Making multiple-locus association mapping on a genome-wide scale routine

Andrew W. George[1] and Joshua Bowden[2]

[1]*Data61, CSIRO, Brisbane, Australia.*

[2]*IM &T, CSIRO, Brisbane, Australia.*

Since the earliest of genome-wide association studies, a key shortcoming in how their data have been analysed has persisted. The strength of association between a marker locus and trait is measured for each locus separately, on a locus-by-locus basis. Multiple-locus methods that map multiple locus-trait associations simultaneously have been available for some time. However, they have attracted little attention. They can be demanding, computationally, and their results are not always easy to interpret. Yet, it is widely accepted that multiple-locus methods are superior, statistically, to locus-by-locus methods. Here, we present our method that makes the multiple-locus analysis of data from genome wide association studies routine. It is formulated within a linear mixed model framework. We call our method AMplus. AMplus produces results faster than competing multiple-locus methods and often with greater statistical power. Also, it is just a little slower than the fastest single-locus linear mixed model implementations. AMplus is freely available as a fully documented R package.

Over the past decade, Genome wide association studies have changed considerably in both their analysis and design. Early gwas followed a case-control design. Association mapping methods were no more complicated than contingency table tests or regression. These designs though

had a tendency to yield spurious findings if there was unrecognised population stratification. This prompted a shift towards family-based designs and score tests, such as the tdt test and its variants (refs). Today, instead of by design, it is through statistical modelling that we account for the effects of population stratification. This has meant that data can be collected from general populations, even if these populations are highly structured. Analysis via sophisticated association mapping methods, such as linear mixed model based approaches, is now almost routine.

What has not changed over the past decade is that it remains common practice to analyse gwas data on a locus by locus basis. This is despite there being several significant problems with analysing data in this way. First, the aim of association mapping is to identify regions of the genome that house genes that are influencing a trait. The identification of these regions from these analyses is not always straightforward. Gwas results are reported, typically, via Manhattan plots that plot the -log10 of the p value for each locus against the map position of the locus. The location of peaks in this plot signify regions of interest. Inferring the exact number of regions of interest can be difficult If the peaks are not well separated. Second, when multiple statistical tests are performed, the probability of wrongly accepting a significant result (type 1 error) is inflated. This is known as the multiple testing problem (refs). Many different solutions have been offered (refs). Yet, there is still no well accepted way of correcting for multiple testing in the context of genome- wide association mapping. Third, many of the traits whose genetic secretes we are trying to discover are complex. There will be multiple loci in linkage equilibrium with genes that influence the trait. Yet, A locus by locus mapping approach only assesses the evidence for association between a single marker locus and trait.

It is somewhat surprising then that multiple locus association mapping methods haven't attracted more attention. Methods based on regularisation techniques, such as ridge regression and lasso, measure all locus-trait associations simultaneously. Here, multiple testing is not an issue. These techniques though are computationally demanding. Also, their results can be difficult to interpret. The strength of association is not measured by a p-value but by the size of the regression coefficient for a locus in the model. More recently, associations have started to be mapped with random forests (refs). Similar to regularisation techniques though, it is not clear how to infer genomic regions of interest from their findings (refs). An multiple locus method that does show promise is the multi-locus linear mixed model method (ref). The best multiple locus model is built with simple variable selection. Results are immediately interpretable but here, computation can be a challenge for large data.

In this paper, we present our new multiple locus method for genome wide association mapping, which we are calling Eagle. Eagle combines the strength of regularisation techniques (being able to fit all locus-trait associations jointly), with simple variable selection (having easy to interpret results). Our method does not require a significance threshold to be set nor regularisation parameters to be fine tuned. Through a clever dimension reduction step, we are able to achieve a computational performance similar to the fastest single locus linear mixed model implementations. Eagle is statistically more powerful than single locus association mapping and is as and often more powerful than most multiple locus methods. Our aim is to make multiple locus association mapping on a genome wide scale routine. To this end, we have created a fully documented R package, that is easy to use, even for non R users. Our package accepts marker data of different formats and

can handle data larger than a computer's memory capacity. It includes detailed error checking and makes heavy use of distributed computing for computation when available. The purpose of this work was to make multiple-locus association mapping on a genome-wide scale, for large data sets, practical and we have built AMplus accordingly.

1 Results

Association Mapping Methods We compared AMplus, in terms of computational and statistical performance, against seven other association mapping methods. We chose methods that had been purpose built for genome-wide analysis, that could handle data from quantitative traits, and where the methods had been implemented in freely available computer software. Two of the methods are based on single locus (or locus-by-locus) models and five are based on multiple locus models. Of the many ways of performing single locus association mapping, we chose GEMMA and FaST-LMM because of their popularity and computational speed. For multiple locus association mapping, we chose bigRR, glmnet, LMM-Lasso, MLMM, and r2VIM. Each takes a different approach to mapping multiple locus-trait associations jointly.

From a practical perspective, how the methods were implemented, in terms of usability, varied greatly. In Table 1, we list seven important features for ensuring the usability of computer software for association mapping. We also identify which software has what features. In forming this list of features, we assumed that the primary purpose of a genome-wide association study is to identify the set of marker loci in true and strongest association with a trait. The level of

documentation varied across the implementations but only a few came with user manuals detailing the format of the input data and how to perform analyses. Most implementations could handle additional fixed effects which is important when accounting for hidden population structure. Few though could cope with marker data larger than the memory capacity of the computer. Similarly, few returned the best set of marker loci in strongest association with a trait as their result. Almost all the software returned results that required further processing, either by identifying the best set of marker loci via the setting of significance thresholds or having to do further analysis to calculate the significance of the results. Almost none of the association mapping software checked for errors in the input data. Only AMplus has all seven features.

Computational Performance: run times and memory usage To asses the computational performance of AMplus, we conducted a large simulation study. We were interested in the impact of study size on performance so we generated data under five different scenarios. These are a GWAS of size 150 individuals and 5000 snp (150 X 5K), 350 individuals and 400000 snp (350 X 400K), 1500 individuals and 50000 snp (1500 x 50K), 2000 individuals and 500000 snp (2000 x 500K), 4000 individuals and 1500000 snp (4000 x 1.5M), and 10000 individuals and 1500000 snp (10000 x 1.5M). We chose these scenarios to reflect some of the different sized GWAS being performed in animals, plants, and humans. For each scenario, we generated 100 replicates of data. A single replicate consists of snp genotype data and quantitative trait data. We obtained the snp data from the publicly available 1000 genome project (phase 3). The quantitative trait data we generated from the snp data by selecting a set of snp loci, assigning allelic effects to these snp, and aggregating these effects for each individual along with random error. The number of snp selected

per replicate follows a Poisson distribution with mean 30. The quantitative trait was generated to have a heritability of 50%. Analyses were performed on a high end desktop computer. It had dual 8-core Xeon processors, three Kepler Tesla GPUs, and 128 Giggabytes of RAM. All implementations except, GEMMA, made use of distributed computing, either explicitly or implicitly through multi-threaded BLAS/LAPACK libraries.

The run times for AMplus against the other software programs, across the five scenarios, is shown in figure 1 To help with interpretibility of the results, in both plots, we have taken the x and y axes to be on a log scale. This means a unit change on the x or y axes is equivalent to a change in the order of magnitude. In the top plot, a point is the median of the ratios of elapse times of the multiple locus method to AMplus for a given scenario. The median is over the 100 replicates. Here, since the median of the ratios are all positive on the log scale, it means that AMplus had a shorter run time than than all the other multiple locus methods. In fact, in some cases, AMplus was over a hundred times, or over two orders of magnitude, faster. Unlike AMplus, the size of data under scenario 10000 X 1.5M was beyond the memory constraints of the other multiple locus implementations. This was also the case for LMM-Lasso, bigRR, and glmnet, for scenario 4000 x 1.5M.

In the bottom plot of figure 1, we compare the median run time of AMplus against the median run times of the single locus methods, FaST-LMM and GEMMA. FaST-LMM was run in two ways. It was run where the genetic similarity matrix was built with all the marker data (FaST-LMM^{all}) or with data on every 500th snp (FaST-LMM^{few}). AMplus was also run in two

ways. The default behavior for AMplus is to make use of CPUs for computing. However, AMplus also has the capacity to harness multiple GPUs (AMplus^{G}PU). From the bottom plot, all the implementations have short run times when analysing data from scenario 150x5K. However, for the other scenarios, AMplus and AMplusGPU have significantly shorter run times than GEMMA and FaST-LMM^{a}ll. FaST-LMM^{f}ew has a shorter run time than AMplus and AMplus^{G}PU but only for scenarios 1500x50K and 10000x1.5M. Furthermore, for scenario 10000x1.5M, the median run time for GEMMA is 4071 minutes, for AMplus is 699 minutes, for AMplus^{G}PU is 447 minutes, and for FaST-LMM^{f}ew is 346 minutes. The very fast single-locus method FaST-LMM^{f}ew is only 29% faster than our multiple-locus method AMplus^{G}PU. It is worth noting though that there is a setup cost to accessing GPU computing, making AMplus^{G}PU most efficient on the larger data.

We also examined the memory usage of the different software programs (figure 1 supplementary materials). AMplus is comparable to the most efficient single locus implementation, FaST-LMM^{f}ew. AMplus is also the only multiple locus program able to analyse data under scenarios ..... and ..... . AMplus can process data larger than the memory capacity of the computer.

Statistical Performance: power, FDR, and significance thresholds As part of assessing the performance of AMplus against the other association mapping methods, we calculated their statistical power and false discovery rates. We want a method to have high statistical power (probability of finding a true positive finding) but low false discovery rate (proportion of findings that are false positives). All the methods, besides AMplus and MLMM, required a threshold to be set in order to identify significant findings. A conservative threshold guards against false discoveries but

reduces power. A anti-conservative threshold increases power at the cost of also increasing the false discovery rate. Only AMplus and mlmm avoid thresholds by treating association mapping as a model selection problem where the best model is found by minimising the model selection statistic, extBIC.

Figures 1 and 1 show the relationship, calculated empirically, between the power and fdr for each method. The multiple-locus (Figure 1) and the single-locus (Figure 1) methods are plotted separately for clarity. AMplus features in both plots for comparison. By varying the significance thresholds for the different methods, we were able to calculate the change in power and fdr. Since AMplus and MLMM do not rely on thresholds, their power and fdr appear as single points in the plots. In practice, the true power and fdr is unknown. However, because we generated data where we knew which loci were acting as quantitative trait loci, we were able to calculate the true power and fdr for the methods.

From figures 1 and 1, the superior performance of AMplus is apparent. When the fdr is low, AMplus has the highest power with Mlmm also performing well. AMplus is noticeably more powerful than the single locus methods. It is possible to set a significance threshold for these other methods that results in higher power. However, this comes at the cost of an intolerably high fdr.

For the methods evaluated here that require a significance threshold, figure 1 shows how threshold size impacts fdr. Each plot corresponds to the results from a different method. A curve within a plot is formed from analysing data with the method and using different threshold values to identify locus-trait associations. The true fdr is then calculated from these findings. We can

see from figure 1 that the threshold impacts the fdr differently across methods. The relationship between threshold value and fdr even differs when data are analysed with the same method but for different scenarios. This demonstrates one of the challenges in using association mapping methods that are reliant upon thresholds. With AMplus, we do not have this challenge. For completeness, for the small sampled size scenarios ... And ....., the fdr for AMplus was 0.07 and 0.09, respectively. With the larger study sizes of ..., ..., and ...., the fdr drops to ..., .., and ..... respectively.

Mouse Study

2 Methods

Outbred Mice Study Phenotypic and genotypic data were collected on a large study population of commercially available outbred mice. The aim of the study was to map, through association, the genetic basis of complex traits in mice. The construction of the population and collection of data are described in detail in [1]. Briefly, the population consists of 2073 adult outbred mice from Charles River Laboratories. Phenotypic data were obtained from 200 traits. These traits fell into the categories of behaviour, tissue, and physiology. Genotype data were collected from 7, 073, 398 SNPs. A reduced set of 359, 559 SNP across the 19 autosomes and sex chromosome was then identified, upon which all genome-wide analyses were based. These SNP tagged all other SNP with MAF $> 0.1\%$ and LD $r^2 > 0.98$. Missing data?

Eagle Approach for Multiple-locus Association Mapping

2.0.1 Overview

Eagle is an approach that identifies genomic regions of interest for a trait via the building of a (linear mixed) model. The model is built iteratively, without need of significance levels or thresholds. The model has two parts, a fixed effects part and a random effects part. The random effects part contains the error and an effect for the snp-trait association across the entire genome. It is from this effect that we are able to identify which snp is in strongest association with the trait. Our model building process is as follows. First we fit the current model to the data. Second, from the random effect for association, we identify which snp is in strongest association with the trait. Third, we move this snp from the random effects to the fixed effects. We do this to stop the neighbours of the chosen snp from being selected in subsequent iterations and to give opportunity to snp in linkage disequilibrium with genes having lesser impact on the trait to be discovered. Fourth, we calculate if the new model is ”better” than the current model. If so, the new model becomes the current model and we repeat the process. Once the model building process is complete, the findings from Eagle are the set of snp that have been moved into the fixed effects part of the model. These snp map the seperate regions of the genome that house genes that are influencing the trait.

2.0.2 The specifics

The Eagle approach for multiple-locus association mapping is based on building a linear mixed model, iteratively, via forward selection. It is a simplification of a whole-genome method for

mapping quantitative trait loci in classic linkage studies [2]. Suppose s iterations of our model building process have already been performed. This means s snp-trait associations have been identified. It also means that s separate genomic regions of interest have been found. To perform the $s+1$th iteration of our model building procedure, we do the following.

First, we fit the current model to the data. Let $S = \{S_1, S_2, \ldots, S_s\}$ be a set of ordinal numbers where S_k is the S_kth ordered snp that was selected in the kth iteration. For example, if three iterations of our model building procedure have been performed and say the 500023rd, 15th, and 420th, snp were selected, then $S = \{500023, 15, 420\}$. Let $y^{(n \times 1)}$ be a vector containing the quantitative trait data. Let $M^{(n \times L)} = [m_1 m_2 \ldots m_L]$ be a matrix containing the genotype data which have been collected from L loci that span the genome. It is common for the columns of this matrix to be in map order but this is not a requirement. The vector $m_j^{(n \times 1)}$ contains the genotypes, from the n individuals, for the jth snp. The genotypes are coded as -1, 0, and 1 corresponding to snp genotypes AA, AB, and BB, respectively.

gocard123

The (current) model is of the form

$$y = X\tau + u_g + e \tag{1}$$

where $X^{(n \times p)}$ is a known design matrix of full rank and may contain data on non-genetic effects such as age and height and/or principal component values to account for population structure (ref). The vector $\tau^{(p \times 1)}$ has p fixed effects parameters. The vector $u_g^{(n \times 1)}$ contains the genetic effects.

The vector of residuals is $e^{(n \times 1)}$ whose distribution is assumed to follow $N(0, \sigma_e^2 I^{(n \times n)})$. So far, this model differs little from standard linear mixed models for association mapping (refs). However, it is how we specify u_g that distinguishes our model from others.

The genetic effects u_g are modelled as

$$u_g = \sum_{k=1}^{s} m_{S_k} a_{S_k} + M_{-S} a_{-S} \tag{2}$$

where $m_{S_k}^{(n \times 1)}$ is the vector of genotypes for the S_kth snp locus which is the kth selected snp, a_{S_k} is the additive effect of the S_kth snp locus, $M_{-S}^{(n \times L-s)}$ is the matrix of snp genotypes with the data for the selected snp in S removed, and $a_{-S}^{(L-s \times 1)}$ is a random effect whose distribution is $a_{-S} \sim N(0, \sigma_a^2 I^{(L-s \times L-s)})$. The terms in the summation on the left hand side are fixed effects. The other term is a random effect. The terms in the summation account for the additive effects of genes that are in linkage disequilibrium with the selected snp. The other term models snp-trait associations along the entire genome, except for those snp that have already been selected. This is a simple genetic model but it is effective for discovering snp-trait associations. It also reduces the computational cost of our model building procedure.

Second, we estimate the parameters of (1) and (2) via residual maximum likelihood estimation.

Third, we identify the $(s+1)$th snp that is in strongest association with the trait, based on the maximum score statistic $t_j^2 = \frac{\widetilde{a}_j^2}{\text{var}(\widetilde{a}_j)}$ where $\widetilde{a}_j$ is the best linear unbiased predictor (BLUP), and var$(\widetilde{a}_j)$ is its variance. This statistic is not only appealing intuitively, where we identify a snp

based on its (random) effect size and accuracy, but is theoretically justified. It follows from outlier detection in linear and linear mixed models (ref).

Fourth, we determine the importance of the $(s+1)$th selected snp via a model selection strategy. We begin by reforming (2) where S now contains the $s+1$ selected snp. We then fit this new model to the data via maximum likelihood and calculate its extended Bayesian information criteria (ExtBIC) [3]. The ExtBIC is a model selection measure that takes into account the number of unknown parameters and the complexity of the model space. It is especially well suited to the model selection problem in genome-wide association studies [3]. If this new model has a larger ExtBIC than the current model, then the $s+1$th selected snp is added to the current model and the above process is repeated. If this new model has a smaller ExtBIC than the current model, then the model building process is complete. The set of snp in strongest association with the trait is the s snp previously identified.

2.0.3 Reducing the dimension of the model

In practice, estimating the parameters of (2) can be demanding, computationally. The vector a_{-S} has $L-s$ random effects where in modern genome-wide association studies, L, the number of snp, can be extremely large. An alternative model is given by Verbyla. They show how to reformulate (2) as ..... where ..... and XXX is calculated via s. v. d. Although it may not be obvious, the two models are equivalent, having identical variance structures. Yet, the computational cost of model (7) compared to model 2, in most modern studies, is much less because the sample size n is often

much smaller than the number of snp L.

Verbyla go on to show how to recover $\widetilde{a}$ from estimates from model (3) with

$$\widetilde{a} = \left[M_{-S}^T (M_{-S} M_{-S}^T)^{-1/2}\right] \widetilde{a}^* \tag{3}$$

and its variance matrix with

$$\text{var}(\widetilde{a}) = M_{-S}^T (M_{-S} M_{-S}^T)^{-1/2} \text{var}(\widetilde{a}^*) (M_{-S} M_{-S}^T)^{-1/2} M_{-S} \tag{4}$$

These values are needed to calculate the score statistic t_j^2 for identifying the snp in strongest association with the trait. Fortunately, when calculating t_j^2, only the diagonal elements of the variance matrix are needed which simplifies the calculation. of (4).

Fortunately

Since $\widetilde{a}$ and var$(\widetilde{a})$ are needed to calculate t_j^2, as described in Verbyla [4,5],

$$\widetilde{a} = \left[M_{-S}^T (M_{-S} M_{-S}^T)^{-1/2}\right] \widetilde{a}^* \tag{5}$$

with variance matrix

$$\text{var}(\widetilde{a}) = M_{-S}^T (M_{-S} M_{-S}^T)^{-1/2} \text{var}(\widetilde{a}^*) (M_{-S} M_{-S}^T)^{-1/2} M_{-S} \tag{6}$$

Fortunately, only the diagonal elements of the variance matrix are needed which simplifies its calculation.

It is not uncommon in genome wide association studies for the number of snp (L) to be many orders of magnitude larger than the sample size (n). As a result, estimating a in (2) can be

demanding, computationally. We avoid having to estimate a directly by instead reformulating (2) as the dimension reduced model

$$u_g = \sum_{k=1}^{s} m_{S_k} a_{S_k} + (M_{-S} M_{-S}^T)^{1/2} a_{-S}^* \tag{7}$$

where a^* is a $n \times 1$ vector of random effects that follows $N(0, \sigma_a^2 I^{(n \times n)})$, and $(M_{-S} M_{-S}^T)^{1/2}$ can be calculated via single value decomposition (ref). Models (2) and (7) are equivalent models [4,5]. They have the same variance structure, despite the random effect in (7) being only of size n. This allows us to use (7) in place of (2) in the above described Eagle approach, affording us considerable computational benefits. Since $\widetilde{a}$ and var$(\widetilde{a})$ are needed to calculate t_j^2, as described in Verbyla [4,5],

$$\widetilde{a} = \left[M_{-S}^T (M_{-S} M_{-S}^T)^{-1/2} \right] \widetilde{a}^* \tag{8}$$

with variance matrix

$$\text{var}(\widetilde{a}) = M_{-S}^T (M_{-S} M_{-S}^T)^{-1/2} \text{var}(\widetilde{a}^*) (M_{-S} M_{-S}^T)^{-1/2} M_{-S} \tag{9}$$

Fortunately, only the diagonal elements of the variance matrix are needed which simplifies its calculation.

1. Nicod, J. *et al.* Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. *Nature genetics* (2016).

2. Verbyla, A. P., Cullis, B. R. & Thompson, R. The analysis of qtl by simultaneous use of the full linkage map. *Theoretical and Applied Genetics* **116**, 95 (2007).

3. Chen, J. & Chen, Z. Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771 (2008).

4. Verbyla, A. P., Taylor, J. D. & Verbyla, K. L. Rwgaim: an efficient high-dimensional random whole genome average (qtl) interval mapping approach. *Genetics Research* **94**, 291–306 (2012).

5. Verbyla, A. P., Cavanagh, C. R. & Verbyla, K. L. Whole-genome analysis of multienvironment or multitrait qtl in magic. *G3: Genes, Genomes, Genetics* **4**, 1569–1584 (2014).

Acknowledgements Put acknowledgements here.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to A.B.C. (email: myaddress@nowhere.edu).

Figure 1 Run times

Figure 2 Memory usage

Figure 3 Power curves for multiple locus methods

Figure 4 Power curves for single-locus methods – Put into sup methods

Figure 5 The impact of threshold on FDR