	0.01	0.26	-0.29	0.21	0.24	**-0.55**	0.22	0.28
Mod	**-0.48**	-0.60	**0.38**	**0.04**	*	-0.05	-0.02	**0.37**	-0.10	0.09	0.05	0.06	**-0.42**	0.02	-0.07	-0.10	0.09	0.03
Sigc	**-0.13**	-0.10	0.06	-0.37	-0.06	*	**-0.34**	**-0.36**	-0.23	**0.33**	-0.03	**0.39**	**-0.39**	-0.21	-0.13	-0.18	**0.33**	**0.58**
Dist	0.66	0.56	0.51	-0.11	0.33	**0.04**	*	**0.56**	0.21	-0.19	-0.17	-0.24	0.07	-0.10	-0.10	0.24	-0.18	-0.20
Year	-0.56	**-0.50**	**-0.43**	**0.22**	**-0.33**	**-0.11**	**0.94**	*	-0.21	-0.05	-0.18	-0.05	-0.24	0.39	**0.33**	**-0.32**	-0.04	-0.02
Mdens	0.47	**0.43**	**0.31**	**-0.10**	0.25	-0.17	-0.69	0.63	*	-0.09	0.11	-0.16	-0.03	**-0.35**	**-0.41**	**0.92**	-0.09	-0.27
Peakp	**0.07**	0.17	0.00	0.06	0.15	**0.06**	-0.16	0.05	-0.17	*	**0.68**	**0.78**	**-0.44**	-0.29	-0.26	-0.10	**1.00**	**0.39**
PosfM1	-0.04	0.06	-0.01	0.04	0.03	-0.09	-0.01	-0.04	-0.10	**0.29**	*	**0.51**	-0.07	-0.23	-0.25	0.15	**0.67**	-0.09
PosfM2	**0.17**	0.05	0.05	0.01	-0.05	**-0.11**	0.00	-0.03	0.07	**0.00**	**0.50**	*	**-0.48**	-0.21	-0.17	-0.19	**0.78**	**0.72**
Fsize_class	**0.59**	-0.25	0.13	0.26	**0.07**	**-0.05**	-0.03	-0.05	-0.13	**0.00**	0.12	**-0.16**	*	0.18	0.18	0.08	-0.44	**-0.51**
Anim	0.65	**0.58**	**0.54**	-0.23	0.40	-0.04	-0.94	0.90	**-0.69**	-0.12	-0.01	0.04	-0.01	*	**0.98**	**-0.51**	-0.29	-0.07
Anim_class	-0.59	**-0.38**	**-0.52**	0.22	-0.39	0.07	0.89	**-0.88**	**0.68**	0.10	0.00	-0.10	0.12	**0.96**	*	**-0.58**	-0.25	0.01
Mden_class	-0.56	**-0.50**	**-0.39**	**0.05**	-0.35	0.04	0.86	**-0.82**	**0.92**	0.21	0.09	-0.10	0.07	**0.85**	**-0.85**	*	-0.11	-0.31
Pos_class	**-0.06**	-0.17	0.00	-0.06	-0.15	**-0.05**	0.16	-0.05	0.17	**1.00**	**-0.27**	**0.01**	0.00	0.12	-0.10	-0.20	*	**0.39**
Mar_dist	**0.04**	-0.08	0.33	0.05	-0.10	**0.20**	-0.25	0.22	-0.23	**-0.03**	-0.49	**0.75**	**-0.03**	-0.28	0.34	0.29	**0.02**	*

Nfam=number of family ;Fsize= average size of the family;Ds=Experimental design;Dipv=dependent variable (EBV-DYD-PTA or DeReg Proof); Mod=model used;SigC=level of significance (p-value) Dist=distance Year=year of the experiment;Mden=marker density;Peakp=peak position of QTL; PosfMf1=flanking marker position1;PosmMf2=flanking marker position 2;Fsize_class=classes of family size; Anim=number of animal in the experiment; Anim_class=classes of animal size for the experiment; Mdenclass=classes of marker density; Pos_class=class of position for QTL; Mar_dist=distance between flanking markers.

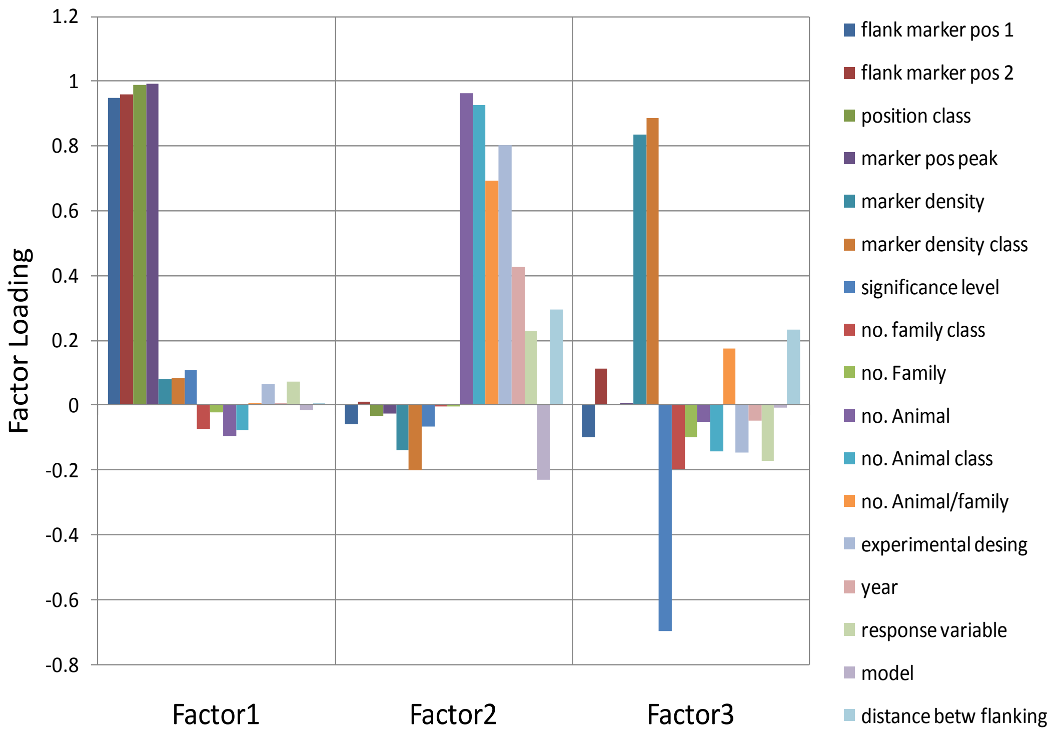


Figure 15. Pattern of Factor loading for the first 3 factor extracted in FA.

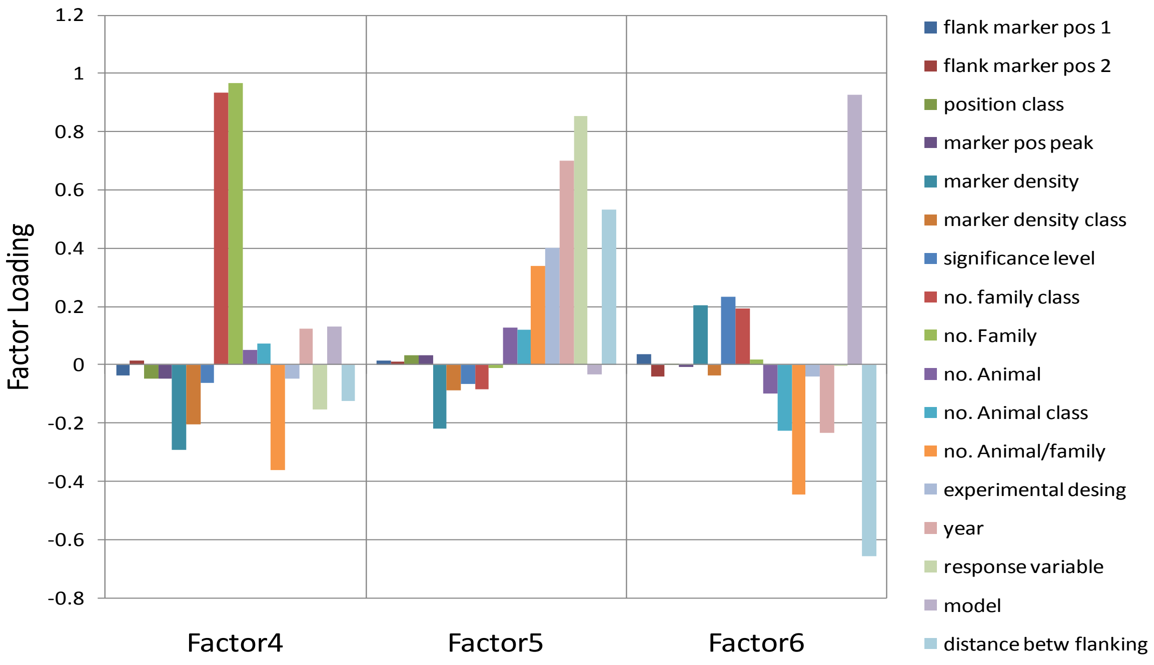


Figure 16. Pattern of factor loading for the factor 4 to factor6.

Number of Animals, year of the experiment, flanking marker distance, QTL peak position, significance level, number of families, density markers were retained as original variables for FA and PCA applications. Model and experimental design were chosen as classificatory variables.

Figure 17 report the *scree* plot of 7 eigenvalues and proportion of variance explained by the factor using FA on the subset of 7 variables chosen to perform FA and PCA.

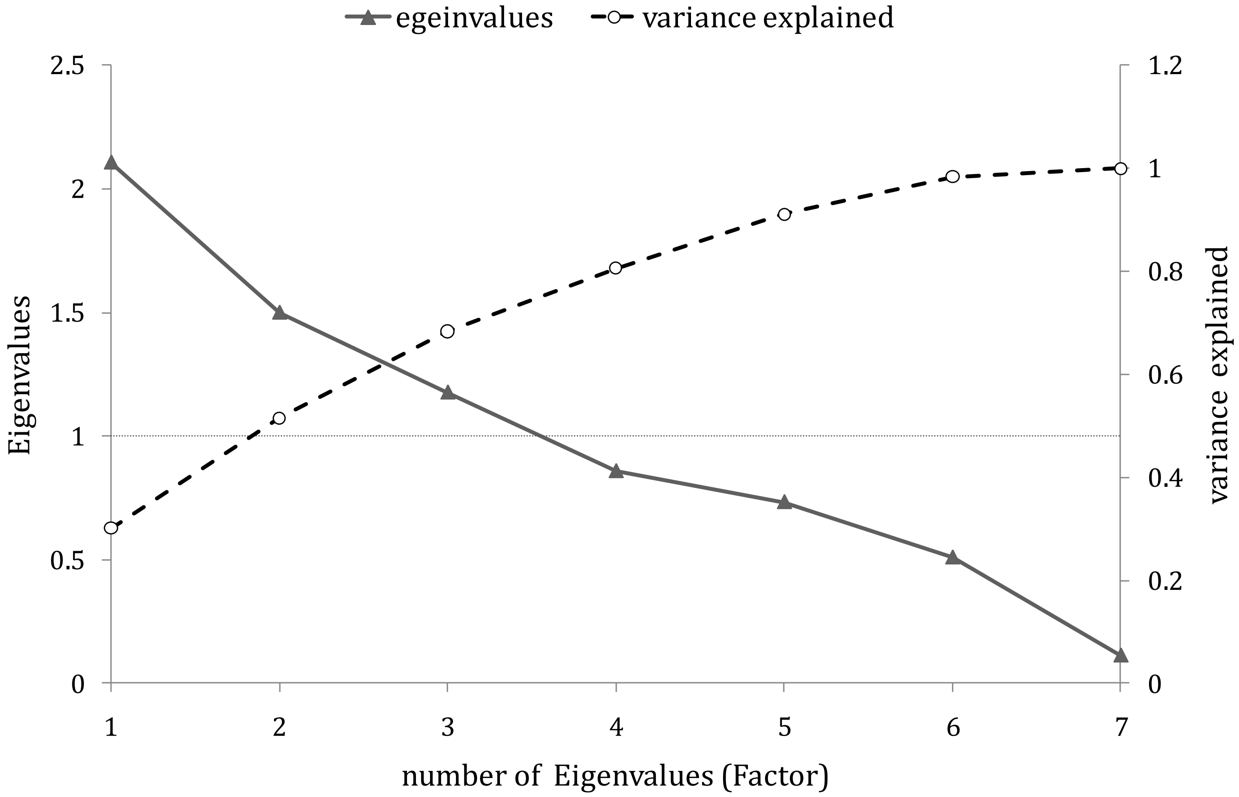


Figure 17. Pattern of Eigenvalues and variance explained by all 7 eigenvalues of correlation matrix and variance explained by 3 common factor retained after rotation of loading coefficient matrix.

Table 8 reports Factor loadings and eigenvectors for the FA and PCA, respectively. Factor analysis is able to explain 68% of the original variability with 3 latent factors: the first factor extracted is highly associated (factor loading of 0.95) to marker location along the chromosome and could be considered as a *marker map index*; the second factor shows the highest factor loadings (>0.70) for the number of animal involved and year of the experiment respectively, and it can be regarded as an indicator of the *dimension of the study*; the third factor is correlated to the significance level of the statistical test (0.78), number of families (0.63) and, negatively, to the marker density (-0.43). It can be named as index of *power of the experiment*. Same patterns can be observed in the eigenvectors of PCA. Four PCs are able to explain about 80% of the original variance. The first two PCs basically underline accurately the same structure found with the first two factors in FA, whereas PC3 and PC4 summarize the structure of F3. The score that each QTL get on each Factor or PC could be useful to classify the original QTL records and make them more comparable once that the redundancy of information has been removed.

Table 8. Factor loading of varimax rotated factor (FA) and eigenvector of PCA (PCA)

Original variable	FA			PCA			
	F1	F2	F3	PC1	PC2	PC3	PC4
Number of Animals	0.14	**0.74**	0.02	-0.04	**0.57**	-0.26	0.29
Year	-0.08	**0.84**	0.00	-0.19	**0.57**	-0.34	-0.22
Flanking Marker Distance	**0.95**	0.05	-0.02	**0.61**	0.28	0.07	-0.02
Peak Position	**0.95**	0.00	-0.04	**0.62**	0.23	0.07	0.02
Significance level	0.21	-0.16	**0.78**	0.03	0.18	**0.74**	-0.38
Number of families	-0.15	0.10	**0.63**	-0.21	0.22	0.47	**0.76**
Density markers	0.36	-0.43	**-0.51**	0.39	**-0.36**	-0.21	0.38
Variance Explained (%)	29	21	18	30	21	17	12

The results of the use of the factor scores calculated for each QTL record as response variable in a mixed model were reported in table 9 and Figure 17.

Table 9 indicate that all the classificatory factors included in the model were highly statistically significant excluding the effect of the model for the factor 2. The estimated least square means of the effect included in the model were reported in figure 18.

Table 9. Results of PROC MIXED

	P-value		
Factor	F1 (*map index*)	F2 (*dimension*)	F3 (*power*)
Experimental Design	0.0094	<0.0001	0.0094
Model	<0.0001	**0.641**	<0.0001
Trait	<0.0001	0.0011	<0.0001

Factor 1 allows to separate DD and GDD from DNA pooling. The estimated factor score for Factor 2 are quite large for DD and DD pool opposite to DD pool that showed a negative values. Apart from the sign, a possible explanation of these values may be found looking at the interpretation of the factor. There is no clear explanation for LS mean of factor 1. If the factor 2 is considered as index of the size of the experiment, it is possible to argue that DD is a design that generally involve larger sample size in order to achieve a reasonable power to detect QTL, while DNA pooling experiment generally reduce the size of the experiment and pool of DNA from milk sample to increase the experimental data point. The lowest figure is for GDD, probably due to the fact that for reach the same power of DD less animal are needed (WELLER *et al.* 1990). Looking at the LS mean for the third factor (*power of the experiment*) it is possible to note that DNA pool and GDD obtained similar values, maybe because the power of this experimental design are comparable. The estimated LS means for the model are quite similar for factor 1 and factor 2 (NS) and basically differs for the 3 factor. As the model indicated with 2 are generally more complex, it might be related to the higher predictive ability of these models. As far as the trait there is no difference or just slight difference among different trait in the values of the LS means for the three factors. Another remark worth to make is the quite large SE of these estimate. This quite complicate the results interpretation of the meaning of the factors.

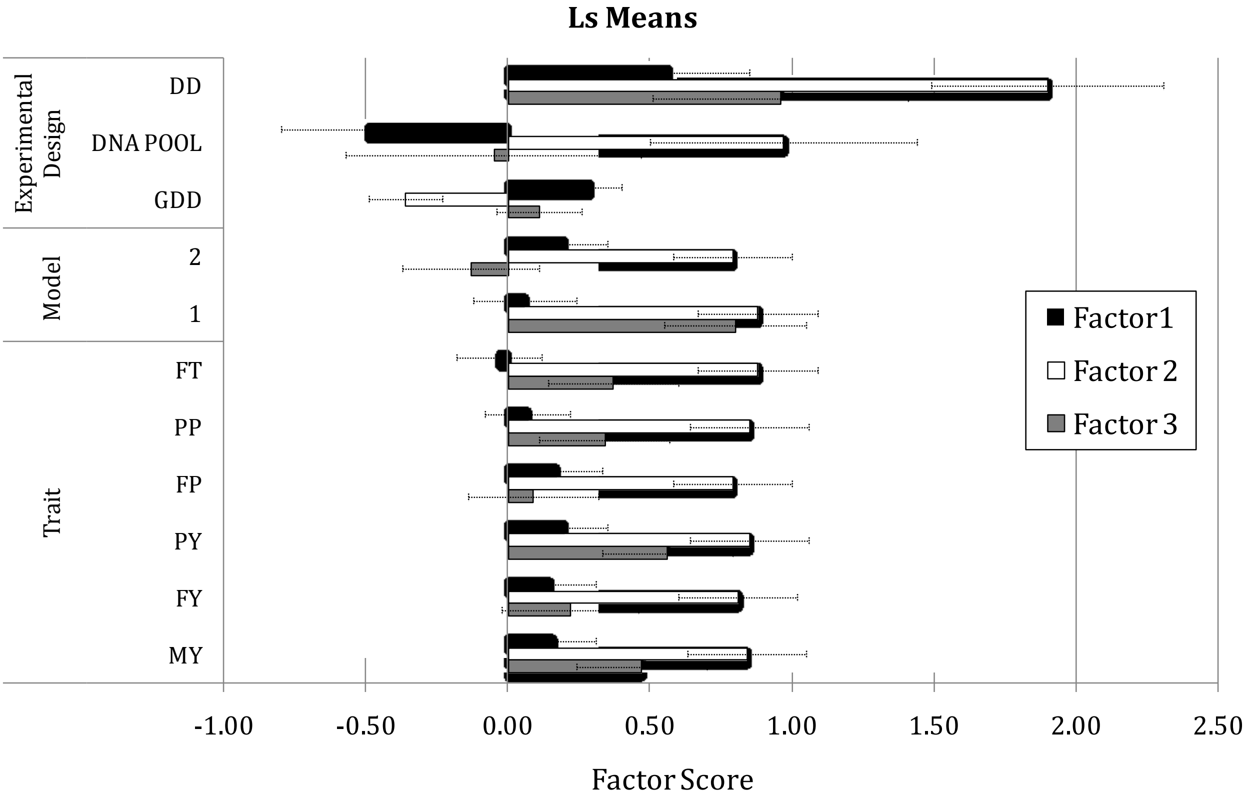


Figure 18. Ls means (and standard error) of three classificatory factor (Experimental design, model and trait) for factor 1, factor2 and factor 3

2.3.3 Distribution of QTL effects

The estimation of the effects of QTLs was carried out on 648 QTLs because of lack of data for most of the articles. Figure 19 reports the fit of the three distributions for all the traits analyzed.

Figure 19a shows the distribution of effects for milk production traits considered together (milk, fat, and protein yields) expressed in genetic standard deviations. The distributions of effects for traits milk yield, fat percentage, protein percentage, fat yield, protein yield and conformation traits (SD) are reported in figures 19b-g. Tables 10 ,11, and 12 reports the parameter when gamma function lognormal or Weibull is used to fit the QTL effect respectively. The gamma distribution fits all the traits except for milk production and the protein percentage (Table 10).

Table 10. Parameters of Gamma distribution used to fit QTL effects

Trait	N	Mean±sd	Scale (σ)	Shape (α)	Goodness of fit
AT (DSg)	148	0.76 ± 0.41	0.24	2.87	*
MY (Kg)	115	189.9 ± 134.2	94.57	2.00	NS
FY (Kg)	79	9.33 ± 4.57	2.24	4.16	*
PY (Kg)	80	7.02 ± 3.49	1.73	4.04	*
FP (%)	87	0.0013 ± 0.0008	0.00059	2.21	*
PP (%)	48	0.0005 ± 0.0003	0.00016	3.23	NS
CT (DSg)	91	0.81 ± 0.33	0.13	6.12	*

* test not significant (p-value>0.05), NS all the test are significant (p-value<0.05) the null hypothesis is rejected.

Lognormal distribution is suitable for fat yield and fat percentage, and milk production traits considered together (Table 11).

Table 11. Parameters of Lognormal distribution used to fit QTL effects

Trait	N	Mean±sd	Scale (ζ)	Shape (σ)	Goodness of fit
AT (DSg)	148	0.77 ± 0.46	-0.416	0.55	*
MY (Kg)	115	189.3 ± 134.2	4.97	0.73	NS
FY (Kg)	79	9.41 ± 5.18	2.10	0.51	*
PY (Kg)	80	7.09 ± 4.02	1.81	0.52	*
FP (%)	87	0.0013 ± 0.00113	-6.87	0.73	*
PP (%)	48	0.0005 ± 0.00032	-7.74	0.57	NS
CT (DSg)	91	0.82 ± 0.33	-0.29	0.43	NS

* test not significant (p-value>0.05), NS all the test are significant (p-value<0.05) the null hypothesis is rejected.

Finally, the Weibull distribution (Table 12) adequately fitted only fat yield and protein and fat percentage. The estimated parameters of mean for the gamma, weibull and lognormal distribution are quite similar across traits and distributions, whilst the gamma distributions present the lower standard deviation.

Table 12. Parameters of Weibull distribution used to fit QTL effects

Trait	N	Mean±sd	Scale (δ)	Shape (c)	Goodness of fit
AT (DSg)	148	0.76 ± 0.41	0.86	1.97	NS
MY (Kg)	115	191.6 ± 138.4	210.3	1.40	NS
FY (Kg)	79	9.37 ± 4.58	10.58	2.15	*
PY (Kg)	80	7.03 ± 3.56	7.94	2.06	*
FP (%)	87	0.0013 ± 0.0009	0.001	1.46	*
PP (%)	48	0.0005 ± 0.0003	0.0005	1.73	NS
CT (DSg)	91	0.81 ± 0.32	0.91	2.69	NS

* test not significant (p-value>0.05), NS all the test are significant (p-value<0.05) the null hypothesis is rejected.

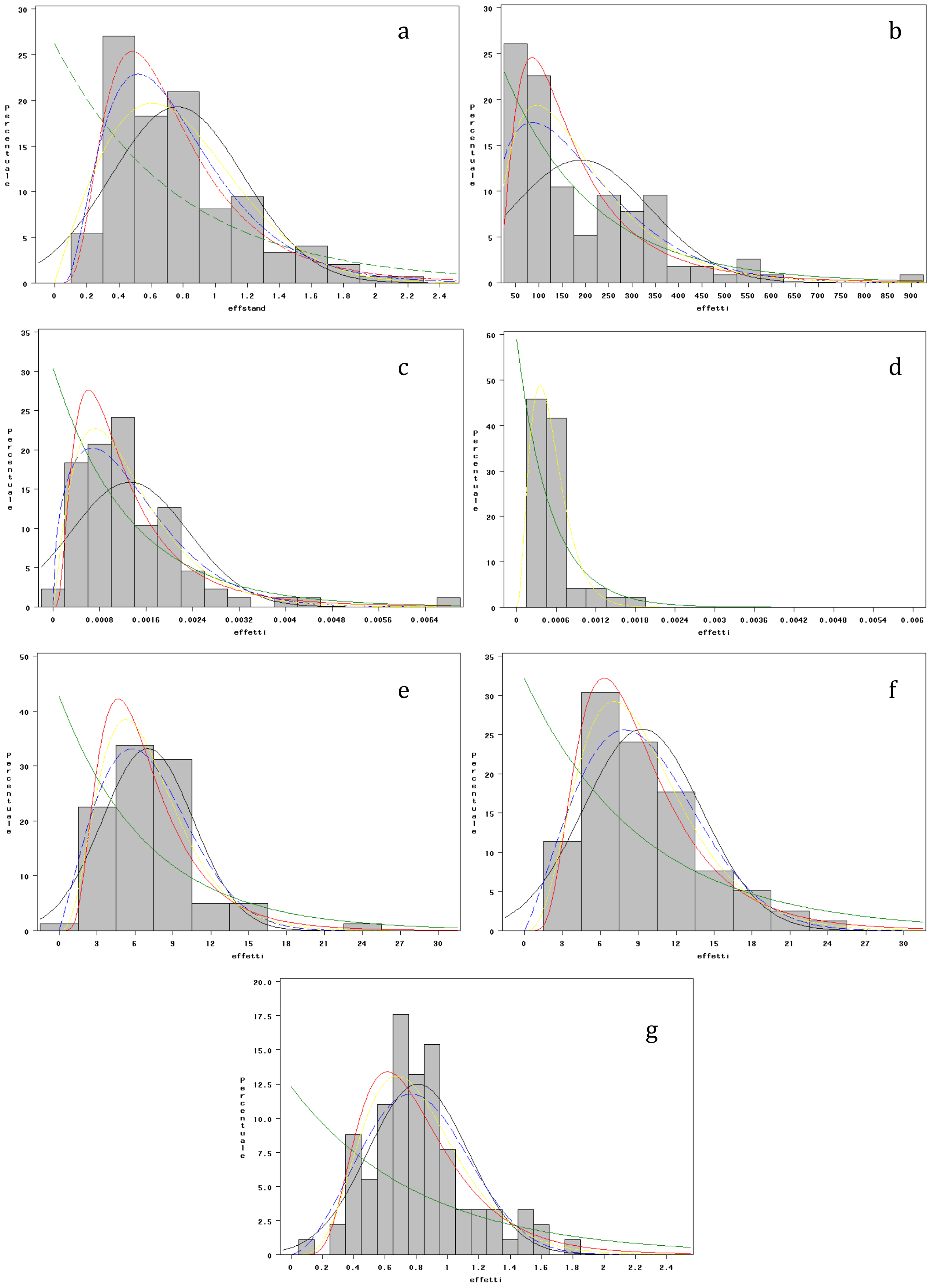


Figure 19. QTL effect fitted with Lognormal (red line), Weibull (blu line), Gamma (yellow line), Normal (black line) and exponential (green line) distribution for all the effect standardized (a), milk yield (b), fat percentage (c), protein percentage (d), fat yield (e), protein yield (f) and conformation traits (g).

These findings confirm, even if only partially, results reported by Hayes and Goddard (2001) for dairy cattle. Indeed, HAYES AND GODDARD (2001) suggested the gamma distribution as the most suitable to describe the distribution of the effects of QTLs. The difference with the results presented in this work regards the estimated parameter values, probably due to the effect of standardization adopted. Results seem to confirm the general hypothesis of the existence of a large number of QTL with small effect and few with great effect. However, the distribution of these effects is different across traits, probably because of the existence of QTLs with large effects on some quantitative traits.

2.4 CONCLUSION

A large amount of information on QTLs has been yielded by researches carried out on dairy cattle during the last fifteen years. More emphasis has been put on production traits, although some reports on milk nutritional quality and functional traits can be found. However, the use of different phenotypes, marker maps, statistical techniques make the comparison of results across studies rather difficult. In any case meta-analysis techniques used for removing redundant information and validating the position and the effects of QTLs give just some indication on the possibility of using this technique to score the QTL according to their reliability. A preliminary analysis of distributions partially confirm the suitability of the Gamma to model the QTL effects. This results seems to confirm the general hypothesis of the existence of a large number of QTLs with small effects and of a few ones with large effects.

ACKNOWLEDGEMENTS

I wish to acknowledge Dr. Rosanna Marino and Dr. Davide Perini (Universita cattolica del Sacro cuore, Piacenza), Dr. Lorraine Pariset (Università della Tuscia, Viterbo) for the helpful contribution to gene data collection. I would like to thank Dr. Raffaele. Mazza for his support on rescaling QTL position.

REFERENCES

AJMONE-MARSAN, P., E. MILANESI, F. SCHIAVINI, R. MAZZA and R. NEGRINI, 2007 Identification of milk protein percentage QTLs in Italian Friesian cattle by selective genotyping two GDD families with AFLP and SSR markers. Italian Journal of Animal Science **6:** 40-42.

ALAIN, K., N. A. KARROW, C. THIBAULT, J. ST-PIERRE, M. LESSARD *et al.*, 2009 Osteopontin: an early innate immune marker of Escherichia colimastitis harbors genetic polymorphisms with possible links with resistance to mastitis, pp. BioMed Central Ltd.

ALLISON, D. B., and M. HEO, 1998 Meta-analysis of linkage data under worst-case conditions: A demonstration using the human OB region. Genetics **148:** 859-865.

ARRANZ, J. J., W. COPPIETERS, P. BERZI, N. CAMBISANO, B. GRISART *et al.*, 1998 A QTL affecting milk yield and composition maps to bovine chromosome 20: a confirmation. Animal Genetics **29:** 107-115.

ASHWELL, M. S., D. W. HEYEN, T. S. SONSTEGARD, C. P. VAN TASSELL, Y. DA *et al.*, 2004 Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle. Journal of Dairy Science **87:** 468-475.

ASHWELL, M. S., and C. P. VAN TASSELL, 1999 Detection of putative loci affecting milk, health, and type traits in a US Holstein population using 70 microsatellite markers in a genome scan. Journal of Dairy Science **82:** 2497-2502.

ASHWELL, M. S., C. P. VAN TASSELL and T. S. SONSTEGARD, 2001 A genome scan to identify quantitative trait loci affecting economically important traits in a US Holstein population. Journal of Dairy Science **84:** 2535-2542.

BAGNATO, A., F. SCHIAVINI, A. ROSSONI, C. MALTECCA, M. DOLEZAL *et al.*, 2008 Quantitative trait loci affecting milk yield and protein percentage in a three-country Brown Swiss population. Journal of Dairy Science **91:** 767-783.

BANOS, G., J. A. WOOLLIAMS, B. W. WOODWARD, A. B. FORBES and M. P. COFFEY, 2008 Impact of Single Nucleotide Polymorphisms in Leptin, Leptin Receptor, Growth Hormone Receptor, and Diacylglycerol Acyltransferase (DGAT1) Gene Loci on Milk Production, Feed, and Body Energy Traits of UK Dairy Cows. Journal of Dairy Science **91:** 3190-3200.

BENNEWITZ, J., N. REINSCH, C. GROHS, H. LEVEZIEL, A. MALAFOSSE *et al.*, 2003 Combined analysis of data from two granddaughter designs: A simple strategy for QTL confirmation and increasing experimental power in dairy cattle. Genetics Selection Evolution **35:** 319-338.

BENNEWITZ, J., N. REINSCH, S. PAUL, C. LOOFT, B. KAUPE *et al.*, 2004a The DGAT1 K232A mutation is not solely responsible for the milk production quantitative trait locus on the bovine chromosome 14. Journal of Dairy Science **87:** 431-442.

BENNEWITZ, J., N. REINSCH, F. REINHARDT, Z. LIU and E. KALM, 2004b Top down preselection using marker assisted estimates of breeding values in dairy cattle. Journal of Animal Breeding and Genetics **121:** 307-318.

BLOTT, S., J. J. KIM, S. MOISIO, A. SCHMIDT-KUNTZEL, A. CORNET *et al.*, 2003 Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics **163:** 253-266.

BOICHARD, D., C. GROHS, F. BOURGEOIS, F. CERQUEIRA, R. FAUGERAS *et al.*, 2003 Detection of genes influencing economic traits in three French dairy cattle breeds. Genetics Selection Evolution **35:** 77-101.

BRAUNSCHWEIG, M.H., 2008. ASSOCIATIONS BETWEEN 2 PATERNAL CASEIN HAPLOTYPES AND MILK YIELD TRAITS OF SWISS FLECKVIEH CATTLE. J APPL GENET 49(1), 2008, PP. 69–74

BRYM, P., S. KAMINSKI and E. WOJCIK, 2005 Nucleotide sequence polymorphism within exon 4 of the bovine prolactin gene and its associations with milk performance traits. J Appl Genet **46:** 179-185.

CERNY, B. A., and H. F. KAISER, 1977 A Study Of A Measure Of Sampling Adequacy For Factor-Analytic Correlation Matrices. Multivariate Behavioral Research **12:** 43 - 47.

CHAMBERLAIN A., MCPARLAN H., BALASINGHAM T., CARRICK M., BOWMAN P.,ROBINSON N., GODDARD M.,2002. MAPPING QTL AFFECTING MILK COMPOSITION TRAITS IN DAIRY CATTLE USING A COMPLEX PEDIGREE, PROCEEDINGS OF THE 7THWORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, MONTPELLIER, FRANCE, 19–23 AUGUST 2002, ISBN 2-7380-1052-0, PAPER 09-08.

CHAMBERLAIN, A. J., H. C. MCPARTLAN and M. E. GODDARD, 2007 The number of loci that affect milk production traits in dairy cattle. Genetics **177:** 1117-1123.

COBANOGLU, O., P. J. BERGER and B. W. KIRKPATRICK, 2005 Genome screen for twinning rate QTL in four North American Holstein families. Animal Genetics **36:** 303-308.

COBANOGLU, O., I. ZAITOUN, Y. M. CHANG, G. E. SHOOK and H. KHATIB, 2006 Effects of the signal transducer and activator of transcription 1 (STAT1) gene on milk production traits in Holstein dairy cattle. J Dairy Sci **89:** 4433-4437.

COHEN, M., M. REICHENSTEIN, A. EVERTS-VAN DER WIND, J. HEON-LEE, M. SHANI *et al.*, 2004 Cloning and characterization of FAM13A1--a gene near a milk protein QTL on BTA6: evidence for population-wide linkage disequilibrium in Israeli Holsteins. Genomics **84:** 374-383.

CONTE, G., M. MELE, S. CHESSA, B. CASTIGLIONI, A. SERRA *et al.*, 2010 Diacylglycerol acyltransferase 1, stearoyl-CoA desaturase 1, and sterol regulatory element binding protein 1 gene polymorphisms and milk fatty acid composition in Italian Brown cattle. Journal of Dairy Science **93:** 753-763.

COPPIETERS, W., A. KVASZ, F. FARNIR, J. J. ARRANZ, B. GRISART *et al.*, 1998a A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sib pedigrees: application to milk production in a granddaughter design. Genetics **149:** 1547-1555.

COPPIETERS, W., J. RIQUET, J. J. ARRANZ, P. BERZI, N. CAMBISANO *et al.*, 1998b A QTL with major effect on milk yield and composition maps to bovine chromosome 14. Mammalian Genome **9:** 540-544.

DE KONING, D. J., 2006 Conflicting candidates for cattle QTLs. Trends Genet **22:** 301-305.

DEKKERS, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim Sci. **82:** E313-328.

DEKKERS, J. C. M., and F. HOSPITAL, 2002 The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics **3:** 22-32.

EWING, B., and P. GREEN, 2000 Analysis of expressed sequence tags indicates 35,000 human genes. Nature Genetics **25:** 232-234.

FONTANESI, L., E. SCOTTI, D. PECORARI, R. ZAMBONELLI, D. BIGI *et al.*, 2005 The BovMAS Consortium: investigation of bovine chromosome 14 for quantitative trait loci affecting milk production and quality traits in the Italian Holstein Friesian breed. Italian Journal of Animal Science **4:** 16-18.

FURBASS, R., A. WINTER, R. FRIES and C. KUHN, 2006 Alleles of the bovine DGAT1 variable number of tandem repeat associated with a milk fat QTL at chromosome 14 can stimulate gene expression. Physiological Genomics **25:** 116-120.

GAUTIER, M., A. CAPITAN, S. FRITZ, A. EGGEN, D. BOICHARD *et al.*, 2007 Characterization of the DGAT1 K232A and variable number of tandem repeat polymorphisms in french dairy cattle. Journal of Dairy Science **90:** 2980-2988.

GEORGES, M., D. NIELSEN, M. MACKINNON, A. MISHRA, R. OKIMOTO *et al.*, 1995 Mapping Quantitative Trait Loci Controlling Milk-Production in Dairy-Cattle by Exploiting Progeny Testing. Genetics **139:** 907-920.

GODDARD, M. E., and B. J. HAYES, 2007 Genomic selection. Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie **124:** 323-330.

GOFFINET, B., and S. GERBER, 2000 Quantitative trait loci: A meta-analysis. Genetics **155:** 463-473.

GRISART, B., W. COPPIETERS, F. FARNIR, L. KARIM, C. FORD *et al.*, 2002 Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res **12:** 222-231.

GRISART, B., F. FARNIR, L. KARIM, N. CAMBISANO, J. J. KIM *et al.*, 2004 Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proc Natl Acad Sci U S A **101:** 2398-2403.

GUILLAUME, F., S. FRITZ, D. BOICHARD and T. DRUET, 2008 Short Communication: Correlations of Marker-Assisted Breeding Values with Progeny-Test Breeding Values for Eight Hundred Ninety-Nine French Holstein Bulls. J. Dairy Sci. **91:** 2520-2522.

GUO, Y., H. CHEN, X. LAN, B. ZHANG, C. PAN *et al.*, 2008 Novel SNPs of the Bovine &lt;i&gt;LEPR&lt;/i&gt; Gene and Their Association with Growth Traits. Biochemical Genetics **46:** 828-834.

HALLÉN, E., A. WEDHOLM, A. ANDRÉN and A. LUNDÉN, 2008 Effect of β-casein, κ-casein and β-lactoglobulin genotypes on concentration of milk protein variants. Journal of Animal Breeding and Genetics **125:** 119-129.

HAYES, B., and M. E. GODDARD, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution **33:** 209-229.

HEYEN, D. W., J. I. WELLER, M. RON, M. BAND, J. E. BEEVER *et al.*, 1999 A genome scan for QTL influencing milk production and health traits in dairy cattle. Physiological Genomics **1:** 165-175.

HUANG, W., C. MALTECCA and H. KHATIB, 2008 A proline-to-histidine mutation in POU1F1 is associated with production traits in dairy cattle. Animal Genetics **39:** 554-557.

KAUPE, B., H. BRANDT, E. M. PRINZENBERG and G. ERHARDT, 2007 Joint analysis of the influence of CYP11B1 and DGAT1 genetic variation on milk production, somatic cell score, conformation, reproduction, and productive lifespan in German Holstein cattle. Journal of Animal Science **85:** 11-21.

KHATKAR, M. S., P. C. THOMSON, I. TAMMEN and H. W. RAADSMA, 2004 Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genetics Selection Evolution **36:** 163-190.

KIM, J. J., and M. GEORGES, 2002 Evaluation of a new fine-mapping method exploiting linkage disequilibrium: a case study analysing a QTL with major effect on milk composition on bovine chromosome 14. Asian-Australasian Journal of Animal Sciences **15:** 1250-1256.

KNOTT, S. A., J. M. ELSEN and C. S. HALEY, 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theoretical and Applied Genetics **93:** 71-80.

KUHN, C., G. FREYER, R. WEIKARD, T. GOLDAMMER and M. SCHWERIN, 1999 Detection of QTL for milk production traits in cattle by application of a specifically developed marker map of BTA6. Animal Genetics **30:** 333-340.

KUHN, C., G. THALLER, A. WINTER, O. R. P. BININDA-EMONDS, B. KAUPE *et al.*, 2004 Evidence for multiple alleles at the DGAT1 locus better explains a quantitative tip trait locus with major effect on milk fat content in cattle. Genetics **167:** 1873-1881.

KRZANOWSKY, W. J. 2003. PRINCIPLES OF MULTIVARIATE ANALYSIS. OXFORD UNIVERSITY PRESS INC., NEW YORK.

KUSS, A. W., J. GOGOL, H. BARTENSCHLAGER and H. GELDERMANN, 2005 Polymorphic AP-1 Binding Site in Bovine CSN1S1 Shows Quantitative Differences in Protein Binding Associated with Milk Protein Expression. Journal of Dairy Science **88:** 2246-2252.

LACORTE, G. A., M. A. MACHADO, M. L. MARTINEZ, A. L. CAMPOS, R. P. MACIEL *et al.*, 2006 DGAT1 K232A polymorphism in Brazilian cattle breeds. Genetics and Molecular Research **5:** 475-482.

LANDER, E., and L. KRUGLYAK, 1995 Genetic Dissection of Complex Traits - Guidelines for Interpreting and Reporting Linkage Results. Nature Genetics **11:** 241-247.

LI, Z., and D. RAO, 1996 Random effects model for meta-analysis of multiple quantitative sibpair linkage studies. Genetic Epidemiology **13:** 377-384.

LIPKIN, E., M. O. MOSIG, A. DARVASI, E. EZRA, A. SHALOM *et al.*, 1998 Quantitative trait locus mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide microsatellite markers: Analysis of milk protein percentage. Genetics **149:** 1557-1567.

LOOFT, C., N. REINSCH, C. KARALL-ALBRECHT, S. PAUL, M. BRINK *et al.*, 2001 A mammary gland EST showing linkage disequilibrium to a milk production QTL on bovine Chromosome 14. Mammalian Genome **12:** 646-650.

LÜ, A., X. HU, H. CHEN, J. JIANG, C. ZHANG *et al.*, 2010 Single nucleotide polymorphisms in bovine &lt;i&gt;PRL&lt;/i&gt; gene and their associations with milk production traits in Chinese Holsteins. Molecular Biology Reports **37:** 547-551.

Macciotta, N. P. P., M. Mele, G. Conte, A. Serra, M. Cassandro *et al.*, 2008 Association Between a Polymorphism at the Stearoyl CoA Desaturase Locus and Milk Production Traits in Italian Holsteins. Journal of Dairy Science **91:** 3184-3189.

Mele, M., G. Conte, B. Castiglioni, S. Chessa, N. P. P. Macciotta *et al.*, 2007 Stearoyl-Coenzyme A Desaturase Gene Polymorphism and Milk Fatty Acid Composition in Italian Holsteins. Journal of Dairy Science **90:** 4458-4465.

Mosig, M. O., E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller *et al.*, 2001 A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. Genetics **157:** 1683-1698.

Nilsen, H., H. Olsen, B. Hayes, E. Sehested, M. Svendsen *et al.*, 2009 Casein haplotypes and their association with milk production traits in Norwegian Red cattle. Genetics Selection Evolution **41:** 24.

Olsen, H. G., L. Gomez-Raya, D. I. Vage, I. Olsaker, H. Klungland *et al.*, 2002 A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. Journal of Dairy Science **85:** 3124-3130.

Olsen, H. G., S. Lien, M. Svendsen, H. Nilsen, A. Roseth *et al.*, 2004 Fine mapping of milk production QTL on BTA6 by combined linkage and linkage disequilibrium analysis. Journal of Dairy Science **87:** 690-698.

Pareek, C. S., U. Czarnik, T. Zabolewicz, R. S. Pareek and K. Walawski, 2005 DGAT1 K232A quantitative trait nucleotide polymorphism in Polish Black-and-White cattle. J Appl Genet **46:** 85-87.

Plante, Y., J. P. Gibson, J. Nadesalingam, H. Mehrabani-Yeganeh, S. Lefebvre *et al.*, 2001 Detection of quantitative trait loci affecting milk production traits on 10 chromosomes in Holstein cattle. J Dairy Sci **84:** 1516-1524.

Riquet, J., W. Coppieters, N. Cambisano, J. J. Arranz, P. Berzi *et al.*, 1999 Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. Proc Natl Acad Sci U S A **96:** 9252-9257.

M. Ron, D. W. Heyen, J. I. Weller, M. Band, E. Feldmesser, *et al.,* 1998. Detection and analysis of a locus affecting milk Concentration in the us and israeli dairy cattle populations. Proc. 6th World Cong pp

Ron, M., M. Cohen-Zinder, C. Peter, J. I. Weller and G. Erhardt, 2006 Short communication: a polymorphism in ABCG2 in Bos indicus and Bos taurus cattle breeds. J Dairy Sci **89:** 4921-4923.

Sanders, K., J. Bennewitz, N. Reinsch, G. Thaller, E. M. Prinzenberg *et al.*, 2006 Characterization of the DGAT1 mutations and the CSN1S1 promoter in the German Angeln dairy cattle population. J Dairy Sci **89:** 3164-3174.

SAS Institute. 1996. User's Guide: Statistics. SAS Inst., Inc., Cary,NC.

Schennink, A., H. Bovenhuis, K. M. Leon-Kloosterziel, J. A. M. van Arendonk and M. H. P. W. Visker, 2009 Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. Animal Genetics **40:** 909-916.

Schennink, A., J. M. L. Heck, H. Bovenhuis, M. H. P. W. Visker, H. J. F. van Valenberg *et al.*, 2008 Milk fatty acid unsaturation: Genetic parameters and effects of stearoyl-CoA desaturase (SCD1) and acyl CoA : diacylglycerol acyltransferase 1 (DGAT1). Journal of Dairy Science **91:** 2135-2143.

Schennink, A., W. M. Stoop, M. H. P. W. Visker, J. M. L. Heck, H. Bovenhuis *et al.*, 2007 DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. Animal Genetics **38:** 467-473.

Schnabel, R. D., J. J. Kim, M. S. Ashwell, T. S. Sonstegard, C. P. Van Tassell *et al.*, 2005 Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. Proc Natl Acad Sci U S A **102:** 6896-6901.

SHEEHY, P. A., L. G. RILEY, H. W. RAADSMA, P. WILLIAMSON and P. C. WYNN, 2009 A functional genomics approach to evaluate candidate genes located in a QTL interval for milk production traits on BTA6. Animal Genetics **40:** 492-498.

SPELMAN, R. J., W. COPPIETERS, L. KARIM, J. A. M. VANARENDONK and H. BOVENHUIS, 1996 Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population. Genetics **144:** 1799-1807.

SZYDA, J., Z. LIU, F. REINHARDT and R. REENTS, 2005 Estimation of quantitative trait loci parameters for milk production traits in German Holstein dairy cattle population. Journal of Dairy Science **88:** 356-367.

THALLER, G., W. KRAMER, A. WINTER, B. KAUPE, G. ERHARDT *et al.*, 2003 Effects of DGAT1 variants on milk production traits in German cattle breeds. Journal of Animal Science **81:** 1911-1918.

VIITALA, S., J. SZYDA, S. BLOTT, N. SCHULMAN, M. LIDAUER *et al.*, 2006 The role of the bovine growth hormone receptor and prolactin receptor genes in milk, fat and protein production in Finnish Ayrshire dairy cattle. Genetics **173:** 2151-2164.

VIITALA, S. M., N. F. SCHULMAN, D. J. DE KONING, K. ELO, R. KINOS *et al.*, 2003 Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle. Journal of Dairy Science **86:** 1828-1836.

WELLER, J. I., Y. KASHI and M. SOLLER, 1990 Power of Daughter and Granddaughter Designs for Determining Linkage between Marker Loci and Quantitative Trait Loci in Dairy-Cattle. Journal of Dairy Science **73:** 2525-2537.

WINTER, A., W. KRAMER, F. A. O. WERNER, S. KOLLERS, S. KATA *et al.*, 2002 Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA : diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content. Proceedings of the National Academy of Sciences of the United States of America **99:** 9300-9305.

ZHANG*, C., B. LIU*, H. CHEN, X. LAN, C. LEI *et al.*, 2009 Associations of a <i>Hinf</i>I PCR-RFLP of <i>POU1F1</i> Gene with Growth Traits in Qinchuan Cattle. Animal Biotechnology **20:** 71 - 74.

APPENDIX-TABLE OF QTLS DETECTED FOR ALL THE DAIRY TRAITS

Trait symbol	Trait	Trait symbol	Trait
ANG	Angularity	PL	Productive Live
BC	Body capacity	PLE	Pleiotropic milk traits
BDPT	Body depth	PP	Protein Percentage
BODY	Body	PY	Protein Yield
BSE	BSE	QFL	Quality feet and leeg
CBA	Calves born alive (%)	RANG	Rump angle
CCT	Canonical conformation t,	RES	Disease resistance
CDPT	Chest depth	RLNG	Rump Length
CE	Calving Ease	RLSR	Reare leg set rear view
CWDT	Chest width	RUH	Rear udder height
DAIR	Dairyness	RUW	Rear udder Width
DPR	Daughter Pregnancy rate	RWDT	Rump width
DSP	Degree of spotting	SCS	Somatic Cell Score
DYS	Dystocia	SIZE	Size
FA	Foot angle	SL	Suspensory legament
BDPT	Body depth	SS	Structurally soundness
FP	Fat Percentage	STAT	Stature
FTP	Front teat placement	STBI	Still birth
FUA	Fore udder attachment	STR	Strength
FY	Fat Yield	TDSV	Teat distance side view
GLNG	Gestation length	TPLA	Teat placement
HD	Heel depth	TW	Thurl width
HOCK	Hocks	TWR	Twinning rate
HS	Height at sacrum	TYPE	Type
IMPL	Implantation	UATT	Udder Attachment
MAST	Mastitis	UBAL	Udder Balance
MSP	Milking speed	UC	Udder Cleft
MY	Milk Yield	UCI	Udder Composite Index
NRR9	Non-Return Rate	UH	Udder Height
OR	Ovulation rate	UWDT	Udder Width
PERS	Persistency	VT	Veterinary treatment

Trait	Breed	Chromosome	Mb	cM	-Log (P-value)	Authors
BSE	HF	BTA_01	80108790	125	3.48	[1]
FER	HF	BTA_01	80109564	62	5.3	[2]
FP	HF	BTA_01	3278784		0	[3]
FP	HF	BTA_01	3278784	24	2.3	[4]
FY	HF	BTA_01	3278784		0	[3]
FY	HF	BTA_01	3278784	22	3	[4]
MY	BR	BTA_01	12014673	15.4	5.93	[5]
MY	HF	BTA_01	3278784		3	[3]
MY	HF	BTA_01	3278784	8	3	[4]
MY	AY	BTA_01	106572226	135	3.54	[6]
PERS	HF	BTA_01	139481488	170	3.51	[7]
PP	BR	BTA_01	134014694	122.1	7.53	[5]
PP	HF	BTA_01	50565088		0	[3]
PP	HF	BTA_01	74946884	64.9	6.21	[8]
PP	HF	BTA_01	3278784	22	2.3	[4]
PY	HF	BTA_01	3278784		3	[3]
PY	HF	BTA_01	154015460	118	5.65	[9]
PY	HF	BTA_01	3278784	11	2.3	[4]
RANG	HF	BTA_01	32587282	40	6.5	[2]
SCS	HF	BTA_01	144026580	125	3	[10]
SCS	DC	BTA_01	144026580		4.61	[11]
TPLA	HF	BTA_01	133500304	119	4.61	[12]
UC	HF	BTA_01	133500304	119	3	[12]
BDPT	HF	BTA_02	27001226	21	4.61	[12]
BODY	HF	BTA_02	27001226	21	4.61	[12]
CCT	HF	BTA_02	116972326		8.11	[13]
CDPT	HF	BTA_02	32534842	40	5.12	[2]
CWDT	HF	BTA_02	131258043	139	7.42	[14]
FP	HF	BTA_02	33504782	29	4.61	[15]
MY	BR	BTA_02	27001226	11.9	6.06	[5]
MY	AY	BTA_02	19162554	44	9.21	[6]
PERS	HF	BTA_02	131258043	139	3	[7]
PL	HF	BTA_02	126230233	101.5	11.51	[16]
PL	HF	BTA_02	63664600	79	4.2	[17]
PP	HF	BTA_02	27001226	16	4.61	[15]
PP	BR	BTA_02	27001226	11.9	4.07	[5]
PP	HF	BTA_02	27001226	41	7.82	[9]
PP	HF	BTA_02	66618342	56.3	9.9	[8]
PY	NR	BTA_02	66618342		0.01	[18]
STAT	HF	BTA_02	27001226	24	4.61	[12]
STR	HF	BTA_02	5896219	3	3	[12]
TDSV	HF	BTA_02	32534842	38	5.55	[2]
TW	HF	BTA_02	5896219	2	3	[12]
UATT	HF	BTA_02	5896219	2	3	[12]
FA	HF	BTA_03	85593845	65	3	[12]
FP	HF	BTA_03	57075572	49	4.61	[15]
FP	HF	BTA_03	15255799	16	7.01	[9]
FP	HF	BTA_03	22409217	34	5.52	[19]
FP	HF	BTA_03	40108772		4.61	[20]
FP	AY	BTA_03	10506133	1	3	[6]
FY	HF	BTA_03	15255799	16	4.99	[9]
FY	NR	BTA_03	8717459	14	4.27	[21]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
MAST	NR	BTA_03	89151424	110	4.27	[22]
AST	DC	BTA_03	104524038		4.61	[11]
MY	HF	BTA_03	30143222	32	4.61	[15]
MY	BR	BTA_03	57075572	68	6.25	[5]
MY	HF	BTA_03	15255799	16	4.76	[9]
MY	AY	BTA_03	57075572	62	9.21	[6]
PP	HF	BTA_03	23673607	36	7.42	[13]
PP	HF	BTA_03	30143222	29	4.61	[15]
PP	BR	BTA_03	44989616	59.4	5.04	[5]
PP	HF	BTA_03	15255799	24	4.61	[2]
PP	HF	BTA_03	15255799	16	8.11	[9]
PP	HF	BTA_03	117410651	115	3.22	[8]
PP	HF	BTA_03	22409217	34	5.3	[19]
PP	AY	BTA_03	10506133	1	9.21	[6]
PY	HF	BTA_03	30143222	39	4.61	[15]
PY	HF	BTA_03	15255799	16	5.18	[9]
SCS	NR	BTA_03	974625	30	0	[22]
SCS	DC	BTA_03	122601311		4.61	[11]
BDPT	HF	BTA_04	5426681	4	3	[12]
FP	SR	BTA_04	99603227	73	4.96	[23]
FY	SR	BTA_04	99603227	79	4.71	[23]
MAST	NR	BTA_04	75297570	82	3.12	[22]
MAST	DC	BTA_04	75297570		4.61	[11]
MY	BR	BTA_04	85616671	87.3	4.02	[5]
MY	SR	BTA_04	99603227	79	4.42	[23]
PL	HF	BTA_04	25416102	32	4.62	[9]
PP	BR	BTA_04	85616671	87.3	9.11	[5]
PP	SR	BTA_04	99603227	95	4.61	[23]
PP	HF	BTA_04	25416102	24.3	7.6	[8]
SCS	NR	BTA_04	114265410	120	0	[22]
SCS	HF	BTA_04	6898946	3.1	0	[24]
STAT	HF	BTA_04	27407874	28	3	[12]
TLNG	HF	BTA_04	70177264	63.9	11.51	[16]
TLNG	HF	BTA_04	70177264		8.52	[13]
ANG	HF	BTA_05	38773802	31	6.5	[14]
BC	HF	BTA_05	38773802	154	7.13	[14]
CDPT	HF	BTA_05	104045902	124	4.61	[2]
DAIR	HF	BTA_05	26701704	46	3	[12]
FP	HF	BTA_05	78209773	87	4.61	[15]
FP	HF	BTA_05	104045902	90	4.74	[9]
FP	NR	BTA_05	111918714	120	4.71	[21]
FY	NR	BTA_05	46178414		6.91	[18]
FY	NR	BTA_05	78209773	115	4.02	[21]
FY	HF	BTA_05	14382552	63	3.65	[19]
HS	HF	BTA_05	104045902	124	11.51	[2]
MY	AY	BTA_05	78209773	98	4.29	[6]
OR	HF	BTA_05	104045902	107	10.56	[25]
PP	BR	BTA_05	14382552	17.3	4.16	[5]
PP	HF	BTA_05	51912073	55.4	3	[8]
PP	HF	BTA_05	7023104	0	3.17	[19]
PY	HF	BTA_05	7023104	0	2.66	[19]
PY	AY	BTA_05	60836475	77	3	[6]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
RANG	HF	BTA_05	104045902	112	4.61	[12]
RLNG	HF	BTA_05	104045902	124	5.91	[2]
RUW	HF	BTA_05	14382552	18.8	8.52	[16]
RWID	HF	BTA_05	104045902	124	8.11	[2]
SCS	HF	BTA_05	45122505	54	4.61	[15]
SCS	HF	BTA_05	104045902	90	6.32	[9]
SCS	SR	BTA_05	7023104		3	[26]
SIZE	HF	BTA_05	38773802	123	6.5	[14]
STAT	HF	BTA_05	104045902	122	5.78	[14]
TLNG	HF	BTA_05	26701704	43	4.61	[12]
TWR	HF	BTA_05	51937550	65	2.91	[27]
TWR	NR	BTA_05	71198741	80	5.38	[28]
TYPE	HF	BTA_05	104045902	109	4.61	[12]
UATT	HF	BTA_05	104045902	112	4.61	[12]
BODY	HF	BTA_06	65910741	85	5.12	[29]
BSE	HF	BTA_06	26679810	60	4.41	[1]
CBA	HF	BTA_06	47949144	58	4.02	[17]
CCT	HF	BTA_06	90235201		8.52	[13]
DAIR	HF	BTA_06	13407380	0	6.03	[14]
DSP	HF	BTA_06	65910741	83	9.21	[10]
FA	HF	BTA_06	55021812	67	3	[12]
FP	HF	BTA_06	34553110	49	4.61	[15]
FP	HF	BTA_06	26679810	0	0	[30]
FP	HF	BTA_06	13407380	41	0	[31]
FP	HF	BTA_06	26679810	46	4.61	[32]
FP	HF	BTA_06	13407380	23	0	[4]
FP	NR	BTA_06	53249618	41	13.82	[21]
FP	NR	BTA_06	34553110		9.21	[33]
FP	HF	BTA_06	44167400	54	3	[34]
FP	AY	BTA_06	75218590	95	3	[6]
FY	HF	BTA_06	55021812	30	0	[30]
FY	HF	BTA_06	13407380	41	0	[31]
FY	HF	BTA_06	1005847	68	3	[35]
FY	HF	BTA_06	26679810	31	0	[32]
FY	HF	BTA_06	44167400		3.77	[61]
FY	HF	BTA_06	52471030	55	9.21	[36]
FY	HF	BTA_06	51320395	56	0	[4]
FY	NR	BTA_06	44167400		9.21	[33]
FY	HF	BTA_06	11830912	7	0	[34]
FY	HF	BTA_06	47949144	9	9.57	[37]
FY	AY	BTA_06	95684727	101	3	[38]
FY	HF	BTA_06	13407380	27	0	[39]
G	HF	BTA_06	93912529	122	4.61	[15]
HS	HF	BTA_06	26679810	54	6.03	[2]
MAST	NR	BTA_06	26679810	37	7.42	[22]
MSP	HF	BTA_06	117551813	160	4.83	[2]
MY	BR	BTA_06	11830912	8.2	4.19	[5]
MY	HF	BTA_06	59094958	32	0	[30]
MY	HF	BTA_06	1005847	91	0	[31]
MY	HF	BTA_06	1005847	71	0	[35]
MY	HF	BTA_06	26679810	40	0	[32]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
MY	HF	BTA_06	26679810	42	3	[4]
MY	NR	BTA_06	43424290	37	3.96	[21]
MY	NR	BTA_06	56279887		4.96	[33]
MY	HF	BTA_06	140491	108	3	[34]
MY	HF	BTA_06	44167400	13	4.61	[40]
MY	HF	BTA_06	47949144	17	7.8	[37]
MY	AY	BTA_06	13407380	70	3.24	[38]
MY	AY	BTA_06	26679810	66	3.69	[6]
MY	HF	BTA_06	26679810	45	0	[39]
PERS	HF	BTA_06	47949144	77	3.91	[7]
PLE	HF	BTA_06	52471030	68	4.2	[35]
PLE	HF	BTA_06	26679810	58	4.61	[32]
PP	HF	BTA_06	75218590	91	8.52	[16]
PP	HF	BTA_06	75218590		9.21	[13]
PP	HF	BTA_06	95684727	106	4.61	[15]
PP	BR	BTA_06	110450962	101.4	3.93	[5]
PP	HF	BTA_06	88531297	98	7.82	[2]
PP	HF	BTA_06	44167400	19	0	[30]
PP	HF	BTA_06	13407380	41	3	[31]
PP	HF	BTA_06	26679810	46	4.61	[32]
PP	HF	BTA_06	26679810	35.5	3.91	[8]
PP	HF	BTA_06	13407380	42	3	[4]
PP	NR	BTA_06	53249618	41	13.82	[21]
PP	NR	BTA_06	34553110		9.21	[33]
PP	HF	BTA_06	110450962	99	3	[34]
PP	HF	BTA_06	44167400	13	9.21	[40]
PP	AY	BTA_06	13407380	71	3.58	[38]
PP	AY	BTA_06	26679810	66	9.21	[6]
PY	HF	BTA_06	15892433	24	4.61	[15]
PY	HF	BTA_06	55021812	29	0	[30]
PY	HF	BTA_06	1005847	41	0	[31]
PY	HF	BTA_06	1005847	71	3	[35]
PY	HF	BTA_06	26679810	31	0	[32]
PY	HF	BTA_06	56279887	58	4.96	[36]
PY	HF	BTA_06	26679810	42	0	[4]
PY	NR	BTA_06	43557420		8.25	[33]
PY	HF	BTA_06	44167400	54	3	[34]
PY	HF	BTA_06	47949144	17	9.72	[37]
PY	HF	BTA_06	44167400	49	0	[39]
QFL	HF	BTA_06	65910741	89	5.81	[29]
RWDT	HF	BTA_06	47949144	62	5.74	[2]
RWDT	HF	BTA_06	65910741	87	4.07	[29]
SCS	NR	BTA_06	1005847	7	0	[22]
SL	HF	BTA_06	65910741	88	4.07	[29]
STAT	HF	BTA_06	26679810	66	3.77	[29]
STBI	HF	BTA_06	44167400	58	4.02	[17]
STR	HF	BTA_06	26679810	70	2.92	[29]
TLNG	HF	BTA_06	93912529	133	3	[12]
TPLA	HF	BTA_06	65910741	88	6.21	[29]
BDPT	HF	BTA_07	70880326	95	3	[12]
CBA	HF	BTA_07	3519	9	3.86	[17]
CE	HF	BTA_07	3519	10	4.02	[17]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
DYS	HF	BTA_07	3519	10	4.02	[17]
FA	HF	BTA_07	70880326	83	4.61	[12]
FER	HF	BTA_07	103078178	120	5.6	[2]
FER	HF	BTA_07	3519	11	8.11	[41]
FP	HF	BTA_07	8362905	29	5.3	[41]
FY	HF	BTA_07	16073893	84	5.12	[2]
FY	HF	BTA_07	103354	30	15.94	[41]
HD	HF	BTA_07	29549377	32	4.83	[2]
MY	HF	BTA_07	18374198	30	4.61	[15]
MY	BR	BTA_07	58272906	72.9	3.83	[5]
MY	HF	BTA_07	95653326	115	5.43	[9]
MY	NR	BTA_07	8362905		0	[18]
OR	HF	BTA_07	5044158	57	7.26	[25]
PL	HF	BTA_07	61803055	71	4.61	[15]
PL	HF	BTA_07	23902610	74	6.38	[41]
PP	BR	BTA_07	16073893	25.4	3.46	[5]
PP	HF	BTA_07	3519	0	6.21	[8]
PY	HF	BTA_07	18374198	30	4.61	[15]
PY	HF	BTA_07	95653326	115	5.55	[9]
PY	NR	BTA_07	8362905		0	[18]
PY	HF	BTA_07	3519	29	8.11	[41]
SCS	HF	BTA_07	40615252		9.21	[13]
SCS	HF	BTA_07	40615252	67	4.61	[15]
SCS	HF	BTA_07	105987669	128	5.74	[9]
SCS	HF	BTA_07	61803055	107	3.69	[17]
SCS	DC	BTA_07	40615252		4.61	[11]
SCS	HF	BTA_07	7256641	60	8.25	[41]
STBI	HF	BTA_07	3519	9	3.86	[17]
TWR	HF	BTA_07	20486029	31	3	[27]
TWR	NR	BTA_07	95653326	109	3.02	[28]
CE	HF	BTA_08	115008873	116	4.61	[12]
DYS	HF	BTA_08	71007272	93	3.22	[17]
FP	HF	BTA_08	92684519		3.52	[42]
MAST	NR	BTA_08	43269174	46	0	[22]
PP	HF	BTA_08	92684519		4.32	[42]
PP	BR	BTA_08	3138003	2.7	3.31	[5]
PP	HF	BTA_08	16349773	19.1	8.11	[8]
PY	HF	BTA_08	92684519		3.21	[42]
RWDT	HF	BTA_08	7642	140	7.26	[2]
SCS	NR	BTA_08	43269174	54	4.26	[22]
SCS	HF	BTA_08	14657850	17	3	[10]
SCS	DC	BTA_08	14657850		4.61	[11]
STBI	HF	BTA_08	71007272	93	3.35	[17]
TWR	HF	BTA_08	92684519	116.7	6.91	[43]
CCT	HF	BTA_09	48829387		9.21	[13]
CDPT	HF	BTA_09	91998	127	4.96	[2]
CE	HF	BTA_09	88747991	96	3	[12]
FP	HF	BTA_09	47964403		0	[3]
FY	HF	BTA_09	53855608	71	3	[3]
FY	HF	BTA_09	9083186	37	3	[39]
MAST	SR	BTA_09	76757514	145	4.61	[26]
MY	BR	BTA_09	17864617	24.1	3.84	[5]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
MY	HF	BTA_09	53855608	71	0	[3]
MY	HF	BTA_09	63072434	60	3.61	[19]
MY	HF	BTA_09	40095010	44	0	[39]
PERS	HF	BTA_09	48829387	56	3.91	[7]
PP	HF	BTA_09	9977152		0	[3]
PP	HF	BTA_09	48829387	44.9	2.53	[8]
PP	HF	BTA_09	48266906	44	2.76	[19]
PY	HF	BTA_09	47964403		3	[3]
PY	HF	BTA_09	9083186	21	0	[39]
RANG	HF	BTA_09	66959287		9.21	[13]
RANG	HF	BTA_09	63072434	58	4.61	[12]
RES	SR	BTA_09	48829387	130	4.61	[26]
SCS	SR	BTA_09	96877725	120	4.61	[26]
SS	HF	BTA_09	76757514	61	3	[12]
STR	HF	BTA_09	76757514	64	3	[12]
UBAL	HF	BTA_09	48829387	48	5.12	[2]
	HF	BTA_10	78024227	80	3.08	[17]
BDPT	HF	BTA_10	21780794	46	4.61	[12]
CE	HF	BTA_10	78024227	87	4.71	[17]
DYS	HF	BTA_10	77769031	83	4.27	[17]
FY	HF	BTA_10	3760237	25	3	[3]
MY	HF	BTA_10	95242673	98	4.61	[15]
MY	BR	BTA_10	94089479	100	7.68	[5]
MY	HF	BTA_10	3760237	25	0	[3]
MY	HF	BTA_10	57843327	62	3	[19]
NRR9	HF	BTA_10	34378167	48	3.19	[17]
PP	HF	BTA_10	15297220	24.7	3.51	[44]
PP	BR	BTA_10	94089479	100	3.76	[5]
PP	HF	BTA_10	3760237		0	[3]
PP	HF	BTA_10	14270604	19.3	3.22	[8]
PP	HF	BTA_10	93046118	85	4.2	[19]
PP	HF	BTA_10	78024227		5.99	[62]
PY	HF	BTA_10	3760237		0	[3]
SCS	HF	BTA_10	87757725	86	4.71	[2]
SCS	HF	BTA_10	34378167	49	3.61	[17]
SCS	DC	BTA_10	87757725		3	[11]
STBI	HF	BTA_10	77769031	80	3.08	[17]
STR	HF	BTA_10	21780794	42	3	[12]
TWR	HF	BTA_10	25659564	41	6.91	[43]
UATT	HF	BTA_10	78024227	116	3	[12]
FP	HF	BTA_11	93600688	106	0	[32]
FY	HF	BTA_11	86753945	90	4.61	[15]
FY	NR	BTA_11	86753945	83	4.27	[21]
MAST	SR	BTA_11	25860118	22	3	[26]
MY	BR	BTA_11	10563705	19.4	7.33	[5]
MY	HF	BTA_11	39676942	67	3	[32]
PERS	HF	BTA_11	93353307	124	6.27	[2]
PP	BR	BTA_11	10563705	19.4	7.38	[5]
PP	HF	BTA_11	39676942	67	3	[32]
PP	HF	BTA_11	10563705	9.5	4.61	[8]
PY	HF	BTA_11	43765564	83	4.61	[15]
PY	HF	BTA_11	93600688	106	0	[32]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
RANG	HF	BTA_11	104015545	146	4.96	[2]
RES	SR	BTA_11	25860118	22	6.91	[26]
SCS	SR	BTA_11	39676942	45	6.91	[26]
STBI	HF	BTA_11	29768216	32	3.22	[17]
ANG	HF	BTA_12	38773802	31	6.5	[14]
CCT	HF	BTA_12	67356645		7.26	[16]
FY	NR	BTA_12	11451506		4.61	[18]
FY	AY	BTA_12	14953932	28	9.21	[6]
MY	BR	BTA_12	47633280	50.4	4.83	[5]
MY	AY	BTA_12	14953932	21	5.07	[6]
PP	BR	BTA_12	67356645	83.6	3.04	[5]
PP	HF	BTA_12	14953932	21.4	2.3	[8]
PP	AY	BTA_12	14953932	10	9.21	[6]
PY	NR	BTA_12	11451506		5.3	[18]
PY	AY	BTA_12	14953932	21	3.83	[6]
TWR	NR	BTA_12	11451506	10	2.76	[28]
BSE	HF	BTA_13	12147149	55	3.11	[1]
CCT	HF	BTA_13	45835686		9.21	[13]
FA	HF	BTA_13	45835686	54	3	[12]
FUA	HF	BTA_13	4443382	0	9.21	[14]
FY	HF	BTA_13	30550744	28	2.67	[19]
HOCK	HF	BTA_13	15494818	53	3.41	[29]
HS	HF	BTA_13	68284220	74	5.52	[2]
MSP	HF	BTA_13	68284220	94	4.83	[2]
MY	HF	BTA_13	59914941	84	4.61	[15]
MY	BR	BTA_13	12147149	23	3.79	[5]
PP	HF	BTA_13	29963861	34	4.61	[15]
PP	BR	BTA_13	15494818	27.6	4.72	[5]
PP	HF	BTA_13	12147149	14.8	20.03	[8]
PP	NR	BTA_13	15494818	32	5.3	[21]
PY	HF	BTA_13	59914941	77	4.61	[15]
RANG	HF	BTA_13	39013400	54	7.42	[2]
RLSR	HF	BTA_13	15494818	54	4.07	[29]
STR	HF	BTA_13	15494818	51	3.44	[29]
TDSV	HF	BTA_13	4443382	8	7.6	[2]
TLNG	HF	BTA_13	12147149	39	5.12	[29]
TYPE	HF	BTA_13	59914941	62	3	[12]
UATT	HF	BTA_13	59914941	63	4.61	[12]
UCI	HF	BTA_13	59914941	64	4.61	[12]
UD	HF	BTA_13	59914941	72	3	[12]
UH	HF	BTA_13	59914941	66	4.61	[12]
UWDT	HF	BTA_13	59914941	63	3	[12]
CCT	HF	BTA_14	9188082		9.21	[13]
FA	HF	BTA_14	51171085	54	3	[12]
FEC	HF	BTA_14	11778542	11	4.61	[15]
FP	HF	BTA_14	33620332		4.01	[42]
FP	HF	BTA_14	9188082	6	9.21	[13]
FP	HF	BTA_14	1198414	4	4.61	[15]
FP	HF	BTA_14	262181	0.3	4.61	[45]
FP	HF	BTA_14	3940258	0	11.51	[2]
FP	HF	BTA_14	3940258	5	6.91	[63]
FP	HF	BTA_14	1198414	0	4.61	[46]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
FP	HF	BTA_14	1198414	0	25.33	[9]
FP	HF	BTA_14	444082	0.01	260.19	[47]
FP	HF	BTA_14	17162375	24	0.69	[32]
FP	HF	BTA_14	444082	1	11.51	[48]
FP	HF	BTA_14	1198414	5	3	[49]
FP	HF	BTA_14	3940258		16.12	[20]
FP	HF	BTA_14	3940258		9.21	[62]
FP	AY	BTA_14	1198414	0	9.21	[6]
FP	FL	BTA_14	1198414	0	9.21	[50]
FTP	HF	BTA_14	33620332		11.51	[16]
FTP	HF	BTA_14	33620332		9.21	[13]
FUA	HF	BTA_14	33620332		11.51	[16]
FUA	HF	BTA_14	33620332		9.21	[13]
FY	HF	BTA_14	9188082	6	9.21	[13]
FY	HF	BTA_14	1198414	4	4.61	[15]
FY	HF	BTA_14	262181	0.3	4.61	[45]
FY	HF	BTA_14	3940258	0	6.81	[2]
FY	HF	BTA_14	1198414	0	10.82	[9]
FY	HF	BTA_14	444082	0.01	56.31	[47]
FY	HF	BTA_14	33620332	42	0	[32]
FY	HF	BTA_14	1198414	6	4.61	[51]
FY	HF	BTA_14	3940258		9.21	[62]
FY	AY	BTA_14	1198414	0	5.3	[6]
MAST	NR	BTA_14	61416682	60	6.32	[22]
MAST	DC	BTA_14	73839050		4.61	[11]
MY	HF	BTA_14	33620332		3.5	[42]
MY	BR	BTA_14	3940258	5.1	2.99	[5]
MY	HF	BTA_14	262181	0.3	4.61	[45]
MY	HF	BTA_14	3940258	5	4.61	[63]
MY	HF	BTA_14	1198414	0	4.61	[46]
MY	HF	BTA_14	65029647	101	5.26	[9]
MY	HF	BTA_14	444082	0.01	67.15	[47]
MY	NR	BTA_14	61416682		3.58	[18]
MY	HF	BTA_14	1198414	0	4.61	[51]
PERS	HF	BTA_14	262181	1	4.61	[7]
PP	HF	BTA_14	73839050	86	7.6	[16]
PP	HF	BTA_14	73839050	6	9.21	[13]
PP	HF	BTA_14	1198415	1	4.61	[15]
PP	BR	BTA_14	1198414	0	4.47	[5]
PP	HF	BTA_14	262181	0.3	4.61	[45]
PP	HF	BTA_14	3940258	0	11.51	[2]
PP	HF	BTA_14	3940258	5	4.61	[63]
PP	HF	BTA_14	1198414	0	3	[46]
PP	HF	BTA_14	1198414	0	5.34	[9]
PP	HF	BTA_14	17162375	21	0	[32]
PP	HF	BTA_14	76576330	79.7	5.81	[8]
PP	AY	BTA_14	7864999	50	9.21	[6]
PY	HF	BTA_14	65029647	74	4.61	[15]
PY	HF	BTA_14	262181	0.3	4.61	[45]
PY	HF	BTA_14	444082	0.01	24.34	[47]
PY	HF	BTA_14	1198414	6	4.61	[51]
RANG	HF	BTA_14	24666782	33	3	[12]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
SCS	HF	BTA_14	33620332	37	4.65	[42]
TPLA	HF	BTA_14	39554394	48	3	[12]
TWR	HF	BTA_14	34157877	67.9	6.91	[43]
UC	HF	BTA_14	39554394	51	4.61	[12]
UCI	HF	BTA_14	33620332		9.21	[16]
BODY	HF	BTA_15	47822871	45	3	[12]
FY	NR	BTA_15	54031734		6.91	[18]
HD	HF	BTA_15	40147387	30	5.78	[2]
PERS	HF	BTA_15	8203267	0	3.91	[7]
PY	NR	BTA_15	54031734		3.69	[18]
RES	SR	BTA_15	71313406		4.61	[26]
SCS	HF	BTA_15	40147387	34	4.61	[15]
SCS	HF	BTA_15	40147387	40	7.6	[2]
STAT	HF	BTA_15	40147387	37	4.61	[12]
TPLA	HF	BTA_15	47822871	52	4.61	[12]
TW	HF	BTA_15	47822871	48	3	[12]
TYPE	HF	BTA_15	47822871	47	4.61	[12]
UATT	HF	BTA_15	40147387	36	4.61	[12]
UC	HF	BTA_15	47822871	55	4.61	[12]
UCI	HF	BTA_15	47822871	45	4.61	[12]
UD	HF	BTA_15	40147387	37	4.61	[12]
FEC	HF	BTA_16	65420458	81	4.61	[15]
MY	BR	BTA_16	37105151	54.1	5.78	[5]
MY	NR	BTA_16	23314		0.62	[18]
PP	BR	BTA_16	18978724	30.2	3.23	[5]
PP	HF	BTA_16	3905436	11.5	7.13	[8]
PY	NR	BTA_16	23314		0.4	[18]
TYPE	HF	BTA_16	23314	1	3	[12]
UD	HF	BTA_16	47816804	61	4.61	[12]
BSE	HF	BTA_17	60983439	144	6.5	[1]
CE	HF	BTA_17	28901254	69	4.61	[12]
FP	HF	BTA_17	41819506	0	3.77	[19]
MY	BR	BTA_17	116972326	92.1	3.63	[5]
MY	HF	BTA_17	41819506	33	2.92	[19]
PERS	HF	BTA_17	41819506	48	3.91	[7]
PL	HF	BTA_17	63940959	68	7.42	[9]
PP	BR	BTA_17	116972326	92.1	3.62	[5]
PY	HF	BTA_17	63940959	96	4.61	[15]
PY	HF	BTA_17	41819506	28	3.15	[19]
RANG	HF	BTA_17	6750519	8	5.3	[2]
TLNG	HF	BTA_17	54265266	78	3	[12]
UH	HF	BTA_17	28901254	69	3	[12]
	HF	BTA_18	38773802	75	6.21	[17]
BODY	HF	BTA_18	1809898	0	3.22	[29]
CCT	HF	BTA_18	7236416		9.21	[13]
CE	HF	BTA_18	5909692	53	3.47	[17]
DPR	HF	BTA_18	17914774	28	4.61	[52]
DYS	HF	BTA_18	5909692	53	3.47	[17]
FEC	HF	BTA_18	38773802	14	4.61	[15]
FP	HF	BTA_18	62032979		3.32	[42]
FUA	HF	BTA_18	38773802	68	7.01	[14]
FY	HF	BTA_18	62032979		4.83	[42]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
FY	HF	BTA_18	63005888	84	4.61	[15]
GLNG	HF	BTA_18	38773802	17	6.91	[14]
MY	HF	BTA_18	62032979		5.15	[42]
MY	BR	BTA_18	42313982	54.7	3.01	[5]
MY	NR	BTA_18	21390609	39	5.81	[21]
NRR9	HF	BTA_18	62032979	111	4.71	[17]
PERS	HF	BTA_18	42313982	88	3.51	[7]
PL	HF	BTA_18	18182789	104	4.27	[17]
PL	HF	BTA_18	21390609	33	4.61	[52]
PP	HF	BTA_18	9741926	10	4.83	[2]
PP	HF	BTA_18	42994567	55	4.96	[8]
PY	HF	BTA_18	62032979		4.11	[42]
PY	HF	BTA_18	42994567	62	3	[15]
PY	NR	BTA_18	42313982	79	4.02	[21]
SCS	HF	BTA_18	62032979	78	4.55	[42]
SCS	HF	BTA_18	17914774	26	4.61	[15]
SCS	HF	BTA_18	65544902	117	6.21	[17]
SCS	HF	BTA_18	21390609	34	4.61	[52]
SCS	DC	BTA_18	38773802		4.61	[11]
SCS	HF	BTA_18	38773802	70	4.58	[14]
STAT	HF	BTA_18	1809898	0	5.3	[29]
STBI	HF	BTA_18	38773802	75	6.21	[17]
UATT	HF	BTA_18	17914774	33	3	[12]
UBAL	HF	BTA_18	54084213	98	7.6	[2]
UCI	HF	BTA_18	17914774	26.2	7.01	[16]
UCI	HF	BTA_18	17914774	29	3	[12]
UD	HF	BTA_18	17914774	36	4.61	[12]
UH	HF	BTA_18	17914774	28	4.61	[12]
BSE	HF	BTA_19	57735277	97	5.17	[1]
C10:	HF	BTA_19	29320502	71	4.61	[53]
C12:	HF	BTA_19	29320502	71	4.61	[53]
C14:	HF	BTA_19	29320502	68	4.61	[53]
C18:	HF	BTA_19	29320502	75	3	[53]
C6:0	HF	BTA_19	29320502	68	3	[53]
C8:0	HF	BTA_19	29320502	71	4.61	[53]
FP	AY	BTA_19	57735277	67	9.21	[6]
FTP	HF	BTA_19	32701737	67	5.74	[14]
FY	HF	BTA_19	62787919	134	6.65	[2]
MY	BR	BTA_19	59427053	95	3.82	[5]
OR	HF	BTA_19	56748922	65	10.56	[25]
PY	HF	BTA_19	57735277	138	4.83	[2]
RANG	HF	BTA_19	57735277	118	6.07	[2]
TLNG	HF	BTA_19	45986139	76	3	[12]
TWR	HF	BTA_19	36257751	57	4.61	[27]
TYPE	HF	BTA_19	29936789		6.91	[16]
UBAL	HF	BTA_19	10888948	30	5.71	[2]
BDPT	HF	BTA_20	29489060	36	3	[12]
BODY	HF	BTA_20	29489060	38	3	[12]
DAIR	HF	BTA_20	15028800	30	3	[12]
FP	HF	BTA_20	12701617	32	6.21	[54]
FP	HF	BTA_20	28328789	44.1	3.45	[55]
FP	HF	BTA_20	29489060	34	4.61	[63]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
FP	HF	BTA_20	41842995		0	[3]
FP	AY	BTA_20	29489060	37	3.36	[56]
FUA	HF	BTA_20	38026113		9.21	[13]
FY	HF	BTA_20	2125027	10	2.3	[54]
FY	HF	BTA_20	38026113	46	5.36	[55]
FY	HF	BTA_20	41842995		0	[3]
FY	HF	BTA_20	2125027	8	3.1	[19]
FY	AY	BTA_20	34302734	41	3.13	[56]
MY	HF	BTA_20	14200868	32	3.82	[54]
MY	HF	BTA_20	43633141	68	4.61	[15]
MY	BR	BTA_20	41842995	55.1	5.74	[5]
MY	HF	BTA_20	38026113	43	8.62	[55]
MY	HF	BTA_20	41842995		0	[3]
MY	HF	BTA_20	23797437	33	2.92	[19]
MY	AY	BTA_20	46038915	89	6.27	[6]
MY	AY	BTA_20	41842995	59	5.78	[56]
PP	HF	BTA_20	12701617	38	2.99	[54]
PP	HF	BTA_20	38026113		8.11	[13]
PP	HF	BTA_20	29489060	40	4.61	[15]
PP	BR	BTA_20	11573135	19.1	4.22	[5]
PP	HF	BTA_20	38026113	48.4	5.76	[55]
PP	HF	BTA_20	23797437	38	5.65	[2]
PP	HF	BTA_20	29489060	34	6.91	[63]
PP	HF	BTA_20	41842995		3	[3]
PP	HF	BTA_20	26586164	31.2	2.41	[8]
PP	AY	BTA_20	46038915	68	3	[6]
PP	AY	BTA_20	12701617	24	4.95	[56]
PY	HF	BTA_20	2125027	10	3	[54]
PY	HF	BTA_20	38026113	43	10.2	[55]
PY	HF	BTA_20	12701617		0	[3]
PY	NR	BTA_20	38026113	66	3.3	[21]
PY	HF	BTA_20	6101483	16	4.14	[19]
PY	AY	BTA_20	12701617	31	6.91	[56]
RANG	HF	BTA_20	6101483	8	3	[12]
RLNG	HF	BTA_20	23797437	34	5.81	[2]
RWDT	HF	BTA_20	23797437	24	5.78	[2]
SCS	HF	BTA_20	15028800	29	4.61	[15]
STR	HF	BTA_20	29489060	38	4.61	[12]
TPLA	HF	BTA_20	2125027	2	4.02	[29]
TW	HF	BTA_20	29489060	38	4.61	[12]
UBAL	HF	BTA_20	23797437	30	5.12	[2]
FA	HF	BTA_21	2094150	6	3.96	[29]
FP	HF	BTA_21	20054369		3.46	[42]
FY	HF	BTA_21	10367149		4.01	[42]
MY	HF	BTA_21	10367149		6.65	[42]
MY	BR	BTA_21	57556325	62.7	3.72	[5]
MY	HF	BTA_21	34036511	56	7.26	[9]
MY	AY	BTA_21	8909340	24	9.21	[6]
PERS	HF	BTA_21	2094150	0	3.22	[7]
PL	HF	BTA_21	20054369		3.28	[42]
PL	HF	BTA_21	20054369		11.51	[16]
PL	HF	BTA_21	59291339	85	5.28	[9]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
PP	HF	BTA_21	20054369		3.13	[42]
PP	BR	BTA_21	9394817	13.5	4.04	[5]
PP	HF	BTA_21	22766375	32.3	4.61	[8]
PY	HF	BTA_21	10367149		4.41	[42]
PY	HF	BTA_21	34036511	56	8.52	[9]
SCS	HF	BTA_21	22766375	33	5.99	[9]
SCS	DC	BTA_21	22766375		3	[11]
TWR	HF	BTA_21	2094150	2	4.61	[43]
PERS	HF	BTA_22	54321063	82	3.22	[7]
PP	HF	BTA_22	46132838	77	4.61	[15]
PP	BR	BTA_22	4782800	2.9	8.76	[5]
PP	HF	BTA_22	56718263	76.1	5.81	[8]
PY	HF	BTA_22	18850289	30	4.61	[15]
RANG	HF	BTA_22	46132838		9.21	[13]
RANG	HF	BTA_22	33485870	60	3	[12]
SCS	HF	BTA_22	46132838	80	4.61	[15]
SCS	HF	BTA_22	3375978	0	5.99	[9]
SCS	DC	BTA_22	32628727		4.61	[11]
STAT	HF	BTA_22	46132838	72	3	[12]
	AY	BTA_23	65502	1	3	[57]
CCT	HF	BTA_23	39932757		9.21	[13]
CE	HF	BTA_23	43918282	62	3	[12]
FA	HF	BTA_23	39932757	84	3.82	[29]
FP	HF	BTA_23	27119120		3.68	[42]
FP	HF	BTA_23	48643240	67	0	[32]
FY	HF	BTA_23	19935307		3.27	[42]
FY	HF	BTA_23	18058703	22	0	[32]
FY	HF	BTA_23	27119120	28	2.36	[19]
MSP	AY	BTA_23	19935307	53	3	[57]
MY	BR	BTA_23	34858817	52.3	4.4	[5]
MY	HF	BTA_23	48643240	67	0	[32]
MY	AY	BTA_23	65502	4	9.21	[6]
PL	HF	BTA_23	19935307		3.81	[42]
PP	HF	BTA_23	39932757		4.61	[42]
PP	BR	BTA_23	28066444	42.9	2.91	[5]
PP	AY	BTA_23	13796393	27	9.9	[57]
PP	HF	BTA_23	26374724	34	0	[32]
PP	HF	BTA_23	65502	7.2	7.42	[8]
PP	AY	BTA_23	46038915	21	3.75	[6]
PY	HF	BTA_23	39932757		4.61	[42]
PY	HF	BTA_23	26374724	34	0	[32]
QFL	HF	BTA_23	39932757	84	5.52	[29]
RUH	HF	BTA_23	39932757	84	3.1	[29]
SCS	HF	BTA_23	27119120		4.61	[42]
SCS	HF	BTA_23	27501397	41	4.61	[15]
SCS	HF	BTA_23	33375215	52	5.45	[9]
SCS	SR	BTA_23	48643240		3	[26]
SCS	HF	BTA_23	13796393	18	3	[10]
SCS	HF	BTA_23	33375215		4.61	[20]
SCS	DC	BTA_23	13796393		4.61	[11]
TLNG	HF	BTA_23	39932757	82	5.12	[29]
TWR	HF	BTA_23	1942523	0	3.01	[27]

Trait	breed	Chromosome	Trait	cM	-Log (P-value)	Authors
TWR	NR	BTA_23	13796393	30	2.8	[28]
UATT	HF	BTA_23	65502	16	3	[12]
UCI	HF	BTA_23	65502	17	3	[12]
UD	HF	BTA_23	27501397	49	3	[12]
VT	AY	BTA_23	26374724	34	3	[57]
BDPT	HF	BTA_24	6924621	11	3	[12]
BODY	HF	BTA_24	6924621	14	3	[12]
CE	HF	BTA_24	6924621	22	3	[12]
PP	BR	BTA_24	7100282	8.1	3.08	[5]
PP	HF	BTA_24	31275642	33.9	2.66	[8]
STR	HF	BTA_24	6924621	16	3	[12]
UATT	HF	BTA_24	44208720	48	3	[12]
UCI	HF	BTA_24	44208720	51	3	[12]
UD	HF	BTA_24	58327320	56	3	[12]
FA	HF	BTA_25	32930967	39	3	[12]
MAST	SR	BTA_25	7044904		3	[26]
MY	AY	BTA_25	32930967	70	9.21	[6]
PERS	HF	BTA_25	32930967	73	3	[7]
PP	BR	BTA_25	11948951	14.4	3.91	[5]
PP	AY	BTA_25	792146	0	3	[6]
PY	AY	BTA_25	11948951	44	5.95	[6]
RES	SR	BTA_25	7044904		6.91	[26]
SS	HF	BTA_25	13184868	7	3	[12]
FP	HF	BTA_26	33950591		3.14	[42]
FP	HF	BTA_26	34484440	38	3.04	[19]
FP	AY	BTA_26	16342189	15	3	[6]
FY	HF	BTA_26	33950591		5.6	[42]
FY	HF	BTA_26	40946785	57	11.51	[2]
FY	HF	BTA_26	10183115	14	6.91	[58]
FY	HF	BTA_26	16342189	3	4.61	[19]
MY	BR	BTA_26	3238814	2.8	9.16	[5]
MY	HF	BTA_26	16342189	15	3.17	[19]
PP	BR	BTA_26	18820032	27	5.69	[5]
PP	HF	BTA_26	18820032	24.8	4.61	[8]
PY	HF	BTA_26	40946785	57	7.82	[2]
PY	HF	BTA_26	41702323	64	6.91	[58]
PY	HF	BTA_26	16342189	11	3.44	[19]
SCS	HF	BTA_26	33950591		4.36	[42]
SCS	HF	BTA_26	18820032	0	7.13	[9]
SCS	HF	BTA_26	33950591	39.7	0	[24]
SCS	DC	BTA_26	41702323		2.3	[11]
TLNG	HF	BTA_26	39215378	31	3	[12]
CE	HF	BTA_27	29881182	36	4.61	[12]
DAIR	HF	BTA_27	24204135		9.21	[13]
DAIR	HF	BTA_27	24204135	32	3	[12]
DAIR	HF	BTA_27	29881182	40	4.61	[59]
FEC	HF	BTA_27	40821483	62	4.61	[15]
FP	HF	BTA_27	46111584		3.04	[42]
FP	HF	BTA_27	29881182	46	3	[59]
FY	HF	BTA_27	5199747	5	4.61	[15]
FY	HF	BTA_27	18479665	21	3	[59]
MAST	NR	BTA_27	30072300	45	4.2	[22]

Trait	breed	Chromosome	Mb	cM	-Log (P-value)	Authors
MAST	DC	BTA_27	39065010		4.61	[11]
MY	HF	BTA_27	46111584		3.86	[42]
MY	BR	BTA_27	5199747	0	4.21	[5]
MY	HF	BTA_27	18479665	21	3	[59]
MY	AY	BTA_27	17759641	29	9.21	[6]
PP	HF	BTA_27	46111584		3.55	[42]
PP	BR	BTA_27	18262877	55.8	3.52	[5]
PP	HF	BTA_27	18262877	15	11.74	[8]
PP	HF	BTA_27	39065010	52	0	[59]
PY	HF	BTA_27	46111584		3.54	[42]
PY	HF	BTA_27	18479665	21	3	[59]
PY	AY	BTA_27	17759641	29	3	[6]
SCS	HF	BTA_27	5199747	8	5.52	[17]
SCS	HF	BTA_27	38664639	54	0	[59]
STAT	HF	BTA_27	12434722	6	3	[12]
FA	HF	BTA_28	30799797	48	3	[12]
FEC	HF	BTA_28	30799797	48	4.61	[15]
IMPL	HF	BTA_28	6096065	4	7.82	[2]
MY	HF	BTA_28	17830350	33	4.61	[15]
MY	BR	BTA_28	34662525	49.4	3.97	[5]
PP	HF	BTA_28	40177670		7.82	[13]
PP	BR	BTA_28	34662525	49.4	4.55	[5]
SS	HF	BTA_28	17830350	26	3	[12]
TPLA	HF	BTA_28	35158600	4	5.12	[2]
UATT	HF	BTA_28	9383790	8	3	[12]
UC	HF	BTA_28	6096065	4	5.91	[2]
UCI	HF	BTA_28	17830350	26	4.61	[12]
UH	HF	BTA_28	17830350	25	3	[12]
UWDT	HF	BTA_28	9383790	16	3	[12]
FA	HF	BTA_29	35638708	34	3	[12]
MS	HF	BTA_29	10024112	20	6.21	[29]
MY	HF	BTA_29	10024112	1	4.61	[15]
MY	HF	BTA_29	11748185	0	5.13	[9]
MY	NR	BTA_29	37259742		4.61	[18]
MY	AY	BTA_29	17774915	34	9.21	[6]
PP	BR	BTA_29	21307382	24.2	3	[5]
PP	HF	BTA_29	2688744	0.9	2.81	[8]
PY	HF	BTA_29	11748185	10	4.61	[15]
PY	HF	BTA_29	11748185	0	5.71	[9]
PY	AY	BTA_29	17774915	28	3	[6]
SCS	HF	BTA_29	46403679	50	4.61	[15]
SS	HF	BTA_29	35638708	34	3	[12]
UH	HF	BTA_29	17774915	16	3	[12]
UWDT	HF	BTA_29	17774915	13	3	[12]
BSE	HF	BTA_X	1124718	58	7.82	[1]
FY	HF	BTA_X	70009047		3.27	[60]

REFERENCES

1. Zhang, C., et al., *Mapping of multiple quantitative trait loci affecting bovine spongiform encephalopathy.* Genetics, 2004. **167**(4): p. 1863-1872.
2. Boichard, D., et al., *Detection of genes influencing economic traits in three French dairy cattle breeds.* Genetics Selection Evolution, 2003. **35**(1): p. 77-101.
3. Georges, M., et al., *Mapping Quantitative Trait Loci Controlling Milk-Production in Dairy-Cattle by Exploiting Progeny Testing.* Genetics, 1995. **139**(2): p. 907-920.
4. Nadesalingam, J., Y. Plante, and J.P. Gibson, *Detection of QTL for milk production on Chromosomes 1 and 6 of Holstein cattle.* Mammalian Genome, 2001. **12**(1): p. 27-31.
5. Bagnato, A., et al., *Quantitative trait loci affecting milk yield and protein percentage in a three-country Brown Swiss population.* Journal of Dairy Science, 2008. **91**(2): p. 767-783.
6. Viitala, S.M., et al., *Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle.* Journal of Dairy Science, 2003. **86**(5): p. 1828-1836.
7. Harder, B., et al., *Mapping of quantitative trait loci for lactation persistency traits in German Holstein dairy cattle.* Journal of Animal Breeding and Genetics, 2006. **123**(2): p. 89-96.
8. Mosig, M.O., et al., *A whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion.* Genetics, 2001. **157**(4): p. 1683-1698.
9. Heyen, D.W., et al., *A genome scan for QTL influencing milk production and health traits in dairy cattle.* Physiological Genomics, 1999. **1**(3): p. 165-175.
10. Reinsch, N., et al., *A QTL for the degree of spotting in cattle shows synteny with the KIT locus on chromosome 6.* Journal of Heredity, 1999. **90**(6): p. 629-34.
11. Rupp, R. and D. Boichard, *Genetics of resistance to mastitis in dairy cattle.* Vet Res, 2003. **34**(5): p. 671-88.
12. Ashwell, M.S., et al., *Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle.* Journal of Dairy Science, 2005. **88**(11): p. 4111-4119.
13. Ashwell, M.S., C.P. Van Tassell, and T.S. Sonstegard, *A genome scan to identify quantitative trait loci affecting economically important traits in a US Holstein population.* Journal of Dairy Science, 2001. **84**(11): p. 2535-2542.
14. Schrooten, C., et al., *Whole genome scan to detect quantitative trait loci for conformation and functional traits in dairy cattle.* J Dairy Sci, 2000. **83**(4): p. 795-806.
15. Ashwell, M.S., et al., *Detection of quantitative trait loci affecting milk production, health, and reproductive traits in Holstein cattle.* Journal of Dairy Science, 2004. **87**(2): p. 468-475.
16. Ashwell, M.S. and C.P. Van Tassell, *Detection of putative loci affecting milk, health, and type traits in a US Holstein population using 70 microsatellite markers in a genome scan.* Journal of Dairy Science, 1999. **82**(11): p. 2497-2502.
17. Kuhn, C., et al., *Quantitative trait loci mapping of functional traits in the German Holstein cattle population.* Journal of Dairy Science, 2003. **86**(1): p. 360-368.
18. Lillehammer, M., et al., *A genome scan for quantitative trait locus by environment interactions for production traits.* Journal of Dairy Science, 2007. **90**(7): p. 3482-3489.
19. Plante, Y., et al., *Detection of quantitative trait loci affecting milk production traits on 10 chromosomes in Holstein cattle.* J Dairy Sci, 2001. **84**(6): p. 1516-24.
20. Ron, M., et al., *Short communication: a polymorphism in ABCG2 in Bos indicus and Bos taurus cattle breeds.* J Dairy Sci, 2006. **89**(12): p. 4921-3.
21. Olsen, H.G., et al., *A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle.* Journal of Dairy Science, 2002. **85**(11): p. 3124-3130.
22. Klungland, H., et al., *Quantitative trait loci affecting clinical mastitis and somatic cell count in dairy cattle.* Mammalian Genome, 2001. **12**(11): p. 837-842.
23. Lindersson, M., et al., *Mapping of serum amylase-1 and quantitative trait loci for milk production traits to cattle chromosome 4.* Journal of Dairy Science, 1998. **81**(5): p. 1454-1461.
24. Longeri, M., et al., *Short communication: Quantitative trait loci affecting the somatic cell score on chromosomes 4 and 26 in Italian Holstein cattle.* Journal of Dairy Science, 2006. **89**(8): p. 3175-3177.
25. Kirkpatrick, B.W., B.M. Byla, and K.E. Gregory, *Mapping quantitative trait loci for bovine ovulation rate.* Mammalian Genome, 2000. **11**(2): p. 136-139.
26. Holmberg, M. and L. Andersson-Eklund, *Quantitative trait loci affecting health traits in Swedish dairy cattle.* Journal of Dairy Science, 2004. **87**(8): p. 2653-2659.

27. Cruickshank, J., et al., *Evidence for quantitative trait loci affecting twinning rate in North American Holstein cattle.* Animal Genetics, 2004. **35**(3): p. 206-212.
28. Lien, S., et al., *A primary screen of the bovine genome for quantitative trait loci affecting twinning rate.* Mammalian Genome, 2000. **11**(10): p. 877-882.
29. Hiendleder, S., et al., *Mapping of QTL for body conformation and behavior in cattle.* Journal of Heredity, 2003. **94**(6): p. 496-506.
30. Chen, H.Y., et al., *Detection of quantitative trait loci affecting milk production traits on bovine chromosome 6 in a Chinese Holstein population by the daughter design.* Journal of Dairy Science, 2006. **89**(2): p. 782-790.
31. Freyer, G., et al., *Multiple QTL on chromosome six in dairy cattle affecting yield and content traits (vol 119, pg 60, 2002).* Journal of Animal Breeding and Genetics, 2002. **119**(3): p. 200-200.
32. Kucerova, J., et al., *Multitrait quantitative trait loci mapping for milk production traits in Danish Holstein cattle.* Journal of Dairy Science, 2006. **89**(6): p. 2245-2256.
33. Olsen, H.G., et al., *Fine mapping of milk production QTL on BTA6 by combined linkage and linkage disequilibrium analysis.* Journal of Dairy Science, 2004. **87**(3): p. 690-698.
34. Ron, M., et al., *Multiple quantitative trait locus analysis of bovine chromosome 6 in the Israeli Holstein population by a daughter design.* Genetics, 2001. **159**(2): p. 727-35.
35. Freyer, G., et al., *Search for pleiotropic QTL on chromosome BTA6 affecting yield traits of milk production.* Journal of Dairy Science, 2003. **86**(3): p. 999-1008.
36. Kuhn, C., et al., *Detection of QTL for milk production traits in cattle by application of a specifically developed marker map of BTA6.* Animal Genetics, 1999. **30**(5): p. 333-340.
37. Szyda, J., et al., *Estimation of quantitative trait loci parameters for milk production traits in German Holstein dairy cattle population.* Journal of Dairy Science, 2005. **88**(1): p. 356-367.
38. Velmala, R.J., et al., *A search for quantitative trait loci for milk production traits on chromosome 6 in Finnish Ayrshire cattle.* Animal Genetics, 1999. **30**(2): p. 136-43.
39. Wiener, P., et al., *Testing for the presence of previously identified QTL for milk production traits in new populations.* Animal Genetics, 2000. **31**(6): p. 385-95.
40. Spelman, R.J., et al., *Quantitative trait loci analysis for five milk production traits on chromosome six in the Dutch Holstein-Friesian population.* Genetics, 1996. **144**(4): p. 1799-1807.
41. Weller, J.I., et al., *Detection and analysis of quantitative trait loci affecting production and secondary traits on chromosome 7 in Israeli Holsteins.* Journal of Dairy Science, 2008. **91**(2): p. 802-813.
42. Ashwell, M.S., et al., *Detection of loci affecting milk production and health traits in an elite US Holstein population using microsatellite markers.* Animal Genetics, 1997. **28**(3): p. 216-222.
43. Cobanoglu, O., P.J. Berger, and B.W. Kirkpatrick, *Genome screen for twinning rate QTL in four North American Holstein families.* Animal Genetics, 2005. **36**(4): p. 303-8.
44. Ajmone-Marsan, P., et al., *Identification of milk protein percentage QTLs in Italian Friesian cattle by selective genotyping two GDD families with AFLP and SSR markers.* Italian Journal of Animal Science, 2007. **6**: p. 40-42.
45. Bennewitz, J., et al., *The DGAT1 K232A mutation is not solely responsible for the milk production quantitative trait locus on the bovine chromosome 14.* Journal of Dairy Science, 2004. **87**(2): p. 431-442.
46. Fontanesi, L., et al., *The BovMAS Consortium: investigation of bovine chromosome 14 for quantitative trait loci affecting milk production and quality traits in the Italian Holstein Friesian breed.* Italian Journal of Animal Science, 2005. **4**: p. 16-18.
47. Kim, J.J. and M. Georges, *Evaluation of a new fine-mapping method exploiting linkage disequilibrium: a case study analysing a QTL with major effect on milk composition on bovine chromosome 14.* Asian-Australasian Journal of Animal Sciences, 2002. **15**(9): p. 1250-1256.
48. Kuhn, C., et al., *Evidence for multiple alleles at the DGAT1 locus better explains a quantitative tip trait locus with major effect on milk fat content in cattle.* Genetics, 2004. **167**(4): p. 1873-1881.
49. Riquet, J., et al., *Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 9252-7.
50. Winter, A., et al., *Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA : diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(14): p. 9300-9305.
51. Looft, C., et al., *A mammary gland EST showing linkage disequilibrium to a milk production QTL on bovine Chromosome 14.* Mammalian Genome, 2001. **12**(8): p. 646-650.
52. Muncie, S.A., J.P. Cassady, and M.S. Ashwell, *Refinement of quantitative trait loci on bovine chromosome 18 affecting health and reproduction in US Holsteins.* Animal Genetics, 2006. **37**(3): p. 273-275.

53. Morris, C.A., et al., *Fatty acid synthase effects on bovine adipose fat and milk fat.* Mammalian Genome, 2007. **18**(1): p. 64-74.
54. Arranz, J.J., et al., *A QTL affecting milk yield and composition maps to bovine chromosome 20: a confirmation.* Animal Genetics, 1998. **29**(2): p. 107-115.
55. Blott, S., et al., *Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition.* Genetics, 2003. **163**(1): p. 253-266.
56. Viitala, S., et al., *The role of the bovine growth hormone receptor and prolactin receptor genes in milk, fat and protein production in Finnish Ayrshire dairy cattle.* Genetics, 2006. **173**(4): p. 2151-2164.
57. Elo, K.T., et al., *A quantitative trait locus for live weight maps to bovine Chromosome 23.* Mammalian Genome, 1999. **10**(8): p. 831-835.
58. Gautier, M., et al., *Fine mapping and physical characterization of two linked quantitative trait loci affecting milk fat yield in dairy cattle on BTA26.* Genetics, 2006. **172**(1): p. 425-436.
59. Van Tassell, C.P., T.S. Sonstegard, and M.S. Ashwell, *Mapping quantitative trait loci affecting dairy conformation to chromosome 27 in two Holstein grandsire families.* J Dairy Sci, 2004. **87**(2): p. 450-7.
60. Sandor, C., et al., *Linkage disequilibrium on the bovine X chromosome: characterization and use in quantitative trait locus mapping.* Genetics, 2006. **173**(3): p. 1777-86.
61. Ron M., Heyen D.W., Weller J.I., Band M., Feldmesser E., Pasternak H., Da Y.,Wiggans G.R., VanRaden P.M., Ezra E., Lewin H.A., Detection and analysis of a locus affecting milk concentration in the US and Israeli dairy cattle populations, in: Proceedings of the 6th World Congress on Genetics Applied to Livestock Production, Armidale, Vol. 26, 1998, 6 WCGALP Congress Office, University of New-England, Armidale, pp. 422–425.
62. Chamberlain A., McParlan H., Balasingham T., Carrick M., Bowman P., Robinson N., Goddard M., Mapping QTL affecting milk composition traits in dairy cattle using a complex pedigree, Proceedings of the 7thWorld Congress onGenetics Applied to Livestock Production, Montpellier, France, 19–23 August 2002, ISBN 2-7380-1052-0, Paper09-08.

CHAPTER 3

EVALUATION OF SOME FACTORS OF VARIABILITY OF ACCURACY OF GENOMIC BREEDING VALUES

ABSTRACT

With the aim to assess the effect of some factors of variability of accuracy of direct genomic value prediction (DGV) a simulation of animal population have been carried out. Two thousand animal of training (animal with genotypes and phenotypes) and 6000 animal of prediction spanning over 3 generation have been used to assess the effect on DGV accuracy of the following factors: *i*) heritability, *ii*) number of marker, *iii*) number of Daughter per bull to calculate DYD, *iv*) number of QTL and *v*) generation of random mating. DGV accuracies during the generations showed a downward trend, higher drop have been reported to single marker vs haplotype. The DGV accuracy increased with the heritability. The influence of the density of markers on DGV accuracy was positive as well. The analysis of the effect of different number of daughters per bull on the accuracy of DGVs showed that the number of daughters per bull selection schemes currently used in progeny testing (50 to 120 daughters per bull tested) is sufficient to obtain good accuracy of genomic prediction. Furthermore, the number of QTL had a limited positive effect on accuracy of genomic prediction, whereas the number of generation of random mating does not show any clear pattern.

Key words: Direct genomic values, accuracy of prediction, genomic selection, DYD

3.1 Introduction

In the recent years, the use of marker assisted selection programs in livestock has been constrained by poor knowledge on causal mutations affecting the expression of traits of economic interest (Dekkers 2004). Dense SNP maps allowed the prediction of Direct Genomic Breeding Values (DGV) based on the estimation of SNP genotype effects on the considered trait using a genome-wise approach (Meuwissen *et al.* 2001). Briefly, Genomic Selection (GS) rely on the segmentation of the genome using a dense marker map in thousands of bits, each contributing to the explanation of part of the genetic variance of a quantitative trait. The effect of each segment is estimated in a training population (animal with phenotypes and genotypes). Then effects are used to predict the breeding values of prediction population (animals without phenotypes).

Possible advantages of GS over the conventional selection are the reduction of the generation interval, the increase of accuracy of the female side of the pedigree (Schaeffer 2006) and the reduction of costs for progeny testing (Konig *et al.* 2009). One major issue in DGV estimation is represented by the large number of predictors (for example 50K SNPs for cattle) and the relatively small number of records available. Furthermore, accuracy of genomic prediction are affected by the methods used to estimate the marker effects but also from the genetic model adopted and its variability factor. The study of the factor affecting the variability of DGV is by simulation may be useful to improve the accuracy of genomic prediction.

3.1.2 Factor affecting accuracy of genomic prediction

There are many factor that affecting the DGV estimates. Accuracy of genomic prediction is generally done measuring the correlation between the DGV and the true breeding value (TBV) alternatively, evaluation may be done on the squared correlation ( )(VanRaden *et al.* 2009). Additional criteria to evaluate the genomic prediction is generally the bias of prediction measured by the regression coefficient $b_{TBV,GEBV}$ between phenotype and DGV.

Some feature of the reference population may affect the accuracy of genomic prediction. Number of animals in the reference population (Meuwissen *et al.* 2001; Muir 2007; Hayes *et al.* 2009b), number of markers and the level of LD (Calus *et al.* 2008; Solberg *et al.* 2008), heritability of the trait considered (Meuwissen *et al.* 2001; Kolbehdari *et al.* 2007) are for instance factors known to have an effect on the accuracy of DGV. Moreover, additive genetic relationship in the reference population captured by the SNP influence the accuracy of genomic predictions both in simulated and real data (Habier *et al.* 2007; Habier *et al.* 2010). The choice of the statistical model and its parameterization (single markers or haplotypes) affects the accuracy of prediction as well (Calus *et al.* 2008; Hayes *et* *al.* 2009b) showed also similar relationship between accuracy of genomic predictions and number of phenotypic records in the reference population. The accuracy of DGV increases for increasing values of heritability and number of markers. Furthermore, the higher the number of animals with both genotypes and phenotypes the higher DGV accuracy in the prediction population. To assess the extent of some factors known to have an effect on DGV accuracy a simulation was carried out.

The objective of the present work was to evaluate some of the factor that may affect the accuracy of DGV estimation. In particular, the effect of the heritability of the trait, of number of markers, underlying QTLs, daughters per bull used in the calculation of daughter yield deviation (DYD), and number of previous generations of random mating on the accuracy of genomic prediction using a wide range of values were investigated.

3.2 MATERIALS AND METHODS

In order to assess the main factors of variability that affect the accuracy of direct genomic breeding values estimation (DGV) an animal population was simulated using the programming language FORTRAN 95. In attempt to reproduce the main genetic and technical features of a simplified animal population, the simulation was divided into five main stages (Figure 1):
i) creation of a base population; *ii*) random mating of the initial population for n generations; *iii*) estimation of marker effects and DGV calculation in the training population (TRAIN, *n*); *iv*) DGV estimation in the generations of prediction for the animals without phenotypes (PRED, *n + i*) using the haplotype (or SNP) effects estimated in the TRAIN population; *v*) assessment of the accuracy of DGV estimation using correlation between true breeding values of animals (TBV) and DGV.

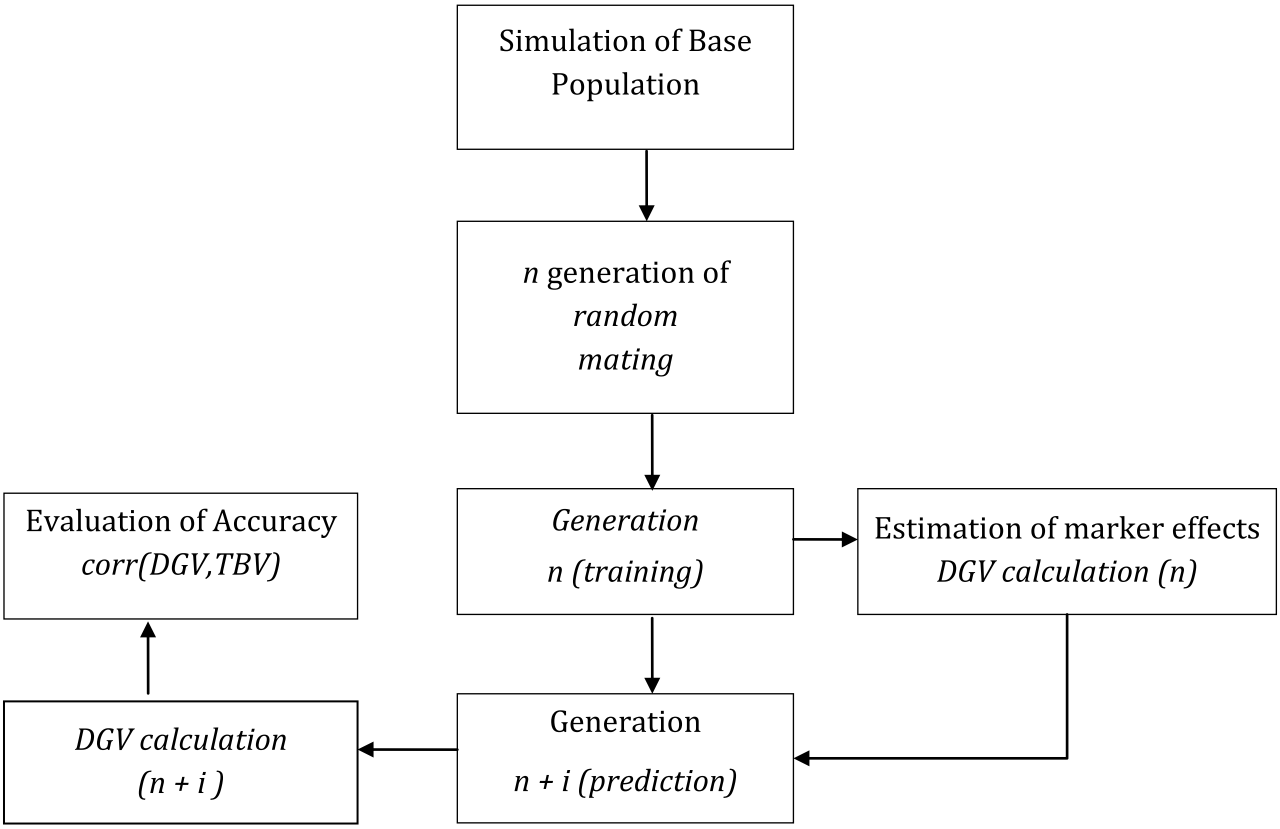

Simulation of Base Population
n generation of random mating
Generation n (training)
Estimation of marker effects DGV calculation (n)
Evaluation of Accuracy corr(DGV,TBV)
DGV calculation (n + i)
Generation n + i (prediction)

Figure 1. Basic scheme of the simulation

3.2.1 Set up the population

A base population of 1000 females and 50 males was generated. For each animal, two chromosomes of one Morgan (M) each were simulated (one of paternal and one of maternal origin). The creation of the chromosomes was carried out following the stages below described.

Assignment of allele frequencies at SNPs and QTLs.

The first step was the random assignment of the genotypes for each SNP and QTL. Both QTLs and SNPs were supposed to be bi-allelic and alleles were coded as 1, 2 (SNP) and 11, 12 (QTL). The allelic frequencies of QTLs and SNPs were sampled from a uniform distribution (0,1). Figure 2 shows the part of the Fortran code used to define the allele frequencies. Then, QTL values were assigned to the QTL alleles assuming that each QTL explained a different proportion of genetic variance. Individual QTL variance was sampled from a gamma distribution with parameters (scale α=1.66; shape β=0.42) according to MEUWISSEN *et al.* (2001) Since the gamma distribution provides only positive values, the signs of the values of QTL effect were randomly assigned. The simulated phenotypic variance was set to 100 and the genetic variance was obtained according to the value of heritability chosen. True breeding values were calculated summing up the QTL effects across the whole genome. The phenotypic values were obtained adding random noise to the TBV, further detail will be provided later.

```
!allelic frequency calculation (SNPs and QTLs)

DO i=1,nmar
  fSNP(i)=0.5
  DO j=1,TimeArray(3)!randomize the random generator
   CALL random_number(fSNP(i))!
  ENDDO
ENDDO

!frequenza del primo allele

DO i=1,nqtl
  DO j=1,TimeArray(3)
       CALL random_number(fall_qtl(i))!
  ENDDO
ENDDO
```

Figure 2 Calculation of allele frequencies of SNPs and QTL in the population.

Generation of chromosomes

Once that simulation parameters were set, the positions of QTLs and SNPs were assigned along the chromosome using odd and even position for SNP and QTL respectively (Figure 3). The base population was created through the construction of 100 male chromosomes and 2000 female chromosomes (sex ratio 1:20) using the parameters generated in the aforementioned steps.

An example of the data files generated to estimate genomic breeding values is provided in figure 4. For each animal, the 2 lines (paternal and maternal chromosome respectively) reports the genotypes at n SNP markers loci. Individual Contents of phenotypes and TBV files have been also reported.

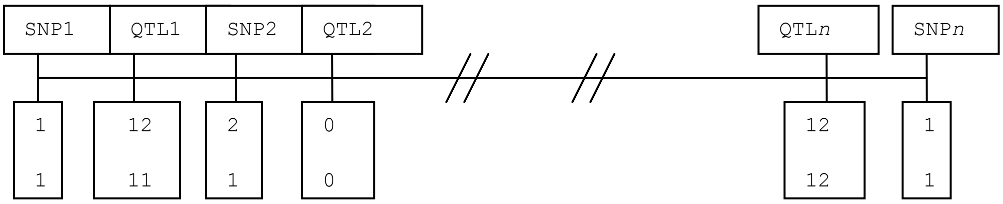
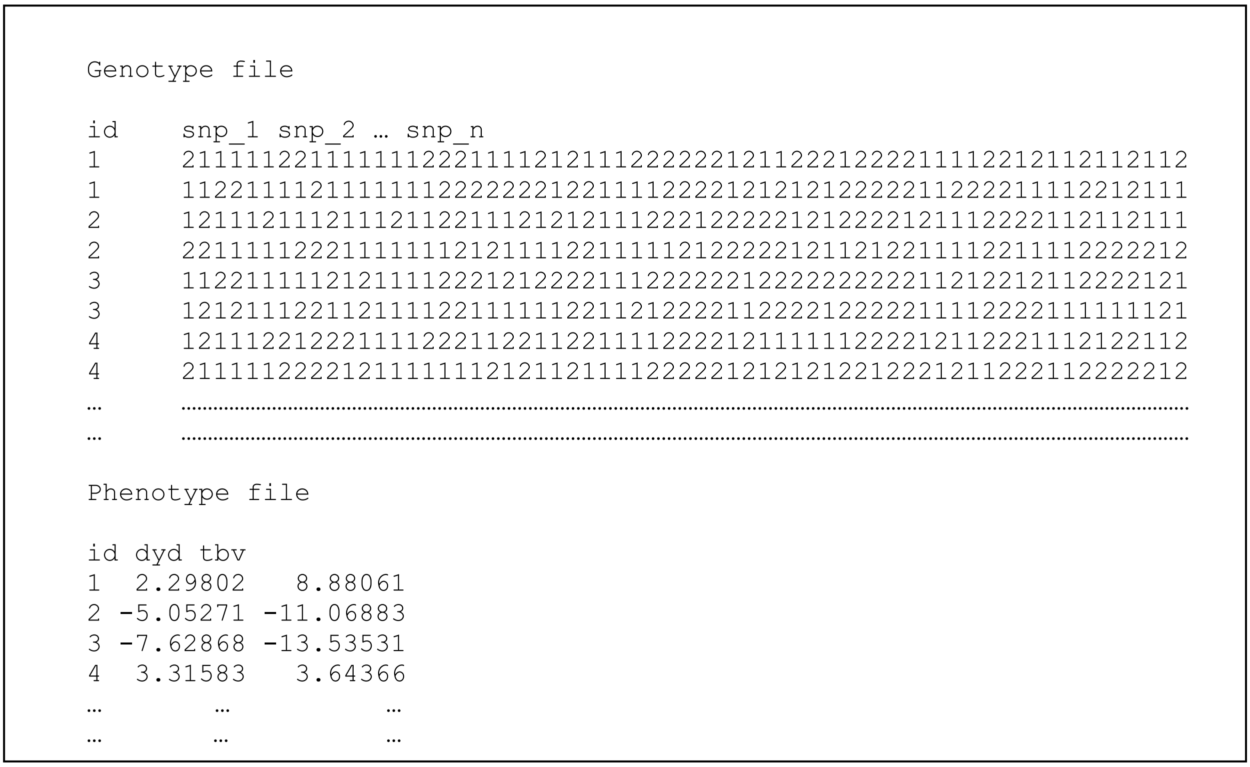


Figure 3. Scheme of a Chromosome



```
Genotype file

id    snp_1 snp_2 … snp_n
1     211111221111111222111121211122222212112221222211112212112112112
1     112211112111111122222221221111222212121212222211222211112212111
2     121112111211121122111212121112221222221212222121112222112112111
2     221111122211111112121111221111121222221211212211112211112222212
3     112211111212111122212122221112222221222222222211212212112222121
3     121211122112111122111111221121222211222212222211112222111111121
4     121112212221111222112211221111222212111111222212112221112122112
4     211111222212111111121211211112222212121212212221211222112222212
…     ...........................................................................
…     ...........................................................................

Phenotype file

id dyd tbv
1  2.29802   8.88061
2 -5.05271 -11.06883
3 -7.62868 -13.53531
4  3.31583   3.64366
…       …          …
…       …          …
```

Figure 4 Example of the files that the program generates to estimate breeding values genomic population simulated.

3.2.2. Generation of Random mating

Once created the base population, the program generated the new animals through iterations until a specific condition was met. This Fortran code (Figure 5) was used to perform a certain number of generations of random mating and to replace animals according the culling rate. Furthermore, for each generation, the program calculates the value of linkage disequilibrium (LD) in the population by calculating the statistic r^2. Firstly, allele frequencies were calculated (denominator of the formula), secondly D^2 was computed counting the haplotype of contiguous SNP (numerator of the formula). Chromosomes of the next generations were created by sampling the first allele at j-th locus with a probability of 50% from parental chromosomes. In the following loci, alleles were sampled on according to the recombination rate (θ). Haldane mapping function was used to calculate the recombination ratio according to the distance between markers. Each QTL was supposed to be in the middle of the interval bracketed by contiguous markers. The animals for the next generation were chosen according to the culling rate (fixed at 50%) by randomly selecting 50% of the population which replaced the culled animals.

```
 cont=1

DO Generation=1,ngen

..............................
..............................

     Mating(crom1male,crom1fem,crom2male,crom2fem)
     Replacing(crom1male,crom1fem,crom2male,crom2fem)
     LD(crom1male,crom1fem,crom2male,crom2fem)

..............................
..............................
 cont = cont +1

ENDDO
```

Figure 5. DO loop to perform the random mating

3.2.3 Estimation of markers effect in the training population

Twenty five generations of random mating were performed. After that, the population size was expanded to 2000 animal. Two thousand male were used as training population to estimate the effects of chromosomal segments. The TBV of selected animals were calculated summing up the values of the chromosome segment, previously assigned (Figure 6).

The next step was to create the phenotypes of the daughters. A value of DYD (Daughter Yield Deviation) expressed as a deviation of the average production of daughters was assigned to each bull. Firstly, half of the genetic contribution from his father (DYgen) was calculated for each daughter, then DYD were calculated sampling from a normal distribution N ~ (DYgen,) where was the phenotypic variance (Figure 6)

```
!TBV Calculation
 DO i = 1,n_animals
       TBV(i) = sum(QTL_val1(i,:))+ sum(QTL_val2(i,:))
 ENDDO

 !CALCULATION OF DYD SAMPLED FROM NORM ~(DYgen,Var_PHEN)

 DO i = 1,n_animals
       DO j = 1,DYDperbull
         DYphen(i,j)= gen_norm(DYgen(i,j),var_Phen)
       ENDDO
 ENDDO

 DO i = 1,n_animals
       DYD(i) = sum(DYPhen(i,:))/DYDperbull
 ENDDO
```

Figure 6. TBV Calculation and bulls' DYD sampling step coded in Fortran

The estimation of the markers effects and DGV calculation for the training generations were carried out using two kind of predictors: haplotypes of contiguous markers (HAP) or marker genotypes (SNP). The HAP approach implies the construction of haplotypes of contiguous markers along the genome. The effect of all possible haplotypes (Figure 7) bracketed by adjacent SNPs were estimated using BLUP methodology and treating the haplotypes as random effects:

Where **DYD** is the vector of phenotypes, μ is the overall mean, **Hap**$_j$ is the random effect of the *j*-th haplotype and **e** the vector of random residual, associated to diagonal covariance matrices **G**~(0,) and **R**~(0,). The contribution of each SNP haplotype to the variance of the trait was assumed to be equal (i.e., σ^2_{gi}= σ^2_g /number of haplotypes,).

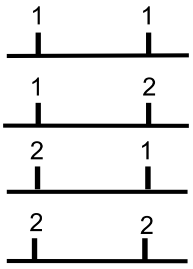


Figure 7 Sample of haplotypes of pair of contiguous markers

In the second approach the DGV were calculated using regression of DYD on individual SNPs using the BLUP method and treating the SNP genotype as random effect. In this case the model used was

where **DYD** is the vector of phenotypes, is the overall mean, **SNP** is the random effect of genotype in the *j*-th SNP and **e** is vector of the random residual. **SNP** and **e** were associated to covariance matrices **G**~(0,) and **R**~(0,) respectively. Also in this case, the contribution of each SNP genotype to the variance of the trait was assumed to be equal.

The DGV for the i-th animal of the training population were calculated using the formula

Where is the overall mean, **Z** is the incidence matrix of predictors (**SNP** or **Hap**) and **ĝ** the vector of solutions (**SNP** or **Hap**) respectively.

3.2.4 Calculation of DGV for prediction generations

Three generations of prediction (PRED1, PRED2 and PRED3) were simulated using random mating and replacement rate of 50% and no selection was performed. Once calculated, TBVs of the generation n+ i (i ≤3), genotypes of the candidate bulls were generated and DGVs were calculated. In the successive generations, DGV were calculated using the values of SNP effects estimated in the TRAIN generation according to:

where the **Z** matrix is the incidence matrix of **HAP** (or **SNP**) for the animals without phenotypes.

3.2.5 Accuracy of DGV

The accuracy of the genomic estimation both in the generation of training and prediction was evaluated using the correlation between TBV and DGV as follow

3.2.6 Simulated scenarios

In the Table 1 are reported the input values used to run the simulation. The population in the base scenario was structured as reported in Table 2. The size of the initial population was 1000 females and 50 males with a sex ratio of 1:20, the replacement rate was 50% each generation (each generation is replaced 50% of animals both males and females). Each bull was evaluated on the basis of the phenotype of 50 daughters and expressed as DYD. Each animal was represented by two chromosomes, one of paternal origin and one of maternal origin, with dimensions of 100 cM. The number of markers per chromosome was equal to 301 which divided the genome into 300 bits. Twenty bi-allelic QTLs were randomly distributed across the chromosome and h^2 of 0.5 were simulated. The rate of recombination, equal to 0.0016, was calculated using the Haldane function $\theta = ½\ (1\text{-}e^{-2d})$, where d= map distance. For each scenario 30 replicates were performed.

Table 1. Inputs of simulation in the base scenario.

Symbol	Parameter of the simulation	Values
Nmarkers	Number of markers	301
Intervals	Number of intervals bracketed by contiguous SNP	300
nQTL	Number of QTL	20
Genomel	Length of the genome expressed in cM	100
Nmales	Number of males	50
Nfemales	Number of females	1000
DYDperbull	Number of daughters per bull	50
Ngenrm	Number of generations of random mating	25
Nrepl	Number of Replicates	30
h^2	Heritability	0.5
Nalleli	number of alleles per QTL	2
Recombination rate	Θ	0.0016

The output of the program are different but the most relevant processed during the simulation were:

i) genotypes and phenotypes of training individuals;

ii) genotypes and phenotypes of prediction individuals;

iii) location and values of QTL;

iv) level of linkage disequilibrium in all generations (r^2).

Final results of the program were the accuracies expressed as correlations between TBV and DGV both for training and prediction generations using both haplotypes and single marker genotypes.

Scenarios

Starting from a population with the characteristics described above, five different scenarios were simulated. In all scenarios the size of animals (1000 males and 50 females) and the number of alleles per QTL were kept constant. The number of markers, number of QTLs, the number of daughters per bull (DYD), the heritability of the trait and the number of generations of random mating were changed at once for each scenario simulated.

The table 3 provide a description of the parameter used in the simulated scenarios.
i) In the first scenario, the effect of the heritability of the traits on the accuracy of DGV was evaluated. Very low heritability values (0.1) up to the extreme values (0.9) were tested; *ii*) in the second scenario the effect of the number of markers on correlations between TBV and DGV was evaluated. From 151 markers to 1001 markers per chromosome were simulated; *iii*) in the third scenario the number of daughters per bulls required to estimate the genetic merit of bull (DYD) was varied (from 10 daughters per bull to 150); *iv*) in the fourth simulation the effect of the number of QTL was tested (from 15 QTL up to 25 with an increment of 5). It was also tested a more extreme scenario with 100 QTLs; *v*) in the fifth scenario the effect of number of previous random mating generation were simulated, let the number of generation of random mating varying from 5 to 100 generations.

Table2. Parameters of the simulation for different scenarios simulated

Parameter	h^2	Markers	DYD	QTLs	N gen RM
h^2	0.1-0.3-0.5-0.7-0.9	301	50	20	25
Markers	0.5	151-201-251-301-351-401-801-1001	50	20	25
DYD	0.5	301	10-30-60-90-120-150	20	25
QTLs	0.5	301	50	10-15-20-25-100	25
N gen RM	0.5	301	50	20	5-15-25-50-100

3.3 RESULTS AND DISCUSSION

The basic assumptions of the simulation have been verified. A first control on the distribution of variability explained by each QTL, assigned by sampling from a gamma distribution (α = 1.66, β = 0.4 according to MEUWISSEN *et al.* (2001) was carried out considering the estimated values of individual marker effects. Figure 8 shows, for a single replicate of a single scenario, the values of the estimated effects for 301 SNPs across the genome simulated. Most of the loci have a value between 0 and 0.7, while only a few SNPs have a value greater than 1 and a marker with a value much larger than all the others were found. A similar situation was described for fat percentage in dairy cattle (GRISART *et al.* 2002). The trait simulated was supposed to be under control of a polygenic system that include a gene with major effect

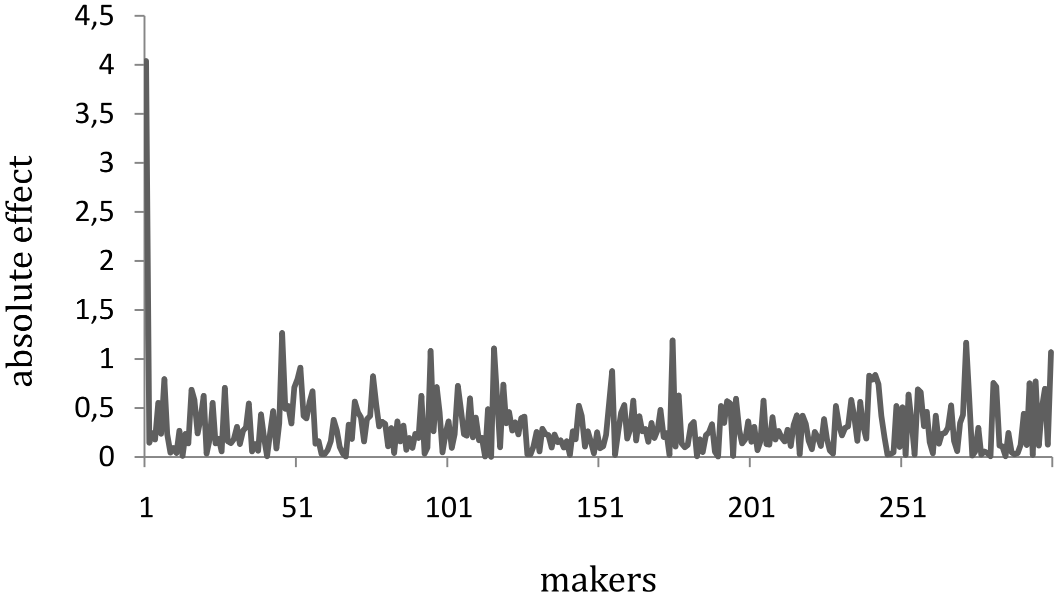


Figure 8 Estimates of effect for all the SNP markers

Correlations between GEBV, TBV were computed both for the generations of training and for those of prediction. Furthermore, the regression of DGV on TBV have been calculated to test the program (Figure 9).

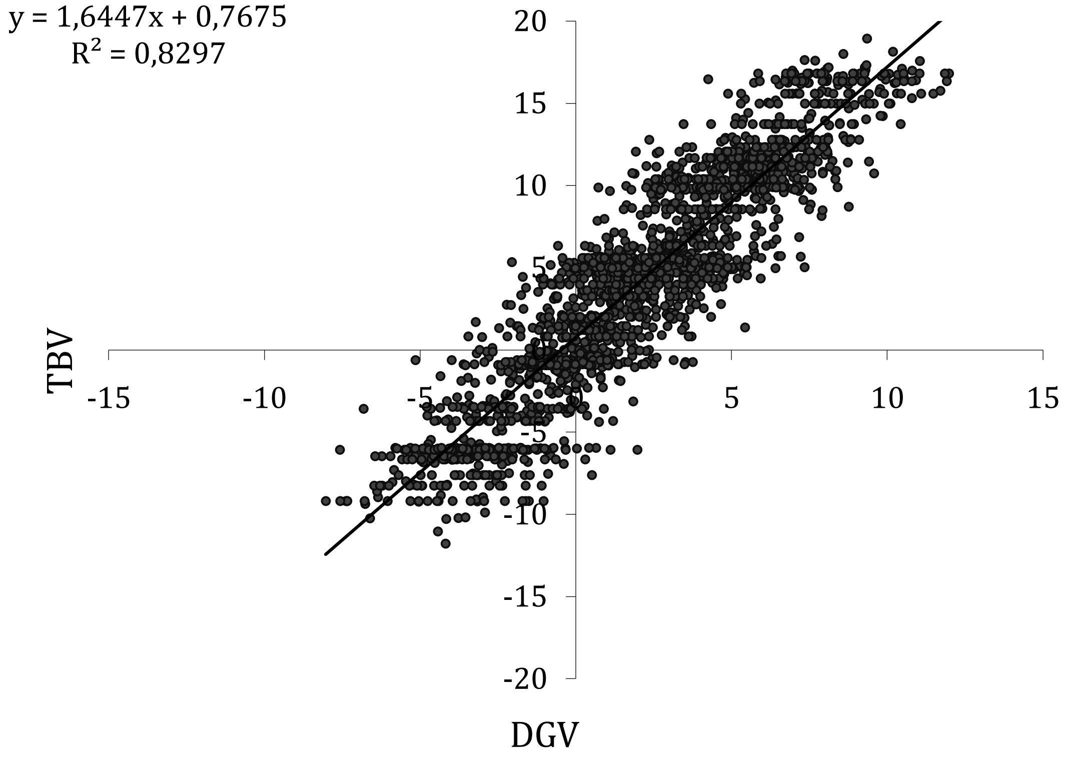


Figure 9 Regression of DGV on true breeding values (TBV) generated by simulation for 2000 animals of training population for the base scenario.

Correlation between DGV and TBV calculated for the training population (table 3) were high and statistically significant. The correlation between DGV and TBV for the generation of prediction are presented in table 4. The correlations in the prediction generation (Table 4) are systematically lower than training, probably due to the fact that the association between marker and QTL was broken down by recombination. The correlation drop from PRED1 to PRED3 even though

Table 3. Correlations (p-value) among DGV, TBV in the generation of training (haplotype above the diagonal, single marker below the diagonal)

	DGV	TBV
DGV	*	0.913 (p<0.001)
TBV	0.899 (p<0.001)	*

Table 4. Correlations (standard deviations) between DGV and TBV in the generations of prediction for the base scenario.

	DGV accuracy		
	PRED1	PRED2	PRED3
HAP	0.829 (0.030)	0.754 (0.046)	0.669 (0.069)
SNP	0.682 (0.059)	0.520 (0.071)	0.404 (0.082)

the loss of accuracy is less when the HAP approach were adopted. The SNP approaches gave lower figure due to the effect of marker density as shown by CALUS *et al.* (2008) that demonstrate at lower marker density the estimates using haplotype are more accurate than single markers.

These results indicate the goodness of the method used to estimate the value of individual segments of the genome (HAP or SNP). The method is able to capture a significant proportion of the variance of the simulated trait, but also the ability of simulation to reproduce the phenomenon under study. The correlation between TBV and DGV in the first generation of PRED1 (Table 4) is lower than the previous generation of predictions, due to allelic recombination, occurred during meiosis, between some markers and QTL associated. The calculation of DGV for the new generations using the values estimated in the previous generation leads to the introduction of an error due to the fact that these segments (haplotypes or markers) are no longer the associated with the QTLs. The decrease in correlation is also confirmed for later generations in agreement to the findings of MUIR (2007). Under the assumptions adopted the simulation can be suitable to describe the inheritance of trait controlled by polygenic complex and, therefore, to study the factors of variability of DGV.

3.3.1 Effect of heritability on accuracy of DGV

Figures 10 and 11 show the correlations between TBV and DGV as function of the heritability for HAP and SNP approach respectively. The correlations show an upward trend for increasing values of heritability and for all generations, with a marked increase for heritability between 0.1 and 0.5. The accuracy reach a plateau for values of heritability higher than 0.5. Similar results were obtained by different authors with simulated data (CALUS and VEERKAMP 2007; KOLBEHDARI *et al.* 2007) and these finding are in agreement with theoretical estimates (DAETWYLER *et al.* 2008).

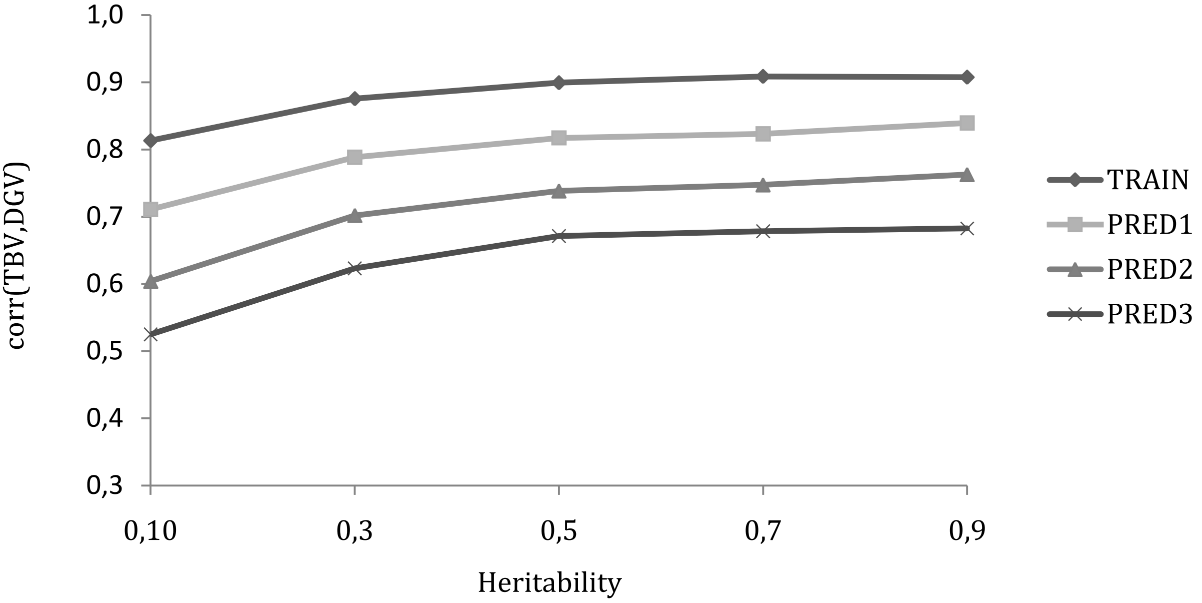


Figure 10. Trends of DGV accuracy as function of the heritability of the trait with the HAP approach. Each line represent a different generation.

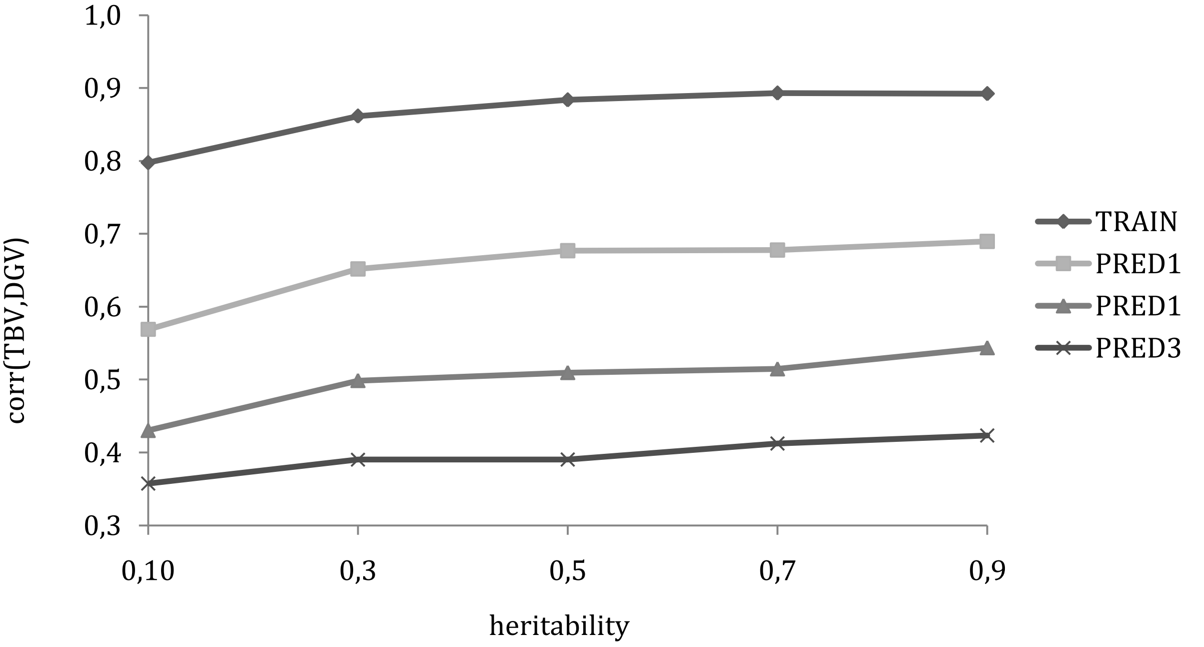


Figure 11. Trends of DGV accuracy as function of the heritability of the trait with the SNP approach. Each line represent a different generation.

DEATWYLER *et al.* (2008) have proposed an analytical approach to predict the accuracy of GEBV in function of certain parameters, according to the following formula:

where is the accuracy in function of $\lambda = N_G/N_P$ (ratio between the number of loci and number of phenotypes) and the observed heritability. Figure 12 shows the trend of theoretical accuracies function of number of available phenotypes (2000) and h^2. These predictions are in agreement with the results obtained with this simulation using the approach based on haplotypes when comparing the number of phenotypes used in the simulation.

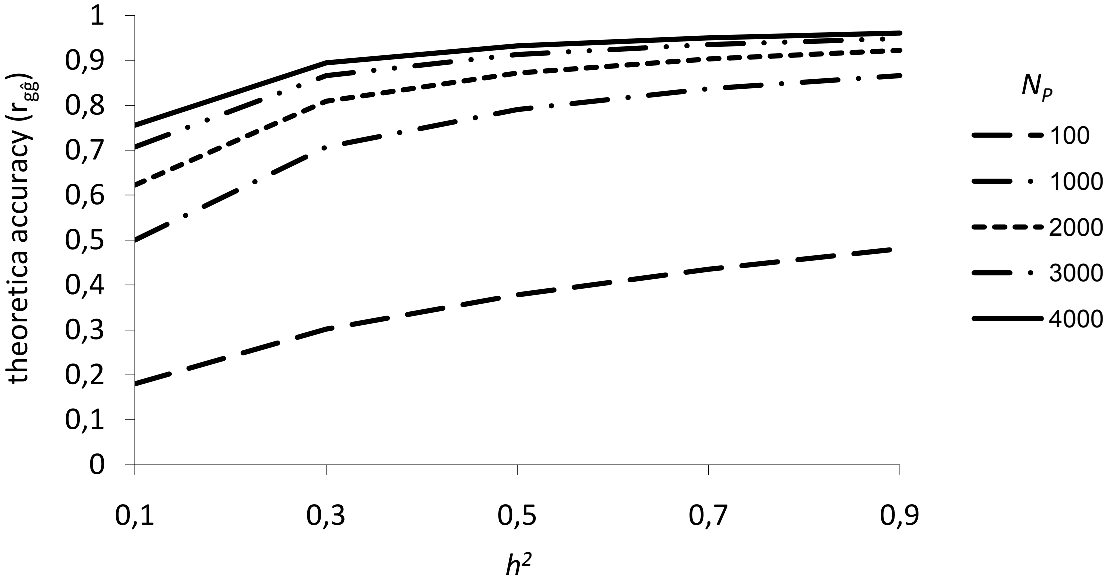


Figure 12 Accuracy of the DGV as function of heritability and number of phenotypes used (N_G = 300 loci) according to (DAETWYLER *et al.* 2008).

The trend of accuracy showed a scale factor, with decreasing correlations, from TRAIN to PRED, and within the prediction generation, from PRED1 to PRED3. This drop in accuracy was probably due to the recombination events that occurs during simulated segregation, and therefore, it became larger through generations. The loss of accuracy is further influenced by the parameterization of the model and the heritability of the trait. The correlations obtained with the SNP approach (Figure 13a) show a downward trend across the generation as well as with the HAP approach (Figure 13b). However, the accuracy in the TRAIN was the same for both approaches and irrespective from heritability (accuracies around 0.8 and 0.9 for heritability of 0.1 and 0.9 respectively both in HAP and SNP). It is worth to notice how the rate of decline of the accuracy is much higher for the SNP approach (from 0.8 to 0.37 and from 0.9 to 0.4 for h^2 of 0.1 and 0.9 respectively) in comparison to the HAP approach (from 0.8 to 0.52 and from 0.91 to 0.68 for h^2 of 0.1 and 0.9 respectively). For the HAP approach seems that the rate of decrease is greater for low heritability values. Whilst in SNP this pattern was not found.

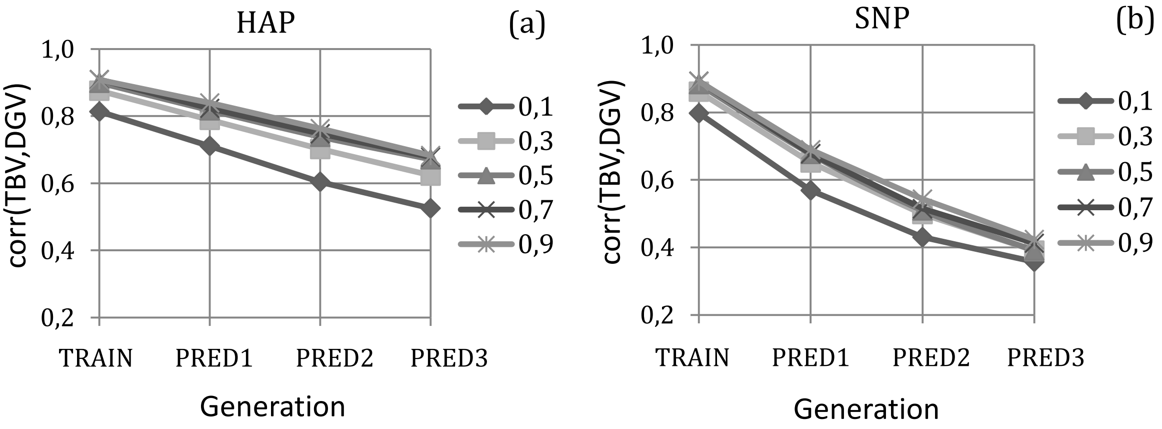


Figure 13 DGV Accuracy over generations for Different h^2 approach with HAP (a) and SNP (b) each line represent a different value of h^2

This may be explained examining previous finding (CALUS *et al.* 2008; CALUS and VEERKAMP 2007; KOLBEHDARI *et al.* 2007; MUIR 2007). In fact, it has been demonstrated that the haplotypes may guarantee a more reliable estimation of DGV at lower marker density in comparison with single markers approaches. Furthermore, CALUS and VEERKAMP (2007) reported that the difference between the two approaches decreased with increasing of LD (the greater difference were found for $r^2 \leq 0.10$) reaching similar accuracies for r^2 values close to 0.2. In the present simulation the average level of LD was found to range between 0.020 to 0.027 – as consequence of the marker density that in the base scenario was about 3 fold less than the density used in CALUS and VEERKAMP (2007). These differences could be explained by this fact.

The average level of LD varies according to the effective population size, for distances between 300 and 400 kb (average distance between markers of this simulation) varies between 0.05 in Human (TENESA *et al.* 2007) and 0.2 in cattle (SVED 1971; ZENGER *et al.* 2007) proposed to calculate the expected level of LD using the following formula:

where N = effective population size, c = distance between markers in Morgan

The r^2 calculated according to the theoretical expectation is equal to 0036 and was 0.01 higher than observed

Summarizing, the DGV accuracies increase at increasing level of h^2 both in TRAIN and PRED generation. Moreover, the average correlations in PRED were systematically lower when SNP was applied that also showed the greater decrease in average correlation across generation. The differences between single marker and haplotypes could be largely due to the low level of LD generated by the simulation, is due to actual size is not comparable to that of livestock populations, which due to lack of selection in the simulated population. To evaluated the effect of the marker numbers in the next paragraph are considered increasing number of markers to estimate the DGV.

3.3.2 DGV accuracy as function number of markers

Figures 14 and 15 show the trend of DGV accuracy depending on the number of markers used for HAP and SNP approaches respectively. The accuracy for generation of training ranges from 0.84 (151 markers) to 0.94 (1001 markers) and 0.81 (151 markers) to 0.94 (1001 markers) when using the single SNP and HAP approach respectively. Whilst, the two approaches for PRED generation gave very different results. The accuracy increases with the number of markers for both methods although to a lesser extent for the single marker approach. It seems that PRED individuals (PRED3 to PRED1 in order) experience a greater advantage in higher marker density rather than TRAIN individuals, both in HAP (figure 14) and SNP methods (figure 15). Furthermore, the accuracy decreases systematically moving from TRAIN to PRED3. The accuracy drop is smaller in the case of the use of haplotypes: from TRAIN to PRED3 average accuracies ranged from 0.81 to 0.36 (101 to 1001 markers) and from 0.94 to 0.51 (101-1001 markers) for SNP approach; accuracies ranges were 0.84-0.53 (101-1001 markers) and 0.94-0.80 (101-1001 markers) for HAP approach instead. The increase in the number of markers, being constant the genome length, implies an increase in the density of markers and an increased level of LD. The differences in accuracy when using haplotypes instead of individual markers are negligible for high levels of r^2 ($r^2 > 0.4$). In contrast to very low levels of r^2 haplotypes provide an higher level of accuracy than individual SPN markers as previously reported (CALUS *et al.* 2008). Variations of the accuracy rely on the number of markers, this finding is in agreement with result found by SOLBERG *et al.* (2008) that showed higher accuracy for higher marker density. However, MUIR (2007) showed a decrease of accuracy as the number of markers increased. Compared to their findings, the results obtained in the present simulation partially overlapped for the generation of training. This can be partly explained by the assumptions made in this simulation.

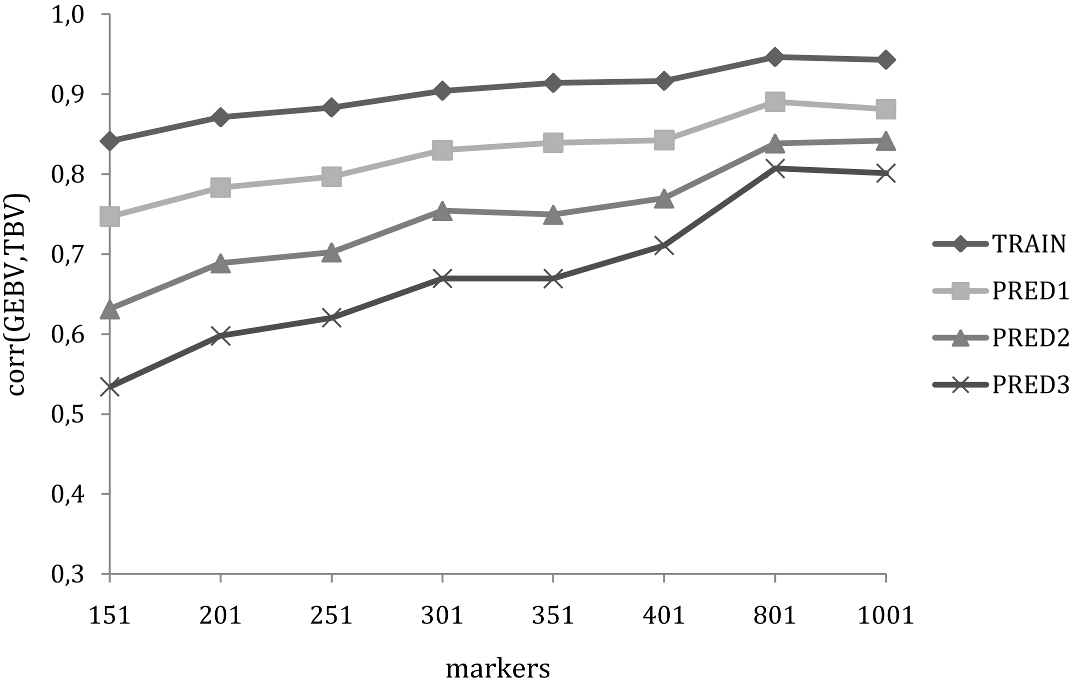


Figure 14 DGV accuracy as function of number of markers with the HAP approach for training and prediction generation

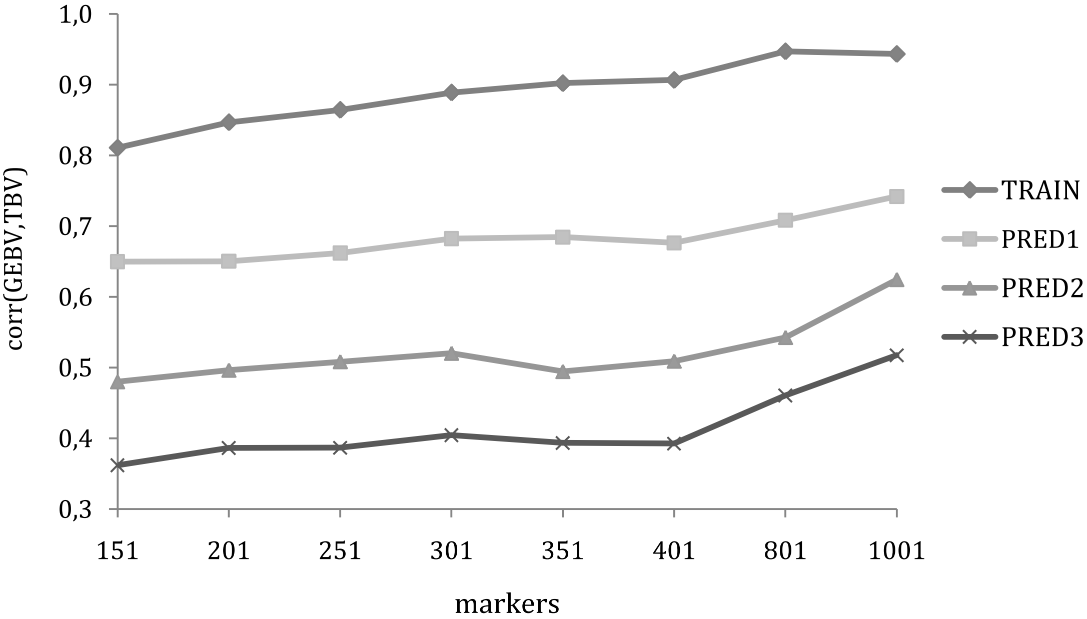


Figure 15 DGV accuracy as function of number of markers with the SNP approach for training and prediction generation

3.3.3 DGV accuracy and number of daughters per bull (DYD)

The figures 16 and 17 show DGV accuracies calculated with different number of daughters per bull (from 10 up to 150) using haplotypes and marker genotypes, respectively. The correlation between TBV and DGV increases markedly up to 60 daughters per bull, beyond this limit increases in accuracy are lower. The average DGV accuracy ranged from 0.76 to 0.90 (TRAIN) and from 0.50 to 0.72 (PRED3) for increasing number of daughters when HAP method were used (figure 16). Although, no difference in the DGV accuracies were found increasing the number of daughters for the TRAIN individuals, marked drop in accuracy were experienced in the PRED generation (figure 17). At least 60 progeny seems to be required to obtain a acceptable level of accuracy. The upward trend of DGV accuracy as function of number of daughter per bull found in the present study confirms the findings of other authors (SCHAEFFER 2007, GUO *et al.* 2010).

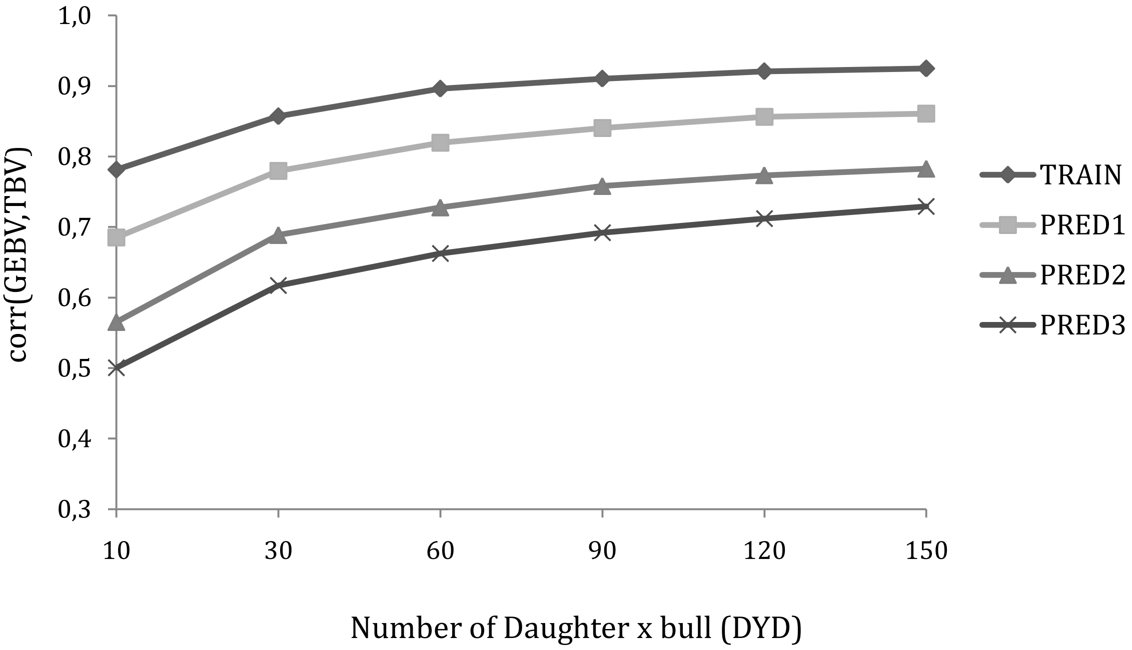


Figure 16 DGV accuracy as function of number of daughter x bull with the HAP approach for training and prediction generation

For the same scenarios, the DGV accuracy in PRED generations using SNP approach had lower values and showed greater decrease across generations (figure 18a) when compared with the HAP approach likewise for the previous scenario (figure 18b).

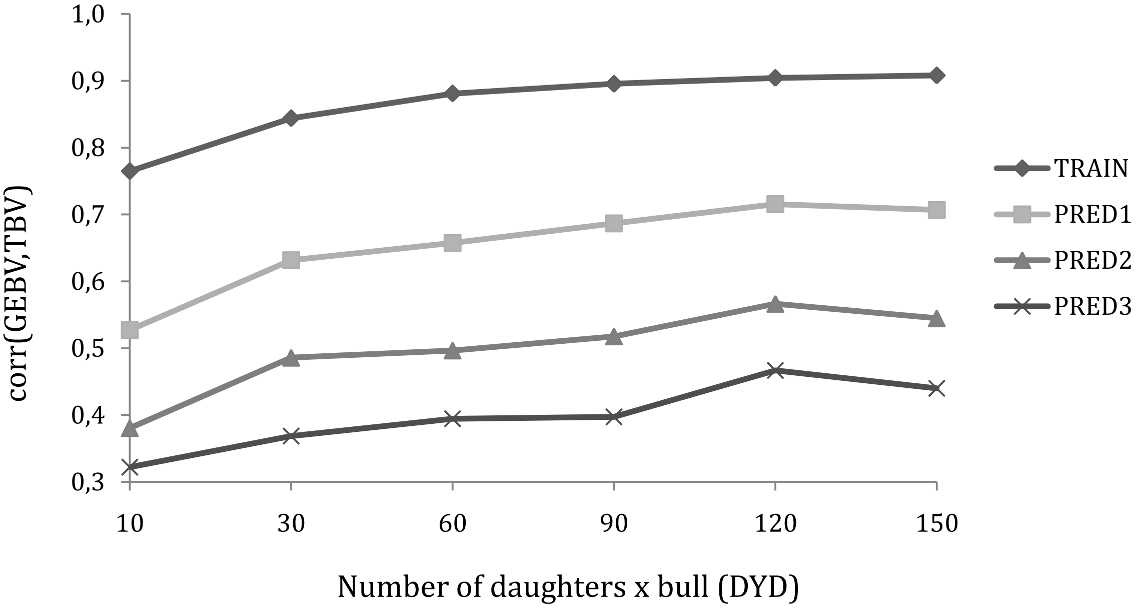


Figure 17. DGV accuracy as function of number of daughter x bull with the SNP approach for training and prediction generation

The residual variance was supposed to be equal for all the DYDs but in actual data each bull may have a different number of daughters. This fact may affect the estimation of DGV if not correctly taken into account, for example weighting the phenotype by number of daughters. (GUO *et al.* 2010) in a simulation study showed that when different number of daughters were considered and DYD were weighted by their reliability, different results across different methods. were obtained In particular, these authors found that DGV accuracy increases around 3-4% (depending on estimation methods) when the DYD were weighted by their reliability.

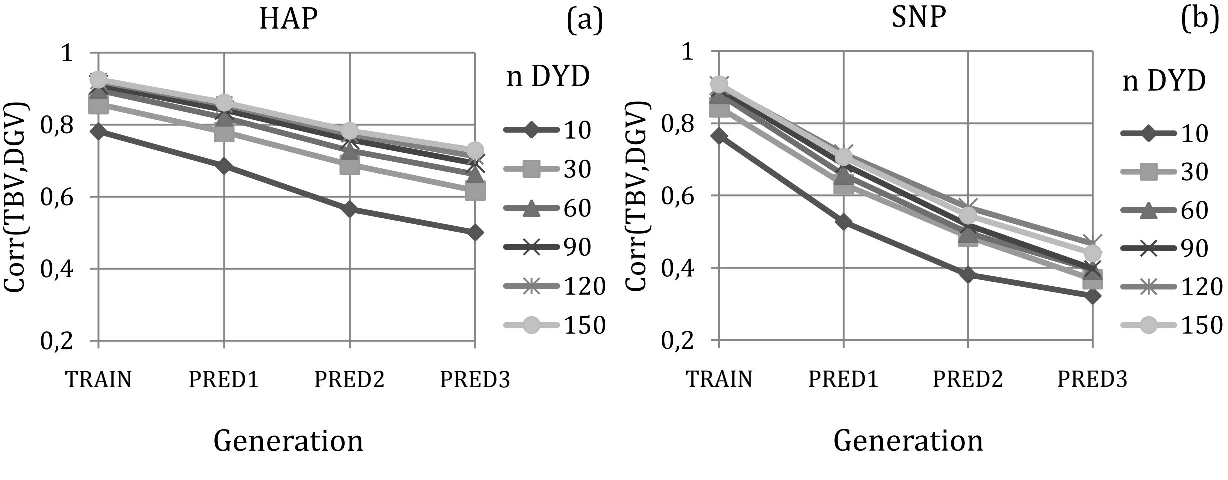


Figure 18. DGV accuracy across generation as function of number of Daughter per bull (from 10 to 150) with the HAP (a) and SNP approach (b) each line represent a different number of daughter per bull.

3.3.4 DGV accuracy and number of QTL

The average DGV accuracies across generations as a function of the number of QTL spread in the genome is reported in figure 19a and figure 19b for the HAP and SNP approach respectively. The accuracy of genomic predictions is slightly affected by the number of simulated QTLs. It is worth to notice that only a slight increase of DGV accuracy for increasing number of QTL (from 10 to 100) for HAP (figure 19 a) and SNP (figure 19b) approach has been detected, even though the increase in accuracy seem to be to some extent greater with the SNP method in comparison to the HAP methods. In any case the differences in accuracies are negligible.

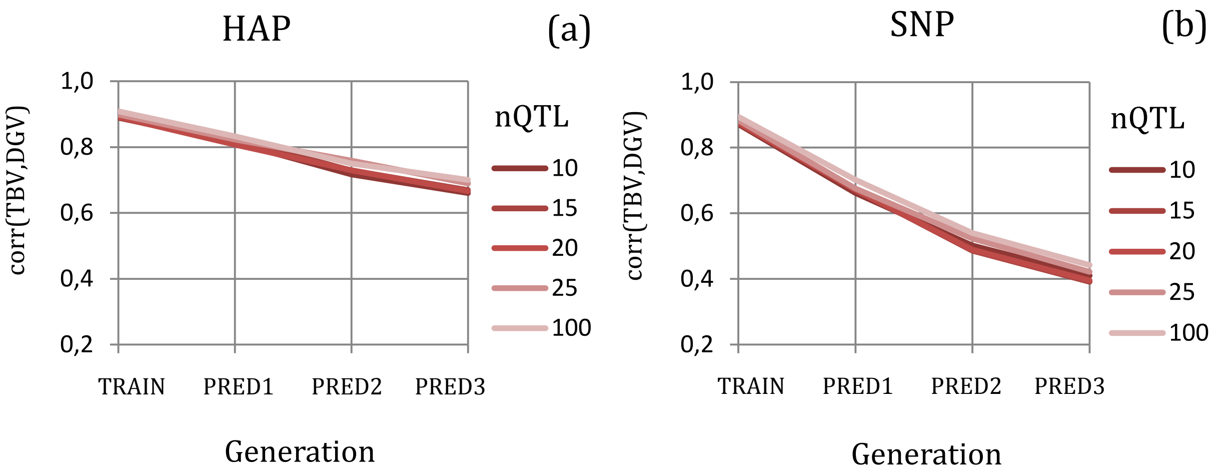


Figure 19 DGV accuracy through generation as function of number QTL using HAP (a) and SNP (b) approach. Each line represent a different number of QTL

The effect of number of QTL on the accuracy seem to be very small. Such results are in agreement with SCHAEFFER (2007) who tested a simulation similar to the one performed in the present study to estimate the effect of main factors of variability of DGV accuracy. Schaeffer found that the effect of the number of QTL on DGV accuracy were only marginal (table 5). These results are quite in accordance with figures obtained by GASPA *et al* (2009) on simulated data. They found an average accuracy of prediction in the training generations of 0.90 when 10 QTLs were simulated. and 0.94 with 20 QTLs. The accuracy in the prediction generations were 0.66 and 0.72 for 10 and 20 QTLs respectively. This results seem to indicate a low influence of number of QTLs in the genome and DGV accuracy. However, the results are difficult to compare because of different number of replicates (30 in the present simulation *vs* 10 replicates) and different methods to generate the base population (use of mutation drift model) In any case the effect of number of QTL seems to be positive associated to higher accuracy.

Table 5. Effect of number of QTL and genome length on the accuracy in 3 prediction generation after 100 generation of random mating. modified from SCHAEFFER (2007).

Genome length	100 cM		200 cM		300 cM	
	nQTL					
generation	40	80	80	160	160	240
101	0.86	0.87	0.86	0.86	0.85	0.85
102	0.81	0.80	0.78	0.77	0.74	0.74
103	0.79	0.78	0.76	0.75	0.72	0.72

3.3.5 DGV accuracy and number of generation of Random mating

The figure 20 shows the trend of accuracy through generation as function of number of previous generation of random mating for HAP (figure 20a) and SNP approach (figure 20b). This step of the simulation is important for the creation of linkage disequilibrium between markers and QTLs. It seems that there is no clear effect of the number of random mating generations. However in the SNP approach there is more variability of response. Being the number of generations of random mating a factor that affect the amount of linkage disequilibrium.

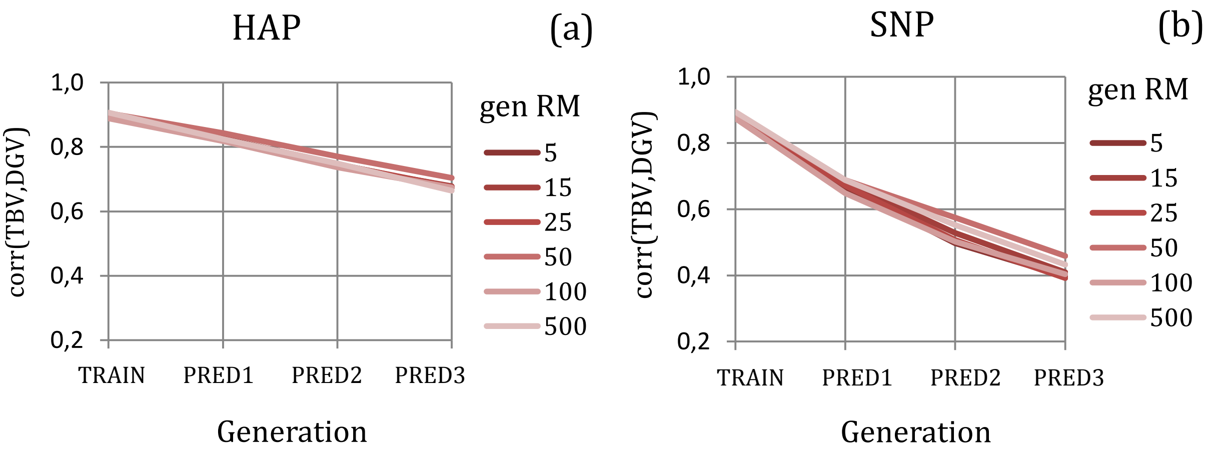


Figure 20 DGV accuracy through generation as function of number of previous generation of random mating (gen RM) using HAP (a) or SNP (b) and approach.

3.4 CONCLUSIONS

The results of this work have highlighted the importance of some factors of variability in the efficiency of genomic selection. The approach used, despite all the limitations that characterize the simulations – such as the simplified assumption of the genetic structure (a pair of chromosomes) – was useful to mimic some features that characterize the genome of livestock species. The simulated population was used to evaluate the influence of various factors on the genetic determinism of the quantitative trait and on the accuracy of the genetic merit of selection candidates measured with markers. Overall, accuracies in the estimation of DGV of this study confirmed the recent literature when BLUP is used in the process of estimating the marker effects and on the features studied in populations of training. In particular, the analysis of variation in the accuracy of the DGV indicates that the best results on traits of high-moderate heritability, whereas poor accuracies were obtained for low heritability traits. Then, it seems important to investigate further the relationships between the heritability and other factors that influence the estimates as, for example, the number of animals used, the prediction model and the pre-selection of more informative markers . The results obtained with different number of markers showed higher accuracy for high density map a and more persistent over the generations. The behavior of the DGV accuracies during the generations is surely of great interest for practical applications of genomic selection. In fact, one of the biggest problems in GS is the identification of the number of generations after which it is necessary re-estimate values of individual. The influence of the density of markers on this parameter, highlighted in this work, deserves further study. However, the analysis of the effect of different number of daughters per bull on the accuracy of DGVs showed that the number of daughters per bull selection schemes currently used in the current progeny test (50 to 120 daughters per bull tested) is sufficient to obtain good accuracy of genomic prediction. Furthermore, the number of QTL (kept constant the number of markers) seems to have a slight positive effect on accuracy of genomic prediction, whereas the number of generation of random mating does not show any effect on the DGV accuracy in the present simulation.

A possible improvement of the simulation could be achieved by introducing a number of generations of selection – introducing a factor that affect the amount of LD – in order to obtain a population with characteristics more similar to livestock populations

ACKNOWLEDGMENTS

Authors would like to acknowledge Prof. Aldo Cappio-Borlino dr. Ezequiel Luis Nicolazzi the for the active contribution and suggestion on writing the simulation.

References

CALUS, M. P. L., T. H. E. MEUWISSEN, A. P. W. DE ROOS and R. F. VEERKAMP, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics **178:** 553-561.

CALUS, M. P. L., and R. F. VEERKAMP, 2007 Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics **124:** 362-368.

DAETWYLER, H. D., B. VILLANUEVA and J. A. WOOLLIAMS, 2008 Accuracy of predicting the genetic risk of disease using a genome-wide approach. Plos One **3:** e3395.

DEKKERS, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim Sci. **82:** E313-328.

GASPA, G.,E. L. NICOLAZZI, R. STERI, C. DIMAURO, AND N. P. P. MACCIOTTA. 2009 Effect of estimation approach and number of QTLs in accuracies of genomic breeding values for simulated data. Proceeding ADSA-ASAS annual joint meeting Montreal 12-15 july 2009. J. Anim. Sci. Vol. 87, E-Suppl. 2/J. Dairy Sci. Vol. 92, E-Suppl. 1 p 315

GRISART, B., W. COPPIETERS, F. FARNIR, L. KARIM, C. FORD *et al.*, 2002 Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res **12:** 222-231.

GUO, G., M. LUND, Y. ZHANG and G. SU, 2010 Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. Journal of Animal Breeding and Genetics**:** no-no.

HABIER, D., R. L. FERNANDO and J. C. M. DEKKERS, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics **177:** 2389-2397.

HABIER, D., J. TETENS, F. R. SEEFRIED, P. LICHTNER and G. THALLER, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution **42**.

HAYES, B. J., P. M. VISSCHER and M. E. GODDARD, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). Genetics Research **91:** 143-143.

KOLBEHDARI, D., L. R. SCHAEFFER and J. A. B. ROBINSON, 2007 Estimation of genome-wide haplotype effects in half-sib designs. Journal of Animal Breeding and Genetics **124:** 356-361.

KONIG, S., H. SIMIANER and A. WILLAM, 2009 Economic evaluation of genomic breeding programs. Journal of Dairy Science **92:** 382-391.

MEUWISSEN, T. H. E., B. J. HAYES and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157:** 1819-1829.

MUIR, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics **124:** 342-355.

SCHAEFFER, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics **123:** 218-223.

SCHAEFFER, L. R. 2007. Factor affecting Accuracy of genomic estimated breeding values. *Internal report*: Department of animal and poultry sciencce. University of Guelph, Canada.

SOLBERG, T. R., A. K. SONESSON, J. A. WOOLLIAMS and T. H. E. MEUWISSEN, 2008 Genomic selection using different marker types and densities. Journal of Animal Science **86:** 2447-2454.

SVED, J. A., 1971 Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology **2:** 125-141.

TENESA, A., P. NAVARRO, B. J. HAYES, D. L. DUFFY, G. M. CLARKE *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome research **17:** 520-526.

VANRADEN, P. M., C. P. VAN TASSELL, G. R. WIGGANS, T. S. SONSTEGARD, R. D. SCHNABEL *et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science **92:** 16-24.

ZENGER, K. R., M. S. KHATKAR, J. A. L. CAVANAGH, R. J. HAWKEN and H. W. RAADSMA, 2007 Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. Animal Genetics **38:** 7-14.

CHAPTER 4

USE OF EIGENVALUES AS VARIANCE PRIORS IN THE PREDICTION OF GENOMIC BREEDING VALUES BY PRINCIPAL COMPONENT ANALYSIS

ABSTRACT

Genome wide selection aims at predicting genetic merit of individuals by estimating the effect of chromosome segments on phenotypes using dense SNP marker maps. In the present paper, principal component analysis was used to reduce the number of predictors in the estimation of genomic breeding values for a simulated population. Principal component extraction was carried out either using all markers available or separately for each chromosome. Priors of predictor variance were based on their contribution to the total SNP correlation structure. The principal component approach yielded the same accuracy of predicted genomic breeding values obtained with the regression using SNP genotypes directly, with a reduction in the number of predictors of about 96% and computation time by 99%. Although these accuracies are lower than those currently achieved with Bayesian methods, at least for simulated data, the improved calculation speed together with the possibility of extracting principal components directly on individual chromosomes may represent an interesting option for predicting genomic breeding values in real data with a large number of SNPs. The use of phenotypes as dependent variable instead of conventional breeding values resulted in more reliable estimates, thus supporting the current strategies adopted in research programmes of genomic selection in livestock.

Key words: SNPs, genomic selection, principal component analysis, eigenvalues.

N. P. P. Macciotta,[*1] G. Gaspa,[*] R. Steri,[*] E. L. Nicolazzi,[§] C. Dimauro,[*] C. Pieramati[†] and A. Cappio-Borlino[*]

Published on Journal of Dairy Science, volume 93, Issue 6, Pages 2765-2774

[*]Dipartimento di Scienze Zootecniche, Università di Sassari, Sassari, Italy 07100
[§]Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza Italy 20100
[†]Centro di Studio del Cavallo Sportivo, Università di Perugia, Perugia, Italy 06100

4.1. Introduction

4.1.1 Use of genome-wide molecular information into breeding program: issues and application

Marker Assisted Selection (MAS) programs have had limited commercial applications till early 2000's due to the fact that most of reported marker-QTL associations had been found within families but were in linkage equilibrium across the population (HAYES and GODDARD 2001; DEKKERS 2004; KHATKAR *et al.* 2004). The availability of genome-wide dense marker maps for several animal species has recently allowed the prediction of genomic breeding values (GEBV) by estimating marker haplotype effects on phenotypes (MEUWISSEN *et al.* 2001; GODDARD and HAYES 2007). Genome wide selection relies on highly dense markers whose effects on phenotypes are estimated on a training population and then used to calculate GEBV both for training individuals and animals with only marker genotypes available (for example, young animals without phenotypes or estimated breeding values). A reduction in generation interval, an increase of accuracy in the cow side of the pedigree and a decrease of selection costs are the expected advantages of an efficient genome wide selection over traditional selection (KONIG *et al.* 2009; SCHAEFFER 2006).

High density SNP maps fulfill the basic requirement of genome wide selection, i.e. the analysis of genome bits having large and persisting population-wide linkage disequilibrium (MUIR 2007). However, the use of dense marker platforms results in a large number of effects to be estimated (many thousands) in comparison with the relatively small amount of phenotypes available (often just a few thousands). Such a data asymmetry raises several statistical issues, such as collinearity among predictors and multiple testing (GIANOLA and VAN KAAM 2008). To cope with such a problem, several methods of reduction of the number of predictors without a large decrease in accuracy have been proposed.

Selection of relevant SNP by single marker regression on phenotypes may improve results in genome-wide association studies (AULCHENKO *et al.* 2007a; LONG *et al.* 2007), but it leads to a decrease of GEBV accuracy (MEUWISSEN *et al.* 2001). Bayesian methods that select SNP by evaluating their individual contribution to the variance of the trait, such Bayes B method (MEUWISSEN *et al.* 2001; FERNANDO *et al.* 2007a; VANRADEN 2008), usually give best GEBV accuracies when simulated data with few QTLs are modeled. However, results on actual data indicate that BLUP estimation, which assumes an equal contribution of all marker intervals to the genetic variance, performs only slightly worse than Bayesian methods in GEBV prediction (HAYES, *et al* 2009; VANRADEN *et al.* 2009).

Moreover in all the above mentioned techniques, markers are selected according to their relevance on the variability of the phenotype analyzed. Consequently, specific sets of markers may be required for different traits (HABIER *et al.* 2009).

4.1.2. Multivariate approach to reduce dimension of SNP data

Multivariate dimension-reduction techniques may offer an alternative approach based on the evaluation of the contribution of each marker locus to the total SNP (co)variance structure. Principal component analysis (PCA) has been used for analyzing complex genetic patterns in human genetics (CAVALLI-SFORZA and FELDMAN, 2003; PASCHOU *et al.* 2007) and for selecting markers in genome-wide association studies. SOLBERG *et al.* (2009) used principal component analysis and partial least squares regression (PLSR) to reduce the dimensionality of predictors in genomic selection. Both principal component (PC) and PLSR showed comparable accuracies with Bayes B when lower marker densities were fitted, whereas the gap between methods increased with the number of markers used. SOLBERG *et al.* (2009) concluded that reduction in computational complexity provided by multivariate methods did not counterbalance their lower accuracy compared to Bayes B. Such considerations are justified by the low cost of calculation time and by the computational speed that can be provided by optimized techniques such as parallel computing. On the other hand, it is reasonable to expect that denser SNP platforms will be very soon available for livestock species and dimensionality will again represent a relevant problem.

In their proposal, SOLBERG *et al.* (2009) regressed phenotypes on principal component scores extracted from the SNP matrix using the single value decomposition approach with an assumption of equal variance of each PC score. The choice of priors of marker effects represents a crucial point for genomic models (DE LOS CAMPOS *et al.* 2009). On the other hand, the ordinary method for calculating PC relies on the eigenvalues of the correlation matrix of starting variables that measure the contribution of each PC to the original variance of predictors. Thus eigenvalues can be used as priors of predictor effect for the calculation of GEBV. It is worth remembering that eigenvalues have been already incorporated in mixed model algorithms to optimize calculations for variance component estimation (DEMPSTER *et al*., 1984; TAYLOR *et al.* 1985).

In the present paper, principal component analysis is used to perform a BLUP prediction of GEBV in a simulated data set to test the ability of this technique to reduce the number of predictors without decreasing GEBV accuracy. Moreover, the feasibility of extracting PC from dense commercially available SNP platforms is tested.

4.2. MATERIAL AND METHODS

4.2.1 Data

The data set was generated for the XII QTLs – MAS workshop (http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html). The base population consisted of 100 individuals (50 males and 50 females). The genome had six chromosomes (total length 6 M), with 6,000 biallelic SNP, equally spaced at a distance of 0.1 cM. A total of 48 biallelic QTL were generated, with positions sampled from the genetic map of the mouse genome. QTL effects were sampled from a gamma distribution with parameters estimated by HAYES and GODDARD (2001). Initial allelic frequencies of both SNP and QTL were set to 0.5. Then 50 generations of random mating followed. Generations 51 to 57 were used to create the experimental population of 5,865 individuals. Generations 51 to 54 (4,665 individuals, TRAIN data set) had pedigree, phenotype, and marker information available. For the last three generations (1,200 individuals, PRED data set) only pedigree and marker information were available. True breeding values (TBV) were considered as the sum of all QTL effects across the entire genome. Phenotypes were generated by adding environmental noise to the TBV. Further details on the simulation can be found in LUND *et al.* (2009).

Polygenic breeding values (EBV), being among the most frequently used dependent variable in GEBV prediction with real data, were also predicted. EBV, additive genetic (σ^2_a) and residual (σ^2_e) variance components were estimated with a single trait animal model that included the fixed effects of sex and generation, and the random additive genetic effect of the animal. The pedigree relationship matrix included 5,939 animals.

4.2.2. PCA analysis.

Principal component analysis aims at synthesizing information contained in a set of *n* observed variables (M_1, ..., M_n) by seeking a new set of *k* ($k<n$) orthogonal variables (PC_1,..., PC_k) named principal components. PC are calculated from the eigen decomposition of the covariance (or correlation) matrix of M. The *j*th PC is a linear combination of the observed variables:

$$PCj = \alpha_{1j}M_1 + \ldots + \alpha_{nj}M_n$$

where coefficients α_{ij} are the elements of the eigenvector corresponding to *j*-th eigenvalue. PC are usually extracted in a descending order of the corresponding eigenvalue that measures the quota of variance of original variables explained by each PC (MORRISON, 1976; KRZANOWSKY, 2003).

A SNP data matrix **M** with m rows (m=5,865, the number of individuals in the entire data set) and *n* columns (*n*=5,925, the number of SNP markers that were found to be polymorphic) was created. Each element (*i*, *j*) corresponded to the genotype at the the *j*-th marker for the *i*-th individual. Genotypes were coded as -1, 0 or 1, according to the notation used by SOLBERG *et al.* (2009).

Data editing is usually recommended when handling dense marker maps (Wiggans *et al.* 2009), either to correct for data quality (i.e. genotyping not successfully performed) or to avoid possible estimation biases due to a severe unbalancement of genotypes. However, considering that in the present simulated data only 288 markers had minor allele frequency (MAF) <0.05, while 47 deviated significantly (P<0.01) from the Hardy-Weinberg equilibrium and this deviation may be attributable to drift, only the 75 monomorphic SNP were discarded from the analysis. Such a choice is, at least partially, supported by results of CHAN *et al.* (2009) that pointed out that SNP attributes commonly considered in SNP data editing, such as MAF or deviation from Hardy-Weinberg equilibrium, have actually a very small effect on overall false positive rate in genome-wide association studies.

PCA was carried out on **M** and the number of PC (*k*) retained for further analysis was based on both the sum of their eigenvalues and the obtained GEBV accuracy. PC extraction was performed either on all SNP simultaneously (PC_SNP_ALL) or separately for each chromosome (PC_SNP_CHROM). Scores of the *k* selected PC were calculated for all individuals. Marker haplotypes may be more efficient than genotypes in capturing marker-QTL association, especially in outbred populations where it may differ between families (CALUS *et al.* 2008). Thus, PCA was performed also on haplotypes constructed from pairs of adjacent marker loci, either using all loci together (PC_HAP_ALL) or separately per chromosome (PC_HAP_CHROM).

4.2.3. Predictor effect estimation and GEBV calculations.

Dependent variables used in the analysis were either phenotypes or polygenic EBV. For the estimation of the effects of predictors, records of the 4,665 individuals of the TRAIN data set were analysed with the following mixed linear model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$$

where **y** is the vector of either phenotypes or EBV, **X** is the design matrix of fixed effects (mean, sex=1,2; generation=1,2,3,4 for phenotypes; only mean for EBV), **b** is the vector of solutions for fixed effects, **Z** is the (*m* x *k*) design matrix of random effects, where each element corresponds to the score of the *k*-th component for the *m*-th animal of the training generations**,** **g** is the vector of solution for random regression coefficients of PC scores, **e** is the random residual. Covariance matrices of random PC effects (**G**) and residuals (**R**) were modeled as diagonal **I**(σ^2_{ai}) and **I**(σ^2_e), respectively.

BLUP methods used for estimating SNP effects usually assume an equal contribution of each SNP locus to the variance of the trait, sampled from the same normal distribution, i.e. $\sigma^2_{aj}=\sigma^2_a/n$ (MEUWISSEN *et al.* 2001; VANRADEN *et al.* 2009). In the present work, two different options were compared. The first is the above mentioned equality of variances. The second starts from the consideration that PC scores were used as predictor variables and their contribution to the original SNP covariance structure is quantified by the corresponding eigenvalue (λ). Thus, variances of PC effects were calculated as $\sigma^2_{aj}=(\sigma^2_a/k)\cdot\lambda_j$.

G matrix diagonality, commonly implemented in BLUP methodologies for estimating SNP marker effects (MEUWISSEN *et al.* 2001; VANRADEN 2008), relies on the assumption that marker effects in a large population are uncorrelated (VANRADEN *et al.* 2009). With the use of PC scores, such an assumption is consistent with the orthogonality between PC (MORRISON, 1976). BLUP solutions were estimated using Henderson's normal equations (HENDERSON 1985).

In order to have a comparison with the most straightforward estimation method, SNP effects were estimated directly by using the same mixed linear model but with **Z** indicating the design matrix of the 5,925 polymorphic SNP genotypes (coded as 0, 1 and 2, i.e. on the basis of the number of alleles). Covariance matrix **G** was assumed to be diagonal as **I**(σ^2_a/n). A Cholesky decomposition was used to solve mixed model equations (HARVILLE, 1997).

Overall mean and effects of PC scores or SNP genotypes ($\hat{\mathbf{g}}$) estimated on the TRAIN data set were then used to predict GEBV both in TRAIN and PRED individuals. as

$$\mathbf{GEBV} = \mu + \mathbf{Z}\widehat{\mathbf{g}}$$

where **GEBV** is the vector of predicted genomic breeding values and **Z** is the matrix of the PC scores or SNP genotypes of all individuals.

Accuracies of prediction where evaluated by calculating Pearson correlations between GEBV and TBV for the PRED generations. Bias of prediction was assessed by examining the regression coefficient of TBV on GEBV (MEUWISSEN *et al.* 2001). Goodness of prediction was evaluated also by the mean squared error of prediction (MSEP) calculated as

$$MSEP = \sum_{i=1}^{n} \frac{\left[TBV_i - GEBV_i\right]^2}{n}$$

where *n* is the number of individuals in the PRED generations, and by its partition in different sources of variation related to systematic and random errors of prediction (TEDESCHI 2006).

4.3 Results

4.3.1 PC analysis

The pattern of eigenvalues of the correlation matrix of SNP genotypes obtained with PCA of all markers simultaneously is reported in Figure 1 (only the first 1,000 eigenvalues are plotted for brevity). A smooth decrease in the amount of variance explained by each successive PC can be observed, with a plateau between 250 and 300 PCs (about 84% of variance explained). A number of principal components between 200 and 300 could therefore be considered adequate for describing the original variance of the system.

GEBV accuracies for different numbers of retained PC (from 50 to 600) using all SNP simultaneously and eigenvalues as variance priors are reported in Figure 2. Accuracy for both training and prediction generations increases till a plateau, reached at about 250-300 PC. Increasing further the number of retained PC does not result in an increase of accuracy, probably due to the small amount of variance explained by each additional variable. Similar results were obtained by Solberg *et al.* (2009) that report best accuracies when 350 PC were extracted from 8,080 bi-allelic markers distributed on 10 chromosomes. However, Solberg *et al.* (2009) found a rather decreasing trend of the correlation between GEBV and TBV for larger numbers of PC. Based on the accuracy of GEBV prediction, 279 PCs (83% of the original variance) were retained in the present work for PC_SNP_ALL and PC_HAP_ALL approaches. In the analysis carried out on individual chromosomes, to keep the same number of predictors of the previous approach, 46 and 47 PC for chromosomes 1-3 and 4-6 were retained, respectively.

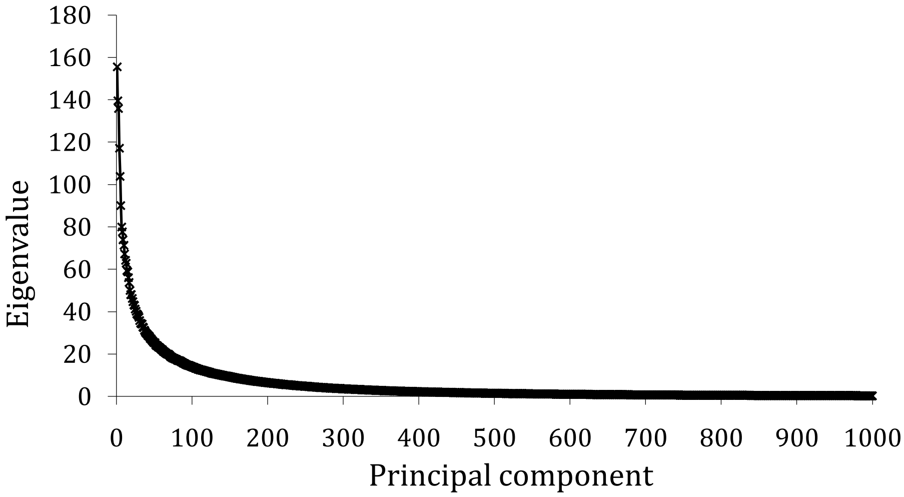


Figure 1. Pattern of the eigenvalues of the correlation matrix of SNP markers.

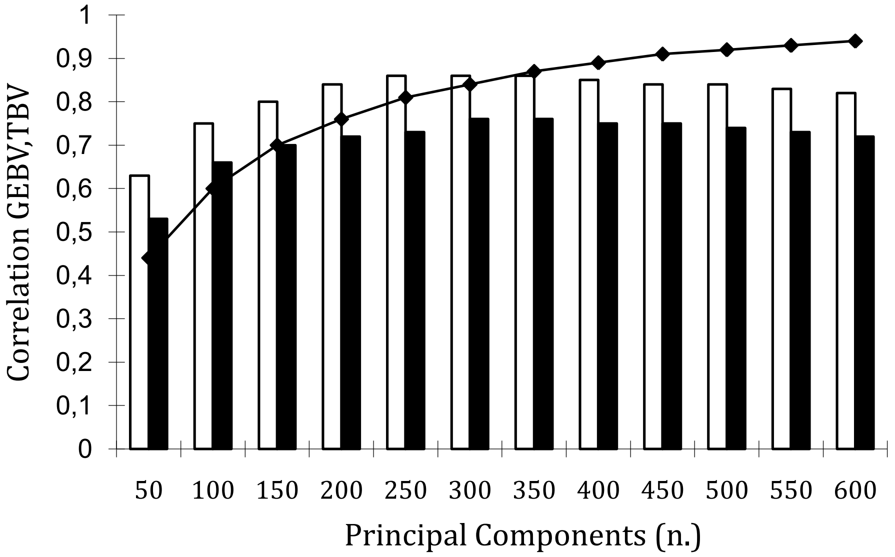


Figure 2. Pattern of correlations between genomic breeding values (GEBV) and true breeding values (TBV) when principal components are extracted from all SNP genotypes simultaneously and eigenvalues are used as priors, for different number of retained PC (white bars=training individuals, black bars=prediction individuals). The continuous line represents the amount of variance explained by the corresponding number of PC.

4.3.2 Accuracy and bias of GEBV

Average GEBV accuracies obtained using phenotypes are, for the three prediction generations, around 0.70 (Table 1) when an equal contribution of PC score on the variance of the trait is assumed, similar to those reported by SOLBERG *et al.* (2009). Accuracies increase by about 10% (to an average of 0.75) when eigenvalues are used in the diagonal of the $\mathbf{G}^{-1}$ matrix of mixed model equations. In general, results are of the same order as in previous literature reports for BLUP estimation on simulated (MEUWISSEN *et al.* 2001; FERNANDO *et al.* 2007b; MEUWISSEN 2009) and real data (VANRADEN *et al.* 2009; HAYES, *et al* 2009). Correlations obtained when all SNP were used as predictors are equal to those obtained with PC with eigenvalues as priors. On the other hand, a remarkable difference in calculation speed between the two methods has been observed: about six hours for the SNP_ALL approach and 3 minutes for the principal components, using a computer with a dual core processor 2.33 GHz and 3.26 MB RAM. Slight differences can be observed between estimates of PC carried on all chromosomes or separately for each of them. Moreover, same results have been basically obtained when genotypes at single markers or haplotypes were used, in agreement with previous reports for high density markers (HAYES *et al.* 2007; CALUS *et al.* 2008;).

GEBV accuracies are larger when phenotypes instead of EBV are used as dependent variables (Table 1). This is particularly evident when all SNP are used as predictors (on average 0.75 vs 0.39). Also the drop of accuracy between TRAINING and PRED generations is more evident for EBV-based predictions (Figures 3 and 4).

Table 1. Pearson correlations between predicted genomic breeding values and true breeding values, for different estimation methods, using either phenotypes or polygenic breeding values (EBV) for the PREDICTION generations and assuming either equal variance contribution for each PC or eigenvalues as variance priors.

Method	Phenotypes	EBV
SNP_ALL	0.76	0.41
	Equal variance	
PC_SNP_ALL	0.69	0.53
PC_SNP_CHROM	0.70	0.55
PC_HAP_ALL	0.68	0.54
PC_HAP_CHROM	0.71	0.56
	Eigenvalues	
PC_SNP_ALL	0.76	0.57
PC_SNP_CHROM	0.73	0.56
PC_HAP_ALL	0.75	0.56
PC_HAP_CHROM	0.73	0.55

(SNP_ALL = all 5,925 SNPs; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PC_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PC_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PC_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome)

These findings are confirmed by values of regression coefficients of TBV on GEBV (Table 2). Moreover, *b* values for methods based on PC are similar to those reported by SOLBERG *et al.* (2009) when equal variances were assumed whereas they are closer to one (about 0.85) when eigenvalues are used as variance priors.

The decomposition of the mean squared error of prediction for some of the considered scenarios is reported in Table 3. MSEP is always smaller (about a half) when GEBV are calculated using phenotypes. Its partition highlights a great relevance of components related to the bias of prediction (i.e. mean bias, inequality of variances) in the approach that fits directly SNP genotypes (about 79%).

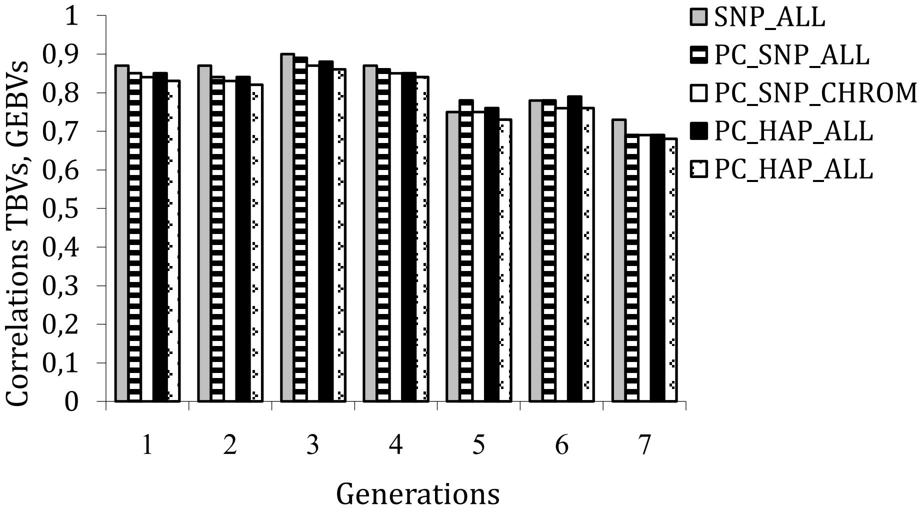


Figure 3. Correlations between genomic breeding values (GEBV) and true breeding values (TBV) in the different approaches when phenotypes were used as dependent variables (SNP_ALL = all 5,925 SNP; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PCA_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PCA_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PCA_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome).

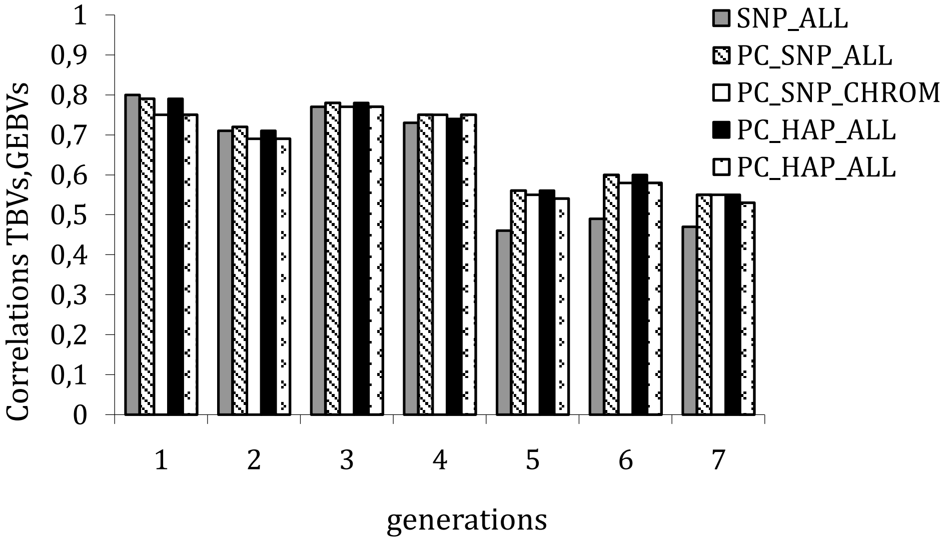


Figure 4. Correlations between genomic breeding values (GEBV) and true breeding values (TBV) in the different approaches when EBV were used as dependent variables (SNP_ALL = all 5,925 SNP; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PCA_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PCA_HAP_ALL = principal components extracted from all SNPS haplotypes simultaneously; PCA_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome).

Table 2. Regression coefficients ($b_{TBV,GEBV}$) of True breeding Value on Predicted Genomic Breeding Value (GEBV) for the different estimation methods using either phenotypes or polygenic breeding values (EBV) for the PREDICTION generations and assuming either equal variance contribution for each PC or eigenvalues as variance priors.

	Trait			
Method	Phenotypes		EBV	
	$b_{TBV,GEBV}$	s.e.	$b_{TBV,GEBV}$	s.e.
SNP_ALL	1.08	0.027	1.15	0.073
	Equal variance			
PC_SNP_ALL	0.63	0.019	1.08	0.049
PC_SNP_CHROM	0.67	0.019	1.13	0.048
PC_HAP_ALL	0.61	0.019	1.08	0.049
PC_HAP_CHROM	0.65	0.018	1.11	0.047
	Eigenvalues			
PC_SNP_ALL	0.88	0.021	1.33	0.055
PC_SNP_CHROM	0.84	0.022	1.28	0.055
PC_HAP_ALL	0.88	0.022	1.32	0.056
PC_HAP_CHROM	0.83	0.023	1.26	0.056

(SNP_ALL = all 5,925 SNPs; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PC_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PC_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PC_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome)

Methods based on PC extraction are characterized by a prevalence (about 80%) of random terms, measured by the random error and by the incomplete covariation. The use of eigenvalues as variance priors results in the lowest MSEP and, compared to the other PC-based method, in a reduction of the slope bias and the highest relevance of random variation. These differences can be clearly seen from the plots of TBV versus GEBV for the PC_SNP_ALL approach using equal (Figure 5a) or eigenvalue-based (figure 5b) variance. The latter shows a regression slope closer to the equivalence line (*y*=*x*) and a smaller value for the intercept, that indicates a smaller systematic underestimation of TBV. The composition of MSEP becomes very similar across the different methods when EBV are used as dependent variables, with a reduced incidence of random components and a larger relevance of unequal variances compared to the phenotype-based estimates (Table 3).

Table 3. Mean squared error of prediction (MSEP) decomposition (%) and coefficient of determination (r^2) for the PREDICTION generations in some scenarios using either phenotypes or polygenic breeding values (EBV) .

		Phenotype	
	SNP_ALL	PC_SNP_ALL1	PC_SNP_ALL 2
MSEP	1.55	1.48	1.02
Mean Bias (U_M)	72.2	53.5	56.9
Unequal variances (U_S)	6.9	0.6	1.9
Incomplete covariation (U_C)	21.9	45.9	41.2
Slope bias (U_R)	0.22	11.1	1.1
Random errors (U_D)	27.6	35.4	42.0
r^2	0.57	0.48	0.57
		EBV	
MSEP	2.96	2.88	2.72
Mean Bias (U_M)	72.0	75.1	74.6
Unequal variances (U_S)	13.9	8.9	11.9
Incomplete covariation (U_C)	14.1	16.0	13.5
Slope bias (U_R)	0.01	0.00	0.7
Random errors (U_D)	27.9	24.9	24.7
r^2	0.17	0.28	0.33

(SNP_ALL= all 5,925 SNPs; PC_SNP_ALL 1= principal components extracted from all SNP genotypes simultaneously and equal contribution of each SNP to the variance of the trait; PC_SNP_ALL 2 principal components extracted from all SNP genotypes simultaneously and contribution of each SNP to the variance of the trait proportional to the eigenvalue

Note that U_M+ U_S+ U_C= U_M+ U_R+ U_D=100%

Actually, the comparison of plots of TBV versus GEBV estimated with the PC_SNP_ALL approach using phenotypes (Figure 5a) or EBV (Figure 5c), clearly shows a reduced range of variability and a higher underestimation (as evidenced by the larger value of the regression intercept) for EBV-based GEBV.

4.3.3. Interpretation of PC

An interesting feature of principal component analysis is the possible technical interpretation of extracted variables. Figure 6 reports score averages for the first two PC that together explain about

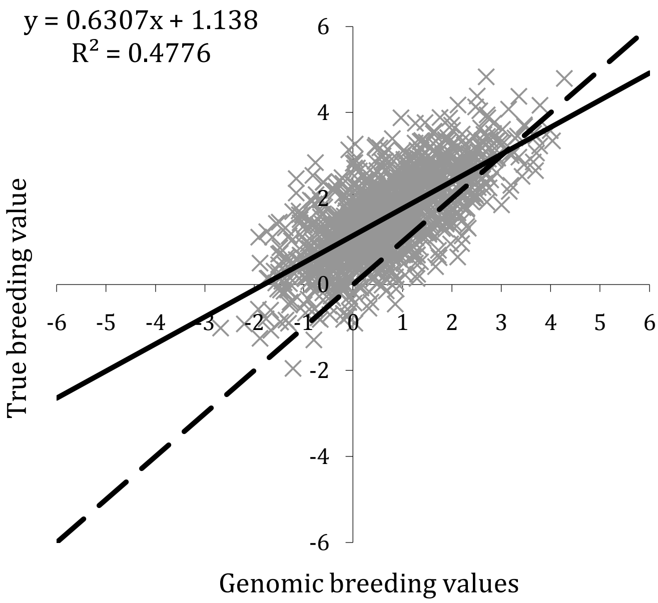


Figure 5a. Plot of true breding values versus genomic breeding values predicted using phenotypes when principal components are extracted from all SNP genotypes simultaneously and variance contribution of the PC scores in the estimation step is assumed equal (continuous line= regression line of TBV on GEBV; dotted line= equivalence line, y=x).

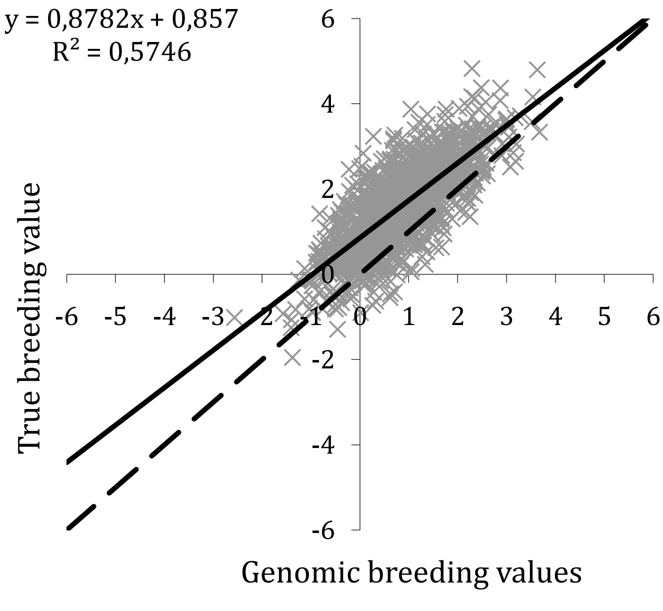


Figure 5b. Plot of true breeding values versus genomic breeding values predicted using phenotypes when principal components are extracted from all SNP genotypes simultaneously and variance contribution of the PC scores in the estimation step is based on their eigenvalues (continuous line= regression line of TBV on GEBV; dotted line= equivalence line, y=x

5% of the original variance of the system, calculated for each generation. Averages of the second PC ranged gradually from negative values for the first three generations to positive for the last three generations. A possible explanation of the ability of the second PC to distinguish individuals of different generations can be found in its negative correlation with the average observed heterozygosity per animal (-0.26) that tends to decrease from older to younger generations (Figure 7).

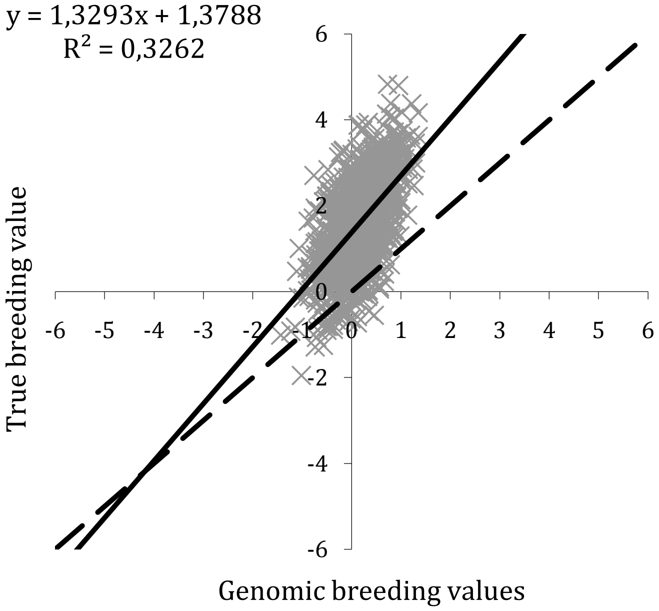


Figure 5c. Plot of true breeding values versus genomic breeding values predicted using phenotypes when all SNP genotypes are used as predictors (continuous line= regression line of TBVs on GEBVs; dotted line= equivalence line, *y* = *x*).

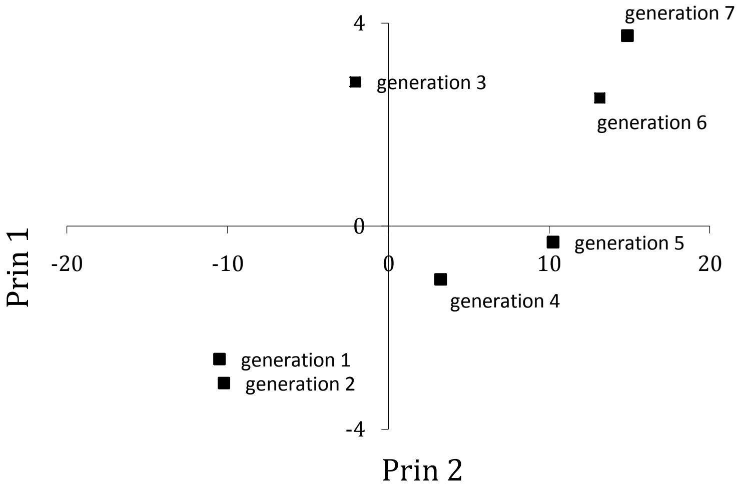


Figure 6. Plot of the average scores of the first two principal components for seven generations.

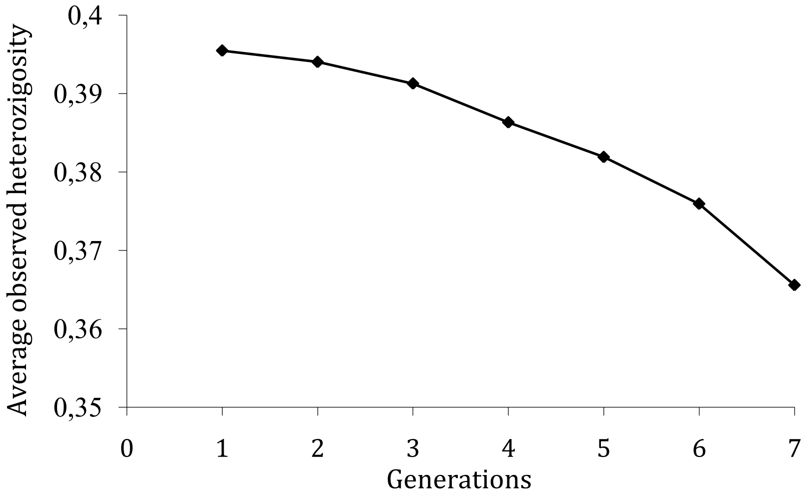


Figure 7. Pattern of the average observed heterozygosity in different generations.

4.4. DISCUSSION

4.4.1 Advantage and issue of using PC to estimate GEBV.

Main objectives of the work are to assess the effect of reducing predictor dimensionality in genomic breeding value estimation using PCA and to test the effect of structuring the variance contribution of PC with their eigenvalues

PCA allows an efficient description of the correlation matrix of biallelic SNP with a markedly smaller number of new variables (4.7%) compared to the original dimension of the system. Such a huge decrease has a straightforward impact on the calculation speed of GEBV, with a reduction of more than 99% of computing time achieving the same accuracy of predicted GEBV using all SNP. Compared to other methods of reduction of predictors where SNP are selected based on their position along the chromosome (VANRADEN *et al.* 2009) or their relevance with the trait considered (HAYES *et al*., 2009), the multivariate reduction approach limits the loss of information because each SNP is involved in the composition of each PC.

GEBV accuracies obtained in the present work agree with a previous report on the use of PCA to estimate genomic breeding values (SOLBERG *et al.* 2009) when an equal contribution of each principal component to the variance of phenotypes is assumed. This approach follows the common BLUP assumption of equality of variance of predictors, usually criticized for its inadequacy to fit the widely assessed distribution of QTL i.e,. many loci with a small effect and very few with large effect (HAYES and GODDARD 2001). However, when eigenvalues are used as prior of PC variance, accuracies increase by about 10%. These figures highlight the importance of an accurate modeling of the variance structure of random effects in GEBV estimation.

Bayesian methods estimate variances of different chromosome segments combining information from prior distribution and data (MEUWISSEN *et al.* 2001). These methods usually give the best performance (accuracies >80%) when simulated data are fitted, whereas results obtained on real data seem to indicate a substantial equivalence with the BLUP approach (HAYES *et al*., 2009; VANRADEN *et al.* 2009). A common explanation is that, in Bayes method, assumptions on prior distributions of parameters are more difficult to infer when real data are handled. The use of eigenvalues as variance priors rely only on data, i.e. the SNPs correlation structure, and does not require assumptions on prior distribution.

A potential drawback in the calculation of GEBV using PCA is represented by PC extraction. In the present work, about 40 minutes were needed to process a SNP data matrix of 5,865 rows and 5,925 columns. The commercially available SNP panel for cattle has 54K marker loci, although about 40K are retained on average after editing (HAYES *et al*., 2009). Such a marked increase of columns, usually not accompanied by a comparable increase of rows (i.e. phenotypic records), may lead to statistical and computational problems if PC are extracted treating all SNP simultaneously. However, results of the present study indicate that PC may be calculated separately for each chromosome, keeping the same GEBV accuracy. It should be remembered that the number of SNP per chromosome is not far from current dairy data (on average 1,200-1,300) (HAYES *et al*., 2009; VANRADEN *et al.* 2009; WIGGANS *et al.* 2009). Thus PCA carried out on individual chromosomes may be of great interest for real data, also considering the substantial biological orthogonality among chromosomes. The availability of denser marker maps (i.e. 500K SNP) will represent a challenge for the method, although the number of PC to be retained does not seem to increase linearly with the number of original variables. Missing genotypes is a potential problem for computation of PCA, which requires data in each cell. Although edits that are normally carried out on SNP data leave only a few missing cells per animal, they are spread across different markers and this may lead to a severe reduction in the number of records. Missing data can be reconstructed using appropriate algorithms as those described by (GENGLER *et al.* 2007) or others implemented in softwares of common use such as PHASE or PLINK.

4.4.2. GEBV accuracy using phenotype or EBV.

Of particular interest is the difference in GEBV accuracy obtained when using phenotypes vs. polygenic EBV as dependent variable. Polygenic EBV are phenotypes corrected for additive relationships among animals based on pedigree information. On the other hand, in GEBV predictions the genetic similarity between animals is accounted for by the specific combination of marker genotypes possessed by each individual. Therefore, the use of EBV as dependent variable in GEBV prediction may be regarded as redundant in terms of exploitation of genetic relationships.

This behavior is particularly evident for the regression using all SNP markers. In this form, the calculation of GEBVs is equivalent to the use of an animal model with the additive genetic effect structured by the genomic relationship matrix (GODDARD 2009). Such a double counting of genetic relationship resulted in a evident reduction of the variability of GEBV compared to true breeding values. From a statistical standpoint, EBV are model predicted values and may not be suitable as dependent variable in further analyses (TEDESCHI 2006). Results of the present study, although obtained on simulated data, may more accurately reflect the reality of genomic selection programmes in cattle. In previous studies, EBV were generally the dependent variable. This is because true breeding values are not available on real data and EBV estimated with a high accuracy (>0.90) may represent a sort of golden standard for cross validations. However, the tendency now seems to move toward the use of partially corrected phenotypes such as de-regressed proofs or Daughter Yield Deviations (HAYES *et al*., 2009; VANRADEN *et al.* 2009).

4.4.3. PC as indicator in population genetics.

Finally, an interesting side product of PCA used to reduce the dimensionality of predictors in genome wide selection is represented by the extraction of synthetic variables that can have a technical meaning. Researches in human and animal genetics have highlighted the role of PC as indicators of population genetic structure: for example, the top eigenvectors of the covariance matrix show often a geographic interpretation (CHESSA *et al.* 2009; PRICE *et al.* 2006). Usually, the meaning of the *i*-th PC in terms of relationship with the original variables is inferred from the structure of its eigenvector. In the present study, such an evaluation was not feasible, probably due to both the relatively small amount of variance explained by each PC and the large number of original variables considered (i.e. the 5,925 SNP). However, one of the top PC was able to reflect the genetic variation among generations, although the discrimination between individuals of different generations was rather fuzzy, as expected, given the small amount of variance explained. However, this last point deserves some additional consideration. An assessed criterion in choosing which PC to retain is to look at their eigenvalues. However, sometimes the PC associated with the largest eigenvalue does not have a defined meaning whereas successive PC characterized by smaller eigenvalues may contain more relevant or biological information (JOMBART *et al.* 2009). In the case of the present work, a meaning of the second PC as indicator of genetic drift, which should be the only reason of variation of genotypic frequencies in the simulated generations (LUND *et al.* 2009) could be hypothesized.

ACKNOWLEDGMENTS

Research funded by the Italian Ministry of Agriculture, grant SELMOL. Authors thank organizers of the XII QTL-MAS workshop for providing simulated data. Discussion with P. Ajmone-Marsan is gratefully acknowledged. Helpful comments and suggestions from the two reviewers are also acknowledged.

REFERENCES

AULCHENKO, Y. S., D.-J. DE KONING and C. HALEY, 2007 Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics **177:** 577-585.

CALUS, M. P. L., T. H. E. MEUWISSEN, A. P. W. DE ROOS and R. F. VEERKAMP, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics **178:** 553-561.

CAVALLI-SFORZA, L. L., and M. W. FELDMAN, 2003. The application of molecular genetic approaches to the study of human evolution. Nature Genetics.

CHAN, E. K. F., R. HAWKEN and A. REVERTER, 2009 The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Animal Genetics **40:** 149-156.

CHESSA, B., F. PEREIRA, F. ARNAUD, A. AMORIM, F. GOYACHE *et al.*, 2009 Revealing the History of Sheep Domestication Using Retrovirus Integrations. Science **324:** 532-536.

DE LOS CAMPOS, G., D. GIANOLA and G. J. M. ROSA, 2009 Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. J. Anim Sci. **87:** 1883-1887.

DEKKERS, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim Sci. **82:** E313-328.

DEMPSTER, A.P., C.M. PATEL, M.R. SELWYN AND A.J. ROTH. 1984. Statistical and computation aspects of mixed model analysis. Appl. Stat. 33:203-214.

FERNANDO, R. L., D. HABIER, C. STRICKER, J. C. M. DEKKERS and L. R. TOTIR, 2007a Genomic selection. Acta Agriculturae Scandinavica, Section A - Animal Science **57:** 192 - 195.

FERNANDO, R. L., D. HABIER, C. STRICKER, J. C. M. DEKKERS and L. R. TOTIR, 2007b Genomic selection. Acta Agriculturae Scandinavica Section a-Animal Science **57:** 192-195.

GENGLER, N., P. MAYERES and M. SZYDLOWSKI, 2007 A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal **1:** 21-28.

GIANOLA, D., and J. B. C. H. M. VAN KAAM, 2008 Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics **178:** 2289-2303.

GODDARD, M., 2009 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136:** 245-257.

GODDARD, M. E., and B. J. HAYES, 2007 Genomic selection. Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie **124:** 323-330.

HABIER, D., R. L. FERNANDO and J. C. M. DEKKERS, 2009 Genomic Selection Using Low-Density Marker Panels. Genetics **182:** 343-353.

HARVILLE, D. A. 1997. Matrix algebra from a statistician's perspective. Springer-Verlag, New York

HAYES, B., and M. E. GODDARD, 2001 The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution **33:** 209-229.

HAYES, B. J., P. J. BOWMAN, A. J. CHAMBERLAIN and M. E. GODDARD, 2009 Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci **92:** 433-443.

HAYES, B. J., A. J. CHAMBERLAIN, H. MCPARTLAN, I. MACLEOD, L. SETHURAMAN *et al.*, 2007 Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genetics Research **89:** 215-220.

HENDERSON, C. R., 1985 Best Linear Unbiased Prediction Using Relationship Matrices Derived from Selected Base Populations. Journal of Dairy Science **68:** 443-448.

JOMBART, T., D. PONTIER and A. B. DUFOUR, 2009 Genetic markers in the playground of multivariate analysis. Heredity **102:** 330-341.

KHATKAR, M. S., P. C. THOMSON, I. TAMMEN and H. W. RAADSMA, 2004 Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genetics Selection Evolution **36:** 163-190.

KONIG, S., H. SIMIANER and A. WILLAM, 2009 Economic evaluation of genomic breeding programs. Journal of Dairy Science **92:** 382-391.

KRZANOWSKY, W. J. 2003. Principles of multivariate analysis. Oxford University Press Inc., New York.

LONG, N., D. GIANOLA, G. J. M. ROSA, K. A. WEIGEL and S. AVENDAÑO, 2007 Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of Animal Breeding and Genetics **124:** 377-389.

LUND, M., G. SAHANA, D.-J. DE KONING, G. SU and O. CARLBORG, 2009 Comparison of analyses of the QTLMAS XII common dataset. I: Genomic selection. BMC Proceedings **3:** S1.

MEUWISSEN, T. H., 2009 Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol **41:** 35.

MEUWISSEN, T. H. E., B. J. HAYES and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157:** 1819-1829.

MORRISON, F. 1976. Multivariate statistical methods. McGraw-Hill, New York.

MUIR, W. M., 2007 Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics **124:** 342-355.

PASCHOU, P., E. ZIV, E. G. BURCHARD, S. CHOUDHRY, W. RODRIGUEZ-CINTRON *et al.*, 2007 PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. Plos Genetics **3:** e160.

PRICE, A. L., N. J. PATTERSON, R. M. PLENGE, M. E. WEINBLATT, N. A. SHADICK *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics **38:** 904-909.

SCHAEFFER, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics **123:** 218-223.

SOLBERG, T. R., A. K. SONESSON, J. A. WOOLLIAMS and T. H. E. MEUWISSEN, 2009 Reducing dimensionality for prediction of genome-wide breeding values. Genetics Selection Evolution **41:** -.

TAYLOR, J. F., B. BEAN, C. E. MARSHALL and J. J. SULLIVAN, 1985 Genetic and Environmental Components of Semen Production Traits of Artificial Insemination Holstein Bulls. Journal of Dairy Science **68:** 2703-2722.

TEDESCHI, L. O., 2006 Assessment of the adequacy of mathematical models. Agricultural Systems **89:** 225-247.

VANRADEN, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science **91:** 4414-4423.

VANRADEN, P. M., C. P. VAN TASSELL, G. R. WIGGANS, T. S. SONSTEGARD, R. D. SCHNABEL *et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science **92:** 16-24.

WIGGANS, G. R., T. S. SONSTEGARD, P. M. VANRADEN, L. K. MATUKUMALLI, R. D. SCHNABEL *et al.*, 2009 Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J Dairy Sci **92:** 3431-3436.

CHAPTER 5

GENERAL DISCUSSION

The discovery of chromosomal regions affecting quantitative traits date 1923 with the experiment of (SAX 1923) who demonstrated the association between a gene associated to the locus that regulates the color of the seeds of *phaseoulus vulgaris* and the weight of the seeds. (THODAY 1961), with a publication entitled *The location of polygenes* gave a contribution to the formulation of the concept of Quantitative Trait Locus (QTL). The basic idea of improving the accuracy of the estimation of the genetic merit of an individual, i.e. his estimated breeding values (EBV), by using genetic markers (blood group in cattle) known to be associated to an effect on phenotype was already developed in the sixties (NEIMANN-SORENSEN and ROBERTSON 1961). However, the systematic use of DNA variation became feasible only some decades after, when the discovery of some classes of polymorphisms as restriction fragment polymorphisms (RFLPs), minisatellites and microsatellite marker) allowed to map gene with a discrete effect on quantitative traits. From the first extensive genome scan carried out by (GEORGES *et al.* 1995), several genome investigations have been performed in livestock species across several countries using different breeds. These studies mainly focused on QTL mapping by tracing the inheritance of microsatellite markers in group of progeny of sires that had different phenotypic expression according different experimental designs. Theoretically, the discovery of DNA regions that affect the traits of economic interest should have resulted in the use of marker information into marker assisted selection (MAS) programmes. Approaches to estimate the genetic merit of individuals in MAS schemes using BLUP have been proposed by FERNANDO and GROSSMAN (1989). Moreover, approaches based on selection index theory have been suggested to combine classical polygenic EBVs with marker information (LANDE and THOMPSON 1990) in order to maximize the response to selection. In spite of the large number of QTL mapping studies that have been carried out and the amount of markers found to be associated to quantitative traits, commercial implementations of MAS have been limited for several year for three main reasons: the relatively small amount of variance explained by the detected QTLs (except few limited cases); the nature of most of the markers found to be associated with phenotypic differences, which are in linkage equilibrium in the population; the imprecise estimate locations of the QTL, mapped using the linkage analysis.

In the present chapter, the limit of application of MAS in cattle and results and the drawback of using meta-analysis to refine the position of QTLs will be discussed. In the second part of the chapter it will be discussed the evolution of high throughput sequencing technology that allowed to sequence the entire genome and to develop panel of tens of thousands of single nucleotide polymorphisms (SNP) used also to provide estimation of genomic breeding values. The genomic selection (GS) as defined by MEUWISSEN (2007) could be seen as special case of MAS on wide scale. Theoretical formulations and statistical methods to deal with such high number of marker data have been initially proposed by MEUWISSEN *et al.* (2001) thereafter, many statistical model have been proposed in literature to capture the genetic variance due to the SNPs. As the model adopted, as well as other factors, influence the accuracy of genomic prediction, in the second part it will be presented the factors that affect the accuracy. In the third part the use of a particular statistical model (PCA) to cope with the curse of dimensionality that affect the genomic selection implementation will be treated.

5.1 QTL DETECTION AND MAS PROGRAM.

In the original preposition, QTL mapping had to enable the identification of genes underlying economic quantitative traits in alternative to the candidate gene approach. Initially, the focus on QTL mapping was on milk production traits, but later the expectation was higher specially for low heritability traits, functional and reproductive traits, as genetic improvement of these traits according to the classical quantitative genetic model were slow and costly (DEKKERS 2004). The theoretical advantage of MAS scheme over traditional selection has also been confirmed by computer simulation. According to MEUWISSEN AND GODDARD (1996) the expected genetic gain increases using markers in breeding scheme were worthwhile to justify cost of genotyping. These authors found that and large extra rates of genetic gain (up to 64% for traits measured on the candidate, but high also for trait measured in the relatives of candidates), especially in the case of continuous detection of QTL.

After more than 15 years of QTL mapping studies just few example of unambiguous associations of gene with phenotypic expression have been reported in literature. DGAT1 in dairy cattle is an example of one polymorphism that explain around 50% of the genetic variance for milk fat content followed by GHR that explained about 10% of genetic variance of milk yield. Other markers which have been associated with production or functional traits explained very low percentage of genetic variance. This fact led to a substantial inefficiency of MAS, considering the cost of genotyping especially for functional traits, for which the expected advantage were higher. The low genetic variance explained by markers have been pointed out also by MANOLIO *et al.* (2009) in a recent study where the authors made some hypotheses about the causes of low genetic variance captured by SNP markers. In particular, they referred to missing heritability in human height (very high heritable traits with h^2 estimates greater than 80%) where the SNP assay capture only the 5% of the genetic variance. Causes of these fail may be found in the structure of Genome (Copy number variation (CNV) of gene which have not taken into account in SNP association studies) and the rare variants of allele having a great contribution to the genetic variance but very low allelic frequency (in GWAS and GS studies allele with MAF < 0.2 or 0.5 are discarded).

Despite the expected benefit of the use of markers in breeding scheme, MAS on large scale had a limited application till the early 2000's. Successful examples of MAS in its original formulation are reported for German (BENNEWITZ *et al.* 2004b) and French experiences (GUILLAUME *et al.* 2008b). Although the number of sire genotyped in the French MAS program was quite high (16,000 animal genotyped for 43 markers underlying 12 QTL region) the average accuracy of MAS-EBVs was just around 4% higher than in case of polygenic EBVs for unphenotyped animals (GUILLAUME *et al.* 2008b). In the German MAS programme, around 5 thousand animals were genotyped and 13 markers underlying 3 QTL region affecting milk production traits were used. Whole pedigree and top down approach to preselect bulls before entering in the progeny test scheme have been proposed in France and German respectively, both using the FERNANDO and GROSSMAN (1989) MAS-BLUP procedure.

Another reason that may to explain the limited MAS implementation is that QTLs have mostly been mapped with very large confidence interval (CI), due to the low marker density and use of LE markers (KHATKAR *et al.* 2004). Furthermore, in some cases the low power of the experiment together with the high cost of genotyping and low percentage of variance explained by QTL compromised the application of MAS program as originally designed. As shown by SCHROOTEN (2003) and reported in figure 1, for low heritability the power to detect QTLs under a certain size was low. The Increase of the sample size allow to have more power to detect QTL of small size. Statistical techniques that allow to fully exploits pedigree information from maternal line have been developed, they have been barely applied (DE KONING 2006; RON and WELLER 2007).

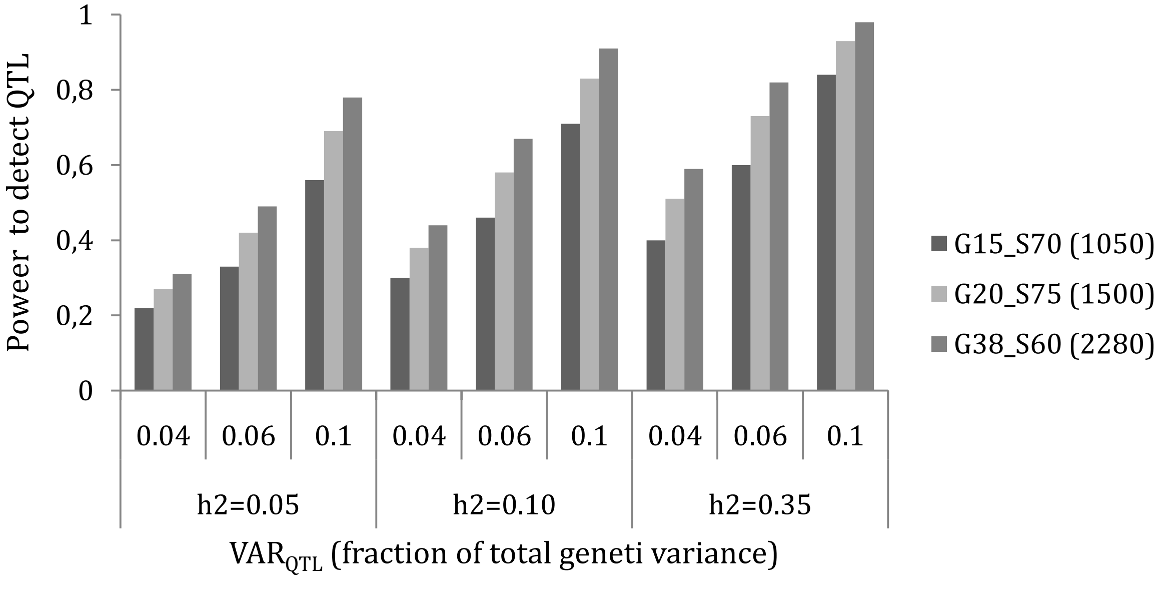


Figure 1. Power to dected QTL as function of variance due to QTL and heritability for different size granddaughter design (Gx_Sy (N) where x=no. Grand Sire, y = no. of sire, N= no. of animal) (from Schrooten 2003)

To refine the position of QTL under 2-3 cM CI, statistical methods have been proposed (identity by descend QTL mapping (RIQUET *et al.* 1999) and combined Linkage disequilibrium and linkage analysis approach (FARNIR *et al.* 2002; MEUWISSEN *et al.* 2002). Likewise the previous case these techniques had not a broad diffusion, at least in animal breeding field, probably for their relevant theoretical and computational complexity.

Meta-analysis techniques may lead to a better exploitation of a large amount of data that have been produced in more than a decade of research in this field. The meta-analysis of QTL mapping studies is a quite complicate tasks due to the great heterogeneity of published results: i) some values are missed (sometimes the effects of the QTL, or the statistical significance); ii) size of the experiment; iii) experimental design used; iv) different breed used; v) different genetic map, vi) genetic model and vii) statistical methodology used. Moreover, many hypotheses are generally tested simultaneously and multiple testing issues arose. Different authors addressed this issue in different ways (Bonferroni correction, False Discovery Rate (FDR), Proportion of False positive (PFP), Q-values, Bootstrapping, Permutation test, ecc.). All these difference in data processing and hypothesis testing further complicate the comparison of QTL across studies. All these problem should be taken into account when the comparison of results is carried out. For instance, the main question is generally :"Are QTLs found in different studies on the same chromosomal region the same?". In other scientific fields, meta-analysis is generally carried out combining data from different experiments with the aim to increase the power of the *meta-analysis experiment* and to extend the range of variability of the variable under exam, excluding the random effect of the study. Thus, meta-analysis could be also used in a predictive fashion. The case of QTL is a quite difficult case and different approaches have been used in literature instead (KHATKAR *et al.* 2004; SMARAGDOV *et al*, 2006).

The approach developed in chapter 1 of the present thesis follows a different logic from other published meta-analysis but challenges the same statistical issues of QTL meta-analysis, specially for the high heterogeneity of QTL data. The aim of the present study was to find some latent variables able to aggregate the information provided by different sources of variability (7 variables that characterize each QTL record were used) able to measure the reliability of a specific QTL to be of interest for a particular trait. Multivariate factor analysis was used at this purpose. Three common factor were found associated *marker map index*, the *dimension of the study* and the *power of the experiment.* However, the use of these variables to classify the QTL do not help to address the issue of scoring unambiguously the QTL on the basis of their "reliability".

The methodology suggested by GOFFINET and GERBER (2000) rely on the identification of the consensus position for *n* QTL detected in *n* published reports. The method evaluates the goodness of fit criteria (Likelihood Ratio Test (LRT), Aikake Information Content (AIC), ecc) of the data with 1,2 or *n* different QTLs hypothesized (in the worst case scenario 1 different QTL for each study). This approach has been adopted by KHATKAR *et al.* (2004) to find consensus on position of QTLs. They found the aforementioned constraints due to the incomplete information provided in the published papers, and about half of the published reports were discarded because they did not meet the basic requirement to be analyzed. They summarized the information of 55 papers aggregating QTLs for milk production traits in the most frequent analyzed chromosomes in dairy cattle (BTA3, BTA6, BTA9, BTA14, BTA20). In particular, in chromosome 6 they found two precise regions (49 cM and 87 cM) that aggregated several QTLs across different studies. Nevertheless, this approach led to a too wide 95% CI of the QTL position in most cases (shorter CI were found for QTL mapped in more precise way in the original studies). For the implementation on MAS program this approach may be useful to reduce the number of effect and variance components to estimate (considering for example less genomic regions). Conversely, this information may not be exploited in the whole population because the CI were still too large. The drawbacks presented here about QTL meta-analysis were also found in another study (SMARAGDOV *et al*, 2006) with different methodology but with similar conclusion on refinement of QTL position.

5.2 SIMULATION OF GENOMIC SELECTION

None of the approach of meta-analysis lead to a fully exploitation or this large amount of marker data. Solution came out just after the bovine genome sequence were available, due to the development of high throughput sequencing technology (VAN TASSELL *et al.* 2008). The huge availability of genetic markers – 54 K SNP chip available for cattle at relatively low cost per SNP and 800 K SNP-chip in project – allow to use LD markers to map QTL, or to predict directly the genomic breeding values (DGV) of genotyped animal with no phenotypic records (MEUWISSEN *et al.* 2001). DGV are then combined with the traditional EBV in a genome enhanced breeding value (GEBV) as shown by VANRADEN *et al.* (2009). The DGV estimation led to the development of different models with different predictive ability.

The genome-wide approach for estimating the breeding values of selection candidates allowed to overcome the issues of the position and effect of QTL. In fact, each marker is likely to be in LD with at least one QTL and so for that reason the position is no longer a key element. From the predictive point of view and breeding values estimation, the knowledge of the position of QTL may be negligible in a genome-wide approach. Conversely, if this tool is used to map QTL, the dense genetic map allow to have more precise location of QTL on genome. In both of cases *ad hoc* statistical methods have been developed to deal with the great asymmetry of data matrix (tens of thousands of predictors and hundreds/thousands of experimental data point). Even if there are cases where the number of genotyped animals approaches the number of markers, as happens in the US genomic programme with almost 40000 animal for around 40000 SNP after editing (VANRADEN AND TOOKER, 2010) have been overcome increasing the sample size. However, the population size is generally lower than this figure, especially in Europe where national breed association handle genomic selection programmes independently. The issue of sample size is of course enhanced in the case of minor breeds

The size of the population as well as the heritability of the trait and the statistical model used to predict DGV affect heavily the accuracy of the genomic prediction. The first optimistic estimates (MEUWISSEN *et al.* 2001; SCHAEFFER 2006), have been resized when real data come out across different breed in several country. The DGV accuracy drop is limited for population where the training set is quite high. The main factors affecting the accuracy of genomic predictions according to the study presented in the simulation presented in the current thesis were: i) the marker density, ii) the heritability iii) the number of daughter per bulls required to compute the daughter yield deviation, and finally the iv) choice of the predictors used in the statistical model (haplotype or single marker genotype), even though this latter factor is indirectly linked to the marker density as shown by CALUS *et al.* (2008). In all scenarios simulated the BLUP was used to estimate the marker or haplotype effects. However the choice of an appropriate statistical model affect the accuracy of prediction as well.

5.3 PCA APPLIED TO GENOMIC SELECTION

Several models have been tested in simulations but mainly BLUP, Bayesian method, multivariate and non parametric methods have been used. Several simulations shown that Bayesian methods performed better than BLUP and other methods in DGV predictions, especially for high density marker maps (MEUWISSEN *et al.* 2001; SOLBERG *et al.* 2008). However the application of these bunch of methods to real data showed that the differences among methods were less relevant than in simulated dataset across studies. In particular BAYES, G-BLUP , PCA or PLSR approach perform in most case similarly.

In the present thesis, principal component (PC) analysis was presented with the aim to reduce the number of predictors in the estimation of genomic breeding values for a simulated population. Priors of predictor variance were based on their contribution to the total SNP correlation structure. The PC approach yielded the same accuracy of predicted genomic breeding values obtained with the regression using SNP genotypes directly, with a reduction in the number of predictors of about 96% and computation time by 99%. The high number of predictors generally used in genomic selection require a selection or reduction of dimensionality of the data matrix. In some case (ordinary least squares models) the high number of predictors in comparison with the data point do not allow to have enough degree of freedom (MEUWISSEN *et al.* 2001). In any case a pre-selection step may be required also for other methods. The BLUP approach can overcome the problem of degree of freedom, adding some penalization (lambda) to the solution allowing to shrunk toward 0 the domine of the solutions (figure 2).

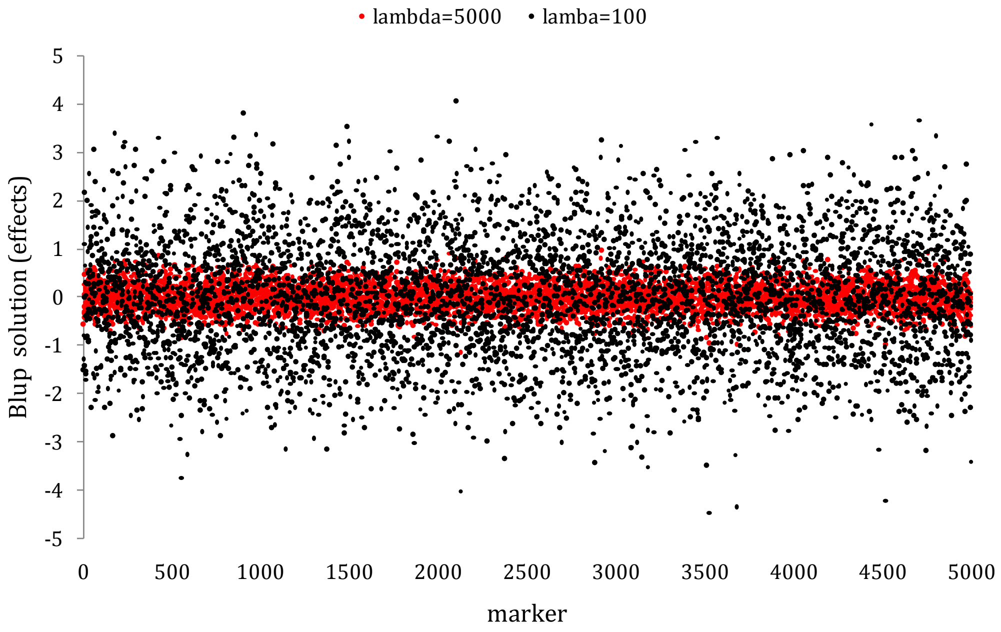


Figure 2. BLUP solution with different values of lambda (100 and 5000) in a simulated dataset with 2000 observation and 5000 marker. Data were simulated using R statistical Package

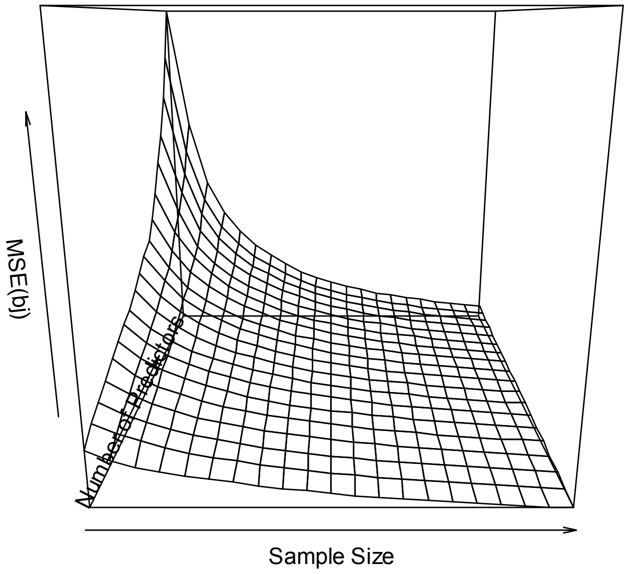


Figure 3. Mean squared error of prediction (MSE) as function of sample size and number of predictors

These results were obtained a simple simulation of a dataset of 2000 individual and 5000 markers, but similar results have been also showed by PIMENTEL *et al.* (2009). However introduction of these penalization introduce bias in the predictions. In figure 3 has been shown the Mean squared error of prediction (MSEP) for a dataset downloaded from in function of number of predictor and sample size. The MSEP decrease as the number of phenotypic record increase, but increase dramatically as soon the number of predictor grow.

The choice of an appropriate values of lambda is discussed in literature and different approach have been adopted (estimation of the genetic and residual variance with REML approach and division of by the number of effect to estimates; the estimation of lambda in a Bayesian context is carried out introducing some prior information about the distribution of this random variable). In figure 4a is reported the MSE for a simulated dataset (BLR package v2.1 http://cran.r-project.org/) for different value of shrinkage factor lambda. Whilst the figure 4b shows as the degree of freedom drop down for higher values of lambda. The figure 4a indicate empirically that a value of lambda around 10 in this case minimize the MSE and allow a sufficient number degree of freedom to estimate the DGV shrinking toward 0 the solution of the system of equation.

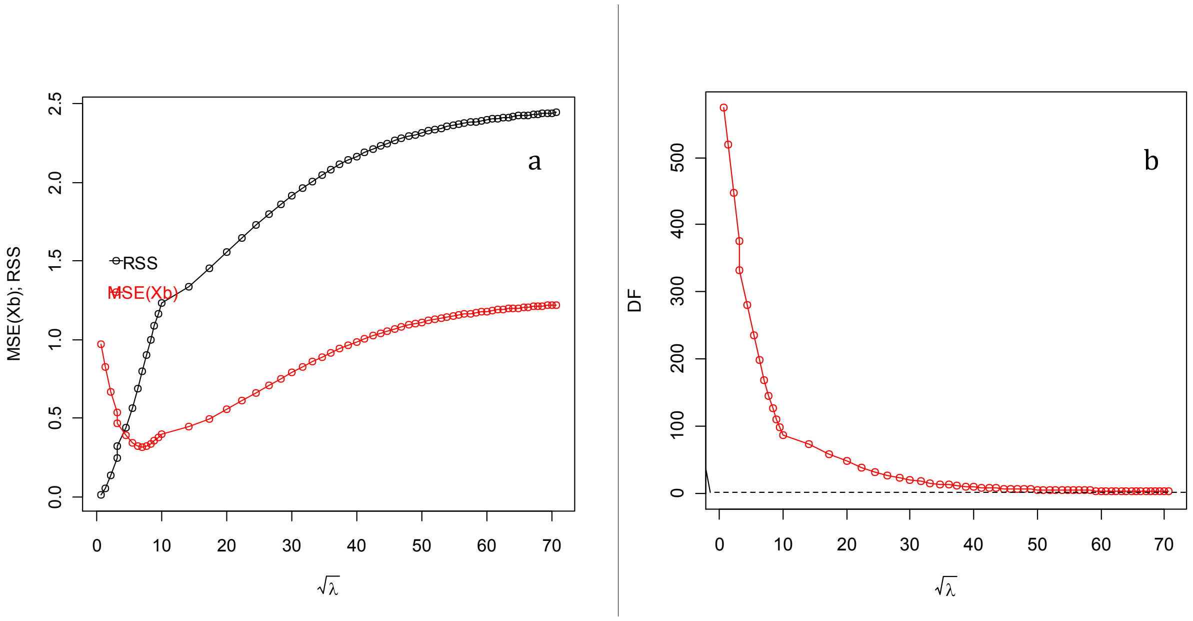


Figure 4. (a) MSE and residual sum of square (RSS) for a dataset of 599 records and 1477 markers as function of lambda parameter (b) trend of degree of freedom as function of lambda (data generated from Wheat lines were recently genotyped using 1447 Diversity Array Technology (DArT) generated by Triticarte Pty. Ltd. (Canberra, Australia; http://www.triticarte.com.au).

The PCA allow to have less bias estimated of DGV, reducing dramatically the number of effects to be estimated in the training generations. Furthermore the use of variables that are uncorrelated among them reduce the problem of collinearity as well.

Some results in real data seems to confirm the goodness of PCA methodology. Figure 5 showed the regression of DGV on calving ease (CE) polygenic EBV of 323 Piedmontese bulls divided into training set (5a) and validation bulls (5b). In both dataset the estimates of PC approach produce estimate that are less biased than whole BLUP approach. The low R^2 for the second regression may be due to the low sample size, and even in this extreme case the PC perform better than the BLUP approach (AJMONE *et al*, 2010)

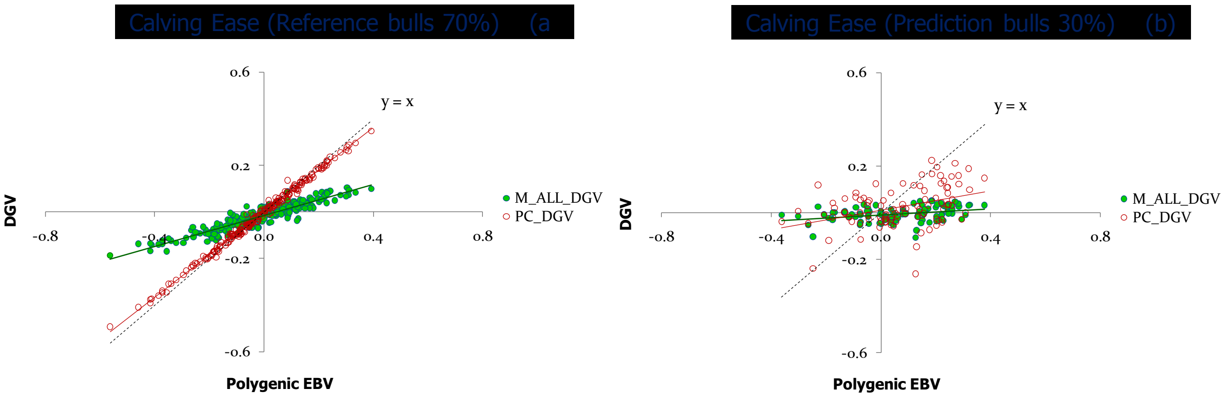


Figure 5. Regression between DGV predicted using the PC (PC_DGV) or all markers (M_ALL_DGV) and polygenic EBV for calving ease bulls in the set of reference (a) and in bulls estimates (b).

The reduction of predictors, so seem to be a way to deal with the major issue on genomic selection and although the accuracies of PC are lower than those currently achieved with Bayesian methods, at least for simulated data, the improved calculation speed together with the possibility of extracting principal components directly on individual chromosomes may represent an interesting option for predicting genomic breeding values in real data with a large number of SNPs. Furthermore the increasing availability of marker data is leading toward a more easy solution of the curse of dimensionality. Even though soon denser chip (800 K) will out in the marker and this will open new challenge in DGV estimations. In this context, PCA may be a suitable solution to speed up the calculation without losing too much in accuracy.

REFERENCES

AJMONE-MARSAN P, N.P.P. MACCIOTTA, M.A. PINTUS, G. GASPA, C. PIERAMATI, E. NICOLAZZI, A. ALBERA A. NARDONE, A. VALENTINI 2010. Accuracies of Direct Genomic Breeding Values for calving ease estimated on Italian Piedmontese bulls with a principal component approach. Poster presented in ISAG 2010- International Society for Animal Genetics Conference. Edinburgh 26-30 July 2010..

BENNEWITZ, J., N. REINSCH, F. REINHARDT, Z. LIU and E. KALM, 2004 Top down preselection using marker assisted estimates of breeding values in dairy cattle. Journal of Animal Breeding and Genetics **121:** 307-318.

CALUS, M. P. L., T. H. E. MEUWISSEN, A. P. W. DE ROOS and R. F. VEERKAMP, 2008 Accuracy of genomic selection using different methods to define haplotypes. Genetics **178:** 553-561.

DE KONING, D. J., 2006 Conflicting candidates for cattle QTLs. Trends Genet **22:** 301-305.

DEKKERS, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. J. Anim Sci. **82:** E313-328.

FARNIR, F., B. GRISART, W. COPPIETERS, J. RIQUET, P. BERZI *et al.*, 2002 Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: Revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics **161:** 275-287.

FERNANDO, R., and M. GROSSMAN, 1989 Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution **21:** 467 - 477.

GEORGES, M., D. NIELSEN, M. MACKINNON, A. MISHRA, R. OKIMOTO *et al.*, 1995 Mapping Quantitative Trait Loci Controlling Milk-Production in Dairy-Cattle by Exploiting Progeny Testing. Genetics **139:** 907-920.

GOFFINET, B., and S. GERBER, 2000 Quantitative trait loci: A meta-analysis. Genetics **155:** 463-473.

GUILLAUME, F., S. FRITZ, D. BOICHARD and T. DRUET, 2008 Short Communication: Correlations of Marker-Assisted Breeding Values with Progeny-Test Breeding Values for Eight Hundred Ninety-Nine French Holstein Bulls. J. Dairy Sci. **91:** 2520-2522.

KHATKAR, M. S., P. C. THOMSON, I. TAMMEN and H. W. RAADSMA, 2004 Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genetics Selection Evolution **36:** 163-190.

LANDE, R., and R. THOMPSON, 1990 Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics **124:** 743-756.

MEUWISSEN, T., 2007 Genomic selection : marker assisted selection on a genome wide scale. Journal of Animal Breeding and Genetics **124:** 321-322.MEUWISSEN, T. H. E., B. J. HAYES and M. E. GODDARD, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157:** 1819-1829.

MEUWISSEN, T. H. E., A. KARLSEN, S. LIEN, I. OLSAKER and M. E. GODDARD, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics **161:** 373-379.

NEIMANN-SORENSEN, A., and A. ROBERTSON, 1961 The Association between Blood Groups and Several Production Characteristics in Three Danish Cattle Breeds. Acta Agriculturae Scandinavica **11:** 163 - 196.

PIMENTEL, E. C., S. KONIG, F. S. SCHENKEL and H. SIMIANER, 2009 Comparison of statistical procedures for estimating polygenic effects using dense genome-wide marker data. BMC Proc **3 Suppl 1:** S12.

RIQUET, J., W. COPPIETERS, N. CAMBISANO, J. J. ARRANZ, P. BERZI *et al.*, 1999 Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. Proc Natl Acad Sci U S A **96:** 9252-9257.

RON, M., and J. I. WELLER, 2007 From QTL to QTN identification in livestock - winning by points rather than knock-out: a review. Animal Genetics **38:** 429-439.

SAX, K., 1923 The Association of Size Differences with Seed-Coat Pattern and Pigmentation in PHASEOLUS VULGARIS. Genetics **8:** 552-560.

SCHAEFFER, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics **123:** 218-223.

SOLBERG, T. R., A. K. SONESSON, J. A. WOOLLIAMS and T. H. E. MEUWISSEN, 2008 Genomic selection using different marker types and densities. Journal of Animal Science **86:** 2447-2454.

THODAY, J. M., 1961 Location of Polygenes. Nature **191:** 368-370.

VAN TASSELL, C. P., T. P. L. SMITH, L. K. MATUKUMALLI, J. F. TAYLOR, R. D. SCHNABEL *et al.*, 2008 SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth **5:** 247-252.

VANRADEN, P. M., C. P. VAN TASSELL, G. R. WIGGANS, T. S. SONSTEGARD, R. D. SCHNABEL *et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science **92:** 16-24.

BLR package v2.1 http://cran.r-project.org/)