Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Accuracy of Quantitative Trait Nucleotide (QTN) Prediction By Surrounding SNPs

Marzieh Heidaritabar

Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics
461
Uppsala 2011

Master's Thesis, 30 hp

Erasmus Mundus Programme – European Master in Animal Breeding and Genetics

Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences

Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Accuracy of Quantitative Trait Nucleotide (QTN) Prediction By Surrounding SNPs

Marzieh Heidaritabar

Supervisors:

Prof. Theodorus Meuwissen, NMBU, Department of Animal and Aquacultural Sciences
Freddy Fikse, SLU, Department of Animal Breeding and Genetics

Examiner:

Erling Strandberg, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: EX0556
Programme: Erasmus Mundus programme
– European Master in Animal Breeding and Genetics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2011
Name of series: Examensarbete / Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics, 461
On-line publication: http://epsilon.slu.se

Key words: Genomic selection, quantitative trait nucleotides, quantitative trait loci, single nucleotide polymorphisms

Education and Culture
Erasmus Mundus

EUROPEAN MASTER OF SCIENCE IN ANIMAL BREEDING AND GENETICS

Accuracy of Quantitative Trait Nucleotide (QTN) Prediction By Surrounding SNPs

Marzieh Heidaritabar

Main supervisor:
Prof. Theodorus Meuwissen
Department of Animal and Aquaculture Sciences
Norwegian University of Life Sciences
Ås, Norway

Co-supervisor:
Freddy Fikse
Swedish University of Agricultural Science
Uppsala, Sweden

Preface

This thesis was written for fulfillment of the Double Masters Degree entitled with European Masters in Animal Breeding and Genetics which is funded by the European Union. This was done during the period 1th of January 2011 to 30th of May 2011 at the department of Animal and Aquaculture Sciences at Norwegian University Of Life Science.

The support for my work was provided by the Norwegian University of Life Science and Swedish University of Agricultural Science.

Norwegian University of Life Science

Marzieh Heidaritabar

Date: 06/06/2011

Acknowledgements

I would like to express my great gratitude to my supervisor Professor Theodorus Meuwissen for his continuous support and his great efforts. Without his encouragement, I would not have come to the end of the thesis.

Special thanks to my co-supervisor Freddy Fikse for his comments on the manuscript.

I thank to Professor Johan A.M. van Arendonk for organizing the Program, thanks also to the people in the three universities I visited for their support and kindness.

My deepest gratitude goes to my family for their love and support throughout my life.

I am most grateful to Alireza, whose love and continued support enabled me to overcome the frustrations which occurred in the process of doing this thesis.

Last but not least, thanks to God, the merciful and the passionate, for providing me the opportunity to step in the excellent world of science.

Table of Contents

List of Figures

List of Tables

Abbreviations

Acc	Accuracy
CV	Cross Validation
DGAT1	diacylglycerol acyltransferase 1
EBV	Estimated Breeding Value
G-BLUP	Genomic Best Linear Unbiased Prediction
GEBV	Genomic Estimated Breeding Value
GHR	Growth Hormone Receptor
GS	Genomic Selection
GWAS	Genome Wide Association Studies
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MAS	Marker Assisted Selection
QTL	Quantitative Trait Loci
QTN	Quantitative Trait Nucleotide
SD	Standard Deviation
SE	Standard Error
SNP	Single Nucleotide Polymorphism

Abstract

Information from thousands of markers distributed across the genome can be used for a new selection method in animal breeding and genetics. This method which is called genomic selection estimated the genomic breeding value based on the estimation of marker effects covering the whole genome. For a successful application of genomic selection, accuracy of the prediction is an important factor that should be considered. Because it is important for genetic progress. Quantitative trait nucleotides (QTN) are polymorphisms that give useful information about gene function and QTL architecture. So prediction its effects and estimation its accuracy enhances rates of genetic gain. In this study, we investigated the accuracy of QTN prediction by neighboring markers, using the simulated data. Our dataset consisted of 1040 markers which were assigned to one chromosome of 500 genotyped animals. Method G-BLUP was used to estimate marker effects, and the accuracy of QTN prediction was estimated by using cross validation. As the accuracy can be affected by different number of surrounding SNPs, it was predicted at various number of surrounding markers ranging from 10 to 100 markers. In general, the accuracy of QTN prediction increased by increasing the number of flanking SNPs from 10 to 60 SNPs. Further increase in number of SNPs resulted in a very small increase in accuracy in case of heritability 1 and 0.8 and a very small decrease in case of heritability 0.5. We also investigated the effect of other factors on accuracy such as Minor Allele Frequency cutoff threshold, heritability and number of phenotypes in the training set. We analyzed four data sets; data set with no selection of markers, data sets with different cutoff thresholds for MAF (0.02, 0.05 and 0.1) in order to get the effect of MAF on accuracy. We observed the minimum SNP MAF of 0.02 is more appropriate for genomic selection studies. After filtering the data with the cutoff threshold of 0.02 for MAF, QTN could be predicted with 100 flanking SNPs, with a

maximum accuracy of 0.777. This is the maximum accuracy in the absence of any environmental effects. We also observed that there is a relationship between the accuracy of QTN prediction and the heritability of the phenotype. The accuracy of QTN prediction dropped when the heritability of phenotype decreased. In general, when we estimated the accuracy by 100 surrounding SNPs and heritability decreased from 1 to 0.8 and from 0.8 to 0.5, the decrease in accuracy was 4.6 and 11%, respectively. In another analysis, when 50% of animals were masked, it means that the number of phenotypes decreased in training set, the accuracies were lower in comparison to 20% masking. When 50% of animals were masked, with 100 surrounding SNPs, the reduction of 6 and 9.25% was observed, when heritability of phenotype was 1 and 0.5, respectively.

1. Introduction

Genomic selection (GS) is a marker based method that predicts breeding values for quantitative traits on the basis of a large number of molecular markers, which cover the whole genome. Using enough dense genome wide marker maps, a large part of genetic variance is expected to be explained by the markers, and all quantitative trait loci (QTL) are in linkage disequilibrium (LD) with at least one marker (Meuwissen *et al.* 2001). This approach has become possible particularly because of the discovery of high-throughput genotyping methodology, the development of the novel science bioinformatics, the identification of polymorphisms at DNA level and identification of new single nucleotide polymorphisms (SNPs) (Hocquette *et al.* 2007). SNPs are the most frequent type of DNA variation in the genome and using them are preferred over other genetic markers due to two reasons: they have low mutation rate, which makes them more useful for investigations of the history of populations and the second reason is that it is easy to genotype them (Romualdi *et al.* 2002; Youngerman *et al.* 2004).

For a successful application of genomic selection, accuracy of the prediction is a key issue that should be considered (Goddard and Hayes 2009). Since the suggestion of genomic selection method by Meuwissen *et al.* (2001), many studies using simulated data have been done on this area (Habier *et al.* 2007; Calus *et al.* 2008). Also, research about the accuracy of genomic predictions has been performed in some animal species such as dairy cattle (Hayes *et al.* 2009) and chicken (Gonzalez-Recio *et al.* 2009). Meuwissen *et al.* (2001) predicted breeding values based on haplotype effects and found the maximum accuracy of 0.85 from marker data alone. Kolbehdari *et al.* (2007) also conducted a simulation study and obtained the same results. They found the accuracy of around 0.80.

According to the studies that have been done in genomic selection so far, by direct selection on QTLs, the genetic gain can be increased (Weller and Ron 2011). Till now, in all major livestock species, genome scans of QTL have been completed. The genotype of QTL is determined just for a few QTL with relatively large effects. The confidence interval for QTL location by linkage analysis still spans hundreds of genes. Detection of a polymorphism which is the casual mutation underlying the QTL, decreases the confidence interval obtained during linkage analysis. This polymorphism is called quantitative trait nucleotide (QTN) (Ron and Weller 2007).

Till now, the methods that were applied to genomic selection did not need the QTN prediction and estimation of its accuracy. As QTN give useful information about gene function and QTL structure and helps us to understand the mechanisms through which the trait is influenced (Weller and Ron 2011), prediction of its effect and estimation of its accuracy enhances rates of genetic gain.

In this study, we applied genomic selection to simulated data to predict the QTN effect, from surrounding SNPs. The method which is used for QTN prediction is Best Linear Unbiased Prediction (BLUP). Then, accuracies of QTN prediction will be assessed. Prediction of QTN with high accuracy would increase the rate of genetic gain. Since it is expected that the difference in number of SNPs surrounding the QTN is an important factor, accuracy was obtained at various number of surrounding markers, ranging from 10 to 100 markers. Other factors that can affect on genomic selection accuracy such as Minor Allele Frequency, heritability and size of test set and training set are also investigated. We analyzed a data set of 1040 SNP markers which was allocated to one chromosome and genotyped on 500 animals. These analyses were performed for 25 replicates.

Specifically; the aim of our study is:

- To obtain the accuracy of QTN prediction by surrounding SNPs
- To study the accuracy of QTN prediction by increasing number of SNPs surrounding the QTN
- To study the accuracy of QTN prediction using different cutoff thresholds for low Minor Allele frequencies
- To study the accuracy of QTN prediction with different levels of masking
- To study the accuracy of QTN prediction with different heritabilities

2. Literature Review

2.1. Quantitative Genetics

Quantitative genetics is the study of traits that indicate a continuous range of values and also the study of the mechanisms of those traits. Many genes control quantitative traits resulting in a continuous distribution of genetic values. The loci that control quantitative traits are named quantitative trait loci (QTL).

There are three kinds of quantitative traits. The first type is continuous traits in which a continuous phenotype expression of the traits can be distinguished. Some examples are milk yield traits and growth rate. The second type is qualitative traits. For these traits, the phenotype expression is in a discrete form. The pattern of inheritance for these traits is monogenetic, which means that just a single gene controls the trait. The environment has little effect on the phenotype of these traits. Some examples are number of eggs and blood type. The third type is threshold traits. These are continuous traits that have only two or a few phenotypic classes, but their inheritance is determined by the effects of multiple genes. Also, the environment affects the phenotypic expression of these traits. Some examples are twining in cattle, fertility, mastitis in dairy cows and human genetic diseases.

2.2. Genetic markers

Genetic markers are loci whose alleles can be used to keep track of a chromosomal region during the transmission form parent to offspring. So, according to this definition, genetic markers are polymorphic. Molecular biology provides us with a wide range of new genetic markers. From the beginning of the concept of genetic markers, three kinds of markers such as morphological

markers, protein based markers (like blood groups) and DNA based markers have been utilized in various fields to identify the variation among genotypes (Liu 1998).

Some examples of genetic variation that happen at the DNA level are base substitutions (mostly SNPs), insertions and deletions of nucleotides and rearrangement of DNA segments around a locus of interest (Liu and Cordes 2004). These variations or polymorphisms that exist among individuals in the population for specific regions of the DNA have been detected by molecular techniques. These polymorphisms can be used for making genetic maps. They can also be used for evaluation of differences between genetic markers in the expression of a trait in a family that may show a direct effect of these differences in terms of genetic determination on the trait (Stein *et al.* 1996).

Among DNA markers, Single Nucleotide Polymorphism (SNP) are the newest that have been developed. Although it is just a bi allelic marker, its use has become common. A SNP marker can be developed when at a particular position in the genome, a single nucleotide differs between animals. For such a position to be considered as a SNP, the least frequent allele should have a frequency of 0.01 or more (Vignal *et al.* 2002).

There was a great development in the technology for SNPs genotyping. In many livestock species, many SNPs markers have been discovered and there are also more SNPs discovered each day because of the development of high-throughput genotyping technologies such as DNA sequencing technologies. For instance in human 500 000 SNPs are available nowadays. Nowadays, in animal breeding and genetics, SNPs are used in a new type of breeding value prediction method. This method which is called Genomic selection uses the genomic information of animals for prediction of breeding value.

2.3. Marker assisted selection (MAS)

Over the past 50 years, genetic progress through artificial selection has been an important contributor to the great advances in productivity in plant and animals which are of agricultural importance. So far, most selection were based on selection of individuals with superior phenotypes and the genetic structure of the selected traits was unknown (Dekkers and Hospital 2002). So, there were many problems regarding traditional selection. Some examples of these issues are: this method is not very efficient when the traits are difficult to measure or have a low heritability. Some traits are expressed late in life. In addition, when the selection objective includes several traits with unfavorable genetic correlation, the traditional selection is not very efficient (Schwerin *et al.* 1995). Using molecular techniques could solve some of the problems of traditional selection and made it possible to generate dense maps (Kinghorn *et al.* 1994). By discovering and analyzing the molecular genetics of traits in animal and plant populations, the genetics of quantitative traits were better understood. These genetic markers can be used to increase the genetic gain of livestock through marker assisted selection (MAS) approach (Dekkers and Hospital 2002). The purpose of MAS is to increase genetic gain, in terms of both accuracy and speed (Muir 2007). Application of MAS is mainly beneficial in situations where the accuracy of selection is low, for instance for the traits with low heritability, sex-limited traits, traits that are visible late in life, traits that are expensive to measure or can only be measured on relatives (Meuwissen and Goddard 1996).

Much research has been done by using MAS. However, its implementation has been limited and enhancement in genetic gain has been small (Dekkers 2004). Factors that affect genetic gain include intensity of selection, accuracy of selection, genetic standard deviation and generation

interval. Marker information mainly influences the accuracy of selection. Therefore, in MAS approach, if the accuracy of selection increases, genetic gain also increases.

In general, extra gains from MAS are because of the accuracy QTL prediction, the accuracy of existing estimated breeding values, the proportion of the genetic variance which is explained by genetic markers and reduction of generation interval (Goddard and Hayes 2002).

Besides the benefits of MAS in breeding programs, its implementation faces some problems. One issue is that for each trait, separate markers are usually required. If the markers and the QTL are linked, the linkage phase variants cause the markers to be incorrect in some families. Also, another issue that should be considered is that the linked QTL may have pleiotropic effects on other traits (Dekkers and Hospital 2002). One possible solution to these problems is a new method of selection called genomic selection that was first suggested by Meuwissen *et al.* (2001).

2.4. Genomic Selection

Genomic selection is a form of marker assisted selection, in which dense genetic markers are used that cover the whole genome. The effects of dense genetic markers, across the whole genome, are summed up in order to get the genomic estimated breeding value (GEBV) (Meuwissen *et al.* 2001). Dense markers should be distributed across the genome in equal spacing without prior knowledge of QTL positions. Finally, the available genetic variation that is in linkage disequilibrium (LD) with these markers is captured. The method works better and the LD will be stronger when the markers are more dense (Muir 2007). Genomic selection can change the structure of animal breeding programs. For example, in dairy cattle, it is used to select bull calves for progeny testing. Therefore, the sires of sons can be selected based on

markers and it is not needed to do the progeny testing. This resulted in reduction in cost about 92% for a breeding program. In addition, using genomic selection decrease the generation interval, because it is possible to genotype the markers at early age (Schaeffer 2006).

2.5. Quantitative Trait Nucleotide (QTN)

According to the research that has been done so far, by direct selection on QTLs, genetic gain can be increased (Weller and Ron 2011). Till now, in all major livestock species, genome scans of QTL have been completed. However, it is not enough to just detect the QTL for using it in breeding programs. So, in order to perform a very successful breeding program for the QTL, it is required to identify a type of specific polymorphism. This polymorphism that is responsible for the observed variation of QTL is called quantitative trait nucleotide (QTN). After identification of a QTN, if both of its alleles are segregating, it can be used by enhancing the frequency of the favorable allele within a breed. Also, the favorable alleles can be increased by introgression of the allele into breeds in which the allele is absent (Ron and Weller 2007).

Until 2000, QTNs could be identified just for plants, microbes and organisms that are used as model. Four QTNs have been identified in dairy cattle genome. The first QTN that was discovered in dairy cattle was found in a QTL that affect fat and protein percentages of milk. This QTN was found in the centromeric region of bovine chromosome 14. This QTN contributed a lysine to alanine substitution at the gene that encodes diacylglycerol acyltransferase 1, and was called DGAT1 (Grisart *et al.* 2002; Winter *et al.* 2002). Another QTN that has been identified in cattle is ABCG2 that has significant effects on milk yield and milk composition (Olsen *et al.* 2007). Blott *et al.* (2003) found another QTN by using fine mapping. This QTN which is called GHR was a phenylalanine to tyrosine substitution at the bovine growth hormone receptor gene.

This QTN was the direct cause of the effects that detected before. This QTN is also responsible for milk production traits. The fourth QTN discovered in bovine genome was osteopontin (SPP1) (Schnabel *et al.* 2005). It has been found that expression of the SPP1 influence the expression of milk protein genes, that shows the regulatory role for the gene product of SPP1 in lactation (Sheehy *et al.* 2009).

So far, the methods that were applied to genomic selection did not need the QTN identification and estimation its accuracy. As QTN gives useful information about gene function and QTL structure and helps us to understand the mechanisms through which the trait is influenced (Weller and Ron 2011), prediction of its effect and estimation its accuracy enhances rates of genetic gain.

2.6. Accuracy of genomic selection

The correlation between true breeding value and the estimated breeding value is measured as accuracy. Reliability is the square of accuracy. The implications of obtaining high accuracies for animals at early age are profound. In simulations that have been performed, it has been showed that the accuracy of genomic EBV for a bull calf can be as high as the accuracy of an EBV after the progeny test is done (Hayes 2008).

2.7. Factors affecting the accuracy of genomic selection

In a study in dairy cattle breeding program, Hayes *et al.* (2008) reported that four parameters influence the accuracy of genomic estimated breeding value (GEBV). The first one is linkage disequilibrium (LD) between the markers and QTL which is quantified with the parameter r^2

(Hill and Robertson 1968). Calus *et al.* (2008) reported that by increasing r^2 from 0.1 to 0.2, the accuracy of GEBV increased from 0.68 to 0.82.

The second factor that influences the accuracy is number of animals in reference set. These animals have both phenotypes and genotypes from which the SNPs effects are predicted. With more phenotypic observations per SNP allele, the accuracy of estimated breeding value will be greater. Saatchi *et al.* (2010) investigated the impact of heritability and number of phenotypic records on accuracy of genomic selection in a simulation study. They found that by increasing the number of animals in the training set, the accuracy increased in the test set particularly for low heritable traits. They obtained accuracies 0.573, 0.639 and 0.706 for 500, 1000 and 2000 records, respectively. Meuwissen *et al.* (2001) used different number of individuals in training sets and estimated the accuracy of genomic breeding values. By using G-BLUP model for analysis and by using data sets with different sizes of 500, 1000, and 2200 for training set, they obtained accuracies of 0.579, 0.659 and 0.732, respectively. In their study, the heritability of the trait was 0.5. De Roos *et al.* (2009) applied genomic predictions to multi-breed populations in a simulation study. They showed that combining two populations, by adding animals from a second population to the reference set, increased the accuracy of genomic predictions in the first population. They reported that this increase was most advantageous for the traits with lower heritability.

The third factor is heritability of the trait that is under selection. One advantage of higher heritability is that fewer numbers of animals are needed for genomic prediction. Calus and Veerkamp (2007) estimated the accuracies of GEBV for juvenile selection candidates in range of 0.38 to 0.55 for low heritability traits (with heritability 0.1) and 0.73 to 0.79 for the traits with heritability 0.5. They also showed that the accuracies are dependent on the number of phenotypes

in the training set particularly for the traits with low heritability. However, they showed that increasing the number of animals in reference set had limited effect for the high heritable trait. De Roos et al. (2009) showed that when the heritability of phenotypes is low, it is essential to have many individuals to estimate the marker effects. Having many individuals to estimate marker effects is more beneficial for low heritability traits than for high heritability traits.

The fourth factor is the distribution of QTL effects. If the number of QTLs that has very small effect contributing to the variation of the trait, is high, a large number of phenotypic records are needed to predict these effects accurately.

The first two factors are controllable in the experiments. However, the last two ones are not. In another study, Villumsen *et al.* (2008) reported that reliabilities are expected to depend on some factors such as on statistical method that is used, SNP frequencies, size of the data, on marker density, and on genetic event such as recombination and LD structure.

2.7.1. Minor Allele Frequency

SNPs are bi allelic and have two alleles. The less common allele of a SNP in a population has lower frequency. This frequency is defined as Minor allele frequency (MAF). When in a population, MAF for a SNP is less than 0.005, it is said that the SNP is monomorphic (http://www.experts123.com/q/what-is-the-frequency-of-a-certain-allele-in-the-following-populations.html). It means that just a single form or allele can be identified. In genome-wide association studies (GWAS), monomorphic SNPs are not informative, because there is not any genotypic difference (Chan *et al.* 2008). When a SNP is monomorphic in a population, no heterozygote individual can be found for that SNP in that population. When a SNP is monomorphic, it dose not mean that it is monomorphic in all populations. It might be

polymorphic in a different population. Wiggans *et al.* (2009) found thousands of SNPs that were monomorphic in Jerseys or Brown Swiss. However, they were polymorphic in Holstein breed. They concluded that in selection of SNPs for the genomic evaluations, the breed specific SNP sets should be considered.

When the plan is to select the SNPs for genomic evaluations, one important issue that is needed to be considered is MAF. There should be cutoff threshold for MAF. Wiggans *et al.* (2009) suggested a cutoff threshold of 0.02 for MAF. They believe that in genomic evaluations the purpose is to maximize the accuracy and SNPs with MAF less than 0.02 do not contribute to the accuracy. One reason of removing them is to decrease the computational challenges and another reason is to increase the stability of predictions of the effects of those SNPs that remained in the analysis (Wiggans *et al.* 2009).

ShiYi *et al.* (2009) investigated the effect of different cutoff thresholds (0.01, 0.05 and 0.1) on the resolution of Haplotypes map. They concluded that SNP allele frequency is one of the most important factors influencing the resulting HapMap. According to their findings, the average number of total haplotypes discovered decreased when cutoff values increased. So, they defined a cutoff threshold of 0.01 for MAF.

3. Materials and Methods

3.1. Simulation of the data

In this study, forward simulations were utilized (Hoggart *et al.* 2007). There were some assumptions about the population model which includes: the Fisher–Wright idealized population model (Falconer and Mackay 1996) and the model assumed was the infinite-sites mutation model (Kimura 1969). The mutation frequency was 2×10^{-8} per nucleotide per generation. The effective population size was 1000. For obtaining a population which is in balance of drift-mutation, 10,000 generations forward simulations were performed. According to the Haldane mapping function, the recombination frequency was assumed to be 10^{-8} per nucleotide per generation.

After simulation of 10000 generations of the whole-genome sequence of one chromosome, 1040 SNPs were generated. These 1040 marker alleles genotyped on 500 animals were used for the analysis. So, the data structure had 1000 rows and 1040 columns (number of markers). Two rows were allocated to each animal, it means that each locus had two alleles.

3.2. Data information

For this analysis, one chromosome with 1040 SNP markers was used. The number of records were 500 animals. Each animal had two alleles for each marker. The heritability of phenotype was 1.

The alleles were converted to genotypes by using Matlab Program. For individuals that had allele 1 from both parents, the genotype was set as 1. For those that had allele 1 and 2 from the parents, the genotype was set to be 2 (heterozygote) and for the individuals that had allele 2 from both parents, the genotype was set as 3. Then, the genotypes were standardaized to obtain a total

genetic variance of 1. For genotypes 11, 12, and 22, the standardized genotype of individual *i* for marker j (X_{ij}) is $-2p_j/\sqrt{h}$, $(1-2p_j)/\sqrt{h}$ and $2(1-p_j)/\sqrt{h}$,respectively. Where, p_j is allele frequency at locus *j* and $h = 2p_j(1-p_j)$ i.e. hetrozygosity.

3.3. Training data

Masking the phenotype is a way for making different cross validations. It means that the phenotype will be covered and considered as unknown for a defined number of animals. For making the first cross validation (CV1), as there were 500 animals in the dataset, phenotypes of the first 100 animals were masked as testing sample. For the second CV (CV2), phenotypes of the second 100 animals were masked as testing sample. This method of masking continued to generate five training samples and five test samples. As this method was without replacement, every phenotype was masked once in the training data. The training and test samples were not overlapping. In this method of masking, 20% of animals were masked. So, when 20% of animals were masked, we had 5 training sets and 5 test sets. Each training set included 400 animals and each test set included 100 animals. We also produced additional training sets by masking 50% of animals. This was done to test the effect of increasing the masked animals on accuracy of QTN prediction. When 50% of all animals were masked, there were only 250 individuals with genotypes in the training data set. Also, there were 250 individuals in each test set.

3.4. Statistical Model

G-BLUP statistical model was used to estimate the SNP effects. G-BLUP estimates the effects of the markers by best linear unbiased prediction, which assumes each marker explains an equal

3.6. Prediction

As it was mentioned above, of the 1040 loci, 25 loci that had MAF > 0.05 were randomly selected as QTN. The values for QTNs were considered as observation vector in the model. The selectively neutral SNPs in the sample, excluding the QTN, were taken to be surrounding markers (half of the SNPs were upstream and half were downstream of the QTN). The SNP effects (surrounding marker effects) were obtained from the training sets. Estimates of SNP effects were used to predict QTN effect in test sets. BLUP method was used for prediction (the model was explained in previous section).

3.7. Accuracy

Following evaluation from G-BLUP method, the accuracy of QTN predication was calculated as the correlation between true and predicted QTN effects.

3.8. Standard error

Standard errors of the cross validation predictions were calculated using the following formula:

$$\text{s.e} = \frac{SD}{\sqrt{n}}$$

Where SD is the standard deviation of cross validation predictions (accuracy predictions) and n is the number of cross validations that were performed. Standard error was used as a measure of error to observe how much the values of accuracies for each cross validation differ from each other in a standardized form.

3.9. Software used

Matlab (2008a) was used to analyze the data. It was used for genotyping the data, for splitting the data into training and test sets (cross validations), for filtering the data and also for analyzing the data by G-BLUP method.

4. Results

4.1. Effect of number of surrounding SNPs

Figure 1 shows accuracies of QTN prediction estimated with 10 to 100 surrounding SNPs in four analyses. Analyses include: analysis of the complete data set, when there was no selection for markers and when markers with MAF < 0.02, < 0.05 and < 0.10 were selected and removed from the analysis (In this study, the threshold 0.02, 0.05 and 0.1 for removing MAF are called cutoff thresholds for MAF). The accuracies were calculated based on average predictive accuracies of 25 replicates. Because there were 5 cross validations for each replicates, the mean was therefore an average of 125 values. Cross validations were performed to assess the performance of the model. The standard errors which were based on the variance between the replicate means can be found in tables in appendix 2.

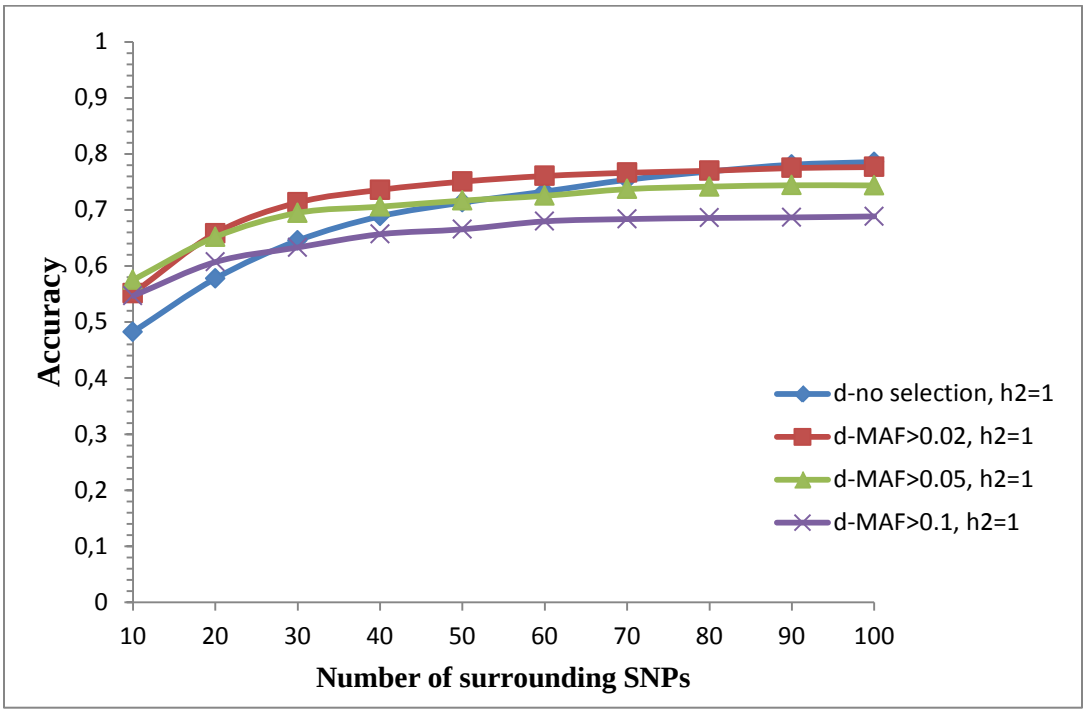

Figure1. Comparison of accuracies of QTN prediction estimated with 10 to 100 surrounding SNPs, when there was no selection for markers and when markers with MAF < 0.02, < 0.05 and < 0.10 were selected and removed. (heritability was 1)

According to Figure 1, in absence of any environmental effects (h^2=1), the accuracy of QTN prediction was affected by increasing the number of SNPs surrounding the QTN and also by removing the SNPs with different MAF from the analysis. As it can be observed, by increasing the number of flanking markers, the accuracy increased. Accuracies varied from 0.482 to 0.786 with different number of surrounding SNPs in all analyses. In all analyses, when there were 10 surrounding SNPs for prediction, the accuracy was lowest and when the number reached to 100, the highest accuracies were obtained. For example, when there was no selection for the markers, a low accuracy of 0.482 was realized for the model with l0 surrounding markers which increased to 0.786 for a model with 100 surrounding markers. When cutoff thresholds for MAF were 0.02, 0.05 and 0.10, the ranges of accuracies varied from 0.551 to 0.777, 0.575 to 0.744 and 0.547 to 0.689, respectively (see the tables in Appendix 2). When moving from 10 to 100 flanking SNPs, an increase of 41, 29 and 26% in accuracy was observed for these data sets with cutoff threshold 0.02, 0.05 and 0.1, respectively.

Except the curve corresponding to the accuracies in case of no selection for markers, the other curves were almost flat between 60 to 100 SNPs. It means that any increase beyond 60 surrounding SNPs was very small. Since the curves were flat, we expect hardly any improvement in accuracy, by increasing the SNP panel beyond 100 surrounding SNPs. For exception case (no selection), because the increase was very small after 80, it can be found that our expectation regarding to gradual decrease in accuracy by adding more than 100 surrounding SNPs is true.

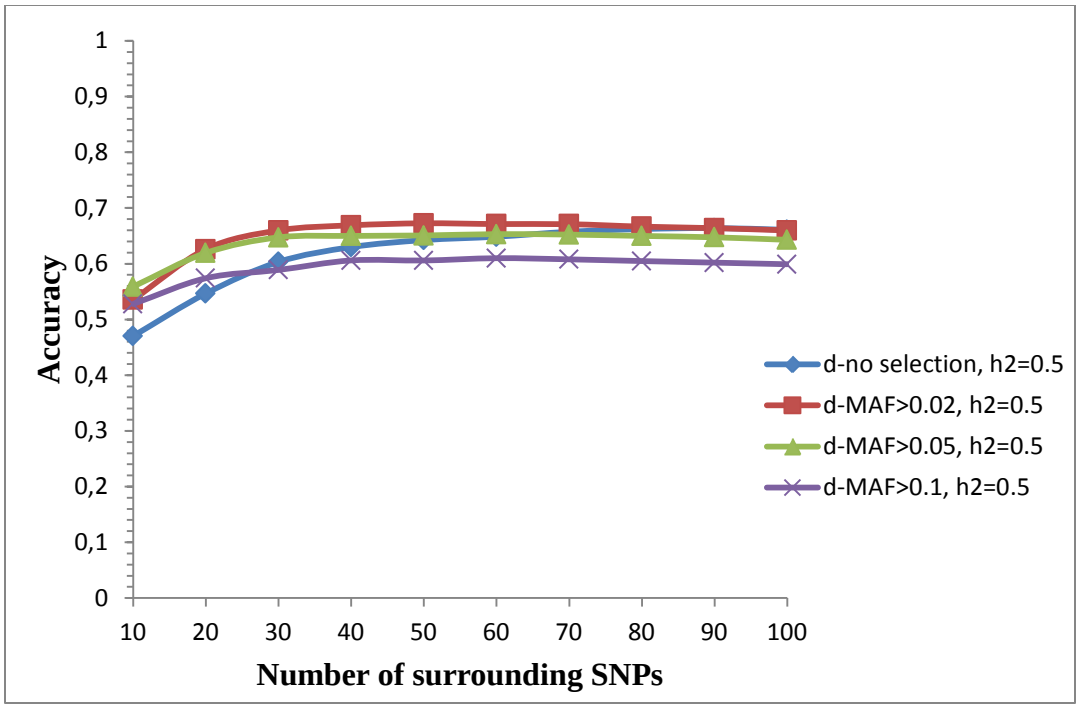

Figure 2. Comparison of accuracies of QTN prediction estimated with 10 to 100 surrounding SNPs, when there was no selection for markers and when markers with MAF < 0.02, < 0.05 and < 0.10 were selected and removed. (heritability was 0.5)

When heritability of phenotypes decreased to 0.5, in the situation of no selection for the markers, the accuracy increased from 0.470 to 0,664 by increasing the number of flanking markers from 10 to 90 and after 90 it had a small decrease. In situation of removing markers with MAF < 0.02, < 0.05 and < 0.10, by increasing the number of flanking markers from 10 to 60, the accuracy increased, for example from 0.536 to 0.67, 0.559 to 0.653 to 0.528 to 0.61, respectively. From 60 to 100 it showed a very small decrease. Here, by observing the gradual decrease in accuracy when moving from 60 to 100 surrounding SNPs for all curves, except the curve with no selection, we expect hardly an improvement in accuracy by adding more than 100 surrounding SNPs. For the exception case (no selection), again we expect that the accuracy decreases with increasing more than 100 flanking SNPs, because, even in this case, we observed a very small decrease in accuracy after 90 flanking SNPs.

Although with a heritability of 0.5 the trend of increasing accuracy with increasing number of SNPs did not change a lot, the accuracies decreased by lowering heritability.

When h^2=0.8, the accuracies of QTN prediction followed the pattern of accuracies when h^2=1. In general, with heritability 0.8, the accuracies increased with increased number of SNPs, but were lower compared to the accuracies for phenotypes with heritability 1. More explanation for result of heritability 0.8 is described in next sections. The figure and the table for the result of heritability 0.8 are in appendix 1 and 2.

4.2. Effect of MAF cutoff threshold

According to Figure 1 and 2, we understand that some markers with small or with no variation should be removed from analysis. Because the highest accuracies were obtained when markers with MAF < 0.02 were excluded. With 80 and less surrounding markers for QTN prediction, using the data set with MAF cutoff threshold 0.02 yielded higher accuracies than using the complete data set (no selection) (Figure 1). For example, with 10 flanking markers, an accuracy of 0.482 was realized when the complete data set was used compared to 0.551 when markers with MAF < 0.02 were selected and removed from analysis. It means a relative increase of 14%.

Knowing that some markers should be discarded from analysis, the most appropriate minimum threshold for SNP MAF in genomic selection studies should be determined. From the previous Figures (1 and 2), it seems the best cutoff threshold for MAF is 0.02. Because by shifting cutoff values from 0.02 to 0.05 and 0.1, more markers were removed and therefore more information of those markers were lost. Because the markers that have high average MAF are more informative, due to more variation. Therefore, lower accuracies were yielded by losing those markers. As an

example, when h^2=1 and QTN was predicted by 100 surrounding SNPs, the decrease in accuracy was 4.25% (accuracy 0.777 compared to 0.744) and 7.4% (accuracy 0.744 comapred to 0.689) when the cutoff threshold for MAF increased from 0.02 to 0.05 and from 0.05 to 0.1, respectively. The reduction in accuracies was to a large extent due to loss of markers that are polymorphic.

Here, the curve that shows the accuracies of QTN prediction when there was no selection for the markers was compared to the curve that presents the accuracies in the data set with MAF threshold cutoff 0.02. As it is observed, when there was no selection for markers, the accuracies of QTN prediction were lower with 10 to 80 surrounding SNPs, in comparison to the situation that markers with MAF < 0.02 were ignored from analysis. However, further inclusion of SNPs resulted in the accuracies that were almost the same for both situations and the curves were overlapped. As mentioned above, the reason for the lower accuracies might be due to markers that are monomorphic (MAF < 0.005) and the markers with MAF lower than 0.02. These markers give no information for increasing accuracy. For example, if there were 10 surrounding SNPs, when 5 of them had no information, the accuracies were estimated as if just 5 flanking SNPs surrounded the QTN. As it was proved, prediction of QTN with lower number of SNPs gives lower accuracy. One reason that the accuracies obtained from complete dataset, became identical to the accuracies obtained from dataset in which MAF cutoff threshold was 0.02, (when moving from 80 to 100 neighboring SNPs) could be due to the effect of previous markers that were accumulated and the accuracy reached to its maximum point by adding more genetic markers. So, by increasing number of markers, the number of markers becomes sufficient even if many markers with low information content (low MAF) were included.

By excluding the markers with MAF less than 0.02, the maximum accuracy that were obtained based on mean of accuracies of 25 replicates, was 0.777. Because most replicates yielded the accuracy that ranged from 0.7 to 0.8. In Figure 4, the histogram of the number of replicates that had accuracies in range of 0.6 to 0.7, 0.7 to 0.8, 0.8 to 0.9 and more than 0.9 was shown. As it is clear more than half of the replicates had the accuracy ranged from 0.7 to 0.8.

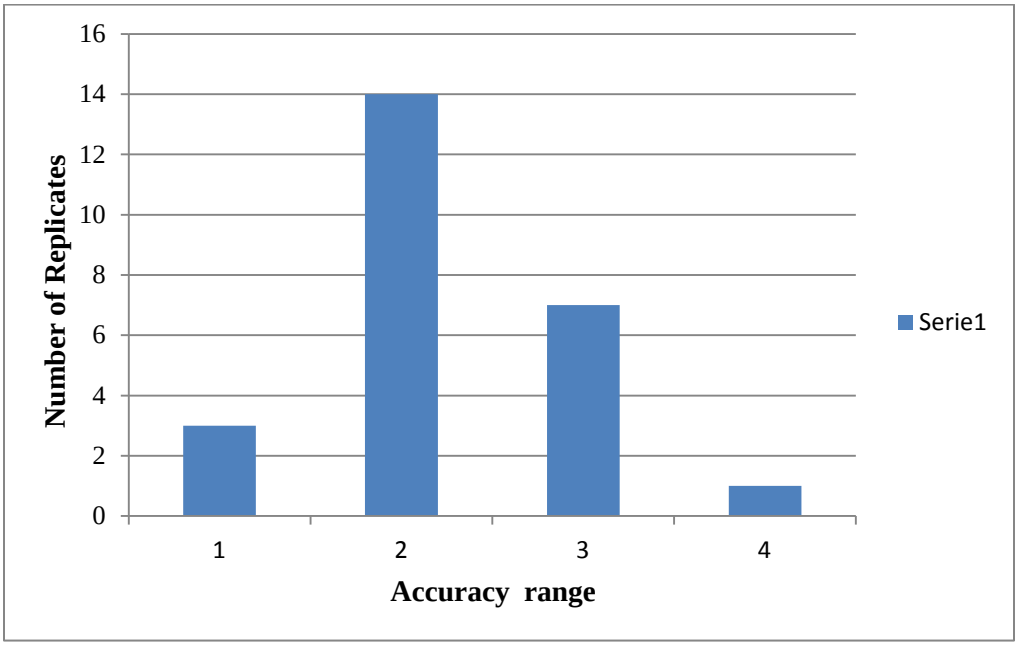

Figure 3: Number of replicates that belong to different ranges of accuracies (**1**: 0.6-0.7, **2**: 0.7-0.8, **3**: 0.8-0.9 and **4**: more than 0.9)

4.3. Effect of different levels of masking

Figure 4 and 5 presents overall average accuracy based on mean accuracies for 25 replicates when 20 and 50% of animals were masked, respectively. Figure 4 shows the comparison of 20 and 50% masking when heritability of phenotype was 1 and Figure 5 shows these comparisons when heritability of phenotype was 0.5. According to these Figures, accuracy for 20% masking

with 400 individuals in the training data set changed compared to 50% masking with 250 individuals in the training data set. Accuracies with 20% masking was higher in comparison to 50% masking. This can be due to higher amount of information in training set which results in better estimates of marker effects. In other words, when more phenotypes are available to estimate marker effects, the accuracy increases.

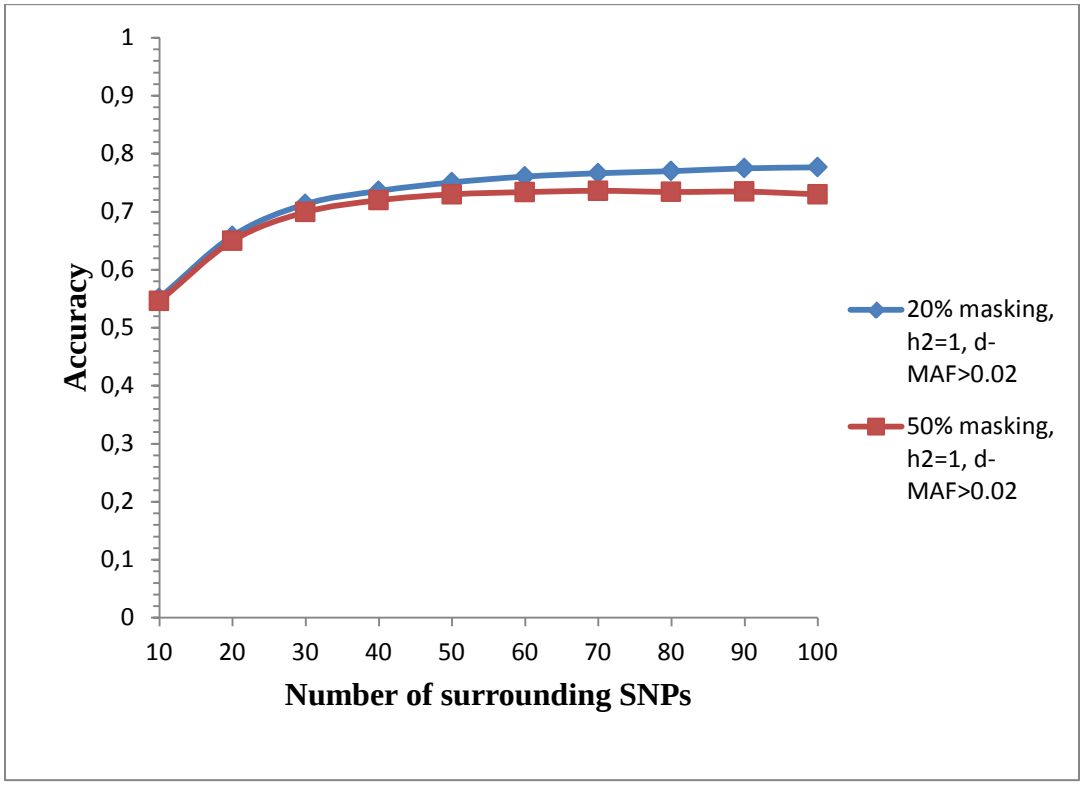

Figure 4. Accuracies of QTN prediction with masking 20% and 50% of all individuals when markers with MAF < 0.02 were excluded and heritability of phenotypes was 1.

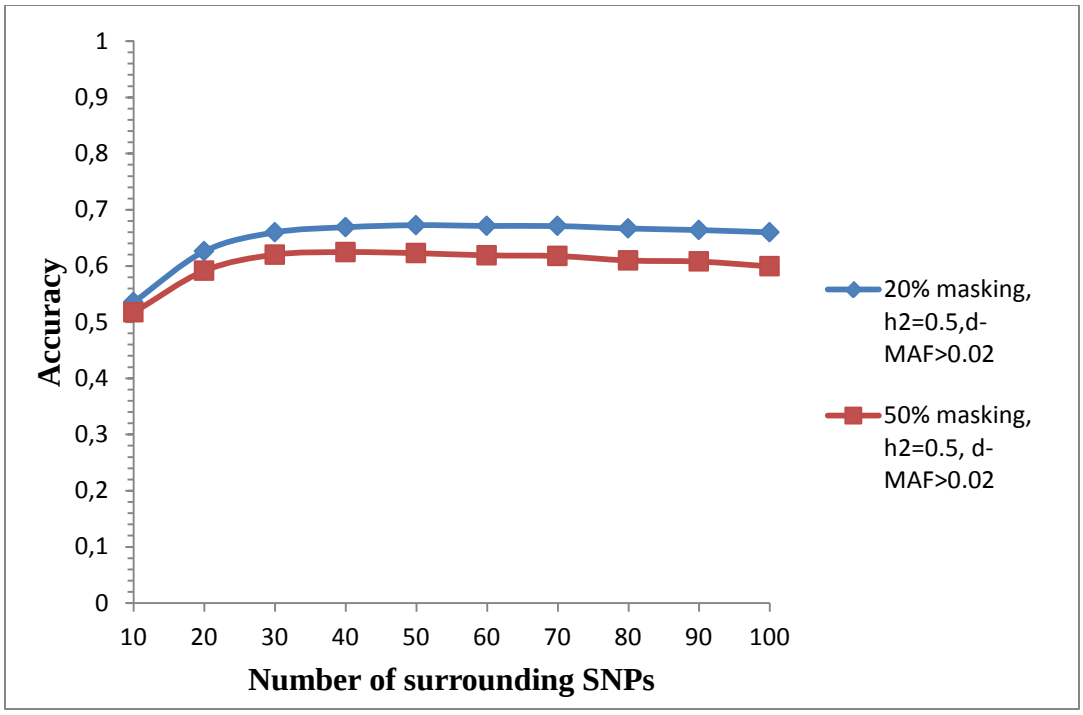

Figure 5. Accuracies of QTN prediction with masking 20% and 50% of all individuals when markers with MAF < 0.02 were excluded and heritability of phenotypes was 0.5.

When 50% of animals were masked, for heritabilities 1 and 0.5, the accuracies of prediction reached to about 0.73 and 0.62 with 50 flanking SNPs, respectively. Further increase in number of surrounding SNPs in the model showed very little change in the accuracies. In addition, for both heritabilities, when the number of surrounding SNPs increased, the differences in accuracies between 20 and 50% masking increased. When 10 SNPs surrounds the QTN, there was very little difference in accuracies (about 0.9%). However, when the number of flanking SNPs reached to 100, for heritabilities 1 and 0.5, the differences were 6% (0.777 to 0.730) and 9.25% (0.66 to 0.599), respectively. So, when heritability was lower, by reducing the size of training set (masking more individuals), the decline in accuracies was bigger (6% versus 9.25% for heritabilities 1 and 0.5, respectively).

Figure 4 also shows that the increase in accuracy of QTN prediction, when moving from 50 to 100 SNPs, was smaller for 50% masking than 20% masking.

For confirmation of the results of 20 and 50% masking, the same analyses were performed on a data set in which cutoff threshold for MAF was 0.05. The curves followed the same pattern. Figures and tables for this analysis can be found in appendix 1 and 2.

4.4. Effect of heritability

Figure 6 to 9 compares the accuracies of QTN prediction with different levels of heritability (1, 0.8 and 0.5) in different data sets. In all analyses, the accuracies decreased with a reduction in heritability, and the trend of decrease was almost the same for all analyses. The accuracies of QTN prediction for the heritability 1, 0.8 and 0.5 of phenotypes, using any of the data sets and any number of surrounding SNPs, ranged between 0.482 to 0.786, 0.478 to 0.741 and 0.47 to 0.67, respectively. In general, with 100 surrounding SNPs, from heritability 1 to 0.8, the decrease in accuracies was 6.5, 4.63, 4 and 5.2% for the analysis with no selection and for analyses in which cutoff thresholds for MAF were 0.02, 0.05 and 0.1, respectively. From heritability 0.8 to 0.5, the decrease in accuracies was 10, 11, 10 and 8.27% for the analysis with no selection and analyses in which SNPs with MAF < 0.02, < 0.05 and < 0.1 were removed, respectively.

The reduced accuracies for lower heritabilities were mainly as the result of larger environmental variance and lower proportion of estimated genetic variance that is explained by the markers.

Since the pattern of reduction in accuracies, by lowering heritability, is the same for all data sets, heritabilities are compared for the dataset in which cutoff threshold for MAF was 0.02. When

decreasing heritability from 1 to 0.8, the accuracies decreased from 0.551 to 0.548 and from 0.777 to 0.741 with 10 and 100 flanking SNPs, respectively. It means a reduction of 0.54 and 4.63%. When reducing heritability from 0.8 to 0.5, the accuracies decreased from 0.548 to 0.536 and from 0.741 to 0.660 with 10 and 100 flanking SNPs, respectively. It means a reduction of 2.2 and 11%. This shows that when heritability was reduced, the reduction in accuracy was larger when the number of surrounding SNPs was large. This might be due to the accumulation of environmental variance with adding more markers. For example, when just 10 neighboring SNPs were used for QTN prediction, the environmental variance is less than the situation that 100 surrounding SNPs were used for prediction.

neighboring SNPs[1]. From this it is concluded that given a dense marker map, QTN can be predicted with a high accuracy.

When h^2=0.8, by moving from 10 to 100 flanking SNPs, the increase in accuracy was 35% (accuracy 0.548 compared to 0.741). So, when h^2=0.8, it was possible to predict the QTN with a high accuracy of 0.741, by combining information from 60 SNPs or more neighboring SNPs. However, when h^2=0.5, by moving from 10 to 60 flanking SNPs, the increase in accuracy was about 25% (accuracy 0.536 compared to 0.67). So, when h^2=0.5, the maximum accuracy was 0.67 by combining information from 60 SNPs.

Heaton *et al.* (2007) developed a set of markers (100 SNPs) with an average MAF bigger than 0.41. These SNPs were from a group of 216 sires. They sequenced the region of about 1000 base pair flanking these SNPs to get additional polymorphisms. They calculated the accuracy of DNA test by using the information of these surrounding nucleotides and found an increase in accuracy of DNA tests by increasing the number of surrounding SNPs. They suggested this set of SNPs for multiple DNA diagnostic uses, for use by researchers, producers and commercial genotyping laboratories.

We applied different cutoff thresholds for MAF (0.02, 0.05 and 0.1) in order to see the effect of various cutoff thresholds for MAF on accuracy. Also, we compared the results of analyzing data sets with different MAF cutoff thresholds to the results of complete data set with no selection, in order to see if removing markers with low MAF has effect on accuracy or not. Our results showed that markers with very low frequencies (less than 0.02) had little impact on accuracy estimation and should be ignored from analysis. However, as it was shown in Figure 1, in case of no selection, with increasing the number of markers, from 10 to 100, the accuracy reached to its

[1] As we proposed cutoff threshold 0.02 for MAF, the results of data set with this threshold are discussed.

in Holstein was 38,416. In another analysis, they used 50% (19,208) or 25% (9,604) of those SNPs. They observed more reliability when more markers were used. For example for milk yield, the coefficient of determination increased 4.44% by increasing the number of markers from 9604 to 19208. In our study, with 100 surrounding SNPs, accuracy increased by around 4.5% when the markers with MAF between 0.02 and 0.05 were included in the analysis.

Wiggans *et al.* (2009) applied a minimum SNP MAF of 0.02 for genomic predictions. They reported that SNPs with a low MAF (less than 0.02) were expected to have a very small effect on genomic evaluation. They suggested this threshold, because the number of genotyped animals have been increased nowadays and has resulted in increase in usefulness of markers with MAF between 0.05 and 0.02. The studies that have been done on human genetics discarded MAF less than 0.05, they also suggested a larger sample size to identify the effects that are very small (Hirschhorn and Daly 2005). Wiggans *et al.* (2010) increased the number of SNPs that were used for genomic evaluations. They used a minimum SNP MAF of 0.01 for Holsteins, Jerseys, or Brown Swiss cows. They reported that it is possible to use SNPs with a low frequency to evaluate the accuracy, because the number of genotyped animals is increasing.

In genome-wide association studies (GWAS), it is common to remove the SNPs with MAF lower than 0.1 (Florez *et al.* 2007). There are some reasons for avoiding MAF less than 0.1. One reason is that markers with low frequencies reduces genotyping rates. Another reason is related to perceptions about the statistical inferences that result from analyzing these SNPs. It means that it is hard to draw conclusion from the results of analysis these SNPs. However, this threshold leads to losing large information of those SNPs (Tabangin *et al.* 2009). Also, ignoring SNPs with low MAF will decrease the ability for detection the polymorphisms causing rare diseases (Gorlov *et al.* 2008). Tabangin *et al.* (2009) investigated the effect of removing the SNPs with

low MAF on the likelihood of getting false positives. They showed that discarding those markers from analysis increased false-positive rates in GWAS.

We also investigated the effect of increasing masked animals in test set on accuracy when the heritability of phenotype was 1 and 0.5. It was shown that 50% masking compared to 20% masking, had a strong influence on the accuracy of QTN prediction. Increasing masked animals from 100 to 250 decreased the accuracies by 6% and 9.25% when heritability was 1 and 0.5, respectively. So, it is concluded that higher accuracy can be achieved by increasing the size of training set. One reason for higher accuracies with masking fewer animals can be explained regarding the size of training sets. When training sets are larger, they give more information for estimation of marker effects. So, with more information, accuracy improves. It should be noted that test sets do not contribute in the estimation of marker effects. Luan *et al.* (2009) used the random masking method on Norwegian red cattle milk data (milk yield, fat yield and protein yield) for calculating the accuracy of GEBV. They found that the accuracies of GEBV for 20% masking with 400 animals in training set was significantly higher in comparison to masking 50% of animals with 250 animals in training set. The accuracies that they obtained for milk yield with heritability around 0.3 were 0.599 and 0.457 for 20 and 50% masking, respectively. In our study, when h^2=0.5 and when 100 SNPs surrounds the QTN, with 20 and 50% masking, the accuracies were 0.660 and 0.599, respectively

We found that there was a relationship between the heritability and accuracy of QTN prediction. The accuracy reduced when the heritability decreased which corresponds to results by Kolbehdari *et al.* (2007). In general, when we estimated the accuracy by 100 surrounding SNPs and heritability decreased from 1 to 0.8 and from 0.8 to 0.5, the accuracies decreased 4.6 and 11%, respectively. Nielsen *et al.* (2009) investigated the effect of heritability on accuracy of

genomic breeding value in aquaculture breeding schemes and showed an increase of about 4% when heritability enhanced from 0.2 to 0.4.

We obtained the accuracy of 0.66 when there were 100 flanking SNPs and when h^2=0.5. Meuwissen et al. (2001) obtained accuracy of 0.579 for G-EBV. The heritability for their phenotype was 0.5 and the size of their data was 500. Luan *et al.* (2009) compared the accuracies of genomic prediction for high and low heritable traits in Norwegian Red cattle and found the accuracies of G-EBV prediction were higher for the traits with higher heritability.

6. Conclusion

It is concluded that with higher number of surrounding SNPs, accuracy of QTN prediction is higher. Since further increase in number of flanking markers beyond 60 SNPs showed a very small increase in accuracy, we do not propose using more than 100 surrounding SNPs for QTN prediction.

The markers that have no variation should be excluded from analysis, because they found not to be useful for QTN prediction. In all analyses, when markers with MAF less than 0.05 and 0.1 were avoided, there was a decreasing trend in the accuracies. According to these results, we showed that a cutoff value of 0.02 for MAF is more appropriate for genomic selection studies. Considering this MAF minimum threshold also provide a tool to avoid monomorphic SNP with MAF < 0.005. However, we showed that by using more flanking markers, even the SNPs with MAF lower than 0.02 can be useful for improvement the accuracy. Because with 80 to 100 surrounding markers the accuracies for the data set with no selection and for the data set with MAF cutoff threshold 0.02 were almost identical (Figure 1). After filtering the data with the determined cutoff threshold for MAF, the QTN could be predicted with 100 flanking SNPs with a high accuracy of 0.777.

The accuracies of prediction with 50% masking followed the trend of accuracies with 20% masking, for different number of surrounding SNPs. There was an increasing trend with increasing the number of flanking SNPs. However, with 50% masking, the accuracies were lower in comparison to 20% masking. So, it is suggested to reduce the size of test sets especially for phenotypes with lower heritability. It is concluded not to mask more than 20% of all animals in test sets. It means that at least 80% of animals should be analyzed in training set.

The accuracy of QTN prediction dropped when the heritabilities of phenotypes decreased. This happened due to larger environmental variance and smaller genetic variance for phenotypes with lower heritabilities.

7. Suggestions

- As this study was performed on simulated data, it is proposed to estimate the accuracy of QTN predictions for the QTNs that have been identified in farm animals so far.
- Because of the limited number of animals in this study, further studies using larger data sets are needed to investigate the minimum cutoff threshold for MAF in genomic selection studies.
- Also, it is suggested to investigate the accuracy of QTN prediction when many QTN are present
- As we used just BLUP method for our analysis, it is also proposed to use different models for QTN prediction and compare the results of these methods in order to understand which model gives highest accuracy and finally performs best

8. References

Blott, S., J.J. Kim, S. Moisio 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:** 253–66.

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. J. Anim. Breed. Genet. **124:** 362-368.

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.

Chan, E.K.F., R. Hawken and A. Reverter, 2008 The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Anim. Genet. **40:** 149–156.

Dekkers, J.C., and F.Hospital 2002, The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. **3:** 22-32.

Dekkers, J. C., 2004 Commercial application of marker and gene assisted selection in livestock: strategies and lessons. J. Anim Sci. **82 E-Suppl:** E313-E328.

De Roos, A. P. W., B. J. Hayes and M. E. Goddard, 2009 Reliability of genomic predictions across multiple populations. Genetics **183:** 1545–1553.

Falconer, D., and T. Mackay, 1996 Introduction to Quantitative Genetics. Longman, London.

Florez, J.C., A.K. Manning, J.Dupuis, J. McAteer, K. Irenze, L. Gianniny, D.B. Mirel,C.S. Fox, L.A. Cupples and B.J. Meigs, 2007 A 100 k genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets.Diabetes. **56:** 3063–3074.

Goddard, M.E., B.J. Hayes, 2002 Optimisation of response using molecular data. Proc. 7th World Congr. Appl. Livest. Prod. **33:** 3-10.

Goddard, M. E., and B.J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programs. Nat. Rev. Genet. **10:** 381-391.

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. Genet. Sel. Evol. **41:** 3.

Gorlov, I. P., O.Y.Gorlova, S.R. Sunyaev, M.R. Spitz and C.I. Amos, 2008 Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J. Hum. Genet. **82:** 100–112.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges and R. Snell, 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.

Habier, D., R. Fernando and J. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics **177:** 2389 - 2397.

Hayes, B. J., 2008 International Postgraduate Course and workshop: Whole Genome Association and Genomic Selection Salzburg, Austria BOKU University of Natural Resources and Applied Life Science.

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.

Heaton, M.P., W.M. Snelling, T. PL. Smith, J. W. Keele, G. P. Harhay, R. T. Wiedmann , G. L.Bennett, B.A.Freking, C.P.VanTassell, T.S.Sonstegard, L.C.Gasbarre, S.S.Moore, B.Murdoch, S.D.McKay, T. albfleisch and W.W.Laegreid, 2007 A Marker Set For Parentage-Based DNA Traceback In Beef And Dairy Cattle. Plant and animal genomes XV conference.

Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics **31:** 423–447.

Hill, W. G., and A. Robertson, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226-23.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. **6:** 95–108.

Hocquette, J. F., S. Lehnert, W. Barendse, I. Cassar-Malek and B. Picard, 2007 Recent advances in cattle functional genomics and their application to beef quality. animal **1:** 159-173.

Hoggart, C. J., M. Chadeau-Hyam, T. G. Clark, R. Lampariello, J. C. Whittaker, et al, 2007 Sequence-level population simulations over large genomic regions. Genetics **177:** 1725–1731.

Kimura, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

Kinghorn, B. P., J. A. M. van Arendonk and J. Hetzel, 1994 Detection and use of major genes in animal breeding. AgBiotech News and Information **6(12):** 297N-302N.

Kolbehdari, D., L. R. Schaeffer and J. A. B. Rabinson, 2007 Estimation of genome wide haplotype effects in half-sib designs. J. Anim. Breed. Genet. **124:** 356-361.

Liu, B. H., 1998 Statistical genomics: linkage, mapping, and QTL analysis, Boca Raton, Fla., CRC Press.

Liu, Z. J., and J. F.Cordes, 2004 DNA marker technologies and their applications in aquaculture genetics. Aquaculture **238:** 1-37.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen and T.H.E. Meuwissen, 2009 The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics **183:** 1119-1126.

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. J. Anim. Breed. Genet. **124:** 356-361.

Nielsen, H.M., A.K. Sonesson, H. Yazdi and T.H.E Meuwissen, 2009 Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes**.** Aquaculture **289:** 259-264.

Olsen H.G., S. Lien, M. Gautier, H. Nilsen, A. Roseth, P.R. Berg, K.K. Sundsaasen, M. Svendsen and T.H.E. Meuwissen, 2007 Genetic support for a quantitative trait nucleotide in the ABCG2 gene affecting milk composition of dairy cattle. BMC Genetics **8:** 32.

Romualdi, C., D. Balding, I. S. Nasidze, G. Risch, M. Robichaux, S. T. Sherry, M. Stoneking, M. A. Batzer and G. Barbujani, 2002 Patterns of Human Diversity, within and among Continents, Inferred from biallelic DNA Polymorphisms. Genome Res. **12:** 602 – 612.

Ron, M., and J.I. Weller, 2007 From QTL to QTN identification in livestock – winning by points rather than knock-out: a review. Anim. Genet. **38:** 429-439.

Saatchi, M., S. R. Miraei-Ashtiani, A. Nejati Javaremi, M. Moradi-Shahrebabak and H. Mehrabani-Yeghane, 2010 The impact of information quantity and strength of relationship between training set and validation set on accuracy of genomic estimated breeding values. African Journal of Biotechnology. **9:** 438-442.

Schaeffer, L.R., 2006 strategy for applying genome wide selection in dairy cattle. J. Anim. Breed. Genet. **118:** 218-223.

Schnabel R.D., J.J. Kim, M.S. Ashwell, T.S. Sonstegard, C.P. Van Tassell, E.E. Connor and J.F. Taylor, 2005 Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. Proceedings of the National Academy of Sciences of the United States of America **102:** 6896–901.

Schwerin, M., G. Brockmann, J. Vanselow and H. M. Seyfert, 1995 Perspectives of molecular genome analysis in livestock improvement. Arch. Tierz. Dummerstorf **38:** 21-31.

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:** 492–8.

ShiYi,X., H.YuanTao, R. ShaoQi, H. WeiJun, H. Bin, Labu, Pubuzhuoma, Gesangzhuogab and W. YiMing, 2009 Effects of cutoff thresholds for minor allele frequencies on HapMap resolution: A real dataset-based evaluation of the Chinese Han and Tibetan population. Chinese Sci Bull. **54:** 2069-2075.

Stein, G. S., J. L. Stain, A.J. van Wijnen and J. B. Lian, 1996 The maturation of a cell. American Scientist **84:** 28-37.

Tabangin, M.E., J.G. Woo and L.J. Martin, 2009 The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proceedings 2009, **3(Suppl 7):** S41.

VanRaden, P.M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor and F. S. Schenkel, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. **92:** 16–24.

Vignal, A., D. Milan, M. Sancristobal and A. Eggen, 2002 A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. **34:** 275-305.

Villumsen, T.M., L.Janss and M.S. Lun, 2009 The importance of haplotype length and heritability using genomic selection in dairy cattle. J. Anim. Breed. Genet. **126:** 3-13.

Weller, J.I., and M. Ron, 2011 Invited review: Quantitative trait nucleotide determination in the era of genomic selection. J. Dairy Sci. **94:** 1082-1090.

Wiggans, G.R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel and C. P. Van Tassell, 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.

Wiggans, G.R., P. M. VanRaden, L. R. Bacheller, M. E. Tooker, J. L. Hutchison, T. A. Cooper and T. S. Sonstegard, 2010 Selection and management of DNA markers for use in genomic evaluation. J. Dairy Sci. **93:** 2287–2292.

Winter, A., W. Kramer, F.A.O. Werner, S. Kollers, S. Kata, G. Durstewitz, J. Buitkamp, J.E. Womack, G. Thaller and R. Fries, 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. Proc. Natl. Acad. Sci. USA **99:** 9300–9305.

Youngerman, S. M., A. M. Saxton and G. M. Pighetti, 2004 Novel single nucleotide polymorphisms and haplotypes within the bovine CXCR2 gene. Immunogenetics **56:** 355–359.

http://www.experts123.com/q/what-is-the-frequency-of-a-certain-allele-in-the-following-populations.html

Appendix 1

Figures:

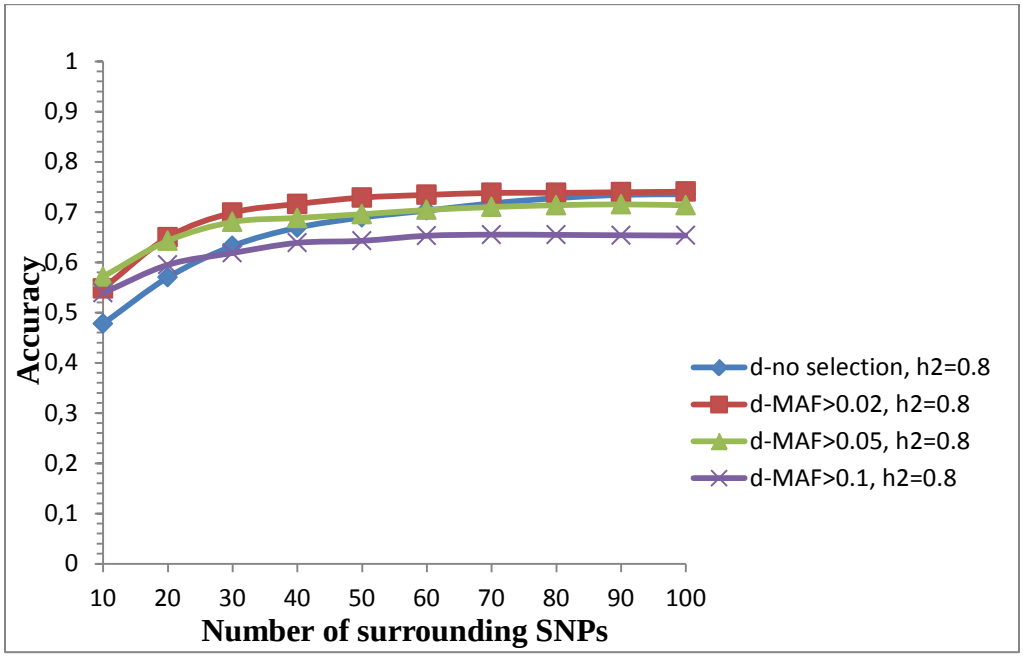

Figure10. Comparison of accuracies of QTN prediction estimated with 10 to 100 surrounding SNPs, when there was no selection for markers and when MAF < 0.02, < 0.05 and < 0.10 were selected and removed. (heritability was 0.8)

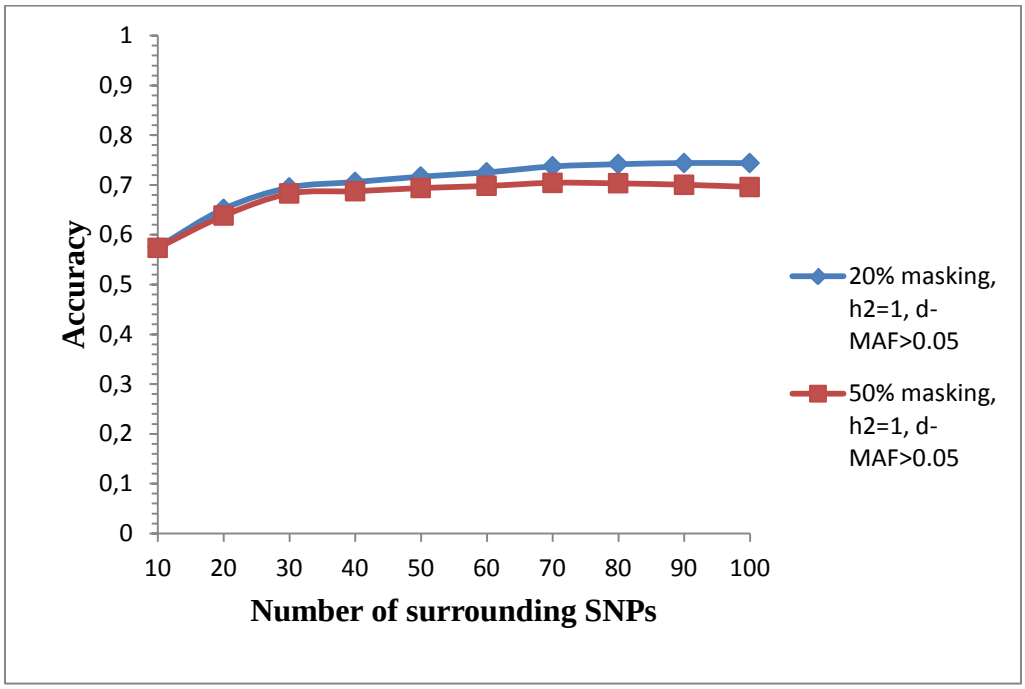

Figure 11. Accuracies of QTN prediction with masking 20% and 50% of all individuals when markers with MAF < 0.05 were excluded and heritability of phenotypes was 1.

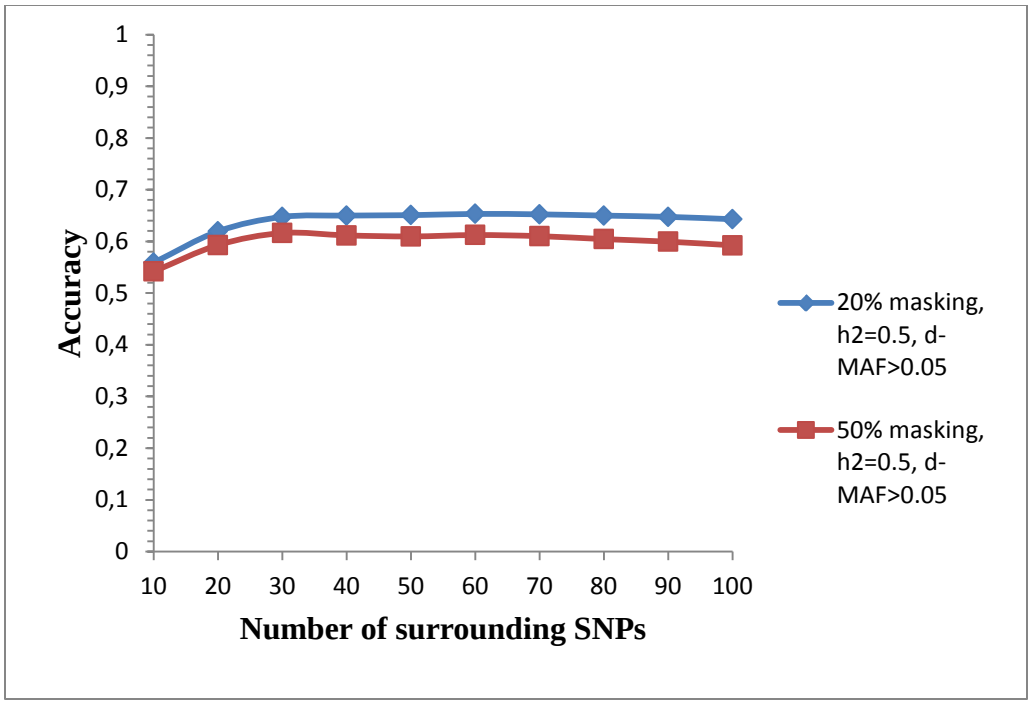

Figure 12. Accuracies of QTN prediction with masking 20% and 50% of all individuals when markers with MAF < 0.05 were excluded and heritability of phenotypes was 0.5.

Appendix 2

Tables:

Table 1: Comparison of different cutoff thresholds for MAF, when heritability of phenotype was 1.

Number of surrounding SNPs	Cutoff thresholds for MAF							
	No selection		0.02		0.05		0.1	
	Acc	SE	Acc	SE	Acc	SE	Acc	SE
10	0.482	0.033	0.551	0.030	0.575	0.028	0.547	0.033
20	0.578	0.027	0.658	0.026	0.652	0.028	0.607	0.029
30	0.645	0.027	0.713	0.024	0.695	0.022	0.633	0.025
40	0.689	0.025	0.736	0.022	0.706	0.020	0.657	0.025
50	0.713	0.025	0.751	0.020	0.717	0.019	0.666	0.024
60	0.733	0.023	0.761	0.019	0.725	0.019	0.680	0.022
70	0.754	0.021	0.766	0.019	0.737	0.019	0.684	0.021
80	0.769	0.020	0.770	0.019	0.742	0.018	0.686	0.022
90	0.781	0.019	0.775	0.018	0.744	0.018	0.687	0.023
100	0.786	0.019	0.777	0.017	0.744	0.020	0.689	0.024

Acc: Accuracy, SE: Standard error

Table 2: Comparison of different cutoff thresholds for MAF, when heritability of phenotype was 0.5.

Number of surrounding SNPs	Cutoff thresholds for MAF							
	No selection		0.02		0.05		0.1	
	Acc	SE	Acc	SE	Acc	SE	Acc	SE
10	0.470	0.032	0.536	0.030	0.559	0.028	0.528	0.032
20	0.547	0.027	0.626	0.027	0.619	0.029	0.574	0.030
30	0.603	0.028	0.660	0.027	0.647	0.025	0.589	0.030
40	0.630	0.027	0.669	0.025	0.650	0.025	0.606	0.029
50	0.642	0.027	0.673	0.025	0.651	0.027	0.606	0.028
60	0.649	0.027	0.671	0.025	0.653	0.028	0.610	0.030
70	0.657	0.026	0.671	0.026	0.652	0.028	0.608	0.031
80	0.662	0.026	0.667	0.026	0.650	0.027	0.605	0.030
90	0.664	0.025	0.664	0.024	0.647	0.027	0.602	0.032
100	0.661	0.026	0.660	0.024	0.643	0.028	0.599	0.032

Acc: Accuracy, SE: Standard error

Table 3: Comparison of different cutoff thresholds for MAF, when heritability of phenotype was 0.8.

Number of surrounding SNPs	Cutoff thresholds for MAF							
	No selection		0.02		0.05		0.1	
	Acc	SE	Acc	SE	Acc	SE	Acc	SE
10	0.478	0.032	0.548	0.029	0.571	0.028	0.539	0.034
20	0.570	0.027	0.650	0.027	0.643	0.029	0.595	0.030
30	0.632	0.028	0.699	0.024	0.680	0.024	0.619	0.026
40	0.669	0.026	0.716	0.021	0.688	0.022	0.639	0.027
50	0.689	0.025	0.729	0.020	0.696	0.022	0.643	0.026
60	0.703	0.024	0.734	0.020	0.704	0.022	0.653	0.025
70	0.717	0.022	0.738	0.020	0.710	0.021	0.655	0.024
80	0.728	0.021	0.738	0.020	0.714	0.020	0.655	0.025
90	0.734	0.021	0.740	0.020	0.715	0.021	0.654	0.025
100	0.735	0.020	0.741	0.019	0.714	0.022	0.653	0.026

Acc: Accuracy, SE: Standard error

Table 4: Accuracies of QTN prediction with 20% and 50% masking of all individuals, when heritability was 1 and cutoff threshold for MAF was 0.02

Number of surrounding SNPs	% of all animals masked			
	20		50	
	Acc	SE	Acc	SE
10	0.551	0.030	0.546	0.026
20	0.658	0.026	0.650	0.020
30	0.713	0.024	0.699	0.014
40	0.736	0.022	0.720	0.015
50	0.751	0.020	0.730	0.015
60	0.761	0.019	0.734	0.016
70	0.766	0.019	0.736	0.016
80	0.770	0.019	0.734	0.014
90	0.775	0.018	0.735	0.016
100	0.777	0.017	0.730	0.012

Acc: Accuracy, SE: Standard error

Table 5: Accuracies of QTN prediction with 20% and 50% masking of all individuals, when heritability was 0.5 and cutoff threshold for MAF was 0.02

Number of surrounding SNPs	% of all animals masked			
	20		50	
	Acc	SE	Acc	SE
10	0.536	0.030	0.517	0.021
20	0.626	0.027	0.591	0.018
30	0.660	0.027	0.620	0.018
40	0.669	0.025	0.625	0.019
50	0.673	0.025	0.623	0.017
60	0.671	0.025	0.619	0.016
70	0.671	0.026	0.618	0.015
80	0.667	0.026	0.610	0.015
90	0.664	0.024	0.608	0.016
100	0.660	0.024	0.599	0.015

Acc: Accuracy, SE: Standard error

Table 6: Accuracies of QTN prediction with 20% and 50% masking of all individuals, when heritability was 1 and cutoff threshold for MAF was 0.05

Number of surrounding SNPs	% of all animals masked			
	20		50	
	Acc	SE	Acc	SE
10	0.575	0.028	0.574	0.023
20	0.652	0.028	0.638	0.026
30	0.695	0.022	0.683	0.019
40	0.706	0.020	0.687	0.016
50	0.717	0.019	0.694	0.015
60	0.725	0.019	0.698	0.015
70	0.737	0.019	0.704	0.014
80	0.742	0.018	0.703	0.016
90	0.744	0.018	0.700	0.017
100	0.744	0.020	0.696	0.016

Acc: Accuracy, SE: Standard error

Table 7: Accuracies of QTN prediction with 20% and 50% masking of all individuals, when heritability was 0.5 and cutoff threshold for MAF was 0.05

Number of surrounding SNPs	% of all animals masked			
	20		50	
	Acc	SE	Acc	SE
10	0.559	0.028	0.542	0.014
20	0.619	0.029	0.593	0.020
30	0.647	0.025	0.616	0.012
40	0.650	0.025	0.612	0.013
50	0.651	0.027	0.610	0.012
60	0.653	0.028	0.612	0.010
70	0.652	0.028	0.610	0.011
80	0.650	0.027	0.604	0.012
90	0.647	0.027	0.599	0.011
100	0.643	0.028	0.592	0.012

Acc: Accuracy, SE: Standard error