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

Andrew W. George[1], Arunas Verbyla[1], Joshua Bowden[2], need to sort out the other authors

[1]*Data61, CSIRO, Australia.*
[2]*IM &T, CSIRO, 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 (GWAS) 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 simple linear 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 Eagle, in terms of computational and statistical performance, against seven other association mapping methods. We chose methods that (mostly) 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 programs. 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 [1] and FaST-LMM [2] because of their popularity and computational speed. For multiple-locus association mapping, we chose bigRR [3], glmnet [4], LMM-Lasso [5], MLMM [6] , and r2VIM [7]. Each takes a different approach to multiple-locus association mapping.

We list some of the key features that make computer programs for association mapping useful in Supplementary Table 1. We also identify which of these features are present in Eagle and the other seven implementations. For example, all but glmnet were purpose-built for association mapping. Not all the computer programs come with a user manual. Most could accomodate data on additional fixed effects but only Eagle and FaST-LMM could deal with data larger than the memory capacity of the computer. Also, if the input data contained errors, we found only Eagle and FaST-LMM tried to diagnose the problem and provided advice on how to fix the errors.

Simulation Study We performed a large simulation study where we sought to answer two questions. First, how does Eagle compare, in terms of run time and memory usage, to competing implementations? Second, how well does Eagle find true associations (power) and avoid false associations (type 1 errors)? We generated data under five different scenarios; a study of size 150 individuals and 5,000 single nucleotide polymorphisms (SNPs) (150 x 5K), 350 individuals and

400,000 SNPs (350 X 400K), 1,500 individuals and 50,000 SNPs (1500 x 50K), 2,000 individuals and 500,000 SNPs (2000 x 500K), 4,000 individuals and 1,500,000 SNPs (4000 x 1.5M), and 10,000 individuals and 1,500,000 SNPs (10000 x 1.5M). We chose these scenarios to mirror 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 data and quantitative trait data. To introduce some of the complexities of dealing with real genotypes into the study, we obtained the SNP genotypes from the publicly available 1000 Genome Project, phase 3 [8]. The quantitative trait data are generated by selecting, randomly, a set of SNP loci, assigning additive allelic effects to these loci, and then aggregating these effects for each individual along with a random error. The number of SNPs selected per replicate follows a Poisson distribution with mean 30. The sizes of the allelic effects across the selected loci are equal. A heritability of 50% is assumed for the trait.

Memory Usage and Run Times We analysed the simulated data with Eagle and the other computer programs, recording their memory usage and run (or elapse) times. The analyses were performed on a high-end desktop computer with dual 8-core Xeon processors and 128 gigabytes of RAM. Not all data generated under the five scenarios could be analysed by all implementations. Memory usage for many of the computer programs was the limiting factor (See **Supplementary Figure 1**). The single-locus program GEMMA was by far the most memory efficient. Not surprisingly, the multiple-locus programs were memory intensive. Most required in excess of the 128 gigabytes of available RAM for the analysis of data generated under 4000 x 1.5M and 10000 x 1.5M. Even FaST-LMM, a single-locus implementation, required more than 128 gigabytes of memory for the analysis of the larger data sets when all the marker data was used to calculate the relationship matrix. Of the multiple-locus programs, only Eagle, with its ability to handle data larger than the memory capacity of the computer, was capable of producing findings for data from our largest scenario 10000 x 1.5M.

The median run times for Eagle and the other computer programs across the six scenarios are shown in Figure 1. The x- and y-axes are on a log scale. This means a unit change on the x- or y-axis is equivalent to a change in the order of magnitude. In answer to our question of how does Eagle compare in terms of run time to competing implementations, Eagle was significantly faster, sometimes by orders of magnitude, than the other multiple-locus implementations and is comparable to the single-locus implementations. For a simulation study with 150 individuals and 5000 SNPs, Eagle produced results in seconds. For the larger simulation scenarios 1500 x 50K and 350 x4 00K, analyses with Eagle took under two minutes. Even for data from a couple of thousand individuals and half a million SNPs (2000 x 500K), the median run time of Eagle was under 14 minutes. For our scenarios where there were thousands of individuals and 1.5 million SNPs, Eagle took just over two hours for the analysis of data from 4000 x 1.5M and just under 12 hours for the analysis of data from 10000 x 1.5M. Towards the final stages of writing this paper, we gained access to a desktop computer with 14-core Xeon processors with 256 gigabytes of RAM. We reran Eagle on data from the largest scenario 10000 x 1.5M to measure the impact on run time. The median run time dropped by more than 70% from just under 12 hours to 3.31 hours.

It is also worth noting that an increase in the number of SNP genotypes in a study does not necessitate, automatically, an increase in the memory usage and run times for its analysis. This is evidenced by the non-monotonically increasing behaviour of the curves in the memory (Supplementary Figure 1) and run-time plots (Figure 1). It is the study dimensions, not number of genotypes, that drives the computing resources required by association mapping programs.

Figure 1 goes around here

Power and False Discovery Rates We calculated, empirically, the statistical power and false discovery rates of Eagle and the other methods across the six scenarios. We were interested in answering the question of how well Eagle finds true SNP-trait associations and avoids false SNP-trait associations. For each replicate, we knew which SNPs had been used in creating the quantitative trait data. These SNPs are in true association with the trait. It is the goal of the single- and multiple-locus association mapping methods to discover these SNPs. By knowing which SNPs are in true association with the trait, we were able to assess the validity of a method's findings. When a replicate was analysed, a method's findings were counted as true if the SNPs were located within 40 kilobase pairs of SNP in true association with the trait. To calculate the power of a method, for each replicate, we divided the number of true SNP findings by the number of SNP that had been used in creating the trait data. We then averaged across the 100 replicates. Similarly, to calculate the false discovery rate of a method, for each replicate, we divided the number of true SNP findings by the number of (true and false) SNP findings. We then averaged across the 100 replicates.

The power and false discovery rate of Eagle and the other multiple-locus methods across the scenarios 150 x 5K, 350 x 500K, 1500 x 50K, and 2000 x 500K are shown in Figure 2. We restricted our attention to these scenarios for the multiple-locus methods because for scenario 4000 x 1.5M, the data could only be analysed by Eagle, MLMM, and r2VIM. For scenario 10000 x 1.5M, the data could only be analysed by Eagle. The power and false discovery rate of Eagle and the two single-locus methods, GEMMA and FaST-LMM, are shown in Supplementary Figure 2. Each plot contains single points and power curves. The single points are the power and false discovery rates for Eagle and MLMM. These two methods treat association mapping as a model selection problem. Their are no significance thresholds to be set. The power curves are for those methods that treat association mapping as a parameter estimation problem. Here, the significance of the findings are assessed against a significance threshold. The power curves in the plot show how power changes with the false discovery rate as the significance threshold is adjusted.

In answer to the question of how well Eagle finds true SNP-trait associations and avoids false SNP-trait associations, it does extremely well. Of the multiple-locus methods, Eagle has the highest power while keeping its false discovery rate low (Figure 2). MLMM also performed well. However, it is when Eagle is compared against single-locus methods that the difference in power is most noticeable. Eagle has much greater power than single-locus methods for finding SNP in true association with a trait while avoiding false associations (Supplementary Figure 2).

Figure 2 goes here

Mouse Data Analysis We were interested in comparing results from Eagle with those from single-locus association mapping for a real data set. We chose to focus on data from a large outbred mouse study [9]. This study was unusual in that it collected and analysed SNP dosages (continuous values from zero to one of expected allele counts) instead of the more common SNP genotypes. Analyses based on dosages rather than discrete genotypes have been shown to have greater power for the detection of genes that are influencing a trait [10]. By converting the dosages into genotypes and analysing the data with the single-locus program FaST-LMM, we obtained a subset of those findings reported in the original study. We then analysed the data with Eagle. Due to Eagles increased power, we found SNP-trait associations not found with the FaST-LMM. However, we were able to confirm the validity of these new findings as they matched what was found in the original study. Having the ability to confirm new findings in a real study was one of the primary motivators for choosing these data for analysis. For the single-locus analysis of the data, we followed the same procedure as originally followed for the analysis of the mouse data [9]. The only differences were that we focused on the autosomal SNP and it was necessary to increase the number of permutations for the controlling of the false discovery rate from 100 to 500.

Eagle was run in two ways; under its default settings (Eagle^{default}) and where we specified the regularisation parameter for model selection (Eagle^{optimal}) Eagle chooses the best model via the extended Bayesian information criteria (extBIC) [11]. The conservativeness of the extBIC can be adjusted by a single regularisation parameter that ranges from zero to one. In the simulation study, this parameter was set to one, its most conservative and default setting. However, there is also opportunity to set the parameter to a value less than one. This increases power but also increases the false discovery rate. For each trait, we used permutation to set the regularisation parameter to give a false discovery rate of 5% .

The genome wide results from the analyses of the mouse data are shown in Figure 3. The mouse study took measurements on 200 traits. When these traits were first analysed in the original study, findings for 45 of these traits were able to be corroborated by prior published evidence. We focused our analyses here on these same 45 traits. For 39 traits, we found SNP-trait associations. For the other six, neither FaST-LMM nor Eagle found any associations. Each plot contains the number of SNP-trait associations that were found and in agreement with the original findings. Neither method found SNP not identified in the original mouse study so neither method found false positives. As we saw in the simulation study, there is a notable difference in the two methods capacity to discover SNP-trait associations. Eagle^{default}, under its default settings, for eight traits found the same number of findings as FaST-LMM and for 28 traits found more findings. Eagle^{optimal}, with its regularisation parameter fine tuned to the trait, for six traits found the same number of findings as FaST-LMM and for 32 traits found more findings. Overall, FaST-LMM, Eagle^{default}, and Eagle^{optimal} found 26, 65, and 95, snp-trait findings respectively. Eagle^{default} and Eagle^{optimal} found two-and-a-half times and over three-and-a-half times, respectively, more SNP-trait associations than what is the established way of analysing these data. Furthermore, these are all findings that were confirmed in the original study.

2 Methods

Outbred Mice Study

Mouse Data The data were obtained from a large genome-wide association study which was performed in outbred mice. The study is described in detail in [9]. Phenotypic and genotypic data were available on 1,887 adult mice. The phenotypic data consisted of measurements from 200 behavioural, tissue, and physiological traits. Of these traits, 43 yielded SNP-trait associations that could be corroborated through other independent published work. It was these 43 traits that were the focus of our real data analyses. Genotypic data were available on 359, 559 (353,697 autosomal) SNPs in the form of marker dosages (expected allele counts that ranged from zero to one). All missing data had been imputed. We converted the dosages into discrete genotypes by clustering around 0, 0.5, and 1, corresponding to SNP genotypes AA, AB, and BB, respectively.

Still working on the rest of the methods ...

1. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821–824 (2012).

2. Lippert, C. *et al.* Fast linear mixed models for genome-wide association studies. *Nature methods* **8**, 833–835 (2011).

3. Shen, X., Alam, M., Fikse, F. & Rönnegård, L. A novel generalized ridge regression method for quantitative genetics. *Genetics* **193**, 1255–1268 (2013).

4. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22 (2010). URL `http://www.jstatsoft.org/v33/i01/`.

5. Rakitsch, B., Lippert, C., Stegle, O. & Borgwardt, K. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics* **29**, 206–214 (2013).

6. Segura, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics* **44**, 825–830 (2012).

7. Szymczak, S. *et al.* r2vim: A new variable selection method for random forests in genome-wide association studies. *BioData mining* **9**, 7 (2016).

8. Consortium, . G. P. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (2010).

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

10. Zheng, J., Li, Y., Abecasis, G. R. & Scheet, P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genetic epidemiology* **35**, 102–110 (2011).

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

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).