

rTASSEL: An R interface to TASSEL for analyzing genomic diversity

Brandon Monier ¹, Terry M. Casstevens ¹, Peter J. Bradbury ^{1,2}, and 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

DOI: [10.21105/joss.04530](https://doi.org/10.21105/joss.04530)

Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: [Frederick Boehm](#)  

Reviewers:

- [@tkchafin](#)
- [@tomsing1](#)

Submitted: 21 June 2022

Published: 10 August 2022

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Summary

The need for efficient tools and applications for analyzing genomic diversity is essential for any genetics research or breeding program. One commonly used tool, TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage), provides many core methods for genomic analyses. Despite its efficiency, TASSEL has limited automation potential for reproducible research and to interact with other analytical tools. Here we present an R package, rTASSEL, that is a front-end to connect to a variety of highly used TASSEL methods and analytical tools. The goal of this package is to create a unified scripting workflow that leverages the analytical prowess of TASSEL, in conjunction with R's data handling and visualization capabilities, without ever having the user switch between these two environments.

Statement of need

As breakthroughs in genotyping technologies allow for increasing available variant resources, methods and implementations to analyze complex traits in a diverse array of organisms are needed. One such resource is TASSEL (Trait Analysis by aSSociation, Evolution, and Linkage). This software suite contains functionality for analyses in association studies, linkage disequilibrium (LD), kinship, and dimensionality reduction (e.g., PCA and MDS) (Bradbury et al., 2007). While initially released in 2001, the fifth version, TASSEL 5, has been optimized for handling large data sets and has added newer approaches to association analyses for many thousands of traits (Shabalin, 2012). Despite these improvements, interacting with TASSEL has been limited to either a graphical user interface with limited workflow reproducibility or a command-line interface with a higher learning curve that can dissuade novice researchers and provide unnecessary intermediate files in an analytics workflow (Zhang et al., 2009). To remediate this issue, we have created an R package, rTASSEL. This package interfaces the analytical power of TASSEL with R's data formats and intuitive function handling.

Approach

Implementation

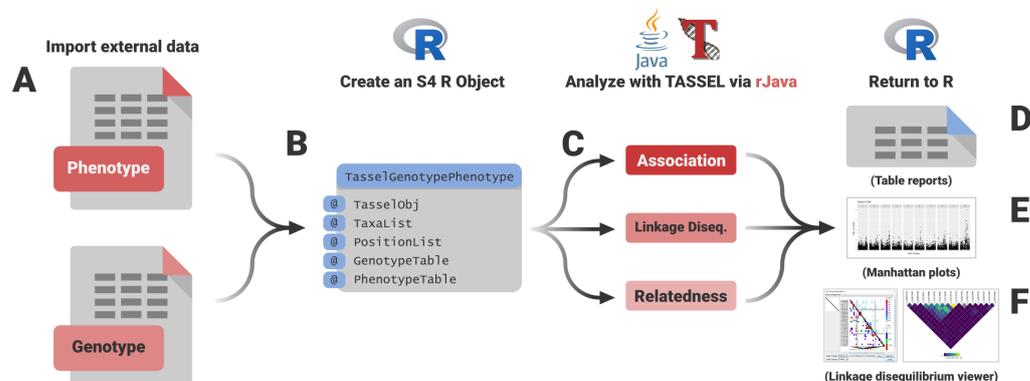


Figure 1: Overview of the *rTASSEL* workflow. Genotypic and phenotypic data (A) are used to create an R S4 object (B). From this object, TASSEL functionalities can be called to run various association, linkage disequilibrium, and relatedness functions (C). Outputs from these TASSEL analyses are returned to the R environment as data frame objects (D), Manhattan plot visualizations (E), or interactive visualizations for linkage disequilibrium analysis (F).

rTASSEL combines TASSEL's abilities to store genotype data as half bytes, bitwise arithmetic for kinship analyses, genotype filtration, extensive forms of linear modeling, multithreading, and access to a range of native libraries while providing access to R's prominent scripting capabilities and commonly used Bioconductor classes (Gentleman et al., 2004; Lawrence et al., 2013; Morgan et al., 2021). Since TASSEL is written in Java, a Java to R interface is implemented via the *rJava* package (Urbanek, 2021).

rTASSEL allows for the rapid import, analysis, visualization, and export of various genomic data structures. Diverse formats of genotypic information can be used as inputs for *rTASSEL*. These include variant call format (.vcf), HapMap (.hmp.txt), and Flapjack (.flpjk.*). Phenotype data can also be supplied in multiple formats. These include TASSEL formatted data sets or R data frame objects (Figure 1A).

Once data is imported, the function `readGenotypePhenotype` is used to construct an S4 object, which is used for all downstream analyses (Figure 1B, Figure 1C). This object contains slots that exclusively hold references to objects held in the Java virtual machine, which can be called with downstream functions. Prior to analysis, genotype objects can be quickly imported and filtered in several ways to help in the reduction of confounding errors. *rTASSEL* can filter genotype objects by either variant site properties (`filterGenotypeTableSites`) or by individuals (`filterGenotypeTableTaxa`).

Association functions

One of TASSEL's most dynamic functionalities is its capability to perform various association modeling techniques. *rTASSEL* allows several types of association studies to be conducted using one primary function, `assocModelFitter`, with different parameter inputs. This allows for implementing both least-squares fixed-effects general linear models (GLM) and mixed linear models (MLM) via the $Q + K$ method (Yu et al., 2006). If no genotypic data is provided to the GLM model, `assocModelFitter` can calculate best linear unbiased estimates (BLUEs). Additionally, fast GLM approaches are implemented in *rTASSEL*, which allow for the rapid analysis of many phenotypic traits (Shabalin, 2012).

Linear models can be specified following the format used by R's `lm` function:

$$y \sim A_1 + A_2 + \dots + A_n$$

where y is phenotype data, and A_n is any covariate or factor data. This formula parameter and several other parameters allow the user to run BLUE, GLM, or MLM modeling. Once association analysis is completed, TASSEL table reports of association statistics are generated as an R list which can then be exported as flat files or converted to data frames (Figure 1D). `rTASSEL` can also visualize association statistics with the function, `manhattanPlot`, which utilizes the graphical capabilities of the package, `ggplot2` (Wickham, 2016) (Figure 1E).

Linkage disequilibrium

`rTASSEL` can also generate linkage disequilibrium (LD) from genotype data via the function `linkageDiseq`. LD is estimated by the standardized disequilibrium coefficient, D' , correlation between alleles at two loci (r^2), and subsequent p -values via a two-sided Fisher's exact test. TASSEL table reports for all pairwise comparisons are generated as data.frame objects, and heatmap visualizations for each given metric are generated via TASSEL's legacy LD Java viewer or `ggplot2` (Figure 1F).

Relatedness functions

For users to run MLM methods, relatedness estimates need to be calculated. `rTASSEL` can efficiently compute this on large data sets by processing blocks of sites at a time using bitwise operations. This can be accomplished using the function `kinshipMatrix`, which will generate a kinship matrix from genotype data. Several methods for calculating kinship in TASSEL are implemented. By default, a "centered" identity by state (IBS) approach is used (Endelman & Jannink, 2012). Additionally, normalized IBS (Yang et al., 2011), dominance-centered IBS (Muñoz et al., 2014), and dominance normalized IBS (Zhu et al., 2015) can be used. `rTASSEL` can either generate a reference object for association analysis or an R matrix object via R's `as.matrix` function for additional analyses. In addition to kinship generation, principal components analysis and multidimensional scaling can be used on genotype data using `rTASSEL` methods, `pca` and `mds`, respectively. Finally, phylogenetic analysis can be performed on genotype data using the `createTree` method which will generate `phylo` objects commonly used by the `ape` package (Paradis & Schliep, 2019). The `createTree` method allows for two clustering methods: neighbor joining or UPGMA (unweighted pair group method with arithmetic mean).

Genomic prediction

The function `genomicPrediction` can be used for predicting phenotypes from genotypes. To do this, `genomicPrediction` uses genomic best linear unbiased predictors (gBLUPs). It proceeds by fitting a mixed model that uses kinship to capture covariance between taxa. The mixed model can calculate BLUPs for taxa that do not have phenotypes based on the phenotypes of lines with relationship information.

Additional resources

More information about various functionalities and workflows can be found on our [project webpage](#). Source code can be found on our [GitHub repository](#). An interactive Jupyter notebook session detailing additional `rTASSEL` workflows can be found on [Binder](#).

Acknowledgements

This project is supported by the USDA-ARS, the Bill and Melinda Gates Foundation, and NSF IOS #1822330. We thank Sara J. Miller, Guillaume Ramstein, and Joseph Gage for their insightful suggestions on this manuscript and pipeline testing.

References

- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633–2635. <https://doi.org/10.1093/bioinformatics/btm308>
- Endelman, J. B., & Jannink, J.-L. (2012). Shrinkage estimation of the realized relationship matrix. *G3: Genes, Genomes, Genetics*, 2(11), 1405–1413. <https://doi.org/10.1534/g3.112.004259>
- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., ... Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), R80. <https://doi.org/10.1186/gb-2004-5-10-r80>
- Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLOS Computational Biology*, 9(8), e1003118. <https://doi.org/10.1371/journal.pcbi.1003118>
- Morgan, M., Obenchain, V., Hester, J., & Pagès, H. (2021). *SummarizedExperiment: SummarizedExperiment container*. <https://bioconductor.org/packages/SummarizedExperiment>
- Muñoz, P. R., Resende, M. F. R., Gezan, S. A., Resende, M. D. V., Campos, G. de los, Kirst, M., Huber, D., & Peter, G. F. (2014). Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics*, 198(4), 1759–1768. <https://doi.org/10.1534/genetics.114.171322>
- Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35, 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- Shabalín, A. A. (2012). Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10), 1353–1358. <https://doi.org/10.1093/bioinformatics/bts163>
- Urbanek, S. (2021). *rJava: Low-level R to Java interface*. <https://CRAN.R-project.org/package=rJava>
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4
- Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1), 76–82. <https://doi.org/10.1016/j.ajhg.2010.11.011>
- Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2), 203–208. <https://doi.org/10.1038/ng1702>
- Zhang, Z., Buckler, E. S., Casstevens, T. M., & Bradbury, P. J. (2009). Software engineering the mixed model for genome-wide association studies on large samples. *Briefings in Bioinformatics*, 10(6), 664–675. <https://doi.org/10.1093/bib/bbp050>

Zhu, Z., Bakshi, A., Vinkhuyzen, A. A. E., Hemani, G., Lee, S. H., Nolte, I. M., Vliet-Ostaptchouk, J. V. van, Snieder, H., Esko, T., Milani, L., Mägi, R., Metspalu, A., Hill, W. G., Weir, B. S., Goddard, M. E., Visscher, P. M., & Yang, J. (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics*, 96(3), 377–385. <https://doi.org/10.1016/j.ajhg.2015.01.001>