Efficient Control of Population Structure in Model Organism Association Mapping

Hyun Min Kang 1,* Noah A. Zaitlen 2, Claire M. Wade 3,4, Andrew Kirby 3,4, David Heckerman 5, Mark J. Daly 3,4, Eleazar Eskin 1,*

Abstract

Genome-wide association mapping in model organisms such as inbred mouse strains is a promising approach for the identification of risk factors related to human diseases. However, genetic association studies in inbred model organisms are confronted by the problem of complex population structure among strains. This induces inflated false positive rates, which can not be corrected using standard approaches applied in human association studies such as Genomic Control or Structured Association. Recent studies demonstrated that linear mixed models successfully correct for the genetic relatedness in association mapping in a maize panel dataset. However, the currently available mixed model methods suffer from computational inefficiency and unknown convergence properties. In this paper, we propose a new method, Efficient Mixed Model Association (EMMA), which corrects for population structure and genetic relatedness in model organism association mapping. Our method is orders of magnitude more efficient than previous mixed model methods while improving the reliability of results with guaranteed convergence properties and global optimization. We applied our EMMA method to in silico whole genome association mapping of inbred mouse strains and Arabidopsis involving hundreds of thousands of SNPs. In spite of limited power of model organism association mapping due to the limited number of inbred strains, we are able to identify significantly associated SNPs, which fall into known QTLs or genes identified through previous studies without an inflation of false positives. The implementation of our EMMA method is publicly available via R package and webserver.

1Computer Science and Engineering, University of California San Diego, La Jolla, California, 92093 USA
2Bioinformatics Program, University of California San Diego, La Jolla, California, 92093 USA
3Broad Institute of Harvard and MIT, Cambridge, Massachusetts, 02141 USA
4Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, 02114 USA
5Microsoft Research, Redmond, Washington, 98052 USA
*Corresponding authors. Emails: h3kang@cs.ucsd.edu, eeskin@cs.ucsd.edu
1 Introduction

With the recent development of high-throughput genotyping technologies, genetic variation in many model organisms such as mice, Arabidopsis, and maize are being discovered on a genome-wide scale[26, 17, 12]. Genome-wide association mapping in model organisms has great potential to identify risk factors for complex traits related to human diseases. Although straight inference from model organisms to human traits may not always apply, model organism association mapping is potentially more powerful than human association mapping because it is possible to reduce the effect of environmental factors by replicating phenotype measurements in genetically identical organisms. In addition, it is often easier and more cost effective to verify associated signals via \textit{in vivo} and/or \textit{in vitro} experiments in model organisms than in human subjects. Moreover, many ongoing genotyping and phenotyping projects in model organisms such as the Mouse Phenome Database (MPD) [31] and Mouse HapMap projects provide publicly available resources to perform \textit{in-silico} mapping of complex traits in model organisms.

However, genetic association studies in inbred model organisms are confronted by the problem of inflated false positive rates due to population structure and genetic relatedness among inbred strains caused by the often complex genealogical history of most model organisms strains. Applying conventional statistical tests of independence between a genetic marker and a phenotype is prone to spurious associations because the marker and the phenotype are likely to be correlated via population structure which violates the independence assumption under the null hypothesis. Recent association or linkage mapping studies in model organisms attempt to avoid inflated false positive rates by designing the studies using recombinant inbred lines[7] generated from a handful of parental strains. However, these studies are limited by the variation present in the parental strains and have long regions between recombinations due to relatively few generations between the recombinant inbred strains and the parental strains.

An alternative approach to reduce the number of false positives is to apply a statistical test that corrects for the bias due to population structure or genetic relatedness. The most widely used methods to reduce such bias in human association mapping are Genomic Control[6] and Structured Association[27]. These are inadequate in the case of model organism association mapping. Genomic Control suffers from weak power when the effect of population structure is large as in model organisms. Structured Association, which assumes a handful number of ancestral populations and admixture, only partially captures the multiple levels of population structure and genetic relatedness in model organisms[3, 35, 34].

Mixed models are another way to correct for complex genetic relatedness in the statistical analysis of quantitative traits under population structure[18]. In addition to uncorrelated random errors of the linear model, the mixed model introduces another non-diagonal variance component that accounts for genetic correlation between the response variables. Although the mixed model has
been extensively studied in the context of longitudinal data analysis[32] and animal inbreeding[16], few studies have applied mixed models to association or linkage mapping in order to correct for population structure. Variance component models, which are similar to mixed models, have been widely applied to the multipoint quantitative-trait linkage analysis with pedigrees[1, 5], but not to the association mapping in inbred model organisms. Recently, Yu et al.[34] and Zhao et. al[35] demonstrated that, in the analysis of maize and Arabidopsis panel samples with multiple levels of genetic relatedness, the association mapping based on a mixed model obtains fewer false positives and higher power than Structured Association and Genomic Control. If combined with Structured Association, the unified mixed model reduces the inflation more robustly. However, the mixed model methods proposed in [34] and [4] have several limitations. First, the kinship matrix inferred from the SPAGeDi software[15] package is not guaranteed to be positive semidefinite, and may therefore violate the non-negative variance component assumption of the mixed model framework. Second, the convergence properties of Henderson’s iterative procedure[16] for estimating best linear unbiased estimator (BLUE) and best linear unbiased prediction (BLUP) are unknown. Even if the method converges, there is no guarantee that the resulting parameters globally maximize the likelihood of observed phenotypes. Third, the mixed model implementation in [34] consumes a substantial amount of computational time, requiring several CPU hours for a single run of only 553 SNPs over 277 strains. This computational cost prohibits performing fine-scaled genome wide association mapping and permutation tests for more accurate multiple hypothesis testing.

In this paper, we propose a new method, Efficient Mixed Model Association (EMMA), which corrects for population structure and genetic relatedness in model organism association mapping. Our method is orders of magnitude more efficient than previous mixed model methods while improving the reliability of results with guaranteed convergence properties and global optimization. Our method takes advantage of the specific nature of the optimization problem in applying mixed models for association mapping, which allows us to substantially increase computational speed and reliability. In model organism association mapping where the kinship matrix or genetic covariance matrix is inferred from genotypes, only the coefficients of the variance components must be estimated. In such cases, the standard approach using Henderson’s iterative procedure[16, 4] imputes unnecessary parameters such as the individual random effects in BLUP. Our method greatly improves the efficiency of the mixed model method by enabling us to perform statistical tests without BLUP estimation, reducing the large number of dimensions that need to be numerically optimized to one. Our method’s efficiency is further increased by avoiding redundant matrix inverses and multiplications in the computation of restricted likelihood by using the properties of singular value decomposition. Due to the simplicity and mathematical tractability of the optimization procedure, its convergence properties are always guaranteed, improving the reliability of our method. Moreover, it is possible to guarantee with very high confidence that our restricted maximum likelihood (REML) estimates, which implicitly takes into account the loss in degrees of freedom associated with fixed effects[30],
are globally optimal even though the likelihood function may not be convex. Our analysis on the maize panel used in [34] shows that our method is orders of magnitude faster than the previous methods while improving the reliability of the results.

We propose three simple methods for the inference of the kinship matrix that guarantee positive semidefiniteness. First, we use the genotype similarity matrix, which is guaranteed to be positive semidefinite, based on the assumption that each nucleotide change has the same level of random effect on the phenotype. Second, we define a kinship matrix by a haplotype similarity matrix based on the assumption that each haplotype has an independent random effect. We propose a third method called phylogenetic control based on the assumption that a phylogenetic tree is a good approximation of the genealogical history of an inbred model organisms. In such cases, the phylogenetic tree may be used as a confounding factor, correcting for complex genetic relatedness between strains. We show that phylogenetic control can be formulated into a linear mixed model, and present an algorithm for inferring the phylogenetic kinship matrix. The matrix is always positive semidefinite and its REML parameters are unique regardless of the choice of root. Our results are consistent with Zhao et. al[35], suggesting that these simpler kinship matrices reduce the false positive rate as effectively as kinship matrices geneated by SPAGeDi.

One of the important questions in the design of model organism association mapping studies is estimating the study power with a limited number of inbred strains. We performed a simulation study of the power of our EMMA method to identify causal SNPs both on a genome-wide scale and within a QTL. Our results show that with a limited number of genetically diverse strains, such as the currently available panel of inbred mice, it is possible to identify causal loci with a genome wide significance only if the locus explains a large portion of phenotypic variance. However, as the number of individual strains grow, the power of these association studies increases dramatically. Our analysis of statistical power in model organism association mapping demonstrates the dramatic effect of using multiple phenotype measurements from each strain. Study designs that do not replicate phenotype measurements and analysis methods that do not take multiple measurements into account suffer a significant decrease in statistical power.

We used our EMMA method to perform a whole genome association mapping study of inbred mouse strains. We analyzed nearly 140,000 mouse HapMap SNPs over 48 strains and three quantitative phenotypes, with QTLs identified by previous studies. We identified significant associations for the three phenotypes, liver weight, body weight, and saccharin preference. Our results also show a significant reduction in the inflation of false positives. Interestingly, many of the significantly associated SNPs fall into the known QTLs, suggesting the results are likely true associations. We also performed genome wide association mapping in 95 Arabidopsis strains for a flowering time phenotype. We were able to identify the previously known FRI gene without observing inflation of p-values. The implementation of our method via R package, as well as the mouse and Arabidopsis association results are are publicly available online at http://snp.ucsd.edu/mouse and
2 Results

2.1 Comparison with Previous Methods over Maize Strains

Figure 1 Comparisons of p-values between the implementation of unified mixed model and our EMMA method. We use the same kinship matrix inferred from SPAGeDi software and population structure coefficients inferred from STRUCTURE from a previous study. (a) Direct comparison of p-values between the SAS implementation of unified mixed model and our method. All p-values are highly correlated, implying that two methods are almost identical in terms of accuracy. (b) Cumulative distribution of observed p-values, which are computed from 553 random SNPs of maize panel data and the flowering time phenotype. Refer to [34] for comparisons. Under the assumption that the SNPs are unlinked and there are few SNP association, the observed p-values are expected to be close to the cumulative p-values.

We applied our EMMA method to the same maize panel data with 553 SNPs and three phenotypes across 277 diverse inbred lines analyzed by the unified mixed model. The kinship matrix is inferred by SPAGeDi software, setting all negative coefficients to zero. The population structure coefficients are estimated from STRUCTURE using 89 microsatellite loci for three subpopulations.

Figure 1(a) shows the comparison of the p-values obtained from the previous unified mixed model (Structured Association (SA) + Mixed Model (MM) method) with those from EMMA for flowering time phenotypes. They are almost identical, differing only due to the differences in the numerical optimization procedure for the estimation of variance components. The cumulative distribution
Figure 2 Cumulative distributions of observed p-values for (a) Mixed Model (MM) only model (Simple regression without kinship matrix) and (b) Structured Association (SA) and Mixed Model (MM) (Structured Association without kinship matrix) for various kinship matrices in the maize panel analysis for flowering time phenotypes. SPAGeDi denotes the kinship matrix inferred from SPAGeDi software with the negative values set to zero as suggested in [34]. Genotype similarity denotes the simple genotype similarity matrix based on the genotype distance. Phylogenetic control denotes the kinship matrix computed from inferred phylogenetic tree.

Of p-values across different models for genetic relatedness, shown in Figure 1(b), are also almost identical to the published results.

While both the SAS and TASSEL (version 1.9.5)[34] implementations of unified mixed model take nearly 2 hours for a single run over these datasets with Intel 2.8GHz Dual Core CPU, the execution time of our mixed model implementation is nearly 50 times faster, taking only three minutes. The results of our method are more reliable because we find the global REML estimate with guaranteed convergence properties. Previous implementations iteratively search for local optima with unknown convergence properties. Possibly due to the instability of the convergence properties, the TASSEL implementation could not compute p-values for several loci in the maize panel analysis.

Since the kinship matrix inferred from SPAGeDi software is not positive semidefinite, we explore other ways to estimate the variance components due to genetic background. We use a genotype similarity matrix and a phylogenetic control matrix which guarantee positive semidefiniteness. Haplotype similarity matrices are not applicable to this dataset due to sparse genotype density.

We compared the goodness-of-fit of these kinship matrices in addition to the SPAGeDi-based kinship matrix over three maize phenotypes using Bayesian Information Criterion (BIC), which provides a measure of how well each model fits the data. Adjusting for the sample size and the number of free parameters, table 1 shows that the goodness-of-fits of the three kinship matrices based on restricted maximum likelihood (REML) estimates are comparable, while all of them were significantly better than not using a mixed model.
Table 1. Goodness of fit of different models and kinship matrices in explaining phenotypic variation of maize quantitative traits

<table>
<thead>
<tr>
<th>Method</th>
<th>Kinship Matrix</th>
<th>Flowering Time -2*(REML)</th>
<th>Flowering Time BIC</th>
<th>Ear Height -2*(REML)</th>
<th>Ear Height BIC</th>
<th>Ear Diameter -2*(REML)</th>
<th>Ear Diameter BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>N/A</td>
<td>1627.5</td>
<td>1638.6</td>
<td>2288.2</td>
<td>2299.3</td>
<td>1278.0</td>
<td>1288.9</td>
</tr>
<tr>
<td>MM</td>
<td>SPAGeDi</td>
<td>1517.4</td>
<td>1534.1</td>
<td>2228.2</td>
<td>2244.8</td>
<td>1247.8</td>
<td>1264.1</td>
</tr>
<tr>
<td>MM</td>
<td>Genotype Similarity</td>
<td>1516.8</td>
<td>1533.4</td>
<td>2230.0</td>
<td>2246.7</td>
<td>1256.7</td>
<td>1273.0</td>
</tr>
<tr>
<td>MM</td>
<td>Phylogenetic Control</td>
<td>1512.4</td>
<td>1529.1</td>
<td>2215.4</td>
<td>2232.1</td>
<td>1240.2</td>
<td>1256.5</td>
</tr>
<tr>
<td>SA</td>
<td>N/A</td>
<td>1511.0</td>
<td>1533.2</td>
<td>2225.9</td>
<td>2248.2</td>
<td>1263.4</td>
<td>1285.2</td>
</tr>
<tr>
<td>SA+MM</td>
<td>SPAGeDi</td>
<td>1475.6</td>
<td>1503.4</td>
<td>2192.8</td>
<td>2220.6</td>
<td>1235.1</td>
<td>1262.3</td>
</tr>
<tr>
<td>SA+MM</td>
<td>Genotype Similarity</td>
<td>1477.9</td>
<td>1505.7</td>
<td>2196.3</td>
<td>2224.1</td>
<td>1244.7</td>
<td>1272.0</td>
</tr>
<tr>
<td>SA+MM</td>
<td>Phylogenetic Control</td>
<td>1469.1</td>
<td>1496.9</td>
<td>2182.3</td>
<td>2210.1</td>
<td>1226.8</td>
<td>1253.1</td>
</tr>
</tbody>
</table>

Note. — Comparison of Restricted Maximum Likelihood (REML) and Bayesian Information Criterion (BIC) of each model with different kinship matrices for maize quantitative traits. The model with the smaller BIC is preferred. 'Simple' denotes simple linear regression without adjustment for population effect. SA is the model using the output from STRUCTURE as covariates. MM is mixed model with different kinship matrices. The descriptions of kinship matrices are the same as in Figure 2.
The cumulative p-value distribution seen in Figure 2 show that the simple genotype similarity matrix corrects for genetic relatedness slightly better than the other two kinship matrices. There is a better reduction of false positive rates, especially within the region of small p-values. Since the simpler kinship matrices show comparable or better goodness-of-fit and false positive reduction results while guaranteeing positive semidefiniteness, we apply only these simple kinship matrices in the following sections.

2.2 Power of Inbred Association Mapping

In order to evaluate the statistical power of inbred model organism association mapping, we performed a simulation study with several model organism datasets using our EMMA method. We simulated an additive effect of causal SNP over the existing phenotypes for mouse, Arabidopsis, and maize strains. Three mouse phenotypes body weight, liver weight, and saccharin preference, described in the Materials and Methods section in detail, had 38, 34, and 24 available strains with multiple measurements respectively. Arabidopsis and maize had 95 and 277 genotyped strains available for flowering time phenotypes, respectively. As the number of currently available inbred mouse strains is limited, the power analysis of Arabidopsis and maize will shed light on how many more genetically diverse strains are needed in order to attain genome wide power in inbred mouse association mapping.

We compared the power of EMMA method using three inbred mouse phenotypes by selecting causal SNPs randomly across the genome and adding a SNP effect proportional to the standard deviations of the phenotypes as shown in in Figure 3. The results show that our method has the power to detect a small quantitative effect (∼5% of the phenotypic variation) when only testing a single marker. When considering the markers in a QTL and correcting for multiple testing, our method can detect a moderate effect (>10% of the phenotypic variation). However, genome wide power is achievable only for very strong effects (>30% of the phenotypic variation), and only in the case of the body weight phenotype where a larger number of strains are available. The power was much lower for liver weight and saccharin preference phenotypes due to the smaller number of strains phenotyped. Our simulation results demonstrate that it is possible to detect associations in a genome-wide study using the current set of inbred mouse strains despite lower power due to the limited number of strains. If the region of interest can be narrowed down through completed QTL studies, the statistical power of the study increases significantly. The body weight phenotype with 38 strains achieves 80% power in genome wide set when the variance explained by SNP is larger than 30%. The saccharin preference phenotype with 24 strains could not achieve genome-wide power even with similarly strong effects. These results show that the number of available strains is critical to power.

We also performed a power analysis of the Arabidopsis and maize datasets containing 95 and 277
Figure 3 Comparisons of the statistical power of the EMMA method across three different inbred mouse phenotypes. Pointwise power denotes the power to identify causal SNPs without multiple hypothesis testing. Region-wide power applies a Bonferroni correction assuming 50 hypothetical tagSNPs in a genomic region. With 20kb between tagSNPs, the QTL covers up to 1Mb. Genome-wide power is the power to achieve genome wide significance using the p-value threshold $10^{-5}$, which is conservative compared to the permutation based genome-wide significance thresholds using the original phenotypes. The phenotypic variation explained by SNP effect is computed assuming a minor allele frequency (MAF) of 0.3.
Figure 4 Comparisons of the statistical power of EMMA method for Arabidopsis and maize panels with flowering time phenotypes. The causal SNPs are randomly selected across all SNPs, and an additive quantitative locus effect is simulated. Pointwise power denotes the power to identify causal SNPs at cutoff p-value of 0.05 without multiple hypothesis testing. Region-wide power applies a Bonferroni correction assuming 50 hypothetical tagSNPs in a genomic region. With a 20kb gap between the SNPs, the length of a QTL covers up to 1Mb. Genome-wide power is the probability of achieving genome wide significance using the threshold p-value $10^{-5}$.

strains respectively. The results show that, with a large number of strains, the genome-wide power is high even for relatively small effects. Since the effect of single marker or gene is likely to be moderate or small for most complex traits, our results highlight the importance of phenotyping a larger number of strains than currently available in order to fully realize the potential of whole genome association mapping in inbred mouse strains.

In order to observe the effect of population structure on the statistical power, we generated simulated phenotypes based on the kinship matrix of 48 inbred mouse strains with different effect of population structure by changing the ratio between the genetic and environmental variance components. Figure 5(a) shows that the power of the studies significantly decreases with the same effect on the phenotypes as the effect of population structure increases. We also tested how having multiple measurements of the phenotype per strain affect the power of association studies. Figure 5(b) shows the simulation results for association in the mouse panel without multiple measurements. Without multiple measurements, the power of the study substantially decreases, especially when the population structure explains small portion of phenotypic variance. An interesting observation is that for the single measurement datasets, statistical power is positively correlated with the effect of population structure, while it is negatively correlated in the multiple measurement datasets. This implies that the improvement of power gained through having multiple measurements for the strains is limited when phenotypic variation is highly correlated with population structure.
Figure 5 Comparisons of the statistical power of the EMMA method applied to inbred mouse association for simulated phenotypes with various effect of population structure and number of multiple measurements. Each causal SNP is generated randomly with minor allele frequency $p=0.03$. Genome-wide power is the probability of achieving genome wide significance using the threshold $p=10^{-5}$.

Figure 6 Genome wide cumulative distribution of observed p-values across 13,416 Arabidopsis SNPs using (a) MM only model (Simple linear regression without kinship matrix) and (b) SA+MM model for various kinship matrices in the association mapping for flowering time phenotypes. 'Simple' denotes the p-values obtained by simple linear regression, 'SA' is Structured Association, which uses the population coefficient obtained from STRUCTURE (k=8) as covariates in the regression model. 'MM(gendist)', 'MM (hapdist)' and 'MM (phylo)' are the mixed models based on genotype distance, haplotype distance, and phylogenetic control, respectively. The two 'SA+MM' methods are combined methods between the SA and MM models.

2.3 High Resolution Genome Wide Association Mapping in Arabidopsis

We applied our EMMA method to perform high resolution genome-wide association mapping between 13,416 Arabidopsis SNPs and a flowering time phenotype. The FRI gene has been shown to
be significantly associated to this phenotype in previous studies[3]. The cumulative distribution of p-values from EMMA is shown in Figure 6. It nearly follows the expected distribution, implying that our mixed model significantly outperforms Structured Association in eliminating the inflation of false positives for this dataset. In the region of nominal p-values, the phylogenetic control and genetic similarity matrix show almost no inflation, suggesting that the significant signals are likely to be true positives.

The REML-based Bayesian Information Content (BIC) was substantially decreased using the mixed model compared to the simple linear model, yielding likelihood ratio test p-values less than $10^{-20}$ across all three kinship matrices. Similarly, introducing a mixed model to Structured Association significantly increased the likelihood with p-values less than $10^{-11}$. Among the three kinship matrices we suggested, phylogenetic control resulted in slightly better REML estimates.

Figure 7 shows the results of a genome wide scan for association mapping. The previously known FRI gene, located in the beginning of chromosome 4, was found to be the most significant with phylogenetic control. Although Structured Association also identifies this locus as significant at a genome-wide level, there are many other genome wide significant signals which are probably false positives caused by the residual inflation of p-values. The fact that only the FRI locus still remains significant in our phylogenetic control method supports that the association found by EMMA is likely to be a true positive.

### 2.4 High Resolution Genome Wide Association Mapping in Inbred Mouse Strains

We performed a high resolution genome wide association mapping study using our mixed model method over inbred mouse strains. We used the Broad Mouse HapMap SNPs, containing nearly 140,000 SNPs and covering approximately 96% of the variation in 48 strains (Daly et. al, unpublished). For phenotypes, we used initial body weight and liver weight phenotypes downloaded from Jackson Laboratory Mouse Phenome Database[31]. In addition, we used a saccharin preference phenotype where statistically significant associations were identified in a previous study[28]. Among 48 genotyped strains, 38, 34, and 24 strains had phenotype values available for body weight, liver weight, and saccharin preference, respectively. Each phenotype has up to 10 repeated measurements across different mouse strains.

The cumulative distributions of observed p-values in Figure 8 shows that, without correcting for population structure, the rate of false positives are very high. In particular, the body weight phenotype has a substantial inflation of false positives. When our mixed model is used, the inflation of the statistics are significantly reduced in all three phenotypes.

Figure 9 shows genome wide association signals for the three phenotypes. Comparing Figure 9(a)
Figure 7 Genome wide scans for association with a flowering time phenotype. Each impulse corresponds to the unadjusted log-p value obtained from our mixed model using (a) no correction for population structure. (b) Structured Association with k=8. (c) A mixed model based on phylogenetic control. The horizontal lines are cutoffs for genome wide significance at a p-value of 0.05 with family wise error control (FWER) method obtained from 10,000 permutation tests. The most significant SNP in the beginning of chr4 is the known FRI gene.
Figure 8 Cumulative distribution of observed p-values across genome wide Broad Mouse HapMap SNPs. Simple linear regression is compared with our mixed model method based on genotype similarity, haplotype similarity, and phylogenetic control. The closer the cumulative distribution is to the ideal line, the smaller the inflation of p-values the statistical test has.
Figure 9 Genome wide scans for association with initial body weight, liver weight, and saccharin preference using different statistical tests. The genome wide significance level is obtained from 10,000 permutation tests. The mixed model is applied using the kinship matrix based on haplotype similarities.
and 9(b), it is obvious that, without correcting for population structure, many false positives are observed at a genome wide level of significance due to inflated p-values. Without correcting for population structure, we were able to identify 2221 SNPs above the threshold of genome wide adjusted p-value 0.05. However, none of them were significant after applying the mixed model. This strongly supports that most of the significant associations without correcting for population structure are false positives. Interestingly, although the strongest signals for the body weight with the mixed model are not genome wide significant, they are concentrated in the region around 114Mb in chromosome 8. This region falls perfectly into the LOD peak of a previously known body weight QTL \( Bwq3 \)[2]. If the region of interest was narrowed down to 1Mb region through a separate study, these signals would be identified significant over the genomic region.

For the liver weight phenotype, we identified a genome wide significant association around the region of 34.5Mb in chromosome 2. This falls into a previously known liver weight QTL \( Lvrq1 \)[29]. The region also contains many relevant QTLs such as organ weight \( (Orgwq2)[20] \), spleen weight \( (Sp1q1)[29] \), heart weight \( (Hrtq1)[29] \), lean body mass \( (Lbm1)[22] \). The pointwise p-value was \( 2.24 \times 10^{-5} \), barely crossing the permutation-based genome wide threshold of \( 2.40 \times 10^{-5} \). The p-value is much stronger when corrected within the QTL, and same locus was identified to be the most significant across both male and female mice.

For the saccharin preference phenotype, we were able to identify a SNP near the \( Tas1r3 \) gene that is perfectly correlated with the SNP previously reported to have significant association with the phenotype \[28\]. The SNP is neither genome-wide significant nor the most significant. We believe this is due to the limited power of the study with small number of strains. The pointwise p-value of the SNP was \( 1.03 \times 10^{-3} \), and the signal is strong enough to be identified as significant in a 1Mb region even though the SNP is only 30kb from the \( Tas1r3 \) gene.

### 3 Discussion

Our maize, Arabidopsis, and inbred mouse strain results show that our mixed model robustly reduces the inflated false positive rate when the phenotypic correlations are corrected by variance components estimated from simple genetic similarity or a phylogenetic tree. The Structured Association method reduced only a portion of inflation of false positives while our mixed model almost completely eliminates the inflation in most datasets. Despite this reduction in p-value inflation due to population structure, there are likely other confounding factors[35] that can cause false positives and there remains a need for effective strategies to resolve these confounding factors.

When population structure explains a large fraction of phenotypic variance among the strains, the power to identify statistically significant quantitative loci decreases. This is one of the main challenges in model organism association mapping where the number of distinct inbred strains are
typically limited. However, the fact that many of the most significant signals in our results fall into previously known QTLs motivates the design of more powerful studies to identify true causal variants.

In order to overcome such limitations, it would be beneficial to utilizes many more inbred strains such as recombinant inbreds or consomics which have a larger number of genetically distinct individual mice. Furthermore, multi-stage association study design which combines independent studies to narrow down the region of interest can be applied to increase the cost-effectiveness of the association mapping as in human disease association. In this case, our results can be directly used in the fine mapping of QTLs obtained from F2 crosses or RI strains, or in filtering regions of interest for the first stage of multi-stage studies.

The efficiency of our EMMA method expedites genome wide association mapping in model organisms. In our inbred mouse association mapping results involving more than a hundred thousand markers, a single EMMA run took only one minute per phenotype.

4 Materials and Methods

4.1 Genotypes and Phenotypes

The maize panel dataset contained 553 SNPs across 277 individual strains. We use the same SNPs, phenotypes, SPAGeDi-based kinship matrix, and STRUCTURE results described in [34]. The Arabidopsis genotypes, phenotypes, and the output from STRUCTURE were obtained from the published datasets[3, 23]. The 13,416 non-singleton SNPs with no more than 10% of genotype calls missing, were tested for association after imputing the missing alleles using HAP[14]. The flowering time phenotypes over 95 strains were log-transformed to fit to a normal distribution. For inbred mouse association mapping, the BROAD mouse HapMap SNP sets were obtained from the Mouse HapMap web site at http://www.broad.mit.edu/~mjdaly/MouseHapMap/. The 106,040 SNPs which have no more than 10% of genotype calls missing were tested after imputing the missing alleles. The initial body weight(MPD10305) and liver weight phenotypes(MPD2907) were downloaded from Jackson Laboratory MPD[31]. They consist of 374 and 308 phenotype measurements over 38 and 34 strains, respectively. The saccharin preference phenotypes consist of 280 phenotype measurements in 24 strains[28].

4.2 Efficient Mixed Model Association (EMMA)

A linear mixed model is typically expressed as the following equation.
\[ \mathbf{y} = G\alpha + X\beta + Z\mathbf{u} + \mathbf{e} \tag{1} \]

where \( \mathbf{y} \) is a \( n \times 1 \) vector of observed phenotypes, \( G\alpha \) is \( n \times p \) matrix of fixed effects (e.g. SNPs) to be tested against null hypothesis, and \( X\beta \) is \( n \times q \) matrix of confounding fixed effects such as mean, sex, age, and population structure coefficients computed from STRUCTURE. \( Z \) is \( n \times t \) incidence matrix mapping each observed phenotype to one of \( t \) inbred strains. \( \mathbf{u} \) is the random effect of the mixed model with \( \text{Var}(\mathbf{u}) = \sigma_g^2 K \) where \( K \) is the kinship matrix, and \( \mathbf{e} \) is a \( n \times n \) matrix of residual effect such that \( \text{Var}(\mathbf{e}) = \sigma_e^2 I \).

Instead of solving mixed model equations by obtaining the best linear unbiased prediction (BLUP) of random effects \( \mathbf{u} \) via Henderson’s iterative procedure, we directly estimate the variance components \( \sigma_g \) and \( \sigma_e \) maximizing the restricted likelihood function. The restricted likelihood is commonly used in mixed models since it avoids a downward bias of maximum likelihood estimates of variance components. Under the null hypothesis, the restricted log-likelihood function can be formulated as follows \cite{33}.

\[
RL = \frac{1}{2} \left[ (n - q) \log(2\pi) - \log |X'X| - \log |X'V^{-1}X| - \log |V - (X - X\beta)'V^{-1}(X - X\beta)| \right] \tag{2}
\]

where \( V = \sigma_g^2(ZKZ' + \delta I) \), \( \delta = \sigma_e^2/\sigma_g^2 \), and \( \beta = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y} \). This equation holds under the alternative hypothesis if \( X \) is replaced with \([G \ X]\) and \( q \) with \( p + q \). For fixed \( \delta \), the \( \sigma_g \) maximizing the restricted likelihood becomes

\[
\hat{\sigma}_g^2 = \frac{(\mathbf{y} - X\bar{\beta})'H^{-1}(\mathbf{y} - X\bar{\beta})}{n - q} \tag{3}
\]

where \( H = ZKZ' + \delta I \). The only remaining variance component \( \sigma_e^2 \), can be estimated by performing one dimensional numerical optimization of restricted likelihood with respect to \( \delta \). Since the restricted likelihood function is not always concave with respect to \( \delta \), a grid search needs to be performed to find globally the optimal solution. Such a procedure may be inefficient due to repetitive computation of matrix inverses and multiplications if \( n \) is large. Our method avoids such a redundancy and greatly improves the efficiency in the optimization procedure. Let us denote \( S = I - X(X'X)^{-1}X' \), then the restricted likelihood function in equation 2 can be rewritten as follows \cite{25}.

\[
RL = \frac{1}{2} \left[ (n - q) \log(2\pi) - \log |X'X| - \log |X'H^{-1}X| - \log |H| - (n - t) \log \sigma_g^2 - \frac{\mathbf{y}'(SHS)^{-g}\mathbf{y}}{\sigma_g^2} \right] \tag{4}
\]
Since $S$ is a singular matrix, the singular value decomposition of $SHS$ can be expressed in the following form,

$$SHS = [U \ W] \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U' \\ W' \end{bmatrix} = UDU'$$ (5)

where $U$ is an $n \times (n - q)$ matrix such that $U'U = I$ and $UU' = S$ and $D$ is a diagonal matrix. More generally, for any $H = ZKZ' + \delta I$, the equation can be expressed as follows,

$$SHS = S(ZKZ' + \delta I)S = U\text{diag}(\lambda_s + \delta)U'$$ (6)

where $\lambda_s + \delta$ are the latent roots of $SHS$. Taking their generalized inverse, the equation becomes

$$(SHS)^{-g} = \{S(ZKZ' + \delta I)S\}^{-g} = U\text{diag}\{(\lambda_s + \delta)^{-1}\}U'$$ (7)

Set $U'y = [u_1 \ u_2 \ \cdots \ u_{n-q}]'$, then $U'y$ will follow a normal distribution with diagonal covariance matrix whose $s$-th diagonal element is $(\lambda_s + \delta)\sigma_g^2$[25]. The restricted likelihood can then be expressed in terms of these latent roots.

$$RL = \frac{1}{2} \left[ (n - q) \log(2\pi) - \log|X'X| - \sum_s \log(\lambda_s + \delta) - (n - t) \log \sigma_g^2 - R/\sigma_g^2 \right]$$ (8)

where

$$R = \sum_s \left( \frac{u_s^2}{\lambda_s + \delta} \right) = y'(SHS)^{-g}y$$ (9)

Using equation 4, the restricted likelihood for any $\delta$ can be computed without repetitive matrix inverses or multiplications. Moreover, it is possible to efficiently compute the derivative of restricted likelihood with respect to $\delta$, enabling us to apply the Newton-Raphson algorithm for optimization.

$$\frac{\partial(RL)}{\partial \delta} = -\frac{1}{2} \sum_s \left( \frac{1}{\lambda_s + \delta} \right) + \frac{(n - q)\sum_s (u_s/(\lambda_s + \delta))^2}{2\sum_s (u_s^2/(\lambda_s + \delta))}$$ (10)

In the application of our method to the various datasets presented in this paper, we transformed the range of $\delta$ from $10^{-10}$ (almost pure population structure effect) to $10^{10}$ (almost pure environmental or residual effect), divided them evenly into 100 regions, and computed the derivatives of restricted likelihood. The global maximum restricted likelihood is searched for by applying the Newton-Raphson algorithm to all the intervals where the signs of derivatives change, and taking the optimal $\delta$ amongst all of the stationary points and endpoints. Since the derivative of the restricted likelihood
function is continuous, such an optimization technique has guaranteed convergence properties as long as the kinship matrix is positive semidefinite. In addition, since the restricted likelihood function usually does not have many stationary points, this technique almost always estimates the global REML parameters.

In the following two sections, we describe different methods to infer a kinship matrix $K$, based on either genetic similarity matrix or phylogenetic tree.

### 4.3 Similarity-based Kinship Matrix

The main assumption underlying the similarity-based kinship matrix is that, under the null hypothesis, each SNP or haplotype causes the same level of random genetic drift on the phenotype. Let $l_{i,j,h} \in \{0, 1\}$ be a binary variable which has a value of one only when the genotype (or haplotype) allele at $j$-th locus (SNP or haplotype block) in $i$-th strain is $h \in 1, \ldots, H_j$, and let $x_{h,j}$ be random variables independently sampled from $N(0, \sigma^2)$. Then the genetic random drift $u_i$ of strain $i$ can be modeled as an accumulation of random effects,

$$u_i = \sum_j H_j \sum_{h=1}^{H_j} l_{i,j,h} x_{h,j} \quad (11)$$

assuming that $x_{h,j}$ denote the random genetic drift caused by allele $h$ at $j$-th locus.

Let $H = \max(H_j)$, and let $L_h$ be the matrix whose element at $(i, j)$ is $l_{i,j,h}$, then the overall genetic drift $u$ is expressed in the following form.

$$u = \sum_h L_h x_h \quad (12)$$

The covariance matrix $u$ becomes $\text{Var}(u) = \sigma^2 \sum_h L_h L_h'$, whose $(i, j)$-th element is eventually proportional to the number of identical by state (IBS) alleles between the $i$-th and $j$-th strains. Thus, $\text{Var}(u)$ is proportional to the distance based kinship matrix $K_d$, which is equivalent to the genotype or haplotype similarity matrix. It is obvious from the equation 12, that the kinship matrix is positive semidefinite. When missing genotypes exists, we estimate $l_{i,j,h}$ to be the square root of the probability of the SNP or haplotype allele at $j$-th locus having the allele $h$. This is so that the random drift for each allele is assigned probabilistically. The window size of the haplotype block is set to five SNPs in our mouse association mapping studies.
4.4 Phylogenetic Control

Evolutionary biologists have tried modeling inter-specific phenotype distribution using various phylogenetic comparative methods (PCMs)[21]. The correlation structure between phenotypes can be effectively captured with phylogenetic trees, and PCMs have been applied to evolutionary analysis of quantitative traits such as gene-expression[24, 13]. Felsentein’s independent contrast (FIC) method[9] models the correlation between phenotypes under the assumption of Brownian motion of phenotypic change along the phylogeny due to random genetic drift. Since random genetic drift occurs within a species as well, in cases where the phylogenetic tree is a good approximation of genealogical history, it is reasonable to apply PCMs such as the FIC method in modeling the phenotypic variation in model organisms.

We followed the Felsenstein’s assumption of Brownian phenotypic changes along the phylogeny. Under this assumption, the branch length between any two nodes is proportional to the phylogenetic covariance of phenotypes. Let $T$ be a phylogenetic tree with $t$ leafs and $m$ edges, and let $z \in \mathbb{R}^m$ be random variables independently sampled from $N(0, \sigma_g^2)$. At each branch $i$ whose length is $b_i$, we represent the amount of random phenotypic changes along the branch as $\sqrt{b_i}z_i$. Let $\Psi_i$ denote the set of branches connecting to a leaf node $i$ from the root. Then the amount of phenotypic changes due to genetic drift is equivalent to $\sum_{e \in \Psi_i} \sqrt{b_e}z_e$. If $G\alpha + X\beta$ is the ancestral mean at an arbitrarily chosen root node, then the phenotype values at the leaf nodes are expressed in the following form,

$$y = G\alpha + X\beta + ZEz + e$$

(13)

where $E$ is an $t \times m$ matrix whose $(i, j)$-th element is $\sqrt{b_j}$ if branch $j$ exists in the path from the root to the leaf node $i$, and zero otherwise. The kinship matrix of random effect $u = Ez$ is $K = EE'$, and is proportional to its covariance. If the root of the phylogenetic tree changes, $E$ is changed into $E + 1_t c^T$, with $1_t$ a vector of ones and another vector $c$. However, the restricted likelihood does not change because $SZ1_t = 0$ always holds.

In the experiments, we computed the genetic distance matrix using the F84 model[19, 10] from the genome wide genotypes, and inferred the phylogenetic tree with the Fitch-Margoliash and least-squared distance method[11].

4.5 Statistical Tests and Multiple Hypothesis Testing

Once the REML variance components $\sigma_g^2$ and $\sigma_e^2$ are estimated, a general F-statistic testing the null hypothesis $M'\alpha = 0$ for an arbitrary $p \times 1$ matrix $M$ can be constructed as suggested in $[34, 18]$

$$F = \frac{\hat{\alpha}'M([M' 0]([GX]'\hat{V}^{-1}[GX])^{-1}[M' 0]')^{-1}M'\hat{\alpha}}{\text{rank}(M)}$$

(14)
with rank($M$) numerator degrees of freedom and $n - p - q$ denominator degrees of freedom, where $\hat{\alpha}$ is computed from the best linear unbiased estimator (BLUE) of fixed effects, 
\[
\begin{bmatrix}
\hat{\alpha} \\
\hat{\beta}
\end{bmatrix} = ([G X]'\hat{V}^{-1}[G X])^{-1}[G X]'\hat{V}^{-1}y.
\]

We use a modified statistic called restricted $F$ statistic, based on the BLUE of $\alpha$ under the linear model integrated over $X$
\[
RF = \frac{\hat{\alpha}'M'(G'\hat{\Lambda}G)^{-1}M)^{-1}\hat{x}}{\text{rank}(M)}
\]

where $\hat{\Lambda} = (S\hat{V}S)^{-q}$ and $\hat{\alpha} = (G'\hat{\Lambda}G)^{-1}G'\hat{\Lambda}y$ For both $F$-tests, the Satterthwaite degree of freedom may also be computed with latent roots, avoiding computationally intensive matrix operations. Likelihood ratio test can also be performed based on the estimated variance components,
\[
LR = (n - q) \ln \left[ \frac{y'\hat{\Lambda}y}{(y - G\hat{\alpha})'\hat{\Lambda}(y - G\hat{\alpha})} \right]
\]

where the statistic asymptotically follows $\chi^2_p$ distribution.

In our results, we applied the restricted $F$-test and pointwise p-values are computed from the asymptotic $F$ distribution. In order to adjust the p-values for multiple hypotheses, we applied standard FWER (Family-Wise Error Rate) control[8] method which controls the probability of observing one or more false positives. The cutoff probability was computed from 10,000 genome wide permutation tests. When permuting the phenotypes, only the SNP matrix is permuted so the correlation between the phenotypes and the kinship matrix is preserved.

References


totypic, pedigree and marker data in self-pollinated crops. *Theoretical and Applied Genetics*, 
112(5), 876–884.

analysis in large pedigrees. *Genetic Epidemiology*, 30(6), 471–84.

997–1004.


step procedures for control of general type i error rates. *Statistical Applications in Genetics and 
Molecular Biology*, 3(13).

125(1).


cable method utilizing estimates of the mutation distance obtained from cytochrome c sequences. 
*Science*, 155.

population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal*, 
44(6), 1054–1064.


Ontario.


25