Center of Life and Food Sciences Weihenstephan

Synbreed

The `synbreed` R package

V. Wimmer, T. Albrecht, H.-J. Auinger, C.-C. Schön

Technische Universität München, Plant Breeding, Freising

Introduction

In many **plant and livestock breeding** programs dense genome-wide markers are used to increase the genetic gain. The prediction of the genetic potential of individuals from their DNA sequence is called **genomic prediction**. Different statistical models and software packages have been used for this purpose. However, there is no comprehensive **software** available where all required data analysis steps including data processing and visualization have been implemented. We present a novel **R package** named `synbreed` which implements methods to build an **analysis pipeline** of genomic prediction data within the freely available **open source** software R [1].

An **unified data object** for genomic prediction or association analysis is created to merge phenotypic, genotypic, marker map and pedigree data. Thereby, the implementation is flexible with respect to data structure and fast with respect to data size. Within this framework, the evaluation of data from a **breeding program** of both plant and animal breeding is feasible and can be **automatized** to a large extend. The `synbreed` package is hosted and developed on

`https://r-forge.r-project.org/projects/synbreed/`

In 2011, the package `synbreed` will be released to CRAN together with a detailed user manual.

Structure

The class `gpData` is used for data storage of all required data sources:

pheno	`data.frame` for one or more traits (with or without replications)
geno	`matrix` with genotypic data in arbitrary coding (by genotypes or by alleles)
map	`data.frame` with chromosome number and position within each chromosome (genetic or physical distance) for markers in `geno`
pedigree	object of class `pedigree` with pedigree structure of individuals
covar	`data.frame` with covariates for individuals

Unique row names/column names are used for individuals/markers to have clear **data queries** and **data merges**. Further methods of the package are implemented based on this structure.

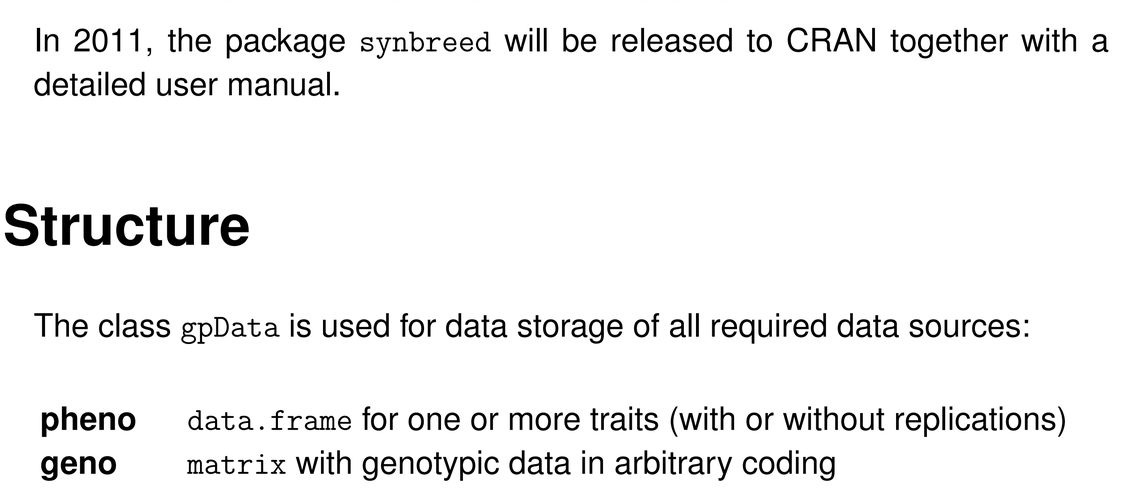

Figure 1: Simulated `maize` data: Marker map for data set (left) and LD decay visualization for 76 markers on the first chromosome (right).

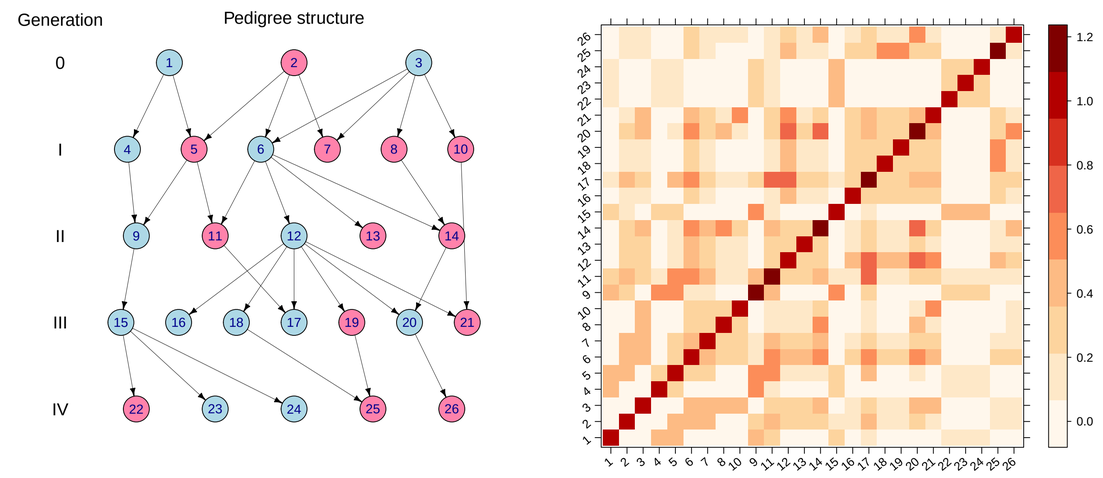

Figure 2: Pedigree structure (left) for 26 individuals (blue=sire, red=dam) and heatmap of the corresponding estimated numerator relationship matrix based on pedigree (right)

Functions

Data processing

- Combining raw data sources to a `gpData` object
- Conversion from and to class `cross` in package `qtl`
- Coding marker data into number of copies of the minor allele
- Preselection of markers according to MAF, % missing values and LD
- Imputation of missing genotypes by fixed value, marginal allele distribution or by family structure for fully homozygous inbred individuals

Data visualization and analysis

- Summary method for classes `gpData`, `pedigree` and `relationshipMatrix`
- Marker map representation for low and high density maps (Figure 1)
- LD computation as r^2 and LD decay visualization as scatterplot or stacked histogram (Figure 1)
- Pedigree tree and kinship visualization of relatedness between individuals (Figure 2)

Statistical models

- Estimation of pedigree based relationship (additive and dominance)
- Marker based relationship (according to vanRaden [2] or Rogers' [3] distance)
- Cross-validation for BLUP, Ridge Regression and Bayesian methods

Data sets

maize Simulated data for a **maize breeding program** using 1250 doubled haploids (DH) with phenotypes and genotypes (Figure 1)

mice Publicly available data set of a **heterogeneous mice stock population** from `http://mus.well.ox.ac.uk/GSCAN` recently analyzed by e.g. [4] comprising 1940 individuals genotyped with 12545 SNP markers and 2527 individuals phenotyped for the traits weight and growth slope

References

[1] R Development Core Team (2010), `http://www.R-project.org/`

[2] vanRaden (2008) J Dairy Sci 91:4414-4423.

[3] Rogers (1972) In: Studies in genetics VII, Univ. of Texas, pp 145-153

[4] Legarra *et al.* (2008) Genetics 180:611-618

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

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

Contact: `Valentin.Wimmer@wzw.tum.de`
`www.synbreed.tum.de`