

University of Nevada, Reno

**Selective genotyping using genome wide association studies for mapping loci
associated to fiber diameter in Merino sheep**

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in

Animal Science

by

Mohamed Goher

Dr. Luis Gomez-Raya/Thesis Advisor

May, 2010



University of Nevada, Reno
Statewide • Worldwide

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

MOHAMED NAGIB EL HELALY GOHER

entitled

**Selective Genotyping Using Genome Wide Association Studies For Mapping Loci
Associated To Fiber Diameter In Merino Sheep**

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

Luis Gomez-Raya, Ph.D, Advisor

Wendy M. Rauw, Ph.D., Committee Member

Matt Forister, Ph.D, Graduate School Representative

Marsha H. Read, Ph. D., Associate Dean, Graduate School

May, 2010

Abstract

The objective of this study was to investigate methods and statistical power for mapping quantitative trait loci (QTL) using selective genotyping and Illumina's 50K BeadChip. A large unrelated population is recorded for a phenotypic trait. Animals with extreme phenotypes are used for genotyping with Illumina's arrays and QTL are mapped by linkage disequilibrium (LD). We carried out computer simulations to compute statistical power using this approach after varying QTL allele frequency, proportion selected in the extremes, and population size. For example, power for a population of 1,000 animals after genotyping the top and bottom 5% (QTL effect of 0.5 phenotypic standard deviations, alpha of 0.01 and allele frequency of 0.1) and assuming maximum LD, was 0.95. The method was tested in a Merino flock with 979 ewes in which fiber diameter (FD) was recorded. Illumina's 50K Bead Chip was used for simultaneous genotyping of 54,241 SNPs in selected animals (24 in each extreme) but only within a year class and breed. A total of 208 tests were significant out of 18,214 SNPs tested at significance level of 0.01. There were more significant tests than expected by chance. The highest significant results were obtained in Chromosomes 1, 14, 15, 21 and 26. Validation of the results must be confirmed in the unselected population.

Acknowledgements

I am deeply indebted to my advisor Dr. Luis Gomez-Raya, for allowing me this opportunity. I am grateful to him for providing me with education in data analysis, power analyses, and computer simulations. I would like to thank Dr. Wendy M. Rauw, and Dr. Matt Forister for serving on my committee. Thank you, Veronica Kirchoff for teaching me most of the lab skills I have acquired and for all extractions of DNA. I also thank to the employees of the Nevada Genomic Center; Dr. Cindy Tungate, Dr. Craig Osborne, and Dr. Kris Kruse for all your advice and all the hard work you did for my project. Special thanks to the employees of Core Laboratories University of Colorado Denver; Dr. Bifeng Gao, Dr. Okyong Cho, and Dr. Ted Shade in helping in SNP's array genotyping.

Special thanks to Bernard Wone for helping and supporting me in running the ASReml program and spent countless hours assisting me. Also I would like to express my thanks to the faculty, staff and graduate students of Animal Science Department for their friendship during our staying in Reno Nevada.

Finally, I would like to give my special thanks to my wife Reham Allam, my daughter Khadija , my son Mansour, my lovely mother Noha Genedy and my greatest father Nagib El Helaly Goher for their patient and love that enabled me to complete this work.

Table of Contents

List of Tables	iv
List of Figures	v
Introduction.....	1
Materials and Methods.....	4
1. Animals and Phenotypic Recording.....	4
2. Statistical Power using Extreme Phenotypes and SNPs	5
3. DNA extraction using QIAGEN DNeasy tissue kit.....	7
4. Illumina bead chip - SNP's array.....	8
5. Data Analysis	10
I- Genome Studio Software for SNP's analysis	10
The input data files of this software were:.....	11
1- SNP table	12
2- Sample table.....	12
4- Full data table.....	15
5- Error table	15
6. Statistical analyses	16
Results.....	17
Statistical Power	17
Discussion.....	27
References.....	29
Appendix:.....	33
i. FORTRAN computer program code to determine the Statistical Power using Chi square test.....	33
ii. FORTRAN computer program code to read the breeding values.....	41
iii. FORTRAN computer program code to determine the analysis of variances (ANOVA) within each SNP for all 96 animals.....	43

List of Tables

Table 1: Statistical power for varying gene effects and population sizes corresponding to year classes existing in a Merino sheep flock. The significance level was $\alpha = 0.01$, SNP allele frequencies were 0.1 or 0.5. Animals were selected among the top and bottom 5%.....	18
Table 2: Statistical power for varying gene effects and population sizes corresponding to year classes existing in a Merino sheep flock. The significance level was $\alpha = 0.05$, SNP allele frequencies were 0.1 or 0.5. Animals were selected among the top and bottom 5%.....	19
Table 3: The most significant SNPs, their position due to Illumina® genotyping, Chromosome number (Chr) and Sequence.....	22
Table 4: The most significant SNPs, their position due to Illumina® genotyping, Chromosome number (Chr) and Sequence.....	23

List of Figures

Figure 1: Distribution of fleece fiber diameter in flock of Merino sheep.	5
Figure 2 : This logical model representation the Illumina Infinium HD Assay Ultra form the manual workflow for use with the 24x1 HD BeadChip. These protocols describe the procedure for preparing 96 DNA samples using 24 HD BeadChip. (Illumina,2008).....	9
Figure 3 : This diagram illustrate the flow of different files throughout the Genome Studio Software.	10
Figure 4: This graph represent the θ and R for the samples in both Cartesian and Polar coordinate graph for Genome Studio Software.	13
Figure 5: The Centroid and the called genotyping. The yellow points are the highest gen calls close to one. (GenomeStudio™ 2008.1 Framework).....	14
Figure 6: Power for varying gene effects (0.1 to 0.5 phenotypic standard deviations) at $\alpha = 0.01$ with selection of 5% in the extremes. Allele frequency of the SNP was either 0.1 or 0.5.	17
Figure 7: Power for varying gene effects (0.1 to 0.5 phenotypic standard deviations) at $\alpha = 0.05$ with selection of 5% in the extremes. Allele frequency of the SNP was either 0.1 or 0.5.	18
Figure 8: This graph demonstrate the effect of different years of Birth (YOB) on fiber diameter in all breed types in the whole population.	20
Figure 9: This graph demonstrate the effect of different breed types on fiber diameter in all years of Birth (YOB) in the whole population.	20
Figure 10: Distribution of F-values after selection of extremes in 18,214 SNPs at 2 and 45 degrees of freedom.	21
Figure 11 : This graph shows the position and F-value of the all SNPs for chromosome 1 to 9 in Merino sheep.....	24
Figure 12: This graph shows the position and F-value of all SNPs for chromosome 10 to 18 in Merino sheep.....	25
Figure 13: This graph shows the position and F-value of all SNPs for chromosome 19 to 26 and the X chromosome in Merino sheep.	26

Introduction

Efforts to map Quantitative Trait Loci (QTL) affecting production traits in farm animals were initiated in 1992 (Andersson et al.) and 1995 (Georges et al.). The first methodologies used to map QTL were consisting in the use of linkage methods. These methods are based on making use of linkage disequilibrium either generated after crossing two inbred lines (Andersson et al., 1992) or existing within outbred families (Georges et al., 1995). In the latter case, DNA marker alleles in the progeny are tested for association with phenotypic performance. Linkage analyses methods require low number of DNA-markers compared to more general association methods that utilize linkage disequilibrium present at the population level. Genome wide association methods do not require a family structure but require very heavily dense maps of DNA markers to allow for detection of associations. Just recently, vast amounts of DNA-markers are being generated in farm animals. Single Nucleotide Polymorphisms (SNPs) are single point mutations that are suitable for large scale genotyping using microarrays. Genome Wide Association Studies (GWAS) consist in the use of a large number of SNPs (microarrays) to detect associations to loci affecting performance or diseases. A large number of SNPs generated for ovine, bovine or porcine can be typed simultaneously in the same animal in a very short period of time, making this technology a powerful tool in genomic studies. Linkage disequilibrium maintained at the population level for loci at short physical distances, together with the block structure of the genome, allow for detecting those associations without requiring a family structure.

Selective genotyping in linkage analyses was first described by Lebowitz et al. (1987). Detection of QTL is facilitated by scoring large families for a quantitative trait and then genotyping only individuals with extreme phenotypes. In this way, only extreme phenotypes containing much information are used for QTL detection, which may result in increased statistical power for a given genotyping effort. For crosses between inbred lines, Darvasi and Soller (1992) showed that selective genotyping can result in a considerable reduction of the number of genotypes for a given power. To our knowledge, selective genotyping has not been used together with GWAS. However, the same benefits in terms of increased statistical power for a given number of genotyping would apply for GWAS. Using the extreme genotypes may reduce genotyping costs, which are rather large (Illumina's SNP array costs over \$250 per individual), by lowering the number of individuals being typed.

Fiber diameter is an important character determining wool quality in sheep (Bray, 1955; Von Bergen, 1963; Lang, 1964; Whan, 1970; Hunter and Gee, 1980). Fiber diameter is measured in microns (1/25,400 of an inch) by an instrument called the Optical Fiber Diameter Analyzer (OFDA, Baxter et al., 1991). Wool finer than 25 microns is used for garments, while coarser grades are used for carpets or rugs (2010 Australian Wool Exchange Ltd). About 90% of the wool fiber in sheep is made up of keratin intermediate filament (IF) and keratin-associated proteins (KAP). The keratin IF proteins from 8- to 10-nm diameter filaments that are embedded in a matrix of KAPs. Type I keratin IF and type II keratin IF are paired forming the basic unit of the filament. Wool keratin IF type I genes are 4–5 kb with 6 introns, whereas the type II genes are 7–9 kb with 8 introns (Powell, 1997).

In spite of its economical importance, little is known of the genetic architecture of fiber diameter in sheep. Jenkins et al. (1998) carried out a cross of Merino x Romney aimed at mapping loci associated to wool traits. However, they were unable to detect QTLs for fiber diameter in this cross. On the contrary, linkage between the gene coding for KAP6 and wool fiber diameter has been reported (Parsons et al, 1994). More recently, Roldan et al., (2010) used a half-sib family structure in Merino sheep to map QTL associated to wool traits but they were not able to detect loci with genome-wise associations to fiber diameter.

QTL detection using crosses between inbred lines in sheep is time consuming due to the long generation interval, and expensive because F2 animals are not likely productive. GWAS and selection of extremes could be carried out just after recording the phenotypes and might be a cost efficient approach for mapping since only a few animals are genotyped.

A first objective of this study is to compute statistical power for the use of GWAS with extreme phenotypes for mapping QTL in farm populations. A second objective is to make use of genome-wide association studies to identify loci affecting fiber diameter in Merino sheep, which is the primary criterion for determining trading price (value), processing performance, and end use of wool from sheep (Stobart et al, 1986).

Materials and Methods

1. Animals and Phenotypic Recording

Protocols for handling of animals and recording of traits were approved by the institutional Animal Care and Use Committee of the University of Nevada–Reno.

A flock of 979 ewes which were housed in Yerington, Nevada at the Rafter 7 Ranch, was used for this study. The ewes were crossbreeds of Marino and Rambouillet: 485 animals were $\frac{1}{2}$ Marino, 177 animals were $\frac{7}{8}$ Marino and 317 animals were full blood Marino. Ear notches were sampled from all the 979 animals and kept in the freezer at -17°C for DNA extraction. Phenotypic performances were recorded in January 2005 for each animal. Records included breed, year of birth and day of birth, as well as pedigree information. Wool samples were collected for each individual on January 2 and right after returning from the rangelands on March 18, 2005. Samples were analyzed with the OFDA2000 instrument (Baxter, 2001) for mean fiber diameter (FD), coefficient of variation of fiber diameter (CVFD), and staple length (SL). Figure 1 shows the distribution of fiber diameter in the herd. Ewes were shorn on March 22 and 23, 2005. Wool was weighed individually (greasy fleece weight; GFW); wool on head, legs, belly, and tail was not included. DNA samples of 48 ewes with extreme phenotypes were used for SNP microarray typing.

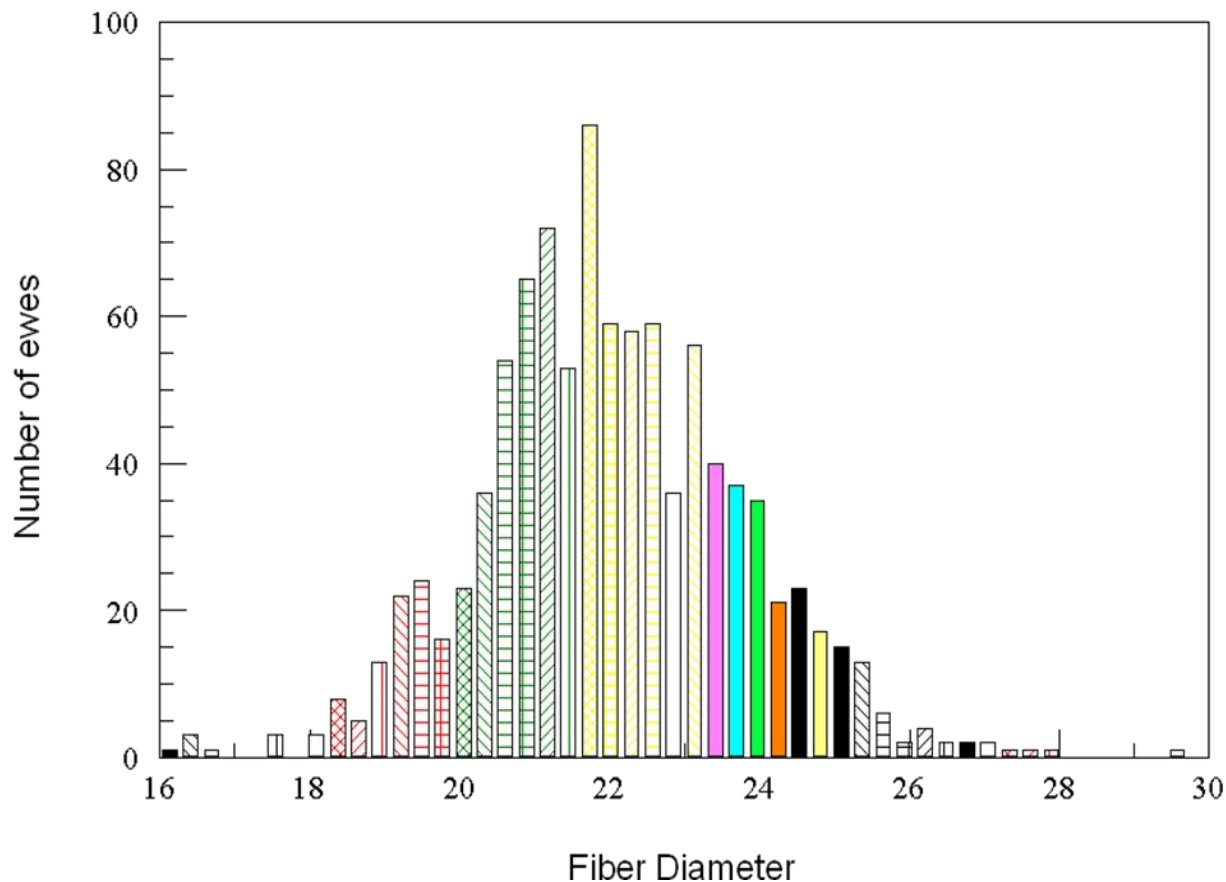


Figure 1: Distribution of fleece fiber diameter in flock of Merino sheep.

2. Statistical Power using Extreme Phenotypes and SNPs

Statistical power is the probability of getting a statistically significant result given that there is a biologically real effect in the population being studied. If a particular test is not statistically significant, it is because there is either no effect or the experimental design makes unlikely that a biologically real effect would be detected with that experiment. It might be due to low number of data and/or poor allocation of the animals to the experimental treatments. Power analysis can distinguish between these alternatives, and

is therefore a critical component of designing experiments and testing results (Toft, 1983; Peterman, 1990; Fairweather, 1991; Taylor, 1993; Thomas, 1996).

A Monte Carlo simulation was carried out to compute power when using extreme phenotypes and SNPs. The phenotype of the i -th individual was generated by

$$P_i = \mu + EN_i + G_i,$$

where $EN_i = y \times (VE)^{1/2}$, $G_i = x \times (VA)^{1/2}$, x and y are deviates drawn from the standard normal distribution, μ and VA are the population mean and genetic variance for fiber diameter, and VE is the environmental variation.

The population generated in silico was sorted according to the phenotype, P_i . The allele frequency in the top and the bottom ranking animals was estimated for each replicate. The allele frequency in both extremes was compared using a χ^2 test. Under the null hypothesis, the allele frequency of the SNP is the same. Under the alternative hypothesis, the allele frequency is different in the two extremes. The percentage of replicates with χ^2 test values larger than the threshold at the significance level, α , was our estimate of the power of the test. Statistical power was computed for a heritability of 0.5 and alternative samples sizes for an SNP with an effect of 0.1, 0.2, 0.3, 0.4, and 0.5 phenotypic standard deviations. The simulated allele frequency was either 0.1 or 0.5. The significance level (α) of the test was 0.01 and 0.05. Each simulation set was replicated 10,000 times.

3. DNA extraction using QIAGEN DNeasy tissue kit:

Genomic extractions were performed using the DNeasy tissue kit according to the manufacturer's instructions (Qiagen, Chatsworth, CA). Each tissue sample was cut to approximately 25 mg, and placed in a 1.5 ml microcentrifuge tube. A 180 μ l ATL buffer and 20 μ l of proteinase K were added and the sample was incubated overnight at 55°C in a shaking water bath for the tissue to be completely lysed. After that, 200 μ l of buffer AL was added to the sample and mixed thoroughly by vortexing and incubated at 70°C for 10 minutes. After that, 200 μ l ethanol (100%) was added to the samples and mixed by vortexing. The mixture was pipetted into DNeasy Mini spin column, placed in a 2 ml collection tube, and was centrifuged at 6000 x g (8000rpm) for 1 minute. Both the flow-through and the collection tube were discarded. The DNeasy Mini spin column was placed in a new 2 ml collection tube where 500 μ l buffer AW1 was added. This was centrifuged at 6000 x g (8000rpm) for 1 minute, after which both the flow-through and the collection tube were discarded. New 2 ml collection tubes were added to the DNeasy Mini spin column where 500 μ l buffer AW2 was added and then centrifuged at 20000 x g (14000rpm) for 3 minutes. Again both the flow-through and the collection tube were discarded. The DNeasy mini spin column was placed in a clean 1.5 ml microcentrifuge tube and 200 μ l of buffer AE was added directly onto the DNeasy membrane and incubated for 1 minute at room temperature. It was then centrifuged at 6000 x g (8000rpm) for 1 minute for elution. Stranded DNA was quantified using a fluorescent nucleic acid stain

(PicoGreen®) at Nevada Genomic center and read on a Nano-drop spectrophotometer reader. Extracted genomic DNA was stored at -20° C until SNP's were analyzed.

4. Illumina bead chip - SNP's array:

The Illumina® Infinium® HD Assay Ultra protocol was used for DNA analysis of 54241 SNPs. These procedures were carrying out through the Core Lab of University Colorado, Denver as follows in the logical model:

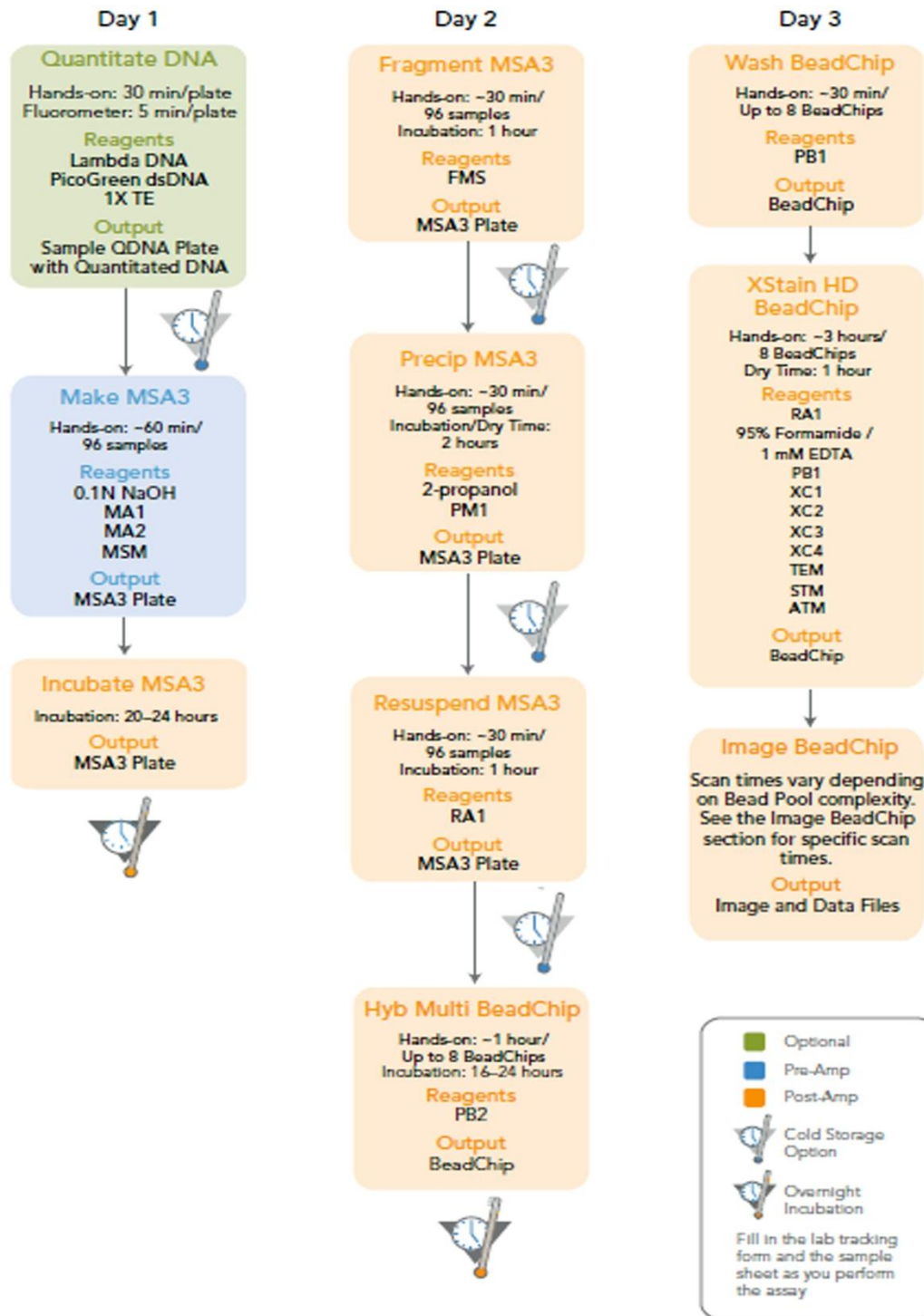


Figure 2 : This logical model representation the Illumina Infinium HD Assay Ultra form the manual workflow for use with the 24x1 HD BeadChip. These protocols describe the procedure for preparing 96 DNA samples using 24 HD BeadChip. (Illumina,2008)

5. Data Analysis.

I- Genome Studio Software for SNP's analysis:

This program was used to visualize and inspect the data generated by all of Illumina's platforms from the SNP's array. The resulting report file was used for the analysis of variance to detect the significant SNPs.

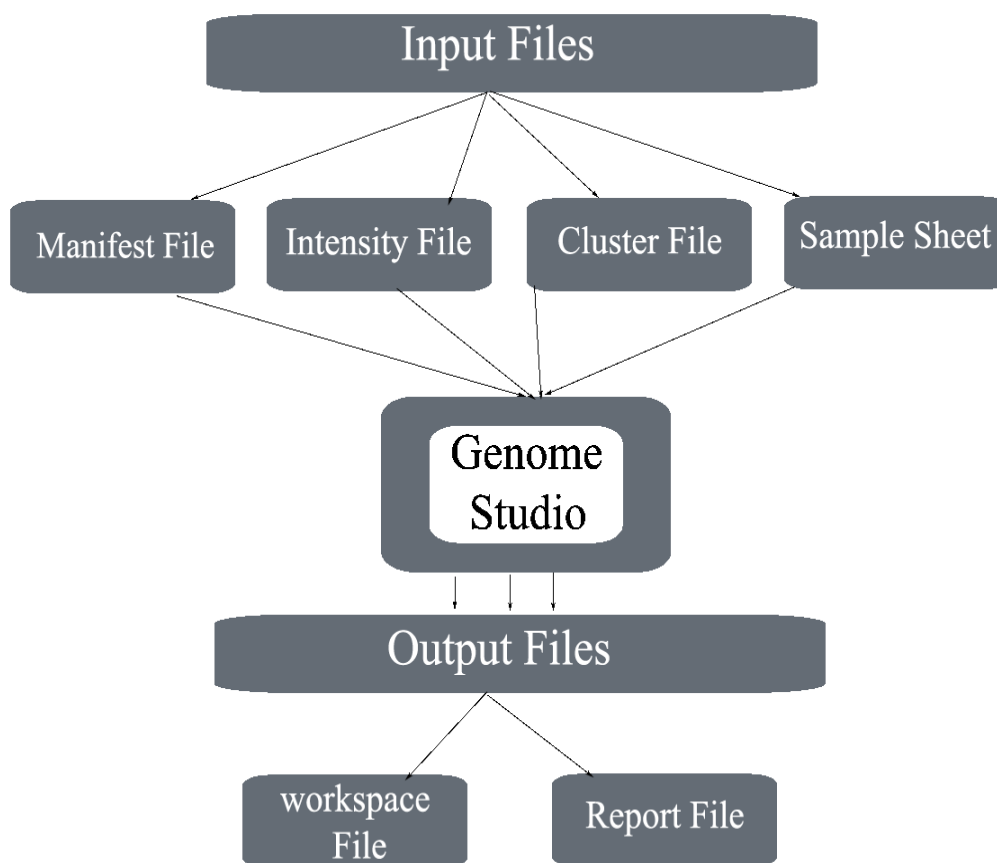


Figure 3 : This diagram illustrate the flow of different files throughout the Genome Studio Software.

The input data files of this software were:

- 1- **Manifest file** with extension (*.opa or *.bpm) which are the list of loci included and in the array in addition contain information and annotation for the ovine bead type, including the allelic identity of the locus, the DNA sequence surrounding the locus, the reference sequence (SNPID number) and the bead type used for each locus (ovine bead).
- 2- **Intensity file** with extension (*.idat) those files were generated by the system and saved in a folder named with the barcode number of the array, standard position designation (given to create unique name for each sample on the array), and it contain the red and green single intensity for each bead type that is being read by the iSan system.
- 3- **Cluster file** with extension (*.egt) that assigns and generates genotypes based on fluorescence intensity data which define the range of red and green signal intensity for AA, AB & BB genotypes. These are given to create unique names for each sample on the array that is used an approximate Gen scores call rate for routine quality control.
- 4- **Samples sheet** with extension (*.csv) were used to import information about the DNA samples that include array barcode and position for each sample, sample ID or name, source of micro-titer plate location of samples on the Beadchip, replicate information, sample's origin and phenotypic data.

When running the program for the analysis, multiple panels with multiple tabs were used which included:

- 1- **SNP table** that was used to determine whether a locus should be excluded from the analysis: (a) Index of the SNP for identification within the manifest, (b) Name of the locus assigned in the manifest, (c) The chromosome on which the locus is located, (d) The chromosome position of the locus, (e) Top Genomic Sequence on the Illumina-defined strand around the SNP, (f) Number expected cluster for a locus given that:
 - 1= for nonopolymorphic probes
 - 2 = for mitochondrial DNA and Y loci
 - 3 = for any other loci
- 2- **Sample table:** a list of all samples included in the experiment. It also contains lists of several fluorescence metrics called from the bead scan and two metrics that describe the performance of all SNP's across the entire samples. These are the 10th (p10GC) and 50th (p 50 GC) percentile gen call scores which were used for direct sample to sample comparison.
- 3- **SNP graph:** plots all samples for the active selected SNP in the full data table or SNP table. It is viewed either in Cartesian or polar coordinates. In Cartesian coordinates, the fluorescence intensity in one channel is plotted on the Y axis and the fluorescence intensity of the other channel is plotted on the X axis; in Polar coordinate it uses the horizontal axis to represent normalized θ , which is

the angle above the Cartesian X axis. The symbol “ θ ” is the angle of the deviation from pure “A” signals, where zero represents pure A signal and one represents pure “B” signals; the vertical axis which is “R” represents the sum of the normalized intensity of the red and green signals.

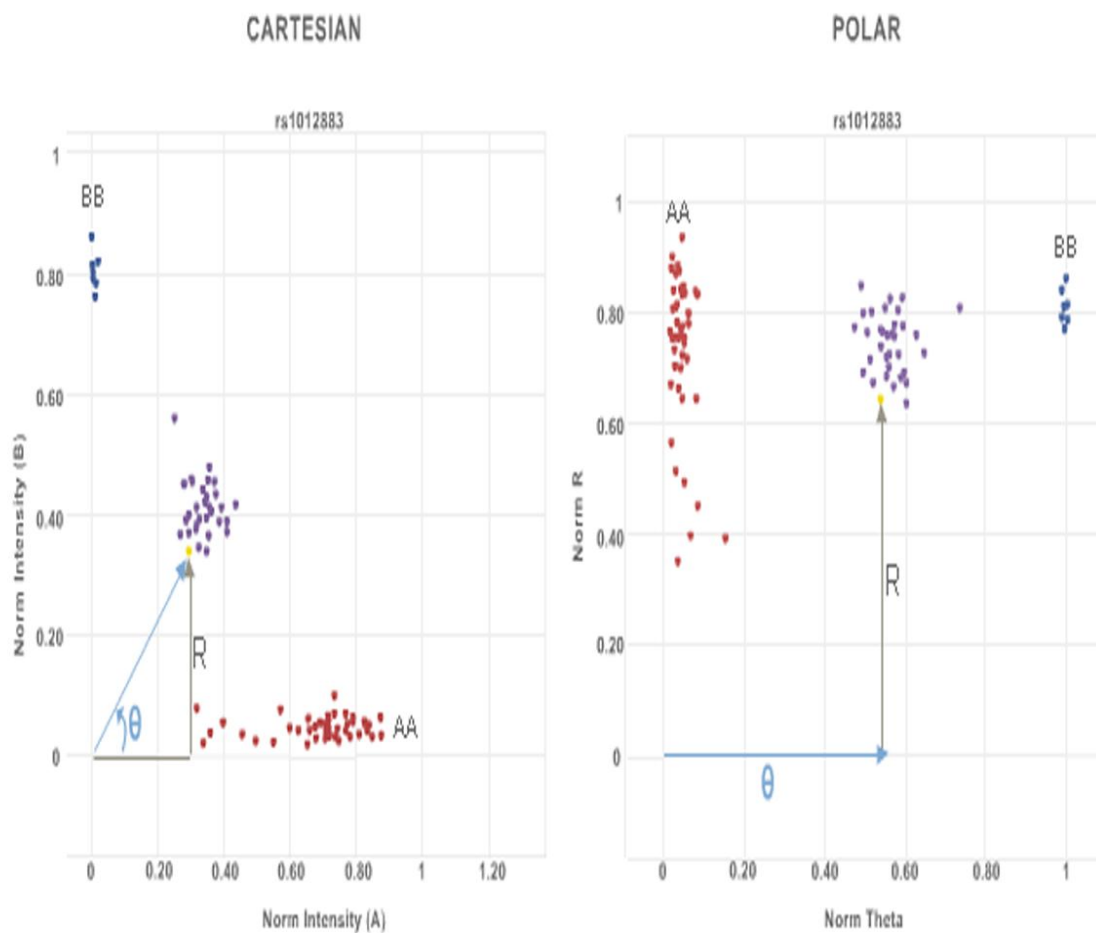


Figure 4: This graph represent the θ and R for the samples in both Cartesian and Polar coordinate graph for Genome Studio Software.

In the SNP graph, data points that fall within the call region are assigned to genotype calls. Any data points outside these regions are no calls, meaning that they are not assigned to genotype the call score ranges from zero to one, with one representing the

greatest confident in the call. The cross at the center of the call region is called the centroid. At that location the genotype calls are made with the highest confidence. Data points farther from the centroid are assigned genotype calls with less confidence, setting the threshold of data at a score of 0.15. Therefore, any data point with scores below 0.15 results in “no call”.

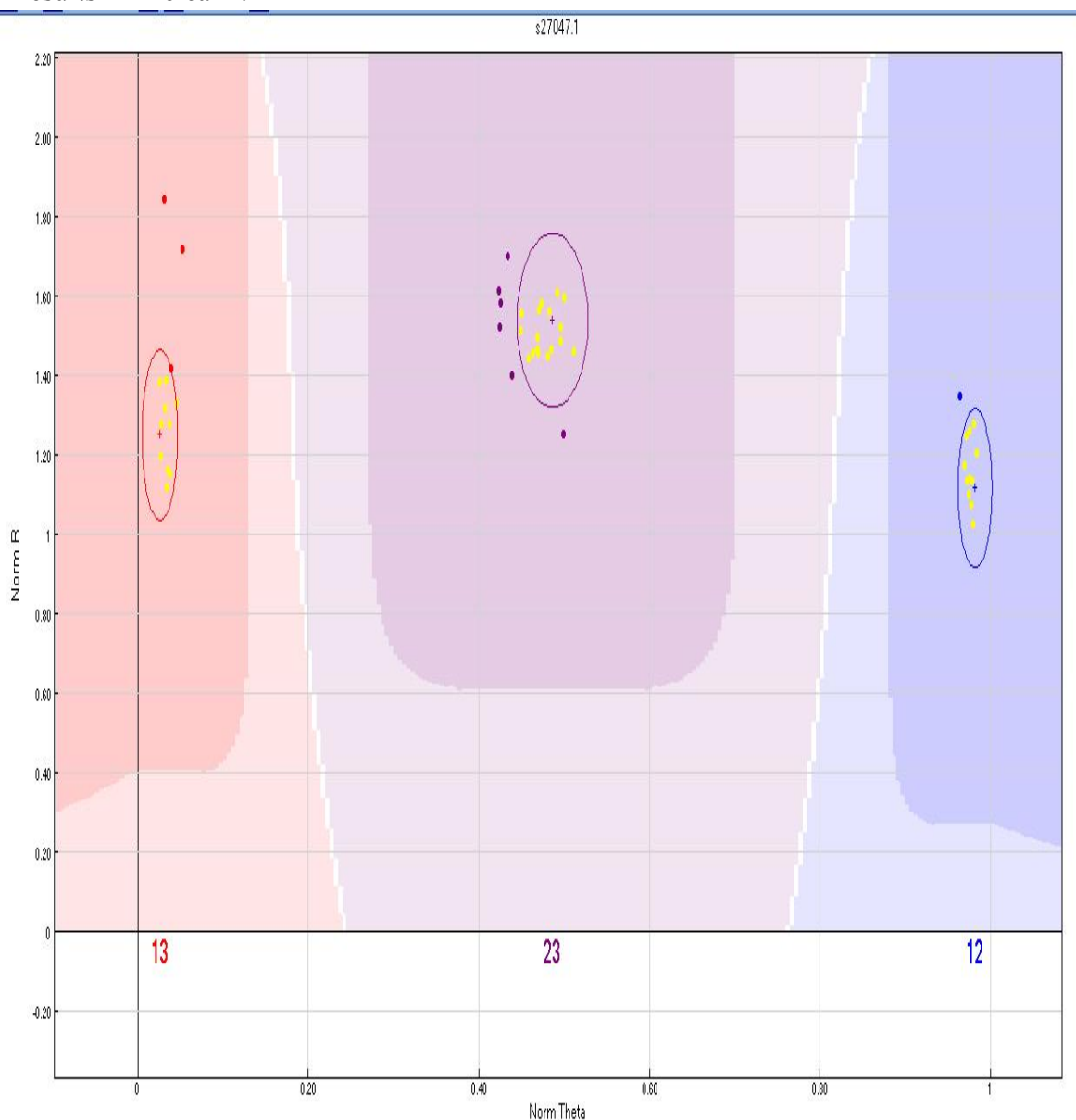


Figure 5: The Centroid and the called genotyping. The yellow points are the highest gen calls close to one. (GenomeStudio™ 2008.1 Framework)

4- **Full data table** contains two groups: (A) Annotation Column Data that provides reference information for all samples. For example, the “Position Column” that has the chromosomal position of the loci in the genome and “Address column” that contains a list of bead type associated with particular loci; also it is used to map each locus to a stripe on the array. (B) Empirical Column Data that lists observed data for each SNP per sample. This was used as a comprehensive source of information about the project. For example, “Gen Train Score” is the measure of the predictive power of the genotype calls made at that locus ranging from zero to one, with one being the best, and “Gen Call Score” refereeing to the predictive power of a specific sample genotyping call made for that SNP based on distance of the data points from the cluster centroid. The gen call histogram is used to evaluate the sample quality of the SNP since in high quality samples the majority of the scores bars are close to one.

5- **Error table:** that displays replicate or heredity errors.

A call rate for the SNPs was determined using the Genome Studio program which enables the detection and measurement of copy number variation. The resulting data (SNP’s position, Gen Call rate and Genotyping) from the Genome Studio software were exported for further statistical analysis.

6. Statistical analyses of SNPs associated to fiber diameter

The top and bottom 24 full Merino ewes for the year class 2003 was genotyped for an Illumina array with 54,241 SNPs. Data for each SNPs was stored in a Sun Ultra 5 unix machine. A one way analysis of variance was used to detect association of the three genotypes with fiber diameter. A computer program was written to sequentially compute analysis of variance for each single SNP. The following formulas were used to determine the F-value for ANOVA:

$$F = \frac{\frac{SS_{Trt}}{(k-1)\sigma^2}}{\frac{SS_{error}}{k(n-1)\sigma^2}} = \frac{MS_{Trt}}{MS_{error}}$$

with $SS_{Total} = SS_{Trt} + SS_{error}$, $SS_{Trt} = \sum_{i=1}^g n_i (\bar{y}_i - \bar{y}_{..})^2$, $SS_{error} = \sum_{i=1}^g \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$, where SS_{Total} = the total sum of squares, SS_{Trt} = sum of squares of the genotype effect, SS_{error} = the sum square of the error, \bar{y}_i = the group mean, $\bar{y}_{..}$ = the overall mean. y_{ij} = the deviation of a response from each group mean, MS_{Trt} = mean square for treatment effect, MS_{error} = mean square for error. The F- value was tested with 2 and 45 degrees of freedom.

Results

Statistical Power

Statistical power after selection of the top and bottom 5% from a population consisting of 1000 individuals is depicted in Figures 7 and 8 for significance levels of 0.01 and 0.05, respectively. Allele frequencies of the SNP of 0.1 and 0.5 were considered. The power using selection of extremes for a population of 1000 animals is enough to detect SNPs with an effect of at least 0.3 phenotypic standard deviations. Power is higher at intermediate allele frequencies.

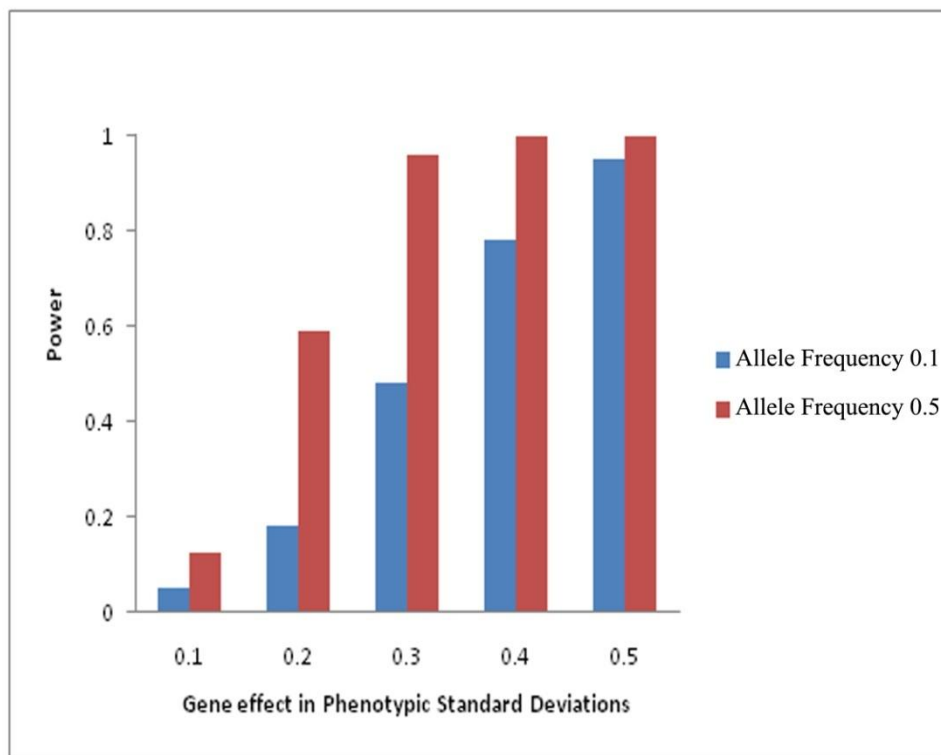


Figure 6: Power for varying gene effects (0.1 to 0.5 phenotypic standard deviations) at $\alpha = 0.01$ with selection of 5% in the extremes. Allele frequency of the SNP was either 0.1 or 0.5.

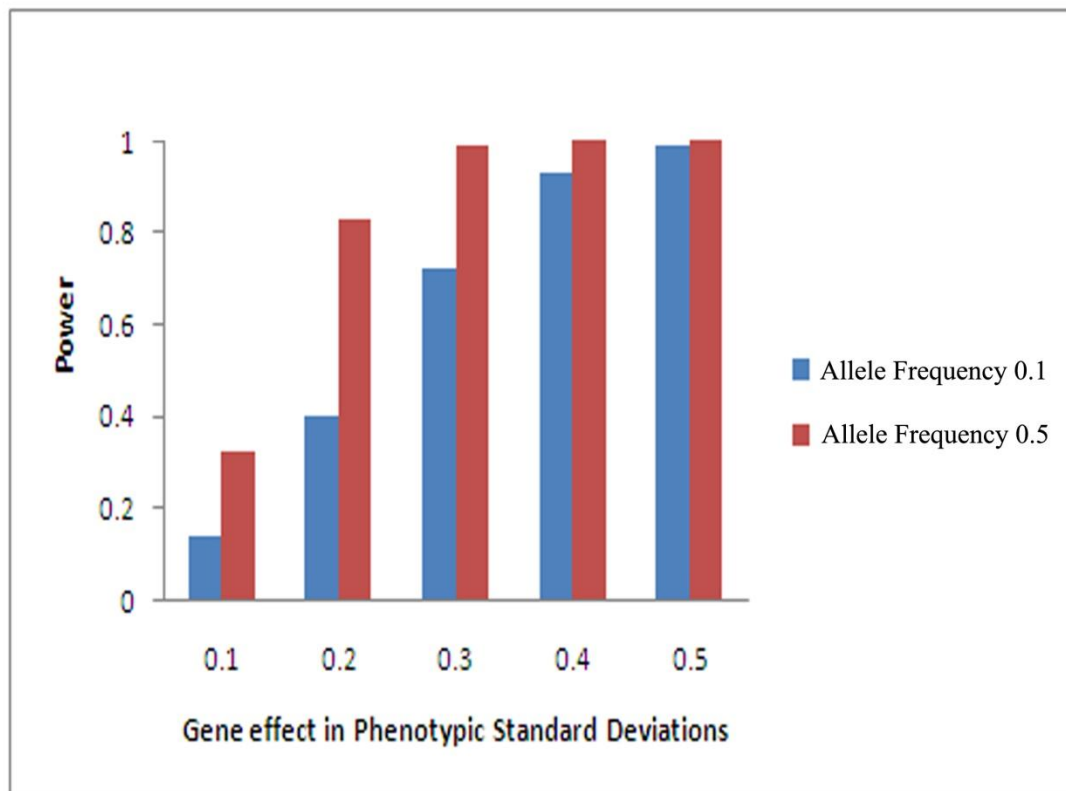


Figure 7: Power for varying gene effects (0.1 to 0.5 phenotypic standard deviations) at $\alpha = 0.05$ with selection of 5% in the extremes. Allele frequency of the SNP was either 0.1 or 0.5.

Year	Allele frequency NO. of Animals/PSD	0.1					0.5				
		0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
1998	36	0.01	0.01	0.01	0.01	0.01	0.05	0.05	0.05	0.05	0.05
1999	147	0.01	0.01	0.06	0.18	0.32	0.05	0.1	0.32	0.63	0.86
2000	183	0.01	0.06	0.13	0.26	0.42	0.05	0.16	0.44	0.73	0.89
2001	172	0.01	0.04	0.11	0.26	0.42	0.05	0.16	0.42	0.73	0.91
2002	178	0.01	0.04	0.11	0.26	0.42	0.05	0.16	0.42	0.73	0.91
2003	201	0.01	0.06	0.13	0.32	0.52	0.3	0.21	0.5	0.81	0.94

Table 1: Statistical power for varying gene effects and population sizes corresponding to year classes existing in a Merino sheep flock. The significance level was $\alpha = 0.01$, SNP allele frequencies were 0.1 or 0.5. Animals were selected among the top and bottom 5%.

Year	Allele frequency NO. of Animals/PSD	0.1	0.1	0.1	0.1	0.1	0.5	0.5	0.5	0.5	0.5
		0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
1998	36	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
1999	147	0.03	0.11	0.26	0.39	0.58	0.11	0.31	0.62	0.84	1
2000	183	0.06	0.18	0.34	0.52	0.68	0.11	0.37	0.7	0.89	0.94
2001	172	0.06	0.16	0.32	0.52	0.65	0.11	0.37	0.68	0.91	1
2002	178	0.06	0.16	0.32	0.52	0.68	0.11	0.37	0.68	0.89	0.94
2003	201	0.06	0.16	0.34	0.58	0.73	0.13	0.47	0.76	0.94	1

Table 2: Statistical power for varying gene effects and population sizes corresponding to year classes existing in a Merino sheep flock. The significance level was $\alpha = 0.05$, SNP allele frequencies were 0.1 or 0.5. Animals were selected among the top and bottom 5%.

The Merino flock available for this study was consisting of ewes from different year classes and also with varying breed composition. Selection of extremes must be used in animals belonging to the same year class and with the same breed composition to avoid spurious significant effects. We computed power for sizes of the year classes of the available Merino flock. The results for $\alpha=0.01$ and $\alpha=0.05$ are in Tables 1 and 2, respectively. Power of detection is large enough to detect SNPs associated to fiber diameter in the largest year classes. We used the largest year class, 2003, for large scale genotyping using Illumina's microarray and the top and bottom 24 animals selected for fiber diameter from the same breed (Full blood Merino sheep).

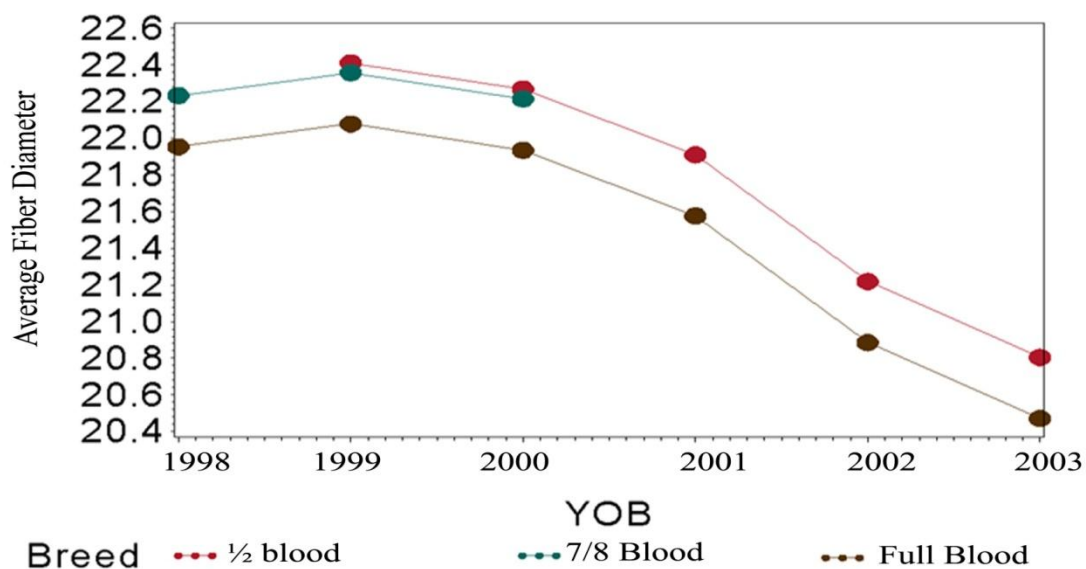


Figure 8: This graph demonstrate the effect of different years of Birth (YOB) on fiber diameter in all breed types in the whole population.

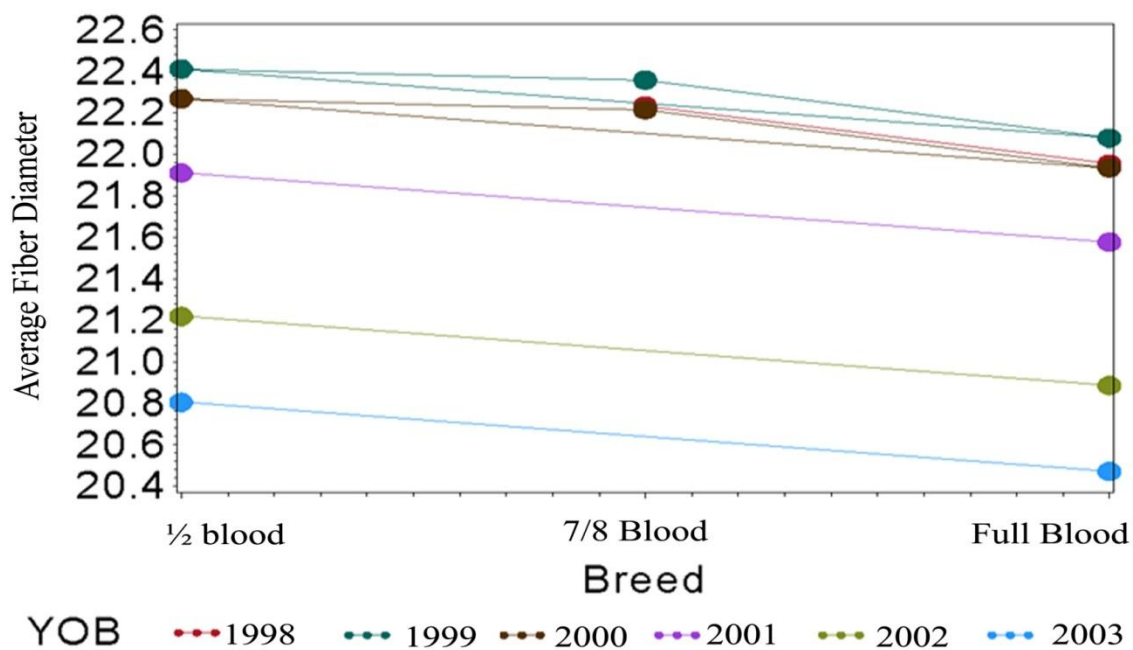


Figure 9: This graph demonstrate the effect of different breed types on fiber diameter in all years of Birth (YOB) in the whole population.

After scoring all SNPs, removing homozygotes, eliminating SNPs with a call rate lower than 0.90, the total of SNPs remaining was 18,214. We carried out a one way ANOVA with each SNPs. The distribution of the F-values is depicted in Figure 9. The threshold in that figure indicates the value beyond which SNPs were declared significant. Because of the multiple testing we would expect about 182 significant SNPs just by chance. We obtained 208 significant results which indicate that some of them could be truly significant. The F-values of all SNPs per chromosome are depicted in Figure 10. SNPs in chromosomes 1, 14, 15, 21 and 26 had the highest significant results.

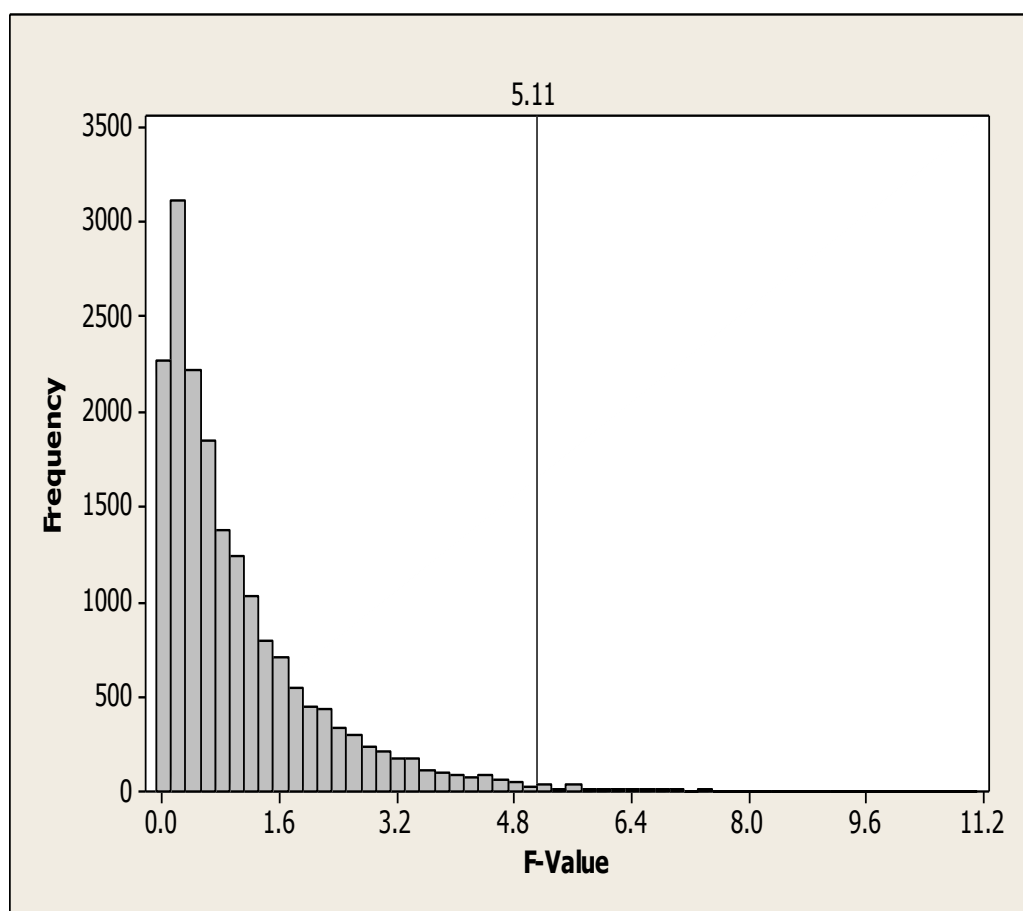


Figure 10: Distribution of F-values after selection of extremes in 18,214 SNPs at 2 and 45 degrees of freedom.

Fvalue	SNP Position	SNP Name	Chr	Sequence
8.791	1475013 47	OAR1 _1475 01347. 1	1	AGAAAAGTAAGTATAATGGCGCAGAGACAGGGTGTATTTTTAGGAT AAGGGATTATATC[T/C]GGAGAGCAGAACTTCTAAATTACCAATAA TGTTGTGTATTTAATTTAAANTATTATAAT
11.056	2429293 35	s2302 2.1	1	TTCCCCAGTCCAACCTTCACCAGCTCCAACCTCTCCAAGGCTGGGACA TAGCAGGGCCAGG[A/G]GAGACCAGATGTAGCTGTTTTAAAGGGATG CGGATGGAGATTGTCTGTGCAAGCCTAGTGG
8.582	2115584 0	OAR2 _2115 5840.1	2	AAATTTTATTAGTTTAGCTCTCTGTAAGCATTCTCATGCATCGTTAA CATTACACATTGC[A/G]AAGAAGTCTTTTCCTTAAATTGTATATATTT GGCTATAAATTTGAGACTGTAGTAAATTG
9.355	2147185 34	OAR2 _2147 18534. 1	2	GCGATCTGAAAAGTTAACAAGATTGTTTCTAGGTCTTGTGTATGGTC ACCTTGTAAGGA[A/G]TACTGCCATAAGATCTCAGTGCCCATGTTT ATATTTGTATTTATGACAAGTTGTACTAAT
8.233	1775578 47	OAR3 _1775 57847. 1	3	AGGAAAGAATGGAAGGAATACTTANGAAATAAGTGGTGTACATT AAGAAGGAGCCTAAG[A/G]TGATTTGTCAGATTCTAGGTTTCAGTGG TTGGGTAGTCGTTAACCAGAAGTAGAGACTAT
8.590	5679952 2	OAR3 _5679 9522.1	3	AGACCTTTTCCTTCCTAACTTTACTTTTTAAGGGCATGAAAAAAAAAT CACTTAGCTTACT[T/G]CAGAACTGTCATATATTTACTTTTTGTTTAA CTTTTTTTGCTTCTACCTTGATCTCATA
8.272	1109901 17	OAR6 _1109 90117. 1	6	CCAGACTGATGGATTTTGATGACAGAAAGTACTGAAGTATTGTTGG AGCACAAGAGCTCA[T/C]GGGCACAAGAGCCAGTGTGACATGGT TAGAGTTGAGGTTCAATGGNAAAGCAGGGGAC
9.131	5709000 2	OAR6 _5709 0002.1	6	TCAATTCCTTAGAGGTTCTCAGTTTTATGCAAGGAGTTAAATTGGGA TTCTGCCTTGATT[A/G]TGGTTAAGATAATTTATCTTGACTTCACCAT TGTGAAATATCAACTGTCAATTCAGATAT
8.439	6526462 3	OAR7 _6526 4623.1	7	CTGTACTGGGCAGGTCAGAGCGCTGGAACAGGAAGAGGAAATGAA TCTTGGGCAAAAAGTG[T/C]ATAAAGACCTTTCCAAGGAAGACATTCT GCATTCCTGACATCCCCCTGAGGGATGCTGAT
9.432	2473704 8	OAR8 _2473 7048.1	8	GAAACAATAACTTATGGATGCCTGTAATTATACTGTCATGTAAA TTAGTTGTGGCTG[A/G]TTGGTTGGACTCAAGGCGAGAATATTTTCT AATACTTCTGCCACAGAGCATCTAGACAG
9.653	7372845 8	OAR8 _7372 8458.1	8	ATGGTTAAAGTTTTAGCAATACTGTCTCCTCTTTTCAAAGAAATAGT GTTCTCTTTCAA[T/C]GTCGGAAACATTTTGAAGAATTTTACCTTT GGATTAGACTGGTCTGGGAAAGTTAATT
8.938	1007908 76	OAR9 _1007 90876. 1	9	TTCGATGCCTAGTAGGGTTTTGCAAGCCCCTACCTTGAGTAATCAGCT AATTCTTGGTCAA[T/C]GATGGTCAATGGTCAATCCCCTGTGTCCTA GAGCCGTGAAAGAGGCATCTTGCCACCTG
10.809	2043106 8	OAR1 4_204 31068. 1	14	AGAGCTGATAAGGAAGGTTTTGCTAGAGAATCATGTATCTCTTCTTT TTCCCTCTGACC[A/G]GTGAGATGCAATTCTCTATTACTCTGNNNNN NN

Table 3: The most significant SNPs, their position due to Illumina® genotyping, Chromosome number (Chr) and Sequence.

Fvalue	SNP Position	SNP Name	Chr	Sequence
10.076	6456514 8	s465 20.1	15	CATGAATGAGACCTGCCATGGATCAAGTGCAGGGC TGGCCCTGAAGATGAGACTGTTGGA[T/C]ACCAAAG CACAGTCTGTTGGACACCAAAGCACACCAAAGCAT GGTCCTTTCTGCTCTTGC
8.699	6191070 6	OAR 17_6 1910 706.1	17	CTCTTCTGGTTCTTCTGGTTTCTGTCCTGGATAGTTG AGGAGAGGGACAATAACCATGAT[A/C]GTTTCTCTT TTCAAATAATCTGCTGTTTGTCTAAACAGGAAAA AATTACTCCAAGGTCA
8.722	6192416 3	OAR 17_6 1924 163.1	17	ACTTTGACCATCAGTAACCATCAGCCTCCAGCTAAT ATAAGACATGGTCAAGATTTTATC[T/C]AGAGGAGG TTTCTCCAACAAGTCTTCGTAGGGGGTGTTCCTTC TTCCCTTTCAACCCCC
9.947	6644217 6	OAR 18_6 6442 176.1	18	TCTTCTCTGAGATTATACTAGTTAAAGTCATTGCCT GTAAAGCCAGTATCCTGATATAAA[T/G]GTGAAGAA AACCATTCCCTTTATCCCCTTCATGGCTGTTTAAGT ATCTTGAAGTATAAG
9.089	3993024 8	OAR 19_3 9930 248.1	19	CTCTGGTCTGAACGTGTTTACCTGAAGCAGGAATG AATGACATTCATGCCATGTGGTTCA[A/C]GCAGGGC TCCTCTGTTGGTTCTCCCACAGCAGCCTTAGGATTG GCGCAGTGTGGGCACG
10.166	4773877 3	s441 29.1	21	CTTAGAGCTCAGACCCAGTGGAGATGGGATGTGTG CGGGGCCACCAAGGAGACGTGCTCA[T/C]TGTGCCT GGCAGGAGCAGACAGGATACCAGAAGCACCAGTTC AGGCAGCTCCTGGGGGCA
11.047	4169507 1	s432 78.1	26	TGATAATCCATTGCCTTCTCACCAGGAGGCCTTTGG GTTTCTCTCTAATCGTAATACCGA[A/G]TGTCAGCAT GGAAGAGGNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNN

Table 4: The most significant SNPs, their position due to Illumina® genotyping, Chromosome number (Chr) and Sequence.

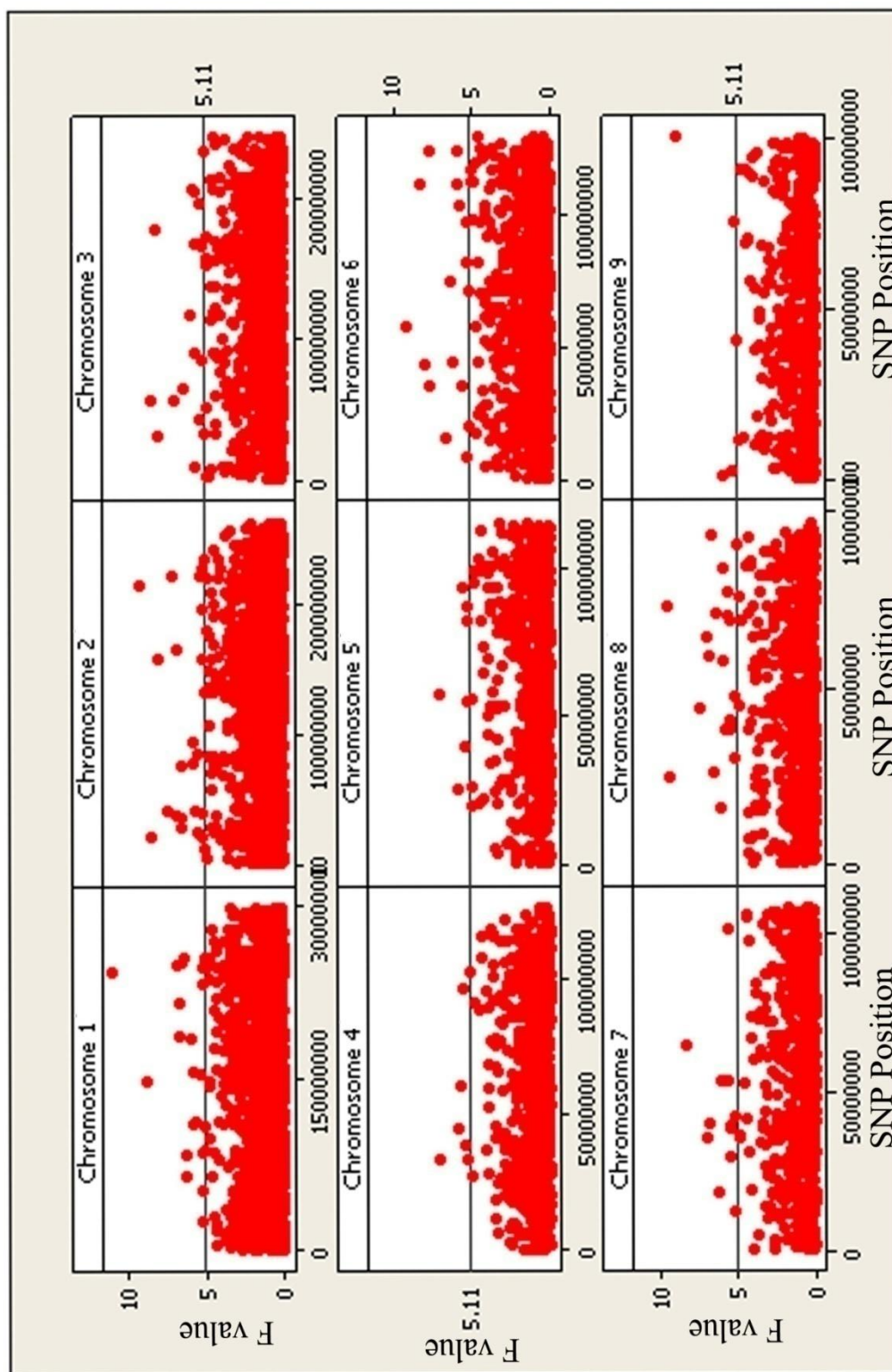


Figure 11 : This graph shows the position and F-value of the all SNPs for chromosome 1 to 9 in Merino sheep.

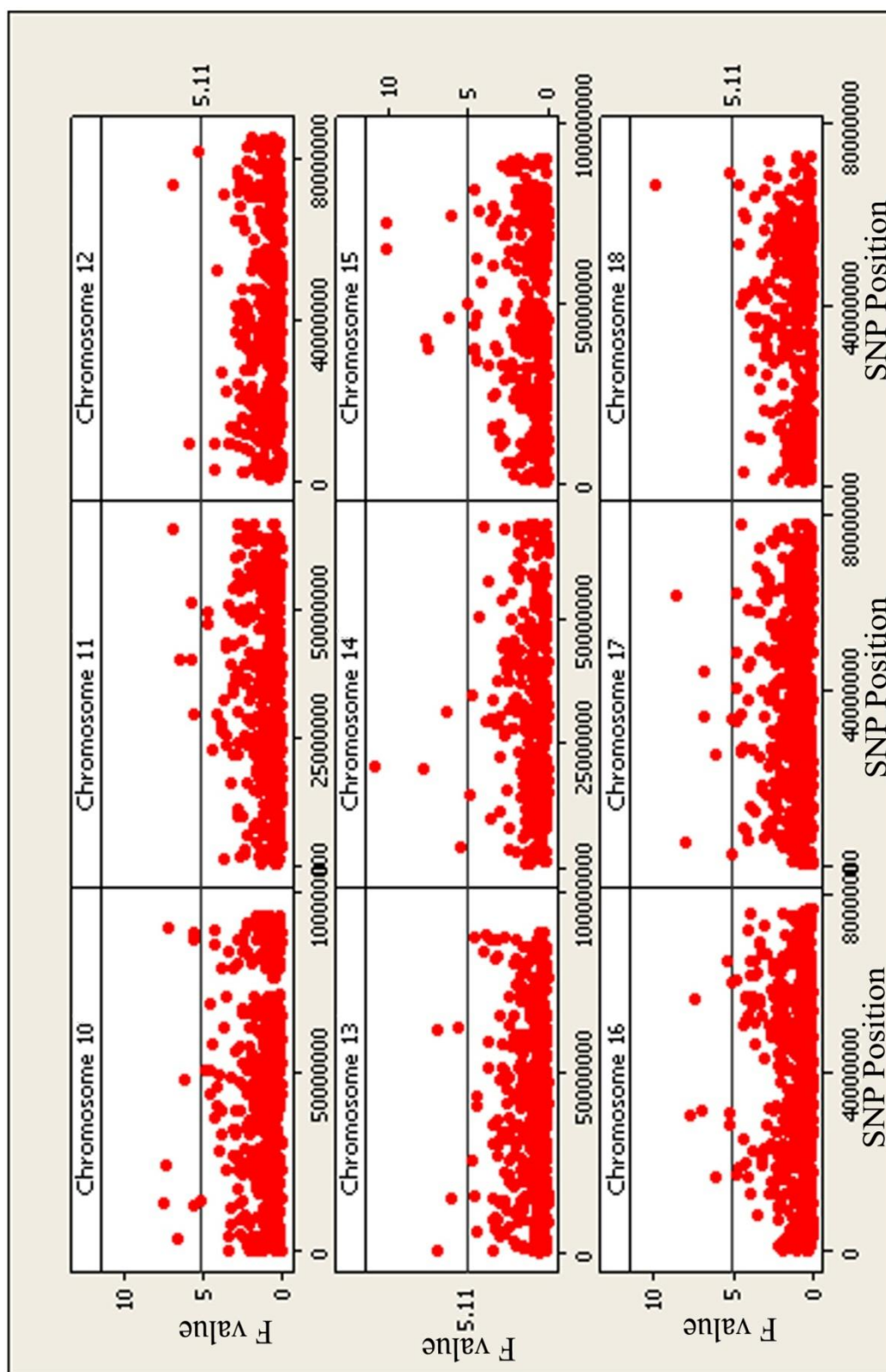


Figure 12: This graph shows the position and F-value of all SNPs for chromosome 10 to 18 in Merino sheep.

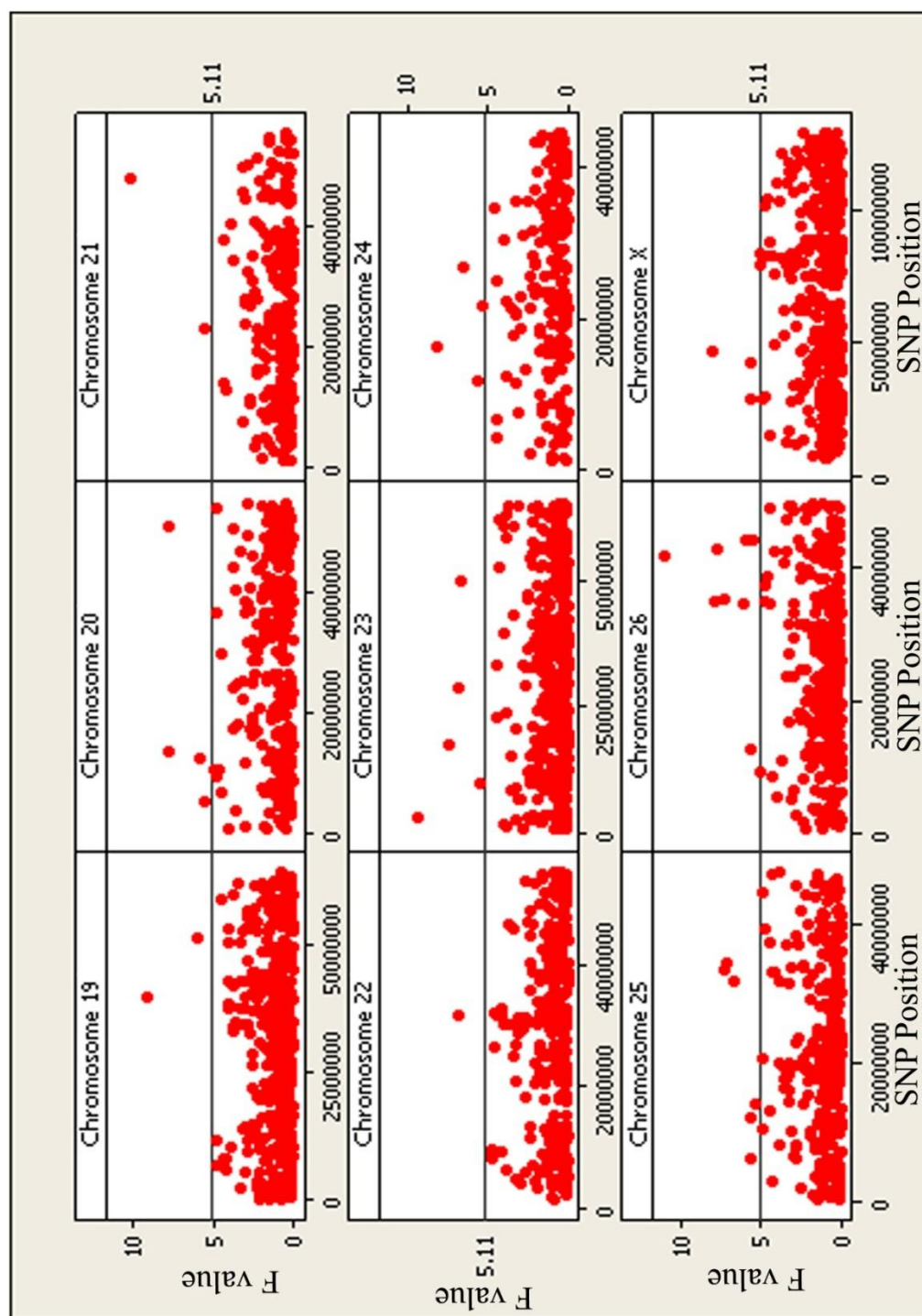


Figure 13: This graph shows the position and F-value of all SNPs for chromosome 19 to 26 and the X chromosome in Merino sheep.

Discussion

Our results indicate that Statistical power for selective genotyping is sufficient for detection of SNPs linked to QTL with medium to large size. We showed that even for populations of approx 250 animals, power of detection is high after selecting the top and bottom 5% of the population. There are two main factors influencing power when using extreme phenotypes. On one hand, the number of individuals genotyped is a decisive factor to achieve power. On the other hand, selection pressure influences power. The higher the percentage of selection in each extreme, the higher the power of detection we obtained. However, if the population is not very large then high selection pressure would yield small numbers of genotyped individuals and, therefore, reduced power. Thus, selection of extremes is limited by the population size scored for the quantitative trait. Only in relatively large populations, high pressure of selection of extremes can provide enough individuals to achieve high power.

We applied selection of extremes for fiber diameter in a Merino sheep flock together with GWAS. We obtained 208 significant results out of 18,214 tests. With a significance level of 0.01 the expected rate of false positives is 182. Therefore, we obtained an "excess" of significant results of 26 tests. These results suggest that some of those significant results may have a truly biological effect on fiber diameter. The highest significant results were obtained in Chromosomes 1, 14, 15, 21 and 26. KRTAP6 and KRTAP8 are candidate genes for fiber diameter mapped to OAR1. These genes have been associated to fiber diameter (Pearson, et al., 1994). DNA-markers in OAR6 and OAR25 have been associated to fiber diameter in previous work (Ponz, et al., 2001; Bidiniost et al., 2008;

Allain et al., 2006). Our results also showed that SNPs on OAR6 might be associated to fiber diameter.

GWAS uses a very large number of SNPs and, therefore, significant genome wide associations are difficult to obtain due to the huge number of tests performed. A next step in this study would be to confirm the detected associations but using an independent sample. For the situation developed in this study, the unselected population would be the resource population to validate the results obtained in GWAS. At this point, only the highly significant SNPs in this study would be used for further testing. It would reduce the problems associated with multiple testing.

One of the limitations of selective genotyping is that detection is limited to the trait being selected. This applies to traditional linkage as well as associations studies as proposed in this research. If several traits are of interest then research should be aimed to make the most economical efficient experimental design. If the traits are genetically correlated then selection of extremes might still be provide power of detection for QTL affecting multiple traits.

References

- Allain D, Schibler L, Mura L, Barillet F, Sechi T, Rupp R, Casu S, Cribiu E and Carta A
2006. QTL detection with DNA markers for wool traits in a sheep backcross Sarda X
Lacaune resource population. 8th World Congress on Genetics Applied to Livestock
Production, Belo Horizonte, MG, Brasil, pp. 5-7. Retrieved August 18, 2006, from
http://www.wcgalp8.org.br/wcgalp8/articles/paper/5_191-252.
- Andersson, L., C. S. HALEY, H. ELLEGREN, S. A. KNOTT, and M. JOHANSSON,
1994. Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science*
263:1771-1774
- Baxter, B. P., M. A. Brims, and T. B. Taylor. 1991. Description and performance of the
Optical Fiber Diameter Analyzer (OFDA). Report No. 8. IWTO Tech. Comm. Mtg., Nice,
France.
- Bidinost F, Roldan DL, Doderio AM, Cano EM, Taddeo HR, Mueller JP and Poli MA
2008. Wool quantitative trait loci in Merino sheep. *Small Ruminant Research* 74, 113-
118
- Bradley, D. R., R. L. Russell, and C. P. Reeve. 1996. Statistical power in complex
experimental designs. *Behaviour Research Methods, Instruments, and Computers* 28:319-
326.
- Bray, R. J. 1955. *Wool Characteristics in Relation to Manufacturers' Requirements*.
International Wool Secretariat, London.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edition. Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey.

Darvasi, A., and M. Soller. 1992. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theor. Appl. Genet.* 85:353–359

Fairweather, P. G. 1991. Statistical power and design requirements for environmental monitoring. *Australian Journal of Marine and Freshwater Research* 42:555-567.

GenomeStudio™ 2008.1 Framework. User Guide. An Integrated Platform for Data Visualization and Analysis, Part # 11318815, Rev. A

Georges, M., D. Nielsen, M. Mackinnon, A. Mishra, and R. Okimoto. 1995. Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* 139:907-920

Hunter, L., and E. Gee. 1980. The effects of staple crimp, resistance to compression and fiber diameter and length characteristics on the physical properties of wool worsted yarns. *Proc. 6th Quin. Int. Wool Text. Res. Conf., Pretoria, South Africa*, IIL327

Illumina. 2008. Infinium® HD Assay Ultra, Manual, Catalog # WG-901-4005, Part # 11328095 Rev. B

Jenkins ZA, Dodds KG, Henry HM, Beattie AE, and Montgomery GW. QTLs for wool production traits identified in Merino * Romney backcross flock. *Proc XXVIth Int Conf Anim Genet 1998 Auckland New Zealand*, p. 101.

Kraemer HC, S.Thiemann. *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA; 1987.

Lang, W. R. 1964. The technical relevance of wool quality. *Wool Technol. and Sheep Breed.* 11(2):89.

Lebowitz, R. J., M. Soller, and J. S. Beckmann. 1987. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theor. Appl. Genet.* 73:556–562.

Lipsey, M. W. 1990. *Design Sensitivity: Statistical Power for Experimental Research.* Sage, Newbury Park, California.

Parsons Y. M., Cooper D. W., and Piper L. R., 1994. Evidence of linkage between high-glycine-tyrosine keratin gene loci and wool fibre diameter in a Merino half-sib family. *Anim Genet* 25: 105–108.

Peterman, R. M. 1990. Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fish and Aquatic Sciences* 47:2-15.

Ponz R, Moreno C, Allain D, Elsen JM, Lantier F, Lantier I, Brunel JC and Pe´rez-Enciso M. 2001. Assessment of genetic variation explained by markers for wool traits in sheep via a segment mapping approach. *Mammalian Genome* 12,569-572.

Powell B. C. Molecular genetics. In: *The Genetics of Sheep.* 1997. Edited by Piper L and Ruvinsky A. Oxon, UK: CAB International, page 149–181.

Roldan D. L., A. M. Doderio, F. Bidinost, H. R. Taddeo, D. Allain, M. A. Poli and J. M. Elsen. 2010. Merino sheep: a further look at quantitative trait loci for wool production. *The Animal Consortium* doi:10.1017/S1751731110000315

Standards. (2010). Retrieved from Australian Wool Exchange Ltd:
<http://www.awex.com.au/standards/1pp-certification.html>

Stobart R. H., W. C. Russell, S. A. Larsen, C. L. Johnson and J. L. Kinnison. 1986.
Sources of Variation in Wool Fiber Diameter *Journal of Animal Science* . 62:1181-1186

Taylor, B. L., and T. Gerrodette. 1993. The uses of statistical power in conservation
biology: the Vaquita and Northern Spotted Owl. *Conservation Biology* 7:489-500.

Thomas, L., and F. Juanes. 1996. The importance of statistical power analysis: an
example from *Animal Behaviour*. *Animal Behaviour* 52:856-859.

Toft, C. A., and P. J. Shea. 1983. Detecting community-wide patterns: estimating power
strengthens statistical inference. *The American Naturalist* 122:618-625.

Von Bergen, W. 1963. What the manufacturer requires in raw wool. *Wool Technol. and
Sheep Breed.* 101:43.

Appendix:

i. FORTRAN computer program code to determine the Statistical Power using Chi square test:

```

real*8 X, Y, Z, Fa1, Fa2

integer iseed,gs1(100000),gs2(100000),nind

real*8 ggs1(100000), ggs2(100000), PS, PSL

real*8 EN(100000), G(100000), P(100000)

real*8 Psort(100000), PP(100000)

real*8 Gsort(100000), Ensort(100000),PPP(100000)

real*8 GG(100000), EEN(100000),mean

real*8 VE,VA,SD,PSD,domi,per,perc

real*8 ngs11H,ngs22H, ngs21H, ngs12H

real*8 ngs11L,ngs22L, ngs21L, ngs12L

real*8 FA1H, FA1L, FA2H, FA2L

real*8 E1,E2,chi2

open(1,file='extre.output')

c nrep = number of replicates

nrep=1000

per=0.05

nind=1000

GP=0.4

mean=20.2643

domi=0

var=29.45

SD=SQRT(var)

c HERITABILITY = h2

h2=0.5

Fa1=0.1

c perc=50

c nind=(perc/per)

```

```
c   per= (perc/nind)
c   print*, 'Per' , per
   print*, 'nind' , nind
   PS=nind*per
   ns=ps
   PS1=(1.-per)*nind
   nsl=ps1+1
iseed=1193784321
Fa2 = 1. - Fa1
c   print*, 'high group ', ns
c   print*, 'low group ', nsl ,nind
   nx1=0
   nx2=0
   ny1=0
   ny2=0
   jj=100
   lflag=0
   do 999 iii=1,nrep
c   if (iii.eq.jj) then
c   print*, 'nrep=', jj
c   endif
   kk=iii/1000
   if (lflag.eq.0.and.kk.eq.1) then
   print*, 'nrep=', iii
   lflag=1
   kka=kk
   else
   if (kk.ne.kka) then
   print*, 'nrep=', iii
```

```
kka=kk
endif
endif

c   if (iii.eq.5000) then
c   print*, 'nrep= 5000'
c   endif
c   if (iii.eq.10000) then
c   print*, 'nrep=10000'
c   endif
c   if (iii.eq.15000) then
c   print*, 'nrep= 15000'
c   endif
c   if (iii.eq.20000) then
c   if (iii.eq.25000) then
c   print*, 'nrep=25000'
c   endif
c   if (iii.eq.30000) then
c   print*, 'nrep=30000'
c   endif
c   if (iii.eq.35000) then
c   print*, 'nrep= 35000'
c   endif
c   if (iii.eq.40000) then
c   print*, 'nrep=40000'
c   endif
c   gs1=0
do i=1,nind
gs1(i)=0
```

```
      Call uniform (iseed,x)
c     print*,x
      if (x.gt.Fa1) then
c     print*,x, ' x=1'
c     nx1=nx1+1
      gs1(i)=1      else
c     print*, x, ' x=2'
      gs1(i)=2
c     nx2=nx2+1
      endif
c     gs2=0
c     CALL RANDOM (ISEED,y)
      Call uniform (iseed,y)
c     print*,y
      gs2(i)=0
      if (y.gt.Fa1) then
c     print*,y, 'y=1'
      gs2(i)=1
      else
c     print*,y, 'y=2'
      gs2(i)=2
      endif
      enddo
      do i=1,nind
c     print*,gs1(i), gs2(i)
      enddo
c     print*, nx1,nx2
c     print*, ' '
c     print*, ' '

```

```

c   VA=0
c   G(i)=0
c   EN(i)=0
c   P(i)=0
   VA=var*h2
c   print *, ' The additive variance of the trait is ',VA
   VE=0
   VE=(var-(var*h2))
c   print *, ' The environmental variance of the trait is ',VE
c   print *, ' The mean is ',mean
   do i=1,nind
call random (iseed,x)
G(i)=X*DSQRT(VA)
call random (iseed,y)
EN(i)=y*DSQRT(VE)
P(i)=EN(i)+G(i)+mean
   if (gs1(i).eq.1.and.gs2(i).eq.1) then
   P(i)=P(i)+GP*sd
   endif
   if (gs1(i).eq.2.and.gs2(i).eq.2) then
   P(i)=P(i)-GP*sd
   endif
   if (gs1(i).eq.1.and.gs2(i).eq.2) then
   P(i)=P(i)+domi*GP*sd
   endif
   if (gs1(i).eq.2.and.gs2(i).eq.1) then
   P(i)=P(i)+domi*GP*sd
   endif
c   write (6,1001) G(i), EN(i), P(i)

```

```

c 1001 format (3f10.4)
    enddo
c   print*, "
c   print*, "
c   print*, "
    do i=1,nind
        GG(i)=0
        EEN(i)=0
        ggs1(i)=0
        ggs2(i)=0
    enddo
    PP=P
c   Print*, 'Genetic   Enviromental   Phynotypic   A1 A2'
    do i=1,nind
        Psort(i)=0
c   print*,P(i)
    enddo
    call ssort (P,Psort,nind)
    do i=1,nind
c   print*,P(i),PP(i)
    enddo
    j=0
    do 112 i=1,nind
        do 113 k=1,nind
            if (P(i).eq.PP(k)) then
c   write (6,1000) G(k),EN(k), P(i)
c 1000 format (3f10.4)
                j=j+1
                PPP(j)=P(i)

```

```

GG(j)=G(k)
EEN(j)=EN(k)
ggs1(j)=gs1(k)
ggs2(j)=gs2(k)
go to 112
endif
113 continue
112 continue
ngs11H=0
ngs22H=0
ngs12H=0
ngs11L=0
ngs22L=0
ngs12L=0
do i=1,ns
c   print*,GG(i),EEN(i),PPP(i), ggs1(i),ggs2(i)
   if (ggs1(i).eq.1.and.ggs2(i).eq.1) ngs11H=ngs11H+1
   if (ggs1(i).eq.2.and.ggs2(i).eq.2) ngs22H=ngs22H+1
   if (ggs1(i).eq.1.and.ggs2(i).eq.2) ngs12H=ngs12H+1
if (ggs1(i).eq.2.and.ggs2(i).eq.1) ngs12H=ngs12H+1
   enddo
c   print*, "
c   print*, "
do i=ns1,nind
c   print*,GG(i),EEN(i),PPP(i), ggs1(i),ggs2(i)
   if (ggs1(i).eq.1.and.ggs2(i).eq.1) ngs11L=ngs11L+1
   if (ggs1(i).eq.2.and.ggs2(i).eq.2) ngs22L=ngs22L+1
   if (ggs1(i).eq.1.and.ggs2(i).eq.2) ngs12L=ngs12L+1
   if (ggs1(i).eq.2.and.ggs2(i).eq.1) ngs12L=ngs12L+1

```

```

        enddo
c    print*,"
c    print*,' '
c    print*,' F1-1 F2-2 F1-2 and F2-1'
c    print*,'High', ngs11H, ngs22H, ngs12H
c    print*,'Low', ngs11L, ngs22L, ngs12L

FA1H=0
FA1L=0
FA2H=0
FA2L=0

FA1H=(2*ngs11H)+(ngs12H)
FA1L=(2*ngs11L)+(ngs12L)
c    print*,'Allele 1 H'
c    print*,FA1H
c    print*,'Allele 1 L'
c    print*,FA1L
c    print*,"
c    print*,"
FA2H=(2*ngs22H)+(ngs12H)
FA2L=(2*ngs22L)+(ngs12L)
c    print*,'Allele 2 H'
c    print*,FA2H
c    print*,'Allele 2 L'
c    print*,FA2L
if((FA1H+FA1L).gt.0.and.(FA2H+FA2L).gt.0) then
    E1=(FA1H+FA1L)/2.
    E2=(FA2H+FA2L)/2.
c    if(E1.lt.5.or.E2.lt.5)go to 112
chi2=(((FA1H-E1)**2)/E1)+(((FA1L-E1)**2)/E1)

```

```

        chi2= chi2+(((FA2L-E2)**2)/E2)+(((FA2L-E2)**2)/E2)
    endif
c    print*, "
c    print*, 'Chi2', chi2
    write(1,199) chi2
199  format (f15.5)
999  continue
    stop
    end

```

ii. FORTRAN computer program code to read the breeding values:

```

character*1 gel(48,2),geh(48,2)
character*30 aid(60000),name
integer id(60000)
integer IndH(240),IndL(240),isnp,h,l
real*16 BVH(240),BVL(240)
real*16 gc
open(11,file='GENO.high.txt')
open(12,file='GC.high.txt')
open(1,file='GENO.low.txt')
open(2,file='GC.low.txt')
open(3,file='SNPnumber.txt')
open(10, file='BVH.prn')
open(30, file='BVL.prn')
i=1
11  read (10,*,end=500) IndH(i),BVH(i)
c    print*, IndH(i),BVH(i)
    i=i+1
    go to 11

```

```

500 continue
      j=25
50  read (30,*,end=51) IndL(j),BVL(j)
c   print*,j, IndL(j),BVL(j)
      j=j+1
      go to 50
51  continue
      k=1
12  read (3,*,end=112) id(k),aid(k)
c   print*,id(k),aid(k)
      k=k+1
      go to 12
112 h=0
      na=k-1
1   continue
      nflag=0
      read (11,*,end=78) isnp,((geh(k,h),k=1,24),h=1,2)
      read (12,*,end=78) isnp,gc
78  read (1,*,end=88) isnp,((gel(k,h),k=1,24),h=1,2)
      read (2,*,end=88) isnp,gc
      do ko=1,24
          write(8,1000) isnp,IndH(ko),gc, BVH(ko),geh(ko,1),geh(ko,2)
1000 format (i5,2x,i5, 2(1x,f10.6),2 (a2,1x))
          enddo
          do ko=1,24
              write(8,1000) isnp,IndH(ko),gc, BVH(ko),geh(ko,1),geh(ko,2)
1000 format (i5,2x,i5, 2(1x,f10.6),2 (a2,1x))
              enddo
          do ko=1,24

```

```

      koo=ko+24
      write(8,1000) isnp,koo,gc,BVL(koo),gel(ko,1),gel(ko,2)
      enddo
      go to 1
88    continue
      stop
      end

```

iii. FORTRAN computer program code to determine the analysis of variances (ANOVA) within each SNP for all 96 animals :

```

character*2 GE1(60000),GE2(60000)
integer AG,AC,AT,CG
integer Ind,isnp , isnpa
real*16 BV(60000),gc(60000)
real*16 ngeGG, ngeCC, ngeCG,ngeAG, ngeGA
real*16 ngeAC, ngeGC, ngeAA, ngeTT, ngeAT
real*16 sgeGG, sgeCC, sgeCG, sgeAG, sgeGA
real*16 sgeAC, sgeGC, sgeAA, sgeTT, sgeAT
real*16 sgeGG2, sgeCC2, sgeAG2
real*16 sgeAC2, sgeGC2, sgeAA2, sgeTT2, sgeAT2
real*16 sgt, ngt, F, TE1, TE2, TE3,TE4, TE5, TE6, TE7, TE8
real*16 TTE1, TTE2, TTE3, TTE4, TTE5, TTE6, TTE7, TTE8
real*16 GMGG, GMAA, GMAG, GMAC, GMGC, GMCC, GMTT, GMAT
real*16 GMGG2, GMAA2, GMAG2, GMAC2, GMGC2, GMCC2, GMTT2, GMAT2
real*16 OM, SST, SSE, MST, MSE
real*16 egGG, egAA, egAG, egAC, egGC, egCC, egTT, egAT
open(12,file='input.anova')
c open(12,file='input')
nfla=0

```

```

1  continue

   na=i-1

   nflag=1

   i=1

11 read (12,1000,end=500) isnp, Ind, gc(i), BV(i), GE1(i), GE2(i)

1000 format (i5,2x,i5, 2(1x,f10.6),2 (a2,1x))

   if (i.eq.1) then

c  print*, isnp, Ind, gc(i), BV(i),GE1(i),GE2(i)

   endif

c  if (gc(i).lt.0.15) go to 1

   if (nflag.eq.1)then

       if (GE1(i).eq.' G'.and.GE2(i).eq.' G')then

           sgeGG2=sgeGG2+BV(i)*BV(i)

           sgeGG=sgeGG+BV(i)

           ngeGG=ngeGG+1

       endif

       if (GE1(i).eq.' A'.and.GE2(i).eq.' A')then

           sgeAA2=sgeAA2+BV(i)*BV(i)

           sgeAA=sgeAA+BV(i)

           ngeAA=ngeAA+1

       endif

       if (GE1(i).eq.' G'.and.GE2(i).eq.' A')then

           sgeAG2=sgeAG2+BV(i)*BV(i)

           sgeAG=sgeAG+BV(i)

           ngeAG=ngeAG+1

       endif

       if (GE1(i).eq.' A'.and.GE2(i).eq.' G')then

           sgeAG2=sgeAG2+BV(i)*BV(i)

           sgeAG=sgeAG+BV(i)

```

```

ngeAG=ngeAG+1
endif
if (GE1(i).eq.' A'.and.GE2(i).eq.' C')then
sgeAC2=sgeAC2+BV(i)*BV(i)
sgeAC=sgeAC+BV(i)
ngeAC=ngeAC+1
endif
if (GE1(i).eq.' C'.and.GE2(i).eq.' A')then
sgeAC2=sgeAC2+BV(i)*BV(i)
sgeAC=sgeAC+BV(i)
ngeAC=ngeAC+1
endif
if (GE1(i).eq.' G'.and.GE2(i).eq.' C')then
sgeGC2=sgeGC2+BV(i)*BV(i)
sgeGC=sgeGC+BV(i)
ngeGC=ngeGC+1
endif
if (GE1(i).eq.' C'.and.GE2(i).eq.' G')then
sgeGC2=sgeGC2+BV(i)*BV(i)
sgeGC=sgeGC+BV(i)
ngeGC=ngeGC+1
endif
if (GE1(i).eq.' C'.and.GE2(i).eq.' C')then
sgeCC2=sgeCC2+BV(i)*BV(i)
sgeCC=sgeCC+BV(i)
ngeCC=ngeCC+1
endif
if (GE1(i).eq.' T'.and.GE2(i).eq.' T')then
sgeTT2=sgeTT2+BV(i)*BV(i)

```

```

sgeTT=sgeTT+BV(i)
ngeTT=ngeTT+1
endif
if (GE1(i).eq.' A'.and.GE2(i).eq.' T')then
sgeAT2=sgeAT2+BV(i)*BV(i)
sgeAT=sgeAT+BV(i)
ngeAT=ngeAT+1
endif
nflag=0
isnpa=isnp
go to 11
endif
if (isnpa.eq.isnp) then
c print*, isnpa,isnp,GE1(i)
if (GE1(i).eq.' G'.and.GE2(i).eq.' G')then
sgeGG2=sgeGG2+BV(i)*BV(i)
sgeGG=sgeGG+BV(i)
ngeGG=ngeGG+1
endif
if (GE1(i).eq.' A'.and.GE2(i).eq.' A')then
sgeAA2=sgeAA2+BV(i)*BV(i)
sgeAA=sgeAA+BV(i)
ngeAA=ngeAA+1
endif
if (GE1(i).eq.' G'.and.GE2(i).eq.' A')then
sgeAG2=sgeAG2+BV(i)*BV(i)
sgeAG=sgeAG+BV(i)
ngeAG=ngeAG+1
endif

```

```
if (GE1(i).eq.' A'.and.GE2(i).eq.' G')then
  sgeAG2=sgeAG2+BV(i)*BV(i)
  sgeAG=sgeAG+BV(i)
  ngeAG=ngeAG+1
endif

if (GE1(i).eq.' A'.and.GE2(i).eq.' C')then
  sgeAC2=sgeAC2+BV(i)*BV(i)
  sgeAC=sgeAC+BV(i)
  ngeAC=ngeAC+1
endif

if (GE1(i).eq.' C'.and.GE2(i).eq.' A')then
  sgeAC2=sgeAC2+BV(i)*BV(i)
  sgeAC=sgeAC+BV(i)
  ngeAC=ngeAC+1
endif

if (GE1(i).eq.' G'.and.GE2(i).eq.' C')then
  sgeGC2=sgeGC2+BV(i)*BV(i)
  sgeGC=sgeGC+BV(i)
  ngeGC=ngeGC+1
endif

if (GE1(i).eq.' C'.and.GE2(i).eq.' G')then
  sgeGC2=sgeGC2+BV(i)*BV(i)
  sgeGC=sgeGC+BV(i)
  ngeGC=ngeGC+1
endif

if (GE1(i).eq.' C'.and.GE2(i).eq.' C')then
  sgeCC2=sgeCC2+BV(i)*BV(i)
  sgeCC=sgeCC+BV(i)
  ngeCC=ngeCC+1
```

```

endif
if (GE1(i).eq.' T'.and.GE2(i).eq.' T')then
  sgeTT2=sgeTT2+BV(i)*BV(i)
  sgeTT=sgeTT+BV(i)
  ngeTT=ngeTT+1
endif
if (GE1(i).eq.' A'.and.GE2(i).eq.' T')then
  sgeAT2=sgeAT2+BV(i)*BV(i)
  sgeAT=sgeAT+BV(i)
  ngeAT=ngeAT+1
endif

i=i+1
go to 11
else
i=1
isp=isnpa
isnpa=isnp
backspace 12
endif

c  AG= ngeAA+ ngeGG+ ngeAG
501  continue
if (( ngeAA+ ngeGG+ ngeAG).eq.48)then
if (ngeGG.gt.0)then
GMGG= (sgeGG/ngeGG)
GMGG2= (sgeGG2/ngeGG)
egGG=GMGG2-GMGG*GMGG
egGG=egGG*ngeGG

```

```

endif

if (ngeAA.gt.0)then
GMAA= (sgeAA/ngeAA)
GMAA2= (sgeAA2/ngeAA)
egAA=GMAA2-GMAA*GMAA
egAA=egAA*ngeAA
endif

if (ngeAG.gt.0)then
GMAG= (sgeAG/ngeAG)
GMAG2= (sgeAG2/ngeAG)
egAG=GMAG2-GMAG*GMAG
egAG=egAG*ngeAG
endif

sgt=sgeGG+sgeAA+sgeAG
ngt=ngeGG+ngeAA+ngeAG
OM= (sgt/ngt)

SSE=egGG+egAA+egAG
print*, sgt,OM ,ngt
TE1= GMGG-OM
TE2= GMAA-OM
TE3= GMAG-OM
TTE1= ngeGG*(TE1*TE1)
TTE2= ngeAA*(TE2*TE2)
TTE3= ngeAG*(TE3*TE3)
SST = TTE1+TTE2+TTE3
print*, ' SST =',SST
print*, ' SSE =',SSE

```

```

DFG= 2

DFN= 45

MST= (SST/DFG)

MSE= (SSE/ DFN)

F=(MST/ MSE)

print*, 'DFG=', DFG, 'DFN=', DFN

print*, 'MST=', MST, 'MSE=', MSE

print*, 'F=',F

c  print*, "
   print*, 'SNP ', isp
      print*, "

c  print*, GMGG, GMAA, GMAG
   print*, "

c  print*, GMGG2, GMAA2, GMAG2
   print*, "

c  print*, ngeGG, ngeAA, ngeAG
   print*, egGG, egAA, egAG

c  call sanova (sgeAA, ngeAA, eg1, sgeGG, ngeGG, eg2, sgeAG, ngeAG, eg3)

endif

if (( ngeAA+ ngeCC+ ngeAC).eq.48)then

if (ngeAA.gt.0) then

GMAA= (sgeAA/ngeAA)

GMAA2= (sgeAA2/ngeAA)

egAA=GMAA2-GMAA*GMAA

egAA=egAA*ngeAA

endif

if (ngeCC.gt.0)then

GMCC= (sgeCC/ngeCC)

GMCC2= (sgeCC2/ngeCC)

```

```

egCC=GMCC2-GMCC*GMCC
egCC=egCC*ngeCC
endif
if (ngeAC.gt.0)then
  GMAC= (sgeAC/ngeAC)
GMAC2= (sgeAC2/ngeAC)
  egAC=GMAC2-GMAC*GMAC
  egAC=egAC*ngeAC
endif
sgt=sgeAA+sgeAC+sgeCC
ngt=ngeAA+ngeAC+ngeCC
OM= (sgt/ngt)
SSE=egAA+egAC+egCC
print*, sgt,OM ,ngt
TE2= GMAA-OM
TE4= GMAC-OM
TE6= GMCC-OM
TTE2= ngeAA*(TE2*TE2)
TTE4= ngeAC*(TE4*TE4)
TTE6= ngeCC*(TE6*TE6)
SST = TTE2+TTE4+TTE6
print*, ' SST =' ,SST
print*, ' SSE =' ,SSE
DFG= 2
DFN= 45
MST= (SST/DFG)
MSE= (SSE/ DFN)
F=(MST/ MSE)
print*, 'DFG=', DFG, 'DFN=', DFN

```

```

print*,'MST=', MST, 'MSE=', MSE
print*,'F=',F
    print*,"
print*,'SNP ', isp
print*,"
c  print*, GMAA, GMCC, GMAC
    print*,"
c  print*, GMAA2, GMCC2, GMAC2
    print*,"
c  print*, ngeAA, ngeCC, ngeAC
    print*,egAA,egCC,egAC
c  call sanova (sgeAA,ngeAA,eg1,sgeCC,ngeCC,eg2,sgAC,ngeAC,eg3)
    endif
c  AT= ngeAA+ ngeTT+ ngeAT
    if (( ngeAA+ ngeTT+ ngeAT).eq.48)then
        if (ngeAA.gt.0)then
            GMAA= (sgeAA/ngeAA)
            GMAA2= (sgeAA2/ngeAA)
            egAA=GMAA2-GMAA*GMAA
egAA=egAA*ngeAA
        endif
        if (ngeTT.gt.0)then
            GMTT= (sgeTT/ngeTT)
            GMTT2= (sgeTT2/ngeTT)
            egTT=GMTT2-GMTT*GMTT
            egTT=egTT*ngeTT
        endif
        if (ngeAT.gt.0)then
            GMAT= (sgeAT/ngeAT)

```

```

GMAT2= (sgeAT2/ngeAT)
egAT=GMAT2-GMAT*GMAT
egAT=egAT*ngeAT
endif
sgt=sgeAA+sgeTT+sgeAT
ngt=ngeAA+ngeTT+ngeAT
OM= (sgt/ngt)
SSE=egAA+egTT+egAT
print*, sgt,OM ,ngt
TE2= GMAA-OM
TE7= GMTT-OM
TE8= GMAT-OM
TTE2= ngeAA*(TE2*TE2)
TTE7= ngeTT*(TE7*TE7)
TTE8= ngeAT*(TE8*TE8)
SST = TTE2+TTE7+TTE8
print*, ' SST = ' ,SST
print*, ' SSE = ' ,SSE
DFG= 2
DFN= 45
MST= (SST/DFG)
MSE= (SSE/ DFN)
F=(MST/ MSE)
print*, 'DFG=', DFG, 'DFN=', DFN
print*, 'MST=', MST, 'MSE=', MSE
print*, 'F=',F
print*, "
print*, 'SNP ', isp
print*, "

```

```

c  print*, GMAA GMTT, GMAT
   print*, "
c  print*, GMAA2, GMTT2, GMAT2
   print*, "
c  print*, ngeAA, ngeTT, ngeAT
   print*, egAA, egTT, egAT
c  call sanova (sgeAA, ngeAA, eg1, sgeTT, ngeTT, eg2, sgeAT, ngeAT, eg3)
   endif
CG=ngeGG+ ngeCC+ ngeGC
if (( ngeGG+ ngeCC+ ngeGC).eq.48)then
if (ngeGG.gt.0)then
GMGG= (sgeGG/ngeGG)
GMGG2= (sgeGG2/ngeGG)
egGG=GMGG2-GMGG*GMGG
egGG=egGG*ngeGG
endif
if (ngeCC.gt.0)then
GMCC= (sgeCC/ngeCC)
GMCC2= (sgeCC2/ngeCC)
egCC=GMCC2-GMCC*GMCC
egCC=egCC*ngeCC
endif
if (ngeGC.gt.0)then
GMGC= (sgeGC/ngeGC)
GMGC2= (sgeGC2/ngeGC)
egGC=GMGC2-GMGC*GMGC
egGC=egGC*ngeGC
endif
sgt=sgeGG+sgeGC+sgeCC

```

```

ngt=ngeGG+ngeGC+ngeCC
OM= (sgt/ngt)
SSE=egGG+egGC+egCC
    print*, sgt,OM ,ngt
TE1= GMGG-OM
TE5= GMGC-OM
TE6= GMCC-OM
TTE1= ngeGG*(TE1*TE1)
TTE5= ngeGC*(TE5*TE5)
TTE6= ngeCC*(TE6*TE6)
SST = TTE1+TTE5+TTE6
print*, ' SST = ' ,SST
print*, ' SSE = ' ,SSE
DFG= 2
DFN= 45
MST= (SST/DFG)
MSE= (SSE/ DFN)
F=(MST/ MSE)
print*, 'DFG=', DFG, 'DFN=', DFN
print*, 'MST=', MST, 'MSE=', MSE
print*, 'F=',F
print*, "
print*, 'SNP ', isp
print*, "
c  print*, GMGG, GMCC, GMGC
    print*, "
c  print*, GMGG2, GMCC2, GMGC2
    print*, "
c  print*, ngeGG, ngeCC, ngeGC

```

```
c  print*,egGG, egCC,egGC
c  call sanova (sgeGG,ngeGG,eg1,sgeCC,ngeCC,eg2,sgGC,ngeGC,eg3)
endif

TE1=0
TE2=0
TE3=0
TE4=0
TE5=0
TE6=0
TE7=0
TE8=0

TTE1=0
TTE2=0
TTE3=0
TTE4=0
TTE5=0
TTE6=0
TTE7=0
TTE8=0

sgt=0
ngt=0
OM=0
SSE=0
SST=0
DFG=0
DFN=0
MST=0
MSE=0
F=0
```

egGG=0

egAA=0

egCC=0

egAG=0

egGC=0

egAC=0

egTT=0

egAT=0

ngeGG=0

ngeAA=0

ngeAG=0

ngeAC=0

ngeGC=0

ngeCC=0

ngeTT=0

ngeAT=0

sgeGG=0

sgeAA=0

sgeAG=0

sgeAC=0

sgeGC=0

sgeCC=0

sgeTT=0

sgeAT=0

sgeGG2=0

sgeAA2=0

sgeAG2=0

sgeAC2=0

sgeGC2=0

```
sgeCC2=0
sgeTT2=0
sgeAT2=0
i=i+1
if (nfla.eq.1) stop
go to 11
500 continue
nfla=1
go to 501
end
```