# The RepeatABEL package - a tutorial

Lars Rönnegård

November 3, 2015

## Introduction

This vignette gives an introduction to the RepeatABEL package. The package performs GWAS for data where there are repeated observations on individuals that may be related. Various random effects can be fitted. Random polygenic effects are fitted by default, but also other random effects can be fitted including spatial effects. A model without the fixed SNP effect is fitted using the hglm package for estimating variance components (see *Ronnegard, Shen & Alam (2010) hglm: A package for fitting hierarchical generalized linear models. The R Journal 2:20-28*). These variance component estimates are subsequently used in the GWAS where each marker is fitted one at a time. Thus, this is similar to using the `polygenic_hglm` function and subsequently the `mmscore` function in GenABEL. The package consists of three main functions `rGLS`, `preFitModel` and `simulate_PhenData`.

## Theory for the `rGLS` function

The `rGLS` function is the main function of the package and by default fits a linear mixed model including permanent environmental effects $\boldsymbol{p}$ and polygenic effects $\boldsymbol{g}$ with correlation matrix given by the genomic relationship matrix (aka kinship matrix) in

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{Zp} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is the studied trait, $\boldsymbol{\beta}$ are fixed effects, $\boldsymbol{e}$ are residuals with $\boldsymbol{e} \sim N(0, \sigma_e^2)$ having residual variance $\sigma_e^2$. Furthermore, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are model matrices. The random effects are assumed multivariate normal such that $\boldsymbol{p} \sim N(0, \boldsymbol{I}\sigma_p^2)$ and $\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$ where $\boldsymbol{G}$ is the genomic relationship matrix. Thus the estimated (co)variance matrix for this model is

$$\hat{\boldsymbol{V}} = \boldsymbol{ZGZ}^T\hat{\sigma}_g^2 + \boldsymbol{ZZ}^T\hat{\sigma}_p^2 + \boldsymbol{I}\hat{\sigma}_e^2. \tag{2}$$

Subsequently, a linear model is fitted (using generalized least squares, GLS) for each marker where the covariate $x_{SNP}$ is coded as 0,1,2:

$$\boldsymbol{y} = \boldsymbol{X\beta} + x_{SNP}\beta + \boldsymbol{\varepsilon} \tag{3}$$

with

$$\boldsymbol{\varepsilon} \sim N(0, \sigma_\epsilon^2\hat{\boldsymbol{V}}). \tag{4}$$

A Wald test is used to compute the P value for SNP effect $\beta$.

The computations are made fast by applying a eigen-decomposition of $\hat{V}$ and using the built-in `qr` function in R to fit the linear models (3).

## Using the `rGLS` function

The following example illustrates the use of the function. There are two data objects to be included in the input: a GenABEL object including the genotypic information and a data frame including the phenotypic information. The name of the ID variable in the phenotype data frame should be ”id” (otherwise specify `id.name` equal to the ID variable name).

In this example there are 360 observations from 100 individuals, and there are 5792 SNP to be tested.

```
> library(RepeatABEL)
> #GenABEL object including IDs and marker genotypes
> data(gen.data)
> #Phenotype data with repeated observations
> data(Phen.Data)
```

The data frame `Phen.Data` includes the trait value y, two covariates (age and sex) and the ID of the individuals. We wish to include age and sex as fixed effects so the function input is

```
> GWAS1 <- rGLS(y ~ age + sex, genabel.data = gen.data,
    phenotype.data = Phen.Data)
[1] "GRM ready"
[1] "Variance component estimation ready"
[1] "Rotation matrix ready"
[1] "Rotate LMM started"
[1] "Rotate LMM ready"
```

The computations in this function consists of four parts: construction of the GRM, variance component estimation using the `hglm` function, rotating the GLS using eigen-decomposition transforming it to an ordinary least squares (OLS) problem, and finally fitting an OLS for each SNP. How far the computations have come is shown in the output.

The class of the output object GWAS1 is `gwaa.scan`, so we can apply the generic functions `summary()` and `plot()` defined by the GenABEL package.

```
> summary(GWAS1)
Summary for top 10 results, sorted by P1df
          Chromosome Position Strand A1 A2       effB
rs120315           1  3725352      +  T  C  2.1696316
rs9670687          1  4779800      -  C  A -2.0412127
rs891586           2  8024318      +  A  G -1.6669267
rs1352451          2  8022651      -  A  C  1.6169429
rs1252282          2  8167984      -  A  T -0.9982089
rs9922492          1  3729983      +  C  T -1.8586037
rs7633966          1  3721952      +  C  T  1.8586037
rs4378234          1  4800348      -  G  T -1.9012073
rs7499832          3 10242953      -  G  C -1.0897248
rs6561272          1  3715538      -  T  G -1.9073305
            se_effB chi2.1df         P1df Pc1df
rs120315  0.4258979       NA 3.501204e-07    NA
rs9670687 0.5351582       NA 1.366120e-04    NA
rs891586  0.4478499       NA 1.975997e-04    NA
rs1352451 0.4564078       NA 3.959643e-04    NA
rs1252282 0.2825403       NA 4.109058e-04    NA
rs9922492 0.5325691       NA 4.832322e-04    NA
```

```
rs7633966 0.5325691       NA 4.832322e-04    NA
rs4378234 0.5509388       NA 5.588239e-04    NA
rs7499832 0.3182487       NA 6.167705e-04    NA
rs6561272 0.5726170       NA 8.656541e-04    NA
```

### Extracting the genotypic variance, a 95% CI and the heritability

From the GWAS1 object, the estimated variance components for the prefitted model, not including SNP effects, can be extracted as follows.

```
> est.hglm <- GWAS1@call$hglm
> cat("Genotypic and permanent env. variance components:","\n",
        est.hglm$varRanef,", resp.","\n",
        "The residual variance is", est.hglm$varFix,".","\n")
Genotypic and permanent env. variance components:
 1.376368 1.354224 , resp.
 The residual variance is 3.090352 .
```

Furthermore, the standard errors for the estimated variance components are computed on a log scale. Thereby, 95% confidence intervals can be computed for the variance components. Here the computations for the genotypic variance are shown.

```
> logVCE <- est.hglm$SummVC2[[1]]
> cat("Estimate and SE for the genotypic variance on a natural log scale:",
        "\n", logVCE[1],"and",logVCE[2],", resp.", "\n \n",
        "Confidence interval: [", exp( logVCE[1] - 1.96*logVCE[2] ) , "," ,
        exp( logVCE[1] + 1.96*logVCE[2] ) , "]" , "\n")
Estimate and SE for the genotypic variance on a natural log scale:
 0.3194482 and 0.2494617 , resp.

 Confidence interval: [ 0.8440896 , 2.244299 ]
```

The heritability for this example is computed as:

```
> cat("Heritability:",
        est.hglm$varRanef[1]/(est.hglm$varFix + est.hglm$varRanef[2]),
        "\n")
Heritability: 0.3096737
```

## Using the preFitModel function

The function preFitModel is used for variance component estimation and increases the flexibility of modeling in the rGLS function.

## Fitting the same model as above

To start with we have a look at how the model in the previous section can be fitted in two steps using the `preFitModel` function and thereafter the `rGLS` function. Note that the results are exactly the same as in the previous section.

```
> #The same results can be computed using the preFitModel as follows
> fixed=y ~ age + sex
> Mod1 <- preFitModel(fixed, random=~1|id,
      genabel.data = gen.data,
      phenotype.data = Phen.Data,
      corStruc=list( id=list("GRM","Ind") ))
[1] "GRM ready"
> GWAS1b <- rGLS(fixed, genabel.data = gen.data,
    phenotype.data = Phen.Data, V = Mod1$V)
[1] "Rotation matrix ready"
[1] "Rotate LMM started"
[1] "Rotate LMM ready"
> summary(GWAS1b)
Summary for top 10 results, sorted by P1df
          Chromosome Position Strand A1 A2       effB
rs120315           1  3725352      +  T  C  2.1696316
rs9670687          1  4779800      -  C  A -2.0412127
rs891586           2  8024318      +  A  G -1.6669267
rs1352451          2  8022651      -  A  C  1.6169429
rs1252282          2  8167984      -  A  T -0.9982089
rs7633966          1  3721952      +  C  T  1.8586037
rs9922492          1  3729983      +  C  T -1.8586037
rs4378234          1  4800348      -  G  T -1.9012073
rs7499832          3 10242953      -  G  C -1.0897248
rs6561272          1  3715538      -  T  G -1.9073305
            se_effB chi2.1df         P1df Pc1df
rs120315  0.4258979       NA 3.501204e-07    NA
rs9670687 0.5351582       NA 1.366120e-04    NA
rs891586  0.4478499       NA 1.975997e-04    NA
rs1352451 0.4564078       NA 3.959643e-04    NA
rs1252282 0.2825403       NA 4.109058e-04    NA
rs7633966 0.5325691       NA 4.832322e-04    NA
rs9922492 0.5325691       NA 4.832322e-04    NA
rs4378234 0.5509388       NA 5.588239e-04    NA
rs7499832 0.3182487       NA 6.167705e-04    NA
rs6561272 0.5726170       NA 8.656541e-04    NA
```

The only information transferred from the `preFitModel` function to `rGLS` is the estimated (co)variance matrix `Mod1$V`. The `corStruc` option specifies the correlation structure to be applied on each random effect. In the example we wish to fit polygenic effects and permanent environmental effects. The former requires a correlatioon structure given by the GRM whereas the permanent environmental effects are iid. Consequently, `corStruc=list( id=list("GRM","Ind")`

`))` is specfied.

## A model having several different random effects

In this (fake) example there are 60 observations on each nest and there are 6 different nests. We wish to include these as random effect too, and to start with they are modelled as independent random effects.

```
> # In this example there are 6 nests and 60 observations per nest
> Phen.Data$nest <- rep(1:6, each=60)
> # A model including polygenic effects,
> # permanent environmental effects, and nest effect as random
> Mod2 <- preFitModel(fixed, random=~1|id + 1|nest,
    genabel.data = gen.data, phenotype.data = Phen.Data,
    corStruc=list( id=list("GRM","Ind") , nest=list("Ind")) )
> GWAS2 <- rGLS(fixed, genabel.data = gen.data,
    phenotype.data = Phen.Data, V = Mod2$V)
```

## Spatial modelling

This example shows how random effects having a spatial correlation structure can be included to account for populaton structure. The spatial model used is a Conditional AutoRegressive (CAR) model and the input spatial information is given by a neighborhood matrix. (A neighborhood matrix has a non-zero value for an element (i,j) where the subject (nest in our example) i and j come from neighboring locations. The diagonal elements are zero.) Here the neighborhood matrix (`D`) is a 6 × 6 matrix

```
> D= matrix(0,6,6)
> D[1,2] = D[2,1] = 1
> D[5,6] = D[6,5] = 1
> D[2,4] = D[4,2] = 1
> D[3,5] = D[5,3] = 1
> D[1,6] = D[6,1] = 1
> D[3,4] = D[4,3] = 1
```

where for instance nests 1 and 2 are defined as neighbors. The model including polygenic effects, permanent environmental effects and a spatial correlation between nests is then fitted as

```
> Mod3 <- preFitModel(y ~ age + sex, random=~1|id + 1|nest,
    genabel.data = gen.data, phenotype.data = Phen.Data,
    corStruc=list( id=list("GRM","Ind") ,
    nest=list("CAR")), Neighbor.Matrix=D )
> GWAS2b <- rGLS(fixed, genabel.data = gen.data,
    phenotype.data = Phen.Data, V = Mod3$V)
```

## Using the `simulate_PhenData` function

The third function included in the RepeatABEL package is a simulation function where one can simulate phenotypic data having repeated observations. The input genotype information is given by a GenABEL object.

Suppose for instance we want to simulate 4 observations from each individual and the three variance components (polygenic, permanent env., residual) are 1.

```
> VC.poly <- VC.perm <- VC.res <- 1
> n.obs <- rep(4, nids(gen.data))
```

In this example, the GenABEL object `gen.data` is used as input and an additive genetic effect of 2.0 is simulated at the location of SNP number 1000. The phenotype data is then simulated as

```
> Phen.Sim <- simulate_PhenData(y ~ 1,
     genabel.data = gen.data,
     n.obs = n.obs, SNP.eff = 2, SNP.nr = 1000,
     VC = c(VC.poly, VC.perm,VC.res))
```

The simulated data can then be fitted as

```
> GWAS.sim1 <- rGLS(y ~ 1, genabel.data = gen.data,
    phenotype.data = Phen.Sim)
```

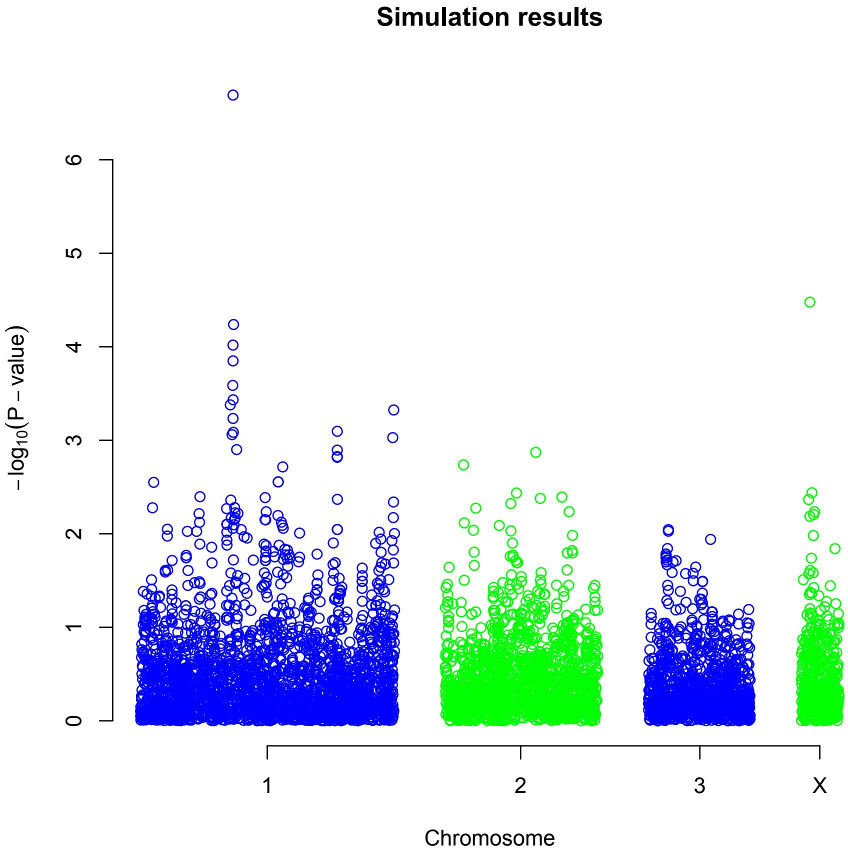
The data set `gen.data` includes information on sex, so we can also add a fixed sex effect to our simulations. Here a sex effect of 1.0 is simulated

```
> Phen.Sim <- simulate_PhenData(y ~ sex,
    genabel.data = gen.data, n.obs = n.obs, SNP.eff = 2,
    SNP.nr = 1000, VC = c(VC.poly, VC.perm, VC.res), beta = c(0,1))
```

where `beta` is a vector specifying the simulated values of the fixed effects and since we fit an intercept and a sex effect the length of `beta` is two. The data can then be fitted as

```
> GWAS.sim1 <- rGLS(y ~ sex, genabel.data = gen.data,
    phenotype.data = Phen.Sim)
> plot(GWAS.sim1, main="Simulation results")
```

The produced Manhattan plot is given below


Simulation results
−log₁₀(P − value)
0
1
2
3
4
5
6
1
2
3
X
Chromosome

## Simulating year effects using the simulate_PhenData function

Year effects can be added to the simulated response after running the simulate_PhenData function.

```
> Phen.Sim <- simulate_PhenData(y ~ sex,
    genabel.data = gen.data, n.obs = n.obs, SNP.eff = 2,
    SNP.nr = 1000, VC = c(VC.poly, VC.perm, VC.res), beta = c(0,1))
> ####### PRODUCE YEAR EFFECTS FOR EACH INDIVIDUAL
> sd.year = 1 #Standard Deviation of Year Effects
> beta <- rnorm(max(n.obs), 0, sd.year) #Simulated Year Effects
> year.effects <- years <- NULL
> for (i in 1:length(n.obs)) {
      yr.i <- sort(sample(1:max(n.obs), n.obs[i]))
      years <- c(years, yr.i)
            year.effects <- c(year.effects, beta[yr.i])
    }
> ########################
> #### A FUNCTION TO ADD A VARIABLE TO A LIST
> add.var <- function(x, add.new, new.name) {
      x[[length(x) + 1]] <- add.new
            names(x)[length(x)] <- new.name
            return(x)
    }
> #######################
> #ADDS A NEW PHENOTYPE WITH YEAR EFFECTS ADDED
> Phen.Sim <- add.var(Phen.Sim, Phen.Sim$y + year.effects, "y.yrs")
> #ADD YEAR AS FACTOR
> Phen.Sim <- add.var(Phen.Sim, as.factor(years), "Years")
> #######################
> #RUN THE ANALYSIS
> GWAS.sim1 <- rGLS(y.yrs ~ sex + Years, genabel.data = gen.data,
    phenotype.data = Phen.Sim)
```

## Simulating binary data using the simulate_PhenData function

Binary data can be simulated using a threshold model where a Gaussian response is first simulated and all values greater than a specified threshold $\tau$ are 1's on the observed scale. The binary data can be simulated and analyzed using the following code.

```
> Phen.Sim <- simulate_PhenData(y ~ sex,
    genabel.data = gen.data, n.obs = n.obs, SNP.eff = 2,
    SNP.nr = 1000, VC = c(VC.poly, VC.perm, VC.res), beta = c(0,1))
> tau = 0
> Phen.Sim$y_observed <- as.numeric(Phen.Sim$y > tau)
> GWAS.sim1 <- rGLS(y_observed ~ sex, genabel.data = gen.data,
    phenotype.data = Phen.Sim)
```