Norwegian University of Life Sciences
Faculty of Veterinary Medicine and Biosciences
Department of Chemistry, Biotechnology and
Food Science

Master Thesis 2015
30 credits

Evaluation of GWAS Method Performance Focusing on Population Stratification and Cryptic Relatedness

Yonatan Ayalew Mekonnen

Table of Contents

Acknowledgment

First, I would like to thank the Almighty GOD for sending miracles and helping me reach my goal. I thank his mother Holy Virgin Mary for being right beside me all the way here.

This thesis would not have been possible without the advices and guidance from my supervisor, professor Torgeir Rhoden Hvidsten. Your constructive comments have helped me a lot in shaping my research and have given me a great lesson, thank you so much! Next, I would like to thank my friend Teshome Mulugeta for the ideas and discussions we shared.

I also forward my gratitude for my spiritual family at the Ethiopian Orthodox Church for their prayers and thoughts during my study. A word of appreciation also goes to the Ethiopian community in Ås who have warmed my stay during my study. Biniyam, Lidya, Amare, Tesfaye, Aweke and Diakon Gebreyohannes, thank you very much for the wonderful time we had together. I also want to thank Diakon Zebene for keeping company.

Next, a sincere gratitude goes to my father, Ayalew Mekonnen and my mother Aselefech Haile for their love and support from the start of my education. My siblings, Meron, Yishak, Tenagne, Alemtsehay, Dereje and Endalkachew, thank you so much for your care.

Special thanks goes to my wife Bethelhem Legesse. I am so grateful having you in my life and thank you for your support, patience and encouragements. My little angle Rediet, you were not born when I was enrolled as a master student and then you came and sing your refreshing songs. Thank you so much both of you for making me happy.

Abstract: Genetic association studies are primarily used to identify genes associated with complex disease. It can be conducted by genotyping intentionally selected or randomly chosen markers. Numerous statistical and computational algorithms have been developed in the past to analyze the genome wide association study (GWAS) dataset. These are classified as parametric, non-parametric and Bayesian methods. However, there are methodological and computational challenges related with population stratification and the vast volume of data generated by chip and sequencing based technologies. The packages, SNPRelate and GenABEL, are built to overcome this burden. SNPRelate uses parallel computing and loads genotypes block by block to optimize high-speed cache memory. It is designed for principal component analysis (PCA) and identity by descent (IBD) analyses which are used for correcting population structure. Whereas, GenABEL incorporates genome wide rapid association using mixed model and regression (GRAMMAR). It is developed to overcome the limitation of efficiently storing, handling and analyzing data in GWAS by integrating a data format called gwaa.data. In order to evaluate and compare these packages, this study obtained PLINK formatted data from heritable dog osteosarcoma study. PLINK data format is then changed into a genomic data structure (GDS) file format for SNPRelate and gwaa.data file for GenABEL. Using GenABEL, data analysis was performed by ignoring population structure and taking into account population structure. In SNPRelate, LD based pruning is performed prior to PCA and IBD calculation. For three dog breeds, the first and the second PCs have almost 50% of the information. IBD interpretation of PCA indicate that Irish wolfhounds are inbred compared to the other two dog breeds. PCA correction on population structure has the most accurate estimates compared with genomic control and PCs as a predictor correction methods. Comparing SNPRelate and GenABEL, SNPRelate method used for PCA calculation is faster and allows larger data sets than GenABEL which use EIGENSTAR for PCA calculation.

Keywords: GenABEL, GWAS, IBD, SNPRelate, parallel computing, PCA, population structure.

1. Introduction

Genetic association studies are primarily used to identify genes controlling susceptibility to complex disorder. This can be accomplished by testing the correlation between disease status (phenotype) and genetic variation (genotype). Initially, disease genes were identified by genotyping affected families by using genetic markers across the genome and evaluating the segregation of genetic markers across multiple families (pedigree). This approach is called genome wide linkage analysis and was preliminarily used to identify disease genes which follow a monogenic (*i.e*. a trait that is controlled by a single gene) type of Mendelian inheritance [14]. These variants have low frequency due to natural selection. However, they have high penetrance and the markers within 10-20cM of the actual disease causing allele will co-segregate with diseases eminence [8]. Genome wide linkage analysis has a limitation to detect genetic variants that has modest effect on the disease. In other words, the linkage analysis approach has a weakness when it comes to detecting alleles that have low penetrance. Candidate gene resequencing approach is a practical alternative to linkage analysis. In this analysis, genes are selected based on linkage or other evidence associated with the trait (disease) for further study. Then, the selected genes are resequenced using disease and control groups. Candidate genes are obtained by comparing the disease and control groups for the richness or deleted variants in the disease cases. However, this approach is laborious and expensive.

Now a days, Genome wide association studies (GWAS) are usually used to carry out association studies [8]. In association studies, single nucleotide polymorphism markers (SNPs) are predominantly used, but other markers also exists such as microsatellites, insertions/deletions, tandem repeats (VNTRs) and copy number variants (CNVs). In the past years, the vast volume of data generated in chip and sequencing based GWAS had faced significant challenges in analytical and computational processing.

Genome wide association study

GWAS analysis is performed by examining the genome for causal genetic variants without prior information of the location of these variant. GWAS can be conducted by genotyping intentionally selected or randomly chosen markers (SNPs) in a case-control population [8]. The corrected p-value (*i.e.* significance measure by false positive rate) is then computed for each statistical test. The marker (SNP) should pass the significant threshold in order to have a significant association with the trait of interest (*i.e*. an association of a single locus with a trait). This approach is considered to be unbiased and reliable since it does not require prior knowledge regarding the function and/or location of the causal genes (see Figure 1).

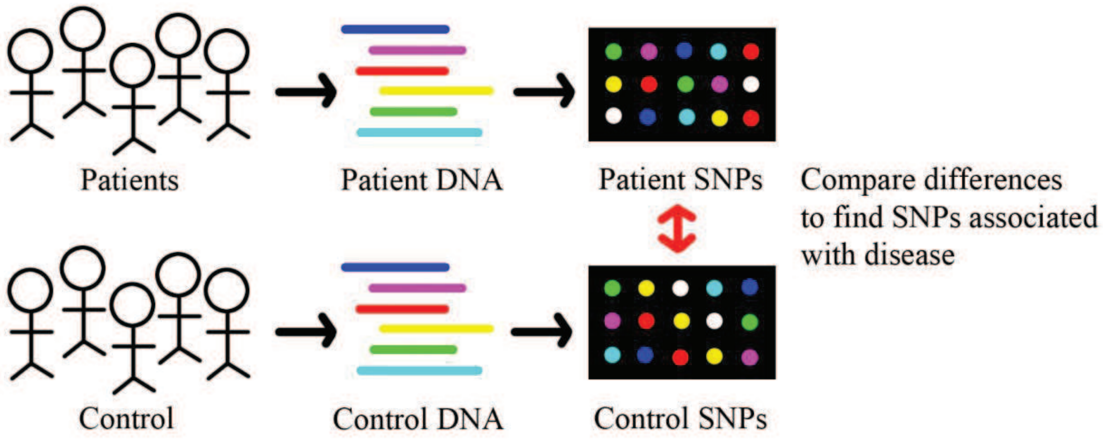

Figure 1. GWAS used to test the association between a SNP and a trait of interest (e.g. Disease). (http://cubocube.com/dashboard.php?a=344&b=462&c=1)

There are direct and indirect association of SNPs with a given trait of interest. The first type is when the genotyped SNPs are directly associated with the trait. The second type is that the genotyped SNPs are not directly associated with the trait rather they act as a tag SNPs, that is, a representative SNP for a genomic region where influential SNPs are located. In other words, the tag SNP and influential SNP are in linkage disequilibrium (LD). LD measures the degrees of association between two loci. Depending on the distance of LD, mapping at centimorgan (cM) for long distance or base pair gene distance for short LD could be applied. Because of these two types of associations between SNPs and traits, significant SNPs identified by genome wide association studies (GWAS) are not considered actual variants. That is why the results of GWAS require additional procedures to map the precise location of actual SNPs [14]. It has been shown that identifying disease genes using association studies is more powerful than linkage studies [9].

Genome wide association studies have been widely and successful used to identify common genetic variant associated with complex traits. To analyze GWAS datasets, there have been numerous statistical procedures and computational algorithms developed in the past decade classified by three fundamental statistical methods. These are parametric, non-parametric and Bayesian methods. Among many parametric models, logistic regression is dominantly used for the detection of interacting gene effects for dichotomous traits (*i.e.* the traits that take an either/or form but not both. E.g. sick/healthy). Combinatorial partitioning method is among the most commonly used non-parametric methods, which is used for detecting quantitative traits by partitioning of multi-locus genotypes based on the corresponding inter-individual variation. Bayesian methods are used to model and test interactions among SNPs for case/control study. However, this method is not used for higher-order interaction due to its computational burden of Monte Carlo Markov Chain algorithms (*i.e.* largely Bayesian analysis depends on) and sample size [1].

Basically, GWAS performs scanning by testing each marker individually [11]. In other words, traits are analysed separately (univariate analysis) by searching for signal of association at a specific loci across the studied traits. However, multivariate (MV) approach (multiple correlated traits) could be beneficial for several reasons. Multivariate analysis provides cross-trait covariance information due to genetic correlation between different traits. In addition, multivariate analysis reduces the burden of analysing all traits individually since it can perform a single test for association with multiple traits [4, 5, 6]. It is also widely believed that a single genetic variant could be associated with multiple traits which lead to the conclusion that multivariate GWAS is more appropriate in a biological context compared to univariate approach [3, 10]. Individual loci may also interact to control a certain trait *epistatically*. The R package VariABLE is developed to analyze interacting loci by applying the variance heterogeneity test [23]. Some of the multivariate and univariate methods and applications are listed in Table 1 [2].

Table 1. Some of the multivariate and univariate methods and applications.

Methods	Application	Output
Multi variate (MV)-PLINK	Use additive model	F-statistic and p-value
MV-SNPTEST	Use method called “expected”	Expected genotype counts (dosages)
MultiPhen	Use likelihood ratio test (LRT)	p-value per trait and p-value for LRT
MV-BIMBAM	Use two different approaches: 1) testing for association between multivariate traits, all partitioned in the group of directly affected traits and genotype; and 2) considers all different possible partitions of traits into different categories of traits (directly affected, indirectly affected, unaffected).	Summarized by log10 Bayes Factor (BF) that evaluates presence of any Multivariate Genome-Wide Association between QTL and trait
PCHAT	Use splitting in a training set and test set. In addition, so called ‘bagging’ is performed, in which bootstrap samples are drawn from training sample and optimal linear combination of traits is averaged across bootstrap samples.	Association result is expressed as p-value.
TATES	Requires correlation matrix. Fitting linear models	p-value corrected for traits correlation.
Univariate meta-analysis (UV_MA) and univariate principal component analysis (UV-PCA).	Uses univariate results per trait as input files and use p-values direction of effect as input for meta-analysis. PCA performed. Using first PC in univariate analysis	Overall z-statistic and p-value

Generally, association studies can be classified in to two types: single locus association study and multiple locus association study [7].

Single locus test

A statistical test is conducted to analyze each SNP individually for the association to a phenotype. Different statistical tests are required based on whether the traits are quantitative or dichotomous (case/control). If the traits are quantitative, the generalized linear model (GLM) approach, usually analysis of variance (ANOVA), is applied. For dichotomous traits (case/control), logistic regression is often used [12]. Genotypic data can be encoded to test association between allele and phenotype (*i.e.* allelic association) or genotype and phenotype (*i.e.* genotypic association). Genotype classes could also be modelled as dominant, recessive, multiplicative or additive [13].

Let us consider two alleles, T and t, for a dominant model. The presence of one or two copies of T allele could increase risk of getting T allele controlling character. But for a recessive model, only two copies of T could increase the risk. For the multiplicative model, for example, if 4x is the value of T allele controlling character then for two copies of T allele, there is 16x. This means that for Tt, there is k value of T allele controlling character and for TT, there is K^2 character values. When considering additive model, if 4x is for Tt, then TT would have 8x. This means that the risk for having T controlling character for Tt is K and for TT is 2K. Among these genotypic models, the additive form is commonly practiced in GWAS.

Multi-locus test

Multi-locus testing approaches require the examination of every pair-wise combination of SNPs for association with the trait. Basically, multilinear regression (*i.e*. a multivariate analysis approach which models trait values as a function of autonomous variable vectors corresponding to genotypes of multiple loci) is used in multilocus association study. This approach is computationally challenging even when applying efficient algorithms. To tackle this problem, SNPs are filtered based on their results from single SNP analysis. The significant SNPs in the single SNP analysis are used to find interactions. However, this approach would undermine the role of epistatic loci, (specially those alleles with marginal effects individually and could not be detected by a statistical test) since the subsets are selected based on their main effect. Limiting the analysis to SNPs that are involved in a biological network such as biochemical pathways or protein families is another approach to detect interactions and is referred to as bio-filtering approach. This approach uses different types of publicly available data sources for screening. For testing interaction, logistic regression is used most commonly in several statistical methods such as INTERSNP and multifactor dimensionality reduction (MDR) [12].

However, there are methodological and computational challenges related to creating robust statistical model for association studies in complex trait. Specially, when dealing with larger data sets, population stratification and scaling problem remains a challenge for the computation infrastructure. The more preferred way to deal with these issues is splitting the problems into smaller parts (parallelization), sending each to different CPUs and finally combining the results (out puts) together [20].

Population stratification and covariance analysis

The test statistics could be affected by factors like age, sex and geography. Covariate adjustment should therefore be applied to minimize the effect of such confounding factors. Usually, in GWAS analysis, there is lack of a full genealogy (*i.e.* traces of lines of decent) of the population due to population structure, family structure and cryptic relatedness. If the population and sample structures (family structure and cryptic relatedness) are not properly corrected in the model, GWAS may face a significant number of false positives. Genomic Control (GC) is one of the methods to handle the problems of population stratification. However GC has limitation due to other confounders such as family structure and cryptic relatedness. Structured Association (SA) and Principal Component Analysis (PCA) are among other approaches to correct false positives due to stratification. Now a day, combining the three methods (GC for adjusting residual inflation, SA for removing closely related sample and PCA for correcting broad sample structure) has become the preferred approaches by some researchers. In human population, allele frequency is significantly different across subpopulations (ethnicity) [12]. In order to avoid population stratification, the method STRUCTURE/EIGENSTART is used to compare allele frequencies to HapMaps subpopulations. The samples would be excluded if similarity is found or covariate analysis could be conducted [12].

Multiple testing correction approaches

Bonferroni correction is used to change the threshold value (α) = 0.05 in which p-value is measured against, into α/k (0.05/k) where k is the number of statistical tests performed. This approach is, however, considered as highly conservative since it assumes that markers are independent and ignores linkage disequilibrium among markers. False discovery rate (FDR) is an alternative approach to adjust α which controls the proportion of false positives [15]. Another complimentary approach is permutation tests in which the phenotypes of each individual are reassigned into another individual by altering the genotype-phenotype maps of the data. Each reassigned steps are considered as one possible sampling and the process is repeated N times. Software packages such as PLINK, PRESTO and PERMORY are developed to do permutation tests. Genome wide significance notion is another approach which is commonly used. This approach is based on linkage disequilibrium (LD) information. The number of autonomous genomic regions would therefore determine the number of corrected statistical test for hypothesis testing at the genomic level [12].

Linear mixed model (LMM) approaches for association studies

Mixed model approaches have been applied in linkage analysis [16]. The model was initially developed for animal model. The Variance components of the genetic effects are additive and polygenic effects which is expressed as:

$$\text{У} = \mu + \alpha + g + e$$

where μ is overall mean, ɑ is additive genetic effects, g is polygenic effects and e is residual effects. However, with larger data set and sample size, it becomes difficult to apply variance components for random effect estimation. In order to tackle this problem, LMM based approaches were implemented in GWAS and the model is:

$$\text{У} = X\beta + g + e$$

where X is the matrix of fixed effect (overall mean, covariance, SNPs), g and e are polygene and residual effects, respectively. The variance of g is dependent on kinship matrix, Var (g) = $K\delta_g^2$ and K denoted kinship matrix quantifying genetic similarity across individuals. Therefore, population structure, family structure and cryptic relatedness are included in K. LMM based approaches applied in GWAS is used to correct false positive inflation and it could be applied for both single and multi-loci analysis [17].

In this thesis, we evaluate the performance of R packages SNPRelate and GenABEL. The goal is to evaluate and compare these packages on their population stratification and cryptic relatedness dealing performance.

SNPRelate:

Since SNPRelate is primarily designed to do PCA and IBD analysis, it is provided with the GDS data format to run efficiently. The package gdsfmt and SNPRelate has advantages compared to previous methods in terms of efficient data storage technique and implementation for PCA and IBD analysis. One of the challenges in GWAS analysis is the computational burden due to big data size for data processing and memory limitation. For instance, in PLINK, all SNP genotypes has to be loaded into memory and it could be the main limitation for PLINK analysis. However, SNPRelate overcomes this problem by allowing access to data as needed without loading all data into memory. SNPRelate use parallel computing and have an R interface to utilize high speed memory cache by blocking the computations. This means that the algorithms in SNPRelate packages are optimized to load genotypes block by block without the limitation of the number of SNPs (bearing in mind the limitation of main memory). These packages are developed to facilitate principal component and identity by descent (IBD) analysis in general.

GenABEL:

GenABEL uses EIGENSTRAT that incorporates SA and genomic kinship matrix for adjusting possible population stratification. For larger data set analysis involved in GWAS, there is a need to store, handle and analyze the data efficiently in addition to correcting population structure. In standard R data, GWAS data storage is not efficient. GenABEL, which implements genome wide rapid associations using mixed model and regression (GRAMMAR) [23], is developed to overcome such limitation by integrating a special data format called gwaa.data for efficient data storage, handling and for fast GWA analysis for case–control data. Since R is supported by a wide-range of statistical analysis and graphical facilities, developing GenABEL as an R library enables to facilitate not only the analysis of GWAS, but also result presentation supported by graphs and figures as well [21, 22].

2. Methods

SNPRelate

The package gdsfmt, which is needed to load SNPRelate, is used to provide efficient memory usage and file management independent of the platform. SNPRelate is used to perform principal component analysis (PCA) and identical by descent (IBD) (*i.e.* similarity of alleles due to the same ancestry) calculations which are numerically intensive. The algorithms kernels are written in C/C++. PCA is a statistical method used to convert a set of observations described by several dependent variables (correlated variables) in to a set of new orthogonal variables (*i.e.* linearly uncorrelated variables) called principal components [18]. This means, it identifies PCs based on genetic correlations among individuals representing the population [30]. PCA analysis has two purposes. First, PCA is used to classify the data in the way that reflects the internal structure of the data according to how much of the information they have explained and stored in the data. Second, PCA is used to reduce the number of variables into a smaller set of components while maintaining the data variability. However, PCA might not give an optimal solution. Since it is a dimension reduction technique, it will lose information if too few principal components are used. Therefore, as an alternative method, hierarchical clustering analysis is proposed to determine clusters. Hierarchical clustering analysis is based on the individual dissimilarity which is directly related to co-ancestry coefficient (kinship coefficient). Agglomerative clustering algorithm is used for the analysis based on individual dissimilarity (distance). The average dissimilarity between individuals is used to draw a tree of the dissimilarity between clusters.

Zheng *et al*., 2012 provided an alternative interpretation of PCA based on relatedness measure as the probability of set of genes which are identical-by-descent (descended from a single ancestral origin (gene)). Hence, Population structure could also be adjusted by pair wise relatedness analysis (i.e. identical by descent (IBD) analysis). To do identical by descent calculation, the reference population is needed. Using allele frequency, in order to estimate the relatedness of the individuals in the population, is analogous to changing the reference population back in time. For relatedness analysis, maximum likelihood estimation (MLE) and method of moments (MoM) are commonly used in a homogeneous population.

For our analysis, data were obtained from heritable dog *osteosarcoma* study (http://www.broadinstitute.org/ftp/pub/vgb/dog/OSA_GenomeBiology2013paper). We used 169,010 SNPs and

543 samples (267 greyhounds, 135 Rottweilers and, 141 Irish wolfhounds) with almost equal male-female and case-control proportion. All three dog breeds are genetically distinct populations. To do the analysis, PLINK data formats was changed in to GDS file format. Then, linkage disequilibrium (LD) based SNP pruning was applied to filter SNP that are in linkage equilibrium. For the diagnosis and correction of population stratification, fixation index (f_{st}), identity by state (IBS) (*i.e.* identical alleles but have no identical origin) and PCA was applied. The performance of efficient memory usage and speed was compared with other methods.

Data formats

For the purpose of efficient memory usage, the gdsfmt package uses the genomic data structure (GDS) file format to store annotation data and SNP genotypes. This file format is able to encode up to four SNP genotypes in each byte and therefore reduces file size and the time required to access data. In the GDS file format, only the data that is being analyzed is retained in memory since it is supported with data blocking. Data blocking can be defined as an algorithm used to analyze the data structure by preventing interference from other processes. It is an optimization technique that reduces usage of memory bandwidth by allowing full cache use [19]. The raw data format used in the analysis was PLINK binary file format. In order to process the data with SNPRelate, the PLINK file format had to be changed into the GDS file format. The function snpgdsBED2GDS provided by SNPRelate is used to convert PLINK files into GDS files format.

Data analysis

After the data conversion, linkage disequilibrium (LD) based pruning of SNPs was performed to evade SNP clusters in PCA and IBD calculation (see Figure 2). In the dataset paper, they have indicated that 98% of SNPs are in LD. Therefore, we used LD threshold of 0.98. For PC analysis, the genetic covariance matrix was calculated from genotypes followed by creating correlation coefficients between sample loadings and genotypes for individual SNP. Then SNP eigenvectors (loadings) of the new dataset was approximated after SNP eigenvectors (loadings) was calculated (see Figure 2; flow chart for computing). For the first 16 PC components, the percentage of variation explained by them was estimated. Plots for PCA were performed using the highest scoring eigenvectors. Plots were also made to show the correlation between eigenvectors and SNP genotypes. Fixation index (f_{st}) was calculated by the method of Weir & Cockerham (1984) to measure the degree of differentiation between case and control population. IBD calculation was performed using both method of moments (MoM) by Purcell *et al*., 2007 and maximum likelihood estimation (MLE) by Milligan, 2003; Choi *et al*., 2009 for relatedness analysis. MLE are more accurate than MoM. But it is slow compared to MoM due to its computational burden. Identity by state (IBS) estimation was also performed using individuals in the sample by creating an nxn matrix of averaged genome wide IBS pair wise identity using the snpgdsIBS function.

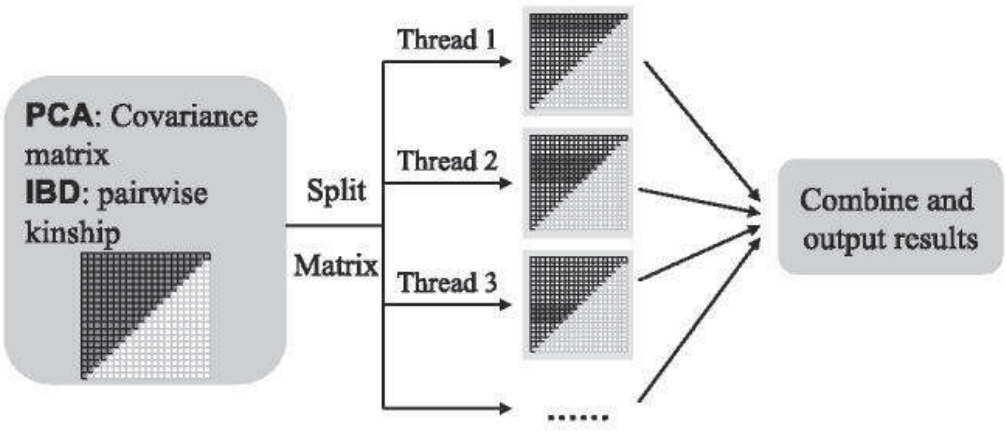

Figure 2. Parallel computing flow chart of PCA and IBD analysis [28]

GenABEL

The association tests are carried out using the package GenABEL in R and data was obtained from heritable dog *osteosarcoma* study (http://www.broadinstitute.org/ftp/pub/vgb/dog/OSAGenomeBiology2013paper). We used 184 genotyped SNPs for 432 samples with equal proportion of male and female. That is, 124 SNPs in greyhounds (174 cases and 110 controls) and Rottweilers (64 cases and 32 controls); and 60 SNPs in Irish wolfhounds (22 cases and 30 controls). GenABEL was tested in the presence of population stratification for its efficiency of storage, handling and fast analysis of GWAS data. In order to detect and adjust population stratification; genomic control, multi dimensional scaling (MDS) and PCA were used for comparison purpose. Since the data type was PLINK formatted, it had to be converted into GenABEL raw format using the convert.snp.ped function. However, this dataset lack 'sex' as a variable and the GenABEL converting function requires this variable[1]. To solve this problem, the 'sex' variable is created at random in the phenotype dataset but not used for the analysis. The converted file, which belongs to the gwaa.data class, is developed to facilitate GWA analysis and is used to store GWA data. After the PLINK data format conversion, gwaa data is loaded into R using load.gwaa.data function.

[1] In the dataset paper [27], as they have stated in their analysis, they did not detect any significant association between sexes. Therefore, they exclude the variable in the dataset they have uploaded.

Data analysis was performed first by ignoring the presence of population stratification (*i.e.* the presence of allele frequency difference between populations due to ancestral difference). A genome scan was performed using the glm() function which implements a maximum likelihood estimation (MLE) method which is computationally intensive. For genome wide significance, we use α= 0.05 (95% confidence interval) rather than Benferroni correction. Because Benferroni correction is highly conservative for SNPs tested in dog breeds due to extensive LD occurrence [27]. Association tests taking into account population structure is more preferable since we have three different dog breeds as one population. Therefore, correcting the population structure using components (PCs) as a predictor is one of the methods. Both scanning methods, glm and qtscore was applied and the results were compared. The general linear model (GLM) parameters are estimated by MLE and hence glm scan is slow compared to qtscore. The second method for correcting population structure is genomic control in which it uses corrected p-values (*i.e.* uncorrected p-value multiplied by the number of comparisons) for test statistics. However, it is not recommended to use this method for admixed population (*i.e.* population with mixed ancestry) due to its conservative nature. The third method chosen to correct population structure is PCA. In order to apply this method, GenABEL integrate EIGENSTRAT which enables to test the association along with correcting population structure. The implementation is performed using ‘egscore’ function and plots are drawn. The comparison between PCs as a predictor and correction with PCA methods were made.

3. Results

Three dog breeds; Irish wolfhounds, greyhounds and Rottweilers were used in our analysis. All three dog breeds are genetically distinct population.

SNPRelate

In our analysis using SNPRelate, PCA and IBD analysis were performed in genomic SNP data.

PCA analysis using SNPRelate

As it is shown in figure3, the three dog population are genetically distinct and the variation of top two PCs are; in the upper right corner (Irish wolfhounds) has high values for both components whereas the upper left population (Greyhounds) has relatively higher values in component one (comp1) compared to the lower left population (Rottweilers) which has higher values for comp2 only. This could be interpreted as Irish wolfhounds breed is more inbred compared to the other two breeds. The correlation between SNP genotypes and eigenvectors are also shown in figure 6.

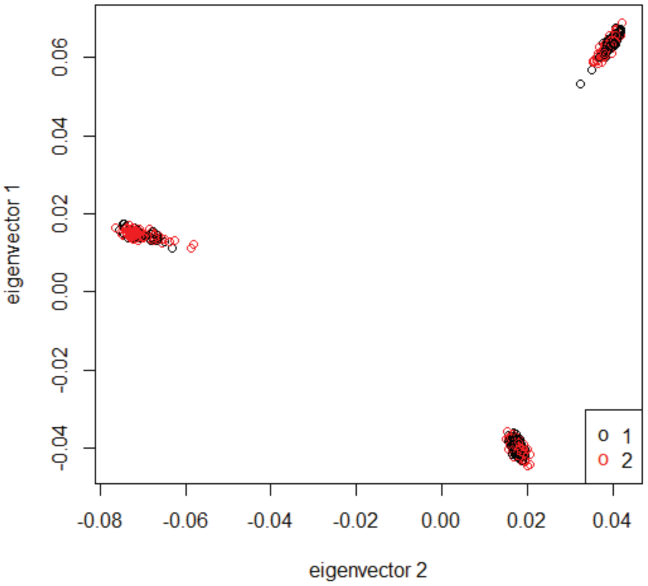

Figure 3. Principal component analysis using the first two eigenvectors where 1(black) is control and 2 (red) is case.

Looking into the first 6 components, the proportion of variance explained from component 1 to 6 is; 15.84, 14.49, 0.57, 0.53, 0.46, 0.45. The total variance explained by them is less than 33% of the total. However, the first and the second principal components account for the largest proportion of variance as shown in figure 4.

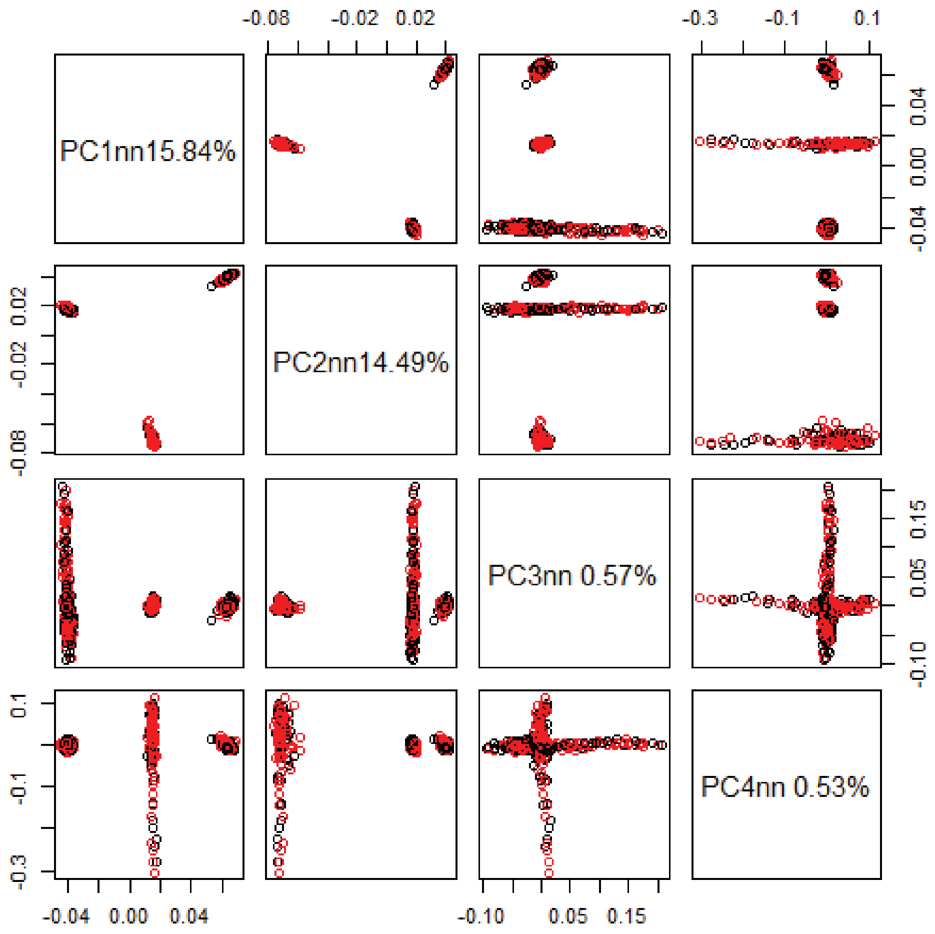

Figure 4. Principal component analysis of 543 samples. Pairwise plots of the first four eigenvectors and proportion of variance explained by each is given along the diagonal.

The first two components contain more than 50% of the information as seen in figure 5. The other components explain a smaller proportion (for example, comp 3 shown in figure 6). Therefore, it is sensible to reduce the dimensions in two dimensions by choosing comp 1 and 2.

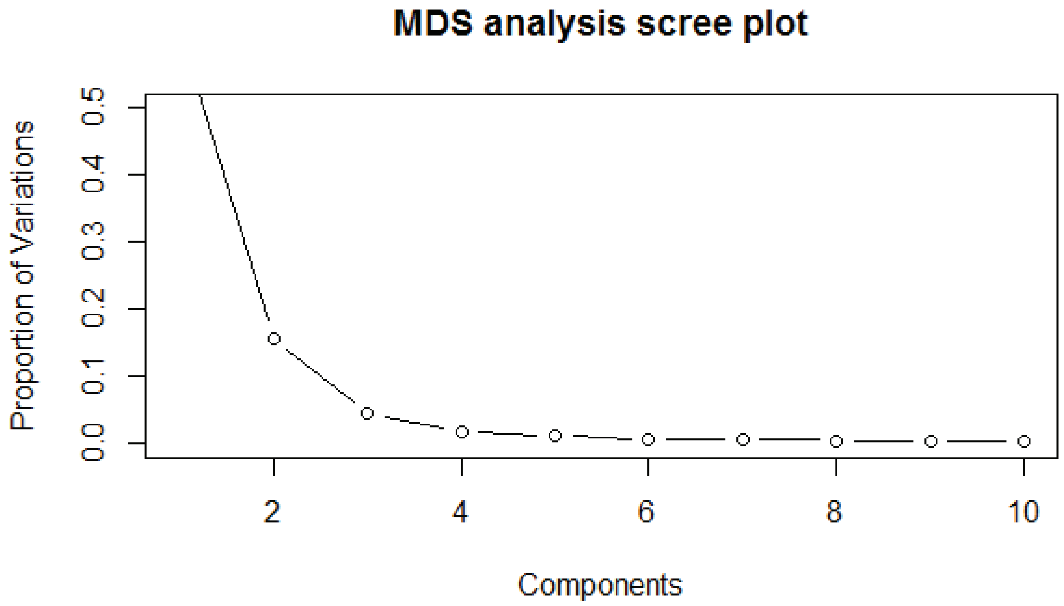

Figure 5. Scree plots of the number of components explaining the proportion of variation.

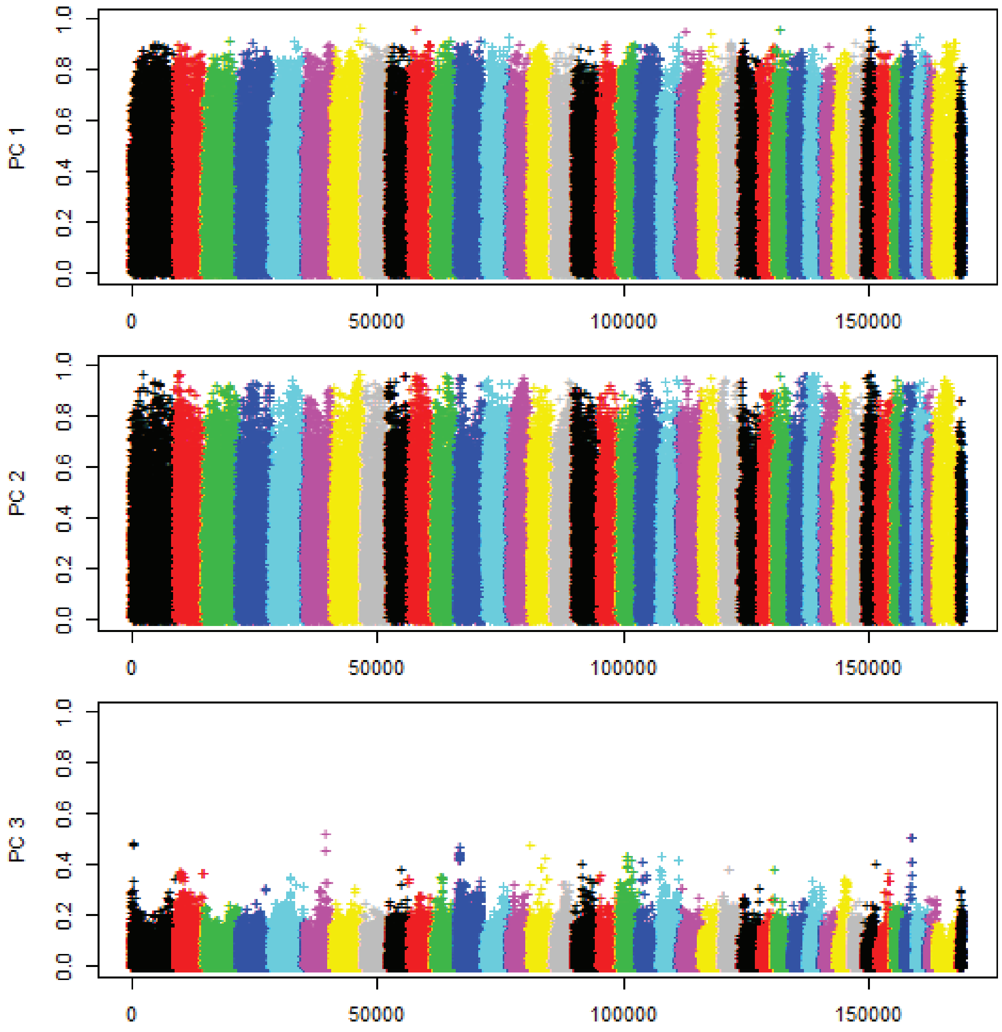

Figure 6. The correlation between SNP genotypes and eigenvectors.

When comparing the running time of these two methods, SNPRelate is relatively faster than GenABEL as shown in table 2. GenABEL took half of the running time of SNPRelate to calculate PCA for 184 markers (fewer marker sets) and 432 samples whereas SNPRelate takes twice the running time of GenABEL for 169,010 SNPs (larger marker sets) and 543 samples[2]. This means that SNPRelate is faster than GenABEL since it only doubles the time required for PCA analysis by GenABEL while using very large number of SNP sets.

[2] We used a pedigree file for GenABEL in which the number of markers is usually less than the number of subjects since only few markers are typed.

Table 2. Running time of SNPRelate and GenABEL on dual-core Intel processor (2.4GHz and 4GB RAM) where m, s and ms are minute, second and millisecond respectively.

Methods	Runing time
SNPRelate	00m.12s.86ms
GenABEL	00m.06s.35ms

Fixation index (f_{st}) estimation was 9.85232e-06 which implies that the two populations case and controls are interbreeding freely (no evidence to support that the two populations do not share any genetic diversity).

Hierarchical Clustering Analysis

Hierarchical clustering analysis was conducted using the full set of SNPs and 543 individual based on individual dissimilarity matrixes. Different colours (black-Rottweilers, red-Irish wolfhounds, and green-greyhounds) represent different populations (breeds) as shown in figure 7.

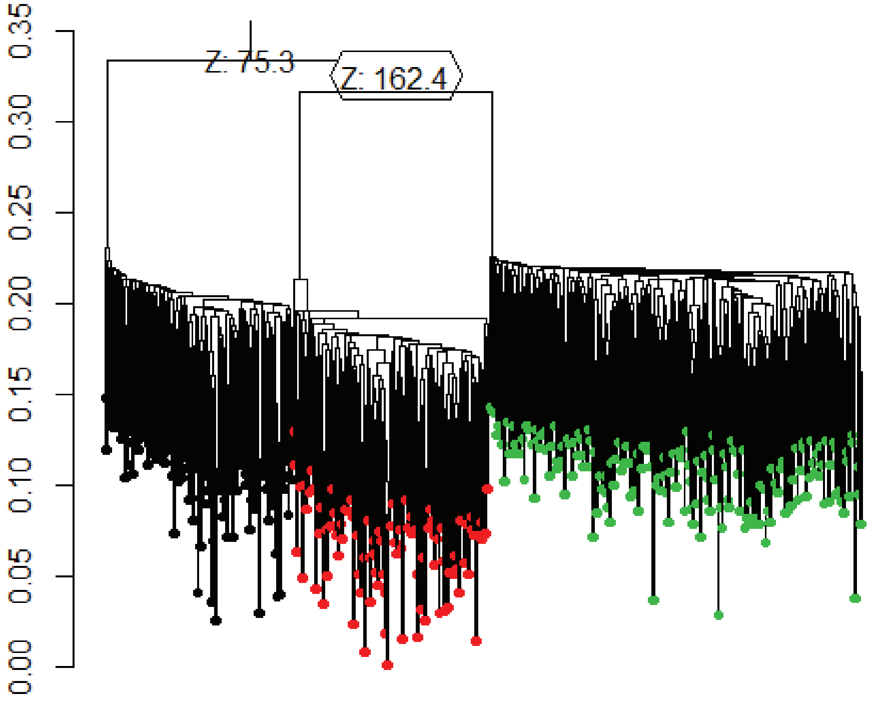

Figure 7. Hierarchical cluster analysis of the three dog breeds.

Hierarchical clustering is a complimentary method to PCA and has more power for clustering analysis than PCA since PCA is a dimension reduction method and might lose information.

GenABEL

Correcting population using genomic control and PCA

There are 30 significant loci found before genomic control using both glm and qtscore (uncorrected p-values) (blue circles) with genomic inflation factor(λ) of 3.5 as shown in Figure 8 . After genomic control (corrected p-values), 16 loci are significant and λ is 1 (green circles in Figure 8).

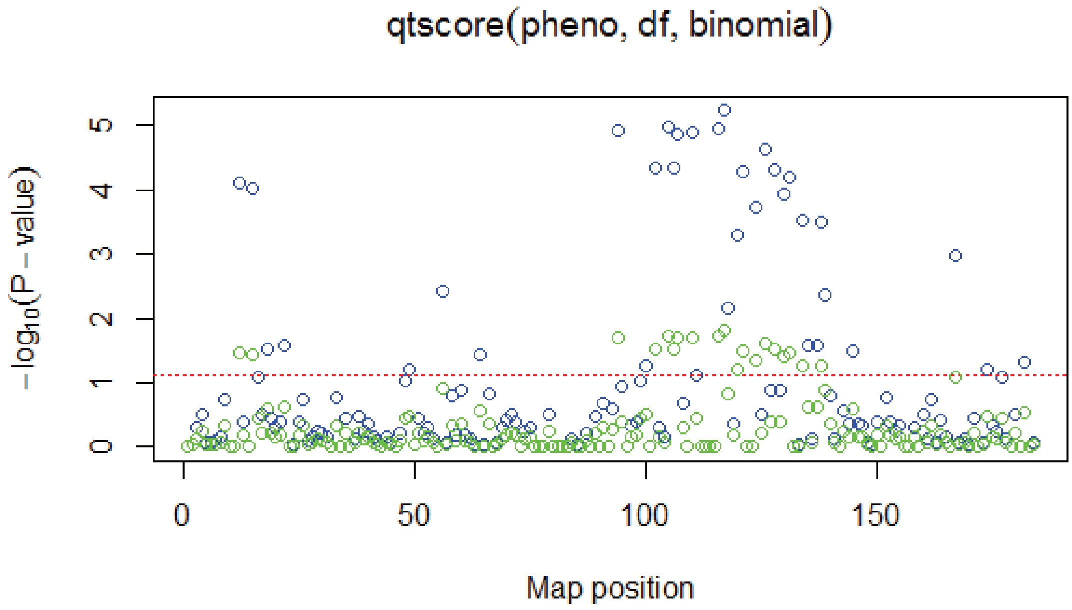

Figure 8. –log_{10}(P-value) of GWAS scan using raw data (blue circles) and after genomic control (green circles) (red line is the threshold value (p=0.05)).

PCA correction for population structure was performed using markers that are not in linkage disequilibrium (LD). EIGENSTRAT, built-in GenABEL, is able to test the association along with correcting population structure. Therefore, 33 loci are found to be significant using PCA corrected population with λ of 3.16 as shown in figure 9. PCs as a predictor correction method was also applied and 8 loci are found significant by using glm scan and 2 loci using qtscan with λ of 1.

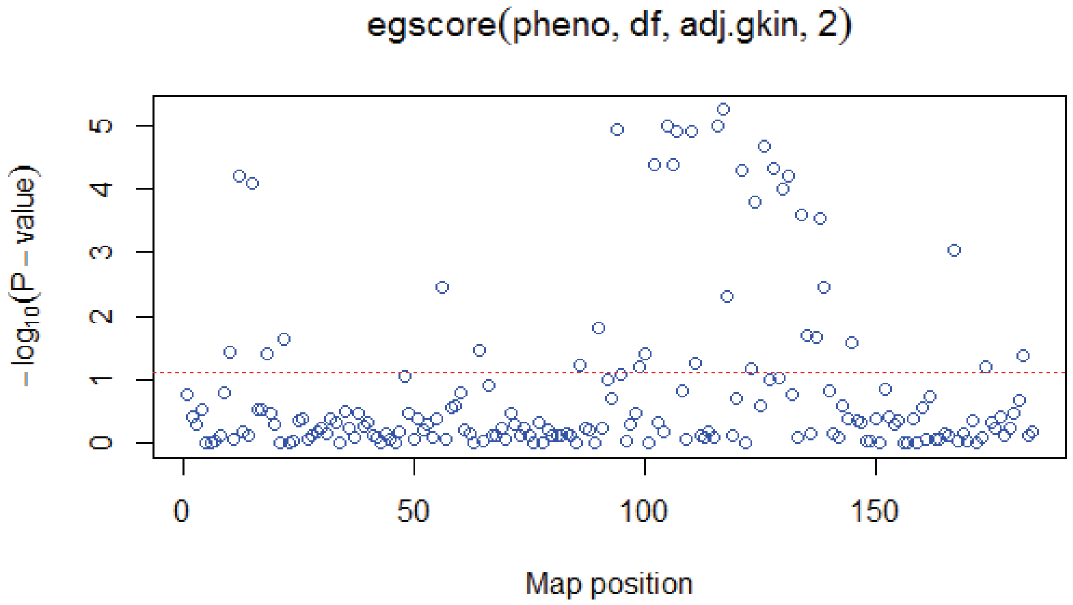

Figure 9. $-log_{10}$*(P-value) GWAS scan using PCA corrected population structure (red line is the threshold value (p=0.05))*

4. Discussion

Up to the current application, GWAS is the primary tool widely used to investigate and analyze the genetic architecture of a disease or a trait. Since genome wide analysis has involved numerous computations and applications, a faster and efficient algorithm is needed to carry out the task given. PCA and IBD analysis is two methods that reduce the dimensions in order to address false positive associations due to the presence of population structure and cryptic relatedness. However, PCA analysis is confronted with a computational burden mainly on larger sample and SNP analysis which requires efficient numerical implementation and memory management. In order to solve this limitation, Xiuwen Z. *et.al.*, 2012, developed R packages; gdsfmt, for the efficient memory and file management independent of the platform; and SNPRelate, for efficient GWAS calculations for PCA and IBD.

In SNPRelate, the calculation for covariance matrix and pair wise IBD are performed on multi-core multiprocessing computer simultaneously without overlapping as shown in Figure 2. These packages are advantageous for loading genotypes block by block without limiting the number of SNPs. However, the size of the main memory could be the limiting factor which holds covariance matrix or IBD coefficient matrix. The performance of SNPRelate was compared with PCA and IBD calculating algorithms, EIGENSTART and PLINK. Our result was consistent with earlier works by Xiuwen Z. *et.al*, 2012, in which the performance of PCA and IBD was faster in SNPRelate compared to GenABEL which incorporates EIGENSTART for PCA calculation (see table 2). The reason why SNPRelate is faster than EIGENSTART is that it uses multi-threaded local alignment search for eigenvector and eigenvalue calculations whereas EIGENSTART use uniprocessor. This would increase the computational performance for larger number of sample size. SNPRelate is also unique for extracting sample and SNP loadings while correcting for population stratification [4]. In addition, SNPRelate performs genotype-PC correlation in order to test whether a local region of the genome reflects the correlation structure [28]. However, except the difference on the speed of calculation, EIGENSTART and SNPRelate have the same accuracy [28].

The genomic interpretation of PCA in terms of relatedness is the reflection of the probability of gene sets that are identical by descent (IBD). This means that based on the relatedness measures, PCA can be interpreted as the probability of set of genes which are identical-by-descent (descended from a single ancestral origin (gene)) [28]. In our analysis Irish wolfhounds is more inbred compared to the other two breeds and our result is consistent with the work of Karlsson, E.K. *et al*., 2013.

The occurrence of large proportion of false positive associations in GWAS analysis could be tackled by the implementation of PCA for diagnosis and correction of population structure and IBD for relatedness diagnosis between pair of samples. However, for a larger data set analysis involved in GWAS, there is a need to store, handle and analyze the data efficiently in addition to correcting population structure. In standard R data, GWAS data storage is not efficient. GenABEL is developed to overcome such limitation by integrating a special data format called gwaa.data for efficient data storage and handling and qtscore for fast GWA analysis for case–control data. GenABEL is also able to perform data quality control (QC) and analysis faster than the previous methods. During QC analysis, using PCA correction for population stratification has the most accurate estimation compared to incorporating PCs as a predictor and genomic control. Because, using PCA correction, 33 loci has been identified which is the same as the dataset paper[27].

Incorporating PCs as a predictor with smaller λ adjusts for genotypes only whereas PCA correction adjusts both genotypes and phenotypes for PCs and calculates their correlation after applying correction. This makes 'PCs as a predictor' method less accurate although it has smaller λ than PCA correction. When we look the genomic control test statistic inflation control, it uses the value of the observed test statistics divided by the genomic inflation factor (λ) with corrected p-value (Pcd1df). For λ calculation, previous analysis uses the ratio of median observed X^2 and expected X^2 test statistics. However, GenABEL uses the ratio of regression coefficient (slope) of observed X^2 and expected X^2 which makes it a bit conservative. Due to this nature, genomic control is not recommended to use for admixed population; it may not correct the population efficiently.

5. Conclusion

Advancement in chip and sequencing based technologies has created a tsunami of data where one needs to have a robust statistical model to do the analysis. In addition, an efficient memory use is also required to withstand the wave. Although GenABEL is efficient for its fast QC, data analysis and memory use compared to previous methods, it incorporates EIGENSTART for PCA calculation. Sticking on the first two PCs would then reduce the number of variables which is critical to avoid the problem of multicollinearity, large standard errors and inaccurate prediction caused by maintaining all covariates. A systematic selection of number of variables into a smaller set of variables while maintaining the data variability is also reduces computational burden. The methodology SNPRelate used for PCA calculation is faster and allows much larger data sets than EIGENSTART. Therefore, incorporating SNPRelate methodology in to GenABEL for correcting population structure and cryptic relatedness would enhance the performance of GenABEL in the future.

6. References

1. Gao, H., Wu, Y., Li, J., Li, H., Li, J., & Yang,. (2014). Forward LASSO analysis for high-order interactions in genome-wide association study. *Briefings in Bioinformatics* 15(4):552–561.

2. Galesloot, T.E., van Steen, K., Kiemeney, L.M., Janss, L.L., & Vermeulen, S.H. (2014). A Comparison of Multivariate Genome-Wide Association Methods. *PLoS ONE* 9(4): e95923. doi:10.1371/journal.pone.0095923

3. Chavali S., Barrenas, F., Kanduri, K., &., Benson, M. (2010). Network properties of human disease genes with pleiotropic effects. *BMC Systems Biology* 4: 78.

4. Zhu, W., & Zhang, H. (2009). Why Do We Test Multiple Traits in Genetic Association Studies? *Journal of the Korean Statistical Society* 38: 1–10.

5. Allison, D.B., Thiel, B., St Jean, P., Elston, R.C., &., Infante, M.C. & Schork, N.J. (1998). Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *The American Journal of Human Genetics* 63: 1190–1201.

6. Klei, L., Luca, D., Devlin, B., & Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology* 32: 9–19.

7. Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* 33: 497–507.

8. McCarthy, M.l., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little J., loannidis, J.P., & Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Rev. *Genet.* 9, 356–369.

9. Ott, J.m Kamatani, Y., & Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics* 12:465.

10. Büchel, F., Mittag, F., Wrzodek, C., Zell, A., & Gasser, T., & Sharma, M. (2013). Integrative Pathway-Based Approach for Genome-Wide Association Studies: Identification of New Pathways for Rheumatoid Arthritis and Type 1 Diabetes. *PLoS ONE* 8(10): e78577. doi:10.1371/journal.pone.0078577

11. Cantor, R.M., Lange, K. & Sinsheimer, J.S. (2010) Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am J Hum Genet* 86: 6–22.

12. Holzinger, E. R., Dudek, S. M., Frase, A. T., Pendergrass, S. A., & Ritchie, M.D. (2013). ATHENA: the analysis tool for heritable and environmental network associations. *Bioinformatics* 30 (5): 698-705.

13. Bush, W.S., & Moore, J.H. (2012). Genome-Wide Association Studies. *PLOS Computational Biology* 8(12): e1002822. doi:10.1371/journal.pcbi.1002822

14. Lewis, C.M. (2002). Genetic association studies: design, analysis and interpretation. *Briefings Bioinformatics* 3(2):146-53.

15. Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine* 9:811-818.

16. Goldgar, D.E. (1990). Multipoint analysis of human quantitative genetic variation. *The American Journal of Human Genetics* 47: 957-967.

17. Li, G., & Zhu, H. (2013). Genetic Studies: The Linear Mixed Models in Genome-wide Association Studies. *The Open Bioinformatics Journal* 7: 27-33.

18. Abdi. H., & Williams, L.J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: *Computational Statistics* 2: 433–459. doi:10.1002/wics.101

19. Hummel, S.F. (1990). SMARTS - Shared-memory Multiprocessor Ada Run Time Supervisor. Technical Report 495.

20. Cavuoti, S., Garofalo, M., Brescia, M., Pescape, A., Longo, G., & Ventre, G. (2013). Genetic Algorithm Modeling with GPU Parallel Computing Technology. Neural Nets and Surroundings, Proceedings of 22nd Italian Workshop on Neural Nets, WIRN 2012; Smart Innovation, Systems and Technologies, Vol. 19: 29-39, Springer. doi: http://link.springer.com/chapter/10.1007%2F978-3-642-35467-0_4

21. Aulchenko, Y.S., Ripke S., Isaacs A., & van Duijn C.M. (2007). GenABEL: an R package for genome-wide association analysis. *Bioinformatics* 23(10):1294-6.

22. Aulchenko, Y.S., de Koning, D.J., & Haley, C. (2007). Genomewide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. *Genetics* 177(1):577-85.

23. Struchalin, M.V., Amin, N., Eilers, P.H., Duijn, C. M. & Aulchenko, Y.S. (2012). An R package "VariABEL" forgenome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity. *BMC Genetics* 13(4). doi:10.1186/1471-2156-13-4

24. Cockerham, W. and Clark, C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38 (6): 1358. doi:10.2307/2408641. ISSN 0014-3820.

25. Purcell. S., Neale, B., Todd, B. K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., Bakker, P.I., Daly, M.J., & Sham P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses*. Am J Hum Genet* 81: 559–575.

26. Milligan, B.G. (2003). Maximum-likelihood estimation of relatedness. *Genetics* 163: 1153- 1167.

27. Karlsson, E.K., Sigurdsson, S., Ivansson, E., Thomas, R., Elvers, I., Wright, J., *et.al*. (2013). Genome-wide analyses implicate 33 loci in heritable dog osteosarcoma, including regulatory variants near CDKN2A/B. *Genome Biology* 14: R132. doi:10.1186/gb-2013-14-12-r132

28. Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C. and Weir, B.S., 2012. A high performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinfomatics*. 28: 3326–3328. doi:10.1093

29. Choi, Y., Wijsman E.M., Weir, B.S., 2009. Case-Control Association Testing in the Presence of Unknown Relationships. *Genetic Epidemiology* 33 (8): 668–78.

30. Wang, D., Sun, Y., Stang, P., Berlin, J. A., Wilcox, M. A., & Li, Q., 2009. Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. BMC Proceedings, 3(Suppl 7):S109. doi: 10.1186/1753-6561-3-S7-S109

Appendix

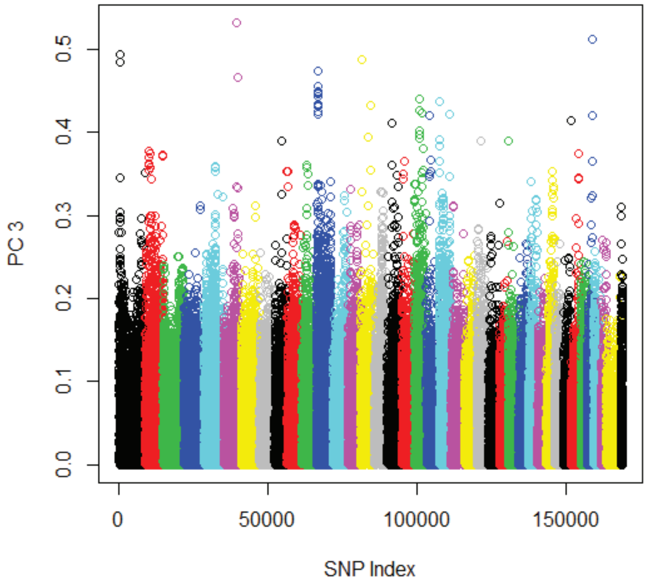

Figure A1. The correlation plot between eigenvector and genotype representing genome wide correlation from PCA joint ancestry analysis.

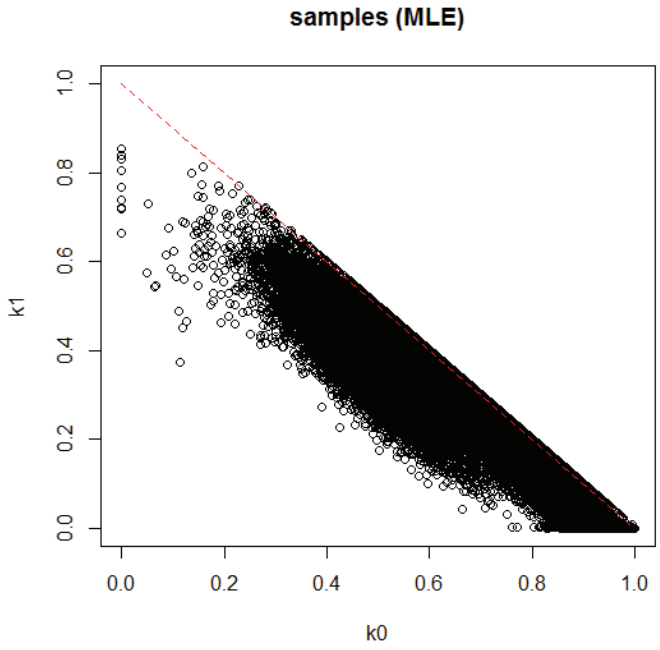

Figure A2. Relatedness estimates of all three dog breeds using IBD coefficient by MLE method. The black circle represents pair of samples.

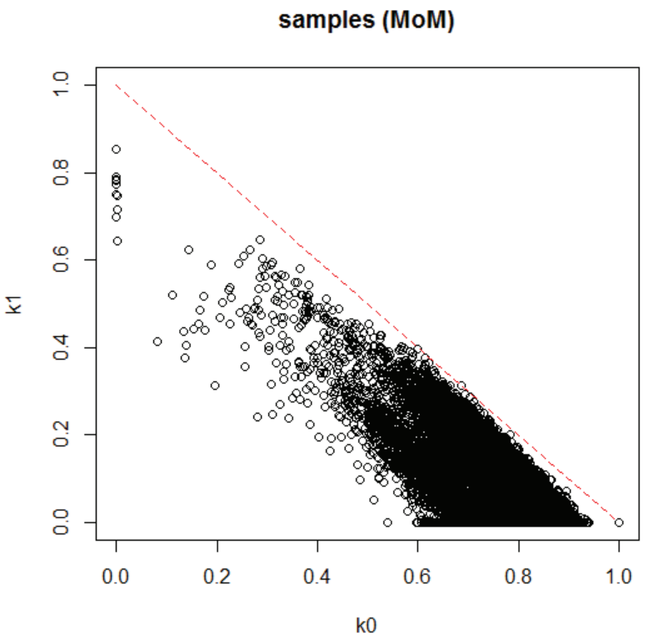

Figure A3. Relatedness estimates of all three dog breeds using IBD coefficient by MoM method. The black circle represents pair of samples.

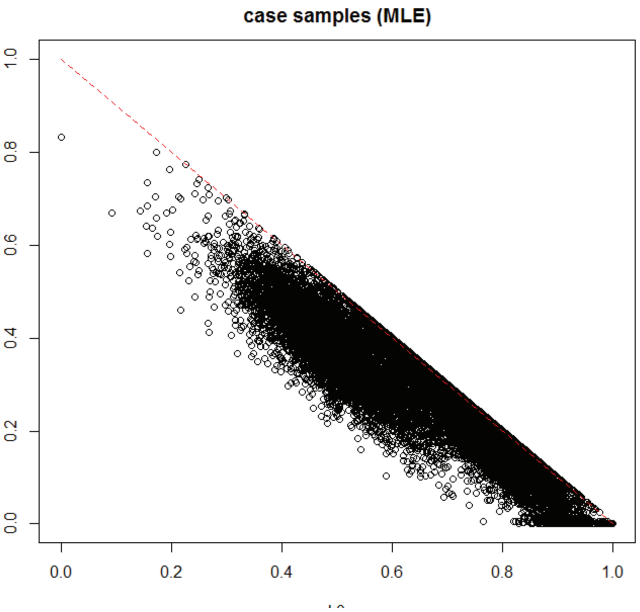

Figure A4. Relatedness estimates of all case dog breeds using IBD coefficient by MEM method. The dots represent pair of samples.

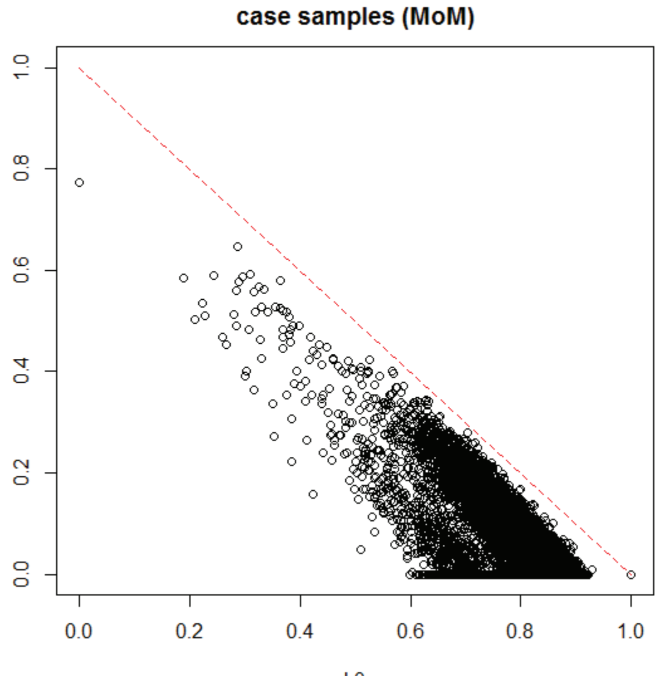

Figure A5. Relatedness estimates of all case dog breeds using IBD coefficient by MoM method. The dots represent pair of samples.

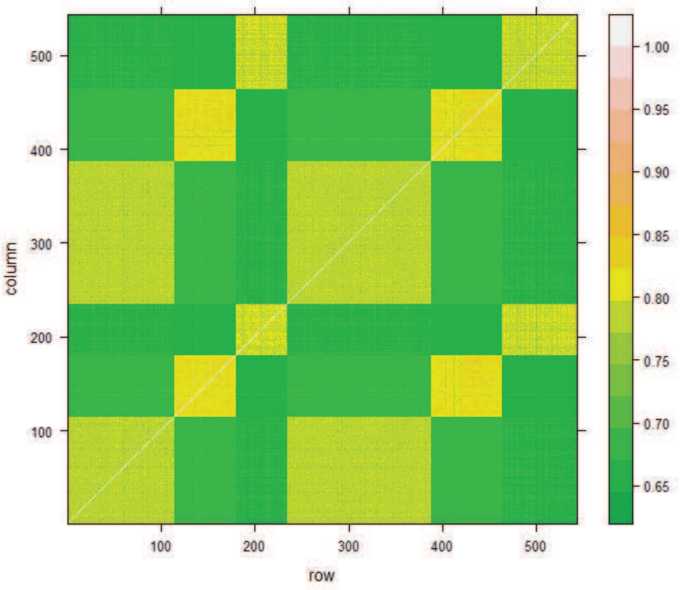

Figure A6. Heat plots of IBS. The extent of IBS increases across the color gradients (from green to red).

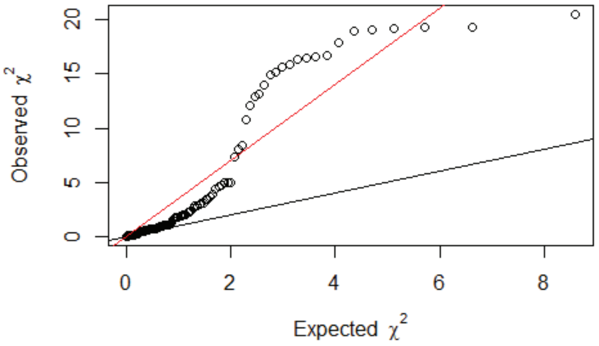

Figure A7. Q-Q plot before population stratification correction applied (black line is slop (assuming no inflation) and red line is fitted line).

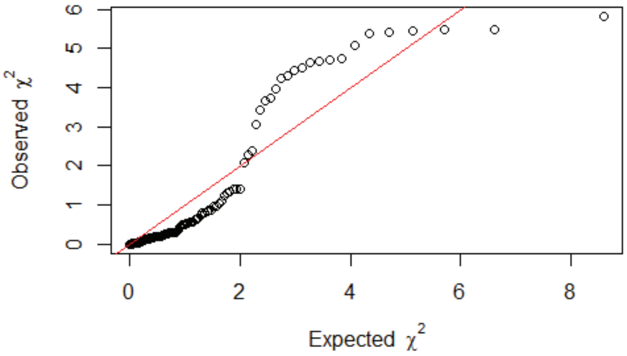

Figure A8. Q-Q plot after population stratification correction applied using genomic control.

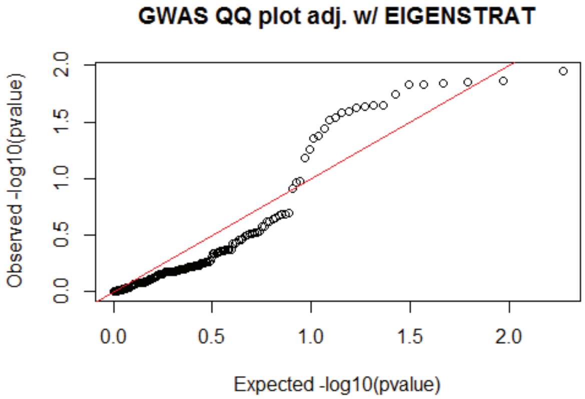

Figure A9. Q-Q plot after population stratification correction applied using PCA.

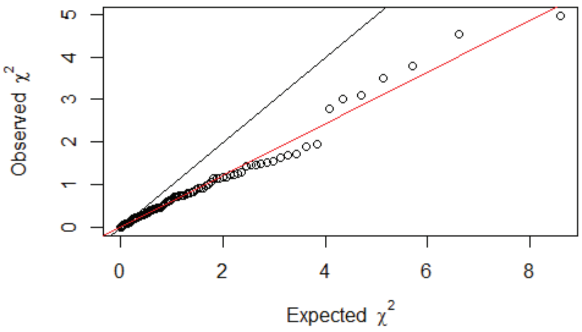

Figure A10. Q-Q plot after population stratification correction applied using PCs as a predictor.

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no