Supplementary material for: rTASSEL: an R interface to TASSEL for association mapping of complex traits

Brandon Monier[1], Terry M. Casstevens[1], Peter J. Bradbury[1,2], Edward S. Buckler[1,2]

1. Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853
2. United States Department of Agriculture-Agricultural Research Service, Robert W. Holley Center for Agriculture and Health, Ithaca, NY 14853

Methods

To achieve benchmarking results, several data sets were used. Genotypic and phenotypic maize data consisting of 279 samples, 3093 variant sites, and 1 measured trait were utilized for the analysis of variant call format (VCF) import, generalized linear model (GLM) association, mixed linear model (MLM) association, and kinship generation times (Flint-Garcia *et al.*, 2005). To illustrate the effectiveness of the fast association method, 100 simulated RNA expression traits for the prior genotype data was used. Trait data was generated using the `makeExampleDESeqDataSet()` function from the R package `DESeq2` (Love *et al.*, 2014). A larger genotypic data set consisting of 1,210 samples and 2,255,405 variant sites was also utilized for large VCF import and kinship generation times. All benchmarks were generated using the `microbenchmark()` function from the R package `microbenchmark` (Mersmann, 2019).

All benchmarks sans large VCF import and kinship generation times were evaluated 100 times and recorded on a workstation running 16 GB of RAM and 4 cores on an Intel® Core™ i5-6500 CPU with a clock speed of 3.20 GHz and. Large VCF import and kinship generation benchmarks were evaluated 10 times and recorded on a workstation running 256 GB of RAM and 12 cores on an Intel® Xeon® CPU E5-2643 v3 with a clock speed of 3.40GHz.

Figures

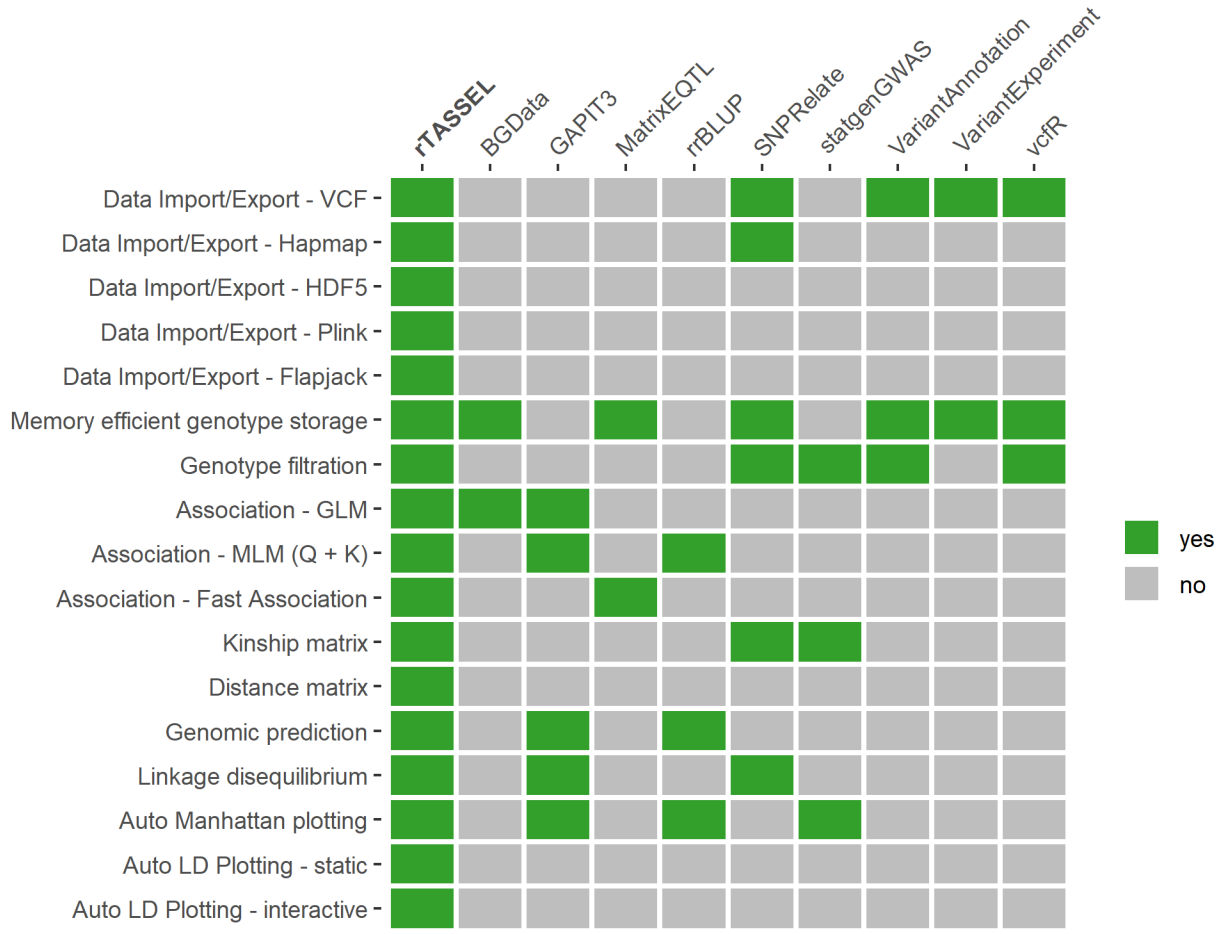

Figure S1: Feature comparisons of rTASSEL with other R packages. Features of rTASSEL (y-axis) are compared with other commonly-used R packages (x-axis). Packages that contain a specified feature are highlighted green (yes) and grey (no) if they do not contain a feature or are limited in scope. Association features for packages are based on if said package contains methods for generalized linear models, mixed linear models utilizing the "Q+K" method (Yu *et al.*, 2006), or multi trait fast association methods (Shabalin, 2012). Kinship and distance matrix features denote if a package can return an *n x n* matrix of values for further use. Packages that contain plotting features indicate if the package contains an automated plot feature instead of using base or grid-based R graphics (R Core Team, 2020) in conjunction with data output. The packages used for this comparison are BGData (Grueneberg and Campos, 2019), GAPIT3 (Wang and Zhang, 2020), MatrixEQTL (Shabalin, 2012), rrBLUP (Endelman, 2011), SNPRelate (Zheng *et al.*, 2012), statgenGWAS (Rossum and Kruijer, 2020), VariantAnnotation (Obenchain *et al.*, 2014), VariantExperiment (Liu *et al.*, 2020), and vcfR (Knaus and Grünwald, 2017).

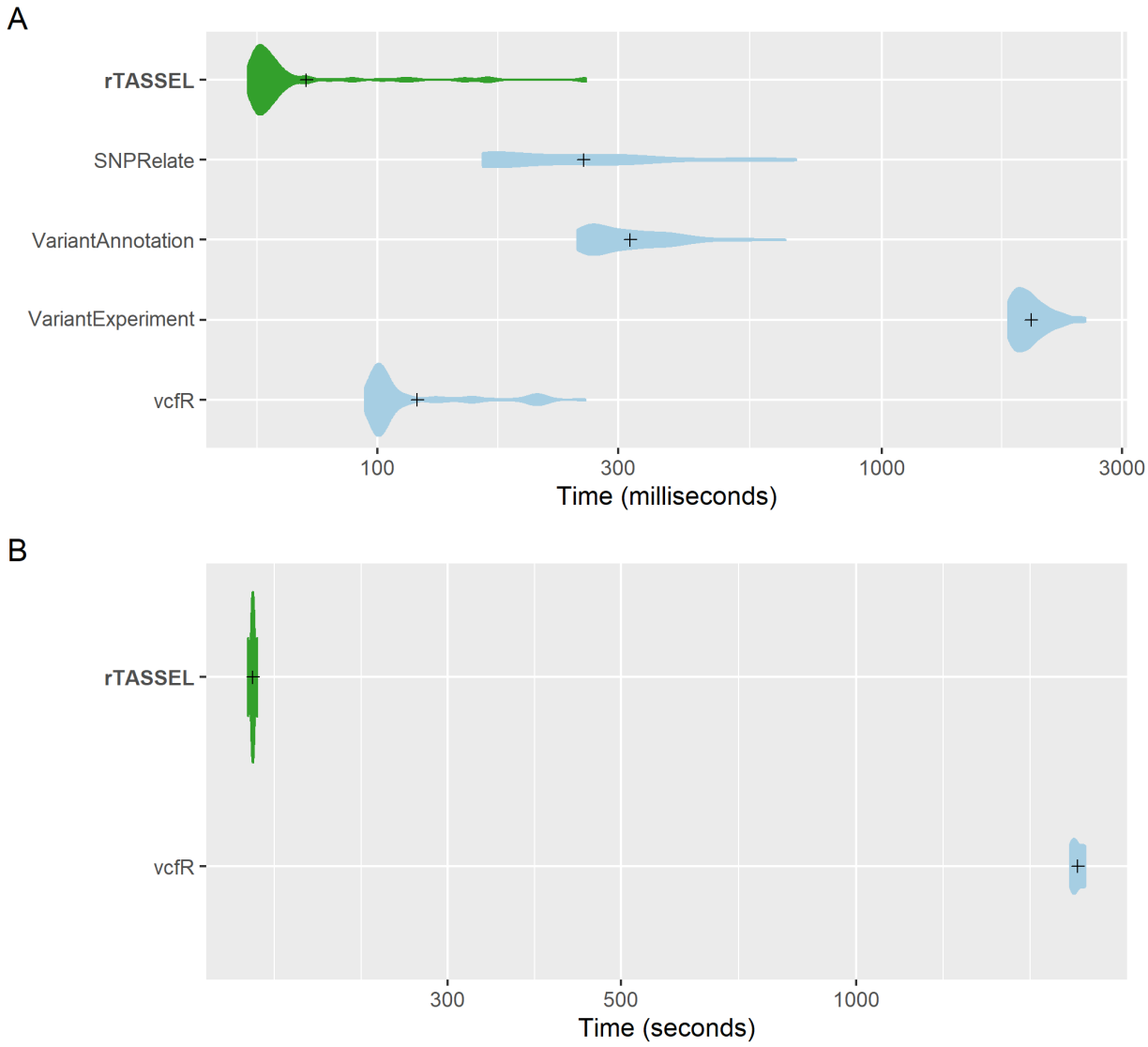

Figure S2: VCF import time comparisons of genotypic data. A distribution of replicated benchmark evaluations with recorded means (cross shapes) are plotted for rTASSEL and several R packages: SNPRelate (Zheng *et al.*, 2012), VariantAnnotation (Obenchain *et al.*, 2014), VariantExperiment (Liu *et al.*, 2020), and vcfR (Knaus and Grünwald, 2017). Import times are recorded for 279 samples x 3093 variant sites (A) and 1,210 samples x 2,255,405 variant sites (B).

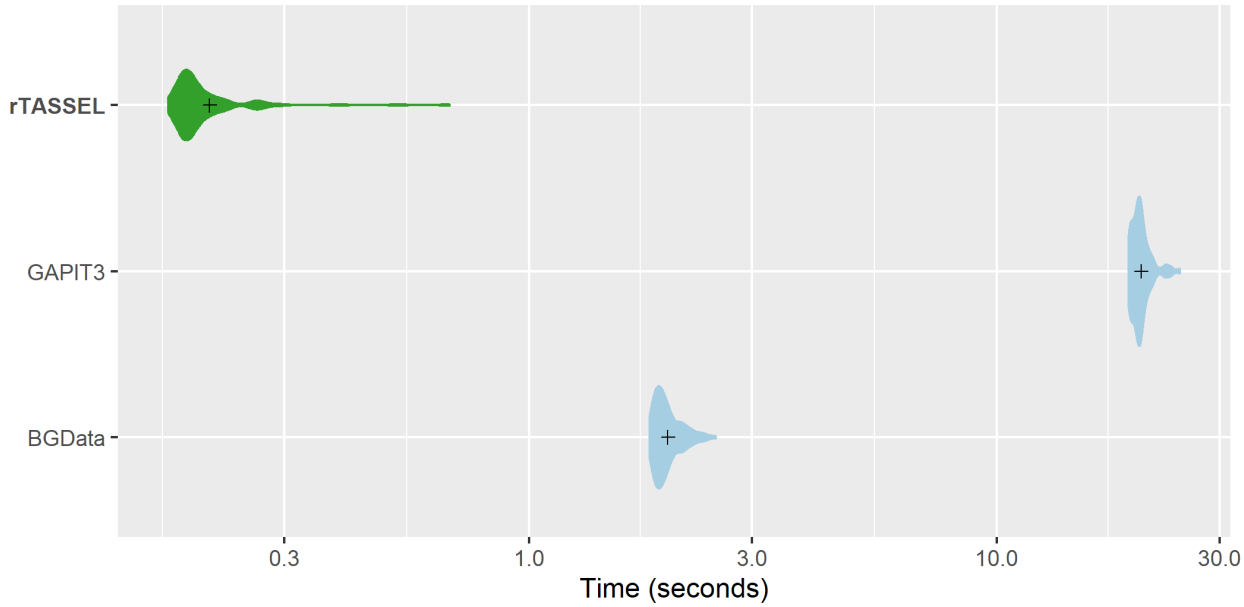

Figure S3: GLM association time comparisons. A distribution of replicated benchmark evaluations with recorded means (cross shapes) are plotted for rTASSEL and the R packages GAPIT3 (Wang and Zhang, 2020) and BGData (Grueneberg and Campos, 2019). Import times are recorded for 279 samples x 3093 variant sites and 1 measured trait.

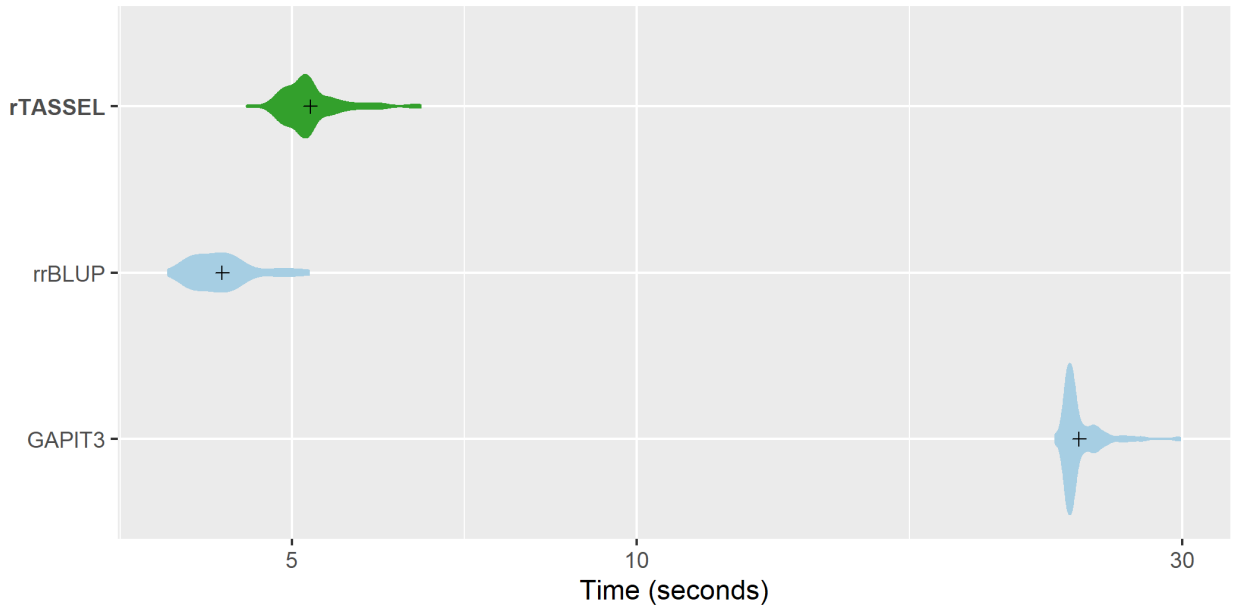

Figure S4: MLM association time comparisons. A distribution of replicated benchmark evaluations with recorded means (cross shapes) are plotted for rTASSEL and the R packages rrBLUP (Endelman, 2011) and GAPIT3 (Wang and Zhang, 2020). Import times are recorded for 279 samples x 3093 variant sites and 1 measured trait.

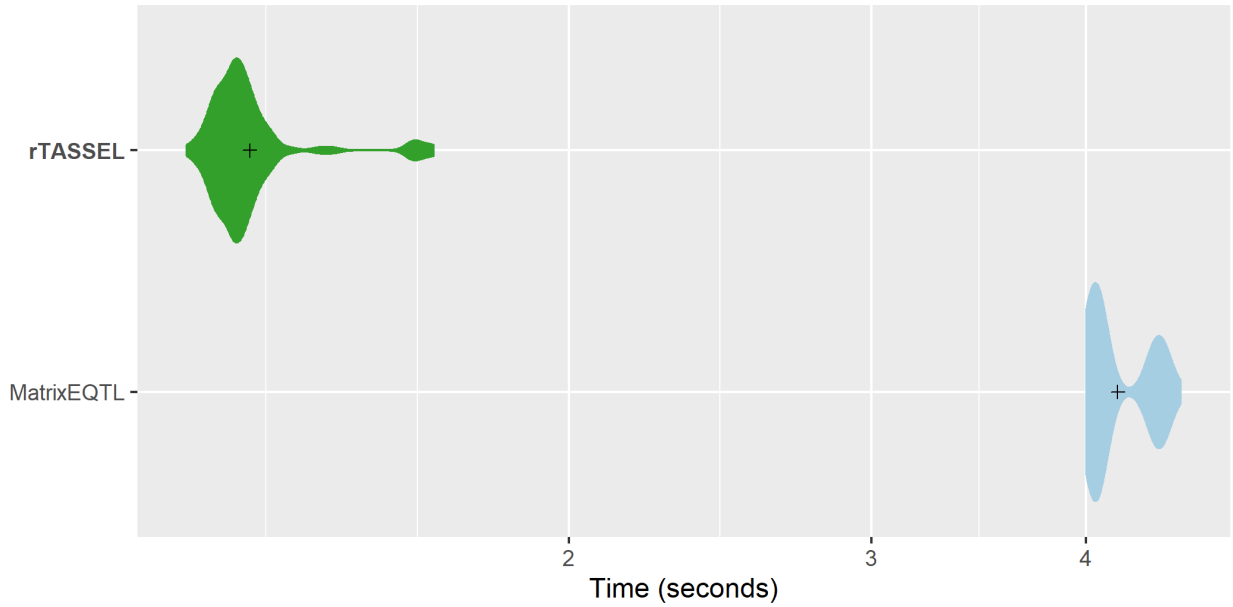

Figure S5: Fast association time comparisons. A distribution of replicated benchmark evaluations with recorded means (cross shapes) are plotted for rTASSEL and the R package MatrixEQTL (Shabalin, 2012). Import times are recorded for 279 samples x 3093 variant sites and 100 simulated RNA expression traits.

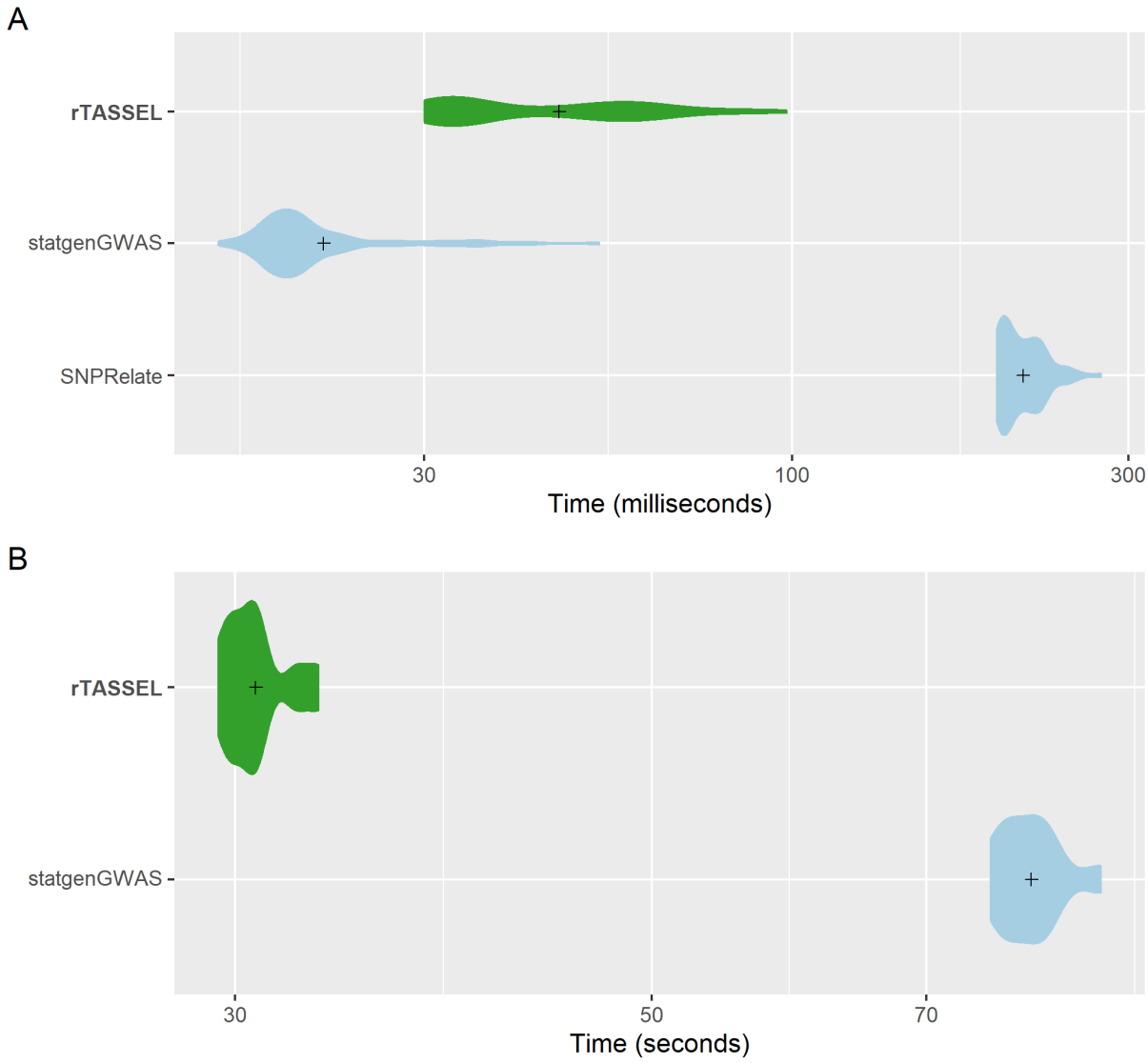

Figure S6: Kinship matrix (IBS) generation time comparisons of genotypic data. A distribution of replicated benchmark evaluations with recorded means (cross shapes) are plotted for rTASSEL and the R packages statgenGWAS (Rossum and Kruijer, 2020) and SNPRelate (Zheng *et al.*, 2012). Generation times are recorded for 279 samples x 3093 variant sites (A) and 1,210 samples x 2,255,405 variant sites (B).

References

Endelman,J.B. (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, **4**, 250–255.

Flint-Garcia,S.A. *et al.* (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.*, **44**, 1054–1064.

Grueneberg,A. and Campos,G. de los (2019) BGData - A Suite of R Packages for Genomic Analysis with Big Data. *G3 Genes Genomes Genet.*, **9**, 1377–1383.

Knaus,B.J. and Grünwald,N.J. (2017) VCFR: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.*, **17**, 44–53.

Liu,Q. *et al.* (2020) VariantExperiment: A RangedSummarizedExperiment Container for VCF/GDS Data with GDS Backend.

Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Mersmann,O. (2019) microbenchmark: Accurate Timing Functions.

Obenchain,V. *et al.* (2014) VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, **30**, 2076–2078.

R Core Team (2020) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Rossum,B.-J. van and Kruijer,W. (2020) statgenGWAS: Genome Wide Association Studies.

Shabalin,A.A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

Wang,J. and Zhang,Z. (2020) GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *bioRxiv*.

Yu,J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Zheng,X. *et al.* (2012) A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. *Bioinformatics*, **28**, 3326–3328.