*Bioinformatics*
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original Paper

Genetics and Population Analysis

# Eagle: multi-locus association mapping on a genome-wide scale made routine

**Andrew W. George [1,*], Arunas Verbyla [2] and Joshua Bowden [3,]**

[1]Data61, CSIRO, Brisbane, 4102, Australia and
[2]Data61, CSIRO, Atherton, 4883, Australia and
[3]IM&T, CSIRO, Brisbane, 4067, Australia.

*To whom correspondence should be addressed

Associate Editor: Alfonso Valencia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

## Abstract

**Motivation:** We present Eagle, a new method for multi-locus association mapping. The motivation for developing Eagle was to make multi-locus association mapping "easy" and the method-of-choice. Eagle's strengths are that it a. is considerably more powerful than single-locus association mapping b. does not suffer from multiple testing issues c. gives results that are immediately interpretable and d. has a computational footprint comparable to single-locus association mapping.
**Results:** By conducting a large simulation study, we will show that Eagle finds true and avoids false SNP-trait associations better than competing single- and multi-locus methods. We also analyse data from a published mouse study. Eagle found over 50% more validated findings than the state-of-the-art single-locus method.
**Availability and Implementation:** Eagle has been implemented as an R package, with a browser-based Graphical User Interface (GUI) for users less familiar with R. It is freely available via the CRAN website at https://cran.r-project.org. Videos, Quick Start guides, FAQs, and Demos are available via the Eagle website http://eagle.r-forge.r-project.org
**Contact:**
**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Over the past decade, genome-wide association studies (GWASs) have changed considerably in both their analysis and design. Early studies 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 (Cardon and Palmer, 2003). This prompted a shift towards family-based designs and score tests, such as the transmission/disequilibrium test (TDT) and its variants (Spielman and Ewens, 1996). Today, instead of by design, it is through statistical modelling that we account for the effects of population stratification (Price *et al.*, 2010). 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 (Yu *et al.*, 2006; Zhao *et al.*, 2007).

What has not changed is that it remains common practice to analyse genome-wide association study (GWAS) data on a locus-by-locus basis. This is despite there being several significant problems with analysing data in this way. First, for each SNP, a hypothesis test is performed. The null hypothesis is that there is no association between the SNP and trait. The alternative is that the SNP is in association with the trait. It is straight forward to guard against wrongly rejecting the null hypothesis (or making a type 1 error) if only a single hypothesis test is being performed. However, the analysis of GWAS data with locus-by-locus methods necessitates conducting a large number of correlated hypothesis tests, simultaneously. This leads to an increased risk of type 1 errors. To deal with this challenge, many different solutions have been offered (Storey and Tibshirani, 2003; Li and Ji, 2005; de Bakker *et al.*, 2005). Second, 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 -log$_{10}$ of the $p$ value for each locus against the map position of the locus. The $p$ value is obtained from the hypothesis test. The location of peaks in this plot identify genomic regions of interest. Inferring the exact number of regions though can be difficult if the peaks are not well separated. Third, many of the traits whose genetic secrets we are trying to discover are complex. There will be multiple SNPs in linkage disequilibrium with genes that are influencing 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 multi-locus association mapping methods haven't attracted more attention. Methods based on regularisation techniques, such as ridge regression (Shen *et al.*, 2013) and lasso (Rakitsch *et al.*, 2013), measure all locus-trait associations simultaneously. These techniques though are computationally demanding. Also, the strength of association is not measured by a $p$ value but by the size of the regression coefficient for the SNP in the model. Further processing is required before the results can be interpreted (Cho *et al.*, 2010; Rakitsch *et al.*, 2013). More recently, associations have started to be mapped with random forests (Szymczak *et al.*, 2016). Similar to regularisation techniques though, it is not clear how to infer genomic regions of interest from their findings. A multi-locus method that does show promise is the multiple-locus linear mixed model method (Segura *et al.*, 2012). The best multi-locus model is built with forward and backward stepwise selection. Results are immediately interpretable in that the SNP closest to the genes underlying the trait are identified but computation does become challenging for large datasets.

In this paper, we present our new multi-locus method for genome-wide association mapping, which we are calling Eagle. Eagle combines the strength of regularisation techniques (being able to fit all SNP-trait associations jointly), with forward selection giving easy-to-interpret threshold-free results. We are able to achieve a computational performance similar to the fastest single-locus linear mixed model implementations through a dimension reduction step. Our aim was to make multi-locus association mapping on a genome-wide scale routine. To this end, we have implemented Eagle within an R package of the same name. Our package accepts marker data of different formats, can handle data larger than a computer's memory capacity, and makes heavy use of parallel computing for computation when available.

## 2 Methods

### 2.1 Mouse Data

The data were obtained from a large genome-wide association study that was performed in outbred mice (Nicod *et al.*, 2016). Phenotypic and genotypic data were available on 1,887 adult mice. The phenotypic data included raw and adjusted (for fixed effects) measurements from 200 behavioural, tissue, and physiological traits. Of these traits, 45 yielded SNP-trait associations that could be corroborated through other independent published work. It was these 45 traits that were the focus of our real data analyses. As in the original study (Nicod *et al.*, 2016), our analyses were based on the adjusted traits. 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. We focused our analyses on the autosomal SNPs.

### 2.2 Eagle Approach for Multi-locus Association Mapping

Eagle is a method for multi-locus association mapping on a genome-wide scale. It is based on linear mixed models. It differs from most other single- and multi-locus association mapping methods in that Eagle treats association mapping as a model selection problem (Ball, 2001; Broman and Speed, 2002; Yi *et al.*, 2005). The "best" model is found via forward selection. It makes use of a modified form of the Bayesian information criterion, BIC, for model selection. A "best" model is built iteratively. At each iteration, a hypothesis test is performed. Only a small number of iterations are needed in building the "best" model. Consequently, Eagle does not suffer from multiple testing issues. In contrast, for single-locus methods, multiple testing is an issue because each SNP is assessed separately, culminating in the need for a large number of hypothesis tests to be performed. Eagle reports as its findings only those SNPs that are in strongest linkage disequilibrium with the genes influencing a trait. The methodological foundation for Eagle comes from a whole-genome linkage analysis method that was developed for mapping quantitative trait loci in experimental crosses (Verbyla *et al.*, 2007).

Let $S = \{S_1, S_2, \ldots, S_s\}$ be a set of $s$ ordinal numbers where $S_k$ is the $S_k$th ordered SNP that was selected in the $k$th iteration of the model building process. Suppose three iterations ($s = 3$) have been performed and say the 500023rd, 15th, and 420th SNP were selected. Then $S = \{500023, 15, 420\}$. Let $\boldsymbol{y}^{(n \times 1)}$ be a vector containing $n$ measurements of the quantitative trait. Let $\boldsymbol{M}^{(n_g \times L)} = [\boldsymbol{m}_1 \boldsymbol{m}_2 \ldots \boldsymbol{m}_L]$ be a matrix containing the genotype data which have been collected from $L$ loci that span the genome on $n_g$ groups/lines/strains. Here, $n \geq n_g$ meaning that a single or several trait measurements may be taken of the same group/line/strain. It is common for the columns of $\boldsymbol{M}$ to be in map order but this is not a requirement. The vector $\boldsymbol{m}_j^{(n_g \times 1)}$ contains the genotypes for the $j$th SNP. The genotypes are coded as -1, 0, and 1 corresponding to SNP genotypes AA, AB, and BB, respectively.

The specifics of the Eagle method are as follows. Eagle builds the "best" model iteratively, via forward selection. 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, we first fit the current model to the data. The (current) model is of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u}_g + \boldsymbol{e} \qquad (1)$$

where $\boldsymbol{X}^{(n \times p)}$ and $\boldsymbol{Z}^{(n \times n_g)}$ are known design matrices with $\boldsymbol{X}$ being of full rank and $\boldsymbol{Z}$ containing zeros and ones that assign the appropriate genetic effect to each measurement. The vector $\boldsymbol{\tau}^{(p \times 1)}$ has $p$ fixed effects parameters including the intercept. The vector $\boldsymbol{u}_g^{(n_g \times 1)}$ contains the genetic effects. The vector of residuals is $\boldsymbol{e}^{(n \times 1)}$ whose distribution is assumed to follow $N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}^{(n \times n)})$. So far, this model differs little from standard linear mixed models for association mapping (Yu *et al.*, 2006; Zhao *et al.*, 2007) However, it is how we specify $\boldsymbol{u}_g$ that distinguishes our model from the others.

The genetic effects $\boldsymbol{u}_g$ are modelled as

$$\boldsymbol{u}_g = \sum_{k=1}^{s} \boldsymbol{m}_{S_k} a_{S_k} + \boldsymbol{M}_{-S}\boldsymbol{a}_{-S} \qquad (2)$$

where $\boldsymbol{m}_{S_k}^{(n_g \times 1)}$ is the vector of genotypes for the $k$th selected SNP, $a_{S_k}$ is the additive effect of the $k$th selected SNP, $\boldsymbol{M}_{-S}^{(b \times L-s)}$ is the matrix of SNP genotypes with the data for the SNP in $S$ removed, and $\boldsymbol{a}_{-S}^{(L-s \times 1)}$ is a random effect whose distribution is $\boldsymbol{a}_{-S} \sim N(\boldsymbol{0}, \sigma_a^2 \boldsymbol{I}^{(L-s \times L-s)})$. The terms in the summation on the left hand side are fixed effects. They account for the additive effects of those SNPs that have been found to be in association with the trait. The other term is a random effect. It accounts for the joint effect of the yet-to-be-identified SNP that are in association with the trait. This is a simple genetic model but it is effective for discovering SNP-trait associations.

Second, we estimate the parameters of (1) and (2) via restricted maximum likelihood (REML). For complex models, REML can be computationally demanding. However, our model only contains a single random effect ($\boldsymbol{a}_{-S}$). Here, highly efficient single-dimension optimisation via spectral decomposition is possible (Kang *et al.*, 2008).

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 justifiable, theoretically (Verbyla *et al.*, 2012).

Fourth, we determine the importance of the $(s+1)$th selected SNP via a model selection strategy (Verbyla *et al.*, 2007). 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) (Chen and Chen, 2008). 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 well suited to the model selection problem in genome-wide association studies (Chen and Chen, 2008). It is different to the model selection measure used in (Verbyla *et al.*, 2007). 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 $s$ SNPs previously identified.

#### 2.2.1 Reducing the dimension of the model:

In practice, estimating the parameters of (2) can be demanding, computationally. The vector $\boldsymbol{a}_{-S}$ has $L-s$ random effects where in modern genome-wide association studies, $L$, the number of SNPs, can be extremely large. An alternative model is given by Verbyla (Verbyla *et al.*, 2012, 2014). They show how to reformulate (2) to be a model with a random effect with only $n$ elements

$$\boldsymbol{u}_g = \sum_{k=1}^{s} \boldsymbol{m}_{S_k} a_{S_k} + (\boldsymbol{M}_{-S}\boldsymbol{M}_{-S}^T)^{1/2}\boldsymbol{a}^*_{-S} \tag{3}$$

where $\boldsymbol{a}^* \sim N(\boldsymbol{0}, \sigma_a^2 \boldsymbol{I}^{(n_g \times n_g)})$, and $(\boldsymbol{M}_{-S}\boldsymbol{M}_{-S}^T)^{1/2}$ can be calculated via singular value decomposition (Golub and Van Loan, 2012). Although it may not be obvious, the two models are equivalent, having identical variance structures. Yet, the computational cost of model (3) compared to model (2) is much less, due to the random term in model (3) having only $n$ instead of $L-s$ effects needing estimating.

Verbyla (Verbyla *et al.*, 2012, 2014) go on to show how to recover $\widetilde{\boldsymbol{a}}$ from estimates from model (3) with

$$\widetilde{\boldsymbol{a}} = \left[\boldsymbol{M}_{-S}^T(\boldsymbol{M}_{-S}\boldsymbol{M}_{-S}^T)^{-1/2}\right]\widetilde{\boldsymbol{a}}^* \tag{4}$$

where its variance matrix is

$$\text{var}(\widetilde{\boldsymbol{a}}) = \boldsymbol{M}_{-S}^T(\boldsymbol{M}_{-S}\boldsymbol{M}_{-S}^T)^{-1/2}\text{var}(\widetilde{\boldsymbol{a}}^*)(\boldsymbol{M}_{-S}\boldsymbol{M}_{-S}^T)^{-1/2}\boldsymbol{M}_{-S} \tag{5}$$

These values are needed in order 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 (5).

### 2.3 Comparison Methods

#### 2.3.1 Multi-locus methods:

We compare the computational and statistical performance of Eagle against five multi-locus methods. They are bigRR (Shen *et al.*, 2013), LMM-Lasso (Rakitsch *et al.*, 2013), glmnet (Friedman *et al.*, 2010), MLMM (Segura *et al.*, 2012), and r2VIM (Szymczak *et al.*, 2016). All but glmnet have been purposely designed for genome-wide association mapping. BigRR, LMM-Lasso, and glmnet are regression-based regularisation methods. BigRR is based on generalised ridge regression, LMM-Lasso is based on lasso, and glmnet is based on elastic net. Regularisation methods make parameter estimation possible in models where the number of predictors is far greater than the number of samples. They allow the strength of association between all the SNPs and trait to be measured within a single model, simultaneously. A limitation of these methods though is that the statistical significance of the SNP effects cannot be easily determined. Due to the adaptive nature of the estimation procedures, to do this analytically is challenging and is an area of active research (Lockhart *et al.*, 2014). Instead, we calculate significance empirically via stability selection (see below).

MLMM is closest in philosophy to Eagle. It too is based on building the best model via stepwise selection, within a linear mixed model framework, and uses the extBIC as one of its model selection criterion. However, there are differences between the two methods. MLMM does not make use of dimension reduction. Also, how SNP are selected to enter the model differs between the two methods. Eagle identifies a SNP of interest from its score statistic (see Section 2.2 for details). This score statistic was originally developed for outlier detection in linear (mixed) models but it is being used by Eagle to identify SNP with unusually large random effects. MLMM instead uses the statistical significance of a SNP, when treated as a fixed effect in the model. This involves fitting a separate linear mixed model for each candidate SNP, a potentially computationally expensive exercise. However, MLMM does this in a clever and efficient way via the Gram-Schmidt process. Both are R packages but there is a significant difference in computational performance (see Results). Note, even though a hypothesis test is being performed for each SNP by MLMM, it does not suffer from multiple testing issues. Neither the null nor the alternative hypothesis is being accepted or rejected. Only the hypothesis yielding the most significant association is of interest.

R2VIM differs to the other four methods in that it is a non-parametric model-free approach. It implements random forests but where multiple parallel runs are performed. Each run leads to different random forests being created. A relative importance score is calculated, within a run, for each SNP. This is done by dividing a SNP's importance score by the minimum importance score observed across all the SNPs within a run. Only those SNPs with relative importance scores above a certain threshold across all the runs are deemed to be significant. Unfortunately, the relationship between threshold value and false positive rate is unknown. The threshold could be found empirically via permutation but the computational cost is high, restricting the size of data that can be analysed.

#### 2.3.2 Single-locus methods:

We also compare the performance of Eagle against two single-locus methods, GEMMA (Zhou and Stephens, 2012) and FaST-LMM (Lippert *et al.*, 2011). Both are based on linear mixed models. The models have a single fixed effect for the SNP, other fixed effects, a single random effect to account for familial relatedness (or polygenic background), and an error. The significance of the SNP effect in the model is a measure of the strength of association. They are of the same computational complexity (Zhou and Stephens, 2012), and produce exact results. Both perform a single spectral decomposition of the relationship (or similarity) matrix $K$, use an eigenvector matrix to rotate the data, and reformulate the (residual) log likelihood for easier computation. They do differ in their estimation procedure. GEMMA implements Newton-Raphson. FaST-LMM implements Brent's algorithm. Newton-Raphson is more complicated but has better convergence properties than Brent's algorithm. Both methods are state-of-the-art and have been implemented in highly efficient computer programs.

## 2.4 Generation of Simulation Data

The data are generated via data perturbation (Zhao *et al.*, 2007). Data perturbation amalgamates real with simulated data to generate replicates. It is a way of introducing greater realism into a simulation study. Here, the genotype data are real, the quantitative trait data are simulated. The SNP genotypes are drawn, according to the specifications of a particular simulation scenario, from data collected from the 1000 Genome Project, version 3 (Consortium *et al.*, 2010). Six different scenarios are considered. These scenarios differ in their sample size and number of SNPs (see Results for details). Here, across scenarios, the SNP data differs. Across replicates within a scenario, the SNP data are the same. For each scenario, 100 replicates are generated.

To generate the trait data $\boldsymbol{y}$, first, $q$, the number of SNPs that are to be assigned a quantitative value is drawn from a Poisson distribution with mean 30. Second, $q$ SNP are selected randomly. Third, we assume an additive model for the SNPs. The SNP genotypes AA, AB, and BB are assigned the values -1, 0, and 1, respectively. Fourth, the SNP effects are summed across the $q$ selected loci, for each individual, to generate a $\boldsymbol{g}^{(n\times1)}$ vector of genetic values where $n$ is the number of individuals. Fifth, $\boldsymbol{e}^{(n\times1)}$, a vector of residuals, is drawn from a normal distribution where $e_i \sim N(0, \sigma_e^2)$ and $\sigma_e^2$ is the residual variance that has been set to yield a trait with heritability 0.5. Sixth, the trait data are formed as $\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{e}$. In forming $\boldsymbol{y}$, we have purposely not included any other environmental variables such as age, sex, or experimental design effects. This is because not all the methods were implemented to handle the inclusion of additional fixed effects. A two-stage modelling approach is often adopted to deal with this situation, but we chose not to introduce this complexity into the analyses.

## 2.5 Stability Selection

Stability selection (Meinshausen and Bühlmann, 2010) is a subsampling strategy with a range of applications. It is used here to estimate, empirically, the statistical significance of the results from LMM-Lasso, glment, and bigRR analyses of the simulated data. These three regularisation methods give the effect sizes of the SNPs, but not their significance as their results. Stability selection was chosen over permutation and other sampling procedures because of its reduced computational cost.

The stability selection procedure for LMM-Lasso and glmnet is as follows. For a particular scenario, we begin by finding, via a binary search, the value of the regularisation parameter that yields 20 to 30 non-zero SNP effects. We know that 20 to 30 SNP-trait associations is a reasonable number of findings to expect from the analysis of a replicate in the simulation study. The regularisation parameter though could have been tuned to give any reasonable number of non-zero SNP effects. This tuning was done for each of the six scenarios but only for a single replicate, selected at random, from within a scenario. It is not necessary to tune the regularisation parameter on every replicate when replicates are generated under the same (trait, sample size, and number of SNP) conditions within a scenario.

Once a suitable value for the regularisation parameter had been found, for the replicate whose SNP results are to be assigned statistical significance, we subsample repeatedly, 100 data sets of size $n/2$. A larger number of data sets and/or larger sized data sets could have been chosen here but we found these changes to have little impact on the final significance estimates. The subsamples are drawn without replacement. Also, the matching of trait to genotype is preserved in the subsamples. A subsample differs to the replicate in size only. The subsamples are analysed with LMM-Lasso (glmnet) with its regularisation parameter fixed to the tuned value found previously. From the analysis of a subsample, a binary vector, of length the number of SNP, is recorded as the result where a one (zero) means the SNP had a non-zero (zero) effect size. Calculating a SNP's statistical significance is now a simple task. We calculate the vector sum of the binary vectors over all 100 subsamples. This vector sum will contain elements in the range of 0 to 100. By dividing each element in this vector sum by the number of subsamples upon which the sum is calculated (which is 100), we obtain empirical probabilities. These probabilities measure the strength of evidence for the SNPs to be in association with the trait.

For bigRR, stability selection is implemented in a different way. Unlike LMM-Lasso and glmnet, bigRR yields non-zero SNP effects for all the SNPs. Also, there is no need to tune the regularisation parameter for bigRR as an optimal value is found as part of its analysis procedure. We still draw 100 subsamples of size $n/2$, without replacement, and each subsample is analysed with bigRR. However, from each analysis, we order the SNPs according to the absolute size of their SNP effect estimates from bigRR. A binary vector, of length the number of SNPs is then formed where a one (zero) means the SNP is (not) in the top 20 ordered SNPs. Calculating the significance of the SNPs then proceeds as described above.

## 2.6 Implementation

Eagle has been implemented as an R package of the same name. Much of the computation though is performed outside of R via C++ functions that utilise Eigen C++ routines. Eagle has been purpose built to rely heavily on calls to BLAS and LAPACK, mathematical libraries common to most computer systems. By making use of multi-threaded BLAS and LAPACK libraries, many of the calculations in Eagle are parallelised. We have gone to great lengths to make Eagle easy-to-use. Tutorials, videos, How-To guides, and a link to our server for demonstrating Eagle on some test data are available on the Eagle website http://eagle.r-forge.r-project.org. Eagle is available for download from the CRAN website.

# 3 Results

## 3.1 Association Mapping Methods

We compared Eagle, in terms of computational and statistical performance, against seven other association mapping methods. We chose methods that almost all 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 or packages. Two of the methods are based on single-locus (or locus-by-locus) models and five are based on multi-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 multi-locus association mapping, we chose bigRR, glmnet, LMM-Lasso, MLMM, and r2VIM. Each takes a different approach to multi-locus association mapping. A summary of the key attributes of the different computer programs/packages is given in Supplementary Table 1 (see Methods for further details).

## 3.2 Simulation Study

A large simulation study was performed where we sought to answer two questions. First, how well does Eagle find true associations (power) and avoid false associations (type 1 errors)? Second, how does Eagle compare, in terms of run time and memory usage, to competing implementations? Data were generated under six different scenarios; a study of size 150 individuals and 5,000 single 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). These scenarios reflect, at least in some cases, the sizes of study being performed in animals, plants, and humans.

For each scenario, 100 replicates were generated. A single replicate consisted of SNP and quantitative trait data (see Section 2.4). Extra realism was introduced into the simulation study through the drawing of the SNP genotypes from the 1000 Genome Project, phase 3 (Consortium *et al.*, 2010). The quantitative trait data were generated by selecting, randomly, a set of SNPs and assigning these loci additive allelic effects. Random errors were then drawn from a normal distribution with variance set to give a heritability of 50% for the trait. For each individual, a quantitative trait value was obtained by summing its random error and additive allelic effects. The number of randomly selected SNPs follows a Poisson distribution with mean 30. The size of the allelic effects across the selected loci are equal, because the SNP genotypes AA, AB, and BB are assigned the values -1, 0, and 1, respectively (Section 2.4) .

Analyses by the eight programs/packages of a replicate proceeded as follows. They were all run at their default settings. Eagle and MLMM were the easiest of the programs/packages to implement. The only parameters requiring specification were the amount of available memory and number of CPUs for Eagle and the number of chunks for MLMM. MLMM breaks its matrices into blocks or chunks, reducing its memory footprint but at the cost of increased computation. Their results were also immediately interpretable. Their findings were the set of SNPs in strongest association with the trait. Each SNP in this set identified a separate genomic region of interest, whose position was given by the map location of the SNP.

BigRR, LMM-Lasso, and glmnet required more effort to implement. They are based on regularisation methods and as such, all the SNPs were fitted simultaneously in a regression framework. The difficulty was in calculating the significance of the SNP effects. To do this analytically is challenging. We instead opted for stability selection (see Methods), an empirical approach for calculating significance.

R2VIM is different from the rest in that it is a nonparametric approach for association mapping. It is based on random forests. Three important parameters needed to be set. These were the number of trees, the number of variables for building a tree, and the minimum size of a terminal node. Ideally, these parameters would be "tuned" on a replicate-by-replicate basis (Boulesteix *et al.*, 2012). However, this was not practical here. We instead used the same settings as in (Szymczak *et al.*, 2016) where the number of trees was set to 1000, the number of variables was set to 20% of the number of SNPs, and the minimum size of a node was set to 10% of the sample size. A relative importance measure was calculated for each SNP measuring its strength of association with the trait.

FaST-LMM and GEMMA implement single-locus association mapping. FaST-LMM was run in two ways. One way was where a subset of the SNP data were used in calculating the similarity (or relationship) matrix. Here, FaST-LMM is highly efficient, computationally. The other was where calculation of the similarity matrix was based on all the SNP data. The $p$ values of the SNP were reported as their results.

The results from all but Eagle and MLMM required post-processing before the findings were interpretable. The SNPs were placed in map order, a significance threshold was set, peak regions containing SNPs with significance measures above the threshold were identified, and the SNP with the largest significance measure in each of the peak regions was recorded as a finding.

### 3.3 Power and False Positive Rates

Here, we answer the question of how well Eagle finds true SNP-trait associations and avoids false SNP-trait associations. We do this by estimating the power and false positve rates of Eagle and the other methods for the six scenarios. Since, for a replicate, we knew which SNPs were assigned additive effects, we knew the SNPs that were in true association with the trait. We will refer to these SNPs as being true SNPs. By knowing the true SNPS, we were able to assess the validity of a method's findings. A finding was counted as true if it was positioned within 40 kilobase pairs of the location of a true SNP. This was the average (confidence interval) distance used by Nicod *et al.* (2016) for determining if their findings were close to candidate genes and whose mouse data we analyse below.

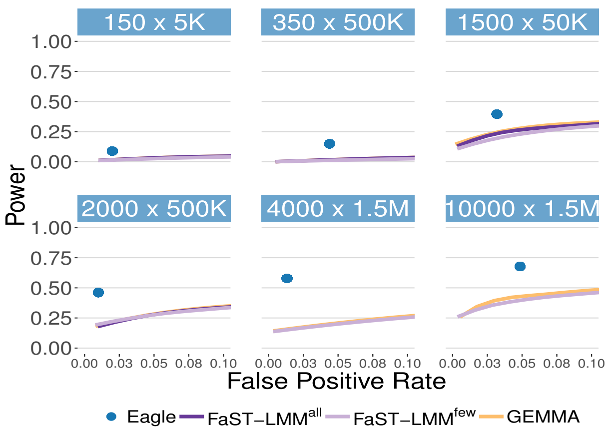



**Fig. 1.** Power verse false positive rates for Eagle and the single-locus methods GEMMA and FaST-LMM. FaST-LMM was run where all the SNP data are used to estimate the relationship matrix (FaST-LMM$^{all}$) and where genotype data from every five-hundredth SNP are used to estimate the relationship matrix (FaST-LMM$^{few}$). Eagle has substantially higher power than the single-locus methods.

When a replicate was analysed, we obtained an estimate of the power of the method by taking the number of findings that were found to be true and dividing by the number of true SNPs. We also obtained an estimate of a method's false positive rate. It is the number of findings that were found to be false divided by the number of true SNPs. Both these estimates varied with replicate. The power (false positive rate) of a method, for a scenario, was found by taking the median of the power (false positive rate) estimates over the 100 replicates.

The power and false positive rates of Eagle and the other multi-locus methods across the scenarios 150 x 5K, 350 x 500K, 1500 x 50K, and 2000 x 500K are shown in Supplementary Figure 1. We restricted our attention to these scenarios because not all multi-locus methods could cope with the size of data in the other scenarios. Each plot contains single points and power curves. The single points are the power and false positive 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 variable selection 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 positive rate as the significance threshold is adjusted. The power and false positive rate of Eagle and the two single-locus methods, GEMMA and FaST-LMM, are shown in Figure 1.

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 multi-locus methods, Eagle had the highest power while keeping its false positive rate low (Supplementary Figure 1). MLMM also performed well. However, it was when Eagle was compared against single-locus methods that the difference in power was most noticeable. Eagle had much higher power than single-locus methods for finding SNP in true association with a trait while avoiding false associations (Figure 1).

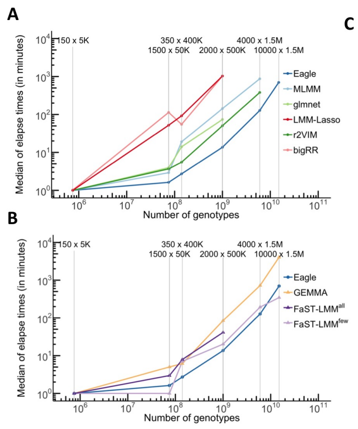
## 3.4 Memory Usage and Run Times

Memory usage and run (or elapse) times were recorded for Eagle and the other computer programs/packages across the simulation scenarios. 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 six scenarios could be analysed by all implementations. Memory usage for many of the computer programs/packages was the limiting factor (see Supplementary Figure 2). The single-locus program GEMMA was by far the most memory efficient. Not surprisingly, the multi-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, when all the SNP data were being used to calculate the similarity matrix, ran out of memory for the larger scenarios. Of the multi-locus programs/packages, 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/packages across the six scenarios are shown in Figure 2. The x- and y-axes are on a log scale. 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 multi-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 of 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 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 high-end sever with 14-core Xeon processors and 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 12 hours to 3.31 hours.

## 3.5 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 (Nicod *et al.*, 2016). 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 (Zheng *et al.*, 2011). 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 Eagle's increased power, we found SNP-trait associations not found with 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.




C

| Method | Name | 150x5K | 1500x50K | 350x400K | 2000x500K | 4000x1.5M | 10000x1.5M |
|---|---|---|---|---|---|---|---|
| Multiple | Eagle | 0.08 | 1.62 | 2.71 | 13.65 | 127.63 | 699.55 |
| | MLMM | 0.15 | 2.91 | 19.04 | 143.01 | 870.84 | |
| | glmnet | 0.11 | 3.95 | 14.06 | 74.03 | | |
| | r2VIM | 0.09 | 3.66 | 5.51 | 50.59 | 380.52 | |
| | bigRR | 1.01 | 113.35 | 54.99 | 1030.61 | | |
| | LMM-Lasso | 0.57 | 52.08 | 92.20 | 1031.85 | | |
| Single | GEMMA | 0.02 | 5.02 | 6.17 | 84.83 | 723.33 | 4071.60 |
| | FaST-LMM$^{few}$ | 0.01 | 0.80 | 7.07 | 20.16 | 193.90 | 346.19 |
| | FaST-LMM$^{all}$ | 0.03 | 2.96 | 7.90 | 41.27 | | |

**Fig. 2.** Median run times, in minutes, for the analysis of simulation study data from the six scenarios. Eagle is compared against five other multi-locus programs/packages (A) and two single-locus programs (B). The x- and y-axes are on a log scale for improved aesthetics. Eagle has the lowest run-times of the multi-locus programs/packages, sometimes by orders of magnitude. Eagle can even produce results faster than single-locus programs. The median run times for the programs/packages across the scenarios are given in the table (C). The entries in a blue font correspond to the lowest run-time for a scenario. FaST-LMM$^{all}$ is where calculation of the similarity matrix is based on all the SNP data. FaST-LMM$^{few}$ is where calculation of the similarity matrix is based on a subset of the SNP data.

We repeated the single-locus analyses as first performed (Nicod *et al.*, 2016) but some exceptions. We focused on autosomal SNPs, our analyses were based on SNP genotypes rather than SNP dosages, we sought to control the false positive rate not false discovery rate of the methods, and significance thresholds were found empirically via permutation (Doerge and Churchill, 1996).

We ran Eagle in three ways. Eagle chooses the best model via the extended Bayesian information criteria (extBIC) (Chen and Chen, 2008). The conservativeness of the extBIC can be adjusted by a single regularisation parameter $\gamma$ that ranges from zero to one. In the simulation study, this parameter was set to one, its most conservative and default setting. The mouse data were also analysed under this setting (Eagle$^{default}$). An alternate (Chen and Chen, 2008), less conservative way of setting $\gamma$ is to let $\gamma = 1 - \frac{1}{(2\kappa)}$ with $\kappa = \frac{log(L)}{log(n_g)}$ where $L$ is the number of loci that span the genome, and $n_g$ is the number of individuals/groups/lines/strains in the study (Eagle$^{alt}$). However, our preferred way is to set the $\gamma$ parameter for each trait via permutation (Eagle$^{optimal}$). We used 100 permutations to set $\gamma$ to give a false positive rate of 5%. This only took six times as long as a single analysis of the data. This is because the marker data need only be read once, and only the trait data changes across permutations leading to other computational efficiencies. This permutation method has been implemented within the Eagle package.

The genome wide results from the analyses of the mouse data are shown in Figure 3, with the Manhattan plots of the single-locus analysis shown in Supplementary Figure 3. The mouse study recorded measurements on 200 traits. Of these, in the original study, 45 were able to have their findings corroborated by previously published work. We focused our analyses here on these same 45 traits. Overall, FaST-LMM$^{Bonf}$, FaST-LMM$^{perm}$, Eagle$^{default}$, Eagle$^{alt}$, and Eagle$^{optimal}$ found 47, 68, 37, 67, 106, SNP-trait findings, respectively, across 39 traits. No associations were found by FaST-LMM and Eagle for the other six traits. Eagle$^{alt}$ and Eagle$^{optimal}$ also found SNP-trait associations not found in the original study. This is despite their analyses being based on the SNP genotype data

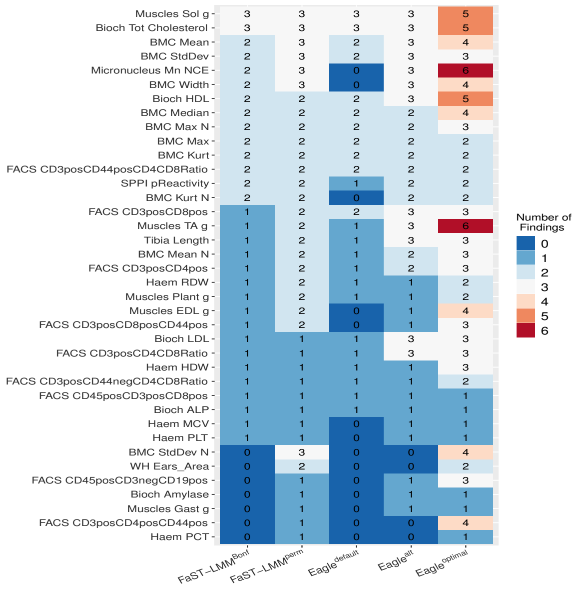



**Fig. 3.** Genome-wide association mapping results from analyses of the mouse data for the single-locus method FaST-LMM and the multi-locus method Eagle. Genome-wide significance thresholds for FaST-LMM were calculated, via the Bonferroni correction (FaST-LMM$^{Bonf}$) and permutation (FaST-LMM$^{perm}$), at the 5% significance level. Eagle was run under three settings; its default setting (Eagle$^{default}$), an alternate less conservative setting based on the number of SNPs and sample size (Eagle$^{alt}$), and where the model selection had been optimised for a false positive rate of 5% (Eagle$^{optimal}$). The number of SNP-trait associations found are reported in the cells.

and the original study being based on SNP dosage data. Eagle$^{alt}$ found two and Eagle$^{optimal}$ found seven new findings (Supplementary Table 2). These new findings all involved SNPs whose association had been confirmed for other related traits in the original study.

In the simulation study, Eagle outperforms single-locus association mapping. Here, Eagle$^{default}$, where $\gamma = 1$, finds less associations than FaST-LMM. Why the discrepancy in performance? The answer lies in the conservativeness of Eagle. With the added genetic complexity implicit within the mouse data, Eagle is more conservative when $\gamma$ is set to one than in the simulation study. However, the relative results of the simulation study remain true. For similar false positve rates, Eagle is superior to single-locus association mapping. As a case in point, here FaST-LMM$^{perm}$ found 68 SNP-trait associations with a false positive rate of 5%. Eagle, with the same false positive rate (Eagle$^{optimal}$) found 106 SNP-trait associations, more than a 50% increase in findings.

## 4 Discussion/Conclusion

Eagle is a new linear mixed model based method (and R package) for multi-locus association mapping. It advances the state of association mapping in several ways. First, its computational footprint is much smaller than other multi-locus implementations. Eagle makes multi-locus analysis practical, even when the datasets are large. Second, the results from Eagle are immediately interpretable. They are the set of SNPs in strongest association with the trait where each SNP identifies a separate genomic region of interest. Third, it treats association mapping as a model selection problem, avoiding multiple testing issues. As we saw in the simulation study, Eagle has considerably higher power than single-locus methods but is comparable in run time. Also, when analysing the mouse data, Eagle found more than 50% the SNP-trait associations than with single-locus association mapping, the method of choice. Furthermore, because we converted the SNP dosages of the original study into genotypes and focused our analyses on these data, the validity of the extra findings were able to be confirmed against the original findings. These extra findings were all found to be true.

Eagle outperformed the other multi-locus methods in our simulation study. However, we are cognisant of the fact that we made several implementation choices that impact our conclusions. For instance, we chose to calculate the significance of the SNP effects from bigRR, LMM-Lasso, and glmnet via stability selection. Permutation and its variants (Browning, 2008; Pahl and Schafer, 2010) are also equally valid empirical approaches. Stability selection though has the advantage of being based on repeated sampling of only a proportion (50% in our case) of the data. Also, when analysing the (sub)samples, it was not necessary to calculate the entire solution path for a method. Instead, analyses are performed for a fixed value of the regularisation parameter, greatly reducing the amount of computation required. For r2VIM, an R package implementing random forests, we had to decide on the minimum size of a terminal node, the number of trees, and number of potential variables. The setting of these parameters greatly affects performance. We acknowledge that in the hands of an expert, r2VIM could be fine-tuned for a better balance of computational and statistical performance. However, we would like to think that the parameter settings we used are sensible since they match the values in the original r2VIM publication (Szymczak *et al.*, 2016).

Eagle's computational speed does come at a cost. It is a weakness shared by all of the methods considered here, although in different ways. Eagle cannot handle extra random effects which are sometimes needed when more advanced study designs are employed. One solution is to adopt a two-stage analysis procedure. In the first stage, a single linear mixed model is fitted to the data. Much of the modelling complexity, including the extra random effects, is captured in this first-stage model. In the second stage, Eagle is run not on the original trait data but adjusted trait data which are obtained from the first stage analysis. Even though this is a well accepted practice, it is approximate (Gogel *et al.*, 2018). A better solution is to fit a single model to the data. Although not specifically designed for association mapping, WGAIM (Verbyla *et al.*, 2007), upon which Eagle is based, and RWGAIM (Verbyla *et al.*, 2012) are two R packages where this is possible. The difficulty is that for large datasets and/or complex models, run time and memory usage can become limiting factors for analysis.

Upon submitting our paper for review, a more recent multi-locus association mapping method, FarmCPU (Liu *et al.*, 2016), was brought to our attention. It is a statistically unorthodox approach. Instead of working with a single model, results are passed back and forth between two models, a fixed effects model and a random effects model. Measures of association are obtained from the fixed effect model, which in turn help define pseudo-QTN from the random effects model. Conversely, pseudo-QTN found from the random effects model are passed back to the fixed effect model to better refine the measures of association. The method involves multiple rounds of genome-wide testing. Out of interest, we reanalysed the mouse data with FarmCPU. Runtimes were around five times longer for FarmCPU than Eagle. Interestingly, FarmCPU found the same number of associations, 106, as Eagle$^{optimal}$, but the findings were only the same for 11 of the traits. For 14 traits, Eagle found more associations. For the other 14 traits, FarmCPU found more associations. Where the new findings from Eagle could be confirmed, of the 18 new associations found by FarmCPU, we were unable to confirm 10 using the results from the original study. Also, the two methods differ significantly in their implementation. FarmCPU is not an R package but a set of R scripts. We had no problem in using the scripts but this may not be true for non-R users. In contrast, Eagle has been developed for ease-of-use. Its browser-based GUI makes it accessible to all.