

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

Andrew W. George¹, Arunas Verbyla², and Joshua Bowden³

¹Data61, CSIRO, Brisbane, 4102, Australia.

²Data61, CSIRO, Atherton, 4883, Australia.

³IM &T, CSIRO, Brisbane, 4067, Australia

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 100% 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 web-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>.

Contact: andrew.george@csiro.au

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

57 the exact number of regions though can be difficult if the peaks are not well
58 separated. Third, many of the traits whose genetic secrets we are trying to
59 discover are complex. There will be multiple SNPs in linkage disequilibrium with
60 genes that are influencing the trait. Yet, a locus-by-locus mapping approach
61 only assesses the evidence for association between a single marker locus and
62 trait.

63 It is somewhat surprising then that multi-locus association mapping methods
64 haven't attracted more attention. Methods based on regularisation techniques,
65 such as ridge regression (Shen *et al.*, 2013) and lasso (Rakitsch *et al.*, 2013),
66 measure all locus-trait associations simultaneously. These techniques though
67 are computationally demanding. Also, the strength of association is not mea-
68 sured by a p value but by the size of the regression coefficient for the SNP in the
69 model. Further processing is required before the results can be interpreted (Cho
70 *et al.*, 2010; Rakitsch *et al.*, 2013). More recently, associations have started to be
71 mapped with random forests (Szymczak *et al.*, 2016). Similar to regularisation
72 techniques though, it is not clear how to infer genomic regions of interest from
73 their findings. A multi-locus method that does show promise is the multiple-
74 locus linear mixed model method (Segura *et al.*, 2012). The best multi-locus
75 model is built with forward and backward stepwise selection. Results are imme-
76 diately interpretable in that the SNP closest to the genes underlying the trait
77 are identified but computation does become challenging for large datasets.

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

89 2 Methods

90 2.1 Mouse Data

91 The data were obtained from a large genome-wide association study that was
92 performed in outbred mice (Nicod *et al.*, 2016). Phenotypic and genotypic
93 data were available on 1,887 adult mice. The phenotypic data included raw
94 and adjusted (for fixed effects) measurements from 200 behavioural, tissue, and
95 physiological traits. Of these traits, 43 yielded SNP-trait associations that could
96 be corroborated through other independent published work. It was these 43
97 traits that were the focus of our real data analyses. As in the original study
98 (Nicod *et al.*, 2016), our analyses were based on the adjusted traits. Genotypic
99 data were available on 359,559 (353,697 autosomal) SNPs in the form of marker
100 dosages (expected allele counts that ranged from zero to one). All missing
101 data had been imputed. We converted the dosages into discrete genotypes by
102 clustering around 0, 0.5, and 1, corresponding to SNP genotypes AA, AB, and
103 BB, respectively. We focused our analyses on the autosomal SNPs.

104 2.2 Eagle Approach for Multi-locus Association Mapping

105 Eagle is a method for multi-locus association mapping on a genome-wide scale.
106 It is based on linear mixed models. It differs from most other single- and multi-
107 locus association mapping methods. Eagle treats association mapping as a
108 model selection problem. Here, each candidate model postulates a different set
109 of SNPs in association with the trait. The "best" model is the model that
110 contains those SNPs closest to the genes that are underlying the trait.

111 ball2001bayesian,broman2002model,sillanpaa2002model,yi2005bayesian

112 Here, each candidate model contains a SNP or SNPs

113 Here, the "best" model has the set of SNP closest to the genes that are
114 influencing a trait. Here, the "best" model is the model that has identified the
115 set of SNP closest to the genes that are underlying the trait.

116 Consequently, we do not have to contend with having to construct signifi-
117 cance thresholds. Eagle also reports as its findings only those SNPs that are
118 in strongest linkage disequilibrium, and hence closest to the genes influencing

119 a trait. The methodological foundation for Eagle comes from a whole-genome
 120 linkage analysis method that was developed for mapping quantitative trait loci
 121 in experimental crosses (Verbyla *et al.*, 2007).

122 Let $S = \{S_1, S_2, \dots, S_s\}$ be a set of s ordinal numbers where S_k is the
 123 S_k th ordered SNP that was selected in the k th iteration of the model building
 124 process. Suppose three iterations ($s = 3$) have been performed and say the
 125 500023rd, 15th, and 420th SNP were selected. Then $S = \{500023, 15, 420\}$. Let
 126 $\mathbf{y}^{(n \times 1)}$ be a vector containing n measurements of the quantitative trait. Let
 127 $\mathbf{M}^{(n_g \times L)} = [\mathbf{m}_1 \mathbf{m}_2 \dots \mathbf{m}_L]$ be a matrix containing the genotype data which
 128 have been collected from L loci that span the genome on n_g groups/lines/strains.
 129 Here, $n \geq n_g$ meaning that a single or several trait measurements may be taken
 130 of the same group/line/strain. It is common for the columns of \mathbf{M} to be in map
 131 order but this is not a requirement. The vector $\mathbf{m}_j^{(n_g \times 1)}$ contains the genotypes
 132 for the j th SNP. The genotypes are coded as -1, 0, and 1 corresponding to SNP
 133 genotypes AA, AB, and BB, respectively.

134 The specifics of the Eagle method are as follows. Eagle builds the "best"
 135 model iteratively, via forward selection. Suppose s iterations of our model build-
 136 ing process have already been performed. This means s SNP-trait associations
 137 have been identified. It also means that s separate genomic regions of interest
 138 have been found. To perform the $s + 1$ th iteration, we first fit the current model
 139 to the data. The (current) model is of the form

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

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

148 The genetic effects \mathbf{u}_g are modelled as

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

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

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

163 Third, we identify the $(s+1)$ th SNP that is in strongest association with the
 164 trait, based on the maximum score statistic $t_j^2 = \frac{\tilde{a}_j^2}{\text{var}(\tilde{a}_j)}$ where \tilde{a}_j is the best
 165 linear unbiased predictor (BLUP), and $\text{var}(\tilde{a}_j)$ is its variance. This statistic is
 166 not only appealing intuitively, where we identify a SNP based on its (random)
 167 effect size and accuracy, but is justifiable, theoretically (Verbyla *et al.*, 2012).

168 Fourth, we determine the importance of the $(s+1)$ th selected SNP via a
 169 model selection strategy (Verbyla *et al.*, 2007). We begin by reforming (2)
 170 where S now contains the $s+1$ selected SNP. We then fit this new model to the
 171 data via maximum likelihood and calculate its extended Bayesian information
 172 criteria (extBIC) (Chen and Chen, 2008). The extBIC is a model selection
 173 measure that takes into account the number of unknown parameters and the
 174 complexity of the model space. It is well suited to the model selection problem
 175 in genome-wide association studies (Chen and Chen, 2008). It is different to the
 176 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 \mathbf{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

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

where $\mathbf{a}^* \sim N(\mathbf{0}, \sigma_a^2 \mathbf{I}^{(n_g \times n_g)})$, and $(\mathbf{M}_{-S} \mathbf{M}_{-S}^T)^{1/2}$ 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 $\tilde{\mathbf{a}}$ from estimates from model (3) with

$$\tilde{\mathbf{a}} = \left[\mathbf{M}_{-S}^T (\mathbf{M}_{-S} \mathbf{M}_{-S}^T)^{-1/2} \right] \tilde{\mathbf{a}}^* \quad (4)$$

where its variance matrix is

$$\text{var}(\tilde{\mathbf{a}}) = \mathbf{M}_{-S}^T (\mathbf{M}_{-S} \mathbf{M}_{-S}^T)^{-1/2} \text{var}(\tilde{\mathbf{a}}^*) (\mathbf{M}_{-S} \mathbf{M}_{-S}^T)^{-1/2} \mathbf{M}_{-S} \quad (5)$$

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

200 the calculation of (5).

201 **2.3 Comparison Methods**

202 **2.3.1 Multi-locus methods:**

203 We compare the computational and statistical performance of Eagle against
204 five multi-locus methods. They are bigRR, LMM-Lasso, glmnet, MLMM, and
205 r2VIM. All but glmnet have been purposely designed for genome-wide associ-
206 ation mapping. BigRR, LMM-Lasso, and glmnet are regression-based regular-
207 isation methods. BigRR is based on generalised ridge regression, LMM-Lasso
208 is based on lasso, and glmnet is based on elastic net. Regularisation methods
209 make parameter estimation possible in models where the number of predictors is
210 far greater than the number of samples. They allow the strength of association
211 between all the SNPs and trait to be measured within a single model, simulta-
212 neously. A limitation of these methods though is that the statistical significance
213 of the SNP effects cannot be easily determined. Due to the adaptive nature of
214 the estimation procedures, to do this analytically is challenging and is an area
215 of active research (Lockhart *et al.*, 2014). Instead, we calculate significance
216 empirically via stability selection (see below).

217 MLMM is closest in philosophy to Eagle. It too is based on building the
218 best model via forward selection, within a linear mixed model framework, using
219 the extBIC. However, there are differences between the two methods. MLMM
220 does not make use of dimension reduction. Also, it builds its "best" model
221 differently to Eagle. Eagle uses a score statistic in which to identify SNPs to
222 enter the model. MLMM uses the statistical significance of the SNPs. This
223 involves fitting a separate linear mixed model for each unidentified SNP. Both
224 are R packages but there is a significant difference in computational performance
225 (see Results).

226 R2VIM differs to the other four methods in that it is a non-parametric model-
227 free approach. It implements random forests but where multiple parallel runs are
228 performed. Each run leads to different random forests being created. A relative
229 importance score is calculated, within a run, for each SNP. This is done by
230 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 Stability Selection

Stability selection (Meinshausen and Bühlmann, 2010) is a subsampling strategy with a range of applications, including the estimation of the significance of effects in regression models where the number of parameters (p) can be much larger than the number of samples (n). We chose stability selection over permutation and other sampling procedures because of its low computational cost. We employ stability selection to estimate, empirically, the significance of the SNP effects from the regularisation method’s analyses of the simulated data.

To obtain significance estimates via stability selection, we do the following.

For LMM-Lasso and glmnet, we begin by performing a preliminary analysis of a replicate where we tune the regularisation parameter to yield 20 to 30 non-zero SNP effects. This tuning of the regularisation parameter is done for data under each simulation scenario but it is not necessary to do it for every replicate within a scenario. It is sufficient to pick a single replicate at random upon which to tune the regularisation parameter. Here, we have the luxury of knowing the genetic conditions under which the simulated trait data are generated. We know that 20 to 30 SNP-trait associations is a reasonable number of findings to expect. However, stability selection estimates are robust (Meinshausen and Bühlmann, 2010) and the regularisation parameter can be tuned to any reasonable number of non-zero effects. We sample repeatedly, without replacement, from the replicate. We draw 100 replicate subsets of size $n/2$. Each replicate subset is then analysed with LMM-Lasso and glmnet, with their regularisation parameter set to the value found in the preliminary analysis. A (probability) estimate of the statistical significance of a SNP effect is obtained by counting the number of times the SNP have a non-zero effect size over all the replicates divided by the number of replicates (which was 100).

For bigRR, we modify our stability selection procedure slightly. There is no need to tune the regularisation parameter for bigRR as an optimal value is found as part of its analysis. As described above, we draw 100 replicate subsets of size $n/2$ and analyse these data with bigRR. We then order the SNPs according to the absolute size of their SNP effects and record the top 20 SNPs. A (probability) estimate of the statistical significance of a SNP effect is then obtained by counting the number of times the SNP is recorded divided by the number of replicates.

2.5 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 the scenario, from data collected from

the 1000 Genome Project, version 3 (Consortium *et al.*, 2010). Across scenarios, the SNP data differs. Across replicates within a scenario, the SNP data are the same.

To generate the trait data \mathbf{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 $\mathbf{g}^{(n \times 1)}$ vector of genetic values where n is the number of individuals. Fifth, $\mathbf{e}^{(n \times 1)}$, a vector of residuals, is drawn from a normal distribution where $e_i \sim N(0, \sigma_e^2)$ and σ_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 $\mathbf{y} = \mathbf{g} + \mathbf{e}$. In forming \mathbf{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.

2.6 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. 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 (Zhou and Stephens, 2012) and FaST-LMM (Lippert *et al.*, 2011) because of their popularity and computational speed. For multi-locus association mapping, we chose bigRR (Shen *et al.*, 2013), glmnet (Friedman *et al.*, 2010), LMM-Lasso (Rakitsch *et al.*, 2013), MLM (Segura *et al.*, 2012), and r2VIM (Szymczak *et al.*, 2016). 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

350 of SNP and quantitative trait data. Extra realism was introduced into the
 351 simulation study through the drawing of the SNP genotypes from the 1000
 352 Genome Project, phase 3 (Consortium *et al.*, 2010). The quantitative trait
 353 data were generated by selecting, randomly, a set of SNPs and assigning these
 354 loci additive allelic effects. Random errors were then drawn from a normal
 355 distribution with variance set to give a heritability of 50% for the trait. For
 356 each individual, a quantitative trait value was obtained by summing its random
 357 error and additive allelic effects. The number of randomly selected SNPs follows
 358 a Poisson distribution with mean 30. The size of the allelic effects across the
 359 selected loci are equal.

360 Analyses by the eight programs/packages of a replicate proceeded as follows.
 361 They were all run at their default settings. Eagle and MLMM were the easiest
 362 of the programs/packages to implement. The only parameters requiring speci-
 363 fication were the amount of available memory and number of CPUs for Eagle
 364 and the number of chunks for MLMM. Their results were also immediately in-
 365 terpretable. Their findings were the set of SNPs in strongest association with
 366 the trait. Each SNP in this set identified a separate genomic region of interest,
 367 whose position was given by the map location of the SNP.

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

374 R2VIM is different from the rest in that it is a nonparametric approach for
 375 association mapping. It is based on random forests. Three important parame-
 376 ters needed to be set. These were the number of trees, the number of variables
 377 for building a tree, and the minimum size of a terminal node. Ideally, these pa-
 378 rameters would be "tuned" on a replicate-by-replicate basis (Boulesteix *et al.*,
 379 2012). However, this was not practical here. We instead used the same settings
 380 as in (Szymczak *et al.*, 2016) where the number of trees was set to 1000, the
 381 number of variables was set to 20% of the number of SNPs, and the minimum
 382 size of a node was set to 10% of the sample size. A relative importance measure

383 was calculated for each SNP measuring its strength of association with the trait.

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

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

395 **3.3 Power and False Discovery Rates**

396 Here, we answer the question of how well Eagle finds true SNP-trait associations
397 and avoids false SNP-trait associations. We do this by estimating the power and
398 false discovery rates of Eagle and the other methods for the six scenarios. Since,
399 for a replicate, we knew which SNPs were assigned additive effects, we knew the
400 SNPs that were in true association with the trait. We will refer to these SNPs as
401 being true SNPs. By knowing the true SNPs, we were able to assess the validity
402 of a method's findings. A finding was counted as true if it was positioned within
403 40 kilobase pairs of the location of a true SNP. When a replicate was analysed,
404 we obtained an estimate of the power of the method by taking the number of
405 findings that were found to be true and dividing by the number of true SNPs.
406 We also obtained an estimate of a method's false discovery rate. It is the number
407 of findings that were found to be false divided by the number of (true and false)
408 findings found by the method. Both these estimates varied with replicate. The
409 power (false discovery rate) of a method, for a scenario, was found by taking
410 the median of the power (false discovery rate) estimates over the 100 replicates.

411 The power and false discovery rates of Eagle and the other multi-locus meth-
412 ods across the scenarios 150 x 5K, 350 x 500K, 1500 x 50K, and 2000 x 500K are
413 shown in Supplementary Figure 1. We restricted our attention to these scenar-

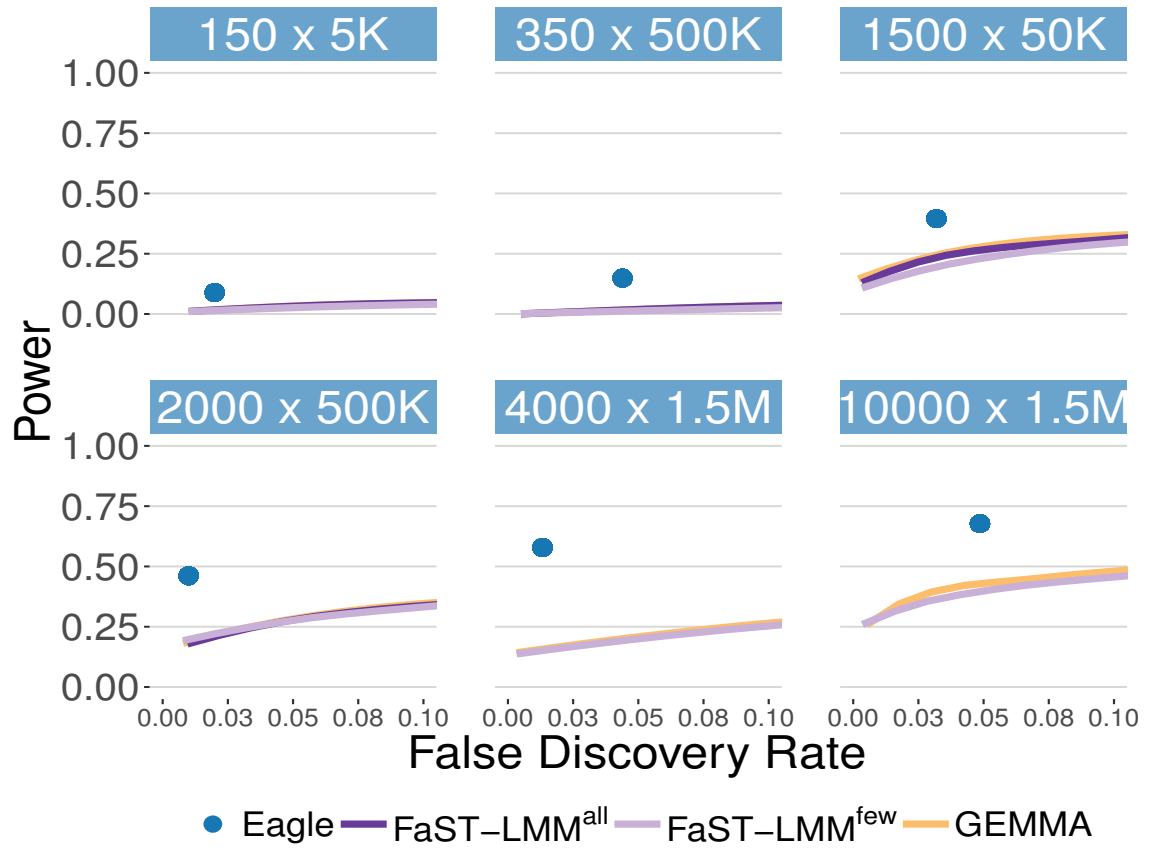
ios because not all multi-locus methods could cope with the size of data in the other scenarios. Each plot contains single points and power curves. The single points are the power and false discovery rates for Eagle and MLMM. These two methods treat association mapping as a model selection problem. Their are no significance thresholds to be set. The power curves are for those methods that treat association mapping as a 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 discovery rate as the significance threshold is adjusted. The power and false discovery 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 discovery 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

Figure 1: Power verse false discovery 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.



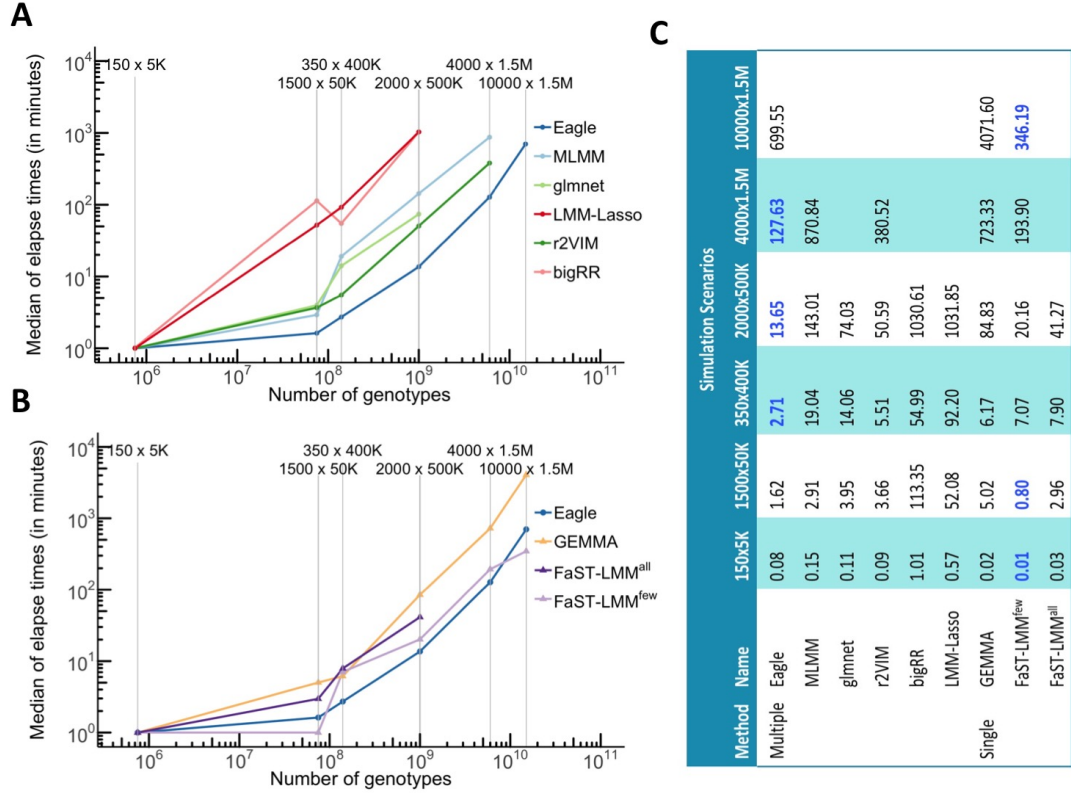
larger scenarios. Of the multi-locus programs/packages, only Eagle, with its ability to handle data larger than the memory capacity of the computer, was capable of producing findings for data from our largest scenario, 10000 x 1.5M.

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 x 400K, 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 server 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

Figure 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.



476 those findings reported in the original study. We then analysed the data with
 477 Eagle. Due to Eagle’s increased power, we found SNP-trait associations not
 478 found with FaST-LMM. However, we were able to confirm the validity of these
 479 new findings as they matched what was found in the original study. Having the
 480 ability to confirm new findings in a real study was one of the primary motivators
 481 for choosing these data for analysis.

482 We repeated the single-locus analyses as first performed (Nicod *et al.*, 2016),
 483 except that we focused on autosomal SNPs and our analyses were based on SNP
 484 genotypes rather than SNP dosages. In the original analysis, a genome-wide
 485 threshold that gave a false discovery rate of 5%, was found via permutation. We
 486 followed the same empirical procedure but increased the number of permutations
 487 from 100 to 500 for more accurate thresholds.

488 We ran Eagle in three ways. Eagle chooses the best model via the extended
 489 Bayesian information criteria (extBIC) (Chen and Chen, 2008). The conserva-
 490 tiveness of the extBIC can be adjusted by a single regularisation parameter γ
 491 that ranges from zero to one. In the simulation study, this parameter was set
 492 to one, its most conservative and default setting. The mouse data were also
 493 analysed under this setting (Eagle^{default}). An alternate (Chen and Chen, 2008)
 494 , less conservative way of setting γ is to let $\gamma = 1 - \frac{1}{(2\kappa)}$ with $\kappa = \frac{\log(L)}{\log(n_g)}$ where
 495 L is the number of loci that span the genome, and n_g is the number of individ-
 496 uals/groups/lines/strains in the study (Eagle^{alt}). However, our preferred way
 497 is to set the γ parameter for each trait via permutation (Eagle^{optimal}). We used
 498 100 permutations to set γ to give a false positive rate of 5%. This only took six
 499 times as long as a single analysis of the data. This is because the marker data
 500 need only be read once, and only the trait data changes across permutations
 501 leading to other computational efficiencies. This permutation method has been
 502 implemented within the Eagle package.

503 The genome wide results from the analyses of the mouse data are shown in
 504 Figure 3. The mouse study recorded measurements on 200 traits. Of these, in
 505 the original study, 45 were able to have their findings corroborated by previously
 506 published work. We focused our analyses here on these same 45 traits. Overall,
 507 FaST-LMM, Eagle^{default}, Eagle^{alt}, and Eagle^{optimal} found 50, 37, 67, and 106,
 508 SNP-trait findings, respectively, across 39 traits. No associations were found by

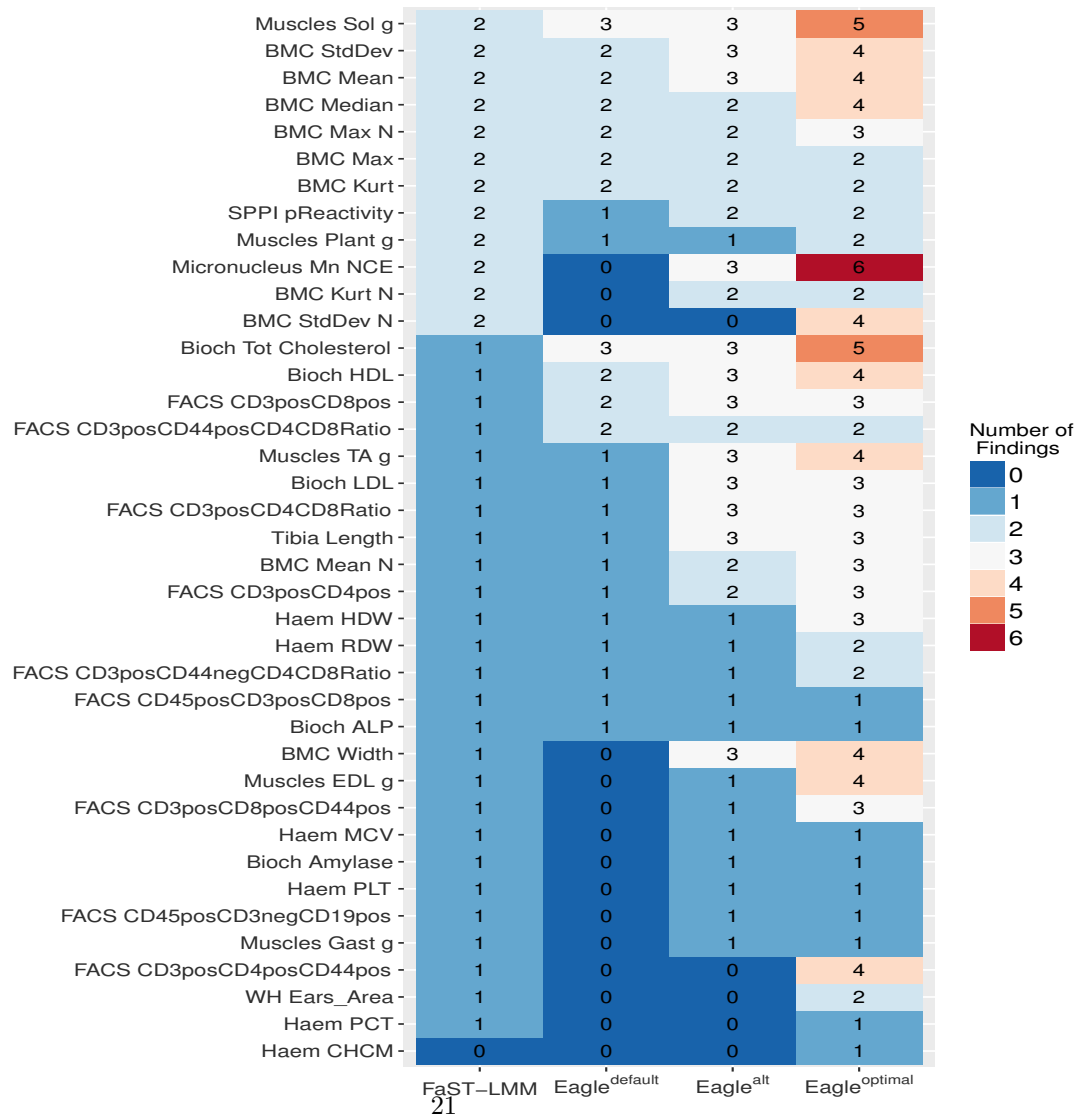
509 FaST-LMM and Eagle for the other six traits. Eagle^{alt} and Eagle^{optimal} also
510 found SNP-trait associations not found in the original study. This is despite
511 their analyses being based on the SNP genotype data and the original study be-
512 ing based on SNP dosage data. Eagle^{alt} found two and Eagle^{optimal} found seven
513 new findings (Supplementary Table 2). These new findings all involved SNPs
514 whose association had been confirmed for other related traits in the original
515 study.

516 In the simulation study, Eagle outperforms single-locus association mapping.
517 Here, Eagle^{default}, where $\gamma = 1$, finds less associations than FaST-LMM. Why
518 the discrepancy in performance? The answer lies in the conservativeness of
519 Eagle. With the added genetic complexity implicit within the mouse data,
520 Eagle is more conservative when γ is set to one than in the simulation study.
521 However, the relative results of the simulation study remain true. For similar
522 false discover rates, Eagle is superior to single-locus association mapping. As
523 a case in point, here FaST-LMM found 50 SNP-trait associations with a false
524 discovery rate of 5%. Eagle, with the same false discovery rate (Eagle^{optimal})
525 found 106 SNP-trait associations, more than a 100% increase in findings.

526 4 Discussion/Conclusion

527 Eagle is a new linear mixed model based method (and R package) for multi-
528 locus association mapping. It advances the state of association mapping in
529 several ways. First, its computational footprint is much smaller than other
530 multi-locus implementations. Eagle makes multi-locus analysis practical, even
531 when the datasets are large. Second, the results from Eagle are immediately
532 interpretable. They are the set of SNPs in strongest association with the trait
533 where each SNP identifies a separate genomic region of interest. Third, it treats
534 association mapping as a model selection problem, avoiding the need for signif-
535 icance thresholds. As we saw in the simulation study, Eagle has considerably
536 higher power than single-locus methods but is comparable in run time. Also,
537 when analysing the mouse data, Eagle found more than double the SNP-trait
538 associations than with single-locus association mapping, the method of choice.
539 Furthermore, these extra findings were all true.

Figure 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. 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.



540 Eagle outperformed the other multi-locus methods in our simulation study.
 541 However, we are cognisant of the fact that we made several implementation
 542 choices that impact our conclusions. For instance, we chose to calculate the
 543 significance of the SNP effects from bigRR, LMM-Lasso, and glmnet via stability
 544 selection. Permutation and its variants (Browning, 2008; Pahl and Schafer,
 545 2010) are also equally valid empirical approaches. Stability selection though has
 546 the advantage of being based on repeated sampling of only a proportion (50%
 547 in our case) of the data. Also, when analysing the (sub)samples, it was not
 548 necessary to calculate the entire solution path for a method. Instead, analyses
 549 are performed for a fixed value of the regularisation parameter, greatly reducing
 550 the amount of computation required. For r2VIM, an R package implementing
 551 random forests, we had to decide on the minimum size of a terminal node,
 552 the number of trees, and number of potential variables. The setting of these
 553 parameters greatly affects performance. We acknowledge that in the hands of
 554 an expert, r2VIM could be fine-tuned for a better balance of computational and
 555 statistical performance. However, we would like to think that the parameter
 556 settings we used are sensible since they match the values in the original r2VIM
 557 publication (Szymczak *et al.*, 2016).

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

572 Over the coming years, computationally, the demand placed upon associa-

tion mapping methods is going to increase. High-throughput array-based technologies continue to decrease the cost of genotyping, permitting ever larger GWASs to be performed. Whole-genome sequencing is also now a reality. Already sequence across entire genomes are being collected for GWASs (Gudbjartsson *et al.*, 2015; Long *et al.*, 2017) culminating in data on millions of SNPs. It is because of this growing demand that we have purposely structured the Eagle package for continued development. We are already experimenting with a GPU-based version of Eagle. Early results suggest that for small to moderate sized datasets (<10,000 samples), there is little improvement in performance over CPU-based computation. However, for larger study sizes, we are seeing up to a 40% decrease in run times. We also have plans for Eagle to run on computer clusters. Structuring Eagle for larger-than-memory calculations was a preemptive step in this direction. GWASs have changed significantly in the past decade but the size and complexity of GWASs is expected to change even more in the coming decade.

Data Availability

The input files for Eagle containing the mouse GWAS data are available for download from <https://doi.org/10.25919/5bc08287717dd>. The original data were obtained from the Heterogeneous Stock Mice website <http://wp.cs.ucl.ac.uk/outbredmice/heterogeneous-stock-mice/>.

Acknowledgements

We would like to thank the IM&T Scientific Computing group at the Commonwealth Scientific and Industrial Research Organisation (CSIRO) for providing access and support to the CSIRO Accelerator Cluster upon which all computation was performed.

598 Funding

599 This research received no specific grant from any funding agency in the public,
600 commercial, or not-for-profit sectors.

601 *Conflict of interest:* The authors declare that they have no competing financial
602 interests.

603 References

- 604 Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of
605 random forest methodology and practical guidance with emphasis on compu-
606 tational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data*
607 *Mining and Knowledge Discovery*, **2**(6), 493–507.
- 608 Browning, B. L. (2008). PRESTO: rapid calculation of order statistic distri-
609 butions and multiple-testing adjusted P-values via permutation for one and
610 two-stage genetic association studies. *BMC Bioinformatics*, **9**(1), 309.
- 611 Cardon, L. R. and Palmer, L. J. (2003). Population stratification and spurious
612 allelic association. *The Lancet*, **361**(9357), 598–604.
- 613 Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model
614 selection with large model spaces. *Biometrika*, **95**(3), 759–771.
- 615 Cho, S., Kim, K., Kim, Y. J., Lee, J.-K., Cho, Y. S., Lee, J.-Y., Han, B.-
616 G., Kim, H., Ott, J., and Park, T. (2010). Joint identification of multiple
617 genetic variants via elastic-net variable selection in a genome-wide association
618 analysis. *Annals of Human Genetics*, **74**(5), 416–428.
- 619 Consortium, . G. P. *et al.* (2010). A map of human genome variation from
620 population-scale sequencing. *Nature*, **467**(7319), 1061.
- 621 de Bakker, P. I., Yelensky, R., Pe’er, I., Gabriel, S. B., Daly, M. J., and Alt-
622 shuler, D. (2005). Efficiency and power in genetic association studies. *Nature*
623 *Genetics*, **37**(11), 1217.

- 624 Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for
625 Generalized Linear Models via Coordinate Descent. *Journal of Statistical*
626 *Software*, **33**(1), 1–22.
- 627 Gogel, B., Smith, A., and Cullis, B. (2018). Comparison of a one-and two-stage
628 mixed model analysis of Australia’s National Variety Trial Southern Region
629 wheat data. *Euphytica*, **214**(2), 44.
- 630 Golub, G. H. and Van Loan, C. F. (2012). *Matrix Computations*, volume 3.
631 JHU Press.
- 632 Gudbjartsson, D. F., Helgason, H., Gudjonsson, S. A., Zink, F., Oddson, A.,
633 Gylfason, A., Besenbacher, S., Magnusson, G., Halldorsson, B. V., Hjartar-
634 son, E., *et al.* (2015). Large-scale whole-genome sequencing of the Icelandic
635 population. *Nature Genetics*, **47**(5), 435.
- 636 Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly,
637 M. J., and Eskin, E. (2008). Efficient control of population structure in model
638 organism association mapping. *Genetics*, **178**(3), 1709–1723.
- 639 Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using
640 the eigenvalues of a correlation matrix. *Heredity*, **95**(3), 221.
- 641 Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heck-
642 erman, D. (2011). FaST linear mixed models for genome-wide association
643 studies. *Nature Methods*, **8**(10), 833–835.
- 644 Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A signif-
645 icance test for the Lasso. *Annals of Statistics*, **42**(2), 413.
- 646 Long, T., Hicks, M., Yu, H.-C., Biggs, W. H., Kirkness, E. F., Menni, C., Zierer,
647 J., Small, K. S., Mangino, M., Messier, H., *et al.* (2017). Whole-genome
648 sequencing identifies common-to-rare variants associated with human blood
649 metabolites. *Nature Genetics*, **49**(4), 568.
- 650 Meinshausen, N. and Bühlmann, P. (2010). Stability Selection. *Journal of the*
651 *Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473.

652 Nicod, J., Davies, R. W., Cai, N., Hassett, C., Goodstadt, L., Cosgrove, C.,
653 Yee, B. K., Lionikaite, V., McIntyre, R. E., Remme, C. A., *et al.* (2016).
654 Genome-wide association of multiple complex traits in outbred mice by ultra-
655 low-coverage sequencing. *Nature Genetics*.

656 Pahl, R. and Schafer, H. (2010). PERMORY: an LD-exploiting permutation
657 test algorithm for powerful genome-wide association testing. *Bioinformatics*,
658 **26**(17), 2093–2100.

659 Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches
660 to population stratification in genome-wide association studies. *Nature Re-*
661 *views Genetics*, **11**(7), 459.

662 Rakitsch, B., Lippert, C., Stegle, O., and Borgwardt, K. (2013). A Lasso multi-
663 marker mixed model for association mapping with population structure cor-
664 rection. *Bioinformatics*, **29**(2), 206–214.

665 Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q.,
666 and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for
667 genome-wide association studies in structured populations. *Nature Genetics*,
668 **44**(7), 825–830.

669 Shen, X., Alam, M., Fikse, F., and Rönnegård, L. (2013). A novel generalized
670 ridge regression method for quantitative genetics. *Genetics*, **193**(4), 1255–
671 1268.

672 Spielman, R. S. and Ewens, W. J. (1996). The TDT and other family-based
673 tests for linkage disequilibrium and association. *American Journal of Human*
674 *Genetics*, **59**(5), 983.

675 Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide
676 studies. *Proceedings of the National Academy of Sciences*, **100**(16), 9440–
677 9445.

678 Szymczak, S., Holzinger, E., Dasgupta, A., Malley, J. D., Molloy, A. M., Mills,
679 J. L., Brody, L. C., Stambolian, D., and Bailey-Wilson, J. E. (2016). r2VIM: A
680 new variable selection method for random forests in genome-wide association
681 studies. *BioData Mining*, **9**(1), 7.

- 682 Verbyla, A. P., Cullis, B. R., and Thompson, R. (2007). The analysis of QTL by
683 simultaneous use of the full linkage map. *Theoretical and Applied Genetics*,
684 **116**(1), 95.
- 685 Verbyla, A. P., Taylor, J. D., and Verbyla, K. L. (2012). RWGAIM: an efficient
686 high-dimensional random whole genome average (QTL) interval mapping ap-
687 proach. *Genetics Research*, **94**(6), 291–306.
- 688 Verbyla, A. P., Cavanagh, C. R., and Verbyla, K. L. (2014). Whole-genome
689 analysis of multienvironment or multitrait QTL in MAGIC. *G3: Genes*,
690 *Genomes, Genetics*, **4**(9), 1569–1584.
- 691 Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F.,
692 McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., *et al.* (2006).
693 A unified mixed-model method for association mapping that accounts for
694 multiple levels of relatedness. *Nature Genetics*, **38**(2), 203.
- 695 Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian,
696 C., Zheng, H., Dean, C., Marjoram, P., *et al.* (2007). An Arabidopsis example
697 of association mapping in structured samples. *PLoS Genetics*, **3**(1), e4.
- 698 Zheng, J., Li, Y., Abecasis, G. R., and Scheet, P. (2011). A comparison of ap-
699 proaches to account for uncertainty in analysis of imputed genotypes. *Genetic*
700 *Epidemiology*, **35**(2), 102–110.
- 701 Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis
702 for association studies. *Nature Genetics*, **44**(7), 821–824.