The 'synbreed' R package

Valentin Wimmer and Christina Lehermeier

Plant Breeding
Technische Universität München

November 8, 2012

Part 1: Introduction and data structure

Summary - synbreed R package

- Add-on for the open source environment for statistical computing
- Three example data sets available in synbreedData package
- Hosted on CRAN: http://cran.r-project.org/web/packages/synbreed/index.html

  R> install.packages("synbreed")
- Latest development version on R-Forge: https://r-forge.r-project.org/R/?group_id=710

  R> install.packages("synbreed",repos="http://r-forge.r-project.
- Recent R version required R $\geq$ 2.15.1
- All operating systems
- Once installed, load the package using

  R> library(synbreed)

Available documentation

- Publication in Bioinformatics (Wimmer *et al.* 2012)
- Package vignette

  ```
  R> vignette("IntroSyn")
  ```
- Website on R-Forge `http://synbreed.r-forge.r-project.org/`
- Manual and help sites:

  ```
  R> help(package="synbreed")
  ```
- Code demonstrations

  ```
  R> demo(package="synbreed")
  ```

Citation

- Please cite the synbreed package in your work, whenever you use it
- Recommended citation

```
R> citation(package="synbreed")

To cite package 'synbreed' in publications use:

  Wimmer V, Albrecht T, Auinger HJ and Schoen CC
  (2012) synbreed: a framework for the analysis of
  genomic prediction data using R. Bioinformatics,
  28: 2086-2087

A BibTeX entry for LaTeX users is

  @Article{,
    title = {synbreed: a framework for the analysis of genomic predi
    author = {Valentin Wimmer and Theresa Albrecht and Hans-Juergen
    journal = {Bioinformatics},
    year = {2012},
    volume = {28},
    number = {15},
    pages = {2086-2087}
```

Design objectives

1. User-friendly interface to analyze genomic prediction data
2. Analysis framework defined by a single, unified data object
3. One (open-source) software package
4. Flexible implementation (plant and animal breeding)
5. Gateway to other software and R packages
6. Teaching tool

Analysis pipeline

1. Data management and storage
2. Data processing: recoding, marker selection and imputing
3. Pedigree and marker-based coefficients of relatedness
4. Fit BLUP and Bayesian models by a unified interface
5. Model validation using cross-validation
6. Prediction of unphenotyped individuals
7. Data visualization

Overview

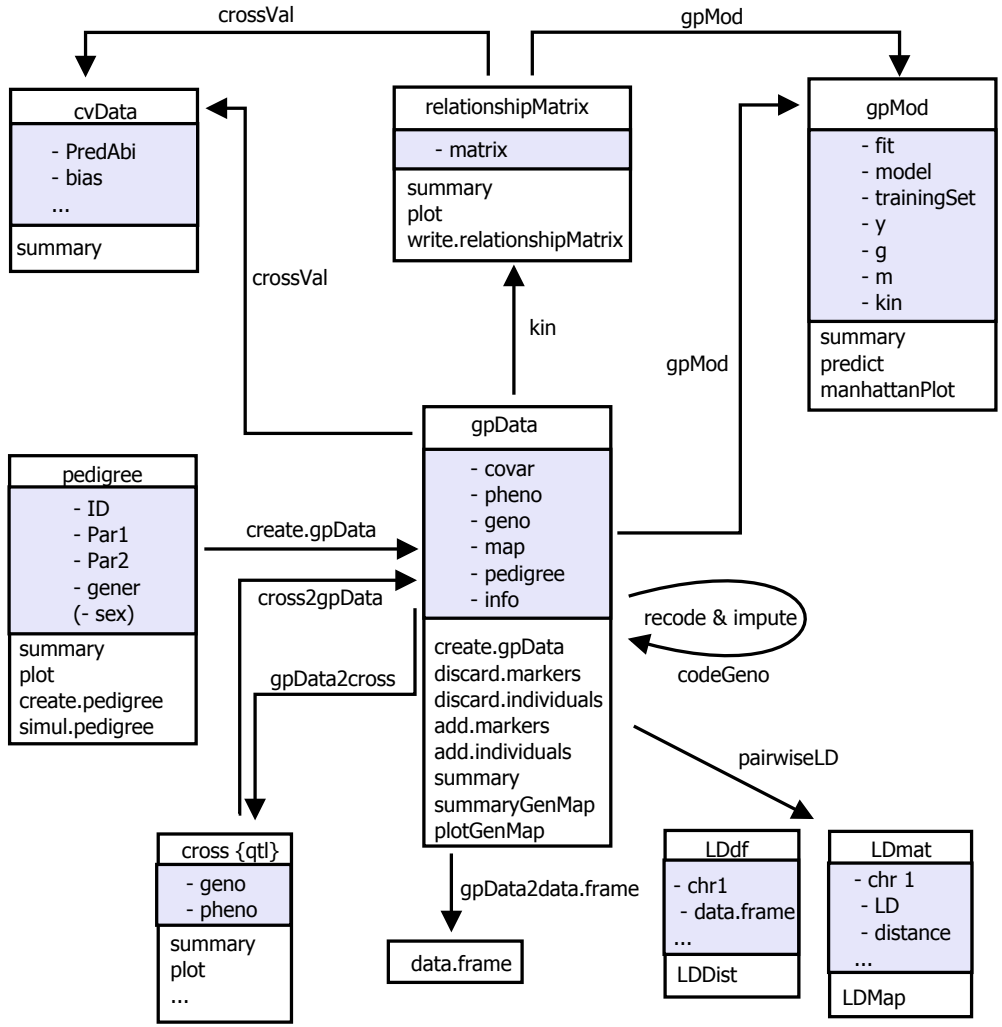


Figure: Classes, functions and methods of the `synbreed` R package

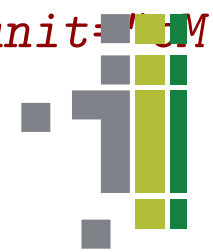
Data structure

- All data for genomic selection are combined in a single object
- Hence, easy data sharing, summary statistics, reduced storage requirements

class gpData

- `pheno` : `array` (3 dimensions) with phenotypes
- `geno` : `matrix` with genotypes (SNP markers)
- `map` : `data.frame` with marker map (chr + position)
- `pedigree` : class "pedigree"
- `covar` : `data.frame` with additional covariate information

```
R> gp <- create.gpData(pheno,geno,map,pedigree,covar,map.unit="cM")
```



Pedigree

ID	Par1	Par2	gener	sex
A	-	-	0	
B	-	-	0	
C	A	B	1	
D	A	C	2	
E	D	B	3	

- first generation = 0
- Create pedigree object

```
R> id <- c("A","B","C","D","E")
R> par1 <- c(0,0,"A","A","D")
R> par2 <- c(0,0,"B","C","B")
R> (ped <- create.pedigree(id,par1,par2))
  ID Par1 Par2 gener
1  A    0    0     0
2  B    0    0     0
3  C    A    B     1
4  D    A    C     2
```

Read-in of own data

- Simulated data from XII QTL-MAS Workshop 2008, Uppsala
- Available from `http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html`

QTLMAS data

- 50 simulated QTLs (explained variance 0 - 5 %)
- 5865 individuals (2778 males, 3087 females)
- 6000 markers on 6 chromosomes (each of length 100cM)

Create object of class gpData - 1

```
R> # Read file TrueEBV.txt with pedigree, trait, and tbv
R> dat <- read.table("TrueEBV.txt",header=TRUE,stringsAsFactors=FALS
R> # Create object of class 'pedigree'
R> ped <- with(dat,create.pedigree(ID=id,Par1=sire,Par2=dam,gener=ge
R> # Phenotypic data
R> pheno <- data.frame(trait=dat$Phenotype,row.names=dat$id)
R> # covar = tbv
R> covar <- data.frame(tbv=dat$GeneticValue,row.names=dat$id)
R> # genotypic data
R> geno <- read.table("genotype_cor.txt",header=FALSE,stringsAsFacto
```

Create object of class gpData - 2

```
R> # gametes to genotypes
R> geno2 <- matrix(data=NA,nrow=nrow(geno),ncol=(ncol(geno)-1)/2)
R> for (j in 1:ncol(geno2)){
+    # combine phased data to a genotype
+    geno2[,j] <- paste(as.character(geno[,2*j]),as.character(geno[
+  }
R> # create map
R> # 6 chromosomes with 1000 markers
R> # dist between adjacent markers = 0.1cM
R> chr <- rep(1:6,each=1000)
R> pos <- rep(seq(from=0,to=99.9,by=.1),times=6)
R> map <- data.frame(chr=chr,pos=pos)
R> # create gpData object
R> qtlMASdata <- create.gpData(pheno=pheno,geno=geno2,map=map,pedig
R> # save data as object of class gpData in Rdata-format
R> save("qtlMASdata",file="qtlMASdata.Rdata")
R> # for loading data, function load() and ls() might be useful
```

Example data sets

Maize data

- Simulated maize breeding program using DH technology
- 1250 DH lines phenotyped for one quantitative trait and 1117 SNPs

Mice data (Valdar *et al.* 2006)

- Heterogeneous stock mice population, publicly available from `http://gscan.well.ox.ac.uk`
- 2527 individuals with 2 traits (weight [g] at 6 weeks age and growth slope between 6 and 10 weeks age [g/day])
- 1940 individuals genotyped with 12545 SNP markers

Cattle data

- 50 individuals genotyped by 7250 SNP markers

The simulated `maize` data

Parameters

- 10 chromosomes of length 160 cM
- 500 segregating biallelic QTL with equal, additive effects
- Doubled-haploid (DH) lines
- 1250 individuals with genotypes (1117 SNPs) and phenotypes
- One quantitative trait evaluated in a testcross in 3 environments
- $h^2 = 0.46$
- Population structure: 25 biparental families of size 50

The maize data

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

object of class 'gpData'
covar
 No. of individuals 1610
         phenotyped 1250
          genotyped 1250
pheno
 No. of traits:    1

     Trait
 Min.   :120.7
 1st Qu.:142.8
 Median :148.9
 Mean   :148.9
 3rd Qu.:154.9
 Max.   :181.8

geno
 No. of markers 1117
 genotypes 0 1
 frequencies 0.339995 0.660005
 NA's 0.000 %
map
 No. of mapped markers  1117
 No. of chromosomes     10

 markers per chromosome

  1   2   3   4   5   6   7   8
 76  96  99 122  85 106 154 130 12

pedigree
Number of
 individuals  1610
 Par 1         219
 Par 2         221
 generations  15
```

Extract parts of the data

An object of class gpData is a `list`, see

```
R> str(maize)
```

- Look a the phenotypic data

```
R> head(maize$pheno[,1,])
```

```
 11360  11361  11362  11363  11364  11365
148.30 145.35 129.44 158.32 150.27 148.75
```

- Look a the genotypic data (individuals 10 to 13, markers 20 to 25)

```
R> maize$geno[10:13,20:25]
```

```
      M20 M21 M22 M23 M24 M25
11369   1   1   1   0   1   1
11370   0   1   1   0   1   1
11371   0   1   1   0   1   1
11372   1   1   1   0   1   1
```

The covar element

Generated within create.gpData, a `data.frame`

```
R> head(maize$covar,n=4)
```

```
     id phenotyped genotyped DH tbv family
1 10910      FALSE     FALSE  0  NA     NA
2 10918      FALSE     FALSE  0  NA     NA
3 10921      FALSE     FALSE  0  NA     NA
4 10924      FALSE     FALSE  0  NA     NA
```

- Column `id`: All names of individuals that either appear in `geno`, `pheno` or `pedigree`
- Column `genotyped`: Has the individual observations in `geno`?
- Column `phenotyped`:Has the individual observations in `pheno`?

Example: Extract all phenotyped individuals

```
R> maize$covar$id[maize$covar$phenotyped]
```

Remove and add markers/individuals

- `discard.individuals`
- `discard.markers`
- `add.individuals`
- `add.markers`

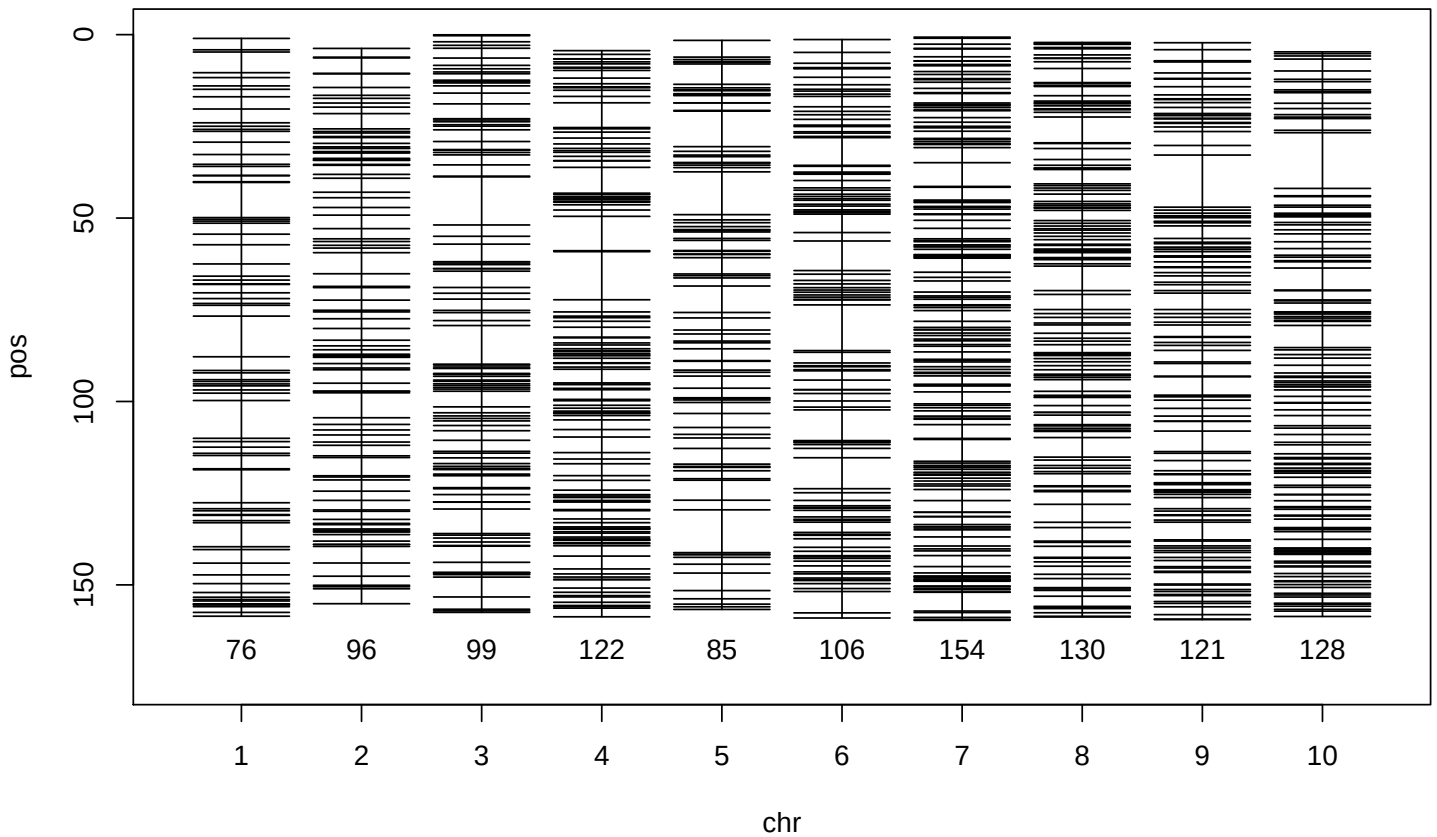
Example: Remove all markers from chromosome 6 to 10

```
R> maizeChr1to5 <- discard.markers(maize,rownames(maize$map)[maize$m
R> summary(maizeChr1to5$map)
```

```
      chr             pos
 Min.   :1.000   Min.   :  0.05
 1st Qu.:2.000   1st Qu.: 35.35
 Median :3.000   Median : 86.27
 Mean   :3.092   Mean   : 80.74
 3rd Qu.:4.000   3rd Qu.:121.44
 Max.   :5.000   Max.   :158.70
```

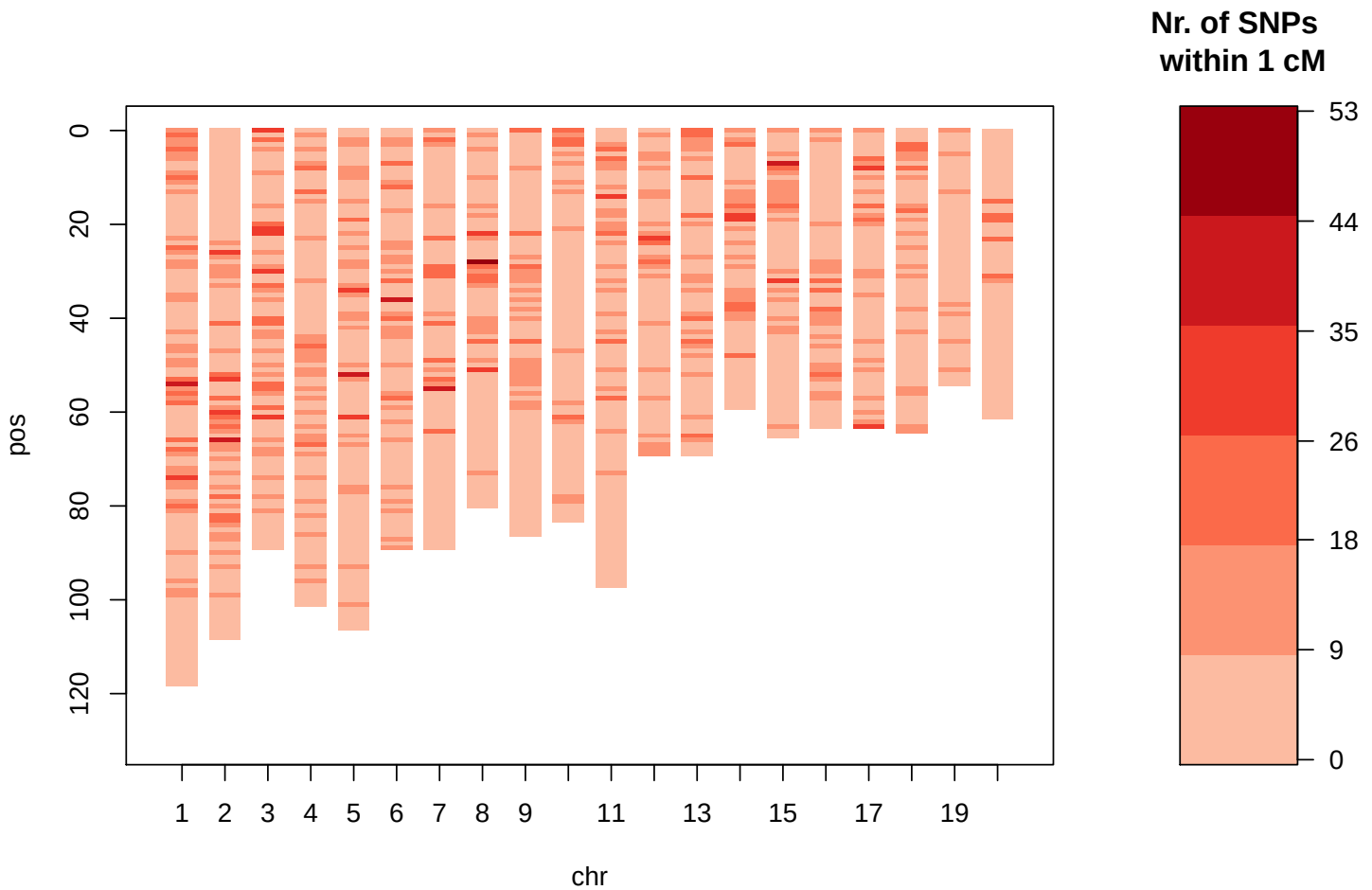
Visualization of marker map

```
R> plotGenMap(maize)
```



Visualization of marker map

```
R> plotGenMap(mice,dense=TRUE,nMarker = FALSE, bw=1)
```



Summary of marker map

```
R> summaryGenMap(maize)
```

	noM	length	avDist	maxDist	minDist
1	76	157.52	2.100267	11.08	0.10
2	96	151.38	1.593474	6.81	0.03
3	99	157.44	1.606531	13.11	0.02
4	122	154.34	1.275537	13.11	0.04
5	85	155.13	1.846786	11.67	0.01
6	106	157.70	1.501905	12.46	0.02
7	154	158.98	1.039085	6.48	0.02
8	130	156.62	1.214109	7.03	0.05
9	121	157.27	1.310583	14.21	0.06
10	128	153.92	1.211969	15.19	0.08
1 - 10	1117	1560.30	1.410027	15.19	0.01

Problems 1 - 1 (Corn borer example)

Please read: `http://www.rise.gs.tum.de/fileadmin/w00bjb/www/Risk_book_Chapters/SchoenWimmer_revised.pdf`

Table: Pedigree, phenotypic values, and marker genotypes for eight simulated maize individuals

Cycle	Individual	Pedigree	Tunnel length [cm]	SNP 1 (0)*	2 (1)	3 (-4)	4 (4)
1	I1	P1 × P2	13	2	2	0	1
1	I2	P3 × P4	17	0	0	0	1
1	I3	-	1	0	1	2	0
2	I4	I1 × I2	17	1	1	0	2
2	I5	I1 × I2	11	1	1	0	1
2	I6	I2 × I3	6	0	1	1	0
2	I7	I1 × I2	-	1	1	0	1
2	I8	I1 × I2	-	1	1	0	0

Problems 1 - 2 (Corn borer example)

1. Transfer the pedigree structure of the 8 individuals into an object of class `pedigree` and plot it.
2. Combine all data of the corn borer example in an object of class `gpData` called `cbData`. Include pedigree, phenotypes and genotypes (SNPs 1 to 4) and add the names of Table 1 for markers and individuals for all objects.
3. Use the `summary` method for this object. Are all details correct?
4. Compute a new object called `cbData2` excluding all individuals without phenotypes.
5. Use this data to compute a single marker regression for each SNP. Which markers are significant at the 5% error rate.
6. Set up a multiple marker regression model using (1) all SNPs and (2) only SNPs 3 and 4. Compare the results and discucss which model you would choose?

Part 2: Processing of marker data

Processing of marker data

Function codeGeno

1. Preselection of markers
2. Recode marker genotypes
3. Impute missing values

```
R> maizeC <- codeGeno(maize,maf=0.05,nmiss=0.1,
+  verbose=TRUE)
step 1  : 0 marker(s) removed with > 10 % missing values
step 2  : Recoding alleles
step 2.1: No duplicated markers discarded
step 5  : 122 marker(s) removed with maf < 0.05
step 6  : No duplicated markers discarded
End     : 995 marker(s) remain after the check
```

Compute pairwise LD measured as r^2 on chr 1

```
R> maizeLD <- pairwiseLD(maizeC,chr=1,type="data.frame")
```

Algorithm of codeGeno

```
R> codeGeno(gpData, impute = FALSE, impute.type = c("fix",
+       "random", "family", "Beagle", "BeagleAfterFamily"),
+       replace.value = NULL, maf = NULL, nmiss = NULL, label.heter
+        keep.identical = TRUE, verbose = FALSE)
```

1. Discard markers with fraction $>$ `nmiss` of missing values
2. Recode alleles as number of the minor alleles, i.e. 0, 1 and 2
3. Replace missing values by replace.value or impute missing values according to `impute.type`
4. Recode of alleles after imputation, if necessary due to changes in allele frequencies by imputed alleles
5. Discard markers with a minor allele frequency of $\leq$ `maf`
6. Discard duplicated markers if `keep.identical=FALSE`
7. Restore original data format (`gpData`, `matrix` or `data.frame`)

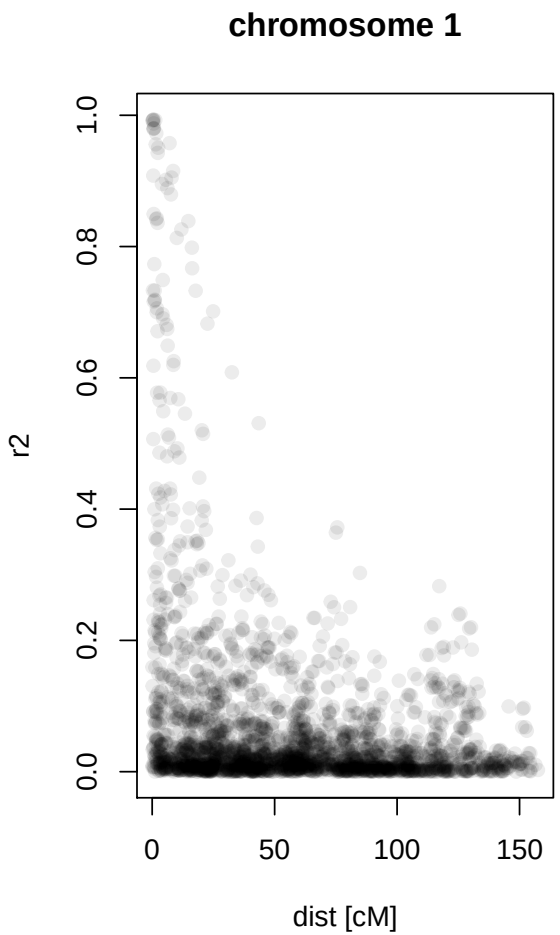
Imputing algorithms

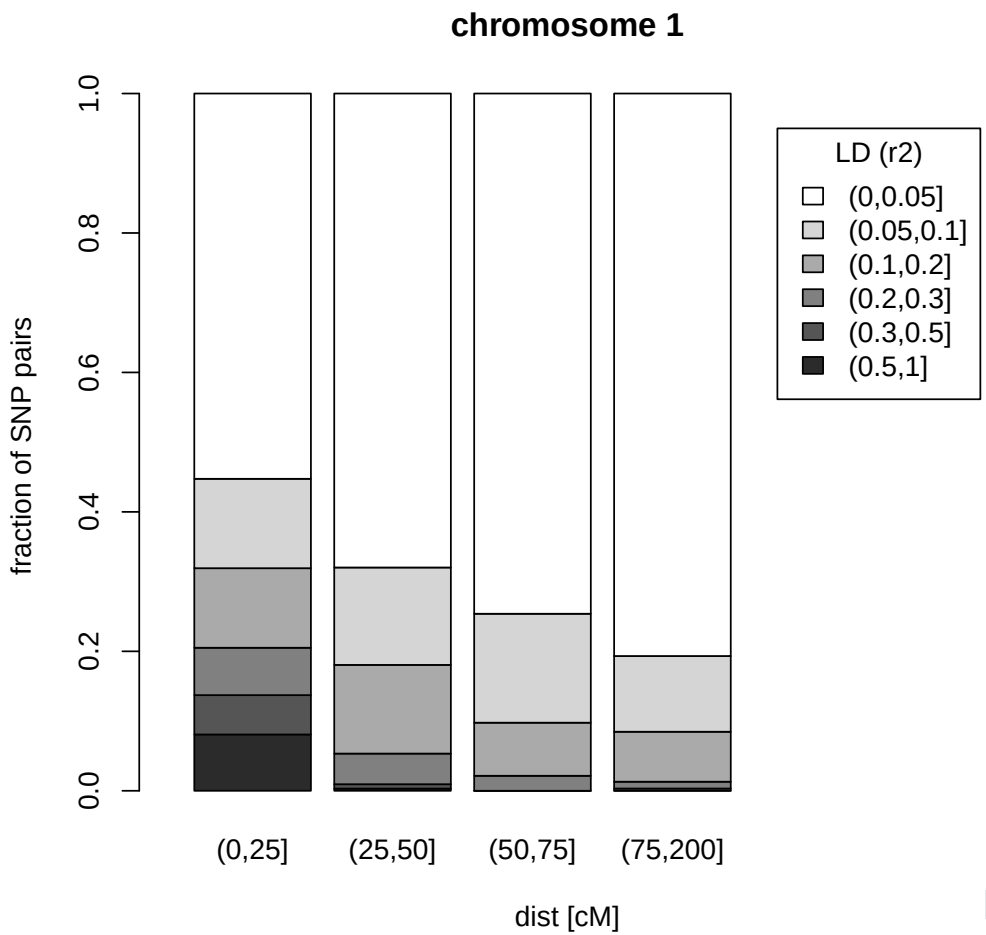
Gaps in the marker matrix can be filled according to

- Beagle (Browning and Browning 2009) (`impute.type = "beagle"`)
- Imputation within families (only for homozygous inbred lines according to Albrecht *et al.* (2011), `impute.type = "family"`)
- Beagle after family (`impute.type = "beagleAfterFamily"`)
- Random imputation according to the marginal allele distribution (`impute.type = "random"`)
- A fixed value (`impute.type = "fix"`)

Visualization of LD decay

```
R> plot(maizeLD); plot(maizeLD,type="bars")
```





Problems 2 - 1

1. Transfer your own data in to class `gpData` **or** use the Arabidopsis MAGIC lines population (Kover *et al.* 2009).
2. How many individuals are genotyped/phenotyped?
3. Make a new object an retain only those individuals which are phenotyped and genotyped.
4. From this object, remove all markers without a map position.
5. Make a visualization of the marker map. What is the largest gap between two markers?
6. Run and retrace the examples of the function `codeGeno`.
7. What are the observed alleles in your genotypic data? Recode your data into the number of copies of the minor allele. In the same step, remove all marker with more than 10% missing values or a MAF < 0.05. How many markers were removed according to these criteria? Use the argument `print.report=TRUE` in `codeGeno` and check the result.

Problems 2 - 2

1. If there are missing values, try to impute them using `Beagle`, if not possible, replace them according to the marginal allele distribution.
2. Make a histogram of the MAF. What is the median and mean of the MAF?
3. Compute the LD as r^2 using the gateway to `PLINK` (only for the first chromosome).
4. What is the minimum/mean/maximum LD between two markers? What is the proportion of markers with $r^2 > 0.20$?
5. Visualize the LD decay using a scatterplot and stacked histograms?
6. Try to add the nonlinear curve according to Hill and Weir (1988) to the scatterplot.
7. Make a LD heatmap for the first chromosome.

Part 3: Prediction and validation

Prediction models

Pedigree-based BLUP *PBLUP*

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e} \\ \mathbf{a} &\sim \mathrm{N}(\mathbf{0}, \mathbf{A}\sigma_a^2) \end{aligned}$$

Marker-based BLUP *GBLUP*

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \\ \mathbf{u} &\sim \mathrm{N}(\mathbf{0}, \mathbf{U}\sigma_u^2) \\ \mathbf{U} &= \frac{(\mathbf{W}-\mathbf{P})(\mathbf{W}-\mathbf{P})^\top}{2\sum_{j=1}^{p} p_j(1-p_j)} \end{aligned}$$

with

- $\mathbf{y}$ Vector of phenotypic records
- $\mathbf{W}$ Marker matrix
- $\mathbf{P}$ Matrix with the allele frequencies p_j
- $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$ vector of residuals

Estimation of relatedness

- Pedigree based (expected) and realized kinship coefficients: function `kin`
  - Additive numerator relationship matrix **A** (default)

    ```
    R> kin(gpData,ret="add")
    ```
  - Dominance relationship matrix **D**

    ```
    R> kin(gpData,ret="dom")
    ```
  - Kinship matrix $\mathbf{K} = \frac{1}{2}\mathbf{A}$

    ```
    R> kin(gpData,ret="kin")
    ```
  - Gametic relationship matrix (dimension $2n \times 2n$)

    ```
    R> kin(gpData,ret="gam")
    ```

Kinship for the 1250 DH lines

```
R> A <- kin(maizeC,ret="kin",DH=maize$covar$DH)
```

Special case

The phenotypes in the `maize` data origin from a testcross of DH lines, hence (Albrecht *et al.* 2011)

- The additive relationship matrix must be replaced by the kinship
- The variance of the marker genotypes is

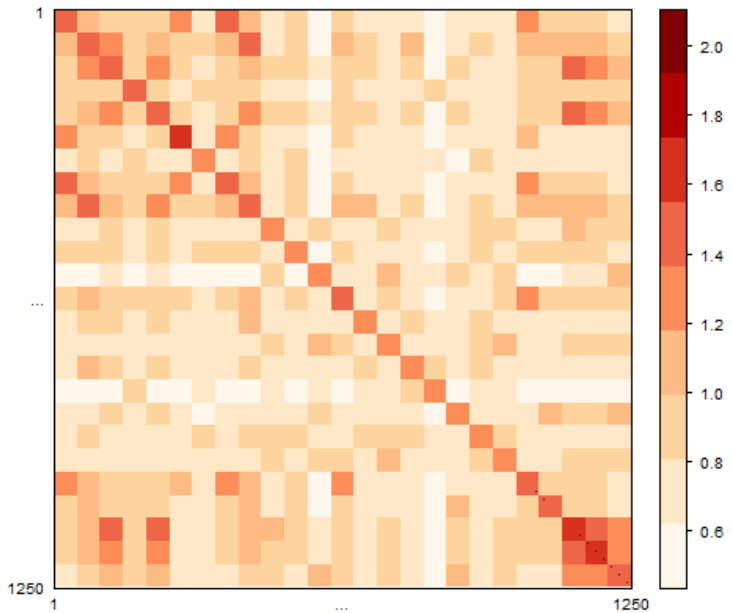
$$4 \sum_{j=1}^{p} p_j(1-p_j)$$

- Thus

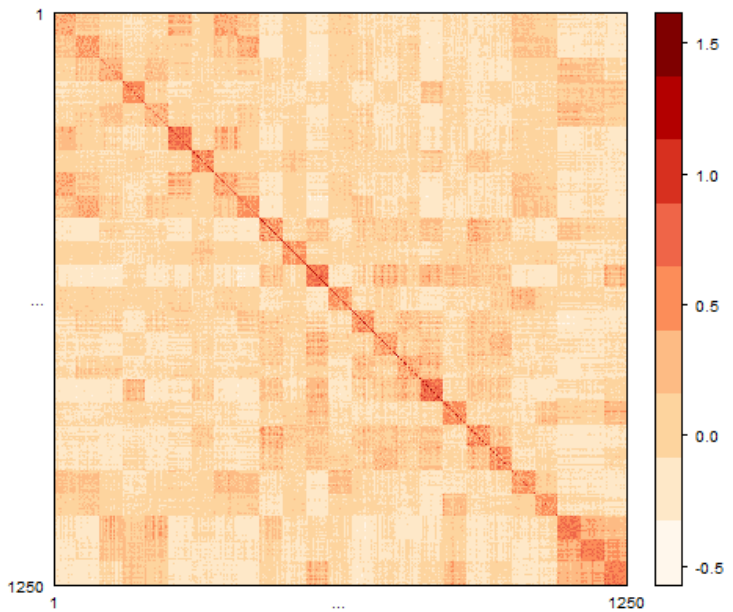
$$\mathbf{U} = \frac{(\mathbf{W}-\mathbf{P})(\mathbf{W}-\mathbf{P})^\top}{4\sum_{j=1}^{p} p_j(1-p_j)}$$

Example

```
R> U <- kin(maizeC,ret="realized")/2
R> plot(A[maize$covar$genotyped,maize$covar$genotyped]); plot(U)
```



(a) Pedigree-based relationship



(b) Marker-based relationship

Equation

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where

$\mathbf{y}$	is a vectors of phenotypes
$\mathbf{X}$	is a design matrix allocating phenotypes to fixed effects
β	is a vector of fixed effects
$\mathbf{Z}$	is a design matrix allocating phenotypes to random effects
$\mathbf{u}$	is a vector of random effects, with $\mathbf{u} \sim \mathrm{N}(0, \mathbf{G}\sigma^2)$
$\mathbf{e}$	is a vector of residuals, with $\mathbf{e} \sim \mathrm{N}(0, \mathbf{R}\sigma^2)$

Expected values and variances

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

$$\mathrm{E}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathrm{Var}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix}\sigma^2$$

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\beta$$

$$\mathrm{Var}(\mathbf{y}) = \mathbf{V} = (\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})\sigma^2$$

Solutions

- Mixed Model Equations (MME):

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

- $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$
  where $\hat{\beta}$ is a generalized least square solution
  $\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
- Difference to the least square estimate of a LM ($\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$) is the decomposition of $\mathbf{V}$
- $\mathbf{V}$ has to be estimated

Solutions

- Mixed Model Equations (MME):

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

- $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$
  where $\hat{\beta}$ is a generalized least square solution
  $\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
- Difference to the least square estimate of a LM ($\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$) is the decomposition of $\mathbf{V}$
- $\mathbf{V}$ has to be estimated

Solutions

- Mixed Model Equations (MME):

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

- $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$
  where $\hat{\beta}$ is a generalized least square solution
  $\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
- Difference to the least square estimate of a LM ($\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$) is the decomposition of $\mathbf{V}$
- $\mathbf{V}$ has to be estimated

Solutions

- Mixed Model Equations (MME):

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

- $\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})$
  where $\hat{\beta}$ is a generalized least square solution
  $\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
- Difference to the least square estimate of a LM ($\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$) is the decomposition of $\mathbf{V}$
- $\mathbf{V}$ has to be estimated

Example from Henderson (1977)

y	time	animal
132	1	1
147	2	2
156	1	3
172	2	4

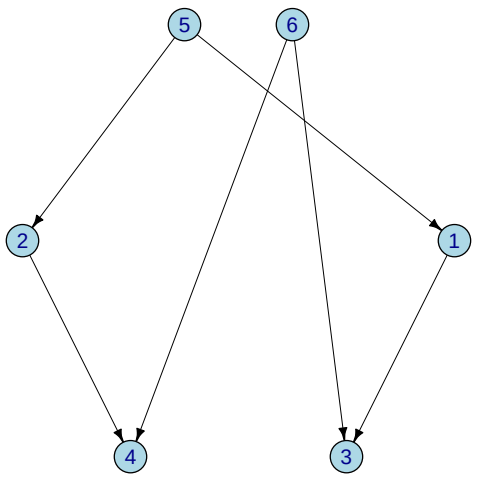


Figure: Pedigree

Equation

$$
\begin{aligned}
y_{ij} &= \beta_i + u_j + e_{ij} \\
\mathbf{y} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}
\end{aligned}
$$

with

$\mathbf{y}' = (y_{11}, y_{22}, y_{13}, y_{24})$	observations
$\beta' = (\beta_1, \beta_2)$	time effects (fix)
$\mathbf{u}' = (u_1, u_2, u_3, u_4)$	additive genetic merit (random) with $\mathbf{u} \sim \mathrm{N}(0, \mathbf{A}\sigma_u^2)$
$\mathbf{e}' = (e_{11}, e_{22}, e_{13}, e_{24})$	residuals (random) with $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma^2)$
$\mathbf{X}, \mathbf{Z}$	design matrices

$$
\begin{bmatrix} 132 \\ 147 \\ 156 \\ 172 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{22} \\ e_{13} \\ e_{24} \end{bmatrix}
$$

Equation

$$
\begin{aligned}
y_{ij} &= \beta_i + u_j + e_{ij} \\
\mathbf{y} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}
\end{aligned}
$$

with

$\mathbf{y}' = (y_{11}, y_{22}, y_{13}, y_{24})$	observations
$\beta' = (\beta_1, \beta_2)$	time effects (fix)
$\mathbf{u}' = (u_1, u_2, u_3, u_4)$	additive genetic merit (random) with $\mathbf{u} \sim \mathrm{N}(0, \mathbf{A}\sigma_u^2)$
$\mathbf{e}' = (e_{11}, e_{22}, e_{13}, e_{24})$	residuals (random) with $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma^2)$
$\mathbf{X}, \mathbf{Z}$	design matrices

$$
\begin{bmatrix} 132 \\ 147 \\ 156 \\ 172 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{22} \\ e_{13} \\ e_{24} \end{bmatrix}
$$

Expected values and variances

$$\mathrm{E}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathrm{Var}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix}\sigma^2 = \begin{bmatrix} \mathbf{A}\frac{\sigma_u^2}{\sigma^2} & 0 \\ 0 & \mathbf{I} \end{bmatrix}\sigma^2$$

Assumption

$$\begin{aligned} h^2 &= \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} = 0.25 \\ &\Rightarrow \quad e.g.\ \sigma_u^2 = 0.25 \text{ and } \sigma^2 = 0.75 \\ &\Rightarrow \quad \mathbf{G}^{-1} = \mathbf{A}^{-1}\frac{\sigma^2}{\sigma_u^2} = 3\mathbf{A}^{-1} \end{aligned}$$

Numerator relationship matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0.25 & 0.5 & 0.125 \\ & 1 & 0.125 & 0.5 \\ & & 1 & 0.3125 \\ & & & 1 \end{bmatrix}; \; 3\mathbf{A}^{-1} = \begin{bmatrix} 4.325 & -1.175 & -2.25 & 0.75 \\ -1.175 & 4.325 & 0.75 & -2.25 \\ -2.250 & 0.750 & 4.50 & -1.50 \\ 0.750 & -2.250 & -1.50 & 4.50 \end{bmatrix}$$

Expected values and variances

$$\mathrm{E}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathrm{Var}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix}\sigma^2 = \begin{bmatrix} \mathbf{A}\frac{\sigma_u^2}{\sigma^2} & 0 \\ 0 & \mathbf{I} \end{bmatrix}\sigma^2$$

Assumption

$$\begin{aligned} h^2 &= \frac{\sigma_u^2}{\sigma_u^2+\sigma^2} = 0.25 \\ &\Rightarrow \quad e.g.\ \sigma_u^2 = 0.25 \text{ and } \sigma^2 = 0.75 \\ &\Rightarrow \quad \mathbf{G}^{-1} = \mathbf{A}^{-1}\frac{\sigma^2}{\sigma_u^2} = 3\mathbf{A}^{-1} \end{aligned}$$

Numerator relationship matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0.25 & 0.5 & 0.125 \\ & 1 & 0.125 & 0.5 \\ & & 1 & 0.3125 \\ & & & 1 \end{bmatrix}; 3\mathbf{A}^{-1} = \begin{bmatrix} 4.325 & -1.175 & -2.25 & 0.75 \\ -1.175 & 4.325 & 0.75 & -2.25 \\ -2.250 & 0.750 & 4.50 & -1.50 \\ 0.750 & -2.250 & -1.50 & 4.50 \end{bmatrix}$$

Expected values and variances

$$\begin{aligned} \mathrm{E}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \mathrm{Var}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} &= \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix}\sigma^2 = \begin{bmatrix} \mathbf{A}\frac{\sigma_u^2}{\sigma^2} & 0 \\ 0 & \mathbf{I} \end{bmatrix}\sigma^2 \end{aligned}$$

Assumption

$$\begin{aligned} h^2 &= \frac{\sigma_u^2}{\sigma_u^2+\sigma^2} = 0.25 \\ &\Rightarrow \quad e.g.\ \sigma_u^2 = 0.25 \text{ and } \sigma^2 = 0.75 \\ &\Rightarrow \quad \mathbf{G}^{-1} = \mathbf{A}^{-1}\frac{\sigma^2}{\sigma_u^2} = 3\mathbf{A}^{-1} \end{aligned}$$

Numerator relationship matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0.25 & 0.5 & 0.125 \\ & 1 & 0.125 & 0.5 \\ & & 1 & 0.3125 \\ & & & 1 \end{bmatrix}; 3\mathbf{A}^{-1} = \begin{bmatrix} 4.325 & -1.175 & -2.25 & 0.75 \\ -1.175 & 4.325 & 0.75 & -2.25 \\ -2.250 & 0.750 & 4.50 & -1.50 \\ 0.750 & -2.250 & -1.50 & 4.50 \end{bmatrix}$$

Solution

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma^2}{\sigma_{\mathbf{u}}^2} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \left[\begin{array}{cc|cccc} 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 5.325 & -1.175 & -2.250 & 0.750 \\ 0 & 1 & -1.175 & 5.325 & 0.750 & -2.250 \\ 1 & 0 & -2.250 & 0.750 & 5.500 & -1.500 \\ 0 & 1 & 0.750 & -2.250 & -1.500 & 5.500 \end{array}\right]^{-} \left[\begin{array}{c} 288 \\ 319 \\ \hline 132 \\ 147 \\ 156 \\ 172 \end{array}\right]$$

Results

$$\hat{\beta} = \begin{bmatrix} 143.89 \\ 159.40 \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{u}} = \begin{bmatrix} -2.07 \\ -2.12 \\ 2.28 \\ 2.33 \end{bmatrix}$$

Solution

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma^2}{\sigma_{\mathbf{u}}^2} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} = \left[\begin{array}{cc|cccc} 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 5.325 & -1.175 & -2.250 & 0.750 \\ 0 & 1 & -1.175 & 5.325 & 0.750 & -2.250 \\ 1 & 0 & -2.250 & 0.750 & 5.500 & -1.500 \\ 0 & 1 & 0.750 & -2.250 & -1.500 & 5.500 \end{array}\right]^{-} \left[\begin{array}{c} 288 \\ 319 \\ \hline 132 \\ 147 \\ 156 \\ 172 \end{array}\right]$$

Results

$$\hat{\beta} = \begin{bmatrix} 143.89 \\ 159.40 \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{u}} = \begin{bmatrix} -2.07 \\ -2.12 \\ 2.28 \\ 2.33 \end{bmatrix}$$

Function MME

```
R> dat <- data.frame(y=c(132,147,156,172),time=c(1,2,1,2),animal=c(
R> ped <- create.pedigree(ID=c(6,5,1,2,3,4),Par1=c(0,0,5,5,1,6),Par
R> gp <- create.gpData(pheno=dat,pedigree=ped)
R> A <- kin(gp,ret="add")
R> (X <- matrix(c(1,0,1,0,0,1,0,1),ncol=2))
     [,1] [,2]
[1,]    1    0
[2,]    0    1
[3,]    1    0
[4,]    0    1
R> (Z <- diag(6)[-c(1,2),])
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    1    0    0    0
[2,]    0    0    0    1    0    0
[3,]    0    0    0    0    1    0
[4,]    0    0    0    0    0    1
```

Function MME

```
R> (AI <- solve(A))
```

```
           5    6           1           2  3  4
5  1.6666667  0.0 -0.6666667 -0.6666667  0  0
6  0.0000000  2.0  0.5000000  0.5000000 -1 -1
1 -0.6666667  0.5  1.8333333  0.0000000 -1  0
2 -0.6666667  0.5  0.0000000  1.8333333  0 -1
3  0.0000000 -1.0 -1.0000000  0.0000000  2  0
4  0.0000000 -1.0  0.0000000 -1.0000000  0  2
```

```
R> RI <- diag(4)
R> res <- MME(X,Z,AI*3,RI,dat$y)
R> res$b; res$u
```

```
[1] 143.8930 159.3976
```

```
[1] -1.675214  3.350427 -2.065980 -2.122054  2.280054
[6]  2.326783
```

Example

- Fit models

```
R> modA <- gpMod(maizeC,model="BLUP",kin=A)
R> modU <- gpMod(maizeC,model="BLUP",kin=U)
```

- Predicted genetic values

```
R> gA <- predict(modA)
R> gU <- predict(modU)
```

- Extract true breeding values

```
R> tbv <- maizeC$covar$tbv[maizeC$covar$phenotyped]
```

- Evaluate correlations with tbv

```
R> cor(gA,tbv)

0.587

R> cor(gU,tbv)

0.856
```

```
R> summary(modU)

Object of class 'gpMod'
Model used: BLUP
Nr. observations 1250
Genetic performances:
   Min.    1st Qu.  Median   Mean    3rd Qu.  Max
-19.4200 -3.4210 -0.2841  0.0000  3.2830 15.3000
--
Model fit
Likelihood kernel: K = (Intercept)

Maximized log likelihood with kernel K is  -3223.837

Linear Coefficients:
             Estimate Std. Error
 (Intercept)  148.921      0.197

Variance Coefficients:
             Estimate Std. Error
       kinTS   53.055      7.359
       In      48.577      2.287
```

Prediction of unphenotyped individuals

Discard last 50 individuals from the data set

```
R> last50 <- rownames(maizeC$pheno)[1201:1250]
R> maizeC2 <- discard.individuals(maizeC,last50)
```

Fit *modU* using the variance-covariance structure from the whole data set

```
R> modU24 <- gpMod(maizeC2,model="BLUP",kin=U)
```

Prediction for the last 50 individuals

```
R> g <- predict(modU24,rownames(maizeC$pheno)[1201:1250])
```

Model cross-validation

- Prospects for GS derived by out-of-sample performance
- Cross-validation as assumption-free method
- Divide data set in k mutually exclusive subsets
- $k-1$ form the estimation set (ES), kth subset is used as independent test set (TS)
- Model validation by
  - Predictive ability $r(\hat{g}_{TS}, y_{TS})$
  - Prediction bias
- Sampling schemes: random, within family, across family (Albrecht *et al.* 2011)

Example

```
R>  cv.maize <- crossVal(maizeC,cov.matrix=list(U),k=5,Rep=2,Seed=1
R> summary(cv.maize)
Object of class 'cvData'

 5 -fold cross validation with 2 replications
     Sampling:                    random
     Variance components:         committed
     Number of random effects: 1
     Number of individuals:       1250
     Size of the TS:              250 -- 250

Results:
                          Min          Mean +- pooled SE      Max
 Predictive ability:  0.4589       0.5287 +- 0.0079       0.5691
 Bias:                0.8747       1.0061 +- 0.0253       1.1179
```

Bayesian Lasso

The model (de los Campos *et al.* 2009)

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

with the prior distributions (Park and Casella 2008)

$$\begin{aligned}
\beta_j &\sim \mathrm{N}(0, \sigma^2\tau_j^2) \\
\tau_j^2 &\sim \mathrm{Exp}(\lambda^2) \\
\lambda^2 &\sim \mathrm{Ga}(\alpha, \beta) \text{ or } \frac{\lambda}{\lambda_{\max}} \sim \mathrm{Beta}(a, b) \\
\sigma^2 &\sim \chi^{-2}(\nu, S)
\end{aligned}$$

Choice of hyperparameters

According to Pérez *et al.* (2010):

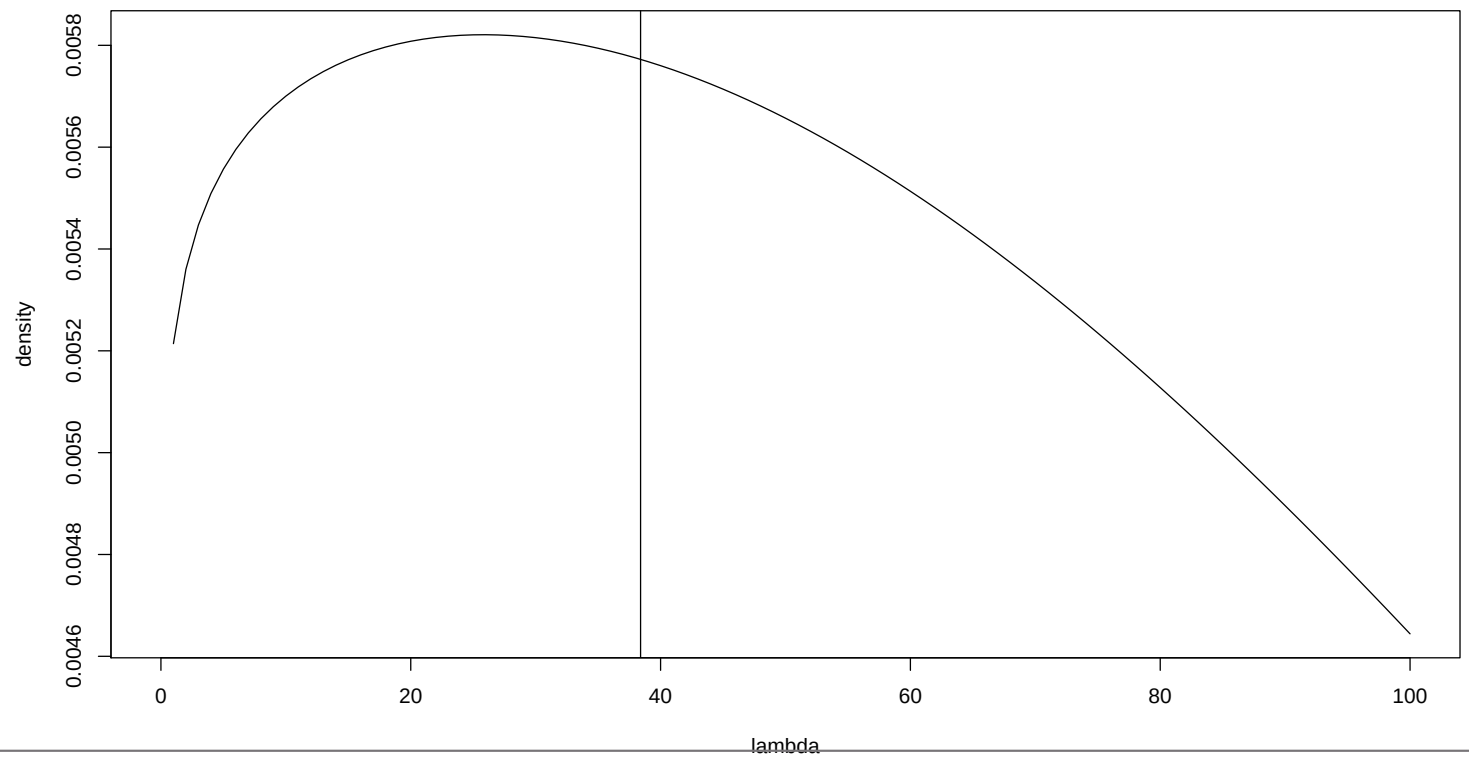
$$\lambda_{start} = \sqrt{2 \sum_{j=1}^{p} \bar{x}_{.j}^2 \frac{(1-h^2)}{h^2}}$$

```
R> y <- maize$pheno[,1,]
R> X <- maize$geno
R> sX2 <- sum(X^2)
R> h2 <- 0.5 # priori expectation
R> (lambdaStart <- sqrt(2*sum(X^2)*(1-h2)/h2/nrow(X)))
```

```
[1] 38.39858
```

Choice of hyperparameters

```
R> lambda <- seq(from=0,to=100,by=1)
R> dens <- dgamma(x=lambda^2,shape=.52,rate=3e-5)*lambda*2 # distril
R> plot(dens~lambda,type='l',ylab="density")
R> abline(v=lambdaStart)
```



Run Bayesian Lasso

Evaluation on the whole data set:

```
R> prior <- list(varE=list(df=3,S=35),lambda = list(shape=0.52,rate=
R> modBL <- gpMod(maizeC,model="BL",prior=prior,nIter=6000,burnIn=10
```

Use CV to evaluate the predictive ability:

```
R> cv.BL <- crossVal(maizeC,k=5,Rep=2,Seed=123,sampling="random",VC
R> summary(cv.BL)
```

Gateway from synbreed to package qtl

- Package qtl for QTL analysis in experimental crosses (Broman *et al.* 2003)
- Main data class cross
- Conversion from gpData to cross

  ```
  R> gpData2cross(gpDataObj)
  ```

- Conversion from cross to gpData

  ```
  R> cross2gpData(crossObj)
  ```

Problems 3 - 1 (Corn borer example)

1. Try to reproduce the results of Illustration 3.1 using function `MME`.
2. Combine the pedigree, the phenotypes and genotypes in an object of class `gpData`.
3. Set up the matrix **A** for the individuals and plot a heatmap of it.
4. Try to reproduce the results of Illustration 4.1 using function `MME`.
5. Set up the matrix **U** for the individuals and plot a heatmap of it. Discuss the the differences with regard to contents compared to the matrix **A**.

Problems 3 - 2

1. Construct a genomic relationship matrix **U** according to Habier *et al.* (2007) and fit a GBLUP model. What are the estimated variance components?
2. Make a manhattan plot of the estimated marker effects.
3. Predict the unphenotyped individuals in your data set using the `predict` method for the GBLUP model. If all individuals are phenotyped, mask 10% of the phenotypes an predict them.
4. Use CV to routinely estimate the predictive ability of the GBLUP model in your data. Commit for each CV model the variance components estimated with the whole data set.
5. What is the definition of the bias in the summary of the CV? Try to interpret the values you obtain with your data.

Problems 3 - 3 (Advanced)

1. Try to fit different types of genomic relationship matrices using the function `kin`. Use them in a linear mixed model as variance-covariance structure (using function `gpMod`) and compare the variance components you obtain. For further connections between the matrices, see Albrecht *et al.* (2011). Use CV to estimate the predictive ability of the different models. What do you observe?
2. Check the help for function `MME`. Try to replicate the results from problem 1. First, you need to extract the necessary parts from the `gpData` object. Next, you need to set up the variance-covariance structure using the **U** matrix and the estimated variance components from problem 1.
3. Use the function `gpData2cross` to convert your object to an object of class `cross` for package `qtl` .
4. Use the function `scanone` of package to scan for QTLs and display the LOD curve you obtain along the genome.

Literature

Albrecht, T., V. Wimmer, H.-J. Auinger, M. Erbe, C. Knaak, *et al.*, 2011 Genome-based prediction of testcross values in maize. Theoretical and Applied Genetics **123**: 339 – 350.

Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: Qtl mapping in experimental crosses. Bioinformatics **7**: 889–890. R package version 1.20-15.

Browning, B. L., and S. R. Browning, 2009 A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics **846**: 210–223. Version 3.3.1.

de los Campos, G., H. Naya, D. Gianola, A. L. José Crossa, E. Manfredi, *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics **182**: 375–385.

Habier, D., R. Fernando, and J. Dekkers, 2007 The impact of genetic relationship information on Genome-Assisted breeding values. Genetics **177**: 2389 – 2397.

Henderson, C., 1977 Best linear unbiased prediction of breeding values not in the model for records. Journal of Dairy Science **60**: 783–787.

Kover, P. X., W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, *et al.*, 2009 A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet **5**: e1000551.

Park, T., and G. Casella, 2008 The bayesian lasso. Journal of the American Statistical Association **103**: 681 – 686.

Pérez, P., G. de los Campos, J. Cross, and D. Gianola, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the blr package in r. The Plant Genome **3**: 106 – 116.

Valdar, W., L. Solberg, D. Gauguier, W. Cookson, J. Rawlins, *et al.*, 2006 Genetic and environmental effects on complex traits in mice. Genetics **174**: 959–984.

Wimmer, V., T. Albrecht, H.-J. Auinger, and C.-C. Schoen, 2012 synbreed: a framework for the analysis of genomic prediction data using r. Bioinformatics **28**: 2086–2087.