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

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

[1]Data61, CSIRO, Australia.
[2]IM &T, CSIRO, Australia.

October 3, 2018

Supplementary Table 1: Implementation and methodology attributes of eight computer programs/packages for genome-wide association mapping.

Attributes	**Eagle**	**bigRR**	**glmnet**	**LMM-Lasso**	**MLMM**	**r2VIM**	**FaST-LMM**	**GEMMA**
Implementation								
Purpose built[a]	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Language	R/C++	R	R	Python	R	R	C++ and Python[b]	C++
GUI	Yes	No	No	No	No	No	No	No
Documentation	Videos, user-manuals, website, R help	R help	Vignettes, R help	Readme.txt, test script	Vignette, R help	R help	Videos, website user-manuals,	User-manual, website
Additional fixed effects[c]	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Types of trait data	Cont.	Cont., binary, count	Cont., binary, count	Cont.	Cont.	Cont., binary	Cont.	Cont., binary
Data larger than memory	Yes	No	No	No	No	No	Yes	No
Methodology								
Model [d]	LMM	HEM	GLMM	LMM	LMM	RF	LMM	LMM, mvLMM Bayesian Sparse LMM
SNPs fitted [e]	All/multiple	All	All	All	Multiple	Multiple	Single	Single
Selection type	Model	Variable	Variable	Variable	Model	Variable	Variable	Variable
Threshold free	Yes	No	No	No	Yes	No	No	No

[a] Specifically created for the analysis of GWAS data.

[b] Separate programs, one written in Python, the other C++

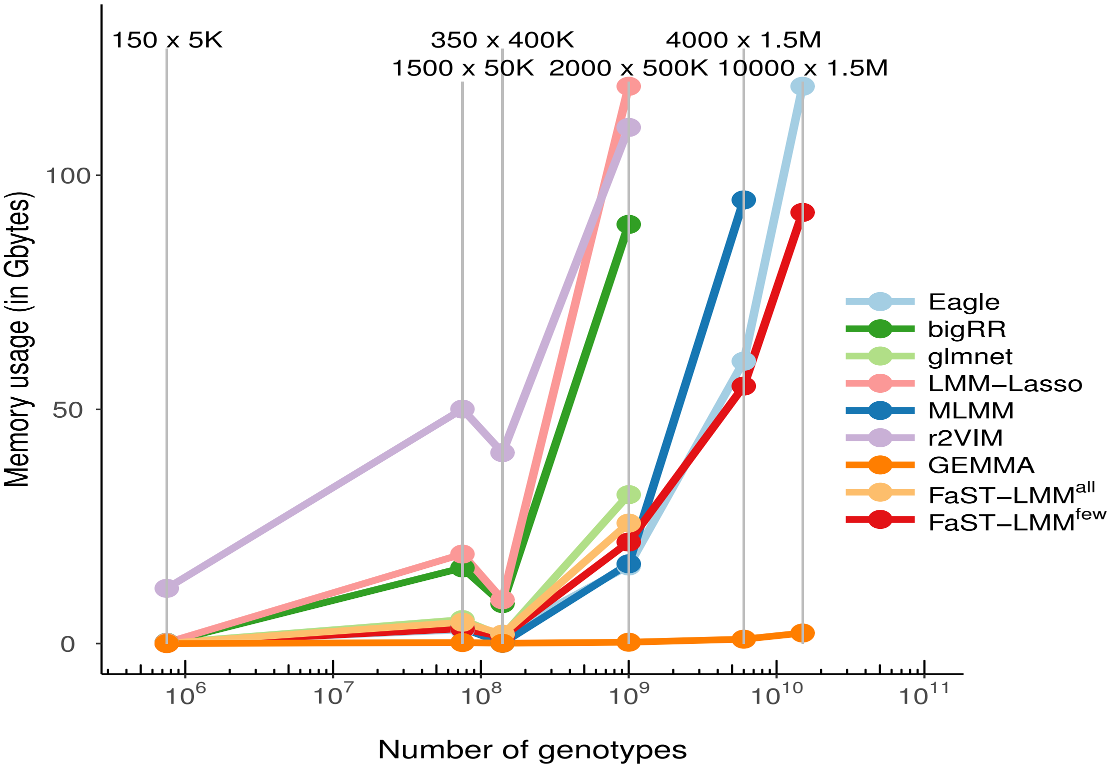
[c] Capacity for additional fixed effects (such as age, sex, and/or population structure effects) to be included directly in the model.

[d] For the different types of model, LMM is linear mixed model. GLM is generalised linear model., GLMM is generalised linear mixed model, and RF is random forests.

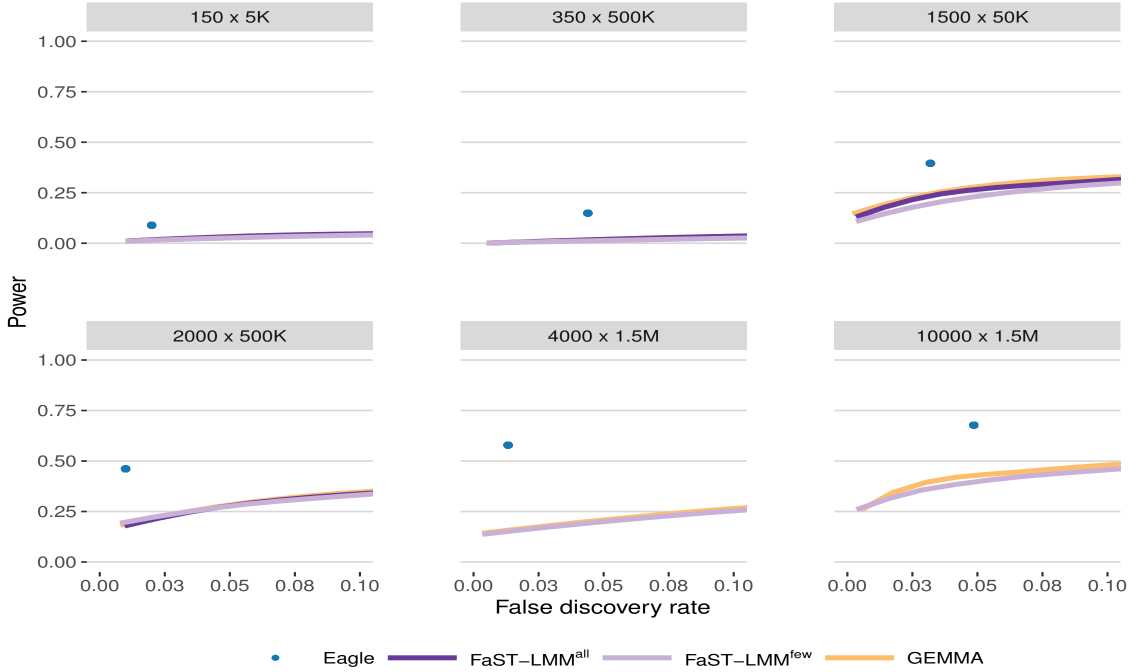
[e] Association is assessed a SNP at a time (single), for multiple SNPs (multiple), or for all SNPs (all). Eagle fits all SNPs but also identifies multiple SNPs (All/multiple) in association with the trait.

Supplementary Figure 1: Memory usage (in gigabytes) of Eagle and the other association mapping programs/packages across the six simulation scenarios. The maximum amount of memory on the computer is 128 gigabytes. The x-axis is on the log scale. GEMMA, a single-locus implementation, had the lowest memory usage. Of the multiple-locus implementations, Eagle had the lowest memory usage. Also, it was the only multiple-locus implementation able to produce results for data under scenario 10000 x 1.5M. This is due to its ability to handle data larger than the available memory of a computer. FaST-LMM was run where all the SNP data are used to estimate the relationship matrix (FaST-LMM^{all}) and where genotype data from every five-hundredth SNP are used to estimate the relationship matrix (FaST-LMM^{few})

.



Supplementary Figure 2: Power verse false discovery rates for Eagle and the single-locus methods GEMMA and FaST-LMM. FaST-LMM was run where all the SNP data are used to estimate the relationship matrix (FaST-LMM^{all}) and where genotype data from every five-hundredth SNP are used to estimate the relationship matrix (FaST-LMM^{few}). Eagle has substantially higher power than the single-locus methods.


150 x 5K
350 x 500K
1500 x 50K
2000 x 500K
4000 x 1.5M
10000 x 1.5M
Power
False discovery rate
Eagle
FaST−LMM^all
FaST−LMM^few
GEMMA