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

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

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

June 27, 2018

Table 1: A summary of the features possessed by the Eagle package and the comparison implementations.

	Computer Software for Association Mapping							
	Multiple-locus						Single-loc	
Features	Eagle	bigRR	glmnet	LMM-Lasso	MLMM	r2VIM	FaST-LMM	C
Purpose built[1]	✓	✓	✗	✓	✓	✓	✓	
Well documented[2]	✓	✗	✓	✗	✓	✗	✓	
Simultaneous fitting of SNPs	✓	✓	✓	✓	✗	✗	✗	
Additional fixed effects[3]	✓	✓	✓	✗	✓	✓	✓	
Data larger than RAM[4]	✓	✗	✗	✗	✗	✗	✓	
Threshold free[5]	✓	✗	✗	✗	✓	✗	✗	
Informative error checking	✓	✗	✗	✗	✗	✗	✓	

[1] Computer software specifically designed for the analysis of data from GWAS.

[2] More than just a readme file or comments in an example file. Those programs with ticks had detailed user manuals.

[3] Ability to accommodate additional fixed effects in the model such as age, sex, and 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.

Table 2: 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		

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})
.