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.

July 10, 2018

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

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, user-manuals, website	User-manual, website
Additional fixed effects[c]	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes adsf
Types of trait data	Cont.	Cont., binary, count	Cont., binary, count	Cont.	Cont.	Cont.	Cont.	Cont.
Data larger than memory	Yes	No	No	No	No	No	Yes	No
Methodology								
Model	LMM		GLMM	GLM	LMM	RF	LMM	LMM
SNPs fitted	All/multiple	All	All	All	Multiple	Multiple	Single	Single
Selection type	Model	Variable	Variable	Variable	Model	Variable	Variable	Variable
Threshold free[5]	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.

Supplementary Table 2: Some of the key features possessed by Eagle and the other seven computer programs/packages for association mapping

	Computer Programs/Packages for Association Mapping							
	Multiple-locus						Single-locus	
Features	Eagle	bigRR	glmnet	LMM-Lasso	MLMM	r2VIM	FaST-LMM	GEMMA
Purpose built[1]	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Well documented[2]	Yes	No	Yes	No	Yes	No	Yes	Yes
Simultaneous fitting of SNPs	Yes	Yes	Yes	Yes	No	No	No	No
Additional fixed effects[3]	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Data larger than RAM[4]	Yes	No	No	No	No	No	Yes	No
Threshold free[5]	Yes	No	No	No	Yes	No	No	No
Informative error checking	Yes	No	No	No	No	No	Yes	No

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

[2] More than just a readme file or comments in an example file. Programs/packages with a tick had a detailed user manual.

[3] Ability to handle additional fixed effects such as age, sex, and/or population structure effects.

[4] Able to deal with data larger than the memory capacity of the computer.

[5] Results reported as the set of SNP closest to the genes influencing a trait. No need to construct thresholds to determine significance of the findings.

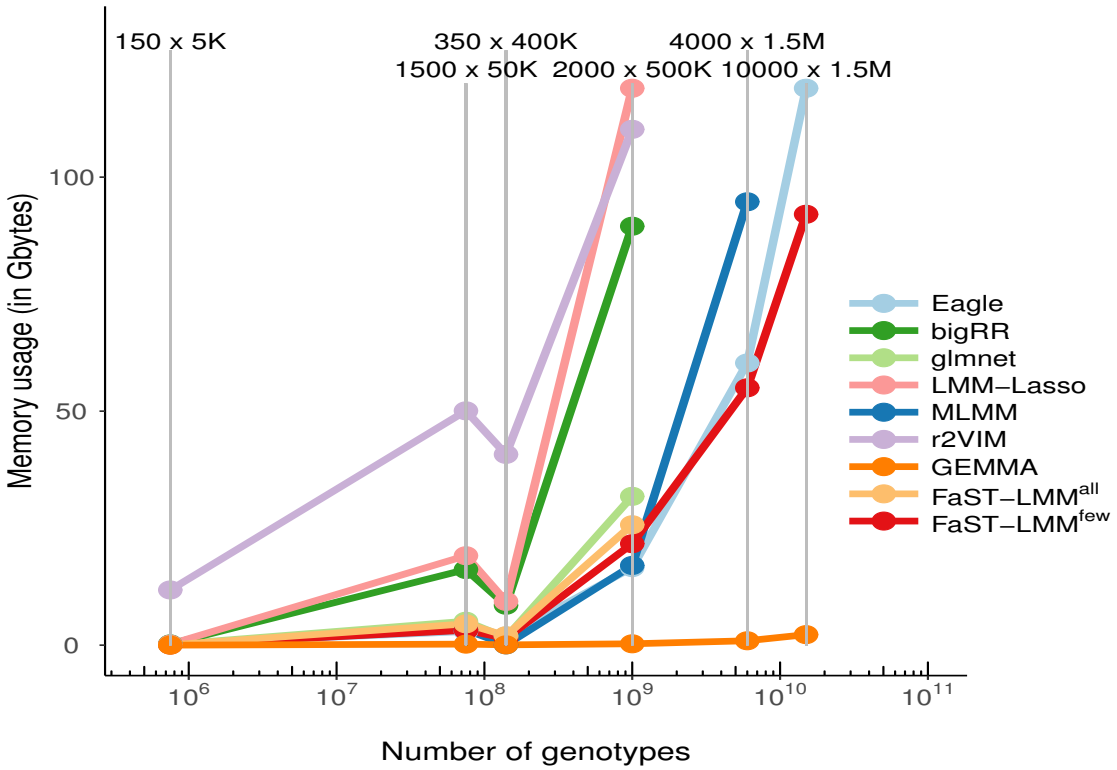
[6] All the programs terminated on errors. However, not all the programs informed the user of the cause and how to fix the errors.

Supplementary Table 3: The median run times (in minutes) of Eagle and the other association mapping programs across the six simulation scenarios.

		Simulation Scenarios					
Method	Name	150 x 5K	1500 x 50K	350 x 400K	2000 x 500K	4000 x 1.5M	10000 x 1
Multiple	Eagle	0.08	1.62	**2.71**	**13.65**	**127.63**	699.5
	MLMM	0.15	2.91	19.04	143.01	870.84	
	glmnet	0.11	3.95	14.06	74.03		
	r2VIM	0.09	3.66	5.51	50.59	380.52	
	bigRR	1.01	113.35	54.99	1030.61		
	LMM-Lasso	0.57	52.08	92.20	1031.85		
Single	GEMMA	0.02	5.02	6.17	84.83	723.33	4071.6
	FaST-LMM^{few}	**0.01**	**0.80**	7.07	20.16	193.90	**346.1**
	FaST-LMM^{all}	0.03	2.96	7.90	41.27		

Supplementary Figure 1: Memory usage (in gigabytes) of Eagle and the other association mapping programs 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.