genipe: An automated genome-wide imputation pipeline with automatic reporting and statistical tools

Supplementary Material

Lemieux Perreault, L.-P. Legault, M.-A. Asselin, G. Dubé, M.-P.

Table S1: Imputation steps performed by *genipe*. The majority of steps are parallelized per chromosome or per genomic segments.

	Step	**Program**	**Parallel**
1	Initial marker filtering	PLINK	No
2	Missing rate	PLINK	No
3	Split by chromosome	PLINK	Yes (chromosome)
4	Check strand	SHAPEIT	Yes (chromosome)
5	Flip	PLINK	Yes (chromosome)
6	Final check strand	SHAPEIT	Yes (chromosome)
7	Final exclusion	PLINK	Yes (chromosome)
8	Phasing	SHAPEIT	Yes (chromosome)
9	Imputation	IMPUTE2	Yes (5Mb segments)
10	Cross validation statistics	genipe	No
11	Merge imputed segments	genipe	Yes (chromosome)
12	Compression (optional)	BGZIP	Yes (chromosome)
13	Imputation statistics and MAF	genipe	No

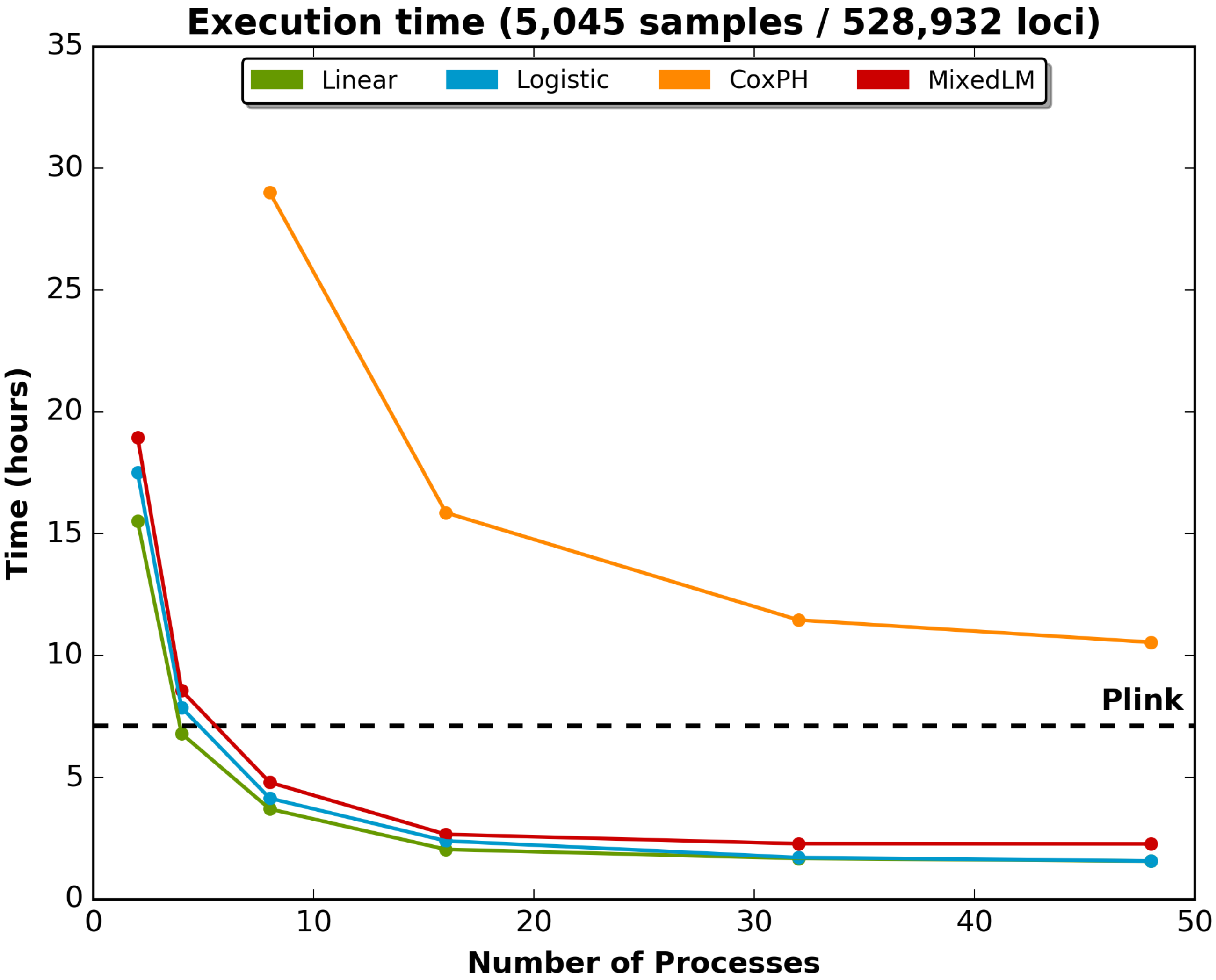

Figure S1: Execution time for typical imputation analysis. Imputation was performed on chromosome 2 for 5,045 samples using *genipe*. A total of 1,170,797 loci were imputed, where 961,019 (82.1%) had sufficient imputation quality. Statistics were computed on loci with minor allele frequency higher than 1% (a total of 528,932 loci). The black dashed line is the execution time for Plink (logistic regression on a single process). The four models (linear, logistic, Cox's proportional hazard and mixed linear model [ten repeated measurements]) were executed using 2, 4, 8, 16, 32 and 48 processes. Cox's proportional hazard analysis was not performed on 2 and 4 processes to save time. An optimization was made so that the linear mixed model could perform as well as a linear or logistic regression. This optimization is the two-step linear mixed model [1]. If the estimated p-value is lower than a user-specified threshold, the standard linear mixed model is used to gather all the required statistics. Figure S2 and S3 show the correlation between the estimated p-value and the real one.

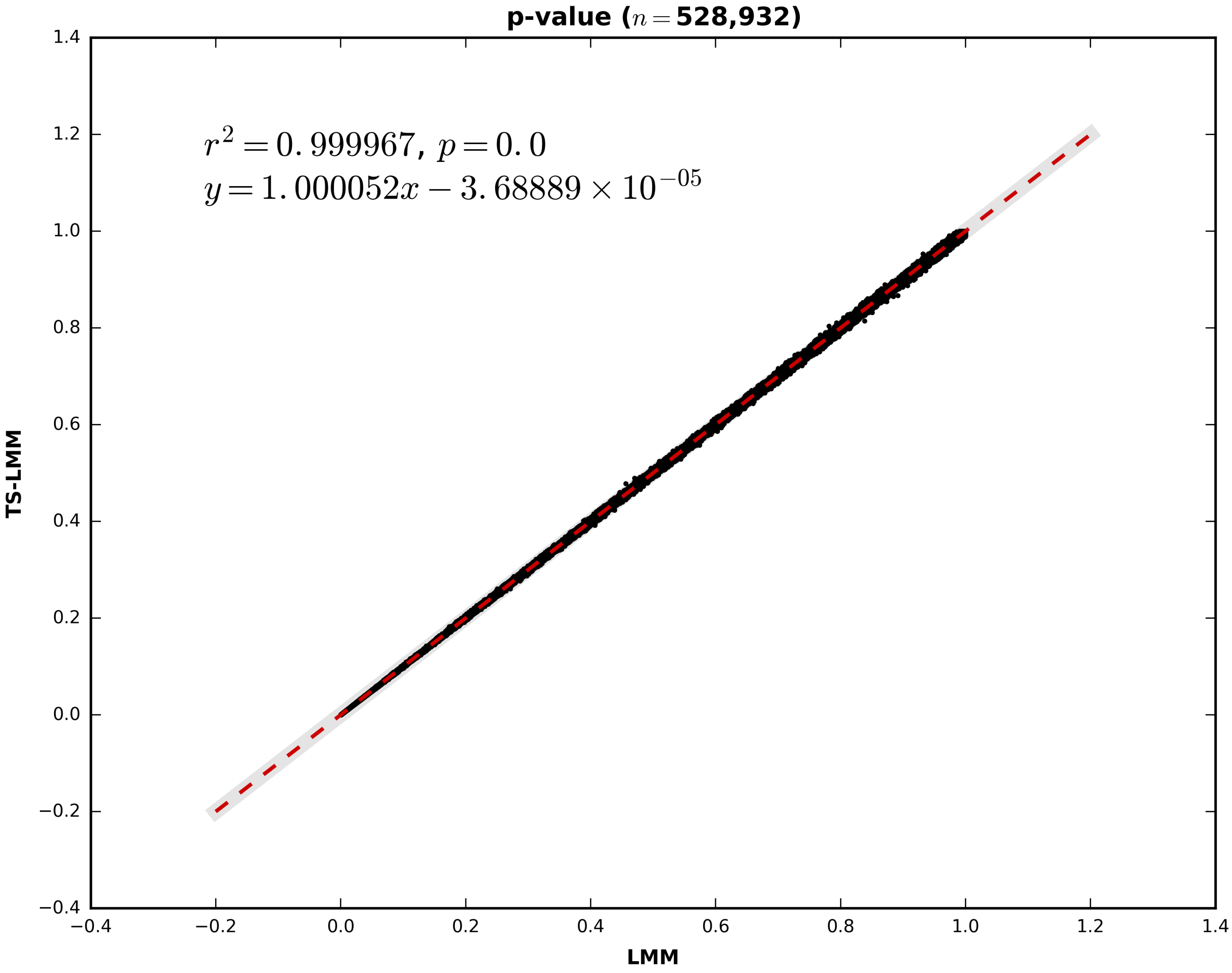

Figure S2: Correlation of the p-values between the standard and the two-step linear mixed models. The standard and two-step linear mixed models were used on the same dataset (*i.e.* 5,045 samples (ten repeated measurements) imputed on chromosome 2, where 528,932 loci had sufficient imputation quality and a minor allele frequency higher than 1%). Each dot represents a p-value. The light-gray bar is the identity line ($y = x$). The red dashed line is the estimated slope of the linear regression (equation at the top-left). The Pearson correlation (r^2) was 0.999967.

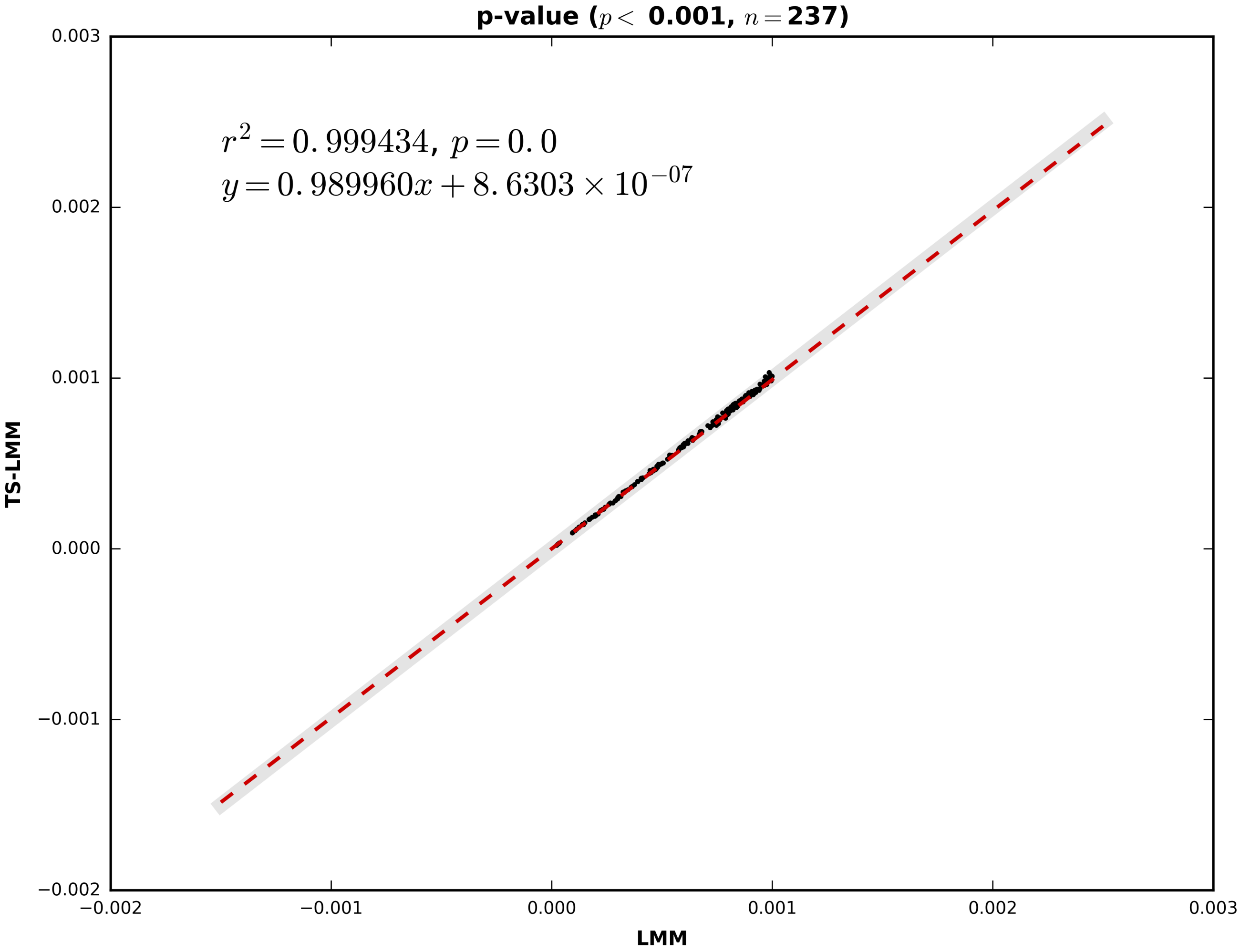

Figure S3: Correlation of the p-values ($< 1 \times 10^{-3}$) between the standard and the two-step linear mixed models. The standard and two-step linear mixed models were used on the same dataset (*i.e.* 5,045 samples (ten repeated measurements) imputed on chromosome 2, where 528,932 loci had sufficient imputation quality and a minor allele frequency higher than 1%). A total of 237 loci had a p-value lower than 1×10^{-3}. Each dot represents a p-value. The light-gray bar is the identity line ($y = x$). The red dashed line is the estimated slope of the linear regression (equation at the top-left). The Pearson correlation (r^2) was 0.999434.

References

[1] Sikorska K, Montazeri NM, Uitterlinden A, Rivadeneira F, Eilers PH, Lesaffre E: **GWAS with longitudinal phenotypes: performance of approximate procedures**. *European Journal of Human Genetics* 2015, **23**(10):1384–1391.