BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 15 2012, pages 2086–2087
doi:10.1093/bioinformatics/bts335

*Genetics and population analysis*

Advance Access publication June 10, 2012

# synbreed: a framework for the analysis of genomic prediction data using R

Valentin Wimmer, Theresa Albrecht, Hans-Jürgen Auinger and Chris-Carolin Schön*

Plant Breeding, Technische Universität München, Emil-Ramann-Straße 4, 85354 Freising, Germany

Associate Editor: Jeffrey Barrett

## ABSTRACT

**Summary:** We present a novel R package named `synbreed` to derive genome-based predictions from high-throughput genotyping and large-scale phenotyping data. The package contains a comprehensive collection of functions required to fit and cross-validate genomic prediction models. All functions are embedded within the framework of a single, unified data object. Thereby a versatile genomic prediction analysis pipeline covering data processing, visualization and analysis is established within one software package. The implementation is flexible with respect to a wide range of data formats and models. The package fills an existing gap in the availability of user-friendly software for next-generation genetics research and education.

**Availability:** `synbreed` is open-source and available through CRAN http://cran.r-project.org/web/packages/synbreed. The latest development version is available from R-Forge. The package `synbreed` is released with a vignette, a manual and three large-scale example datasets (from package `synbreedData`).

**Contact:** chris.schoen@wzw.tum.de

**Supplementary Information:** Supplementary data are available at *Bioinformatics* online.

Received on January 17, 2012; revised on March 28, 2012; accepted on June 6, 2012

## 1 BACKGROUND

High-throughput genotyping technologies delivering thousands of single nucleotide polymorphism markers (SNPs) have become available for many crop and livestock species. In breeding, selection of the best genotypes can be conducted on high-density marker profiles once sufficiently accurate genome-based prediction models have been established. To achieve this, prediction models are developed based on large training populations for which genotypic and phenotypic data are available (Meuwissen *et al.*, 2001).

Implementation of genomic prediction in breeding programs will be advanced through the availability of comprehensive, user-friendly software that covers a wide range of analysis steps. Currently, a researcher is faced with a plethora of different software tools. The program `ASReml` (Gilmour *et al.*, 2009) provides restricted maximum-likelihood estimation procedures for linear mixed models with arbitrary variance–covariance structure. The program `PLINK` (Purcell *et al.*, 2007) implements algorithms for linkage disequilibrium (LD) and identical-by-descent estimation. Within the R software (R Development Core Team, 2012), the package `regress` (Clifford and McCullagh, 2012) fits linear mixed models in which the covariance structure can be expressed as a linear combination of known matrices. However, there is no single software covering the specific needs of genomic prediction. By connecting important analysis tools such as processing of SNP data, estimation of genome- and pedigree-based coefficients of relatedness, different statistical models and cross-validation (CV), we provide a framework for genomic prediction within one software. Where necessary, a gateway to other software is provided to extend the field of applications. This enhances the implementation of customized high-throughput analysis pipelines. An intuitive application is warranted through the consistent use of a unified data object.

## 2 OVERVIEW

The data flow in `synbreed` is guided by a single, unified data object of class `gpData` ('*g*enomic *p*rediction *Data*') which is used for storage of multiple data sources. This includes an array for the phenotypes (individual × trait × replication) and a matrix for the marker genotypes (individual × marker scores). If required, this structure can be extended to include pedigree information and a marker map. All analysis functions are based on this data structure. A key feature of the `synbreed` package is the generality of the class `gpData` which is suitable for a wide range of statistical methods using genotypic and phenotypic data. Moreover, it is very convenient to store and share objects of class `gpData`. Any object of class `gpData` can be converted to class `cross` in the `qtl` R package (Broman *et al.*, 2003) and vice versa or a `data.frame`.

The function `codeGeno` provides algorithms for the processing of SNP data. This includes the transfer of arbitrary coding schemes into the number of copies of the minor allele, preselection of SNPs and imputing of missing values, e.g. using a gateway to `Beagle` (Browning and Browning, 2009). Linear mixed models are used for the prediction of genetic values from genome-wide marker data with

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (1)$$

where $\mathbf{y}$ is the $n \times 1$ vector of $n$ phenotypic records, $\beta$ is the vector of fixed effects and $\mathbf{u}$ is a $n \times 1$ vector of random effects. Observations are allocated to the fixed and random effects by the corresponding design matrices $\mathbf{X}$ and $\mathbf{Z}$. The $n \times 1$ vector $\mathbf{e}$ denotes the residuals with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$. In genome-based best linear unbiased prediction (GBLUP), genetic values are modeled as random effects with $\mathbf{u} \sim N(0, \mathbf{U}\sigma^2_{\mathbf{u}})$, where $\mathbf{U}$ is the realized relationship matrix (Habier *et al.*, 2007) and $\sigma^2_u$ is the genetic variance pertaining to model (1).

*To whom correspondence should be addressed.

The GBLUP model can be fitted with package `regress` through the unified interface of the function `gpMod`. Moreover, we implemented pedigree-based BLUP and established a connection to the `BLR` package for Bayesian regression models (Pérez *et al*., 2010). CV can be used to assess the predictive ability (correlation of predicted and observed phenotypic values) of a statistical model with stratified CV schemes (Albrecht *et al*., 2011). All results generated by the package such as the LD patterns and coefficients of relatedness can be visualized using summary statistics and innovative plot functions.

## 3 EXAMPLE

This example traces the steps from processing raw data to the validation of a prediction model according to the workflow in Supplementary Figure S1. We use simulated data of a maize breeding program. It comprises 1250 doubled haploid (DH) lines fingerprinted for 1117 polymorphic SNPs and a quantitative trait evaluated in testcrosses of DH lines with a single tester.

Step 1 (raw data merge): The object `maize` is given as object of class `gpData` and hence Step 1 is omitted. To load the data, use

```
data(maize)
```

Step 2 (processing and filtering): DH lines are fully inbred lines and hence homozygous for every SNP. Recoding SNP marker genotypes to the number of copies of the minor allele and preselection of SNPs with a minor allele frequency $\geq$0.05 is conducted using

```
maizeC <- codeGeno(maize, maf=0.05)
```

Step 3 (kinship coefficients): Remaining 995 SNPs were used to estimate the realized relationship matrix based on the recoded marker genotypes:

```
U <- kin(maizeC, ret="realized")
```

A heatmap visualization is available by using `plot(U)`, see Supplementary Figure S4.

Step 4 (prediction model): A GBLUP model for the trait is developed using the realized relationship matrix from Step 3. For the prediction of testcross values, the relationship matrix must be replaced by the kinship matrix, i.e. divided by 2.

```
GBLUP <- gpMod(maizeC, mod="BLUP", kin=U/2)
```

Estimates for marker effects are visualized using the `manhattanPlot` function, see Supplementary Figure S5.

Step 5 (model validation): Finally, we estimate the predictive ability of GBLUP using 2-fold CV with five replications each with a random assignment into estimation set (ES) and test set. The estimated variance components $\hat{\sigma}^2$ and $\hat{\sigma}_u^2$ are committed from Step 4 and used to build a prediction model within every ES:

```
cv <- crossVal(maizeC, k = 2, Rep = 5,
 cov.matrix = list(U/2),
 varComp = GBLUP$fit$sigma, Seed=1)
```

By using `summary(cv)`, we obtain an average predictive ability of 0.48 with a range from 0.44 to 0.52.

## 4 CONCLUSION

In the plant breeding community, there is a strong demand for a standard software covering a wide range of analysis steps. The package `synbreed` offers a comprehensive collection of methods required in the analysis of genomic prediction data. This is a step towards automatized analysis pipelines in the analysis of next-generation genotype and phenotype data which are required to bring genomic prediction from theory to practice. Moreover, `synbreed` is a valuable tool for the education of young scientists and breeders.

## ACKNOWLEDGEMENTS

We gratefully acknowledge Larry Schaeffer for providing us R code, Michael Höhle and Yu Wang for valuable discussions and Malena Erbe, Christian Reimer and Ulrike Ober for suggestions and contributions to the package.

*Funding*: This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr Synbreed - Synergistic plant and animal breeding (FKZ 0315528A).

*Conflict of Interest*: none declared.

## REFERENCES

Albrecht,T. *et al*. (2011) Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.*, **123**, 339–350.

Broman,K.W. *et al*. (2003) R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, **7**, 889–890.

Browning,B.L. and Browning,S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.

Clifford,D. and McCullagh,P. (2012) The regress package. R package version 1.3–8.

Gilmour,A.R. *et al*. (2009) *ASREML User Guide Release 3.0*. VSN International Ltd, Hemel Hempstead, UK.

Habier,D. *et al*. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, **177**, 2389–2397.

Meuwissen,T.H.E. *et al*. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Pérez,P. *et al*. (2010) Genomic-enabled prediction based on molecular markers and pedigree using the BLR package in R. *Plant Genome*, **3**, 106–116.

Purcell,S. *et al*. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**, 559–575.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.