Eagle: an R package for multi-locus association mapping on a genome-wide scale

Andrew W. George
Commonwealth Scientific and Industrial Research Organisation

Arunas Verbyla
Commonwealth Scientific and Industrial Research Organisation

Joshua Bowden
Commonwealth Scientific and Industrial Research Organisation

Abstract

Eagle is an R package for multi-locus association mapping on a genome-wide scale. It is unlike other multi-locus packages in that it is easy-to-use for R users and non-users alike. It has two modes of use, command line and GUI. **Eagle** is fully documented and has its own supporting website, http://eagle.r-forge.r-project.org/index.html. **Eagle** is a significant improvement over the method-of-choice, single-locus association mapping. It has greater power to detect SNP-trait associations, does not suffer from multiple testing issues, and there is no need for significance thresholds. It is based on model selection, linear mixed models, and a clever idea on how random effects can be used to identify SNP-trait associations. Through an example with real mouse data, we demonstrate **Eagle**'s ability to bring clarity and increased insight to single-locus findings. Initially, we see **Eagle** complimenting single-locus analyses. However, over time, we hope the community will make, increasingly, multi-locus association mapping their method-of-choice for the analysis of genome-wide association study data.

Keywords: association mapping, linear mixed model, model selection, genome-wide association study.

1. Introduction

The **Eagle** package was developed to meet a shared need in animal, plant, and human genetics. It was built to make multi-locus association mapping easy. Multi-locus association mapping is more powerful, statistically, than single-locus association mapping Wang, Feng, Ren, Huang, Zhou, Wen, Zhang, Dunwell, Xu, and Zhang (2016); Zhang, Jia, and Dunwell (2019). By being able to model the association between multiple single nucleotide polymorphisms (SNPs) and a trait simultaneously, multi-locus association mapping better captures the hidden reality of heritable traits with complex genetic architectures. Yet, multi-locus association mapping is rarely used in practice. Many of the current software implementations are not easy to use, can produce results that can be difficult to interpret, are driven by high-level statistical theory making their inner statistical workings mysterious to non-statisticians, and tend to be computationally ineffecient. **Eagle** does not suffer from these limitations.

Genome-wide association studies (GWASs) have become an important resource for unlocking the genetic secrets of heritable traits. They are a "first step" on the road to revealing the genes active for a trait. Their data are analysed with association mapping methods. The

goal of association mapping is to find the SNPs in strongest association with a trait. These are the SNPs closest to the active genes. However, before these SNPs can be found, there are a number of challenges association mapping must overcome. Modern-day GWASs can collect genotypes on millions or even tens of millions of SNPs but comparatively, GWAS have small sample sizes in the hundreds or thousands of individuals. This is a challenge, statistically, for association mapping. Another statistical challenge is that an association is created between a SNP and trait when the SNP is in close proximity to an active gene. However, familial relatedness, population structure, and environmental effects can also cause a SNP to be associated with a trait. To avoid spurious findings, these competing sources of association need to be accounted for by the association mapping method.

The method-of-choice when analysing GWAS data is to fit a separate linear mixed model (LMM) for each SNP Yu, Pressoir, Briggs, Bi, Yamasaki, Doebley, McMullen, Gaut, Nielsen, Holland *et al.* (2006); Zhao, Aranzana, Kim, Lister, Shindo, Tang, Toomajian, Zheng, Dean, Marjoram *et al.* (2007). The LMM framework is well suited to handling multiple sources of association. Here, a SNP is treated as a fixed effect whose statistical significance is a measure of the strength of association between the SNP and trait. The model includes a random effect for familial relatedness. It may also include other fixed effects for environmental factors and population structure. To a statistician, fitting a separate model for each SNP may seem strange. Surely, it would be better to use, say, variable selection techniques to find the SNP of interest? This is true but it quickly becomes intractable, computationally. This practice is also abetted by several highly efficient and well developed software packages, purpose built for the analysis of GWAS data. Such software includes **PLINK** Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, De Bakker, Daly *et al.* (2007), **TASSEL** Bradbury, Zhang, Kroon, Casstevens, Ramdoss, and Buckler (2007), and **GAPIT** Lipka, Tian, Wang, Peiffer, Li, Bradbury, Gore, Buckler, and Zhang (2012). Still, by analysing each SNP separately, a single-locus model is wrongly assumed. Also, how best to control the type 1 error rate is an issue. SNP data are correlated, with a dependence structure that changes along a genome. Adjusting significance thresholds for multiple testing, appropriately, is non-trivial. Multi-locus association mapping does not suffer from the afore mentioned issues.

Over the past decade, there has been a growing assortment of R packages for multi-locus association mapping. Early R packages, **bigRR** Shen, Alam, Fikse, and Rönnegård (2013), **LMM-Lasso** Rakitsch, Lippert, Stegle, and Borgwardt (2013), and **glmnet** Friedman, Hastie, and Tibshirani (2010), were focused on regularisation techniques but the interpretation of results is difficult. The R package **MLMM** Segura, Vilhjálmsson, Platt, Korte, Seren, Long, and Nordborg (2012) avoided this difficulty by treating association mapping as a model selection problem. **FarmCPU** Liu, Huang, Fan, Buckler, and Zhang (2016) and its much faster cousin **BLINK** Huang, Liu, Zhou, Summers, and Zhang (2019) implement a two model strategy where results are passed back and forth between the two models. The first model measures the strength of a SNP-trait association on a SNP-by-SNP basis. The second model, with the help of results from the first model, identify "important" pseudo quantitative trait nucleotides (pseudo-QTNs). These pseudo-QTNs are then fed back into the first model and the fitting process repeated for improved measures of association. **MRMLM** Wang *et al.* (2016) and **FASTmrEMMA** Wen, Zhang, Ni, Huang, Zhang, Feng, Wang, Dunwell, Zhang, and Wu (2018) also make use of two models but in a staged approach. In the first stage, the strength of association between a SNP and trait is measured for each SNP separately. Here, a SNP is treated as a random effect. Those SNPs which are deemed significant, according to some

threshold, are moved to the second-stage. A single multi-locus model is then formed from those SNPs which were identified in the first stage and fitted to the data. It is true to say that software availability is not the cause for the lack of mainstream acceptance of multi-locus association mapping.

In this paper , we present **Eagle**, an R package for multi-locus association mapping. We created our package to be as fast as single-locus association mapping, to be easy to use even for non-R users, and to give easily interpretable results. Methodologically, it is only a little more complicated than single-locus methods. The "best" LMM is built iteratively. At each iteration, the SNP in strongest association with a trait is identified from the random effects part of the model and moved to the fixed effects part of the model. This process is simple yet ingenious. It simultaneously identifies those regions of the genome that house genes influencing a trait while also accounting for all other SNP-trait associations.

Eagle is not like other multi-locus R packages. R, by default, comes with single-threaded math libraries. By replacing these libraries with their multi-threaded counterparts, certain linear algebra operations become parallelised, implicitly. The **Eagle** package has been structured, purposely, to make extensive use of these implicitly parallelised operations. In the parts of **Eagle** where this has not been possible, we have instead written C++ routines and parallelised the code explicitly through openMP Dagum Leonardo (1998). **Eagle** differs though most from other multi-locus R packages in its ease-of-use for non-R users. Considerable effort has been invested in making **Eagle** equally useable to R and non-R users. The **Eagle** package comes with a browser-based graphical user interface (GUI). A user need only issue a single R command, `OpenGUI()`, to harness the full functionality of **Eagle**. **Eagle** has its own website (`http://eagle.r-forge.r-project.org/index.html`) with instructions on how to install a multi-threaded version of R, quick start guide, tutorials, videos, and answers to frequently asked questions. Users can experiment with **Eagle**, prior to installing the package, by analysing a test data set on our public server (`http://eagle.r-forge.r-project.org/demo.html`).

2. Methodology

Eagle implements a recently developed method for multi-locus association mapping George, Verbyla, and Bowden (2020). It is based on LMMs. Below, some notation, the model, how the dimensionality of the model can be reduced, and the Eagle algorithm is described. For consistency, the same notation as George *et al.* (2020) is used.

2.1. Notation

Suppose genotypes are collected on L loci from n_g individuals/lines/strains. The genotypes are coded as -1, 0, and 1 corresponding to SNP genotypes AA, AB, and BB, respectively. Ideally, missing genotypes are imputed prior to analysis. If not, missing genotypes are set to 0 by **Eagle**. Let $\boldsymbol{M}^{(n_g \times L)} = [\boldsymbol{m}_1 \boldsymbol{m}_2 \dots \boldsymbol{m}_L]$ be the matrix of SNP genotype data where the vector $\boldsymbol{m}_j^{(n_g \times 1)}$ contains the genotypes -1, 0, and 1 for the jth SNP. Furthermore, let $\boldsymbol{y}^{(n \times 1)}$ contain the quantitative trait data. Here, n can be larger than n_g if multiple measurements, as is common in plant studies, are recorded on the same line/strain.

The model is built iteratively. At each iteration, a SNP is selected and moved from the random effects to the fixed effects. Suppose s iterations of the model building process have been performed. Let $S = \{S_1, S_2, \dots, S_s\}$ be a set of ordinal numbers. The number S_k

corresponds to the S_kth SNP in the marker map that was selected in the kth iteration of the model building process. For example, if $S = \{101, 12, 1143\}$, then the 101th, 12th, and 1143th SNP in the marker map were selected in the first, second, and third model selection iterations, respectively.

2.2. Multi-locus model

The standard LMM for association mapping is

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

where $\boldsymbol{X}^{(n \times p)}$ and $\boldsymbol{Z}^{(n \times n_g)}$ are known design matrices, n is the number of observations, n_g is the number of individuals/lines/strains/ with $n_g \leq n$, $\boldsymbol{\tau}^{(p \times 1)}$ is a vector with p fixed effects parameters including the intercept, and $\boldsymbol{u}_g^{(n_g \times 1)}$ is a vector containing the genetic effects. The residuals, $\boldsymbol{e}^{(n \times 1)}$, are assumed to follow a normal distribution with mean 0 and covariance matrix $\sigma_e^2 \boldsymbol{I}^{(n \times n)}$ where σ_e^2 is an unknown residual variance.

In the standard LMM, the genetic effects, $\boldsymbol{u}_g^{(n_g \times 1)}$, is a random term that accounts for familial relatedness Yu *et al.* (2006); Zhao *et al.* (2007). It is assumed to follow a $N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{G}^{(n \times n)})$ where $\boldsymbol{G}$ is a relationship matrix and σ_g^2 is the unknown genetic variance. The relationship matrix is calculated from pedigree records or from SNP data. **Eagle** though models $\boldsymbol{u}_g$ differently and this is where the innovation lies. In the standard model , $\boldsymbol{u}_g$ measures relatedness between individuals but **Eagle** instead measures relatedness between SNP.

The genetic effects 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 SNPs 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 first term on the left hand side are the fixed effects. The second term are the random effects. The fixed effects measure the additive effect of the S already-selected SNPs on the trait. The random effects measure the association between all other $L - s$ SNPs and trait, simultaneously. Here, SNPs are assumed to be uncorrelated to reduce model complexity, making the analysis more manageable. Also, for a working model, it is not uncommon to assume SNP effects are pairwise uncorrelated. Such an assumption has long been made for marker-assisted selection with ridge regression Whittaker, Thompson, and Denham (2000).

2.3. Dimension reduction

In modern genome-wide association studies, the number of loci, L, can be very large, sometimes in the tens of millions. This creates a problem, computationally, when fitting (2) as the vector $\boldsymbol{a}_{-S}$ contains a large number of elements. Fortunately, the dimensionality of (2) can be reduced by orders of magnitude.

The goal is to form an equivalent model for (2) of lower dimension but where the equivalent model has the same variance. The variance structure of $\boldsymbol{u}_g^{(n_g \times 1)}$ is the $n_g \times n_g$ matrix

$\sigma_a^2 \boldsymbol{M}_{-S}\boldsymbol{M}_{-S}^T$. Here, the only unknown is the variance σ_a^2. By taking the matrix square root,

$$\boldsymbol{Z}_e = (\boldsymbol{M}_{-S}\boldsymbol{M}_{-S}^T)^{1/2}$$

an equivalent, dimension-reduced, model for $\boldsymbol{u}_g$ is

$$\boldsymbol{u}_g = \sum_{k=1}^{s} \boldsymbol{m}_{S_k} a_{S_k} + \boldsymbol{Z}_e \boldsymbol{a}^* \tag{3}$$

where $\boldsymbol{a}^*$ is a random effect with only n_g elements and distributed as $N(\boldsymbol{0}, \sigma_a^2 \boldsymbol{I}^{(n_g \times n_g)})$.

The **Eagle** algorithm requires estimates of $\boldsymbol{a}$ and its variance to identify the SNP in strongest association with the trait. These can be recovered from the fitting of the dimension reduced model Verbyla, Taylor, and Verbyla (2012); Verbyla, Cavanagh, and Verbyla (2014) since

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

and its variance matrix is

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

Only the diagonal elements of the variance matrix are needed which simplifies its calculation.

2.4. The Eagle algorithm

Eagle treats association mapping as a model selection problem. The model is built iteratively, via forward selection. At each iteration, from the current model, a new model is formed. This is done by selecting a SNP from the random effects and moving it to the fixed effects. The SNP is selected based on a score statistic. The reasoning behind moving effects from random to fixed is if there are major SNP-trait associations, then at first, they are contained in the genetic background of the model. This gives opportunity for the genetic background to act as a SNP selection mechanism. Major SNP-trait associations are identifiable as outliers when compared to background effects.

Suppose s iterations of the model building process have been performed. The current model is of the form (1) and (3). The vector of genetic effects $\boldsymbol{u}_g$ has s fixed effects for the s discovered SNP-trait associations. The model has been fitted and parameter estimates obtained via maximum likelihood. The vector of random effects $\tilde{\boldsymbol{a}}^*$ and its variance $var(\tilde{\boldsymbol{a}}^*)$ are then computed Robinson (1991).

The following steps are performed for the $(s+1)$th iteration of the model building process.

Step 1: SNP selection. A SNP is selected from the random effects based on the maximum score statistic

$$t_j^2 = \frac{\widetilde{a}_j^2}{\text{var}(\widetilde{a}_j)}$$

where j refers to the jth SNP in the marker map, the j index is over all SNPs except the s SNPs already selected, $\widetilde{a}_j^2$ is a scalar value formed from the square of the best linear unbiased predictor of the jth SNP's random effect, and $\text{var}(\widetilde{a}_j)$ is its variance. These values are recovered from $\widetilde{a}^*$ and $\text{var}(\widetilde{a}^*)$, which were obtained from the fitting of the current model, and equations (4) and (5). By choosing the SNP with the maximum score statistic, we are

selecting the SNP which is in strongest association with the trait, from amongst those SNP whose association is being modelled by the random effects.

Step 2: model building and fitting. A new dimension-reduced model is built, according to (1) and (3), from the trait data $\boldsymbol{y}$, and known matrices $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{M}_{-S}$. Here, S is the set of indexes of the s previously selected SNP and the additional SNP found in the previous step. The model is fitted to the data and parameters estimated via maximum likelihood.

Step 3: model selection. The importance of the $(s+1)$th selected SNP is determined via the extended Bayes information criteria (extBIC, Chen and Chen (2008)). The extBIC is a model selection measure that takes into account the number of parameters and the complexity of the model space. If the extBIC increases, then the new model is accepted and the iterative model building process continues.

Upon completion, S is the set of indexes of the SNP in strongest and measurable association with the trait. Each SNP identifies a different part of the genome housing genes that are influencing the trait.

3. The Eagle package

3.1. Overview

Eagle is an R package for the genome-wide analysis of association data. It can handle data collected from inbred or outbred study populations. The populations can be of arbitrary and unknown complexity. The data can be larger than the memory capacity of the computer. Since **Eagle** is framed within a LMM paradigm, it is best suited to the analysis of data on continuous normally distributed traits. LMMs though can also tolerate non-normal data Schielzeth, Dingemanse, Nakagawa, Westneat, Allegue, Teplitsky, Réale, Dochtermann, Garamszegi, and Araya-Ajoy (2020). A flow chart of the analysis pipeline for **Eagle** is shown in Figure 1. The package contains functions for opening the GUI, inputing the data, performing genome-wide analyses, and for summarising and visualising the results. Non-R users need only be familiar with a single function `OpenGUI()` that opens the GUI.

3.2. Installation

Eagle is available on CRAN. As such, it can be installed in the usual way. **Eagle** though has been designed to make extensive use of implicit parallelisation. Many of the vector, matrix, and linear algebra operations in R link directly to the API's of BLAS (Basic Linear Algebra Subroutines, see Blackford, Petitet, Pozo, Remington, Whaley, Demmel, Dongarra, Duff, Hammarling, Henry *et al.* (2002)) and LAPACK (Linear Algebra Package, see Anderson, Bai, Bischof, Blackford, Demmel, Dongarra, Croz, Greenbaum, Hammarling, McKenney *et al.* (1999)). R, by default, comes with single-threaded versions of these libraries. If these libraries are replaced by their multi-threaded counterparts, such as MKL and openBLAS, parts of R become multi-threaded, implicitly. Detailed instructions for converting R to multi-threaded computation are available on the Eagle website (`http://eagle.r-forge.r-project.org/instruction.html`).

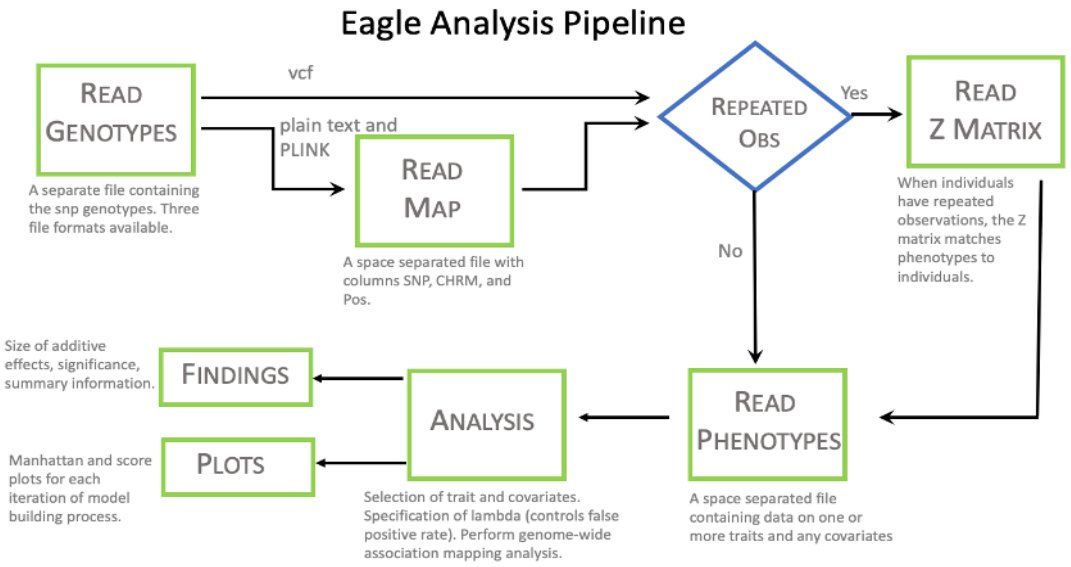

Figure 1: A flow chart of the analysis pipeline for **Eagle**. Each green box corresponds to a separate page of the GUI.

3.3. Data input

There are, potentially, four different types of data required by **Eagle** for input. These are the phenotypic data, the genotypic data, the marker map, and the Z matrix. Whether all four are needed is dependent upon the study design and format of the genotypic data. Each input data type is discussed below.

The phenotypic data consists of observations on one or more traits and any explanatory variables. A trait may have a single observation per individual/line/strain or, as is common in plant studies, may have repeat observations. The data are arranged into columns. The first row contains the column headings. The observations can be space or comma separated. Missing trait and/or explanatory variable values are allowed. The data are read into **Eagle** with the function `ReadPheno()`.

The genotypic data are the genotypes observed on the individuals/lines/strains from the SNPs. Since association studies can collect genotypes on thousands of individuals across millions of SNPS, these data can be extremely large. Fortunately, **Eagle** can handle data beyond a computer's memory capacity. **Eagle** will accept genotypic data that are in variant call format, space delimited ASCII format (the default), or PLINK ped format. The data are read with the function `ReadMarker()`. The argument `type` , which has the value `"vcf"`, `"text"`, or `"PLINK"`, specifies the type of data being read. The argument `availmemGb` tells **Eagle** how much memory, in gigabytes, are available. The order of the SNPs in the input file must correspond to their map order. Ideally, missing genotypes are imputed prior to input but some missing genotypic data can be tolerated.

The marker map consists of the names and locations of the SNPs. The map is specified via three columns of data. The first column contains the SNP labels. The second has the names of the chromosomes upon which the SNPs residue. The third column has their chromosomal

positions. The data are space separated with the first row being the column names. The SNPs are in map order. Missing values are not allowed. The data are read with the function `ReadMap()`. If the genotypic data are in variant call format, a separate marker map file is not needed. A variant call format file contains not only the SNP genotypes but also marker map information.

The Z matrix is needed only if a trait has repeat observations. It is an incidence matrix. As such, it contains zeroes and ones only. The number of rows in the matrix equals the number of rows of phenotypic data. The number of columns equals the number of rows of genotypic data. The data are space separated. The function `ReadZmat()` reads the data.

3.4. Controlling the type 1 error rate

All association mapping methods commit type 1 errors. For some, the type 1 error rate is controlled explicitly. For others, it is implicit to the internal workings of the methodology. In **Eagle**, the conservativeness of the model building process is managed explicitly via the parameter λ. The parameter λ is part of the extBIC. It ranges from zero to one. The conservativeness of the extBIC increases with increasing λ. Although it is possible to set λ analytically, the desired type 1 error rate is not part of the calculation. Instead, an empirical approach is adopted in **Eagle**.

A permutation approach is implemented in the function `FPR4AM()`. It finds the type 1 error rate for discrete values of λ. If n_{perm} permutations are performed and their are n_λ discrete values of λ being considered, potentially this means $n_{perm} \times n_\lambda$ genome-wide analyses are required. For large data sets, this quickly becomes computationally intractable. Fortunately, even though the trait data $\boldsymbol{y}$ changes across replicates, the SNP and explanatory variable data remains the same. This means, for a permutation, only those parts of the **Eagle** algorithm impacted by a change in $\boldsymbol{y}$ need recalculation. Also, through vectorisation, the type 1 error rates corresponding to all n_λ discrete values of λ can be calculated simultaneously, for each permutation.

In the example below, the run time for `FPR4AM()` was 36 seconds and for the analysis it was 20 seconds but the run times are situation specific. Having to run `FPR4AM()` can more than double the computational cost of an analysis but being able to control the type 1 error more than compensates for this cost.

3.5. Association mapping analysis

One of the most important functions in the package, from a user's perspective, is `AM()`. This function implements the methodology presented in Section 2. It is the function that performs association mapping. The function has 12 arguments. The important arguments are as follows. The trait and fixed effects are specified via the arguments `trait` and `fformula`, respectively. The data are passed to `AM()` through the arguments `pheno`, `geno`, `map`, and if required, `Zmat`. The number of threads, for parallel computation, is set with `ncpu`. The type 1 error rate is controlled with `lambda`. It's value is found by running `FPR4AM()`.

As an example, suppose the phenotypic data, SNP data, and marker map have been read with the functions described above and stored in data objects `phenoObj`, `genoObj`, and `mapObj`, respectively. The trait name is 'y'. The explanatory variables of interest are 'cov1' and 'cov2' and the fixed effects part of (1) has the form `cov1 + cov2 + cov1*cov2` where `cov1*cov2`

is an interaction term. Let the λ value that gives a type 1 error rate of 0.05 be 0.78 and it was found with `FPR4AM()`. Then, the function call that performs the analysis is

```
R> AM(trait = "y", fformula = "cov1 + cov2 + cov1*cov2", geno = genoObj,
+     pheno = phenoObj, map = mapObj, ncpu = 8,  lambda = 0.78)
```

The number of threads for parallel computation has been set to 8. After running the function, the SNPs closest to the genes underlying 'y' are reported. Each SNP identifies a different region of interest on the genome.

3.6. Results

Additional analysis information is obtained with the function `SummaryAM()`. Three tables are generated. They are a table of summary information, a findings table, and an effects table. The summary table contains information on items such as the number of cpu that were available, number of samples, the fixed effects formula, number of significant SNP-trait associations, and the λ value at which the analysis was performed. The findings table lists the names, chromosomes, and positions of those SNPs that were found to be in association with the trait. The effects table has the effect sizes, degrees of freedom, Wald statistic values, and p-values of the fixed effects in the model, including the SNPs that were identified as being in association with the trait.

3.7. Visualisation

`PlotAM()` is a interactive function for viewing the strength of association along a chromosome or genome. This is done on an iteration-by-iteration basis. It is useful for better understanding how a model is built, how SNP-trait association varies within a region, and how the strength of association for SNPs changes over the model building process.

The function has the form `PlotAM(AMobj=NULL,itnum=1,chr="All",type="Manhattan")`. An example of its use is given in the Example section.

3.8. Browser-based GUI

To release users from the requirement of having to know R, a GUI was built. Here, a user need only know how to load the package with `library(Eagle)` and start the GUI with `OpenGUI()`. After running `OpenGUI()` , a browser automatically opens to the GUI's home page. By clicking on the tabs in the navigation bar at the top of a page, a user can access pages for reading the input data, for performing analyses, and for summarising/visualising results.

3.9. Help

Detailed help files are available for each of the functions in the package. These help files include many worked examples. Help on a function is accessed in the usual way, with the `library()` function. With the GUI, every page contains a help banner that gives a summary of the functionality contained within the page. Single sentence help descriptions also appear as the mouse cursor hovers over different parts of a page. External to the package, an email

address eaglehelp@csiro.au has been set up to answer any queries . Also, help is available via the website http://eagle.r-forge.r-project.org/index.html.

4. Example

Here, the steps for performing association mapping with **Eagle** are presented. Both modes-of-use are given. That is, via function statements issued at the R command line and via the GUI. For each function statement, a screenshot of its matching GUI page is shown where applicable. The example is for the analysis of a mouse data set. As stated previously, the goal of association mapping is to find the SNPs in strongest association with the genes underlying a heritable trait. The data are real. They were collected from a large GWAS in outbred mice Nicod, Davies, Cai, Hassett, Goodstadt, Cosgrove, Yee, Lionikaite, McIntyre, Remme *et al.* (2016). Many different traits were measured but our focus is on high-density lipoprotein (Bioch.HDL). We chose this trait because from previous analyses, a number of genomic regions of interest across multiple chromosomes have been reported. In the original study Nicod *et al.* (2016), this trait was found to be influenced by the explanatory variables for sex (Sex), batch number (Batch), and average weight (Weight.Average). These same variables are treated as explanatory variables in our analysis. Even though large data sets are not a problem for **Eagle**, we wanted the example data to be easily accessible to R. A way of doing this is to host the data on GitHub (https://github.com/geo047/Example_Data) . GitHub has a file size limit of 10 megabytes, which made it necessary to base the example on a subset of the original data.

4.1. Creating the input files

Three input files were created. These are the files phenoex.dat with the phenotypic data, genoex.dat with the genotypic data, and mapex.dat with the maker map. In the original study, data were collected from 1887 outbred mice on a large number of traits and SNPs. As such, this data set was too large to house on GitHub. So the study size was reduced to 800 randomly selected mice. Our focus was restricted to the analysis of a single trait, Bioch.HDL. The genotypic data was reduced to the genotypes from 70484 SNPs. The SNPs were selected from every 5th locus of the original set where loci with a minor allele frequency of less than 1% have been removed along with loci on the sex chromosome.

The phenotypic data in phenoex.dat is space separated and arranged into 801 rows and four columns. The rows correspond to data on different mice. The columns contain data on Bioch.HDL and the explanatory variables Sex, Batch, and Weight.Average. The first row has the column names.

The genotypic data in genoex.dat has 800 rows and 70484 columns. The data are space separated. The columns are not named, nor are there missing values. Each row contains the genome-wide data for a mouse. Each column contains the genotypes for a SNP. Rows in the two files are assumed to be ordered such that the same row in each file corresponds to data collected on the same mouse. The columns are in marker map order. A numeric coding of 0, 1, and 2 was used for the SNP genotypes, AA, AB, and BB, respectively.

The marker map in mapex.dat is space separated and has 70485 rows and three columns. The first row is the column names. The rows contain map information on the SNPs. The rows are ordered according to a SNP's map order. The first column has the names of the SNPs.

The second column contains the chromosome names upon which the SNPs reside. The third column have the chromosome positions of the SNPs. It is assumed that the row order of the SNPs in this file matches the column order in genoex.dat.

The input files are downloaded and uncompressed from GitHub with the R commands

```
R> DIR <- "https://raw.githubusercontent.com/geo047/Example_Data/master/"
R> download.file(paste0(DIR, "mapex.dat"))
R> download.file(paste0(DIR, "phenoex.dat"))
R> download.file(paste0(DIR, "genoex.dat.zip"))
R> unzip("genoex.dat.zip")
```

4.2. Single-locus association mapping

We begin by analysing these data in the "usual" way, with single-locus association mapping. As stated previously, for a single-locus analysis, a separate LMM is fitted to the data for each SNP. The statistical significance of a SNP, when treated as a fixed effect, is a measure of the strength of association between the SNP and trait. The data were analysed with the R package GAPIT Wang and Zhang (2018). The process was to read in the phenotypic and genotypic information with `read.table()`, convert the explanatory variables into a useable form with `covObj <- model.matrix( ~Sex+Batch+Weight.Average, phenoObj)`, and perform the analysis with the `GAPIT()` function with the argument `model="MLM"` for single-locus association mapping with LMMs.

The single-locus analysis results are shown in the Manhattan plot in Figure 2. The positions of the SNPs on the genome are on the x-axis and the significance scores ($-log_{10}(p\text{-}value)$) of the SNPs are on the y-axis. We would conclude from this analysis that there is a single region of interest on chromosome 5. In fact, with **Eagle**, there are five regions of interest for this trait. These are on chromosomes 1, 5, 6, and 10. The region on chromosome 5 is obvious. The regions on 1 and 6 we might have suspected but lacked the power under a single-locus analysis to confirm. However, the region on chromosome 10 is only revealed after the effects of the other regions have been accounted for. Also, it is not at all obvious from a single-locus analysis that we are dealing with two closely linked regions on chromosome 5.

The **Eagle** analysis is now presented.

4.3. Reading in the data

The function statements for reading the phenotype and map input files are simple. Only the file names need specifying. This is also true if using the GUI. By default, the two input files are assumed space separated. If comma or tab separated, an additional argument is need in the function statement. For comma separated files, `sep=","`. For tab separated files, `sep="\t"`. The GUI has a checkbox for choosing if the file is space or comma separated. A tab separated option is not yet available.

The function statements for reading the two input files are

```
R>  phenoObj <- ReadPheno("phenoex.dat")
R>  mapObj <- ReadMap("mapex.dat")
```

Both the `phenoObj` and `mapObj` are data frame objects.

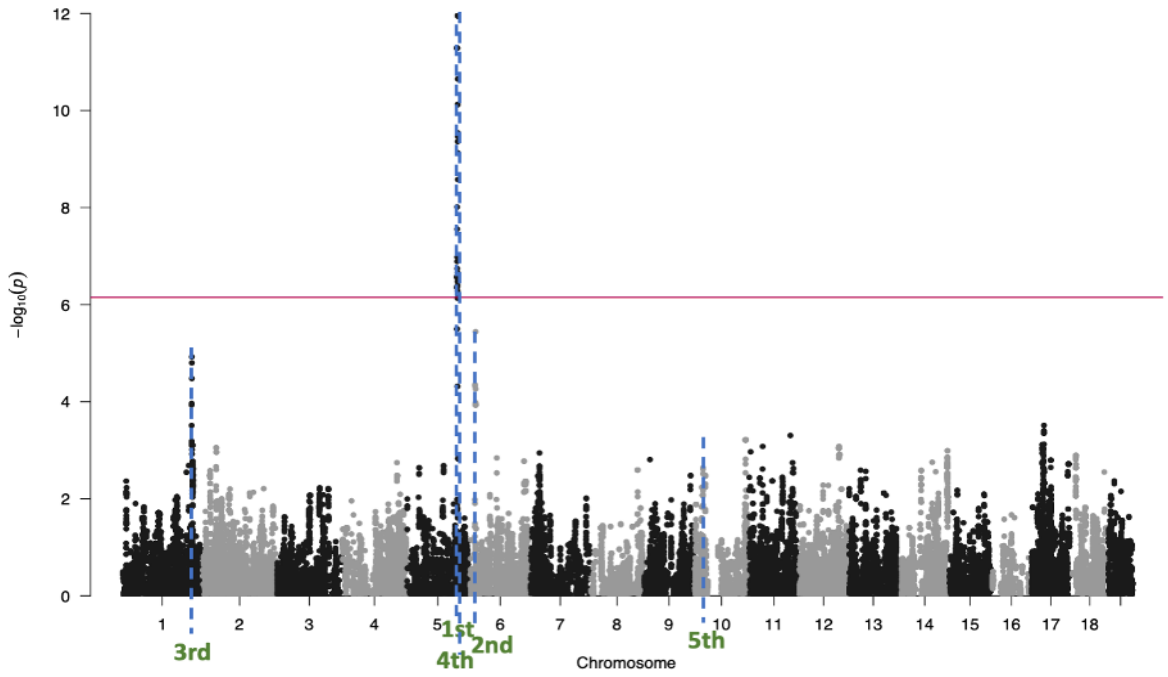

Figure 2: Manhattan plot for the single-locus analysis of the example data. Each point is the strength of association between a SNP and trait. Results are shown for the entire genome. The red horizontal line is the 5% genome-wide significance threshold, calculated via a Bonferroni correction. The blue dashed horizontal lines are the locations of the findings from **Eagle**. The order in which **Eagle** found these findings is given below the blue line. Where single-locus association mapping found only a single region of interest, because of **Eagle**'s increased statistical power, **Eagle** found five regions of interest.

Screenshots of the corresponding GUI pages are shown in Figure 3. The screenshots were taken after the relevant information had been entered and the files uploaded. The output from uploading a file is printed in the right-half of the GUI page. These are the same outputs that appear when running the function statements from the command line.

The function statement for reading the SNP data differs from the two previous input statements. Besides the file name, additional arguments are required. The file type needs specifying. Here, since the marker data are in a space delimited text file, the `type="text"` argument is included in the function statement. Other allowable formats are variant call format (`type="vcf"`) and PLINK ped (`type="PLINK"`). Text files give the user the freedom to select their own coding scheme but how these codes map to the SNP genotypes need specifying. In this example, the file contains the codes 0, 1, and 2 for SNP genotypes AA, AB, and BB, respectively. This means the function statement includes the arguments `AA=0`, `AB=1`, and `BB=2`. Also, it is good practice to set the amount of available memory, in gigabytes, with the `availmemGb` argument. The default is to assume 16 gigabytes of memory.

The `ReadMarker` function statement for this example is

```
R>  genoObj <- ReadMarker("genoex.dat",  type="text",
+                          AA=0, AB=1, BB=2, availmemGb=8)
```

A screenshot of the corresponding GUI page after uploading the file is shown in Figure 4. The output from running the statement is the same as the output shown in the right-half of

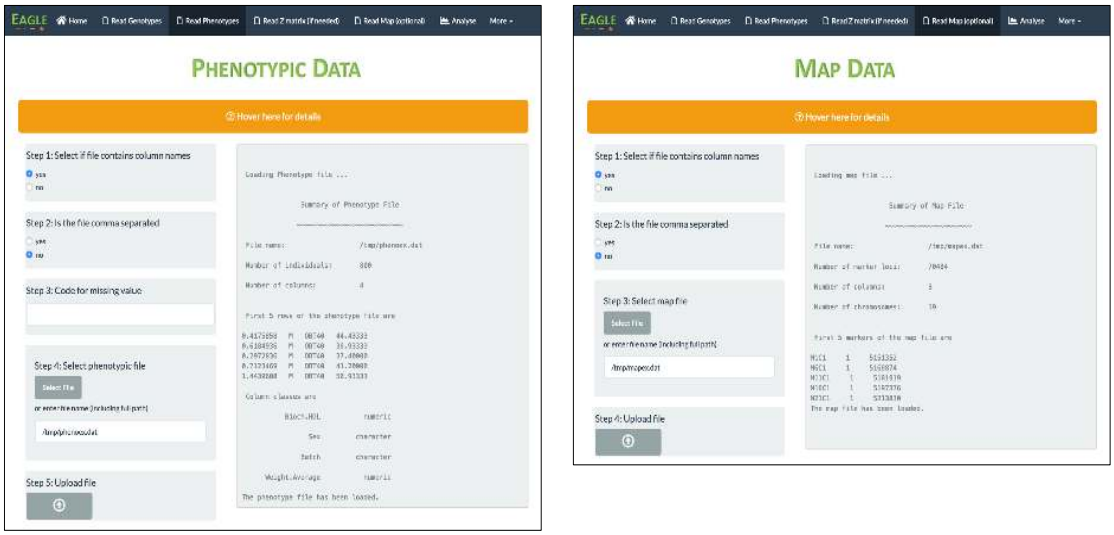

Figure 3: Screenshots of the GUI pages after the phenotypic data (left) and marker map (right) have been uploaded. Any output from the underlying functions is shown in the right-half of a page.

the GUI page. Unlike the other input functions, `ReadMarker()` does not read the data into memory. Instead, the marker data ,and its transpose, are stored on disk in a binary form. By not holding the genotype data in memory, it gives **Eagle** the ability to analyse marker data larger than the memory capacity of a computer. The object returned by `ReadMarker()` is a list object that holds elements containing information on the dimensions of the marker data, the name and location of the reformatted marker data, and the number of phenotypic samples.

4.4. Controlling the type 1 error and performing the analysis

To find the value of `lambda` that will give a 5% type 1 error rate for the analysis, we ran the function `FPR4AM()` . For its arguments, we specified the desired type 1 error rate (`falseposrate=0.05`), the number of permutations (`numreps=100`), the trait name (`trait="Bioch.HDL"`), the fixed effects part of the model (`fformula="Sex+Batch+Weight.Average"`), the phenotypic data (`pheno=phenoObj`), the genotypic data (`geno=genoObj`), the marker map (`map=mapObj`), and the number of processes (`ncpu=8`).

The function statement and its output are

```
R>  fdr <- FPR4AM(falseposrate=0.05, numreps=100,  trait = "Bioch.HDL",
+           fformula="Sex+Batch+Weight.Average",
+           pheno=phenoObj, geno=genoObj, map=mapObj, ncpu=8)

 Setting up null model.
 Calculating variance components  for null model
 Calculating extBIC for null model
```

```
 Calculating matrices that will be used in alternate model.
 Analysing 100 permuations.
\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|/-\|

 Table: Empirical false positive rates, given lambda value for model selection.

  Lambda    |  False Pos Rate
 ----------------------------
0           |  1
0.0526315   |  1
0.1052632   |  1
0.1578947   |  0.98
0.2105263   |  0.97
0.2631579   |  0.82
0.3157895   |  0.59
0.3684211   |  0.36
0.4210526   |  0.23
0.4736842   |  0.16
0.5263158   |  0.06
0.5789474   |  0.03
0.6315789   |  0.01
0.6842105   |  0
0.7368421   |  0
0.7894737   |  0
0.8421053   |  0
0.8947368   |  0
0.9473684   |  0
1           |  0
 -----------------------------
 For a false positive rate of 0.05 set the lambda parameter in
 the AM function to 0.5263158
```

The function concludes by reporting that a `labmda` value of 0.53 (rounded to two significant digits) should be used for this analysis in order to achieve a 5% false positive rate.

To perform multi-locus association mapping of the example data, the function statement and its output are

```
R>  res <- AM(trait= "Bioch.HDL",
+             fformula="Sex+Batch+Weight.Average",
+             pheno=phenoObj, geno=genoObj,  map=mapObj, ncpu=8,
+             lambda=0.53 )
```

```
                    Multiple-Locus Association Mapping
                            Version 2.4.2
```

```
. ,-"-.   ,-"-. ,-"-.   ,-"-. ,-"-. ,-"-. ,-"-.   ,-"-. ,-"-.
 X | | \ / | | X | | \ / | | X | | \ / | | X | | \ / | | X | | \ /
/ \| | |X| | |/ \| | |X| | |/ \| | |X| | |/ \| | |X| | |/ \| | |X|
   `-!-' `-!-"   `-!-' `-!-'   `-!-' `-!-"   `-!-' `-!-'   `-!-' `-
```

```
Iteration 1: Searching for most significant marker-trait association


Number of cores being used for calculation is .. 8
Significant marker-trait association found.

New results after iteration 1 are

          SNP        Chrm           Map Pos     Col Number            extBIC
        -----      ------         ---------     -----------         ---------
    Null Model                                                       1700.22

Iteration 2: Searching for most significant marker-trait association


Significant marker-trait association found.

New results after iteration 2 are

           SNP        Chrm           Map Pos     Col Number            extBIC
         -----      ------         ---------     -----------         ---------
     Null Model                                                       1700.22
     M12008C5            5        124991768          22264           1659.56


  .
  .
  .


 Iteration 7: Searching for most significant marker-trait association


Significant marker-trait association found.

New results after iteration 7 are
```

SNP	Chrm	Map Pos	Col Number	extBIC
Null Model				
M12008C5	5	124991768	22264	1659.56
M1530C6	6	17541026	23521	1652.16
M26336C1	1	171730395	5254	1644.45
M12020C5	5	125044979	22267	1643.63
M11706C10	10	125357987	41402	1643.24
M17665C4	4	134588243	18959	1644.77

Final Results

SNP	Chrm	Map Pos	Col Number	extBIC
Null Model				
M12008C5	5	124991768	22264	1659.56
M1530C6	6	17541026	23521	1652.16
M26336C1	1	171730395	5254	1644.45
M12020C5	5	125044979	22267	1643.63
M11706C10	10	125357987	41402	1643.24

```
Gamma value for model selection was set to  0.53
```

Five snp-trait associations were found. They are listed in a final results table along with their map location, column number in the marker file, and extended BIC value. The search for genes that are influencing the trait can be narrowed to the genomic regions tagged by these five SNPs.

In Figure 4, a screenshot of the Analysis page is shown of how the same analysis can be performed via the GUI. Here, a user chooses a trait for analysis, selects any fixed effects, lets **Eagle** find `lambda` by selecting the `Set automatically` option or specifies their own `lambda` value by selecting `Set manually`, and performs the analysis. The same output from the above two functions is printed in the right-half of the Analysis page.

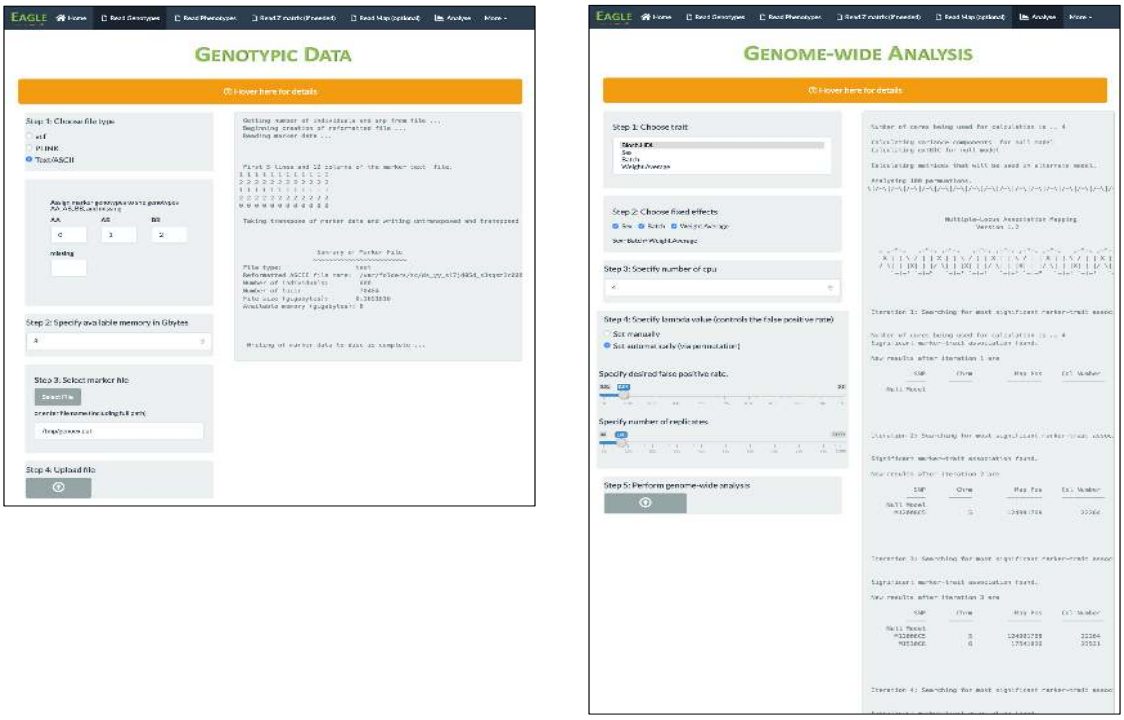

Figure 4: Screenshots of the GUI pages after the genotypic data (left) has been uploaded and the multi-locus association mapping analysis (right) performed. Output is shown in the right-half of the page.

4.5. Summarising the results

A summary of the results is produced with

```
R>  SummaryAM(AMobj=res)
```

where `res` is the list object obtained from `AM()`. Three tables are printed. These same three tables are available within the GUI by going to the Summary page (page not shown). The first table contains summary information such as the number of cpu, trait name, and number of significant snp-trait associations found. The second table gives the names and locations of the SNPs. The third table contains the effect sizes and statistical significances of the explanatory variables and the selected SNPs.

```
 Table 1: Summary Information

-------------------------------------------------------
Number cpu:                                 8
Max memory (Gb):                            8
Number of samples:                          800
Number of snp:                              70484
Trait name:                                 Bioch.HDL
```

```
Fixed model:                              Sex + Batch + Weight.Average
Number samples missing obs:               0
Number significant snp-trait assocs:      5
Lambda value for extBIC:                   0.53
-----------------------------------------------------
```

Table 2: Findings

SNP	Chr	Position	Col index
M12008C5	5	124991768.00	22264
M1530C6	6	17541026.00	23521
M26336C1	1	171730395.00	5254
M12020C5	5	125044979.00	22267
M11706C10	10	125357987.00	41402

Table 3: Size and Significance of Effects in Final Model

	Effect Size	Df	Wald Statistic	Pr(Chisq)
(Intercept)	-2.31	1	35.98	1.995E-09
SexM	1.13	1	507.16	0.000E+00
BatchOBT02	0.00	1	0.00	9.869E-01
BatchOBT03	-0.28	1	1.29	2.561E-01
...				
BatchOBT63	-0.16	1	0.43	5.117E-01
BatchOBT64	-0.11	1	0.21	6.448E-01
BatchOBT65	-0.12	1	0.21	6.491E-01
BatchOBT66	-0.20	1	0.85	3.570E-01
Weight.Average	0.05	1	148.30	0.000E+00
M12008C5	0.35	1	23.30	1.387E-06
M1530C6	0.15	1	29.76	4.878E-08
M26336C1	0.15	1	26.03	3.363E-07
M12020C5	-0.20	1	18.98	1.324E-05
M11706C10	-0.23	1	17.89	2.346E-05

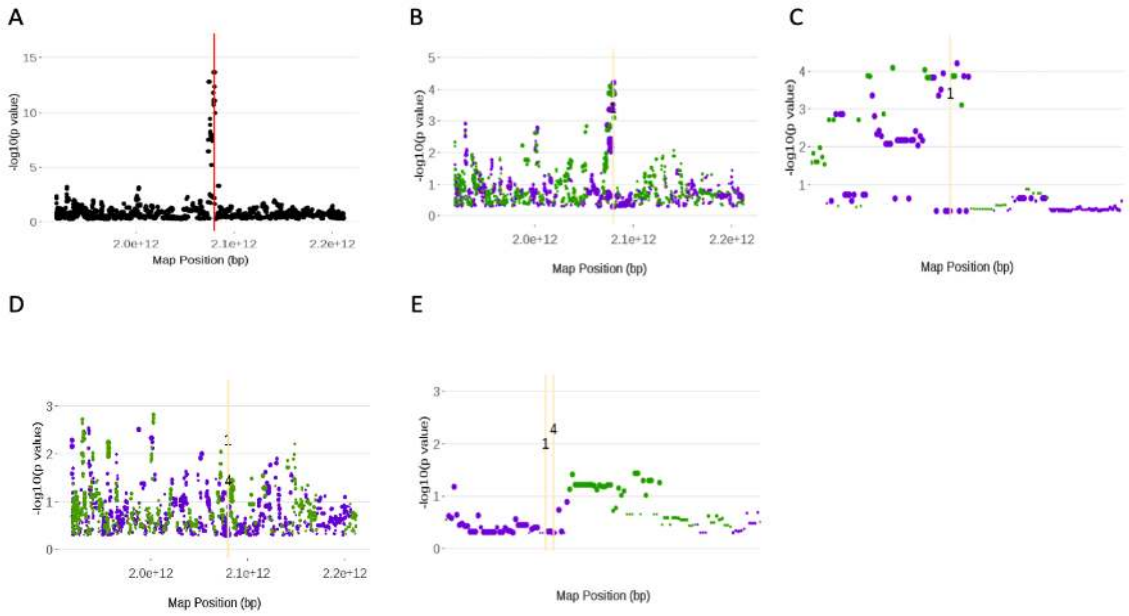

Figure 5: Screenshots of the plots from running the **Eagle** function `PlotAM`. All plots are Manhattan plots and are of chromosome 5. These plots show how the significance of the SNPs change throughout the model building process. The red vertical line is the position of the SNP that has the largest score statistic and strongest association with the trait at that iteration. The orange vertical lines are the positions of SNPs found in previous iterations to be in strongest association with the trait. Green (purple) points denote SNPs that have increased (decreased) in significance from the previous iteration. The size of the point is proportional to the size of the change in significance. (A), (B), and (D) are plots of the first, second, and fifth iteration, respectively, of the model building process. (C) and (E) were created from (B) and (D) respectively by using `PlotAM`'s interactive zoom feature.

4.6. Visualising the findings

Suppose we are interested in viewing how the pattern of significance varies throughout the model building process. Here, we focus on chromosome 5. This chromosome is interesting because it has two closely linked regions housing genes underlying the trait.

Using the function statement

```
R> PlotAM(AMobj=res, itnum=1, chr = "5", type = "Manhattan")
```

the resulting plot, for chromosome 5 and iteration 1, is shown in Figure 5A. Each point is a measure of the strength of association between a SNP and trait. The measure is calculated as the $-log_{10}$ of the p-value of the score statistic (Section 2.4) for the SNP, since `type="Manhattan"`. There is a clear spike towards the middle of the chromosome. In fact, the SNP in strongest association across the entire genome, at iteration 1, was on chromosome 5. Its position is given by the red vertical line.

By the end of the second iteration of the model building process, the SNP which was identified in the first iteration, has been found to be significant. Its effect has been moved from the random to the fixed effects part of the model. This change impacts the significance of the other SNPs. By using the above command but with `itnum` set to 2, the plot in Figure 5B is generated. Here, the SNPs that have increased (decreased) in significance are denoted by green (purple) points. The size of the point is proportional to the size of the change in significance from the previous iteration. Unsurprisingly, the largest changes have occurred around the SNP whose effect is now being treated as a fixed effect. We can see this more clearly by using the zoom feature in `PlotAM()` to focus on the region around the SNP of interest (Figure 5C).

What is interesting about Figure 5C is that there are still several SNPs in strong association with the trait. This suggests that there may be other statistically significant SNP-trait associations here. This is in fact the case, because by the fifth iteration, a second SNP has been found and fitted as a fixed effect. The pattern of association is shown in Figure 5D with the same zoomed region as before shown in Figure 5E. The drop in significance between fitting a single SNP in this region as a fixed effect to fitting two closely linked SNPs as fixed effects is apparent when you compare figures 5C and 5E, noting the change in scales of the y-axes.

5. Summary

The **Eagle** package has been created to make genome-wide multi-locus association mapping easy. The package accepts marker data in different formats, has easy-to-use functions, comes with a user-friendly GUI, and has an interactive plotting function for visualising the model building process. We welcome feedback via `eaglehelp@csiro.au` from users on how the functionality and usability of the package could be even further improved. As we saw in the example, **Eagle** brings clarity to situations where there are tightly linked SNPs in association with a trait. It can also uncover significant SNP-trait associations that are otherwise hidden to single-locus association mapping. At the very least, **Eagle** compliments single-locus association mapping. Ultimately though, with the aid of **Eagle**, our hope is that the genetics community will shift to multi-locus association mapping as the method-of-choice for the genome-wide analysis of association data.

References

Anderson E, Bai Z, Bischof C, Blackford SL, Demmel J, Dongarra J, Croz JD, Greenbaum A, Hammarling S, McKenney A, *et al.* (1999). *LAPACK Users' Guide*. SIAM. doi:10.1137/1.9780898719604.

Blackford SL, Petitet A, Pozo R, Remington K, Whaley CR, Demmel J, Dongarra J, Duff I, Hammarling S, Henry G, *et al.* (2002). "An Updated Set of Basic Linear Algebra Subprograms (BLAS)." *ACM Transactions on Mathematical Software*, **28**(2), 135–151. doi:10.1145/567806.567807.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). "TAS-

SEL: Software for Association Mapping of Complex Traits in Diverse Samples." *Bioinformatics*, **23**(19), 2633–2635. doi:10.1093/bioinformatics/btm308.

Chen J, Chen Z (2008). "Extended Bayesian Information Criteria for Model Selection With Large Model Spaces." *Biometrika*, **95**(3), 759–771. doi:10.1093/biomet/asn034.

Dagum Leonardo MR (1998). "OpenMP: An Industry Standard API for Shared-Memory Programming." *IEEE Computational Science and Engineering*, **5**(1), 46–55. doi:10.1109/99.660313.

Friedman J, Hastie T, Tibshirani R (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software*, **33**(1), 1–22. doi:10.18637/jss.v033.i01.

George AW, Verbyla A, Bowden J (2020). "Eagle: multi-locus association mapping on a genome-wide scale made routine." *Bioinformatics*, **36**(5), 1509–1516. doi:10.1093/bioinformatics/btz759.

Huang M, Liu X, Zhou Y, Summers RM, Zhang Z (2019). "BLINK: a Package for the Next Level of Genome-Wide Association Studies With Both Individuals and Markers in the Millions." *GigaScience*, **8**(2), giy154. doi:10.1093/gigascience/giy154.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012). "GAPIT: Genome Association and Prediction Integrated Tool." *Bioinformatics*, **28**(18), 2397–2399. doi:10.1093/bioinformatics/bts444.

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016). "Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies." *PLoS genetics*, **12**(2), e1005767. doi:10.1371/journal.pgen.1005767.

Nicod J, Davies RW, Cai N, Hassett C, Goodstadt L, Cosgrove C, Yee BK, Lionikaite V, McIntyre RE, Remme CA, *et al.* (2016). "Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing." *Nature Genetics*. doi:10.1038/ng.3595.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, *et al.* (2007). "PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics*, **81**(3), 559–575. doi:10.1086/519795.

Rakitsch B, Lippert C, Stegle O, Borgwardt K (2013). "A Lasso Multi-Marker Mixed Model for Association Mapping with Population Structure Correction." *Bioinformatics*, **29**(2), 206–214. doi:10.1093/bioinformatics/bts669.

Robinson G (1991). "That BLUP is a Good Thing: the Estimation of Random Effects." *Statistical Science*, **6**, 15–51. doi:10.1214/ss/1177011926.

Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Réale D, Dochtermann NA, Garamszegi LZ, Araya-Ajoy YG (2020). "Robustness of linear mixed-effects models to violations of distributional assumptions." *Methods in Ecology and Evolution*, **11**(9), 1141–1152. doi:10.1111/2041-210X.13434.

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, Nordborg M (2012). "An efficient Multi-Locus Mixed-Model Approach for Genome-Wide Aassociation Studies in Structured Populations." *Nature Genetics*, **44**(7), 825–830. doi:10.1038/ng.2314.

Shen X, Alam M, Fikse F, Rönnegård L (2013). "A Novel Generalized Ridge Regression Method for Quantitative Genetics." *Genetics*, **193**(4), 1255–1268. doi:10.1534/genetics.112.146720.

Verbyla AP, Cavanagh CR, Verbyla KL (2014). "Whole-Genome Analysis of Multienvironment or Multitrait QTL in MAGIC." *G3: Genes, Genomes, Genetics*, **4**(9), 1569–1584. doi:10.1534/g3.114.012971.

Verbyla AP, Taylor JD, Verbyla KL (2012). "RWGAIM: an Efficient High-Dimensional Random Whole Genome Average (QTL) Interval Mapping Approach." *Genetics Research*, **94**(6), 291–306. doi:10.1017/S0016672312000493.

Wang J, Zhang Z (2018). "GAPIT version 3: An interactive analytical tool for genomic association and prediction." Preprint on webpage at https://www.researchgate.net/publication/329829469_GAPIT_Version_3_An_Interactive_Analytical_Tool_for_Genomic_Association_and_Prediction.

Wang SB, Feng JY, Ren WL, Huang B, Zhou L, Wen YJ, Zhang J, Dunwell JM, Xu S, Zhang YM (2016). "Improving Power and Accuracy of Genome-Wide Association Studies Via a Multi-Locus Mixed Linear Model Methodology." *Scientific reports*, **6**, 19444. doi:10.1038/srep19444.

Wen YJ, Zhang H, Ni YL, Huang B, Zhang J, Feng JY, Wang SB, Dunwell JM, Zhang YM, Wu R (2018). "Methodological Implementation of Mixed Linear Models in Multi-Locus Genome-Wide Association Studies." *Briefings in bioinformatics*, **19**(4), 700–712. doi:10.1093/bib/bbw145.

Whittaker JC, Thompson R, Denham MC (2000). "Marker-Assisted Selection using Ridge Regression." *Genetics Research*, **75**(2), 249–252. doi:10.1017/s0016672399004462.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, *et al.* (2006). "A Unified Mixed-Model Method for Association Mapping that Accounts for Multiple Levels of Relatedness." *Nature Genetics*, **38**(2), 203. doi:10.1038/ng1702.

Zhang YM, Jia Z, Dunwell JM (2019). "The Applications of New Multi-Locus GWAS Methodologies in the Genetic Dissection of Complex Traits." *Frontiers in Plant Science*, **10**, 100. doi:10.3389/fpls.2019.00100.

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, *et al.* (2007). "An Arabidopsis Example of Association Mapping in Structured Samples." *PLoS Genetics*, **3**(1), e4. doi:10.1371/journal.pgen.0030004.

Affiliation:

Andrew W. George
Data61

Commonwealth Scientific and Industrial Research Organisiation
Brisbane, QLD 4102, Australia
E-mail: Andrew.George@csiro.au