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

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

[1]*Data61, CSIRO, Australia.*

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

We present Eagle, a new method for multiple-locus association mapping. The motivation for developing Eagle was to make multiple 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. doesn't suffer from multiple testing issues c. is threshold free d. gives results that are immediately interpretable and e. has a computational footprint comparable to single-locus association mapping. By conducting a large simulation study, we will show that Eagle finds true and avoids false SNP-trait associations better than competing single- and multiple-locus methods. We also analyse data from a published mouse study. Eagle found over two times more validated findings than the state-of-the-art single-locus method. Eagle has been implemented as an R package, with a web-based Graphical User Interface (GUI) for users less familiar with R.

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 [1]. This prompted a shift towards family-based designs and score tests, such as the transmission/disequilibrium test (TDT) and its variants [2]. Today, instead of by design, it is through statistical modelling that we account for the effects of population stratification [3]. 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 [4,5].

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, the aim of association mapping is to identify re-

gions 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 by testing the statistical significance of a single nucleotide polymorphisms (SNP) when treated as an effect in an appropriate model. 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. Second, when multiple statistical tests are performed, the probability of wrongly accepting a result (type 1 error) is inflated. This is known as the multiple testing problem. Many different solutions have been offered [6–8]. Yet, there is still no well accepted way of correcting for multiple testing in the context of genome-wide association mapping. Third, many of the traits whose genetic 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 multiple-locus association mapping methods haven't attracted more attention. Methods based on regularisation techniques, such as ridge regression [9] and lasso [10], measure all locus-trait associations simultaneously. Here, multiple testing is not an issue. These techniques though are computationally demanding. Also, their results can be difficult to interpret. The strength of association is not measured by a p value but by the size of the regression coefficient for the SNP in the model. More recently, associations have started to be mapped with random forests [11]. Similar to regularisation techniques though, it is not clear how to infer genomic regions of interest from their findings. A multiple-locus method that does show promise is the multiple-locus linear mixed model method [12]. The best multiple-locus model is built with simple forward selection. Results are immediately interpretable but here, computation becomes challenging for large datasets.

In this paper, we present our new multiple-locus method for genome-wide association mapping, which we are calling Eagle. Eagle combines the strength of regularisation techniques (being able to fit all 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 multiple-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.

1 Results

Association Mapping Methods We compared Eagle, in terms of computational and statistical performance, against seven other association mapping methods. We chose methods that 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 multiple-locus models. Of the many ways of performing single-locus association mapping, we chose GEMMA [13] and FaST-LMM [14] because of their popularity and computational speed. For multiple-locus association mapping, we chose bigRR [9], glmnet [15], LMM-Lasso [10], MLMM [12] , and r2VIM [11]. Each takes a different approach to multiple-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).

Simulation Study A large simulation study was performed where we sought to answer two questions. First, how does Eagle compare, in terms of run time and memory usage, to competing implementations? Second, how well does Eagle find true associations (power) and avoid false associations (type 1 errors)? 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. Extra realism was introduced into the simulation study through the drawing of the SNP genotypes from the 1000 Genome Project, phase 3 [16]. 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.

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 CPU for Eagle and the number of chunks for MLMM. 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 [17]. However, this was not practical here. We instead used the same settings as in [11] 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.

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 1). The single-locus program GEMMA was by far the most memory efficient. Not surprisingly, the multiple-locus programs were memory intensive. Most required in excess of the 128 gigabytes of available RAM for the analysis of data generated under 4000 x 1.5M and 10000 x 1.5M. Even FaST-LMM, when all the SNP data were being used to calculate the similarity matrix, ran out of memory for the larger scenarios. Of the multiple-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 1. 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 multiple-locus implementations and is comparable to the single-locus implementations. For a simulation study with 150 individuals and 5000 SNPs, Eagle produced results in seconds. For the larger simulation scenarios 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 desktop computer 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.

Figure 1 goes around here

Power and False Discovery 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 discovery 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. 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 discovery rate. It is the number of findings that were found to be false divided by the number of (true and false) findings found by the method. Both these estimates varied with replicate. The power (false discovery rate) of a method, for a scenario, was found by taking the median of the power (false discovery rate) estimates over the 100 replicates.

The power and false discovery rates of Eagle and the other multiple-locus methods across the scenarios 150 x 5K, 350 x 500K, 1500 x 50K, and 2000 x 500K are shown in Figure 2. We restricted our attention to these scenarios because not all multiple-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 Supplementary Figure 2.

Figure 2 goes here

In answer to the question of how well Eagle finds true SNP-trait associations and avoids false SNP-trait associations, it does extremely well. Of the multiple-locus methods, Eagle had the highest power while keeping its false discovery rate low (Figure 2). 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 (Supplementary Figure 2).

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 [18]. This study was a little 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 [19]. By converting the dosages into genotypes and analysing the data with the single-locus program FaST-LMM, we obtained a subset of those findings reported in the original study. We then analysed the data with Eagle. Due to Eagles increased power, we found SNP-trait associations not found with 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.

We repeated the single-locus analyses, as first performed [18], except that we focused on autosomal SNP and our analyses were based on SNP genotype rather than SNP dosage. In the original analysis, a genome-wide threshold that gave a false discovery rate of 5%, was found via permutation. We followed the same empirical procedure but had to increase the number of permutations from 100 to 500.

Eagle was run in two ways; under its default settings (Eagle^{default}) and where we specified the regularisation parameter for model selection (Eagle^{optimal}). Eagle chooses the best model via the extended Bayesian information criteria (extBIC) [20]. The conservativeness of the extBIC can be adjusted by a single regularisation parameter that ranges from zero to one. In the simulation study, this parameter was set to one, its most conservative and default setting. However, there is also opportunity to set the parameter to a value less than one. This increases power but also increases the false discovery rate. For each trait, we used permutation to set the regularisation parameter to give a false discovery rate of 5% .

The genome wide results from the analyses of the mouse data are shown in Figure 3. The

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

Figure 3 goes here

2 Discussion

Eagle is a new linear mixed model based method (and R package) for multiple-locus association mapping. It advances the state of association mapping in several ways. First, its computational footprint is much smaller than other multiple-locus implementations. Eagle makes multiple-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 and the need for significance thresholds. 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 double the SNP-trait associations than with single-locus association mapping, the method of choice. Furthermore, these extra findings were all true.

Eagle outperformed the other multiple-locus methods in our simulation study. However, we are cognizant 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 [21,22] 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 [11].

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 [23]. A better solution is to fit a single model to the data. Although not specifically designed for association mapping, WGAIM [24], upon which Eagle is based, and RWGAIM [25] 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.

Over the coming years, computationally, the demand placed upon association 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 [26,27] 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.

3 Methods

Mouse Data The data were obtained from a large genome-wide association study that was performed in outbred mice [18]. 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, 43 yielded SNP-trait associations that could be corroborated through other independent published work. It was these 43 traits that were the focus of our real data analyses. As in the original study [18], 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.

Eagle Approach for Multiple-locus Association Mapping Eagle is a method for multiple-locus association mapping on a genome-wide scale. It is based on linear mixed models. It differs from most other single- and multiple-locus association mapping methods. Eagle treats association mapping as a model selection instead of variable selection problem. Consequently, we do not have to contend with multiple testing issues or having to construct significance thresholds. Eagle also reports as its findings only those SNPs that are in strongest linkage disequilibrium, and hence closest to 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 [24].

Let $S = \{S_1, S_2, \ldots, S_s\}$ be a set of s ordinal numbers where S_k is the S_kth ordered SNP that was selected in the kth 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 jth 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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u}_g + \boldsymbol{e} \tag{1}$$

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 [28,29] 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

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

where $\boldsymbol{m}_{S_k}^{(n_g \times 1)}$ is the vector of genotypes for the kth selected SNP, a_{S_k} is the additive effect of the kth 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 residual 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 [30].

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 $\text{var}(\widetilde{a}_j)$ 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 [25].

Fourth, we determine the importance of the $(s+1)$th selected SNP via a model selection strategy [24]. 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) [20]. 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 [20]. It is different to the model selection measure used in [24]. 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.

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 [25,31]. 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 [32]. 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 [25,31] 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).

Comparison Methods

Multiple-locus methods

We compare the computational and statistical performance of Eagle against five multiple-locus methods. They are bigRR, LMM-Lasso, glmnet, MLMM, and r2VIM. 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 [33]. 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 forward selection, within a linear mixed model framework, using the extBIC. However, there are differences between the two methods. MLMM does not make use of dimension reduction. Also,

it builds its "best" model differently to Eagle. Eagle uses a score statistic in which to identify SNPs to enter the model. MLMM uses the statistical significance of the SNPs. This involves fitting a separate linear mixed model for each unidentified SNP. Both are R packages but there is a significant difference in computational performance (see Results).

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.

Single-locus methods

We also compare the performance of Eagle against two single-locus methods, GEMMA [13] and FaST-LMM [14]. 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 [13], 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.

Stability Selection Stability selection [34] 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 selec-

tion over permutation and other sampling procedures because of its low computational cost. We employe 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 [34] 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.

Generation of Simulation Data The data are generated via data perturbation [29]. 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 [16]. Across scenarios, the SNP data differs. Across replicates within a scenario, the SNP data are the same.

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 $\boldsymbol{g}^{(n\times 1)}$ vector of genetic values where n is the number of individuals. Fifth, $\boldsymbol{e}^{(n\times 1)}$, 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 $\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{e}$. 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.

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 (https://cran.r-project.org).

1. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *The Lancet* **361**, 598–604 (2003).

2. Spielman, R. S. & Ewens, W. J. The tdt and other family-based tests for linkage disequilibrium and association. *American Journal of Human Genetics* **59**, 983 (1996).

3. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics* **11**, 459 (2010).

4. Purcell, S. *et al.* Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).

5. Bradbury, P. J. *et al.* Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).

6. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

7. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221 (2005).

8. de Bakker, P. I. *et al.* Efficiency and power in genetic association studies. *Nature Genetics* **37**, 1217 (2005).

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

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

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

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

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

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

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

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

17. Boulesteix, A.-L., Janitza, S., Kruppa, J. & König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 493–507 (2012).

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

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

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

21. Browning, B. L. Presto: rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics* **9**, 309 (2008).

22. Pahl, R. & Schafer, H. Permory: an ld-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics* **26**, 2093–2100 (2010).

23. Gogel, B., Smith, A. & Cullis, B. Comparison of a one-and two-stage mixed model analysis of australia's national variety trial southern region wheat data. *Euphytica* **214**, 44 (2018).

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

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

26. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the icelandic population. *Nature Genetics* **47**, 435 (2015).

27. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nature Genetics* **49**, 568 (2017).

28. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203 (2006).

509 29. Zhao, K. *et al.* An arabidopsis example of association mapping in structured samples. *PLoS*
510 *Genetics* **3**, e4 (2007).

511 30. Kang, H. M. *et al.* Efficient control of population structure in model organism association
512 mapping. *Genetics* **178**, 1709–1723 (2008).

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

515 32. Golub, G. H. & Van Loan, C. F. *Matrix computations*, vol. 3 (JHU Press, 2012).

516 33. Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. A significance test for the lasso.
517 *Annals of Statistics* **42**, 413 (2014).

518 34. Meinshausen, N. & Bühlmann, P. Stability selection. *Journal of the Royal Statistical Society:*
519 *Series B (Statistical Methodology)* **72**, 417–473 (2010).