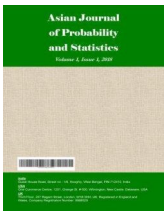


Asian Journal of Probability and Statistics

1(4): 1-31, 2018; Article no.AJPAS.44107

Computational Generalization of Mixed Models on Large-Scale Data with Applications to Genetic Studies

Samson W. Wanyonyi[1], Drinold A. Mbete[2*] and Emile R. Chimusa[3]

[1]Department of Mathematics, University of Eldoret, Box 1125-30100 Eldoret, Kenya.
[2]Department of Mathematics, Masinde Muliro University of Science and Technology, Box 190-50100, Kakamega, Kenya.
[3]Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Medical School Cape Town, South Africa.

Authors' contributions

This work was carried out in collabaration between all authers. Author SWW designed the study, performed the statistcical analysis, wrote the draft of the manuscript. Authors DAM and ERC managed the analysis of the study and literatures searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2018/v1i424550
Editor(s):
(1) Dr. Halim Zeghdoudi, Department of Mathematics, Badji-Mokhtar University, Algeria.
Reviewers:
(1) Essien Archibong Okon, Cross River University of Technology, Nigeria.
(2) Olfa Ben Braïek, University of Monastir, Tunisia.
Complete Peer review History: http://www.sciencedomain.org/review-history/26606

Original Research Article

Received: 15 July 2018
Accepted: 30 September 2018
Published: 10 October 2018

Abstract

Aims: To discuss different LMM-based approaches applied in GWAS and software packages implementation and Classify different computational tools that applies LMM approaches according to their applicability and performance. To identify possible SNPs associated to a particular disease using different computational tools based on LMM approaches. To estimate genetic and residual variance parameters that account phenotypic variation of the disease.
Study Design: Case control study
Place and Duration of Study: The research was carried out in Tanzania at African Institute of Mathematical Science for six months.
Methodology: Linear Mixed Models (LMMs) are widely applied in genomic wide associations studies (GWAS) owing to their effectiveness of correcting hidden relationship, population structure and family structure. This essay is aimed at exploring different mathematical approaches of LMMs in GWAS. These approaches are linear mixed model with inclusion of all markers (LMMi) and linear mixed model with

**Corresponding author: E-mail: dmbete@mmust.ac.ke;*

exclusion of all markers (LMMe) when calculating genetic relationship matrix. LMMi is more efficient as compared to LMMe when applied in studies of randomly ascertained quantitative traits. The LMM approaches are classified based on their applicability and performance. Two computational GWAS tools namely, PLINK and EMMAX were used which were based on LMM approaches to analyze unpublished real data from West Africa (Gambia and Ghana). Genetic and residual variance parameters were estimated that accounted for the phenotypic variation of the disease to be 0.0594 and 0.0723. A total of 338408 variants and 959 people (484 males, 405 females and 70 missing phenotypes) pass filters and quality control using PLINK was used in the study. Among the remaining phenotypes, 864 are cases and 95 are controls. The performance of different mathematical approaches of LMMs and their software implementation, including EMMAX and Plink via the application to a GWAS of tuberculosis (TB) in 959 individuals in West Africa (Ghana and Gambia) was compared. Of these 864 cases of TB and 95 healthy individuals retained after quality control (QC) using Plink, and 329601 autosome single nucleotide polymorphisms (from chromosome 1 to chromosome 22) included in the analysis after 288 duplicands ID individuals removed after QC. The LMM approaches are classified based on their applicability and performance. Two computational GWAS tools, namely Plink and EMMAX were used in the analysis of data. Genetic and residual variance parameters were estimated that accounted for the phenotypic variation of the disease to be 0.0594 and 0.0723.

Results: Result showed that SNPs associated with tuberculosis were $rs7225581$ on chromosome 17 and SNP $rs491412$ on chromosome 13 with both having 0.69% false discovery rate with step up significance value. Plink failed to correct hidden relatedness. Although EMMAX reduced the false positive rate, it still exhibited very low presence of stratification.

Conclusion: This study aimed at understanding and exploring different approaches of mixed models as applied in genetic studies. Overview of genetic variation, advantages, successes and application of mixed models and current challenges of mixed models in GWAS were discussed. Moreover, the study showed that SNPs was associated with a particular disease using computational tools that applies LMM approaches. The summary statistics from PLINK and EMMAX found two causal SNPs associated with the TB. These SNPs were rs7225581 on chromosome 17 and SNP rs4941412 on chromosome 13 with both having 0.69% FDR H. However, PLINK failed to correct hidden relatedness. This phenotypic variation showed that all common single nucleotide polymorphisms (SNPs) expressed approximately 18.52% of phenotypic variation of the disease.

Keywords: Genomic wide association studies (GWAS); mixed linear models approaches; single nucleotide polymorphism (SNP).

1. Introduction and Background

1.1 Overview

A Linear Mixed Model (LMMs) is an extension of the standard linear regression. For formulation and estimation, we consider one of the simplest LMMs [1]:

$$Y = X\beta + Z\mu + \varepsilon \tag{1.1.1}$$

where $Y = (y_1, \ldots\ldots, y_n)^T$ is vector of response on n subjects, X is $n \times p$ matrix of predictor values on the subjects, $\beta = (\beta_1, \ldots\ldots, \beta_n)^T$ is a vector of (unknown) coefficients of X, Z is $n \times K$ matrix of predictor values associated with random effect, $\mu = (\mu_1, \ldots\ldots, \mu_n)^T$ is a vector of random effects and ε is a vector of random errors, where $\mu \sim N(0, \sigma^2{}_\mu)$ and $\varepsilon \sim N(0, \sigma^2{}_e)$ [2,3]. The formulation of this model for continuous variation in human population has been useful in advancing the human genetics variation [4,5]. The model partitions the variation in a quantitative trait into three sources [6]: the effect of a single major gene, residual additive heritable effects of polygenic loci[1], and the independent random effects of the environment.

[1] *polygenic loci refers to any individual locus which is included in the system of genes responsible for the genetic component of variation in a quantitative (polygenic) character*

According to Guo and Thompson [6], model (1.1.1) can be used to unravel the genetic basis of quantitative traits such as height, weight and blood pressure. The model initially used to explain how the genetic component of a quantitative trait, such as weight, height, blood pressure is correlated between relatives [3]. In genetic association studies, extension of this LMM has been studied in order to estimate heritability of traits, capture genetic relatedness and breed values of individuals and location of quantitative trait loci (QTL) [3,7]. This LMM has gained popularity in testing for association in genome-wide association studies (GWAS) because of their demonstrated effectiveness in accounting for relatedness among samples and in controlling population stratification [2] and other confounding factors [8,9,10]. It tackles confounders using measures of genetic similarity to capture the probabilities that the pairs of individuals have causative alleles in common [11], and such measures include those based on identity by descent and the realised relationship matrix (RRM) [12]. According to Jiang et al. [10], the model (1.1.1) is primarily designed for quantitative traits, and when applied to case control data, it imposes a misspecified model on the binary phenotype, which can lead to power loss. This loss of power is due to several model violations dependence between variant used and tested causal variants in estimating kinship dependence between genetic and environmental effects and use of non-continuous traits [13,14]. The use of generalized linear mixed models (GLMMs) resolves this problem of sensitivity due to confounding, but leads to a different difficult issue of estimating parameters due to high dimensions [15].

1.2 Key concepts in Genetic Variation

Genetic variation can be defined as the difference between individuals or the differences between population in terms of trait such as height, skin or color. The variation is a result of subtle differences in DNA caused by either mutation, gene flows, sexual reproduction or genetic drift [3] [16]. This causal relationship between genetic polymorphism within a species and phenotypic differences observed between individuals has been of fundamental biological interest for researchers in the field of GWAS [17]. GWAS has been used to identify causative/predictive factors for a given trait (i.e. the number of loci that contribute and their respective contribution to the phenotype) but it fails in the case of rare variants that cannot be detected due to confounding [18]. The other methods that have been used includes genomic control[4], quantitative trait loci (QTL) [5], principal component analysis (PCA) [6] and LMM. All these methods fail in the case confounding expected LMM that incorporate pairwise genetic relatedness between every pair of individuals in the statistical model directly, reflecting that the phenotypes of two similar individuals are more likely to be correlated than genetically dissimilar individuals [13]. **Linkage disequilibrium (LD):** This is defined as non-random association of alleles at different loci that occurs when genotypes at two loci are not independent of another [19]. **Heritability**: The variance parameter that estimates how much variation in a phenotypic trait in a population is due to genetic variation among individuals in that population. It can also be defined as the extent to which genetic individual differences contribute to individual differences I n observed behaviour (or phenotypic individual differences). The total phenotypic variance $\sigma^2{}_P$ within a population is the sum of genetic variance $(\sigma^2{}_G)$ and environment variance $(\sigma^2{}_\epsilon)$, that is $\sigma^2{}_P = \sigma^2{}_G + \sigma^2{}_\epsilon$. Where $\sigma^2{}_G$ is given as $\sigma^2{}_G = \sigma^2{}_g + \sigma^2{}_D + \sigma^2{}_I$, $\sigma^2{}_g$ is the additive (polygenic) genetic variance, $\sigma^2{}_D$ dominance variance and $\sigma^2{}_I$ is gene-gene interaction (epistatic) variance. Epistatic variance involves interaction between alleles at different loci. A dominance variance is due to non-linear interactions between alternative alleles at the same locus, while additive genetic variance is due to inheritance of a particular allele and depicts the effects of individual alleles on the phenotype. Heritability is grouped into two categories. Broad-sense heritability (H^2) which is defined as the ratio of total genetic variance to total phenotypic variance,

[2] *population stratification is defined as the differences in allele frequencies between cases, and controls due to systematic differences in ancestry rather than association of genes with disease [18].*
[3] *genetic drift is random fluctuations in allele frequencies over time due to sampling effects, particularly in small populations.*
[4] *Genomic control is defined as a method of detecting stratification based on the genome-wide inflation of association statistics.*
[5] *QTL is a section of DNA (the locus) that correlates with variation in a phenotype (the quantitative trait) [20]*
[6] *PCA is a dimensionality reduction technique used to infer continuous axes of variation in genetic data, often representing genetic ancestry.*

$$H^2 = \frac{\sigma^2{}_G}{\sigma^2{}_P} \tag{1.2.1}$$

and narrow- senses heritability (h^2) which is the ratio of additive genetic variance to the total phenotypic variance,

$$h^2 = \frac{\sigma^2{}_g}{\sigma^2{}_P}. \tag{1.2.2}$$

Genotypes-Phenotype -: Genotype is the allelic constitution of an individual. It is part of DNA sequence of the genetic makeup of the cell that gives rise to observable traits (phenotype) of an organism [7]. Different alleles or forms in a genotype are produced by mutations to DNA, and may result into beneficial or detrimental changes. Phenotypes are all the observable traits (characteristics) of an organism [13], usually with emphasis on traits controlled by the genes under examination. The phenotypic observation on every animal is determined by environmental and genetic factors [7] as illustrated in Fig. 1.1.

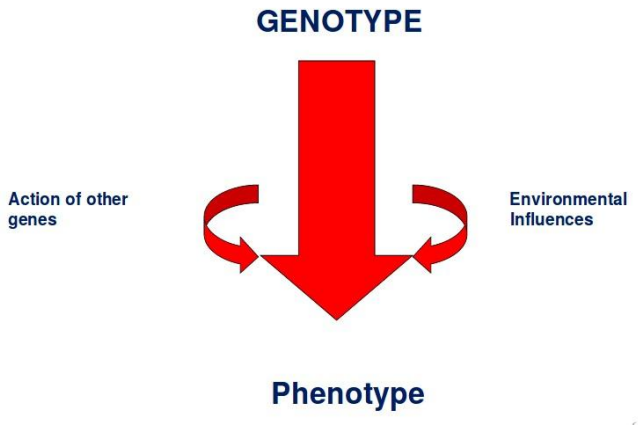


Fig. 1.1. A model defining phenotype

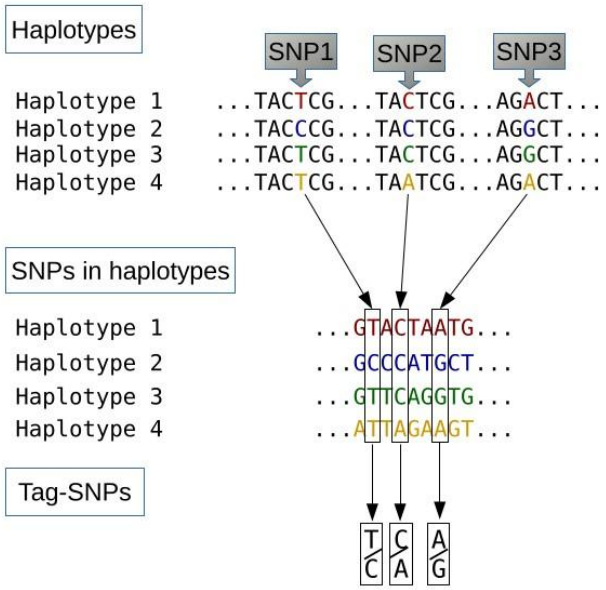
Therefore, the phenotypic observation can be defined by the following model [12]: phenotypic observation = environmental effects + genetical effects + residual effects, or

$$y_{ij} = \mu_i + g_i + \epsilon \tag{1.2.3}$$

where y_{ij} is the record of j of the i^{th} animal; μ_i refers to identifiable non random effects such as year of birth or sex of i^{th} animal; g_i is the sum of the additive (g_a), dominance (g_d) and epistatic (g_e) genetic values of the genotype of animal i; and ϵ_{ij} is the sum of random environmental factors affecting animal i.

Single Nucleotide Polymorphisms (SNPs). SNPs is a single base pair mutation at a specific locus and usually consists of two alleles (where the rare alleles frequency is greater than one percentage) [21]. They are the most abundant type of sequence variation in the human genome. These SNPs are useful in many diverse applications that include disease gene mapping, evolution, pharmacogenetics and forensics in understanding genetic variation [22]. SNPs are formed as a result of various kinds of changes that occur in the nucleotide sequence [21]. For example, these changes can happen because of the errors during DNA replications or radiation of chemicals. The mutation can alter a gene, change its expression or stop a gene from working [21]. This depends on the type of mutation and the location it occurs in the nucleotide sequence [23]. The SNPs can also be used to identify multiple polymorphisms in a region of DNA [21]. Multiple polymorphisms are caused by linkage disequilibrium: particular combinations of alleles in regions that are nearby in the DNA tend to be inherited together, and when inherited together as blocks, they are

called haplotypes [21,23]. In association studies, haplotypes allow us to identify multiple polymorphisms on a short block of DNA just by genotyping a few of them. This has led to the use of tag SNPs [13] as illustrated in Fig. 1.2.

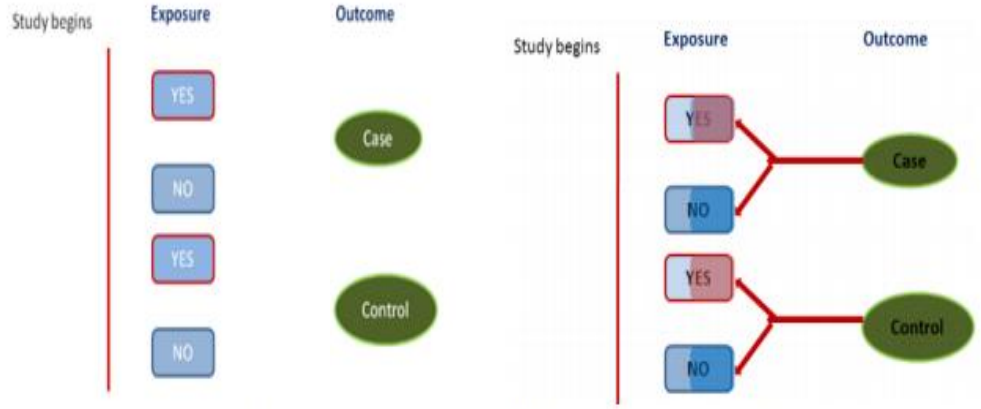


(a) At the top, each line represents a section of DNA with three different single nucleotide polymorphisms.
(b) In the middle, the DNA section has been compressed into only SNPs.
(c) At the bottom the haplotype section has been further compressed into only a few SNPs that are enough to determine which haplotype the individual has [21].

Fig. 1.2 Tag SNPs and haplotypes

There is a haplotype map (Hapmap) project that aims to map the haplotypes in the human genome across different populations [24]. The Hapmap project is intended to design genotyping array chip to measure the tags for world-wide population. This project contributes in advancing human variation and plays a crucial role in GWAS. Recently, genotyping chips can measure a million SNPs with the cost of around one hundred euros per sample [21]. This is the reason why SNPs are widely used in genetics because of the cheap genotyping cost, abundance of SNPs in the genome, and their robustness in capturing the changes in the genome [25]. **Genome wide association studies (GWAS)**. GWAS is an examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait [16,17]. In order to obtain a reliable signal, given the very large number of tests required, association must show a high level of significance to survive the multiple testing correction. Genome wide association studies are particularly useful in finding genetic variations that contribute to common, complex diseases such as schizophrenia and type II diabetes [26]. The main goal of GWAS is to identify loci which harbour causative variants hoping to implicate genes near these loci, thus leading to better understanding of disease and novel therapeutics [7]. In addition, GWAS use genetic risk factors to make predictions about who is at risk and to identify the biological underpinnings of disease susceptibility for developing new prevention and treatment strategies [26]. Using GWAS [21], the statistical association found are false positives or true associations [22,26]. False positives can be partially attributed to spurious associations caused by population structure and cryptic relatedness among individuals in a given cohort or results of false significant test results (type I error) [22]. True associations can result from the linkage disequilibrium: SNP tags a segment of the DNA, locus, which

increases susceptibility for having low or high values of the phenotype or having a disease [21]. A true association can also be a consequence of the SNP causing a direct result, for example a single nucleotide polymorphism can directly increase levels of high-density lipoprotein (HDL) [21]. Genome wide association often use case control designs as illustrated in Fig. 1.3 to identify genetic variants related to a specific phenotype [10,21,28]. This approach is good in capturing common variants but may fail to achieve enough power to detect variants that are rare in the population [10].



(a) Case-cohort study data collection step 1: exposure information not extracted.

(b) Case-cohort study step 2: extract exposure information for case and control.

Fig. 1.3. Case-cohort study

Family-based study is another approach used in a GWAS [21]. This approach provides protection against the heterogeneity of the population as well as genotyping errors since genotypes can be checked against inheritance [29]. The biggest drawback of using a family-based study is that the samples are harder to obtain as families that suffer from a specific disease are needed [21]. Family-based studies are modeled using LMM (1.1.1) to detect SNPs associated to a trait. Researchers make use of LMM to capture fixed effects which represent a candidate SNP (βX), mean or other confounding variables and random effects represent genetic random effect excluding the candidate SNP (u), and they capture the individual relatedness through kinship matrix based on either pedigree or genetic markers. GWAS have the following limitations; first it is difficult to move beyond mere statistical associations to identify the functional basis of the link between a genomic interval and a given complex trait since the case control design of GWAS only indicates the association, rather than causation [30]. Second, SNP associations identified in one population frequently are not transferable to members of other populations [30]. This is because most of SNPs identified by GWAS are inter genetic or intron region [30,31]. Third, the bulk of the heritable fraction of complex traits has not been accounted for in recent GWAS. This is because GWAS do not capture information about rare variants and have limited statistical power to detect small gene–gene and gene–environment interactions [31]. **Genetic relatedness (Kinship) matrix** (GRM)-: Among early steps in GWAS that makes LMM possible is the incorporation of kinship matrix into the statistical model that only needs to be estimated once [32]. This decreases computing time, controls Type-I error and has good properties on the statistical power and the efficiency [33]. In addition, it compels the estimated covariance matrix of the phenotypes to be positive semi-definite [32]. Kinship matrix describes how individuals are genetically related to each other according to their pairwise genotypic similarity [21]. Its structure incorporates population structure, cryptic relatedness and family structure [18]. It has computational advantage when small amounts of data is to be read for each test and used for calculations [32,33]. The estimated covariance matrix of the samples is based upon the variance of the sum of the random effects of the SNPs, summarised by a covariance parameter representing the genotypic additive variance of the phenotype [32,34]. Hence, the covariance matrix depends on the genotypic covariance which is decomposed by the kinship matrix [32]. The kinship matrix can be inferred based on known pedigree relations, genetic markers, or a combination of both [18,33]. Pedigree-based

kinship describes the recent relatedness better while marker-based kinship has a better capture on the distance relations. Although correct use of a marker-based kinship is preferred due to its higher statistical accuracy [33,35], its estimation varies highly in computing time across different available algorithms and definitions. Moreover, the computing time tremendously increases with sample size. The kinship matrix is obtained using the following formula of calculating the relatedness for individual's j and k [21]:

$$K_{ij} = \frac{1}{M}\sum_{i=1}^{M} \frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}, \quad (1.2.4)$$

where M is the set of loci i without missing data for j and k individuals, $x_{ij}$and x_{ik} denote the number of copies of the minor allele for the j^{th} and k^{th} individual in SNP i. p_i is the frequency of the minor allele for SNP i.

1.3 Advantages and success of mixed models in GWAS

The difficult statistical problem in GWAS is to find the genetic location that has a significant effect with respect to trait. The statistical method currently used is LMM. LMMs are used to estimate component of heritability explained by genotype marker and predict complex trait using genetic data [27]. For example, Yang et al. [27] uses LMMs framework to address the problem of missing heritability [13]. The authors find out that the fraction of the heritability of the height explained by asset 500K genotyped SNPs [7] is considerably larger than the heritability explained by genome wide significant hits alone, suggesting that the height is indeed driven by a plethora of common variants with small effect [13,27].

Furthermore, LMMs provide an increase in power to detect causal variants associated with a particular disease by applying a correction that is specific to the sample structure [36]. This aid to unravel rare variants that usually pose a greater challenge for all methods owing to the differential confounding of rare and common variants [36].

LMMs have been used previously with family-based data, and have successfully found genetic variants that affect plasma plasminogen levels (PLG) and estimated heritability for the PLG using a cohort of sibling pairs [21]. They have also been used in testing association studies in many studies ranging from human intelligence to bread wheat [22].

LMMs are used for studies using population samples. Typically, using population sample implies that there needs to be a correction for population stratification. Here, different statistical methods are used to take into account the population structure, for instance, genomic control (GC), structured association (SA) and principle component analysis (PCA). However, these methods are inadequate in the case of model organism association mapping [13,18]. GC suffers from weak power when the effect of the population structure is large as in model organisms. SA or PCA , which assumes a small number of ancestral populations and admixture, only partially capture the multiple levels of population structure and genetic relatedness in model organisms [13]. By contrast, the genetic relatedness matrix (kinship matrix) used in the LMMs captures both population structure and the cryptic relatedness [13,21].

LMMs have been applied in breeding selection, in particular when choosing sires and dames to mate in order to improve trait, or phenotype in the next generation e.g Dairy yield [12].

LMMs are used in predicting disease risk in the Welcome Trust Case and Control Consortium (WTCCC) dataset and quantitative phenotypes in heterogeneous. For example, there are seven common diseases such as bipolar disorder, coronary artery disease, crohn disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes that have been under WTCCC study, which have been previously used for assessing risk

[7] *genotyped SNPs are measurement of genetic variations of SNPs between members of a species that have been determined*

prediction. In the case of prediction of quantitative phenotypes in heterogeneous, the mouse dataset has previously been used to compare phenotype prediction methods [37].

1.4 Current challenges in mixed models in GWAS

1.4.1 False positive association

A LMM is primarily designed for quantitative traits and therefore, when applied to case control data, it imposes a misspecified model on binary phenotype, which can lead to current challenges in Mixed models (CCMM) in GWAS to type 1 error [10,14]. The variance component estimated by various hill-climbing approaches such as the Nelder-Mead simplex algorithm and Newton Raphson algorithm provide only a locally optimal solution, which may cause the statistical inference based on these to be inaccurate [13].

1.4.2 Loss in power

Loss in power by inclusion of the candidate marker. It has been shown that inclusion of the candidate marker in the GRM can lead to loss in power [11,38,39]. This is due to double-fitting of the candidate marker in the model, both as a fixed and random effect when computing GRM and testing genetic association between phenotype and genotype (Yang et al., 2014). Listgarten et al. [40], who referred to this phenomenon as 'proximal contamination,' demonstrates that a LMM with the candidate marker excluded (LMMe) is the mathematically correct approach and provides an elegant and efficient algorithm for LMMe analysis (implemented in FaST-LMM software).

However, owing to the computation time or the memory constraints and the complexities of LD, the LMM with the candidate marker included (LMMi) is more commonly applied in practice [33,36,41]. Yang et al. ([36] derives a new expression, validated by simulations, to quantify the reduction in test statistics when LMMi is applied. For linear regression (LR), the expected means of χ^2 association statistics, Λ_{Mean}, is;

$$\Lambda_{Mean} = 1 + \frac{Nh^2{}_g}{M} \tag{1.4.1}$$

where N denotes the number of samples, M is the number of markers and $h^2{}_g$ denotes the heritability explained by genotyped and/or imputed markers, regardless of the genetic architecture of the traits. For LMMi,

$$\Lambda_{Mean} = 1. \tag{1.4.2}$$

Equation (1.4.2) highlights the dangers of using Λ_{Mean} (or Λ_{Median}) to assess the presence of population stratification or other artifacts. When one observes lower Λ_{Mean} (or Λ_{Median}) values for LMMi than for linear regression it might be concluded that this difference was due to correction for confounding, but this result is in fact expected, even in the absence of any confounding. Finally, for LMMe;

$$\Lambda_{Mean} = 1 + \frac{NMh^2{}_g}{1-\gamma^2 h^2{}_g}, \tag{1.4.3}$$

where $\gamma = \frac{Nh^2{}_g}{M}$ and $M > N$. The ratio of Λ_{Mean} between LMMe and LMMi is

$1 + \frac{NMh^2{}_g}{1-\gamma^2 h^2{}_g}$,

which is consistent for causal, null and all markers.

1.4.3 Using a small subset of markers in the GRM

Here, we briefly discuss how GRM, a method that use a small subset of markers and can compromise correction for stratification. Some researchers in GWAS have advocated choosing a subset of markers to include the GRM when employing mixed linear model association (MLMA) methods [11,39,42]. FaST-LMM [40] uses an equally spaced subset of 4,000 (or 8,000) random markers (*RMs*) in the GRM, motivated by a computational speedup that reduces computational cost to CCMM in GWAS

$O((RMs)^2N)$ when $RMS < N$. The researchers in GWAS make use of local minimum of the genomic control factor or the global maximum of out-of sample prediction when selecting markers (markers with genome-wide significance value). This selection is based on the result of GRM [39,42]. However, some include all the markers when analyzing genetic data, although it is always computational intensive due to high dimensional problems [15]. These are the problems confronted by researchers when performing genetic association analysis [33,41]. Yang et al. [36] evaluates the impact of these choices on both false positive associations and power and finds that there is subtle population stratification, in particular when a few thousand random markers are used and the $_{TM}$ associated markers selected on the basis of the first local minimum of the genomic control factor Λ_{Median}. It concludes that using a small subset of markers in the genetic relatedness matrix can compromise correction for stratification. Based on the methods published so far, they recommend that studies of randomly ascertained quantitative traits in which population stratification is a key concern should generally include all markers (except for the candidate marker and markers in LD with the candidate marker) in the genetic relatedness matrix.

1.4.4 Loss of power in ascertained case-control studies

Most of research in GWAS on mixed model analysis assumes that study samples are randomly ascertained with respect to the phenotype of interest [10,36,43]. This is usually true for quantitative phenotypes but not for case-control studies, which generally oversample disease cases to increase study power [14,28]. Recent work highlights the loss of power that occurs in ascertained case-control studies when genetic or clinical covariates are modelled as fixed effects without accounting for ascertainment [10]. For example, in the case of nonrandom case-control studies, the performance of LMMs deteriorates with increasing sample [43]. However, a subset of these studies has developed new methods to address this problem [36]. For instance, Weissbrod et al. [43] proposes a framework called LEAP (liability estimator as a phenotype) that tests for the association with the estimated latent values corresponding to severity of phenotype. This problem can also be addressed by moving from the LMM to GLMM framework [13] using an appropriate link function f, where instead of assuming model (1.1.1), they assume that

$$f(P(Y = 1|X, Z, \mu)) = X\beta + Z\mu, \tag{1.4.4}$$

where Y represents phenotype vector for individuals, X is fixed effect of data, β is the vector of fixed effects, μ is vector of random effect and Z is random effect data.

1.4.5 Problem of confounding due to sharply designed structure and rare variant in GWAS

Confounding in GWAS arises from population stratification and it has been recognized for many years [9,18]. It is attributed to spatial structure population in conjunction with rare variants, and no current available statistical genetic method could account it [40]. In particular, when simulating the non-genetic cause of disease arising from a sharply defined spatial regions, LMMs and principal component analysis all fail to correct for stratification resulting systemically inflated test statistics and needs geographical information [40]. FastLMM select was developed by Listgarten et al. [40] to address this particular problem of confounding due to sharply designed structure, rare variant and any other types of confounding.

1.4.6 High computational cost

According to Loh et al. [41], LMM analysis is computational expensive despite a series of recent algorithmic advances. The current algorithm requires a total running time of either $O(MN^2)$ or $O(M^2N)$ (where N is the

number of samples and M is the number of SNPs). Therefore, this cost is becoming prohibitive for large cohorts, compelling existing methods to subsample the markers so that $M < N$ [40]. Loh et al. [41] proposes a more efficient mixed model association method, BOLT-LMM [41], which requires only a small number of $O(MN)$ time iterations and increases power by modeling more realistic, non-infinitesimal genetic architectures via a Bayesian mixture prior on marker effect sizes in order to address this problem.

Linear mixed effect models are statistical models that contains both the fixed and random effects which contribute linearly to the response function [13]. They are developed to handle clustered data or data with repeated measurements (longitudinal data). Clustered data are data in which observations are grouped into disjoints classes (data grouped into clusters by a common trait), according to some classification criterion [1]. Observations within the same cluster share common random effects and are statistically dependent [44].

The parameters of mixed effect are categorised into two groups: fixed effects and variance-covariance component [3]. Fixed effects are the average effects of predictors on the response while variance-covariance component is associated with the covariance structure of the random effect and of the error term [44]. The general form of the model is

$$y = X\beta + W\mu + \epsilon \tag{2.1.1}$$

where y is a $n \times 1$ vector of responses, X is a $n \times p$ design matrix for fixed effect, W is a $n \times q$ design matrix for random effect, β is a $n \times p$ vector of unknown fixed parameters and u is a $q \times 1$ is vector of random variable and is the error vector. It is assumed that

$$\mu \sim N(0, G)$$

and

$$\epsilon \sim N(0, R),$$

where G and R are variance-covariance matrix. The variance-covariance matrix R represents the within cluster variances and covariances while G denotes variances and covariances of the between-cluster effect. According to Waterman [44], the vectors μ and ϵ are also assumed to be independent and can be expressed in matrix notation as follows

$$\begin{bmatrix}\mu \\ \epsilon\end{bmatrix} \sim MN\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}G & 0 \\ 0 & R\end{bmatrix}\right) \tag{2.1.2}$$

The expected value of y conditioned on the random effect equals to

$$E[y|\mu] = X\beta + W\mu \tag{2.1.3}$$

and the expected value of y is

$$E[y] = E[E[y|\mu]] = X\beta. \tag{2.1.4}$$

The variance of y can be found using the conditional expectation and conditional variance as

$$Var(y|\mu) = R.$$

The total variance of model (in 2.1.1) is expressed thus:

$$\begin{aligned} Var(y) &= E[Var(y|\mu)] + Var(E[y|\mu]). \\ &= R + WGW'. \end{aligned}$$

It was noticed that conditional distribution of $y|\mu$ follows a normal distribution with mean equal to $X\beta + W\mu$ and variance equal to R under the normal assumption of vectors μ and ϵ . In addition, the distribution y follows a normal distribution with mean $X\beta$ and variance covariance matrix equal to $R + WGW'$.

The variance component model (Random effect model) and mixed effect (ANOVA) model as well as linear model for longitudinal data are special cases of model (in 2.1.1) [45]. The researchers in GWAS make use of this model to unravel the genetic basis of quantitative traits such as height, weight and blood pressure [6].

A fixed effect model (in 2.1.1) is a statistical model that represents the observed quantities in terms of explanatory variables that are treated as if the quantities were non-random [1]. It is formed when random variable is excluded from the model (2.1.1), The general form of the model is as follow

$$Y = X\beta + \epsilon \tag{2.1.5}$$

where

$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ is a $n \times 1$ vector of responses.

2 Likelihood Estimation of Parameters

According to Searle et al. [46], there are different estimation methods for the parameters in model (2.1.1), but maximum likelihood (ML) and restricted maximum likelihood (RML) are most commonly used [1].

Pinheiro [1] shows that when writing the likelihood of y in model (in 2.1.1), it is convenient to factor out the variance of the error term, σ^2, from the variance-covariance matrix of random effect. For example, in the following formula let;

$$G = \sigma^2 K, \tag{2.2.1}$$

where K is called the scaled variance-covariance matrix of the random effects. Then, the total variance in model (2.1.1) can be expressed as follows;

$$\begin{aligned} V = Vary(Y) &= \sigma^2 W'KW + \sigma^2 I, \\ &= \sigma^2 (W'KW + I), \\ &= \sigma^2 H \end{aligned} \tag{2.2.2}$$

where $H = (W'KW + I)$.

We find the maximum log-likelihood function of y in model (in 2.1.1) as follows;

$$l(\beta, \sigma^2|y) = -\frac{1}{2}\left[n \log(2\pi\sigma^2) + \log(|H|) + \frac{1}{\sigma^2}(y - X\beta)^T H^{-1}(y - X\beta)\right],$$

then estimate the values of β and σ^2 that maximizes (the equation in 2.2.3). This can be done by taking the derivative of equation (in 2.2.3) with respect to each parameter then equating it to zero [44]. The values of β and σ^2 that maximize (equation in 2.2.3) are given by the following formulae:

$$\hat{\beta}_{ML} = (X^T H^{-1} X)(X^T H^{-1} y) \tag{2.2.4}$$

$$\hat{\sigma}^2{}_{ML} = \frac{1}{n}(y - X\hat{\beta}_{ML})^T H^{-1}(y - X\hat{\beta}_{ML}) \tag{2.2.5}$$

2.1 Restricted maximum likelihood (RML)

According to Pinheiro [1] and Kang et al. [33], RML estimates of the variance-covariance components are usually preferred to ML estimates in LMMs. This is because RML estimates take into account the estimation of fixed effects when calculating the degrees of freedom associated to the variance components estimates, while ML estimates do not [33]. In this subsection we discuss how to obtain the estimates of parameters for both fixed and random effect model using RML.

RML estimates are defined as ML estimates of the likelihood of a set of $n-p$ linear combination of the response vector y, corresponding to $n-p$ vectors that span the orthogonal compliment of the column space of fixed effects design matrix [1]. According to Thisted [47], one way to define such a set of vector $n-p$ is by considering the QR decomposition of X

$$X = QR = [Q_1, Q_2]\begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1 \tag{2.2.6}$$

where R_1 is the upper triangular matrix, Q_1 and Q_2 are set of orthonormal vectors that span the orthogonal compliment of the column space X [11]. The RML estimates can be obtained from the likelihood of the following equation:

$$y^* = Q_2 y \tag{2.2.7}$$

Equation (in 2.2.7) is simplified further by substituting y from model (in 2.1.1), which leads to

$y^* = Q_2 W\mu + Q_2\epsilon$
(since $Q^T{}_2 X = 0$ equation (in 2.2.6)).

The total variance is now given as

$$\begin{aligned} Var(y^*) &= Q^T{}_2 WGW^T Q_2 + Q^T{}_2 Q_2 R, \\ &= \sigma^2 Q^T{}_2 WGW^T Q_2 + Q^T{}_2 Q_2 \sigma^2 I, \\ &= \sigma^2 (Q^T{}_2 WGW^T Q_2 + I) \end{aligned}$$

(after substituting G from (2.2.1) and $R = \sigma^2 I$),

since $Q^T{}_2 Q_2 = 1$ because Q_2 is orthonormal vectors. Taking $H^* = Q^T{}_2 WGW^T Q_2 + I)$ implies that

$$V\,ar(y^*) = \sigma^2 H^*. \tag{2.2.8}$$

Therefore,

$y^* \sim N(0, \sigma^2 H^*)$.

Letting $n^* = n - p$, we can write the corresponding restricted likelihood as [1]

$$l_R(\beta, \sigma^2 | y) = -\frac{1}{2}\,[n^* \log(2\pi\sigma^2) + \log(|H^*|) + \frac{1}{\sigma^2}(y^* H^{*-1} y^*)]. \tag{2.2.9}$$

The value of σ^2 that maximizes (equation in 2.2.9) is thus;

$$\hat{\sigma}^2{}_R = \frac{1}{n^*}(y^* H^{*-1} y^*). \tag{2.2.10}$$

Notice that the restricted likelihood of equation (in 2.2.9) does not depend upon β and hence no fixed effects RML estimates are available.

These methods ML and RML are widely applied in GWAS. A number of tools have been developed to implement these methods. The GWAS tools include EMMAX [35], EMMA [33], GCTA [48], FAST-LMM [40], Grammar-Gamma method [49], FASTA [50], GEMMA [8] and TASSEL [51].

2.2 Random effect model

A random effect model is a statistical model which assumes that the data being analyzed are drawn from a hierarchy of different populations whose differences relate to that hierarchy [44]. It was proposed by Visscher et al. [52] to model the genetics factors affecting the phenotypes in GWAS. The general model is given as thus

$$y_j = \mu + \sum_{i=1}^{M} w_{ij}\mu_i + \epsilon_j. \tag{2.2.11}$$

Where y_j is a $n \times 1$ vector representing the number of individuals with quantitative traits phenotypes, so $E[w_{ij}] = 0$ and $V\,ar(w_{ij}) = 1$. $\mu_i \sim N(0, {\sigma^2}_\mu)$ is the effect of the i^{th} individual, $\epsilon_j \sim N(0, {\sigma^2}_e)$ is the environment effect which is assumed to be independent and identically distributed across individuals.μ is the mean term, that is typically assumed to be known and equal to 0 and M is the number of causal SNPs. w_{ij} is the value the i^{th} causative SNP of the j^{th} individual after standardizations, and is defined by the following;

$$w_{ij} = \frac{x_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}, \tag{2.2.12}$$

Where x_{ij} is the genotype indicator of the i^{th} SNP ($x_i = 0, 1\ or\ 2$), p_i is the minor allele frequency, and j is the environment effect [52].

The $V\,ar(w_{ij}\mu_i)$ is defined as follows;

$$V\,ar(w_{ij}\mu_i) = V\,ar(w_{ij})Var(\mu_i) = \frac{Var(x_{ij} - 2p_i)Var(\mu_i)}{2p_i(1-p_i)} = Var(\mu_i)$$

since $Var(x_{ij} - 2p_i) = 2p_i(1 - p_i)$ under Handy-Weinberg principle[8] [13].

Defining

$$g_i = \sum_{i=1}^{M} w_{ij}\mu_i$$

as genetic random effect of individuals j. The genetical variance is given by [33]

$$V\,ar(g_i) = {\sigma^2}_g = {\sigma^2}_\mu$$

and

$$Cov(g_i, g_k) = \frac{{\sigma^2}_\mu}{M} \sum_{i=1}^{M} w_{ij}w_{ik} \tag{2.2.13}$$

It follows that

$$y \sim N(0, K{\sigma^2}_g + {\sigma^2}_e),$$

[8] *Handy-Weinberg principle state that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences.*

where

$$K = \frac{1}{M} WW'$$

and W and K can be interpreted as genotypes matrix and the genetic correlation matrix.

This model forms the basis of LMM based association analysis (LMMA) when computing GRM by including the candidate SNP [48]. The variance ${\sigma^2}_g$ and ${\sigma^2}_e$ are estimated using ML or RML implemented in LMMs based approach GWAS tool such as Fast-LMM [50], EMMAX [35] and GEMMA [8].

2.3 Linear mixed model association (LMMA) approaches

LMMA is widely applied in GWAS. It is used to determine the genetic association between phenotypes and genotypes of all individuals [18,48]. The general form of the model is

$$y = 1\mu + X\beta + g + \varepsilon, \tag{2.3.1}$$

where 1 is vector of ones, μ is the mean term, X is a $n \times p$ matrix of (genotype indicator variable of the candidate SNP) predictor values, β is a $p \times 1$ vector of fixed effect of the candidate SNP, g is a $q \times 1$ vector of polygenic effects with each element being the aggregates effect of all SNPs for all individuals and is the error term [18]. It assumes as follows that;

$$g \sim N(0, K{\sigma^2}_g)$$

and

$$\varepsilon \sim N(0, I{\sigma^2}_e),$$

where K is genetic relationship matrix estimated from SNP data as described in equation (1.2.4) and I is a $n \times n$ identity matrix. Therefore,

$$\begin{aligned} Var(y) &= Var(g + \varepsilon) \\ &= {\sigma^2}_g K + {\sigma^2}_e I. \end{aligned}$$

The mean μ in (2.3.1) is set to zero by normalizing the phenotype (y) and the genotype indicator variable (X) [36]. For example, considering as follows;

$$y \sim N(0,1)$$

and

$$x_{ij} = \frac{r_i - 2p_i}{2p_i(1-p_i)},$$

where p_i is the allele frequency and r_{ij} is the genotype indicator of the i^{th} SNP $(r_i = 0, 1 \text{ or } 2)$.

The LMMA is categorised into two thus: LMMA including the candidate SNP (LMMi) and LMMA excluding the candidate SNP (LMMe). They are tools to implement LMM-based approaches. These tools either include or exclude the candidate SNP when computing GRM in (1.2.4). They include EMMAX [35], EMMA [33], GCTA [48], FAST-LMM [40], Grammar-Gamma method [52], FASTA [50], GEMMA [8] and TASSEL [51].

2.3.1 Linear regression analysis

Linear Regression (LR) is widely used in GWAS for analysis of genetic associations between phenotypes and genotype. LR is defined as an approach for modeling the relationship between a dependent variable *y* and one or more independent variables (explanatory variable).

Let N, M and M_c be the cohort (sample) size, the number of markers and the causal markers respectively. Assuming that no population structure or other artefacts and all markers are independent. Yang et al. [36] derives a new expression, validated by simulations to quantify the reduction in test statistics when LMMi is applied. For LR, the expectation of χ^2 association statistics[2] (Λ_{Mean}) LR analyses is as follows;

$$\Lambda_{Mean}\,(LR) = 1 + \frac{Nh^2{}_g}{M} \text{ for all SNP markers,}$$
$$\Lambda_{Mean}\,(LR) = 1 + \frac{Nh^2{}_g}{M_c} \text{for causal SNP markers,}$$
$$\Lambda_{Mean} = 1 \quad \text{for all markers,}$$

which does not depend on the genetic architecture of the trait as opposed to Λ_{Median}. Λ_{Median} is the median of χ^2 association statistics divided by expected median under the null hypothesis of no association [53]. The Λ_{Median} is often slightly lower than Λ_{Mean} when there is highly polygenic trait, but much lower for the case of less polygenic trait or very large sample size [54]. The derivation of LR analysis for LMMi and LMMe is summarized in Table (2.1) below.

For LMMi (in 2.3.2). Here, we closely follow the discussion of Yang et al. [36] on how LMMi is applied in GWAS. LMMi contains two random variables and one fixed effect. These two random variables are as result of inclusion of candidate SNP when computing GRM as well as the random variable associated with all other SNPs. For example, in model (in 2.3.1) g can be defined as

$$g = W\mu + W_t\mu_t$$

where $W\mu$ is the random variable associated with all other SNPs before inclusion of candidate SNP while $W_t\mu_t$ is the random variable associated with candidate SNP after inclusion. The general form of the model is

$$y = X\beta + W\mu + W_t\mu_t + \epsilon \tag{2.3.2}$$

where y is a $n \times 1$ vector of responses, X is a $n \times p$ matrix of fixed candidate SNP genotypes, β is a $p \times 1$ vector of fixed candidate SNP genotypes, μ_t is a $q \times 1$ vector of the random effect of the candidate SNP with W_t being the corresponding genotype matrix, μ is a vector of the random effects of all the other SNPs W being the corresponding genotype matrix and is the error term. It is assumed that as follows that;

$$\mu \sim N\left(0, I\sigma^2{}_\mu\right),$$
$$\mu_t \sim N(0, K\sigma^2{}_g)$$
$$\epsilon \sim N(0, I\sigma^2{}_e),$$

where I is a $n \times n$ identity matrix and K is GRM. Yang et al. [36] shows that if all the SNPs are independent and assuming the variance explained by the candidate SNP is small, then it follows that;

$$\sigma^2{}_g = h^2{}_g$$
$$\sigma^2{}_e = 1 - h^2{}_g$$
$$\sigma^2{}_\mu = \frac{h^2{}_g}{M},$$

where $h^2{}_g$ is the variance explained by all SNPs. Notice that the effect of the candidate SNP is fitted twice, once as fixed effect β_t and once as random μ_t. They also show that the analysis to estimate β_t is equivalent to analysis of the candidate SNP and the phenotype correcting for the effects of all the SNPs. This is done by letting $\alpha^* = W\mu$, then

$$y^* = y - \alpha^*$$
$$= X\beta + W_t\mu_t + (\alpha^* - \hat{\alpha}^*) + \epsilon,$$
$$= X\beta + W_t\mu_t + \epsilon^* \tag{2.3.3}$$

where $\epsilon^* \sim N(0, I\sigma^2{}_{\epsilon^*})$

$$\sigma^2{}_{\epsilon^*} = Var(\epsilon^*) = Var(\epsilon) + Var(\alpha^* - \hat{\alpha}^*),$$
$$= I(1 - \gamma^2 h_g)$$

where γ^2 is the accuracy of squared of predicting α^*, $\gamma^2 = \frac{\vartheta}{(\vartheta+1-\gamma^2 h_g)}$ with $\vartheta = \frac{h^2{}_g}{M}$

The value of β_t and μ_t can be estimated from equation (in 2.3.3) by mixed model equation (MME) [55].

$$\begin{bmatrix} W'_t W_t & W'_t W_t \\ W'_t W_t & W'_t W_t + \frac{\sigma^2{}_{\epsilon^*}}{\sigma^2{}_\mu} \end{bmatrix} \begin{bmatrix} \hat{\beta}_t \\ \hat{\mu}_t \end{bmatrix} = \begin{bmatrix} W'_t y^* \\ W'_t y^* \end{bmatrix}$$

$$\Longrightarrow \quad \begin{bmatrix} \hat{\beta}_t \\ \hat{\mu}_t \end{bmatrix} = \begin{bmatrix} W'_t W_t & W'_t W_t \\ W'_t W_t & W'_t W_t + \frac{\sigma^2{}_{\epsilon^*}}{\sigma^2{}_\mu} \end{bmatrix}^{-1} \begin{bmatrix} W'_t y^* \\ W'_t y^* \end{bmatrix} \tag{2.3.4}$$

Notice that the value of $\hat{\beta}_t = (W'_t W_t)^{-1} W'_t y^*$ is the same as the least square estimate of regression y on W_t after solving MME (in 2.3.4). The estimate value of $\hat{\mu} = 0$ because of fitting the same SNP as fixed effect and random variable [36].

LMMi applies the Wald test to determine the significance of the fixed effect X_t on genetic association between phenotype and genotype. The expected value of the chi-square statistics is expressed thus, [56]

$$E(\chi^2) = \frac{E\left(\hat{\beta}_t{}^2\right)}{Var\left(\hat{\beta}_t\right)},$$
$$= \frac{\hat{\beta}_t{}^2 + \left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)}{\sigma^2{}_\mu + \left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)}.$$

The average χ^2 statistics should be taken across all the SNPs [36] such that

$$\frac{1}{M}\sum_i^M \beta^2{}_i = \frac{h^2{}_g}{M} = \sigma^2{}_\mu$$

so that the mean of the χ^2 association statistics from LMMi analyses is thus

$$\Lambda_{Mean}(LMMi) = \frac{\sigma^2{}_\mu + \left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)}{\sigma^2{}_\mu + \left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)} = 1 \quad \text{at all SNPS}$$

$$\Lambda_{Mean}(LMMi) = \frac{\frac{h^2{}_g}{M_C} + \left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)}{\sigma^2{}_\mu + \left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)} = \frac{\frac{Nh^2{}_g}{M_C} + (1-\gamma^2 h^2 g)}{\frac{Nh^2{}_g}{M} + (1-\gamma^2 h^2 g)} \quad \text{at causal markers}$$

$$\Lambda_{Mean}(LMMi) = \frac{\left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)}{\sigma^2{}_\mu + \left(\frac{\sigma^2{}_{\epsilon^*}}{N}\right)} = \frac{(1-\gamma^2 h^2 g)}{\frac{Nh^2{}_g}{M} + (1-\gamma^2 h^2 g)} \quad \text{at null markers}$$

Where $\gamma^2 \approx \frac{Nh^2{}_g}{M}$ when $M > N$.

The LMMi has become increasingly important in GWAS. The GWAS tools that implement LMMi include EMMAX, EMMA, GCTA-LMMi [48], FAST-LMM [40], GEMMA and TASSEL [51].

For LMMe (in 2.3.3). This method is applied when the candidate SNP is not included in calculating the GRM. The model can be written based on equation (in 2.3.3) as follows;

$$y^* = y - \alpha^* = X\beta + (\alpha^* - \hat{\alpha}^*) + \epsilon \tag{2.3.5}$$

$$= X\beta + \epsilon^* \tag{2.3.6}$$

This implies that

$$y^* = X\beta + \epsilon^* \tag{2.3.7}$$

which is analogue to standard linear model of regressing y^* on X_t [1].

The Wald test is used to determine whether a certain predictor variable X_t is significant or not [56]. The expected chi-squared test statistics is defined by [36,56]

$$E(\chi^2) = \frac{E\left(\hat{\beta}_t^{\,2}\right)}{Var(\hat{\beta}_t)} = \frac{\hat{\beta}_t^{\,2} + Var(\hat{\beta}_t)}{Var(\hat{\beta}_t)} = 1 + \frac{N\hat{\beta}_t}{\sigma^2_{\epsilon^*}},$$

Yang et al. [36] shows that the expectation of the χ^2 association statistics from MLMe analyses is given by

$$\Lambda_{Mean}(LMME) = 1 + \frac{\frac{h^2 g}{M}}{\left(\frac{\sigma^2_{\epsilon^*}}{N}\right)} = 1 + \frac{\frac{Nh^2 g}{M}}{1-\gamma^2 h^2 g} \quad \text{for all markers,}$$

$$\Lambda_{Mean}(LMME) = 1 + \frac{\frac{h^2 g}{M_c}}{\left(\frac{\sigma^2_{\epsilon^*}}{N}\right)} = 1 + \frac{\frac{Nh^2 g}{M_c}}{1-\gamma^2 h^2 g} \quad \text{for causal markers and,}$$

$$\Lambda_{Mean}(LMME) = 1 \quad \text{for null markers.}$$

where $\gamma^2 \approx \frac{Nh^2 g}{M}$ when $M > N$.

Some of the researchers have implemented this model (in 2.3.7) in LMM packages (GWAS tools) to analyse GWAS data [36,41,48,49,50. They include GCTA (GCTA-MLMe or GCTA-LOCO) [48], BOLT-LMM LOCO [41], GRAMMAR-Gamma [49] and FAST-LMM [50].

Table 2.1. Showing expected mean of test statistics for causal, null and all markers [54]

Type of markers	Linear Regression	LMMi	LMMe	LMMe / LMMi
Causal markers (M_c)	$1 + \frac{Nh_g^2}{M_c}$	$\frac{\frac{Nh_g^2}{M_c} + (1-\gamma^2 h_g^2)}{\frac{Nh_g^2}{M} + (1-\gamma^2 h_g^2)}$	$1 + \frac{\left(\frac{Nh_g^2}{M_c}\right)}{1-\gamma^2 h_g^2}$	$1 + \frac{\left(\frac{Nh_g^2}{M}\right)}{1-\gamma^2 h_g^2}$
Null markers ($M - M_c$)	1	$\frac{(1-\gamma^2 h_g^2)}{\frac{Nh_g^2}{M} + (1-\gamma^2 h_g^2)}$	1	$1 + \frac{\left(\frac{Nh_g^2}{M}\right)}{1-\gamma^2 h_g^2}$
All markers (M)	$1 + \frac{Nh_g^2}{M}$	1	$\frac{1 + \left(\frac{Nh_g^2}{M}\right)}{1-\gamma^2 h_g^2}$	$1 + \frac{\left(\frac{Nh_g^2}{M}\right)}{1-\gamma^2 h_g^2}$

where $\gamma = \frac{\left[(1+\vartheta)-\sqrt{(1+\vartheta)^2-4h^2g\vartheta}\right]}{h^2g}$ with $\vartheta = \frac{Nh^2g}{M}$ if $M > N$, $\gamma^2 \approx \frac{Nh^2g}{M}$

2.3.2 Summary of LMMs approach tool. Here, we provide a summary of GWAS tools that use LMMs.

Table 2.2. Summary of different LMMs-based approaches applied by GWAS

Method	APPROACH	Testing	Cost	Applicability	Reference
EMMAX	LMMi	LRT	$O(MN^2)$	Pop	[33]
FAST-LMM	LMMe	LRT	$O(MN^2)$	pop and cc	[40]
GEMMA	LMMi	LRT	$O(MN^2)$	Pop	[8]
FASTA	LMMe	LRT	$O(MN^2)$	pop and cc	[50]
GRAMMAR-Gamma	LMMe	ST	$O(MN^2)$	Pop	[49]
GCTA	LMMi/LMMe	LRT	$O(MN^2)$	pop and cc	[48]
MMM	LMMi	LRT	$O(MN^2)$	pop and cc	[3]
Mendel	LMMi	LRT	$O(MN^2)$	pop and cc	[57]
TASSEL	LMMi	LRT	$O(MN^2)$	pop and cc	[51]
EMMA	LMMi	LRT	$O(MN^2)$	Pop	[33]
PLINK	LMMe	ST	$O(MN^2)$	pop and cc	[58]
BOLT-LMM	LMMi/LMMe	BAYES	$O(MN^{1.5})$	pop and cc	[41]

Where N is the sample size (the number of individuals in one study), M is the number of tested SNPs, cc is case-control cohort, pop is population cohort, LRT is the likelihood ratio test taken in GWAS; ST is the score test in GWAS.

3 Data Analysis and Interpretation

3.1 Introduction

The genomic wide association is an important approach in identifying causal variants associated with diseases. The number of variants to be tested for associations with phenotypes requires a large number of SNPs to be investigated in order to understand GWAS clearly. Moreover, the association test for each SNP is performed, and both *p*-values and linear regression estimate of the effect size (β) in (2.3.7) are obtained. The nominal *p*-values always need to be corrected for multiple testing given the number of tests. This is because significant results can arise by chance with many tests. A *p*-value in GWAS has significance threshold of 5×10^{-8}, which referred to as genome-wide significance is obtained by dividing the usual α of 0.05 by 1 million (the effective number of tests performed). According to Ehret [59], such a Bonferroni correction[9] is always conservative, increasing the credibility of loci with a *p*-value less than 5×10^{-8}. The sample size should be large enough; for instance, highly, significant results can be reached only by analyzing large samples (generally ≥ 1000 participants). This requirement is an important limitation of the method [59].

In this chapter, we use two methods that apply mixed model to identify causal SNPS that cannot be identified by GWAS. This is done by leveraging new GWAS approaches that consider the distinction between inflation from bias and true signal from polygenicity as discussed in chapter 2. These two tools include PLINK and EMMAX. PLINK is one of the most popular tools used in the GWAS [58]. It provides a compact, comprehensive tool-box for GWAS from basic quality control filtering, SNP association testing to advanced features including gene-based analysis, annotation and epistasis test. It uses regression-based

[9] *Bonferroni correction is an adjustment made to p-values when several dependent or independent statistical tests are being performed simultaneously on a single data set [62]*

methods on SNP alleles [32]. Boost algorithm has been implemented in PLINK and it helps in transforming genotype data to Boolean representation that allows fast logic computing [60]. On the basis of the genome-wide average proportion of alleles shared identical by state (IBS) between any two individuals [13], PLINK offers tools to cluster individuals into homogeneous subsets, perform classical multidimensional scaling (MDS) to visualize substructure, provide quantitative indices of population genetic variation, identify outlying individual population genetic variation, and identify outlying individuals. PLINK uses complete linkage hierarchical clustering to assess population stratification with the use of whole genome SNP data [58].

EMMAX is a statistical test for large scale human or model organism association mapping accounting for the sample structure in GWAS [35]. Both identical-by-descent (IBD) and identical by-state (IBS) kinship matrix are implemented in EMMAX. The marker-based IBS kinship matrix is used to reflect the polygenic background [18,35], which assumes small SNP-effects. The test statistic is a Wald-based F-approximation based on the restricted log-likelihood [32,35,56]. According to Armitage [61], the binary trait can be interpreted as a quantitative difference score. Using this line of reasoning, the developers of EMMAX implement the same analysis for both binary and quantitative trait [32].

Lippert et al. [11] points out that EMMAX, along with its predecessor EMMA, achieves additional computational efficiency (over and above that achieved by simply estimating parameter ${\sigma^2}_g$ and ${\sigma^2}_e$ only by reparametrizing the likelihood in terms of a parameters $\delta = \frac{{\sigma^2}_g}{{\sigma^2}_e}$ and by making use of spectral decompositions. This results in total computational complexity of $O(n^3 + sn + rn)$ where $O(n^3 + rn)$ is the computational complexity at stage 1 and $O(sn)$ at stage two, n is the number of observation of the phenotypes, s is the number of SNP tested and r denotes the number of iterations i.e. the number of evaluations of the likelihood required.

3.2 Materials and Methods

3.2.1 Data description

We analyse unpublished real data enrolled from West Africa (Gambia-Ghana). The data consist a total of 338408 variants and 959 individuals (484 males and 405 females, 70 missing phenotypes). Of these, 864 cases of tuberculosis (TB) and 95 healthy individuals (controls) are retained after performing quality control (QC) using PLINK, 288 duplicants ID individuals are removed after QC, and 329601 autosome SNPs (from chromosome 1 to chromosome 22) are included in the analysis.

3.2.2 Methods

A genome-wide association test was performed on the full dataset of 959 people, which contain related individuals. First, text genotype was converted into numeric genotype score (0,1 or 2) using PLINK v1.90b3b software [58] based on the count of the alleles. Three files were created when making binary PED files (PED files are compact representation of the data and save space and speed up subsequent analysis). These files included: GWAS.bed that contained the raw genotype of the data, GWAS.map that contained two extra columns that provided the names for each SNP, and GWAS.fam that represented the first six columns GWAS.ped (see [58] for more description). The input data was specified in binary format as opposed to the normal text PED/MAP format by making use of this command “plink1.9 - -bfile option”. The input data for our analysis was 329601 autosome SNPs after performing quality control and filtering using PLINK. Genetic association test was performed using PLINK v1.90b3b software and EMMAX.

In order to perform association using EMMAX, the first the marker-based kinship matrix had to be created, IBS matrix from the transposed genotype and phenotype data. The IBS matrix was used to perform genetic association as opposed to the Balding-Nichols (BS). This is because IBS is known to be more robust in constructing the empirical kinship matrix [35]. The empirical kinship matrix was constructed by applying

EMMAX-kin [35] that computed a pair-wise GRM from our dataset which represented the structure of our samples. EMMAX software estimated the contribution of the sample structure to the TB phenotype using variance component model (in 2.2.11). The estimated covariance matrix of phenotype that modeled the effect of genetic relatedness on the TB phenotype was obtained. The genome-wide significance level (*p*-value threshold) was computed using [63].

$$p = \frac{\alpha}{2 \times M}, \tag{3.2.1}$$

where α = 0.05 and M is the number of SNPs tested (from typed or imputed data set). Two types of graphs were plotted: Manhattan plot and Q-Q plots. The Q-Q plot is defined as a graphical representation of the deviation of the observed *p*-values from the null hypothesis. Q-Q plots (as shown in Fig. (3.2a), (3.2b) and (3.2c)) were used to compare the genome-wide distribution of the test statistics with the expected null distribution. Genomic control inflation factor[10] (λ_{GC}) was used to check the presence of stratification. λ_{GC} was calculated using [64].

$$\lambda_{GC} = \frac{Median(\chi^2)}{0.455}, \tag{3.2.2}$$

where 0.455 is the median of χ^2 distribution with one degree of freedom. The acceptable values of λ_{GC} ranged from 0.95 to 1.00 that indicated no presences of population stratification or hidden relationship [56]. The observed χ^2values for each SNP were sorted in descending order and plotted against expected values. Manhattan plots (as shown in Fig. (3.1a), (3.1b) and (3.1c)) are used to represent the $p-$values of the entire GWAS on a genomic scale [59], the *p*-values are normally represented in genomic order by chromosomes and the position of chromosome (on the *x*-axis)). The values on the *y* axis on Manhattan plot represent $-log(p)$ (equivalent to the number of zeros after the decimal points plus 1).

3 Results and Discussion

A total of 338408 variants and 959 people (484 males, 405 females and 70 missing phenotypes) passed filters and quality control using PLINK. Among the remaining phenotypes, 864 were cases and 95 controls. A summary statistic for PLINK and EMMAX was obtained shown in Table 3.1 and 3.2 respectively. PLINK identified SNP rs4127341 ($p-$values = 8.693e-08) on chromosome 1 and $rs1343869$ ($p-$value = 9.948e-08) on chromosome 13 to be the most genomic-wide significance. The threshold *p*-values for genomic-wide association was 7.3875e-08 which was calculated using equation (in 3.2.1), where $M = 338408$. The SNP $rs10916338$ was identified in both cases as shown in Tables (3.1), (3.2) and (3.4), but the Bonferroni value in Table (3.4) showed that this was not significant since 0.278 was greater than the general threshold $p-$values of 0.05 for each SNP, with 0.6% false discovery rate as indicated in Table (3.4).

Table 3.1. Summary statistics from PLINK with top genetic variants with significant association signal

CHR	SNP	Position	P values
1	rs4127341	68144093	8.693e-08
1	rs6588294	68144964	1.030e-07
1	rs10916338	226782546	8.214e-07
10	rs2144861	51979762	4.094e-07
10	rs10883553	102625465	8.807e-07
10	rs2181834	102651241	5.056e-07
10	rs1122556	102696588	7.374e-07

[10] *The genomic control inflation factor is defined as the ratio of the median of the empirically observed distribution of the test statistic to the expected median, hence quantifying the extent of the bulk inflation and the excess false positive rate [64].*

CHR	SNP	Position	P values
10	rs10883567	102745769	2.134e-07
10	rs752974	102752246	9.274e-07
13	rs1343869	41223871	9.948e-08

Key: CHR = Chromosome
SNP = Single Nucleotide Polymorphisms

Table 3.2. Summary statistics from EMMAX with top genetic variants with significant association signal.

CHR	SNP	BP	P values
1	rs10916338	226782546	1.581316e-06
2	rs12328060	49824910	8.848916e-06
8	rs17217757	106682497	1.931775e-06
10	rs10824524	78746132	5.667582e-06
13	rs1343869	41223871	2.168218e-06
13	rs4941412	41402448	1.335220e-06
14	rs11620836	46226436	9.576813e-06
17	rs7225581	56372595	7.351500e-06
1	rs11203368	17539095	1.371354e-05
1	rs6694316	56197709	4.928598e-05

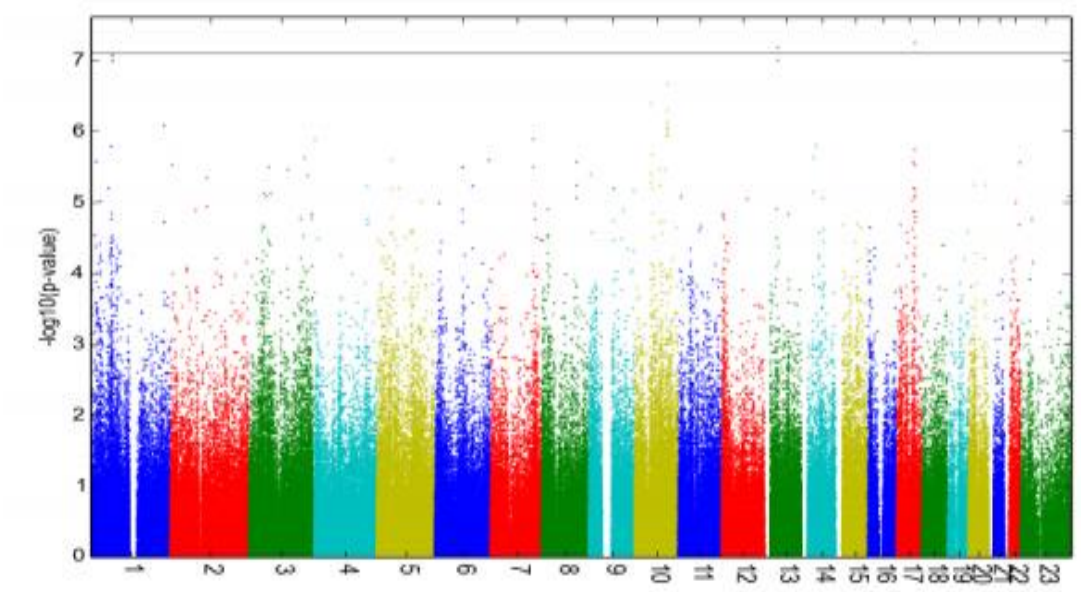
Key: CHR = Chromosome
SNP = Single Nucleotide Polymorphisms

The summary statistics for restricted ML from EMMAX was obtained as shown in Table 3.3. EMMAX estimated genetic variance $\sigma^2{}_g$ and residual variance $\sigma^2{}_e$ of model (in 2.2.11) to be 0.0594 and 0.0723 respectively. Moreover, EMMAX estimated a narrow sense of heritability (pseudo-heritability described in equation (in 1.2.2) to be 1.852e-01. This phenotypic variance showed that all common SNPs expressed approximately 18.52% of phenotypic variation of the disease.

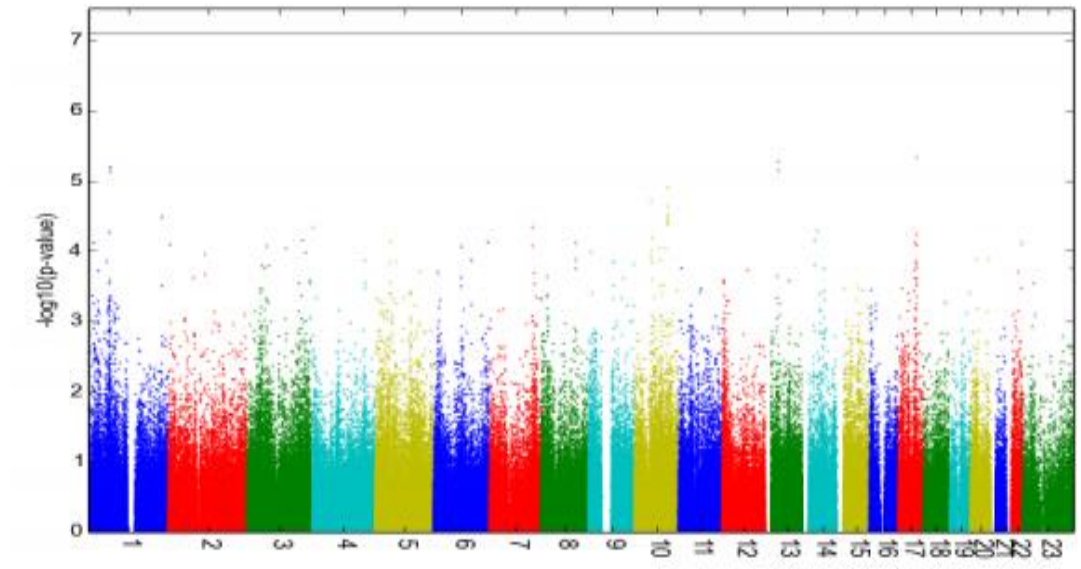
Table 3.3. Summary for restricted maximum likelihood from EMMAX

Parameters	Values
Log-likelihood with variance component (model 2.2.11)	-197.5286
Log-likelihood without variance component (model 2.1.5)	-202.4196
The ratio between variance parameters $\left(\frac{\sigma^2{}_e}{(\sigma^2{}_g}\right)$	1.2163
The genetic variance parameter $(\sigma^2{}_g)$	0.0594
Residual variance $(\sigma^2{}_e)$	0.0723
The pseudo-heritability (h^2)	0.1852

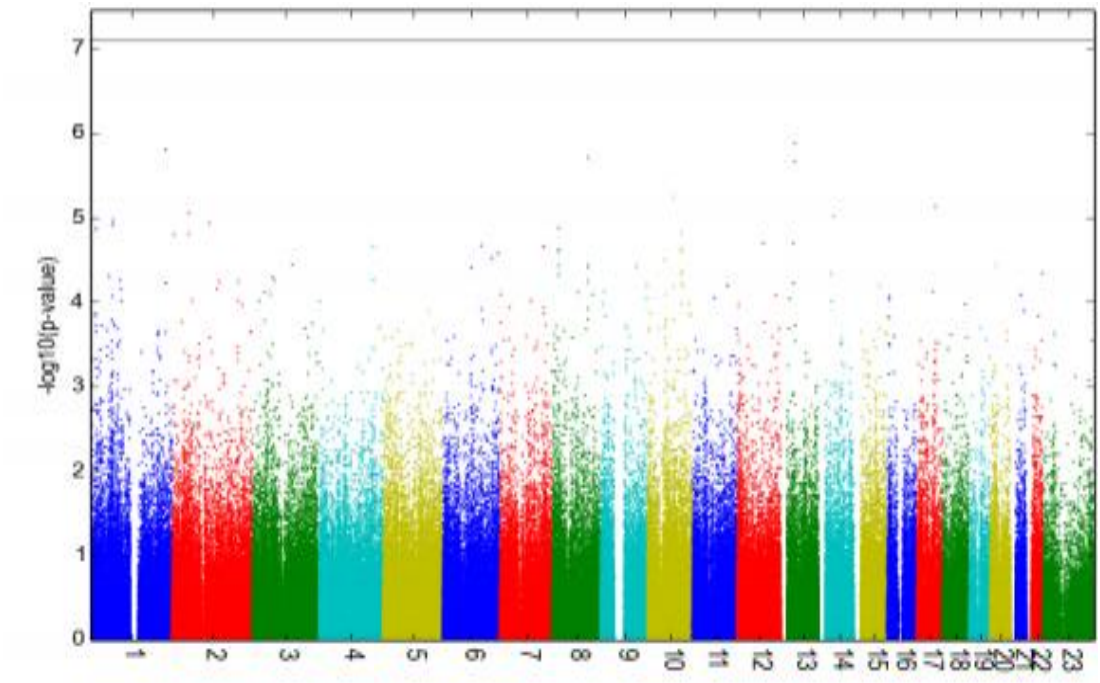
Figs. 3.1a, 3.1b and 3.1c shows Manhattan plots obtained from the association analyses using PLINK and EMMAX. The $-\log_{10}(p)$ for the association between each genetic variants and TB was plotted on the vertical axis, and the genomic control coordinates (the autosomal chromosome) is plotted along the horizontal axis. The gray line in the plots represented genome-wide significance line ($-\log_{10}(7.3875e-08)$). A number of SNPs display *p*-values above the significance threshold. The SNP rs7225581 on chromosome 17 (as shown in Fig. (3.1a)) and SNP rs4941412 on chromosome 13 were both identified using PLINK and EMMAX as most genomic-wide significant. In addition, it was noticed that false positives were reduced when EMMAX is applied as compared to PLINK. This was evidenced from Manhattan plots 3.1a and 3.1c from PLINK and EMMAX respectively.



(a) Fig. 3.1a. Manhattan plot for genetic association from PLINK.



(b) Fig. 3.1b. Manhattan plot for adjusted genetic association from PLINK

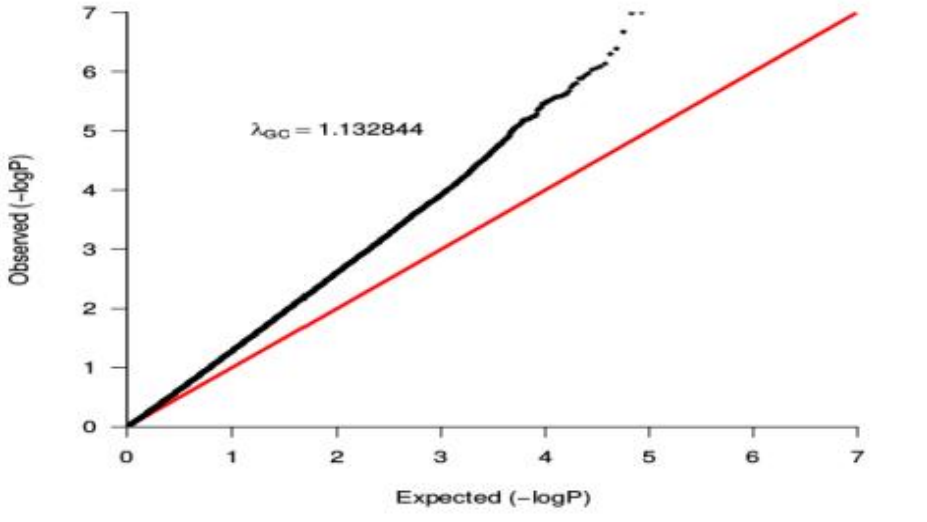


(c) Fig. 3.1c. Manhattan plot for genetic association from EMMAX

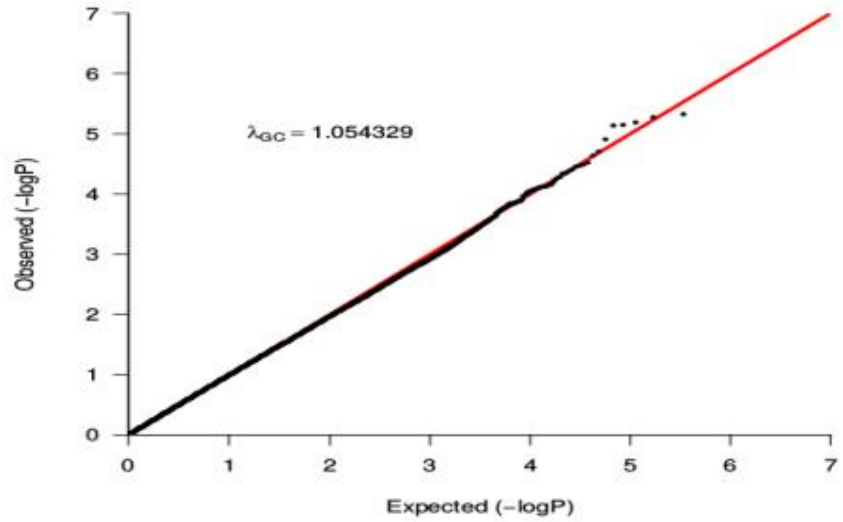
..

Fig. 3.1. Manhattan plot

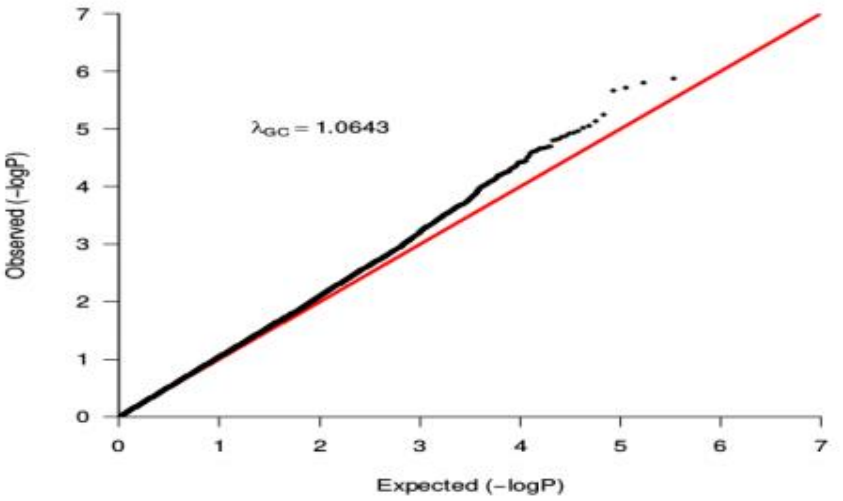
Fig. 3.2. These are Q-Q plots of genetic associations effect that compare the distributions of observed *p*-values with the expected distribution. The genomic control lambda λ_{GC} values indicate the residual population stratification effect.



(a) Fig. 3.2a. Q-Q plot for genetic association from PLINK.



(b) Fig. 3.2b. Q-Q plot for adjusted genetic association from PLINK.



(c) Fig. 3.2c. Q-Q plot genetic association from EMMAX.

Fig. 3.2. Q-Q plots

From Fig. 3.2a, it was noticed that there was an early separation of the expected from the observed. This indicated that many moderately significant *p*-values were more significant than expected under null hypothesis. In addition, $\lambda_{GC} = 1.132844$ suggest the presences of additional relatedness or population stratification that was not well accounted for by known family relationship. This could be attributed to the fact that PLINK did not account for cryptic relatedness in the sample.

The Q-Q plots from EMMAX and PLINK (adjusted genetic association) are shown in Figs. 3.2b and 3.2c. The genomic control lambda is λ_{GC}= = 1.054329 for adjusted plink association, and λ_{GC}= 1.063 for genetic association when using EMMAX. The genomic control lambda for both adjusted PLINK association and EMMAX show little departure from the null expectation, implying that there was still very low levels of existing population stratification.

Table 3.4. Summary statistics from adjusted association using PLINK with top significant genetic variants

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK SS	SIDAK SD	FDR BH
17	rs7225581	5.604e-08	4.713e-06	0.01896	0.01896	0.01879	0.01879	0.006971
13	rs4941412	6.584e-08	5.292e-06	0.02228	0.02228	0.02203	0.02203	0.006971
1	rs4127341	8.693e-08	6.462e-06	0.02942	0.02942	0.02899	0.02899	0.006971
13	rs1343869	9.948e-08	7.119e-06	0.03366	0.03366	0.03310	0.03310	0.006971
1	rs6588294	1.030e-07	7.299e-06	0.03485	0.03485	0.03425	0.03425	0.006971
10	rs10883567	2.134e-07	1.233e-05	0.07222	0.07222	0.06968	0.06968	0.012040
10	rs2144861	4.094e-07	1.969e-05	0.13850	0.13850	0.12940	0.12940	0.019790
10	rs2181834	5.056e-07	2.293e-05	0.17110	0.17110	0.15730	0.15730	0.021390
10	rs1122556	7.374e-07	3.008e-05	0.24950	0.24950	0.22080	0.22080	0.026150
1	rs10916338	8.214e-07	3.251e-05	0.27800	0.27790	0.24270	0.24270	0.026150

KEY:UNADJ = Unadjusted, asymptotic significance value;
GC = Genomic control, is adjusted significance value;
BONF= Bonferroni, is adjusted significance value;
HOLM = Holm is step-down adjusted significance value;
SIDAK SS = Sidak is single-step adjusted significance value;
SIDAK SD = Sidak is step-down adjusted significance value;
FDR H = False discovery rate, step up significance value [65].

4 Discussion

LMM approaches such as those implemented in the package EMMAX and Plink was well demonstrated through using application of real data. These methods offered a convenient and robust approach for analyzing quantitative or binary trait and controlling overall genomic inflation factor to an appropriate level and offer higher power than traditional family association such as those implemented in FBAT [56]. However, from the analysis of real data indicated that, for plink, care may need to be taken to use estimated kinships based on SNP data rather than known pedigree relationship, if one is to avoid any inflation in GWAS test statistics. We therefore believe that our results highlight the concordance between different LMM methods are equally relevant and useful to researchers carrying out GWAS of apparently unrelated individuals as to researchers carrying out family-based studies [56].

Systematic biasness from unrecognized genotyping artifacts or population structure was detected using (Q-Q) plots as shown in Fig. 3.2 of p-values from Plink and EMMAX. The Q-Q plots showed that the observed distribution of the test statistics closely follows the expected (null) distribution implying low levels of existing population stratification. However, when using Manhattan plots, a few p-value fall exactly on gray line in the plot (as shown in Fig. 3.1a). The deviation from gray line was attributed to functional variants in the TB phenotype since standard GWAS association analysis using LMM (such as Plink and EMMAX) or logistic models usually tends to have imperfect asymptotic distribution in the case of rare (1-5%) variants which usually results to SNPS falsely attaining of genome wide significance in standard test (Chimusa et al; 2014). Although in the study in Chimusa et al. [63] in the admixed South African coloured population indicated that the non-African ancestral component confers no risk to TB, it is not completely unreasonable to consider a possible systematic bias arising from using Linkage disequilibrium (LD) scores from African population to calibrate the Bayesian association statistics in BOLT-LMM. This highlighted the need for analyses beyond the standard GWAS and imputation using different LMM approaches. In association analysis using Plink, SNPs $rs10916338$ on chromosome 1 showed substantial association with TB. This SNP was also replicated in analyzing using EMMAX although it was not statistically significance. Moreover, new variants $rs\ 4941412$ on chromosome 13 and $rs\ \ 722558$ on chromosome 17 were detected with a significant association. It was expected after imputation to observe more SNPs in LD to or nearby the identified SNPs, particularly $rs\ 10916338$ on chromosome 1 to either also be associated or moderately associated with TB but was not the case. Therefore, this implies that our result is, in general, inclusive.

The model in (2.2.13) formed the basis of LMM based association analysis when computing GRM by including the candidate SNP [48]. The variance $\sigma^2{}_g = 0.0594$ and $\sigma^2{}_e = 0.0723$ were estimated using RML implemented in LMMs based approach GWAS tool EMMAX. The $\sigma^2{}_e$ showed that 7.23% the environment effect which was assumed to be independent and identically distributed across individuals and $\sigma^2{}_g$ showed 5.94% was the variations across individuals. This implied that $\epsilon_j \sim N(0, 0.0723)$ and $g_i \sim N(0, 0.0594)$ as described in Equation 2.2.13 [33].

LMMA was used to determine the genetic association between phenotypes of all individuals [18,48]. The general form of the was $y = 1\mu + X\beta + g + \varepsilon$ as described in model (2.3.1). LMMi was applied in GWAS. LMMi contained two random variables and one fixed effect. These random variables were as result of inclusion of candidate SNPs when computing GRM as well as the random variable associated with all other SNPs. The general form of the model was as Equation (2.3.2). The variance explained by all SNPs ($h^2{}_g$) was found to be 0.1852. We noticed that the effect of the candidate SNP was fitted twice once as fixed effect and once as random variable. They also showed that the analysis to estimate β_t in model (2.3.3) was equivalent to analysis of the candidate SNP and the phenotype correcting the effect of all the SNPs.

5 Conclusion

This study was aimed at understanding and exploring different approaches of mixed models as applied in genetic studies. Overview of genetic variation, advantages, successes and application of mixed models and current challenges of mixed models in GWAS were discussed. Moreover, SNPs associated with a particular disease using computational tools that applies LMM approaches were identified.

Two mathematical approaches of mixed models LMMi and LMMe which were applied in GWAS and the methods used to estimate parameters, ML and RML were discussed. The ratio of expectation of χ^2 association statistics Λ_{mean} between LMMe and LMMi was found to be consistent for causal, null and all markers as shown in Table 2.1. In addition, LMMi was found to be more efficient than LMMe when all markers were used during genetic association analysis, since at all markers $\Lambda_{mean} = 1$ for LMMi as opposed to LMMe. This indicated that there was no presences of stratification. The genetic variance $\sigma^2{}_g$ and residual variance $\sigma^2{}_g$ that account for phenotypic variation of disease were estimated to be 0.0594 and 0.0723 respectively. The computational tools that applied different LMM approaches based on their applicability and performance were classified as shown in Table 2.2.

The unpublished real data from West Africa (Gambia and Ghana) consisting of 959 individuals (864 cases of TB and 95 controls) were analyzed using the computational tools EMMAX and PLINK. The summary statistics from PLINK and EMMAX found two causal SNPs associated with the TB. These SNPs were rs7225581 on chromosome 17 and SNP rs4941412 on chromosome 13 with both having 0.69% FDR H. However, PLINK failed to correct hidden relatedness as depicted from Q-Q plot (in Fig. 3.2a), but EMMAX reduced false positives as evidenced from Q-Q plot (in Fig. 3.2a), although there was still very low presence of stratification exhibited by EMMAX.

The main challenge in this study was to correct presence of stratification exhibited by EMMAX. It is therefore recommended the extension of LMM-based approach tools in order to asses' rare variant which posed a greater challenge for all methods. It is recommended studies of randomly ascertained quantitative traits to include all markers when computing GRM.

Acknowledgement

Writing this piece of work required tremendous time, energy and patience, and would have been impossible without the help of others. Therefore, I would like to thank everyone who made significant contribution to the success of this piece of work. Specially, I would love to express my gratitude to:

My supervisor Dr Emile Chimusa Rugamika. His outstanding knowledge and expertise in the area of human genetic greatly defined this work. He was a role model and I enjoyed every moment of working with him.

Next Einstein Initiative (NEI) for supporting my entire research project since without them I would not have reached this point.

Drinold A. Mbete. He was dedicated and committed for this work and lend any support whenever I needed. I real enjoyed working with him throughout this project.

My family for encouraging me morally and spiritually and their support throughout my project. Finally, I thank Almighty God for giving me good healthy throughout my project.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Pinheiro JC. Topics in mixed effects models. PhD thesis, University of Wisconsin-Madison;1994.

[2] Golan D, Rosset S. Mixed models for case-control genome-wide association studies: Major challenges and partial solutions; 2011.

[3] Pirinen M, Donnelly P, Spencer CC, et al. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. The Annals of Applied Statistics. 2013;7(1):369-390.

[4] Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Human Heredity. 1971;21(6):523-542.

[5] Morton N, MacLean C. Analysis of family resemblance. 3. complex segregation of quantitative traits. American Journal of Human Genetics. 1974;26(4):489.

[6] Guo SW, Thompson EA. Monte carlo estimation of mixed models for large complex pedigrees. Biometrics. 1994;417-432.

[7] Mrode RA. Linear models for the prediction of animal breeding values. Cabi; 2014.

[8] Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nature Genetics. 2012;44(7):821-824.

[9] Wang K, Hu X, Peng Y. An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. Human Heredity. 2013;76(1):1-9.

[10] Jiang D, Zhong S, McPeek MS. Retrospective binary-trait association test elucidates genetic architecture of crohn disease. The American Journal of Human Genetics. 2016;98(2):243-255.

[11] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. Fast linear mixed models for genome-wide association studies. Nature Methods. 2011;8(10):833-835.

[12] Li G, Zhu H. Genetic studies: The linear mixed models in genome-wide association studies. The Open Bioinformatics Journal. 2013;7(1).

[13] Golan D, Rosset S. Accurate estimation of heritability in genome wide studies using random effects models. Bioinformatics. 2011;27(13):i317-i323.

[14] Fusi N, Lippert C, Lawrence ND, Stegle O. Warped linear mixed models for the genetic analysis of transformed phenotypes. Nature Communications. 2014;5.

[15] Hastie T, Tibshirani R, Friedman J. Unsupervised learning. In The elements of statistical learning. Springer. 2009;485-585.

[16] Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187(2):367-383.

[17] Korte A, Farlow A. The advantages and limitations of trait analysis with gwas: A review. Plant Methods. 2013;9(1):29.

[18] Astle W, Balding DJ. Population structure and cryptic relatedness in genetic association studies. Statistical Science. 2009;451-471.

[19] Slatkin M. Linkage disequilibrium understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics. 2008;9(6):477-485.

[20] Miles C, Wayne M, et al. Quantitative trait locus (qtl) analysis. Nature Education. 2008;1(1):208.

[21] L¨a¨aperi M, et al. Linear mixed models for estimating heritability and testing genetic association in family data; 2015.

[22] Akey J, Sosnoski D, Parra E, Dios S, Hiester K, Su B, Bonilla C, Jin L, Shriver M. Research report melting curve analysis of snps (mcsnp R): A gel-free and inexpensive approach for snp genotyping. Biotechniques. 2001;30(2):358-367.

[23] Strachan T, Read A. Human genetic variability and its consequences. Human Molecular Genetics. 4th ed. New York: Garland Science. 2011;405-440.

[24] Thorisson GA, Smith AV, Krishnan L, Stein LD. The international hapmap project web site. Genome Research. 2005;15(11):1592-1593.

[25] Siva N. 1000 genomes project; 2008.

[26] Bush WS, Moore JH. Genome-wide association studies. PLoS Comput Biol. 2012;8(12):e1002822.

[27] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common snps explain a large proportion of the heritability for human height. Nature Genetics. 2010;42(7):565-569.

[28] Guo F, Hankey J. Modeling 100-car safety events: A case-based approach for analyzing naturalistic driving data. Final Report. Report, (09-UT):006; 2009.

[29] Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. Nature Reviews Genetics. 2011;12(7):465-474.

[30] Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nature Reviews Genetics. 2009;10(4):241-251.

[31] Du Y, Xie J, Chang W, Han Y, Cao G. Genome-wide association studies: inherent limitations and future challenges. Frontiers of Medicine. 2012;6(4):444-450.

[32] Kampert MM. Statistical analysis in genome-wide association studies on genotype-imputed family data: A research strategy to compare various toolsets; 2011.

[33] Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709-1723.

[34] Geldermann H. Investigations on inheritance of quantitative characters in animals by gene markers i. methods. TAG Theoretical and Applied Genetics. 1975;46(7):319-330.

[35] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, Sabatti C, Eskin E, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010;42(4):348-354.

[36] Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics. 2014;46(2):100-106.

[37] Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.

[38] Consortium IMSG, WTCCC, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476(7359):214-219.

[39] Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. Nature Methods. 2012;9(6):525-526.

[40] Listgarten J, Lippert C, Heckerman D. Fast-lmm-select for addressing confounding from spatial structure and rare variants. Nature Genetics. 2013a;45(5):470-471.

[41] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics. 2015;47(3):284-290.

[42] Lippert C, Quon G, Kang EY, Kadie CM, Listgarten J, Heckerman D. The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. Scientific Reports. 2013;3:1815.

[43] Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nature Methods. 2015;12(4):332-334.

[44] Waterman MJT. Linear mixed model robust regression. PhD thesis. Virginia Polytechnic Institute and State University; 2002.

[45] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;963-974.

[46] Searle S, Casella G, McCulloch C. Variance components: Wiley series in probability and mathematical statistics; 1992.

[47] Thisted RA. Elements of statistical computing: Numerical Computation. CRC Press. 1988;1.

[48] Yang J, Lee SH, Goddard ME, Visscher PM. Gcta: A tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011a;88(1):76-82.

[49] Svishcheva GR, Axenovich TI, Belonogova NM, Van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. Nature Genetics. 2012;44(10): 1166-1170.

[50] Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM. Genabel: An r library for genome-wide association analysis. Bioinformatics. 2007b;23(10):1294-1296.

[51] Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al. Mixed linear model approach adapted for genome-wide association studies. Nature Genetics. 2010;42(4):355-360.

[52] Visscher PM, Yang J, Goddard ME. A commentary on `common snps explain a large proportion of the heritability for human height' by yang et al. Twin Research and Human Genetics. 2010;13(06):517-524.

[53] Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997-1004.

[54] Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'connell JR, Mangino M, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011b;19(7):807-812.

[55] Henderson CR. Best linear unbiased estimation and prediction under a selection model. Biometrics. 1975;423-447.

[56] Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell HJ, WTCCC, et al. Comparison of methods to account for relatedness in genome-wide associations studies with family-based data. PLoS Genet. 2014;10(7):e1004445.

[57] Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, Sobel EM. Mendel: The swiss army knife of genetic analysis programs. Bioinformatics. 2013;29(12):1568-1570.

[58] Purcell JS, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. Plink: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007;81(3):559-575.

[59] Ehret GB. Genome-wide association studies: Contribution of genomics to understanding blood pressure and essential hypertension. Current Hypertension Reports. 2010;12(1):17-25.

[60] Li J, Wei Z, Hakonarson H. Application of computational methods in genetic study of inflammatory, bowel disease. World Journal of Gastroenterology. 2016;22(3):949.

[61] Armitage P. Tests for linear trends in proportions and frequencies. Biometrics. 1955;11(3):375-386.

[62] Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z. A super powerful method for genome wide association study. PloS One. 2014;9(9):e107684.

[63] Chimusa ER, Zaitlen N, Daya M, Moller M, van Helden PD, Mulder NJ, Price AL, Hoal EG. Genome-wide association study of ancestry-specic tb risk in the south african coloured population. Human Molecular Genetics. 2014;23(3):796-809.

[64] De Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Human Molecular Genetics. 2008;17(R2):R122-R128.

[65] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological). 1995;289-300.

© 2018 Wanyonyi et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sciencedomain.org/review-history/26606