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

Andrew W. George and Joshua Bowden

November 22, 2016

Abstract

Since the earliest of genome-wide association studies, a key shortcoming in how their data have been analysed has persisted. The strength of association between a marker locus and trait is measured for each locus separately, on a locus-by-locus basis. Multiple-locus methods that map multiple locus-trait associations simultaneously have been available for some time. However, they have attracted little attention. They can be demanding, computationally, and their results are not always easy to interpret. Yet, it is widely accepted that multiple-locus methods are superior, statistically, to locus-by-locus methods. Here, we present our method that makes the multiple-locus analysis of data from genome wide association studies routine. It is formulated within a linear mixed model framework. We call our method AMplus. AMplus produces results faster than competing multiple-locus methods and often with greater statistical power. Also, it is just a little slower than the fastest single-locus linear mixed model implementations. AMplus is freely available as a fully documented R package.

1 Introduction

Over the past decade, Genome wide association studies have changed considerably in both their analysis and design. Early gwas followed a case-control design. Association mapping methods were no more complicated than contingency table tests or regression. These designs though had a tendency to yield spurious findings if there was unrecognised population stratification. This prompted a shift towards family-based designs and score tests, such as the tdt test and its variants (refs). Today, instead of by design, it is through statistical modelling that we account for the effects of population stratification. This has meant that data can be collected from general populations, even if these populations are highly structured. Analysis via sophisticated association mapping methods based on linear mixed models is now almost routine.

What has not changed over the past decade is that it remains common practice to analyse gwas data on a locus by locus basis. This is despite there being several significant problems with analysing the data in this way. First, the aim of association mapping is to identify regions of the genome that house genes that are influencing a trait. The identification of these regions from these analyses is not always straightforward. Gwas results are reported, typically, via Manhattan plots that plot the -log10 of the p value for each locus against the map position of the locus. The location of peaks in this plot signify regions of interest. Inferring the exact number of regions of interest can be difficult If the peaks are not well separated. Second, when multiple statistical tests are performed, the probability of wrongly accepting a significant result (type 1 error) is inflated. This is known as the multiple testing problem (refs). Many different solutions have been offered (refs). Yet, there is still no well accepted way of correcting for multiple testing in the context of genome- wide association mapping. Third, many of the traits whose genetic secretes we are trying

to discover are complex. There will be multiple loci in linkage equilibrium with genes that influence the trait. Yet, A locus by locus mapping approach only assesses the evidence for association between a single locus and trait.

It is somewhat surprising then that multiple locus methods haven't attracted more attention. It is possible to measure all locus-trait associations jointly with methods based on the regularisation techniques, ridge regression and lasso. Here, multiple testing is not an issue. These techniques though are computationally demanding. Also, the results are not easy to interpret. The strength of association is not measured by a p-value but by the size of the regression coefficient for a locus in the model. More recently, associations have started to be mapped with random forests (refs). Similar to regularisation techniques though, it is not clear how to infer genomic regions of interest from their findings (refs). An multiple locus method that does show promise is the multi-locus linear mixed model method (ref). The best multiple locus model is built with simple variable selection. Results are immediately interpretable but here, computation can be a challenge for large data.

Here, we present our new multiple locus method for genome wide association mapping, which we are calling AMplus. AMplus combines the strength of regularisation techniques (being able to fit all locus-trait associations jointly), with simple variable selection (having easy to interpret results). Our method does not require a significance threshold to be set nor regularisation parameters to be fine tuned. Through a clever dimension reduction step, we are able to achieve a computational performance similar to the fastest single locus linear mixed model implementations. AMplus is statistically more powerful than single locus association mapping and is as and often more powerful than most multiple locus methods. Our aim is to make multiple locus association mapping on a genome wide scale routine. To this end, we have created a fully documented R package, that is easy to use, even for non R users. Our package accepts marker data of different formats and can handle data larger

than a computer's memory capacity. It includes detailed error checking and makes heavy use of distributed computing for computation when available. The purpose of this work was to make multiple-locus association mapping on a genome-wide scale, for large data sets, practical and we have built AMplus accordingly.

2 Results

2.1 Association Mapping Methods

We compared AMplus, in terms of computational and statistical performance, against seven other association mapping methods. We chose methods that 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 software. 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 and FaST-LMM because of their popularity and computational speed. For multiple locus association mapping, we chose bigRR, glmnet, LMM-Lasso, MLMM, and r2VIM. Each takes a different approach to mapping multiple locus-trait associations jointly.

From a practical perspective, how the methods were implemented, in terms of usability, varied greatly. In Table 1, we list seven important features for ensuring the usability of computer software for association mapping. We also identify which software has what features. In forming this list of features, we assumed that the primary purpose of a genome-wide association study is to identify the set of marker loci in true and strongest association with a trait. The level of documentation varied across the implementations but only a few came with user manuals detailing the format of the input data and how to perform analyses. Most implementations could handle

additional fixed effects which is important when accounting for hidden population structure. Few though could cope with marker data larger than the memory capacity of the computer. Similarly, few returned the best set of marker loci in strongest association with a trait as their result. Almost all the software returned results that required further processing, either by identifying the best set of marker loci via the setting of significance thresholds or having to do further analysis to calculate the significance of the results. Almost none of the association mapping software checked for errors in the input data. Only AMplus has all seven features.

2.2 Computational Performance: run times and memory usage

To asses the computational performance of AMplus, we conducted a large simulation study. We were interested in the impact of study size on performance so we generated data under five different scenarios. These are a GWAS of size 150 individuals and 5000 snp (150 X 5K), 350 individuals and 400000 snp (350 X 400K), 1500 individuals and 50000 snp (1500 x 50K), 2000 individuals and 500000 snp (2000 x 500K), 4000 individuals and 1500000 snp (4000 x 1.5M), and 10000 individuals and 1500000 snp (10000 x 1.5M). We chose these scenarios to reflect some of the different sized GWAS being performed in animals, plants, and humans. For each scenario, we generated 100 replicates of data. A single replicate consists of snp genotype data and quantitative trait data. We obtained the snp data from the publicly available 1000 genome project (phase 3). The quantitative trait data we generated from the snp data by selecting a set of snp loci, assigning allelic effects to these snp, and aggregating these effects for each individual along with random error. The number of snp selected per replicate follows a Poisson distribution with mean 30. The quantitative trait was generated to have a heritability of 50%. Analyses were performed on a high end desktop computer.

It had dual 8-core Xeon processors, three Kepler Tesla GPUs, and 128 Giggabytes of RAM. All implementations except, GEMMA, made use of distributed computing, either explicitly or implicitly through multi-threaded BLAS/LAPACK libraries.

The run times for AMplus against the other software programs, across the five scenarios, is shown in figure 2.2 To help with interpretibility of the results, in both plots, we have taken the x and y axes to be on a log scale. This means a unit change on the x or y axes is equivalent to a change in the order of magnitude. In the top plot, a point is the median of the ratios of elapse times of the multiple locus method to AMplus for a given scenario. The median is over the 100 replicates. Here, since the median of the ratios are all positive on the log scale, it means that AMplus had a shorter run time than than all the other multiple locus methods. In fact, in some cases, AMplus was over a hundred times, or over two orders of magnitude, faster. Unlike AMplus, the size of data under scenario 10000 X 1.5M was beyond the memory constraints of the other multiple locus implementations. This was also the case for LMM-Lasso, bigRR, and glmnet, for scenario 4000 x 1.5M.

In the bottom plot of figure 2.2, we compare the median run time of AMplus against the median run times of the single locus methods, FaST-LMM and GEMMA. FaST-LMM was run in two ways. It was run where the genetic similarity matrix was built with all the marker data (FaST-LMM^{a}ll) or with data on every 500th snp (FaST-LMM^{f}ew). AMplus was also run in two ways. The default behavior for AMplus is to make use of CPUs for computing. However, AMplus also has the capacity to harness multiple GPUs (AMplus^{G}PU). From the bottom plot, all the implementations have short run times when analysing data from scenario 150x5K. However, for the other scenarios, AMplus and AMplusGPU have significantly shorter run times than GEMMA and FaST-LMM^{a}ll. FaST-LMM^{f}ew has a shorter run time than AMplus and AMplus^{G}PU but only for scenarios 1500x50K and 10000x1.5M. Furthermore, for scenario 10000x1.5M, the median run time for GEMMA is 4071

Figure 1: Run times

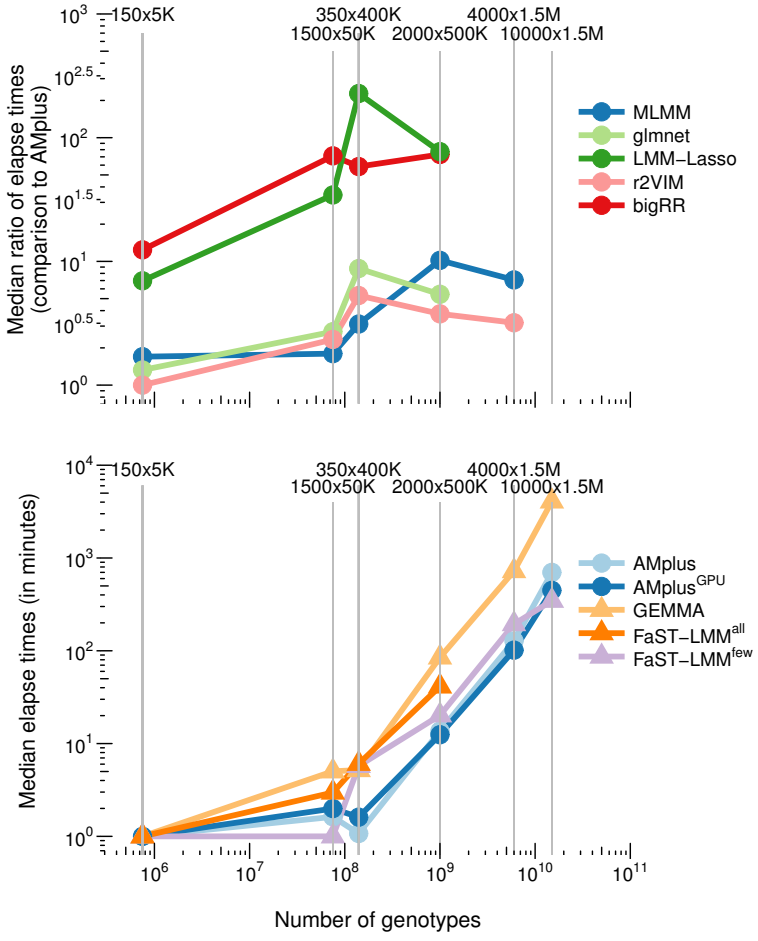

Figure 2: Memory usage

minutes, for AMplus is 699 minutes, for AMplus^{GPU} is 447 minutes, and for FaST-LMM^{few} is 346 minutes. The very fast single-locus method FaST-LMM^{few} is only 29% faster than our multiple-locus method AMplus^{GPU}. It is worth noting though that there is a setup cost to accessing GPU computing, making AMplus^{GPU} most efficient on the larger data.

The memory usage of the software programs is shown in Figure 2.2.

2.3 Statistical Performance: power and FDR

2.4 Application to Arabidopsis data

Figure 3: Power curves