The synbreed R package: A framework for genome-based prediction using R

V. Wimmer, T. Albrecht, H.-J. Auinger, C.-C. Schön Technische Universität München, Plant Breeding, Freising

Motivation

The `synbreed` R package [1] is designed to derive genome-based predictions from high-throughput genotyping and large scale phenotyping data. It contains a **comprehensive collection** of functions required to fit and cross-validate genome-based 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

The `synbreed` package is **open-source** and available through CRAN (see QR code below):

```
http://cran.r-project.org/web/packages/synbreed
```

The latest development version is available from R-Forge:

```
http://synbreed.r-forge.r-project.org
```

The package `synbreed` is released with a vignette (available using `vignette("IntroSyn")`), a manual and three large-scale example data sets from maize, cattle and mice (in package `synbreedData`, also on CRAN).

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 (see Figure 1).

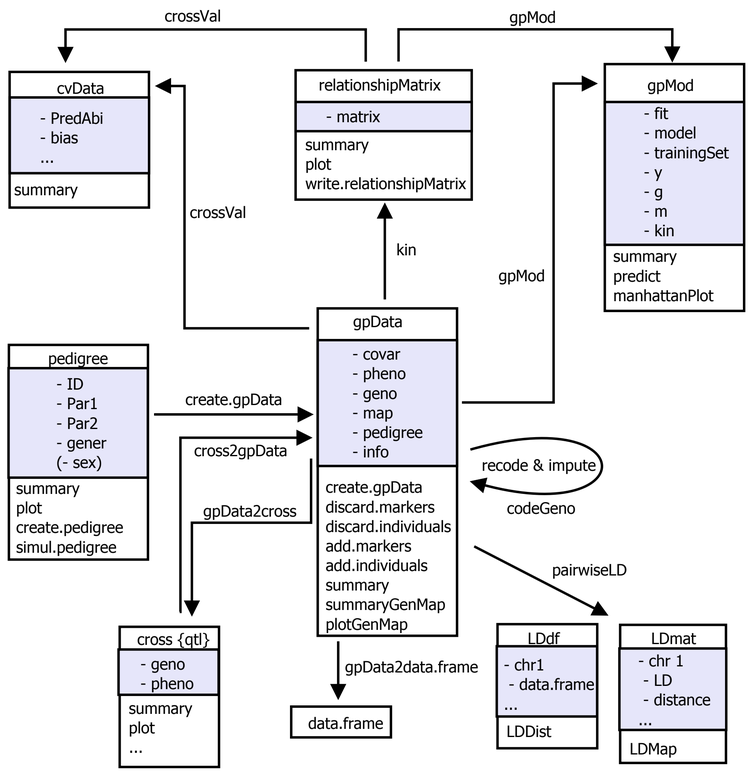


Figure 1: Overview of object classes, methods and functions within the `synbreed` package. Each box indicates a class together with the class name, the elements and the available functions and methods. The arrows indicate the data flow. The origin indicates the input argument and the head is the return value of the function.

References

[1] Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) synbreed: A framework for the analysis of genomic prediction data using R. Bioinformatics

Example

Data: Simulated maize breeding program with 1250 doubled haploid (DH) lines fingerprinted for 1117 polymorphic SNPs and a quantitative trait evaluated in testcrosses of DH lines with one common tester.

Step 1 (Load data):

```
R> library(synbreed)
R> data(maize)
```

Step 2 (processing & filtering): Recoding SNP marker genotypes to the number of copies of the minor allele, i.e. 0, 1 and 2 and preselection of SNPs with a minor allele frequency (MAF) $\geq$ 0.05 is conducted using

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

Recoded marker genotypes were used to estimate pairwise linkage disequilibrium (LD) measured as r^2 by all marker pairs on chromosome 1 using

```
R> LD1 <- pairwiseLD(maizeC,chr=1)
```

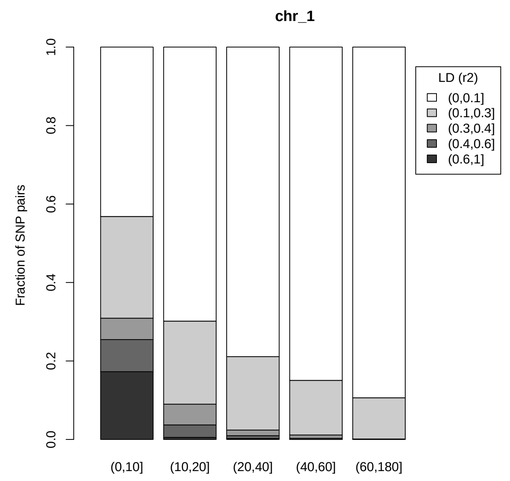
The extent of LD and LD decay is visualized (Figure 2) using

```
R> LDDist(LD1,type="bars",breaks=list(dist=c(0,10,20,40,60,180),
+   r2=c(1,0.6,0.4,0.3,0.1,0)))
```

Step 3 (kinship coefficients): The remaining 995 SNPs were used to estimate the realized relationship matrix for the 1250 DH lines based on the recoded marker genotypes:

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

A heatmap visualization is available by using `plot(U)` (see Figure 3).



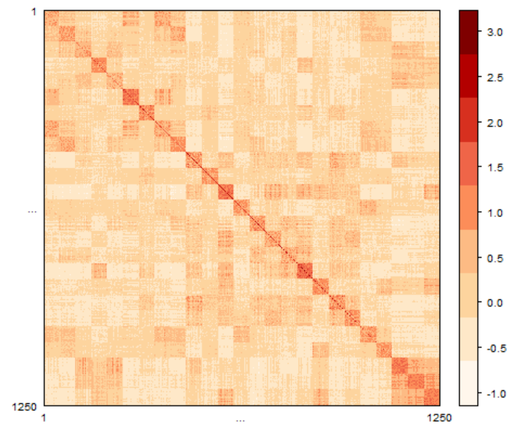


Figure 2: LD decay visualization for chromosome 1 **Figure 3:** Heatmap of the marker-based relationship matrix

Step 4 (prediction model): A GBLUP model for the testcross values 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.

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

Step 5 (model validation): Finally, we estimate the predictive ability of GBLUP using 2-fold cross-validation with 5 replications each with a random assignment into estimation set (ES) and test set. The estimated variance components are committed from step 4 and used to build a prediction model within every ES:

```
R> 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.

SPONSORED BY THE
Federal Ministry of Education and Research

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed - Synergistic plant and animal breeding" (Fund ID 0315528A).

Valentin.Wimmer@wzw.tum.de
www.plantbreeding.wzw.tum.de
www.synbreed.tum.de