Statistical Analysis in Genome-Wide Association Studies on GenoType-Imputed Family Data: A Research Strategy to Compare Various Toolsets

Master Thesis

Leiden University

Mathematics

Specialization Statistical Science

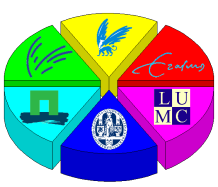
Defended on October 27, 2011

Maarten M.D. Kampert

Thesis Advisors:

Dr. Jouke-Jan Hottenga

Prof. Dr. Jacqueline J. Meulman



Statistical Analysis in Genome-Wide Association Studies on GenoType-Imputed Family Data: A Research Strategy to Compare Various Toolsets

Maarten M.D. Kampert

December 30, 2011

Abstract

Freely available toolsets that can handle genome-wide association (GWA) studies on twin-family data and take into account imputed genotypes are growing in number. However, the documentation that comes with them (if available), does not facilitate the choice for a particular toolset. We propose a research strategy in which we compare ASSOC, EMMAX, MERLIN, PLINK and ProbABEL on feasibility and statistical accuracy for GWA studies on simulated traits. Feasibility comparison was based on install requirements, versatility on data input, command line interface, and help information. The comparison on statistical accuracy was performed on Type-I error, genomic inflation, power, and consistency and efficiency of estimated SNP-effects. We simulated 100 replicates of binary and quantitative phenotypic traits over heritability conditions of 5, 10, 20, 30, 50 and 80%, based on 3 effect-SNPs from 1557 samples from 597 nuclear twin-families from the Netherlands Twin Registry. Analyses on Type-I error and genomic inflation were performed on 7757 pruned and unlinked SNPs that represented the null hypothesis. In the current design PLINK performs best on feasibility and statistical accuracy for the binary trait. On the quantitative trait ASSOC performs best on Type-I error control, EMMAX on statistical power, and PLINK on genomic inflation. Future research is needed for larger sample sizes and larger numbers of causal SNPs to compare the performance of the toolsets on complex traits.

Introduction

Statistical Techniques for GWA studies on family data

Genome-Wide Association (GWA) studies are mostly performed on genotypic data of which the samples are assumed to be unrelated. Applying the standard statistical techniques for such a GWA study on family data gives biased results. Although the genetic effects remain consistent, their standard errors and p-values are biased. A genotype associated with the phenotype could be more frequently present within a family as compared to its population. If this is the case, then the effect size of the genotype is overestimated, because more people have the associated genotype, resulting in underestimated standard errors of the genotype effects and too liberal p-values.

Much research on the statistical techniques in GWA studies has been done to control for family structure [5, 22, 32, 43, 45, 50]. Most popular are techniques derived from the Transmission Disequilibrium Test (TDT) [38], Bayesian statistical analysis [39], a posterior correction on the standard statistical techniques using genomic control, and the Mixed Linear Model (MLM). In the current monograph we will focus on MLM. The techniques derived from TDT or any of its derived methods are omitted because they rely on the amount of heterozygous parent-pairs, making them less powerful than MLM [28]. Furthermore, only the p-value results can be used for meta-analyses, because they do not provide the effect of a Single Nucleotide Polymorphisms (SNP) and its standard error. Bayesian analysis techniques could actually be rewritten as a MLM and have sufficient (or even more) statistical power. Although being advanced in its statistical methods, from the user's perspective, there is no optimal combination on efficient algorithms and its ease of use. Furthermore, Bayesian analysis results should be interpreted differently as compared to the majority of statistical techniques and therefore are hard to implement in meta-analyses from consortia. A last alternative, and the least preferably option, would be to maintain standard statistical techniques, but add measures for genomic control. Adding PCA components as covariates in the model and dividing the inflated test statistics by the residual genomic inflation factor λ_{gc} restores the p-values back to a null distribution for the test of association. This is a method mostly used to control for population admixture as well as cryptic relatedness. However, such a methodology does not grasp family data structure well enough and therefore introduces biased results [34]. In genetics the term "Mixed Linear Models " refer to statistical techniques that could be either the Generalized Linear Mixed Model (GLMM) as well as the Generalized Estimating Equations (GEE) method, or a combination of both. It is confirmed that these techniques have more statistical power and do not only control for family structure, but also control for population admixture and cryptic relatedness as compared to posterior genomic control measures or the techniques derived from TDT [22, 28, 50].

Statistical techniques in the MLM can control for the family structure in the sample through avoidance and relaxation of the assumption of independent identically distributed residuals (i.i.d.). The GLMM is an example in which the

independence assumption is avoided by adding random effects in the model for the clusters in the data. These random effects are assumed to have a distribution of which the (co)variance parameters need to be estimated. These parameters are structured by a known input variable that clusters the data. In the case of twin-family data one could assign a random effect to each nuclear family and a random effect for each monozygotic (MZ) twin-pair. In this way the statistical model allows for a covariance component within the nuclear families plus an extra covariance component for the MZ twins. To summarize, GLMM avoids the problem with random effects because the residual error terms can still be assumed to be i.i.d.. The GLMM method comes with a major computational drawback, however it is difficult (if not impossible) to make the algorithm efficient enough to finish a GWA study within months. The other possibility within MLM is to relax the independence assumption (GEE). In the GEE model, which is the other extreme of the MLM spectrum, correlated residuals are allowed, based on a pre-specified structure, the working correlation matrix. In the working correlation matrix we could specify that the MZ twin-pairs obtain the value 1.0 and other related members within a nuclear family obtain the value 0.5. Then the allowed correlation between MZ twin-pairs will be twice as large as the other related members within a nuclear family. The cost of this method is the loss in statistical efficiency and power, as a consequence relaxing the distributional assumptions, also known as quasi-likelihood based analysis.

MLM tailored to Human Genetics

Nowadays we see a rise of GWA studies on family data using MLM [5, 22, 50, 51]. At first computer and statistical techniques were not able to deal with family data for GWA studies due to lack of an optimal combination of sufficient statistical power, efficient algorithms and computing power. Note that if one test per SNP takes only a second, the GWA study takes 29 days for 2.5 million SNPs. In the next paragraph we'll go deeper into developments on how both MLM and GEE have become less computational demanding in GWA studies, relying on the properties of (human) genetics. Those not familiar with analyses such as MLM - implemented in GWA studies - are recommended to read either the "Appendix" or the online methods of Yang et. al. [22], and for a more technical understanding of MLM the book by McCullogh, Searle and Neuhaus [30].

Estimation of the parameters of MLM in GWA studies can be a computational burden. Note that GWA studies nowadays focus more on the rare and small SNP-effects. Rare SNPs or small effects require larger sample sizes to attain sufficient statistical power. With the knowledge that the total computing time for standard MLMs is a cubic function of the sample size, it is clear why computer science and bio-informatics intensively investigate new statistical techniques that decrease computing time and increase or maintain the statistical efficiency and power [5, 10, 23, 45, 50, 51]. In the next paragraphs we will review these implementations resulting in faster computation. However, with the proliferation of genetic research on this issue, it is very likely that some of these contributions to GWA studies might be overlooked in this monograph.

Kinship matrix

One of the first tailored steps in GWA studies that made MLM possible, is the incorporation of a kinship matrix into the statistical model, which only needs to be estimated once. It decreases computing time, controls Type-I error and has good properties on statistical power and efficiency [23, 40]. Furthermore, it compels the estimated covariance matrix of the phenotypes to be positive (semi-)definite. Hence, the software can always provide p-values and SNP-effects an error messages like "The Hessian Matrix is not positive definite" do no longer appear.

The kinship matrix is defined according to the pairwise genotypic similarity of individuals, so its structure incorporates population structure, family structure and cryptic relatedness. The computational advantage comes with considerably smaller amounts of data to be read for each test and used for calculations. Instead, the estimated covariance matrix of the samples is based upon the variance of the sum of the random effects of the SNPs, summarized by a covariance parameter representing the genotypic additive variance of the phenotype [14]. Hence, the covariance matrix relies on the genotypic covariance which is decomposed by the kinship matrix. A decomposition which is assumed to remain the same for each SNP being tested. E.g. a parent-offspring pair sharing half of the SNPs (kinship = 0.5) will also share half of the genotypic covariance explained in the phenotypic variance independent of the SNP being tested. Therefore, the relation between kinship and genotypic variance needs to be determined on the variance of many random-effect-SNPs. Only then we can reasonably assume that the covariance of the sum of the shared effect-SNPs is approximately half of the genotypic variance indeed. Note the similarity of the kinship matrix with the so-called working correlation matrix in GEE modeling with the difference that distributional assumptions are maintained.

The kinship matrix can be inferred from the known pedigree relations, the markers at hand, or a combination of both. Whereas pedigree based kinship describes the recent relatedness better, 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 [5, 22, 23], its estimation varies highly in computing time across the different available algorithms and definitions. Moreover, the computing time tremendously increases with sample size. Therefore, some researchers state that its use is impractical [28]. Furthermore, Price et al. [34] show that it does not always capture the population structure. One solution would be to use an extra matrix for the population stratification [49], but at the cost of computing time.

EMMA

Another methods that improved computing time is the Efficient Mixed Model Association (EMMA) algorithm, proposed by Kang [23]. In the EMMA algorithm a short-cut is made on Henderson's iterative procedure [15] by rewriting the equations and applying restricted log likelihood (RLL) using latent roots. In this way analyses involve as few as possible computationally intensive matrix inverses and matrix

multiplications during calculations. In addition, EMMA maximizes the individual random effects in the RLL directly from the kinship.

Two stage method

If the kinship matrix is estimated only once for each test, the opportunity appears to estimate the covariance matrix once also, instead of recalculating it for each test. When the contribution of the sample structure to the phenotype is estimated without fixed SNP-effect(s) of interest, this could be achieved with negligible loss in statistical power. Regressing the phenotype on the once only estimated covariance matrix and the covariates of interest will give us the residuals that take into account the family structure of the data. Using these residuals we can test the null hypothesis for each SNP using the standard statistical techniques like the generalized linear model. The method is an extension on the many SNP-effects assumption for the decomposition of the covariance based on the kinship matrix. It does not only assume the decomposition of the covariance to remain the same, but also the absolute value of the genotypic covariance is fixed for each SNP being tested. A method which has been confirmed to be viable for many small effect-SNPs, however associated with a decrease in power when SNPs are present that have a large significant contribution to the genotypic variance of the phenotype [5, 22].

Score test

Relying on the properties of small SNP-effects, a score test could be more reliable than the Wald test or the Likelihood Ratio (LR) test, and the computational burden is lower. The score test is derived under the assumption of the null hypothesis and therefore needs less parameters to be estimated, resulting in less computing time. In addition, it is robust against model deviations in the alternative hypothesis as compared to the LR test or the Wald test in MLM [17, 46]. Although the robustness property of the score test does not hold for large effects, it is very powerful for small SNP-effects [12], which is assumed to be the case in GWA studies.

Imputed Genotype Uncertainty (Soft-called data)

Up till now we briefly described the developments on statistical techniques for detecting the small effect sizes of quantitative trait loci in GWA studies. SNP detection could also be improved using genotype imputation. The posterior probabilities of the SNPs that are not called at the genotyping platforms can be calculated based on the properties of Mendelian inheritance within the family data, and information from reference panels [41, 42] on linkage disequilibrium between SNPs. At first, genotypes were imputed based on the maximum posterior probability, known as a best guess "hard calling" method. Nowadays it is confirmed that the use of an imputation method that extends further on the incorporation of the information from the posterior probabilities, "soft-calling", is associated with better statistical properties [53]. Testing of these soft-called SNPs can amplify

statistical power up to 10%. An increase that especially occurs for SNPs that are harder to tag [29, 37]. Therefore we will not consider best guess genotype imputed data in the current monograph.

Dosage and mixture genotype imputation are the two popular types of soft calling. With p_k denoting the posterior probabilities for a genotype that needs imputation, where k (0, 1, 2) indexes the genotype by its recessive allele counts for a bi-allelic locus:

1. *dosage* - impute the expected genotypic (or allelic) counts, which in the additive cases boils down to $p_1 \times 1 + p_2 \times 2$.
2. *mixture* - no summary, use all 3 possible genotypes with their posterior probabilities for further calculation.

Both have their pros and cons. With dosage data a certain amount of information is lost due to the fact that the true genotypic variance is underestimated. Hence, we risk the association between phenotype and genotype to be masked. In the mixture imputation strategy all information on genotype uncertainty is kept, however it demands more computing time since the genotypic data is at least twice as large. To estimate the SNP-effect parameters, a summation over the three genotype probabilities is needed to integrate out genotype uncertainty. Zheng et. al. [53] state that for most realistic settings of GWAS, such as modest genetic effects, large sample sizes, and informative genotype probabilities imputation accuracies, dosage-based analysis is as powerful as the mixture-based analyses.

Toolsets

We described statistical techniques and the advantages of using soft-called data. The next step is to search for a toolset to perform a GWA study on our soft-called twin-family data. With the growing number on freely available toolsets the choice becomes an elaborate, if not intractable, task. Choosing one toolset in the rapid developing environment of GWA studies is far from an easy job. Furthermore, the documentation explaining the statistical procedures used in each toolset, if available at all, is most times unsatisfactory. For this reason one's skills in statistics need to be up to date, to fully comprehend the statistical procedures implemented in the toolset. Moreover, the number of toolsets and the amount of able to deal with (twin-)family data and genotype uncertainty is still growing.

In the current monograph we present a research strategy in which the performance of freely available toolsets for GWA studies is compared on (twin-)family data taking into account soft-calling. We have selected ASSOC [43, 45], EMMAX[22], MERLIN[1], ProbABEL[5] and PLINK[35] to map the performance on quantitative and binary phenotype. The function descriptions of some of these toolsets still need to be better documented. From what is available in the documentation we give a brief description on the statistical analyses performed by these toolsets, a small overview on these features is presented in Table 1.

ASSOC uses a quasi-likelihood based score test for the quantitative and binary trait on the dosage data. It uses the retrospective likelihood in which the genotype is modeled on the phenotype. To be more specific, in ASSOC the Cochran-Armitage trend (MCA) test is modified such that it is able to control for relatedness using the identity-by-sharing (IBD) kinship matrix [43, 45]. Hence, the (co)variance can be split out for MZ-pairs and sib-pairs. Last, the variance term of the score test is able to take into account the loss of information due to genotype imputation uncertainty using Louis' formula [44].

In EMMAX, the eXpedited version of EMMA, the two stage method is implemented. They have implemented both IBD or identical-by-state (IBS) kinship matrix. We use their marker based identical-by-state kinship matrix to reflect the polygenic background [22], which assumes small SNP-effects. The used test statistic is a Wald-based F-approximation based on the restricted log-likelihood [23]. According to Argmitage's study [3] the binary trait can be interpreted as a quantitative difference score. Using this line of reasoning, the developers of EMMAX implemented the same analysis for both binary and quantitative trait.

The statistical procedures for soft-called family data of MERLIN and PLINK, are not sufficiently documented. Based on the few information PLINK performs a GEE probably allowing for a correlation between all family members, including parents. Within MERLIN the off-line function FastAssoc is implemented. A function known to be able to deal with dosage data performing a MLM score test in which the expected IBD kinship matrix is incorporated [10]. It allows for decomposition of the (co)variance components by MZ twins. Whether the FastAssoc MERLIN-offline function is able to deal with binary data is not documented.

In ProbABEL the functions for the dosage imputed genotypes are well documented. For the quantitative trait a two stage method is combined with a score test that incorporates a marker-based kinship matrix to split the covariance structure. On the binary trait ProbABEL performs a GEE, but it does not allow for any covariance. The diagonal of the working correlation matrix, however, has no restrictions and therefore allows for heterogeneous variances instead of i.i.d. residuals [5].

Table 1: Implemented human genetics in the MLM of the toolsets.

Feature	ASSOC	EMMAX	MERLIN	PLINK	ProbABEL*
Two Stage	–	+	–	–	+
EMMA	–	+	–	–	–
Kinship**	P	M	P	C	None / M
Test	Score	Wald	Score	Wald	Wald / Score
GEE	+	–	–	+	+ / –

Description on the current table: + = feature is present, – = feature is absent.
* If two symbols are given, the first represents the function for the binary trait and the second for the quantitative trait.
** P = kinship based on pedigree, M = marker based kinship, C = nuclear families as cluster in which all pairs have the same correlation, None = no relatedness is being modeled.

Research Strategy

As a research strategy we present a comparison of the above toolsets on feasibility and their statistical accuracy for simulated simple trait phenotypes. One hundred replicates were simulated for both the binary and quantitative traits on three unlinked genotype imputed large-effect-SNPs from 1557 samples from 597 nuclear families from the Netherlands Twin Registry [8]. As feasibility properties we address the install requirements, the versatility, the command language interface and the corresponding help documents. The statistical accuracy comparison is based on estimates of Type-I error, genomic inflation, power and consistency and efficiency of the SNP-effects over heritability conditions of 5, 10, 20, 30, 50 and 80%. Based on this strategy we bring structure to the choice problem on what toolset to use for GWA studies on simple trait having twin family data.

Material and Methods

Genotypic Data

The data on which we compare the toolsets come from individuals who took part in the NTR Biobank study [48]. A study aimed to collect biological samples (DNA, gene expression and biomarkers) in twins and their family members who also participate in the longitudinal phenotypic studies of the NTR. In-between the years 2004 and 2008 participants were visited at their homes between 7:00 and 10:00 am, during which fasting blood samples were collected.

In the current study we used a subsample of 597 nuclear families. Genotyping on these samples was performed on the Affymetrix 6.0 platform. The thresholds for SNPs were MAF $> 1\%$, HWE > 0.00001, missing $> 95\%$ and $0.30 <$ Heterozygosity < 0.35. Samples were excluded from the data if their expected sex and IBD status did not match, if the genotype missing rate was above 5% for the individual or 10% within the nuclear family, or if F-inbreeding coefficient $< .1$. All SNPs were aligned to the positive strand of the Hapmap 2 Build 36 release 24 CEU reference set. SNPs were excluded if allele frequencies differed more than 0.25 with the reference set. This final set was imputed against the reference set using Beagle 3.3 [38]. Bad imputed SNPs were removed based on HWE < 0.00001, MAF $< 1\%$, allele frequency difference > 0.25 against the reference set and a mendelian error rate $> 5\%$ obtained from the best guess data in PLINK v1.07. The final data resulted in a sample of 1557 individuals with 119 MZ-twin pairs.

Simulation Design

To obtain measures on statistical accuracy for the toolsets we simulated 100 replicates of binary and quantitative traits on 1,557 genotypes in GCTA [21] for heritability (H) levels of 0.05, 0.10, 0.20, 0.30, 0.50, and 0.80 using the following model:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \tag{1}$$

where y_i is the quantitative trait of individual i. Each x_{ij} with $j \in \{1, 2, 3\}$, is the dosage effect-SNP `rs4315144` (MAF = 0.3), `rs7956821`(MAF = 0.2), and `rs11830243` (MAF = 0.1), respectively. With the squared Pearson correlation R^2 used as a measure on how well the imputed SNP can be predicted based on the information of the known SNPs, we have an imputation quality of $R^2 = 0.998$ for `rs4315144`, $R^2 = 0.992$ for `rs7956821`, and $R^2 = 0.848$ for `rs11830243`. For the quantitive trait the SNP-effects are $\beta_1 = \beta_2 = 4$ and $\beta_3 = 6$, in the binary case the natural logarithm is taken of these values. A summary on the effect-SNPs can be found in Table 2. Last, $\epsilon_i \sim N(0, \sigma^2_\epsilon)$, with

$$\sigma^2_\epsilon = var\left(\sum_{j=1}^{p=3} x_j \beta_j\right)\left(\frac{1}{H^2} - 1\right). \tag{2}$$

The simulation model for the binary (bn) trait is based on a single threshold model. We set threshold $T = 1000/1557$ to be the pre-specfied proportion of controls and $k = 0.1$ is the pre-specified population disease prevalence. We assign the y_i of an individuals as a case if it exceeds the percentile ${}_{(T-k)\times 100}$ of the underlying marginal normal distribution, and as a control otherwise. In statistical terms this model is similar to using the Probit link, in which the cumulative probability of a normal distribution is modeled [2].

Note that the GCTA simulation approach is not the best one possible; with only three effect-SNPs and the pre-specified parameters one wishes to have in the data, we end up with a maximum of 156 possible cases for the current research strategy. Because the number of 1401 samples defined as controls exceed the pre-specified number of 1000 controls, GCTA "randomly" specifies 401 samples as missing. In other words, for the simulated binary trait we have a sample size equal to $n = 1156$.

Table 2: Information on the effect-SNPs

SNP	MAF	R^2	β_{qt}	β_{bn}	OR_{bn}
rs4315144	0.3	0.998	4	$\ln(4)$	4
rs7956821	0.2	0.992	4	$\ln(4)$	4
rs11830243	0.1	0.848	6	$\ln(6)$	6

The Comparison

The toolsets are compared on both feasibility and statistical accuracy. In the feasibility section of the current paper we present our experience on the install requirements, versatility, command line interface and the available help options. In the same section you can find the description on these feasibility properties. Testing for the statistical accuracy is based on a 5 (toolsets) x 6 (heritability) x 2 (phenotype) design. For each condition we run each of the toolsets on 100 simulated phenotypes to estimate the statistical power, the Type-I error, and the consistency and efficiency of the SNP-effects. We estimate the Type-I error based on $p = 7757$ pruned and unlinked SNPs from Chromosome 12. We used PLINK v1.12 to select these SNPs with the condition that all pairwise SNPs should have $R^2 < .1$ within a window size of 100 SNPs (shifted for 5 SNPs). Last, we used the 800,000 hard called genotyped markers for the kinship matrices to be calculated in EMMAX and ProbABEL.

Statistical Accuracy

The comparison of the toolsets on statistical accuracy is based on estimates of Type-I error, power, and consistency (and efficiency) of SNP-effects.

Binary Trait Type-I error

We compare the toolsets on the average genomic inflation factor ($\bar{\lambda}_{gc}$) and the average number of SNPs ($p\hat{\alpha}$) that have a p-value lower than the Bonferroni corrected α-level of 0.05/7757 as a measures of Type-I error. A toolset performs well if it has an $\bar{\lambda}_{gc}$ close or equal to one and a low number on the average number of ($\bar{p}\hat{\alpha}$). Besides the average we also give the standard deviation of λ_{gc} and the maximum number of Type-I errors in-between brackets in Table 3.

Table 3: Average number of Type-I error SNPs, $p\hat{\alpha}$ (maximum number of Type-I error SNPs) and the average genomic inflation factors $\bar{\lambda}_{gc}$ (standard deviation of λ_{gc}) of the binary trait replicates.

		ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
H = 5%	$p\hat{\alpha}$	0.00(0)	0.19(2)	4.41(11)	1.65(5)	18.56(34)
	$\bar{\lambda}_{gc}$	0.71(0.04)	0.98(0.04)	1.03(0.04)	1.00(0.04)	1.07(0.04)
H = 10%	$p\hat{\alpha}$	0.04(1)	0.59(3)	4.42(11)	1.69(6)	18.57(32)
	$\bar{\lambda}_{gc}$	0.71(0.02)	0.98(0.04)	1.03(0.04)	1.00(0.03)	1.08(0.04)
H = 20%	$p\hat{\alpha}$	0.02(1)	2.89(6)	4.52(11)	1.77(6)	18.65(32)
	$\bar{\lambda}_{gc}$	0.73(0.04)	1.00(0.04)	1.04(0.04)	1.02(0.04)	1.13(0.05)
H = 30%	$p\hat{\alpha}$	0.03(1)	4.02(7)	4.52(11)	1.89(5)	20.59(35)
	$\bar{\lambda}_{gc}$	0.71(0.02)	1.00(0.04)	1.04(0.04)	1.01(0.03)	1.14(0.05)
H = 50%	$p\hat{\alpha}$	0.06(2)	5.30(8)	4.65(11)	2.71(6)	21.76(39)
	$\bar{\lambda}_{gc}$	0.73(0.04)	1.02(0.04)	1.04(0.04)	1.02(0.03)	1.22(0.05)
H = 80%	$p\hat{\alpha}$	0.00(0)	7.21(10)	4.81(11)	2.28(8)	23.71(39)
	$\bar{\lambda}_{gc}$	0.73(0.02)	1.03(0.04)	1.04(0.04)	1.03(0.03)	1.36(0.06)

The toolsets differ on these measures for Type-I error with ProbABEL performing the worst. When the heritability and hence the need for control on family structure becomes higher, ProbABEL yields too liberal p-values and too high values of $\bar{\lambda}_{gc}$. Comparable to ProbABEL, but much less pronounced, is the EMMAX performance of EMMAX that has relatively more false discoveries when the heritability is higher. MERLIN and PLINK show consistent results over the different heritability levels. On average, Plink has in-between 1 and 3 false discovery SNPs out of 7757. ASSOC has the lowest $p\hat{\alpha}$, however, it produces too low estimates of $\bar{\lambda}_{gc}$.

Binary Trait Power

There are differences between the toolsets with respect to the statistical power for the binary trait. In Figure 1, in which the $-\log_{10}$(P-value) boxplots on the effect-SNPs in each heritability condition (H) are shown, we see that MERLIN (dark blue) yields low $-\log_{10}$(P-values) standing apart from the other packages. With the red dotted line indicating the α-level of 5×10^{-8} for standard GWA studies we see low statistical power estimates (< 0.25) for all toolsets in the heritability conditions of H = 5%, 10% and 20%. For heritabilities of 30% and higher sufficient power estimates (> 0.80) are obtained for the effect-SNPs. The difference in performance between the toolsets remains the same over the heritability conditions with EMMAX (orange), ProbABEL (blue), PLINK (green), ASSOC (turquoise) and MERLIN (dark blue) ordered from high to low statistical power, respectively. Although it looks like the order of these toolsets becomes more pronounced as does the increase in range of the $-\log_{10}$(P-values) when heritability becomes higher, it is just an artifact of $-\log_{10}$ transformations on the P-values. The results remain the same when we compare the packages on the $-\log_{10}$(P-values) for the separate effect-SNPs (see Table 4 and Figure 2).

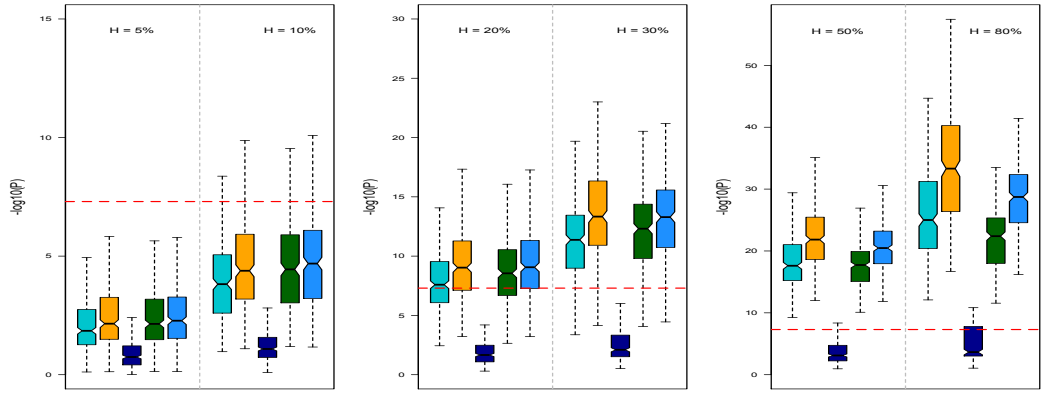


Figure 1: Boxplot on binary trait: Results for combined $-\log_{10}$(P-values) from the analyses performed by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions over the three effect-SNPs. The dotted red line indicates an α-level at 5×10^{-8}. Note the differences in the $-\log_{10}$(P-value) scale on the vertical-axis among heritability conditions.

The results remain the same when we compare the packages on the $-\log_{10}$(P-values) for the separate effect-SNPs (see Table 4 and Figure 2). Unexpected, however, is that the toolsets detect SNP `rs11830243` (slightly) faster than `rs4315144`. Although `rs11830243` has a larger SNP-effect of $\beta = 6$, its effect should contribute less to the trait than `rs4315144` ($\beta = 4$) because of its lowest MAF of 0.1.

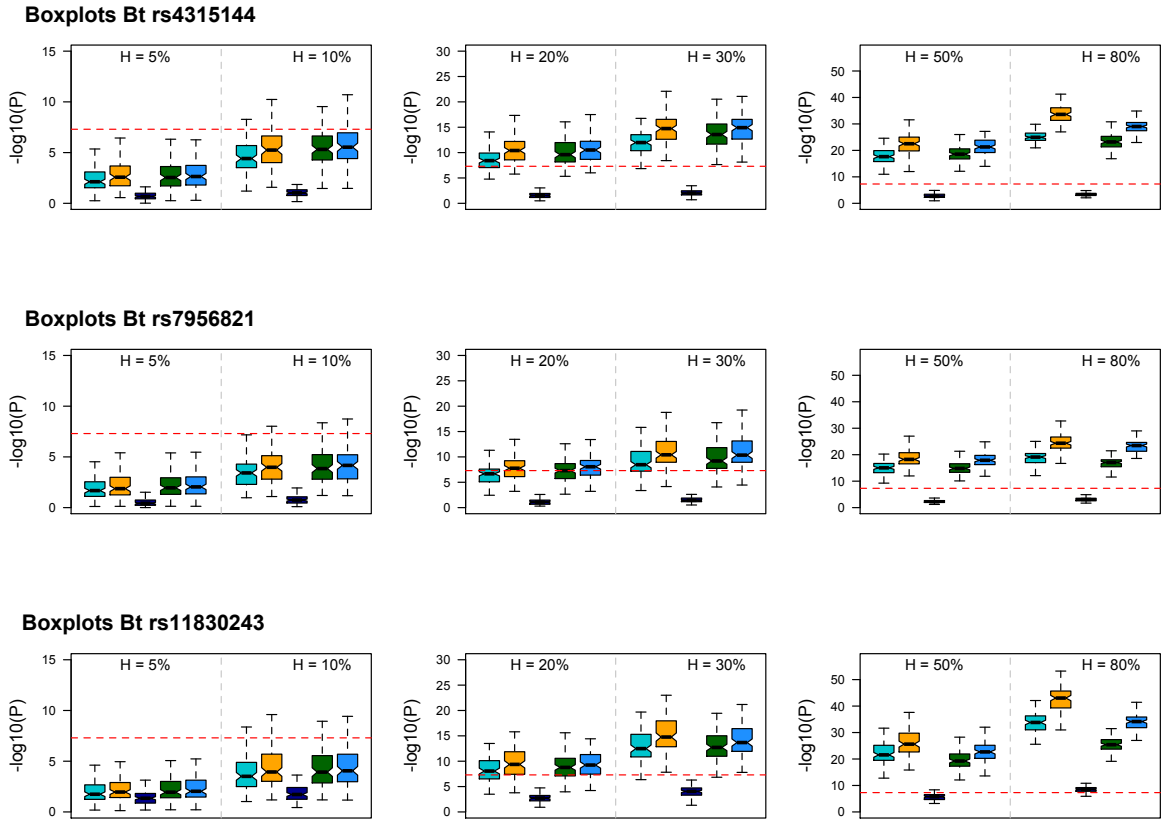


Figure 2: Boxplots on binary trait: Results for $-\log_{10}$(P-values) from the analyses for each SNP separately performed by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions for the separate SNPs. The red dotted line indicates an α-level at 5×10^{-8}. Note the differences in the $-\log_{10}$(P-value) scale on the vertical-axis among heritability conditions.

Table 4: The number of times the effect SNP got detected (out of a 100) against a threshold of $\alpha = 5 \times 10^{-8}$.

Heritability	ASSOC	EMMAX	MERLIN	ProbABEL	ProbABEL
`rs4315144`					
H = 5%	0	2	0	2	2
H = 10%	8	19	0	16	20
H = 20%	73	86	0	81	86
H = 30%	98	100	0	100	100
H = 50%	100	100	0	100	100
H = 80%	100	100	0	100	100
`rs7956821`					
H = 5%	1	1	0	1	1
H = 10%	2	8	0	7	9
H = 20%	30	57	0	50	60
H = 30%	74	92	0	87	97
H = 50%	100	100	0	100	100
H = 80%	100	100	0	100	100
`rs11830243`					
H = 5%	0	1	0	0	0
H = 10%	8	11	0	8	10
H = 20%	63	76	0	70	76
H = 30%	99	100	0	99	100
H = 50%	100	100	9	100	100
H = 80%	100	100	83	100	100

Binary Trait SNP-effects

The estimates of the SNP-effects do not appear to be consistent over 100 simulations. With the red dotted line in Figure 3 indicating no deviation from the true SNP-effect (β_{true}), it is shown that the SNP-effects of ASSOC, PLINK and ProbABEL deviate less from the true SNP-effect when the heritability becomes higher. Whereas we cannot compare the SNP-effects of EMMAX with the other toolsets since it treats the binary trait as quantitative, MERLIN is actually performing the worst. PLINK and ProbABEL yield the same SNP-effects up to the third decimal and overestimate the SNP-effects at the heritability condition of 80%. ASSOC, generating SNP-effects close to PLINK and ProbABEL, does not overestimate the SNP-effect, it actually approaches the β most closely with a sum of the squared error deviations (SS_{true}) of 5.88 in the heritability condition of 80% (see Table 5).

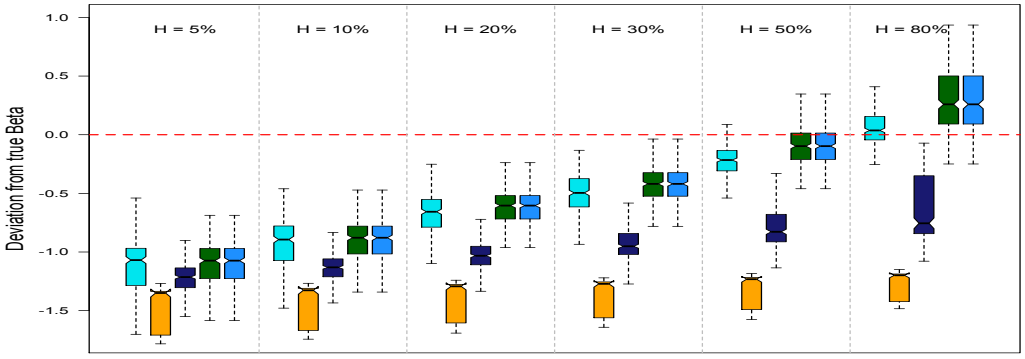


Figure 3: Binary trait boxplots on the deviations of the SNP-effect ($\beta - \beta_{true}$) of the analyses performed by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions (H). The red dotted line indicates $\beta - \beta_{true} = 0$.

Table 5: Sum of squared error deviations from the true SNP-effects (SS_{true}) for the binary trait

Heritability (H)	ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
H = 5%	387.24	655.11	455.00	371.00	371.00
H = 10%	272.98	629.98	391.59	251.29	251.29
H = 20%	146.58	593.56	320.58	120.42	120.42
H = 30%	82.86	567.78	258.13	57.94	57.94
H = 50%	20.95	527.54	190.19	10.96	10.96
H = 80%	5.88	490.11	142.55	45.83	45.83

When considering the separate SNP-effects only, presented in Figure 4 and Table 5, we see that differences between ASSOC on the one hand, and PLINK and ProbABEL on the other hand, are most pronounced for SNP `rs11830243`. ASSOC, PLINK and ProbABEL still give estimates that are most close to statistical the consistency when we consider the separate SNP-effects.

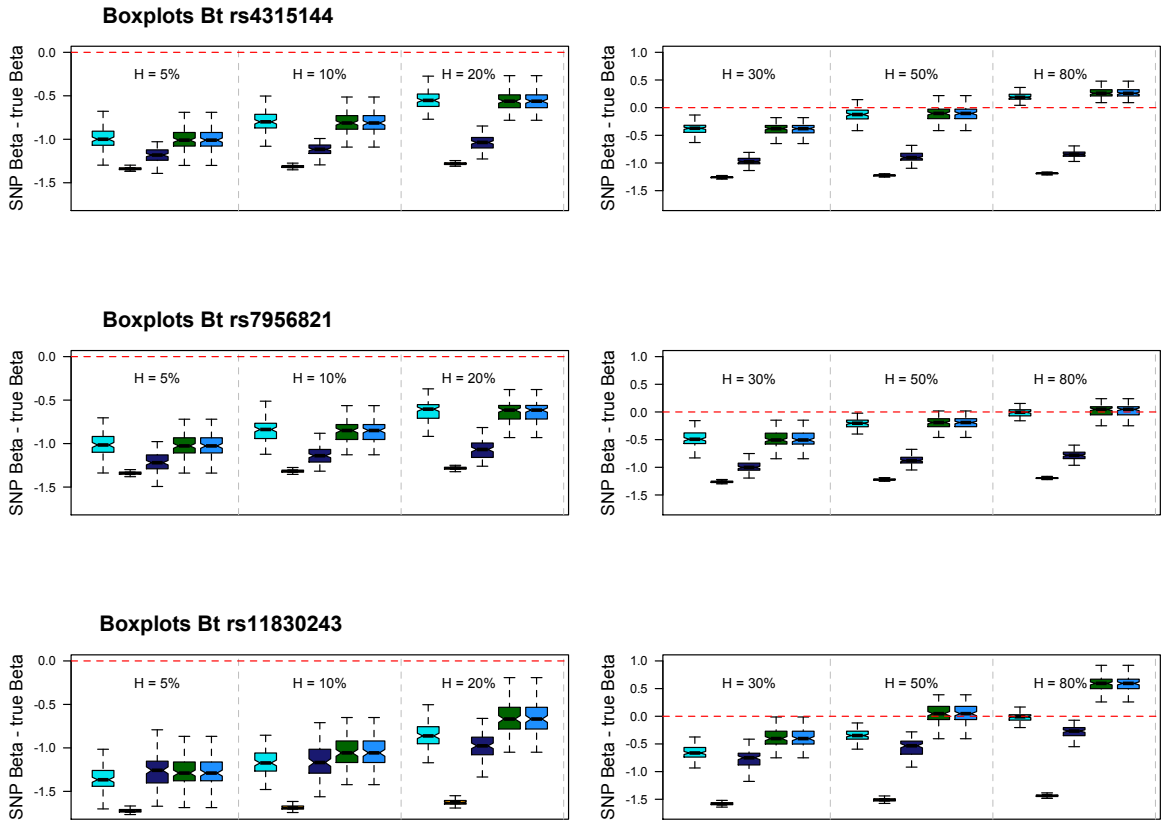


Figure 4: Quantitative trait boxplots on the deviations of the SNP-effects (β) from β_{true} of the analyses performed by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions. The red dotted line indicates $\beta - \beta_{true} = 0$.

Table 6: Mean and standard deviation of the estimated SNP-effects for the binary trait. Note that $\beta_{true} = \log(4) \approx 1.38$ for `rs4315144` and `rs7956821`, and $\beta_{true} = \log(6) \approx 1.79$ for `rs11830243`.

Heritability (H)	ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
`rs4315144`					
H = 5%	0.40(0.13)	0.05(0.02)	0.20(0.09)	0.39(0.12)	0.39(0.12)
H = 10%	0.59(0.12)	0.07(0.02)	0.27(0.07)	0.58(0.12)	0.58(0.12)
H = 20%	0.84(0.12)	0.11(0.02)	0.35(0.09)	0.83(0.12)	0.83(0.12)
H = 30%	1.01(0.11)	0.13(0.01)	0.41(0.08)	1.00(0.11)	1.00(0.11)
H = 50%	1.27(0.10)	0.16(0.01)	0.50(0.09)	1.28(0.12)	1.28(0.12)
H = 80%	1.58(0.08)	0.20(0.01)	0.55(0.07)	1.65(0.10)	1.65(0.10)
`rs7956821`					
H = 5%	0.38(0.15)	0.05(0.02)	0.17(0.12)	0.37(0.14)	0.37(0.14)
H = 10%	0.54(0.13)	0.07(0.02)	0.25(0.11)	0.53(0.13)	0.53(0.13)
H = 20%	0.77(0.12)	0.10(0.02)	0.32(0.11)	0.75(0.12)	0.75(0.12)
H = 30%	0.91(0.13)	0.12(0.02)	0.39(0.09)	0.90(0.13)	0.90(0.13)
H = 50%	1.18(0.10)	0.17(0.01)	0.52(0.09)	1.20(0.11)	1.20(0.11)
H = 80%	1.37(0.08)	0.19(0.01)	0.60(0.08)	1.41(0.10)	1.41(0.10)
`rs11830243`					
H = 5%	0.44(0.14)	0.07(0.03)	0.52(0.18)	0.52(0.17)	0.52(0.17)
H = 10%	0.64(0.15)	0.11(0.03)	0.64(0.18)	0.76(0.18)	0.76(0.18)
H = 20%	0.93(0.13)	0.16(0.03)	0.81(0.15)	1.13(0.17)	1.13(0.17)
H = 30%	1.15(0.12)	0.21(0.03)	1.02(0.17)	1.41(0.16)	1.41(0.16)
H = 50%	1.45(0.12)	0.28(0.03)	1.23(0.16)	1.85(0.19)	1.85(0.19)
H = 80%	1.77(0.08)	0.36(0.02)	1.51(0.12)	2.38(0.15)	2.38(0.15)

Table 7: The average on the standard errors of the SNP-effects given by the toolsets (left), and the standard deviation of the SNP-effects of the binary trait replicates (right).

Heritability (H)	ASSOC*		EMMAX	MERLIN		PLINK		ProbABEL	
`rs4315144`									
H = 5%	0.14	0.13	-	0.16	0.09	0.13	0.12	0.12	0.12
H = 10%	0.14	0.12	-	0.16	0.07	0.13	0.12	0.12	0.12
H = 20%	0.14	0.12	-	0.16	0.09	0.13	0.12	0.12	0.12
H = 30%	0.14	0.11	-	0.16	0.08	0.13	0.11	0.13	0.11
H = 50%	0.14	0.10	-	0.16	0.09	0.14	0.12	0.13	0.12
H = 80%	0.15	0.09	-	0.16	0.07	0.16	0.10	0.15	0.10
`rs7956821`									
H = 5%	0.16	0.15	-	0.19	0.12	0.14	0.14	0.14	0.14
H = 10%	0.16	0.13	-	0.19	0.11	0.14	0.13	0.14	0.13
H = 20%	0.15	0.12	-	0.19	0.11	0.14	0.12	0.13	0.12
H = 30%	0.15	0.13	-	0.19	0.09	0.14	0.13	0.13	0.13
H = 50%	0.15	0.10	-	0.18	0.09	0.15	0.11	0.14	0.11
H = 80%	0.15	0.08	-	0.18	0.08	0.17	0.10	0.14	0.10
`rs11830243`									
H = 5%	0.18	0.14	-	0.26	0.18	0.20	0.17	0.19	0.17
H = 10%	0.17	0.15	-	0.27	0.18	0.19	0.18	0.19	0.18
H = 20%	0.16	0.13	-	0.26	0.15	0.19	0.17	0.18	0.17
H = 30%	0.15	0.12	-	0.26	0.17	0.19	0.16	0.18	0.16
H = 50%	0.15	0.12	-	0.26	0.16	0.20	0.19	0.18	0.19
H = 80%	0.15	0.08	-	0.26	0.12	0.22	0.15	0.19	0.15

*Standard errors of the ASSOC SNP-effects have been back-calculated from the p-value and the SNP-effect using the $\chi^2_{(1)}$-distribution.

From Table 7 we obtain the average on standard errors of the SNP-effects as given by the toolsets as well as the standard deviation of the SNP-effects of the binary trait replicates. Estimates for EMMAX are not displayed because its SNP-effects assume the trait to be quantitative. Noteworthy is that each average on the standard errors obtained for MERLIN is twice as high compared to its standard deviation of the SNP-effects, while the opposite is yielded for the other toolsets. The averages on the standard errors obtained from ASSOC, PLINK and ProbAbel are larger as compared to the standard deviation of the corresponding SNP-effects. PLINK and ProbABEL have lower standard error estimates for SNPs `rs4315144` and `rs7956821`. ASSOC has the lowest standard errors for `rs11830243`.

Quantitative Trait Type-I error

Although all toolsets show worse performance when heritability increases, there are differences in the performance on statistical accuracy when the null hypothesis is true (see Table 8). ASSOC has best control on the number of Type-I error SNPs. However, its average on the genomic inflation factor becomes too low when the heritability is 50% and 80%. PLINK is second when it comes to the control of Type-I error SNPs, and first on the average genomic inflation. The average performance of EMMAX and MERLIN is around the same. One could say however that MERLIN performs worst since it has larger standard deviations on genomic inflation and higher maxima on Type-I error SNPs. ProbABEL is performing worst with a maximum of 145 and 159 Type-I error SNPs (due to non-convergence) for a heritability of 50% and 80%, respectively.

Table 8: Average number of Type-I error SNPs, $p\hat{\alpha}$ (maximum number of Type-I error SNPs), and the average genomic inflation factors $\bar{\lambda}_{gc}$ (standard deviation of λ_{gc}) of the quantitative trait replicates.

		ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
H = 5%	$p\hat{\alpha}$	0.02(1)	1.42(16)	4.56(13)	0.52(4)	0.55(3)
	$\bar{\lambda}_{gc}$	0.94(0.04)	0.97(0.04)	1.01(0.11)	1.00(0.04)	0.97(0.04)
H = 10%	$p\hat{\alpha}$	0.03(1)	3.46(6)	5.17(13)	0.87(3)	1.57(4)
	$\bar{\lambda}_{gc}$	0.93(0.03)	0.98(0.04)	1.01(0.11)	0.99(0.04)	0.97(0.03)
H = 20%	$p\hat{\alpha}$	0.06(1)	4.96(7)	6.04(13)	1.75(4)	3.35(6)
	$\bar{\lambda}_{gc}$	0.92(0.03)	0.98(0.04)	1.01(0.11)	0.99(0.03)	0.98(0.03)
H = 30%	$p\hat{\alpha}$	0.02(1)	6.37(9)	6.64(13)	2.21(5)	4.58(6)
	$\bar{\lambda}_{gc}$	0.91(0.03)	0.98(0.04)	1.02(0.11)	1.01(0.03)	0.98(0.03)
H = 50%	$p\hat{\alpha}$	0.06(2)	7.82(11)	7.42(13)	2.93(8)	7.92(145*)
	$\bar{\lambda}_{gc}$	0.89(0.03)	1.00(0.03)	1.02(0.11)	1.00(0.03)	1.03(0.17)
H = 80%	$p\hat{\alpha}$	0.14(1)	9.94(12)	9.17(13)	4.36(7)	14.23(159*)
	$\bar{\lambda}_{gc}$	0.85(0.02)	1.01(0.03)	1.03(0.11)	1.00(0.02)	1.12(0.13)

* No convergence was achieved for the parameter estimates of the estimated covariance matrix in 2 of the 100 simulation replicates

Quantitative Trait Power

The toolsets perform equally well on the statistical power for the quantitative trait. All packages achieve sufficient power (> 0.80) already at the heritability level of 10%. Due to the similarity between the packages it is difficult to rank the toolsets on statistical power. We can say, however, that results show the highest $-\log_{10}$(P-values) for EMMAX and the lowest for ASSOC (Figure 5 and Table 9).

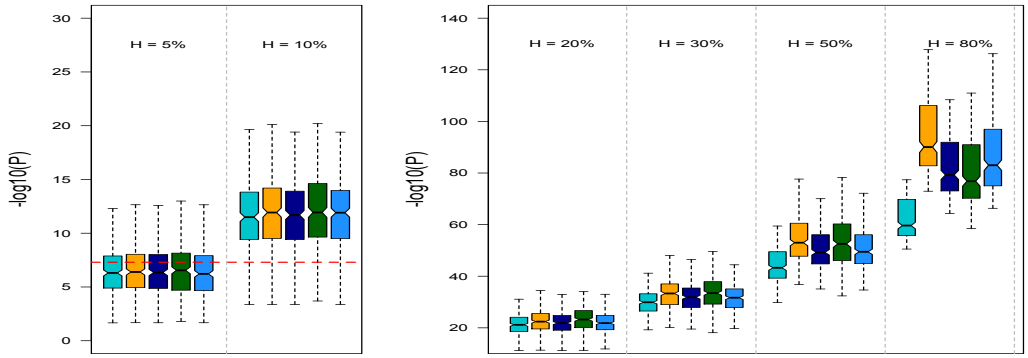


Figure 5: Boxplots on quantitative trait: Results for combined $-\log_{10}$(P-values) from the analyses performed by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions. The red dotted line indicates an α-level at 5×10^{-8}. Note the differences in the $-\log_{10}$(P-value) scale on the vertical-axis among heritability conditions.

Figure 6 shows the results for each SNP separately. The $-\log_{10}$(P-value) results for each effect SNP are similar to those of the results of the effect-SNPs taken together (Figure 5). Hence, highest $-\log_{10}$(P-values) are produced by EMMAX and the lowest by ASSOC, see Figure 6.

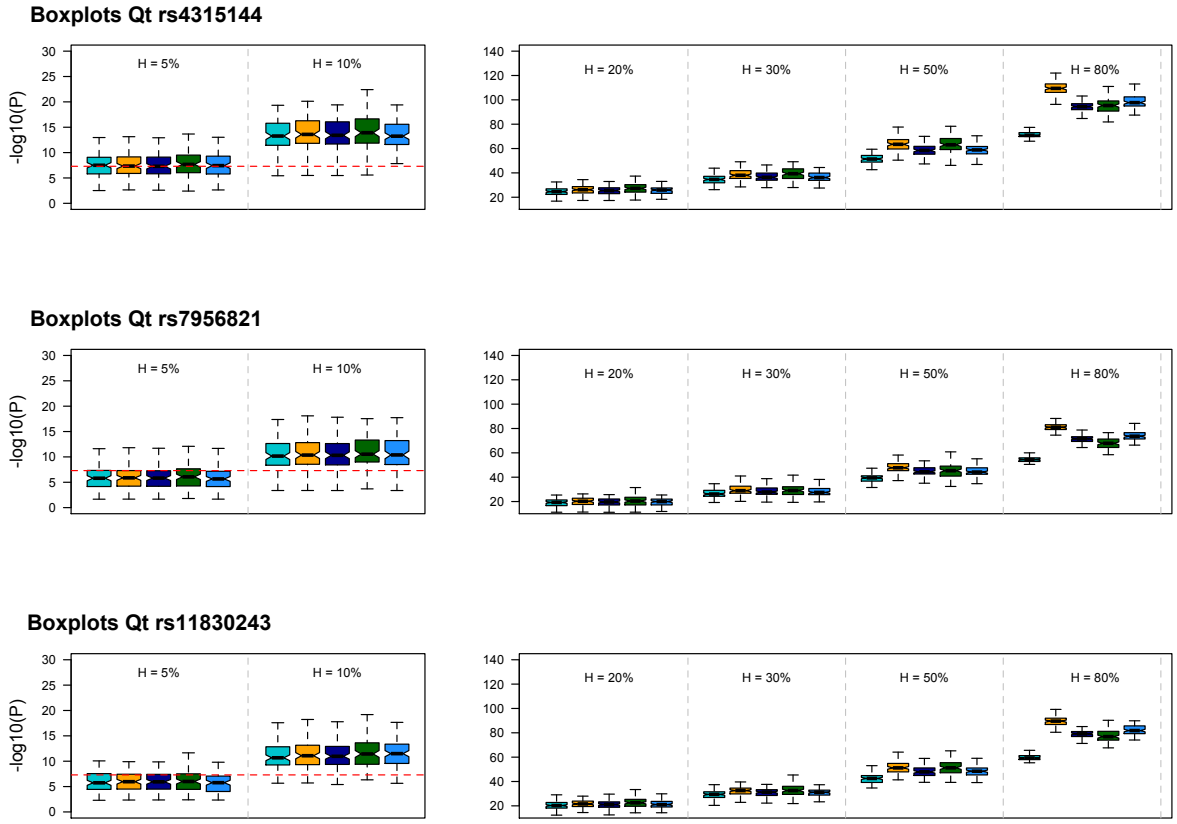


Figure 6: Boxplots on quantitative trait: Results for $-\log_{10}$(P-valRues) from the analyses for each SNP separately by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions for the separate SNPs. The red dotted line indicates an α-level at 5×10^{-8}. Note the differences in the $-\log_{10}$(P-value) scale on the vertical-axis among heritability conditions.

Table 9: The number of times the SNP was detected out of 100 trials against a threshold of $\alpha = 5 \times 10^{-8}$ for the quantitative trait.

Heritability	ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
`rs4315144`					
H = 5%	52	52	50	56	50
H = 10%	99	99	99	99	97
H = 20%	100	100	100	100	100
H = 30%	100	100	100	100	100
H = 50%	100	100	100	100	100
H = 80%	100	100	100	100	100
`rs7956821`					
H = 5%	26	26	24	29	18
H = 10%	89	89	90	87	91
H = 20%	100	100	100	100	100
H = 30%	100	100	100	100	100
H = 50%	100	100	100	100	100
H = 80%	100	100	100	100	100
`rs11830243`					
H = 5%	28	28	28	29	23
H = 10%	96	97	97	95	97
H = 20%	100	100	100	100	100
H = 30%	100	100	100	100	100
H = 50%	100	100	100	100	100
H = 80%	100	100	100	100	100

Quantitative Trait SNP-effects

As can be concluded from Table 10 and Figure 7), the consistency results based on the estimated SNP-effect results on the quantitative trait are very similar the different toolsets. Small differences occur when the heritability becomes higher. EMMAX, MERLIN and ProbABEL slightly overestimate the SNP-effects with a sum of squared error deviations from β_{true} (SS_{true}) of 9.82, 11.00, and 11.83, respectively. ASSOC and PLINK are more consistent on the SNP-effects with the same SS_{true} of 4.02 when the heritability is 80%. If we split out the results on consistency for separate SNPs, results remain the same, see Figure 8 and Table 11.

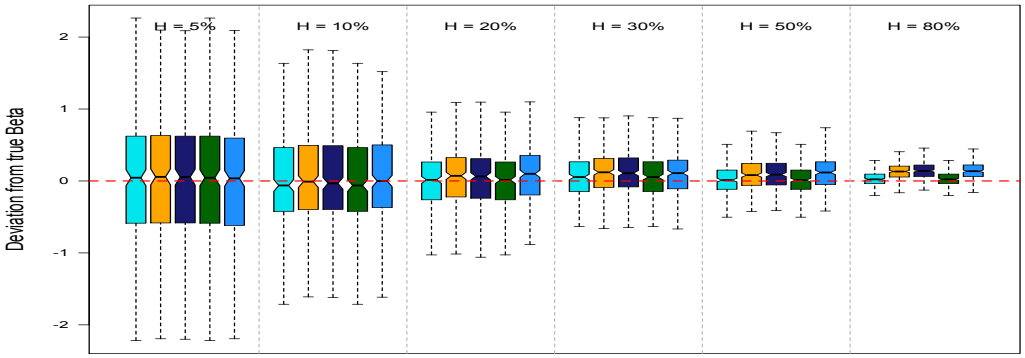


Figure 7: Quantitative trait boxplots on the deviations of the SNP-effects (β) from β_{true} of the analyses performed by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions. The red dotted line indicates $\beta - \beta_{true} = 0$.

Table 10: Sum of squared error deviations from the true SNP-effects (SS_{true}) for the quantitative trait

Heritability (H)	ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
H = 5%	260.53	262.33	262.15	260.52	268.79
H = 10%	125.90	126.39	126.92	125.90	125.80
H = 20%	52.56	55.68	54.96	52.56	58.27
H = 30%	30.90	35.86	35.71	30.90	36.94
H = 50%	15.19	19.21	19.81	15.19	21.85
H = 80%	4.02	9.82	11.00	4.02	11.83

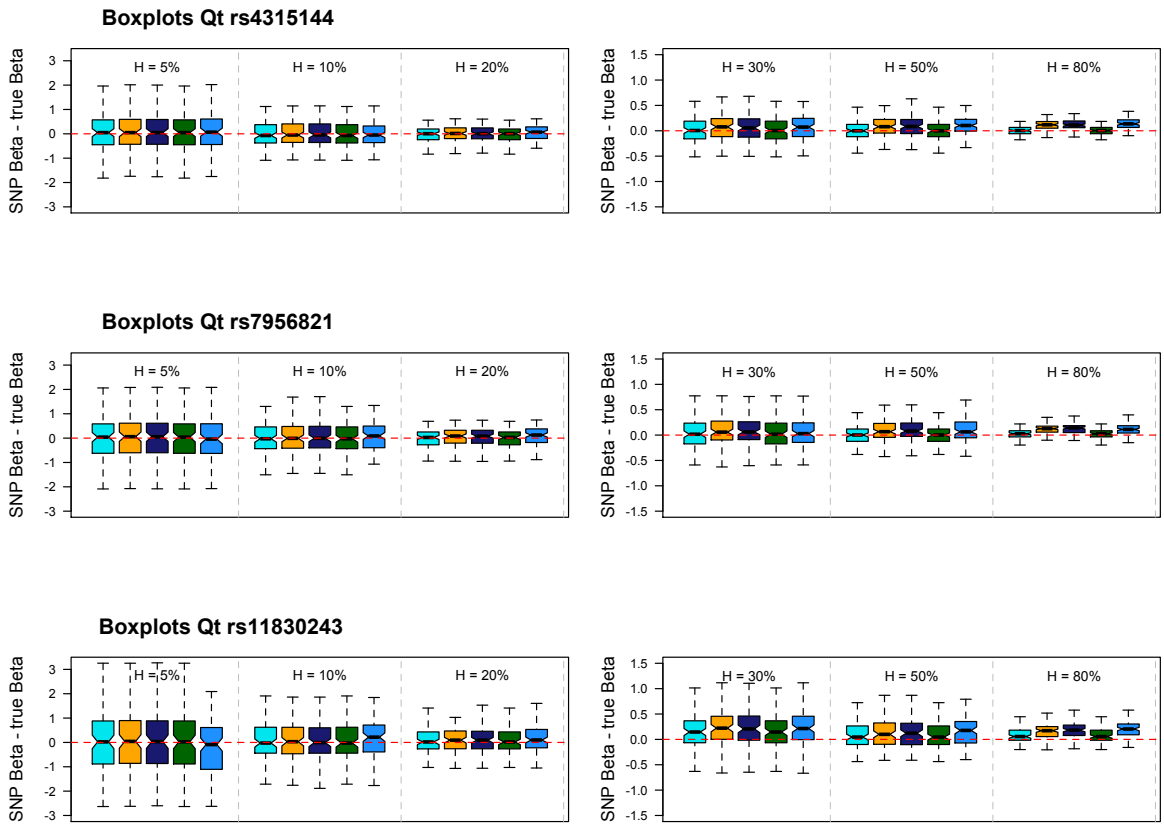


Figure 8: Quantitative trait boxplots on the deviations of the SNP-effects (β) from β_{true} of the analyses performed by ASSOC (turquoise), EMMAX (orange), MERLIN (dark blue), PLINK (green), ProbABEL (blue) in the six different heritability conditions. The red dotted line indicates $\beta - \beta_{true} = 0$.

Table 11: Mean and standard deviation of the estimated SNP-effects for the quantitative trait.

Heritability (H)	ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
`rs4315144`					
H = 5%	4.05(0.74)	4.06(0.74)	4.06(0.74)	4.05(0.74)	4.05(0.76)
H = 10%	3.98(0.50)	4.01(0.51)	4.00(0.51)	3.98(0.50)	3.96(0.53)
H = 20%	3.98(0.31)	4.02(0.32)	4.02(0.32)	3.98(0.32)	4.05(0.32)
H = 30%	4.01(0.24)	4.07(0.25)	4.06(0.25)	4.01(0.24)	4.06(0.24)
H = 50%	4.00(0.18)	4.09(0.20)	4.08(0.20)	4.00(0.18)	4.10(0.21)
H = 80%	4.00(0.08)	4.12(0.09)	4.12(0.10)	4.00(0.08)	4.14(0.10)
`rs7956821`					
H = 5%	4.02(0.87)	4.03(0.88)	4.03(0.88)	4.02(0.87)	3.94(0.84)
H = 10%	4.00(0.63)	4.02(0.63)	4.02(0.63)	4.00(0.63)	4.05(0.59)
H = 20%	3.97(0.39)	4.02(0.30)	4.02(0.40)	3.97(0.39)	4.06(0.40)
H = 30%	4.03(0.29)	4.09(0.30)	4.09(0.30)	4.03(0.29)	4.07(0.31)
H = 50%	4.00(0.21)	4.09(0.20)	4.10(0.20)	4.00(0.21)	4.09(0.23)
H = 80%	4.02(0.09)	4.12(0.10)	4.13(0.11)	4.02(0.09)	4.10(0.11)
`rs11830243`					
H = 5%	5.96(1.15)	5.96(1.15)	5.95(1.14)	5.96(1.15)	5.90(1.19)
H = 10%	6.05(0.78)	6.07(0.79)	6.05(0.79)	6.05(0.78)	6.14(0.79)
H = 20%	6.07(0.52)	6.11(0.54)	6.10(0.53)	6.07(0.52)	6.18(0.54)
H = 30%	6.15(0.38)	6.22(0.39)	6.20(0.40)	6.15(0.38)	6.22(0.40)
H = 50%	6.07(0.27)	6.13(0.29)	6.13(0.29)	6.07(0.27)	6.16(0.28)
H = 80%	6.08(0.14)	6.17(0.15)	6.19(0.15)	6.08(0.14)	6.21(0.15)

The averages on the standard errors of the SNP effects do not to differ across the toolsets, as can be seen in Table 12 below. The higher the heritability of the SNP-effects, the lower the average on the standard errors, as well as the standard deviation of the SNP-effects and the larger the difference.

Table 12: The average on the standard errors of the SNP-effects given by the toolsets (left), and the standard deviation of the SNP-effects (right) of the quantitative trait replicates.

Heritability (H)	ASSOC*		EMMAX*		MERLIN		PLINK		ProbABEL	
`rs4315144`										
H = 5%	0.74	0.74	0.74	0.74	0.74	0.74	0.73	0.74	0.74	0.76
H = 10%	0.53	0.50	0.52	0.51	0.53	0.51	0.51	0.50	0.53	0.53
H = 20%	0.38	0.31	0.38	0.32	0.38	0.32	0.36	0.32	0.38	0.32
H = 30%	0.32	0.24	0.31	0.25	0.32	0.25	0.30	0.24	0.32	0.24
H = 50%	0.26	0.18	0.24	0.20	0.25	0.20	0.23	0.18	0.25	0.21
H = 80%	0.22	0.08	0.18	0.09	0.20	0.10	0.18	0.08	0.19	0.10
`rs7956821`										
H = 5%	0.85	0.87	0.85	0.88	0.85	0.88	0.84	0.87	0.85	0.84
H = 10%	0.61	0.63	0.60	0.63	0.61	0.63	0.59	0.63	0.61	0.59
H = 20%	0.44	0.39	0.43	0.30	0.44	0.40	0.42	0.39	0.44	0.40
H = 30%	0.37	0.29	0.36	0.30	0.37	0.30	0.35	0.29	0.37	0.31
H = 50%	0.30	0.21	0.28	0.20	0.29	0.20	0.27	0.21	0.29	0.23
H = 80%	0.26	0.09	0.21	0.10	0.23	0.11	0.22	0.09	0.22	0.11
`rs11830243`										
H = 5%	1.24	1.15	1.24	1.15	1.24	1.14	1.22	1.15	1.24	1.19
H = 10%	0.89	0.78	0.88	0.79	0.89	0.79	0.86	0.78	0.89	0.79
H = 20%	0.64	0.52	0.63	0.54	0.64	0.53	0.60	0.52	0.64	0.54
H = 30%	0.54	0.38	0.52	0.39	0.53	0.40	0.50	0.38	0.53	0.40
H = 50%	0.44	0.27	0.40	0.29	0.42	0.29	0.39	0.27	0.42	0.28
H = 80%	0.37	0.14	0.31	0.15	0.33	0.15	0.31	0.14	0.32	0.15

*Standard errors of the SNP-effects for ASSOC and EMMAX have been back-calculated from the p-value and the SNP-effect using the $\chi^2_{(1)}$-distribution.

Statistical Accuracy summary

There is no clear winner on statistical accuracy among the toolsets. On the binary trait PLINK is recommended in the current design. Although PLINK is second on statistical power and Type-I error control, it has the best properties on genomic inflation. In EMMAX, the SNP-effects cannot be easily interpreted as the more usual odds-ratio. It needs further calculations from the expected allele frequency obtained from the given SNP-effect to predict for cases and controls. MERLIN is unable to deal with binary trait data and heavily underestimates the SNP-effects in the current design for heritability levels lower than 80%. ASSOC, PLINK and ProbABEL also underestimate the SNP-effects. Note however, for the high heritability level of 80% the effects are overestimated in the current design for PLINK and ProbAbel. At the cost of a low genomic inflation factor, ASSOC comes with better estimates of the SNP-effects with sufficient power.

The differences in statistical accuracy for the toolsets on the quantitative trait are less clear. Small differences in statistical power and SNP-effects occur only when the heritability becomes higher. The number of Type-I error SNPs and the average genomic inflation factor show a positive relation with the heritability conditions. A bias which is most pronounced in the toolset ProbABEL. The estimates on the average number of false discoveries could be biased however due to non-convergence of parameter estimation. An opposite relation, but less pronounced, has been found for ASSOC on the average genomic inflation factor being negatively associated with the heritability conditions. For the heritability levels of 5% and 10%, ASSOC and PLINK perform best. When the heritability levels are at 20% or higher, PLINK performs best.

Feasibility

A summary is presented for each toolset on feasibility properties. The choices for these properties are based on the authors' evaluation from hands-on experience of two empirical GWA studies and the analyses performed on the simulated data for the current paper only, and therefore may be biased. Below we give an overview on the evaluation of install requirements, versatility, command line interface.

Feasibility Properties

Install requirements

The install requirements are evaluated based on the following properties:

> *64 bits operating systems (OS)*: whether the toolset is compatible with Windows XP or higher, Linux Ubuntu 10.04 or higher, or Mac OsX 10.04 or higher. Being able to deal with one OS only yields the symbol – , when able to deal with 2 OS the value of ± is obtained. In the results we give the symbol + to toolsets when it is able to deal with these three (or more) operating systems;
>
> *Source code*: is the source code available? (no = –, upon request = ±, yes = +);
>
> *Executables*: is an executable file available for all operating systems on which the toolset is able to run? (no = –, upon request = ±, yes = +);
>
> *Standalone*: Once compiled or using the executable file directly, does the toolset depends on other toolsets? (no = –, upon request = ±, yes = +);
>
> *Update*: whether an update-check is performed to see whether the newest version is used (no = –, yes = +). The symbol '±' is given to the toolset when it is shown in the output what version is being used;

On the overall PLINK has the highest score since it meets all install requirements. MERLIN comes close, it only misses the update-check. ASSOC has got the source code available upon request and is not standalone since it depends on MERLIN for calculations of the Kinship matrix. EMMAX only runs on Linux 64-bits operating systems and has its source code available on request. Last, ProbABEL does not have executable files available for all the operating systems with which it is compatible and is not standalone. Furthermore, it depends on the statistical package `R` [36] with the library GenABEL [4] to transform the data and calculate the marker-based kinship matrices. For the detailed scores, see Table 13

Versatility

Feasibility on versatility is evaluated on the following properties:

Data input: 1) The amount of files that should be read into the software, 2) whether both long format (SNPs x samples) or ped-format (samples x SNPs) is possible to upload, and 3) the amount of data management that is needed to structure the files from either Mach [26], Beagle [9], or Impute [19] format into the data structure of the toolset. The + symbol is given if the toolset is compatible with both long and ped format from multiple programs directly. A toolset obtains a value of ± if it can handle either long or ped format only, but it reads data output from multiple imputation packages. If otherwise, the symbol – is given.

Kinship: Whether the implemented kinship structure is based on the pedigree (P) and whether it allows for MZ-pairs (P_{mz}), or whether the kinship is marker based (M);

Sample size: The sample size for which we are able to perform a Genome Wide Association, on the 22 autosomal chromosomes, within two days using sequential scripts (and no parallel programming) on a Snow Leopard Mac OS X Version 10.6.7 (2.66GHz Intel Core i7 with 8GB 1067 MHz DDR3). We assign a sample size which is Small (S): for less than 1500; Medium (M): between 1500 and 5000 and Large (L): larger than 5000.

Computing hours: Using the same Mac OS computer we describe the approximate number of hours we needed to perform the whole GWA analysis on 2,500,000 using a sequential script (no parallel programming).

On some relevant versatility properties we did not compare the toolsets. These properties are whether the toolset is [i] able to perform an automatic GWA study on all input SNPs, [ii] can take up extra covariates, and [iii] the analysis code can be easily incorporated for parallel scripting. All toolsets conform to these criteria. An exception, however, occurs for for ProbABEL since it should be taken into account that the dependence on the statistical package `R` with its GenABEL library renders parallel scripting difficulties. To create a standalone package of `R` with the GenABEL library one should be familiar on compiling software programs.

ASSOC scores the highest on versatility. It needs 7 hours to complete the whole GWA analysis on the Mac OS X Version (2.66GHz Intel Core i7 with 8GB 1067 MHz DDR3) for both quantitative and binary trait on the 1557 samples. Moreover, it needs only two input files which it can read in long format as well as ped-format genotypic data. MERLIN approaches closely on versatility, with the only difference that it needs 4 (instead of 2) input files which involves much data management, which is prone to scripting errors. EMMAX also takes 7 hours for the whole GWA analysis but estimates the kinship structure based on the markers. A drawback on the feasibility, however, is that EMMAX, only allows a long-format-file as genotypic data input.

PLINK (for both traits) and ProbABEL (on the binary trait) are not able to deal with monozygotic twins. Whereas ProbABEL does not allow for any covariance between individuals, PLINK assumes the covariances between individuals to be the same. Although both toolsets take 9 hours to complete the whole GWA analysis, we expect both PLINK and ProbABEL (on the binary trait) to be relatively less demanding on computing time when the sample is larger. The reason is that the computing time for the estimation of the kinship matrix from the markers is a quadratic function of the sample size.

ProbABEL for both the binary and quantitative traits can deal with MACH, IMPUTE, PED, and long-format files as data input. However, these files are transformed in the package GenABEL in the statistical package `R`. A statistical package known for not being able to handle large data set files that well. Last, the estimation of the kinship and estimated covariance can take up to 9 hours in ProbABEL for the quantitative trait and it needs at least three hours in addition to test the null hypothesis for each SNP. Hence, we score ProbABEL on the quantitative trait as being only able to deal with small sample sizes. For the detailed scores, see Table 13

Command Line Interface (CLI)

The command language interface (CLI), described as being the interaction possibilities between user and the command line interpreter, is evaluated on the following:

> *Output*: Whether the output is complete and stored in a file. We judge the output to be complete when the name of the SNP, the SNP-effect, standard error of the SNP-effect, test statistic, and the unadjusted p-value are reported. The – symbol is given when it is unattainable to recalculate the p-value from the results given, we give the symbol ± if either the beta/ standard error or p-value needs to be recalculated, the symbol + is given when we assume the output to be complete;
>
> *Log-file*: Whether a log-file is generated as output in which the used function, descriptive measures on the input data and error/warning messages are given (no = –, upon request = ±, yes = +);

PLINK is the only toolset qualifying with the highest scores on all the CLI feasibility properties. For the output-files of ASSOC further calculations are needed to obtain the SNP-effects and its standard errors for the binary phenotypes. The output files from ProbABEL does not give the p-values. The EMMAX output only gives the SNP name, SNP-effect and P-value. MERLIN has better feasibility performance on the CLI since it gives the complete output-file. However, it does not provide a log-file with its results on which we could check what function is used on what specific data. Besides PLINK, only EMMAX provide its output with a log-file. However, from the log-file in EMMAX it is not clear what function code is being used.

Help options

Another property on which we evaluate the toolsets are the help options:

Documented: Are the available functions to perform GWA clearly documented? (no = –, upon request = ±, yes = +);

Tutorial: Whether there is a tutorial with example files available to get a first hand on how to conduct the GWA study using the toolset? (no = –, yes = +);

Manual: Is there extra explanation available in manuel format to grasp the analyses being done? (no = –, upon request = ±, yes = +);

Literature: are the toolsets and or functions being used published and peer reviewed? (no = –, some = ±, yes = +);

Other: are there communities / fora or other communication channels available on which questions can be posed, which would not be answered by the creators of the toolset via e-mail? (no = –, upon request = ±, yes = +).

ProbABEL performs best on the feasibility of help options. It is the only toolset of which the documentation clearly links the formula of the test statistic to the function code in the toolset for the GWA analyses. Moreover, it has active fora and a community on which the developers of the toolset and its statistical procedures post messages and give answers to most questions. Furthermore, only ProbABEL and EMMAX got the specific statistical procedure together with its toolset published in a journal [5, 22]. ASSOC has its statistical procedures published, but there is no published journal-article available the toolset (yet). The toolsets MERLIN and PLINK have been published [1, 35], but the specific statistical procedures on the dosage data have not been published. For MERLIN we eventually assumed that the family based score test [10] is the one implemented in the MERLIN-offline package. It is from correspondence with one of the developers of PLINK that we got to know what specific kind of GEE the toolset is performing. The specific scores on the Help options of each package can be obtained from Table 13.

Multi-purposes

The last property on which we evaluate the toolsets is wether the toolsets can handle more properties relevant for GWA studies on soft-called family data, but less relevant for the current design. The symbol '+' is given to ASSOC because it has the possibility to control for ascertainment and and include X-linked SNPs for the GWA analyses. MERLIN is also able to include X-linked SNPs in the GWA analyses, and therefore is scored with a ±. Last, we score a + for PLINK since it has very good properties for data management and quality control, see Table 13.

Feasibility overview

In the Table 13 below, we summarize our findings on the feasibility properties of each toolset. Noted that each "+" given is not necessarily to be equally weighted. Since the authors value highest diligence on install requirements and the CLI the most for performance of a package on feasibility, PLINK is performing best. However, we are well aware that the reader could come to a different conclusion and is very much welcome to join the discussion on the feasibility performance of toolsets.

Table 13: Feasibility properties of the toolsets.

		ASSOC	EMMAX	MERLIN	PLINK	ProbABEL
Install	OS	+	–	+	+	+
	Open Source	±	±	+	+	+
	Executable	±	+	+	+	–
	Standalone	–	+	+	+	–
	Update	–	–	±	+	±
Versatility	Data Input	+	–	–	±	+
	Kinship[a]	P_{mz}	M	P_{mz}	C	None / M
	Sample Size[b]	L	M	L	M	M / S
	Computing hours	7	7	7	8	14
CLI	Output	±	–	+	+	±
	Logfile	–	+	–	+	–
Help	Documentation	+	+	+	+	+
	Tutorial	+	–	–	+	+
	Manual	–	–	–	–	+
	Literature	±	+	±	±	+
	Fora/Other	–	–	±	±	+
Multipurpose		+	–	+	+	–

[a] Compatibility with the Operating Systems (OS) Linux Ubuntu ..., Windows XP Mac OSx 10.4, or its upgraded versions (– = 1 OS, + = all three OS)
[b] P = kinship based on pedigree, P_{mz} = kinship based on pedigree that allows for MZ twin-paris, and M = marker based kinship.
NOTE: estimates on Service and CLI errors are approximate and based on authors' experience rather than exhaustive testing.

Discussion

In the current monograph we present a research strategy which we compare the performance of toolsets that are able to deal with GWAS on so called soft-called family data for both quantitative and binary traits. Although other studies have reported comparison of performance on toolsets dealing with family data [28, 50], the present one is the first with respect to analyses performed by toolsets that can deal with imputation uncertainty of the genotype as well as MZ twins.

Our results show no toolset clearly outperforms all other toolsets on the feasibility properties and the measures of statistical accuracy that are examined. Overall, we recommend PLINK for GWA on the genotype imputed family data and simple traits in the current design. It performs best on the feasibility properties and it has a good combination of statistical power, false discovery control, and genomic inflation. The recommendation is mainly made for the case of a binary trait phenotype; for the quantitative trait, ASSOC performs equally well for low heritability levels. Last, we recommend not to use ProbABEL or MERLIN for GWA on binary traits when dealing with soft-called family data.

The explanation of the good performance of PLINK on statistical accuracy lies within the modeling of the genotypic relations of the statistical procedures. Our results come from a design in which three causal large-effect-SNPs and residual error determine the phenotypic trait. Except for PLINK and ProbABEL in the binary trait case, the statistical procedures in the toolsets assume many small-effect-SNPs to be able to control for family structure using the kinship matrix. The decomposition of these covariance of these effect-SNPs is not the same as specified structure by the kinship on the 800,000 markers or the IBD estimated from the pedigree. Moreover, there will still be a large proportion of pairs in the sample that will covary on these effect-SNPs while assumed to be unrelated with a kinship coefficient of zero. Hence, the phenotypic variance due to the genotypes does not split up in the expected terms determined by the kinship as is assumed in ASSOC, EMMAX, MERLIN and ProbABEL. The larger the heritability (the effect sizes of the SNPs), the larger the Type-I error when the genetic relations are misspecified. The results of the GEE method in PLINK also show an increase in the Type-I error, but they are lower due to the robustness against misspecification.

Our findings show that it is obvious that MERLIN and ProbABEL are not able to deal with binary traits in the current design. The poor performance of ProbABEL in the binary trait case is due to the fact that it does not model any of the covariance structure among individual pairs, resulting in no control at all on the family structure and hence too liberal p-values. On the performance of MERLIN it is not really clear whether the software is able to deal with genotype imputed family data for the binary trait. Most likely is that in the binary trait case MERLIN behaves as if it was dealing with a quantitative trait. Without any adjustments, as with EMMAX, this leads to a poor performance on statistical accuracy. It would, however, be possible to be more certain on the analyses, by carefully studying the package's source code. However, this task is beyond the scope of the current project.

A point of criticism that could be made on the current research strategy is the

choice of the particular toolsets, the simulation design, and the presentation of the results. There are more freely available packages that can deal with soft-called family data for binary and quantitative traits, such as FaST [27], GCTA [21], OpenMX [7], QTLrel [11], or TASSEL[33]. The reason why we did not choose any of these packages was either because the package was too new (QTLrel and FaST) to be able to include it in the paper, or that the option how to read in the specific data (TASSEL) was not found in time, or we could not get the software running without errors (GCTA). We could not find any other toolsets able to deal with soft-called family data, than the ones we studied and the ones mentioned above.

Although our simulation studies provide a good overview on the performance of the toolset, because we use real genotypic data, it is true that we only have 1557 samples. Nowadays, sample sizes in GWA studies are usually larger to have more statistical power to detect quantitative trait loci. In the current design we analyze few and large SNP-effects over different heritability levels such that the sample size is not a problem. Furthermore, due to the use of a low prevalence and only three effect-SNPs from which we could only simulate 116 cases for the binary trait, the simulation is realistic. Moreover, sample sizes of approximately 1500 could coincide with large costs of genotyping of rare variants and phenotyping of traits. Hence, they will still be analyzed, but mostly with the purpose to upload its results to consortia for meta-analysis.

With a larger sample size it would also have been possible to see whether one's specific choice for a toolset should depend on the sample size. It is important to consider whether the consequences for the statistical accuracy (and feasibility) of the misspecification of the GEE model in PLINK versus the MLM in the other toolsets will become more apparent when the sample size is larger. Also, the "learning" of each statistical procedure of a toolset on the increase in a sample size could be different, leading to differences in the relation between power and sample size for a toolset.

One could remark that the small number of 100 simulated phenotypic traits render unstable estimates of statistical power and Type-I error against low cut-off values of α. That is why we presented only the number of false and true discoveries (instead of Type-I error and power estimates). In addition, the influence of this low number of simulations on the relative performance is less pronounced, since the toolsets underwent exactly the same conditions. We do see, however, that with more simulations, a larger sample size, and more non-effect unlinked SNPs, it would have created the possibility to plot results of statistical power and Type-I errors against an α cut-off value for each package. However, such measures coincide with more costly computing time, especially since stable estimates on power and Type-I error require cross-validation or more simulations.

Our findings in the current design are generalizable to simple traits. An important consideration in generalizing our simulation results to GWA studies is the inclusion of more smaller effect-SNPs and the incorporation of covariates as compared to the current design. Nowadays, practical application of the toolsets on genotype imputed family data will undoubtedly more present on complex phenotypic traits of which it is hypothesized they are explained by many SNPs. Furthermore,

basic covariates such as age and sex are routinely included in the analysis as well. It is for these specific settings that most toolsets have been tailored.

A last note for future studies is on the careful selection of toolsets with a focus on the implementation of advanced statistical techniques in human genetics. Analyses in GWA mainly consider additive SNP-effects, assume random mating, estimate the effects of each SNP separately, use the prospective likelihood (predict probability of the phenotype given the genotype), and most times this is done without the possibility to control for ascertainment. However, it goes without saying that the statistical methodology for GWA is under rapid development, explaining the ongoing growth of the number of packages that deal with genotype imputed family data. Current advanced statistical methods being applied in (human) genetics are:

i optimal scaling to deal with dominance, epistasis and gene-x-environment interaction effects [31, 52],

ii analyses allowing for assortative mating (with the use of an extra covariance matrix) [45],

iii analysis of all SNPs simultaneously in genotype imputed family data using a penalized likelihood [16, 18],

iv the use of a more powerful method such as the joint likelihood that controls for ascertainment automatically [6, 25],

v and by compressing the data to increase computational efficiency [51].

Although its a speculation for future research, the combination of these advanced statistical techniques will become of great importance in the quest of capturing the phenotypic variance due to genetics as obtained from heritability studies.

References

[1] Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin – rapid analysis of dense genetic maps using spare gene flow trees. Nat Genet 30: 97–101.

[2] Agresti A (2002) Categorical Data Analysis. Wiley, New York.

[3] Armitage P (1955) Tests for Linear Trends in Proportions and Frequencies. Bioinformatics 11: 375–386.

[4] Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) Genabel: an r library for genome-wide association analysis. Bioinformatics 23: 1294–1296.

[5] Aulchenko YS, Struchalin MV, Duin CMV (2011) ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics 11: 134–143.

[6] Balliu B, Tsonaka R, van der Woude D, Boehringer S, Houwing-Duistermaat JJ (2011) Modeling The Non-Inherited Maternal Antigens Effect In Multi-Case Families. In: Abstracts of the 20^{th} Annual Meeting of the International Genetic Epidemiology Society Conference, Heidelbeg, Germany.

[7] Boker S, Neale M, Maes H, Wilde M, Spiegel M (2011) OpenMx: An open Source extended structural equation modeling framework. Psychometrika 76: 306–317.

[8] Boomsma DI, de Geus EJC, Vink JM, Stubbe JH, Distel MA, et al. (2006) Netherlands Twin Register: From Twins to Twin Families. Twin Res Hum Genet 9: 849–857.

[9] Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85: 847–861.

[10] Chen WM, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81: 913–926.

[11] Cheng R, Abney M, Palmer AA, Skol AD (2011) QTLRel: an R Package for Genome-wide Association Studies in which Relatedness is a Concern. BMC Genet 12: 66.

[12] Cox DR, Hinkley DV (1974) Theoretical Statistics. Chapman & Hall.

[13] Fitzmaurice G, Laird N, Ware J (2004) Applied Longitudinal Analysis. John Wiley & Sons.

[14] Gelderman H (1975) Investigations on Inheritance of Quantitative Characters in Animals by Gene Markers. Theor Appl Genet 46: 319–330.

[15] Henderson CR (1976) Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics 32: 69–83.

[16] Hoggart CJ, Whittaker JC, De Ioro M, Balding DJ (2008) Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. PLoS Genet 4: e1000130.

[17] Houwing-Duistermaat JJ, Bijkerk C, Hsu L, Stijnen T, Slagboom EP, et al. (2003) A Unified Approach to Modelling Linkage to Quantitative and Qualitative Traits. Ann Hum Genet 67: 457–463.

[18] Houwing-Duistermaat JJ, Uh HW, Tsonaka R (2011) Pathway Analysis for Family Data using Nested Random-Effects Models. BMC proc p. to appear.

[19] Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.

[20] Huber PJ (1967) The behavior of maximum likelihood estimates under non-standard conditions. In: Proceedings of the Fifth Berkeley symposium on mathematical statistics and probability, pp. 221–233, University of California Press, Berkeley.

[21] J Y, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet 88: 76–82.

[22] Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42: 348–356.

[23] Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.

[24] Kenward MG, Roger JH (1997) Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 54: 983–997.

[25] Kraft P, Thomas DC (2000) Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 66: 1119–1131.

[26] Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34: 816–834.

[27] Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, et al. (2011) FaST linear mixed models for genome-wide association studies. NatMethods 8: 833–835.

[28] Manichaikul A, Chen WM, Williams K, Wong Q, Sale MM, et al. (2011) Analysis of family- and population-based samples in cohort genome-wide association studies. Hum Genet pp. 1–13, URL `http://dx.doi.org/10.1007/s00439-011-1071-0`.

[29] Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Genet 11: 499–511.

[30] McCulloch CE, Searle SR, Neuhaus JM (2008) Generalized, Linear, and Mixed Models. Wiley.

[31] Meulman JJ (2003) Prediction and Classification in Non-Linerar Data Analysis: Something Old, Something New, Something Borrowed, Something Blue. Psychometrika 68: 493–517.

[32] Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. Nat Rev Genet 12: 465–474.

[33] PJ B, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, et al. (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

[34] Price A, Zaitlen N, Reich D, Patterson N (2011) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459–463.

[35] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81: 559–575.

[36] R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org`, ISBN 3-900051-07-0.

[37] Spencer CCA, Su Z, Donnely P, Marchini J (2009) Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping chip. PLoS Genet 5: e1000477.

[38] Spielman RS, McGinnis RE, Ewens JW (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506–516.

[39] Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681–690.

[40] Thallman RM, Hanford KJ, Kachman SD, van Vleck LD (2004) Sparse Inverse of Covariance Matrix of QTL EFfects with Incomplete Marker Data. Stat Appl Genet Mol 3: Art. 30.

[41] The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Ann Hum Genet 467: 457–463.

[42] The International Hapmap Consortium (2003) The International HapMap Project. Nature 426: 1061–1073.

[43] Uh H (2011) Testing for genetic association using related subjects dealing with imputed genotypes. submitted for publication .

[44] Uh HW, Houwing-Duistermaat JJ, Putter H, van Houwelingen HC (2009) Assessment of global phase uncertainty in case-control studies. BMC Genet 10: 54.

[45] Uh HW, van der Wijk HJ, Houwing-Duistermaat JJ (2009) Testing for genetic association taking into account phenotypic information of relatives. BMC proc 3: S123.

[46] Wang K (2002) Efficient Score Statistics for Mapping Quantitative Trait Loci with Extended Pedigrees. Hum hered 54: 57–68.

[47] White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50: 1–25.

[48] Willemsen G, de Geus EJC, Bartels M, van Beijsterveldt CEMT, Brooks AI, et al. (2010) The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies. Twin Res Hum Genet 13: 231–245.

[49] Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, et al. (2006) A unified mixed-model methods for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.

[50] Zhang Z, Buckler ES, Casstevens TM, Bradbury PJ (2009) Software engineering the mixed model for genome-wide association studies on large samples. Brief Bioinform 10: 664–675.

[51] Zhang Z, Ersoz E, Lai CQ, Todhunter R, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355–362.

[52] Zheng G, Freidlin B, Li Z, Gastwirth JL (2003) Choice of Scores in Trend Tests for Case-Control Studies of Candidate-Gene Associations. Biometrical Journal 45: 335–348.

[53] Zheng J, Li Y, Abecasis GR, Scheet P (2011) A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes. Genet Epidemiol 35: 102–110.

Appendix

Genome-wide Assocation (GWA) studies which implement Mixed Linear Models (MLM) or General Estimating Equations (GEE) as analysis strategy, usually interpret the polygenic model of the quantitative phenotype as:

$$\text{MLM} \quad : \quad \mathbf{y} = W\boldsymbol{\nu} + X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{3}$$

$$\text{GEE} \quad : \quad \mathbf{y} = W\boldsymbol{\nu} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \tag{4}$$

where $\boldsymbol{y}$ is a vector of the phenotype of size n (number of individuals); $\boldsymbol{\nu}$ is the vector representing non-marker effects (e.g. sex, age); $\boldsymbol{\beta}$ are the marker effects; and $\boldsymbol{\gamma}$ is a vector of size n for unknown random polygenic effects, having a normal distribution with mean of zero and covariance matrix $\mathrm{G} = 2\Phi\sigma_a^2$, where σ_a^2 is an unknown genetic variance, and Φ is the kinship (co-ancestry) matrix with element $\phi_{ij}\,(i, j = 1, 2, ..., n)$ calculated from either a set of genetic markers or pedigrees. W, X and Z are the incidence matrices that include the covariates and SNPs for $\boldsymbol{\nu}$,$\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively, and ϵ is a vector of random residual effects that are normally distributed with zero mean and covariance $\Sigma_{n\times n} = I_{n\times n}\sigma_\epsilon^2$, where $I_{n\times n}$ is the identity matrix and the scalar σ_ϵ^2 is the unknown residual variance. The residual variance in GEE, however, is not completely unknown. In GEE the residuals $\boldsymbol{\epsilon^*}$ are assigned a working covariance matrix $\Sigma^*_{n\times n}$, as we show below. Note that if we define $\boldsymbol{\epsilon^*}$ as the sum of both the unknown random polygenetic effects and the random residual effect from the MLM model (3), the two models appear to be the same. However, there is a difference when modeling the means $\mathrm{E}[\boldsymbol{y}]$ for the individuals when the phenotype trait is not quantitative, as will follow in the sections below.

Parameter estimates in MLM and GEE

The estimation of the parameters in models (3) and (4) is different. Inference in MLM is based on (maximization of) the likelihood:

$$L(\boldsymbol{y}|\boldsymbol{\nu}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_a^2, \sigma_\epsilon^2) \tag{5}$$

which involves distribution assumptions to integrate out the random effects by use of numerically complex algorithms (e.g. adaptive gauss quadrature, nested integrated Laplace approximations) in combination with iterative algorithms such as Expectation Maximization, Fisher scoring, Average information, see McCulloch, Searle and Neuhaus (2008). In the GEE setting, however, distribution assumptions are dropped such that the results depend only on the mean and a pre-specified variance structure in Σ^*. The mean $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{y}]$ is equal to the linear predictor $\boldsymbol{\eta}$ in the quantitative trait model, as is shown below:

$$\boldsymbol{\mu} \;=\; \boldsymbol{\eta} \;=\; W\boldsymbol{\nu} + X\boldsymbol{\beta}. \tag{6}$$

The i, j^{th} elements of s^*_{ij} of $\Sigma^*_{n\times n}$ can be "decomposed" as

$$s^*_{i,j} \;=\; \mathrm{R}_{ij} \times \phi \times \mathrm{var}(\boldsymbol{\mu}), \tag{7}$$

where ϕ is a to be estimated scale parameter such that $\text{var}(y_{ij}) = \phi \times \text{var}(\eta_{ij})$. In the matrix R the variance structure is defined. For example, toolsets using GEE in genetics mostly specify a compound symmetry variance structure within each nuclear family. Suppose the first four rows (and columns) of the R-matrix consist of members from a nuclear family and the fifth does not, R would then look like:

$$\mathrm{R}_{n\times n} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0 & \dots \\ 0.5 & 1 & 0.5 & 0.5 & 0 & \dots \\ 0.5 & 0.5 & 1 & 0.5 & 0 & \dots \\ 0.5 & 0.5 & 0.5 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{8}$$

Another option would be to replace the elements within R with kinship coefficients inferred from the pedigree data or calculated from markers. Finally, the parameters $\boldsymbol{\nu}, \boldsymbol{\beta}, \phi$ are obtained solving the equation

$$\left(\frac{\delta \boldsymbol{\mu}}{\delta \boldsymbol{\theta}}\right)^T \Sigma^{*-1} (\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0}, \tag{9}$$

where $\boldsymbol{\theta}^T$ is the vector $\left[\boldsymbol{\nu}^T | \, \boldsymbol{\beta}^T\right]$, of size $p+1$. The generalized estimating equations (9) does not necessarily has a closed form solution, using iterative algorithms (e.g. quasi scoring procedure).

Hypothesis testing in GWA studies

The goal of GWA studies is the estimation of the fixed effects in $\boldsymbol{\beta}$ for each locus (k) separately. In MLM and GEE, the effect of the genotype at each k^{th} locus can be modeled as a main effect, whereas the relationships among all individuals are taken into account by means of random polygenic effects or a pre-specified variance structure.The null hypothesis for the association test is that $\beta_k = 0$ and the alternative hypothesis is that $\beta_k \neq 0$. Note that for each analysis the equations (3) and (4) change in a sense that the vector $\boldsymbol{\beta}$ containing the marker effects will become β_k with its remaining $\boldsymbol{\beta}_{-k}$ becoming part of the random effects $\boldsymbol{\gamma}$ or residuals $\boldsymbol{\epsilon}^*$, respectively. The estimation of the parameters, however, remains the same.

In MLM the test of the null hypothesis can be performed by either a Wald based approximate F-test or the standard Wald test after the maximization of the likelihood in equation (5). To control the Type-I error, approximate F tests are preferred [24]. The Wald based F-test is based on a distribution with degrees of freedom (df_1) for estimating the marker-effect and degrees of freedom df_2 for estimating the component(s) of random effects σ^2_a. Furthermore, it is based on Restricted Maximum Likelihood (REML) instead of Maximum Likelihood (ML) for the estimation of the variance component. In REML a known error contrast matrix K is implemented in the likelihood such that $\mathrm{E}[\mathrm{K}'\boldsymbol{y}] = 0$, avoiding estimation of the fixed effects. Then, the restricted log likelihood only depends on the unknown components of variance, yielding better estimates of these variance components as compared to the non-restricted likelihood. The standard Wald test, however, is

χ^2-distributed and based on the specification of the degrees of freedom, it does not make such a distinction (comparable to a $df_2 = \infty$, assuming a known σ_a^2), leading to p-values tending to be too liberal. Last, it is advised to use Maximum Likelihood (ML) instead of Restricted Maximum Likelihood (REML) if comparing two nested models with different sets of fixed effects with a likelihood ratio test (LRT). Because REML works with contrasts in such a way to facilitate more accurate estimates of the random effects components, each Log-likelihood for a different set of fixed effects is unique up to a constant. The LRT, then, will not follow a chi-square distribution in its asymptote.

In GEE the calculation of the test statistic relies on quasi-likelihood,of which equation (9) is the derivative. In this perspective the test of the null hypothesis is based on the assumption that the obtained fixed parameter estimate $\hat{\beta}_k$ is approximately normal distributed with mean β_k and the variance of $\hat{\beta}_k$, known as the sandwich estimator, or Huber-White robust variance estimator [20, 47]. A variance term in which the observed covariance matrix (covar($\hat{\boldsymbol{y}}$)) is pre and post-multiplied with the estimated covariance matrix ($\hat{\boldsymbol{\Sigma}}^*$). If we denote $\hat{\boldsymbol{\beta}}$ as the estimated SNP-effects with

$$\text{covar}\left(\hat{\boldsymbol{\beta}}\right) = V_0^{-1} V_1 V_0^{-1} \tag{10}$$

where

$$V_0^{-1} = \left(\frac{\delta \boldsymbol{\eta}}{\delta \boldsymbol{\theta}}\right)^{-1} \Sigma^{*\frac{1}{2}} \Sigma^{*\frac{1}{2}} \left(\frac{\delta \boldsymbol{\eta}}{\delta \boldsymbol{\theta}}\right)^{-T} \tag{11}$$

$$V_1 = \left(\frac{\delta \boldsymbol{\eta}}{\delta \boldsymbol{\theta}}\right)^{T} \Sigma^{*-1} \text{covar}(\boldsymbol{y}) \Sigma^{*-1} \left(\frac{\delta \boldsymbol{\eta}}{\delta \boldsymbol{\theta}}\right) \tag{12}$$

such that

$$\text{covar}\left(\hat{\boldsymbol{\beta}}\right) = \left(\frac{\delta \boldsymbol{\eta}}{\delta \boldsymbol{\theta}}\right)^{-1} \Sigma^{*-\frac{1}{2}} \text{covar}(\boldsymbol{y}) \Sigma^{*-\frac{1}{2}} \left(\frac{\delta \boldsymbol{\eta}}{\delta \boldsymbol{\theta}}\right)^{-T}, \tag{13}$$

than, the variance of an estimated SNP-effect β_k at locus k for the sequential GWA hypothesis testing boils down to

$$\text{var}\hat{\beta}_k = (\boldsymbol{x}_k)^{-1} \Sigma^{*-\frac{1}{2}} (\boldsymbol{y} - \boldsymbol{\eta})(\boldsymbol{y} - \boldsymbol{\eta})^T \Sigma^{*-\frac{1}{2}} (\boldsymbol{x}_k)^{-T}, \tag{14}$$

where $\boldsymbol{x}_k$ is the genotype indicator vector of size n. We obtain the estimate of covar$\hat{\beta}_k$ by plugging in the optimal parameter estimates $\hat{\beta}_k, \boldsymbol{\nu}, \theta, \alpha$ from (9). The null hypothesis is tested with the Wald t-statistic ($df = 1$) by dividing the $\hat{\beta}_k$ by the squared root of the sandwich estimator, equation (14).

Binary trait as phenotype

When dealing with a binary trait as phenotype, the principle of parameter estimation and testing remains the same for both GEE and MLM (although more complex algorithms are needed). A link function $g()$ is required to make sure that the expectation of the phenotypes can be estimated through the linear predictor $\boldsymbol{\eta}$:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta}. \tag{15}$$

For a binary phenotype in GWA studies the link is used the most is the logit link,

$$g(\boldsymbol{\mu}) \;=\; \log\left(\frac{\boldsymbol{\mu}}{1-\boldsymbol{\mu}}\right), \tag{16}$$

such that GEE and MLM become an extension of logistic regression for case-control data, modeling the log(odds). For GEE, the generalized equation (9) and the sandwich estimator (13), still apply. In the MLM setting, however, we have a change in distribution assumptions. Instead of $\boldsymbol{y}$ being multivariate normal distributed, each $y_i|\gamma_i \sim Binomial(1, \mu_i)$. Note that $E[\boldsymbol{y}] \neq E[\boldsymbol{y} \,|\, \boldsymbol{\gamma}] = \boldsymbol{\mu}$. As a consequence, in GWA studies, the $\hat{\beta}_k$ (and $\hat{\boldsymbol{\nu}}$) are interpreted given the subject specific values of the random effects, in contrast to GEE which yields population average parameters. An estimate of the MLM population average (PA) effect of SNP at locus k is obtained as

$$\hat{\beta}_{PA,k} = \frac{\hat{\beta}_k}{\sqrt{0.35\hat{\sigma}_a^2 + 1}}, \tag{17}$$

under the normal distribution assumption of a zero mean and σ_a^2, the (co)variance of the random effects [30].

MLM vs. GEE

In conclusion, MLM and GEE could test a similar null hypothesis in GWA studies, even though they differ on the estimation procedure, testing and interpretation of the marker-effect. The choice for the one or the other is not that obvious. A more simple interpretation of the results is possible with GEE, but it comes at a price of the use of an inefficient sandwich estimator as compared to a parametric estimate in MLM [30]. This is especially true when the number of individuals is small, and the test is performed on a SNP with a rare allele. The rule of thumb for the GEE is: if the model for the mean is specified correctly, and the residual standard deviations are homogeneous, then the results are approximately correct [13]. Even though MLM, given the model is specified correctly, has better performance due to its higher statistical power and efficiency, it needs more computing power and a "messy" transformation of its subject specific parameters to population average parameters when using a binary phenotype [30]. Hence, both statistical techniques have their pro's and con's