BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 19 2007, pages 2633–2635
doi:10.1093/bioinformatics/btm308

Genetics and population analysis

TASSEL: software for association mapping of complex traits in diverse samples

Peter J. Bradbury[1,†,*], Zhiwu Zhang[2,†], Dallas E. Kroon[2,†], Terry M. Casstevens[2], Yogesh Ramdoss[3] and Edward S. Buckler[1,2]

[1]United States Department of Agriculture-Agricultural Research Service, [2]Institute for Genomic Diversity, Cornell University, Ithaca, New York and [3]Cisco Systems, RTP, NC, USA

Received on March 26, 2007; revised on May 11, 2007; accepted on June 2, 2007

Advance Access publication June 22, 2007

Associate Editor: Martin Bishop

ABSTRACT

Summary: Association analyses that exploit the natural diversity of a genome to map at very high resolutions are becoming increasingly important. In most studies, however, researchers must contend with the confounding effects of both population and family structure. TASSEL (Trait Analysis by aSSociation, Evolution and Linkage) implements general linear model and mixed linear model approaches for controlling population and family structure. For result interpretation, the program allows for linkage disequilibrium statistics to be calculated and visualized graphically. Database browsing and data importation is facilitated by integrated middleware. Other features include analyzing insertions/deletions, calculating diversity statistics, integration of phenotypic and genotypic data, imputing missing data and calculating principal components.

Availability: The TASSEL executable, user manual, example data sets and tutorial document are freely available at http://www.maizegenetics.net/tassel. The source code for TASSEL can be found at http://sourceforge.net/projects/tassel.

Contact: pjb39@cornell.edu

1 INTRODUCTION

With advances in genotyping technology, including rapid increases in the number of genetic markers available for quantitative trait loci (QTL) studies (Churchill *et al*., 2004), association analysis has become a viable approach for the dissection of complex traits. A key issue in developing methods for analyzing association data is controlling false positives that arise from population and family structure. One widely used approach, structured association (Pritchard *et al*., 2000; Thornsberry *et al*., 2001), was first implemented in TASSEL (Trait Analysis by aSSociation, Evolution and Linkage) to reduce the risk of false positives arising from population structure. More recently, a unified mixed model method was developed which improves on the previous method by integrating population structure and family relatedness within populations (Yu *et al*., 2006). TASSEL reflects these improvements and offers a variety of data manipulation and results displays. Plant, animal or human geneticists and breeders interested in performing association analysis will find this software useful.

2 ASSOCIATION TOOLS

Because complex traits are usually controlled by multiple QTL, the primary goal of QTL mapping is to find associated markers for each QTL. To achieve this goal, many association studies fit markers into a linear model as fixed effects. Any QTL not associated with a marker will contribute to the residual error, thereby inflating the error term and reducing statistical power. To compound matters further, QTL not physically linked to a marker may cause spurious associations between phenotypes and that marker due to factors, such as selection, population admixture or family structure. A structured association approach can partially correct for these false associations by using a Q-matrix of population membership estimates. Additional improvement can be made by incorporating multiple background QTL as a random effect in a mixed model, which takes into account covariances due to relatedness. The average relationship between individuals or lines can be estimated by kinship (K) calculated either from pedigrees or a suitable number of random markers across the entire genome. A composite approach, Q + K, that combines information from both Q and K, has been shown to be superior (Yu *et al*., 2006) to these former methods.

The Q method for structured association analysis was implemented in TASSEL as a general linear model (GLM) function. Population membership estimates serve as covariates in the model and can be derived using programs such as STRUCTURE (Pritchard *et al*., 2000) or principal components analysis (PCA) (Zhao *et al*., 2007). For each marker-trait combination, GLM finds the ordinary least squares solution as described in Searle (1987). The model can include main effects, interactions, nested effects and covariates. The user can specify

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

F-tests to be calculated, permutations tests to be run and estimates of model effect means to be output.

The test of significance derived from an *F*-distribution assumes that the trait being analyzed has normally distributed residual error. When this is not the case, TASSEL provides two options. The first option provides some transformation functions which may produce roughly normal error terms, while the second utilizes a permutation test to generate *P*-values that are not distribution dependent. Algorithms for conducting such permutation tests are based on the formulation by Anderson and Ter Braak (Anderson and Ter Braak, 2003).

The Q + K method was implemented in TASSEL as a mixed linear model (MLM) function. The statistical model can be described in Henderson's notation (Henderson, 1975) as follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$$

where **y** is the vector of observations; **β** is an unknown vector containing fixed effects including genetic marker and population structure (Q); **u** is an unknown vector of random additive genetic effects from multiple background QTL for individuals or lines; **X** and **Z** are the known design matrices; and **e** is the unobserved vector of random residuals. Each marker allele is fit as a separate class with heterozygotes fit as additional marker classes. The resulting marker effect is not decomposed into additive and dominance effects but simply tested for overall significance. The **u** and **e** vectors are assumed to be normally distributed with null mean and variance of

$$\mathrm{Var}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix}$$

where $\mathbf{G} = \sigma_a^2\mathbf{K}$ with σ_a^2 as the unknown additive genetic variance and **K** as the kinship matrix. TASSEL provides a function to estimate **K** from a set of random markers covering the whole genome. TASSEL also provides a function to import matrix **K** calculated externally from pedigrees by using SAS PROC INBREED (SAS, 2002) or from markers by using software packages such as SPAGedi (Hardy and Vekemans, 2002). Homogeneous variance is assumed for the residual effect, making $\mathbf{R} = \mathbf{I}\sigma_e^2$, where σ_e^2 is the unknown residual variance. The REML estimates of σ_a^2 and σ_e^2 are obtained through the expectation and maximization (EM) algorithm (Laird and Ware, 1982).

3 OTHER FEATURES

In addition to providing association tools, TASSEL permits the analysis of diversity estimates such as average pairwise divergence (π) and segregating sites. Linkage disequilibrium is estimated by the standardized disequilibrium coefficient, D', as well as r^2 and *P*-values. TASSEL also includes a variety of data extraction utilities and visualization tools, such as a sequence alignment viewer, extraction of SNPs and indels from alignments, neighbor-joining cladogram and a variety of data graphing functions.

TASSEL contains several useful data management functions. In addition to importing data in flat file formats, a data browser from the GDPC (Genomic Diversity and Phenotype Connection) project (Casstevens and Buckler, 2004) has been integrated into TASSEL to provide an interface to relational databases. GDPC can utilize multiple data sources, retrieve filtered data and export tab-delimited text files. TASSEL can merge data from different sources into a single analysis data set, impute missing data using a k-nearest-neighbor algorithm (Cover and Hart, 1967) and conduct PCA to reduce a set of correlated phenotypes. Some of these features are shown in Figure 1.

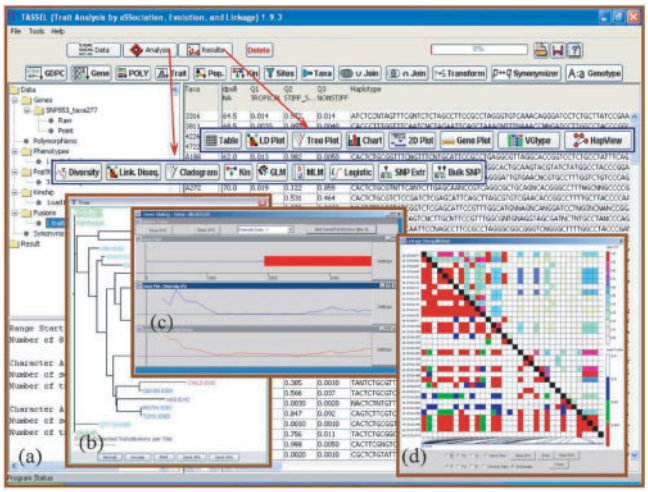

Fig. 1. Illustration of features in TASSEL. (**a**) Main interface with data (default), analysis and result control panels. (**b**) Plotting trees to understand phylogenetic relationships. (**c**) Diversity estimates (silent, non-synonymous, indel and synonymous of gene, π and θ). (**d**) LD plots with positional information.

4 IMPLEMENTATION

This software package was developed in Java, making it compatible with multiple platforms (e.g. Windows, Mac and Linux). The package uses the standard PAL library (http://iubio.bio.indiana.edu/soft/molbio/java/pal/doc/), the COLT library (http://dsd.lbl.gov/~hoschek/colt/) and jFreeChart (http://www.jfree.org/jfreechart/). Database access is achieved by GDPC middleware (http://www.maizegenetics.net/gdpc).

ACKNOWLEDGEMENTS

This project is supported by the USDA-ARS and the National Science Foundation. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

Conflict of Interest: none declared.

REFERENCES

Casstevens,T.M. and Buckler,E.S. (2004) GDPC: connecting researchers with multiple integrated data sources. *Bioinformatics*, **20**, 2839–2840.

Churchill,G. *et al*. (2004) The collaborative cross, a community resource for the genetic analysis of complex traits. *Nat. Genet*., **36**, 1133–1137.

Cover,T. and Hart,P. (1967) Nearest neighbor pattern classification. *Proc. IEEE Trans. Inform. Theory*, **13**.

Hardy,O.J. and Vekemans,X. (2002) SPAGEDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes*, **2**, 618–620.

Henderson,C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

Laird,N.M. and Ware,J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Pritchard,J.K. *et al*. (2000) Association mapping in structured populations. *Am. J. Hum. Genet*., **67**, 170–181.

SAS Intstitute Inc. (2002) *SAS/STAT software, version 9*. SAS Institute, Inc., Cary, NC, USA.

Thornsberry,J.M. *et al*. (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet*., **28**, 286–289.

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

Zhao,K. *et al*. (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genet*., **3**, e4. doi:10.1371/journal.pgen.0030004.