Genetic analysis using the sommer package

Giovanny Covarrubias-Pazaran

2017-04-14

The sommer package has been developed to provide R users with a powerful multivariate mixed model solver for different genetic and non-genetic analysis in diploid and polyploid organisms. This package allows the user to estimate variance components for a mixed model with the advantage of specifying the variance-covariance structure of the random effects and obtain other parameters such as BLUPs, BLUEs, residuals, fitted values, variances for fixed and random effects, etc.

The package is focused on genomic prediction (or genomic selection) and GWAS analysis, although general mixed models can be fitted as well. The package provides kernels to estimate additive (`A.mat`), dominance (`D.mat`), and epistatic (`E.mat`) relationship matrices that have been shown to increase prediction accuracy under certain scenarios. The package provides flexibility to fit other genetic models such as full and half diallel models as well.

Vignettes aim to provide several examples in how to use the sommer package under different scenarios in breeding and genetics. We will spend the rest of the space providing examples for:

1) Heritability (h^2) calculation
2) Half and full diallel designs
3) Genome wide association analysis (GWAS) in diploids and tetraploids
4) Genomic selection
5) Single cross prediction
6) Multivariate genetic models and genetic correlations
7) Multivariate GWAS
8) Specifying heterogeneous variances in univariate mixed models

Background

The core of the package are the `mmer` (matrix-based) and `mmer2` (formula-based) functions which solve the mixed model equations. The functions are an interface to call one of the 4 ML/REML methods supported in the package; `EMMA` efficient mixed model association (Kang et al. 2008), `AI` average information (Gilmour et al. 1995; Lee et al. 2015), `EM` expectation maximization (Searle 1993; Bernardo 2010), and the default `NR` Newton-Raphson (Tunnicliffe 1989). All methods can handle multiple random effects and covariance structures.

The mixed model solved by the algorithms has the form:

$y = X\beta + Zu + \epsilon$

or

$y = X\beta + Zu_1 + ... + Zu_i + \epsilon$

where:

X is an incidence matrix for fixed effects

Z is an incidence matrix for random effects

β is the vector for BLUEs of fixed effects

u is the vector for BLUPs of random effects

ϵ are the residuals

The variance of the response is known to be the random part of the model:

Var(y) = Var($Zu + \epsilon$) = ZGZ + R = V

and with

u ~ MVN(u, G)

ϵ ~ MVN(u, R)

When multiple random effects are present the Z matrix becomes the column binding of each of the Z_i matrices for the i random effects. And the G matrix becomes the diagonal binding of each of the variance covariance structures (K matrices) for the random effects:

$$\mathbf{Z} = \begin{bmatrix} Z_1 & ... & Z_i \end{bmatrix}$$

$$\mathbf{G} = \begin{bmatrix} K_1\sigma_u^2 1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & K_i\sigma_u^2 i \end{bmatrix}$$

The program takes the Zs and Ks for each random effect and construct the neccesary structure inside and estimates the variance components by ML/REML using any of the 4 methods available in sommer; Direct-Inversion Average Information, Expectation-Maximization, Direct-Inversion Newton-Raphson, and Efficient Mixed Model Association. Please refer to the canonical papers listed in the Literature section to check how the methods work. We have tested widely the methods to make sure they provide the same solution when the likelihood behaves well but for complex problems they might lead to slightly different answers. If you have any concer please contact me at cova_ruber@live.com.mx or covarrubiasp@wisc.edu.

1) Marker and non-marker based heritability calculation

The heritability is one of the most popular parameters in the breeding and genetics community. The heritability is usually estimated as narrow sense (h^2; only additive variance in the numerator σ_A^2), and broad sense (H^2; all genetic variance in the numerator σ_G^2).

In a classical experiment with no molecular markers, special designs are performed to estimate and disect the additive (σ_A^2) and dominance (σ_D^2) variance along with environmental variability. Designs such as generation analysis, North Carolina designs are used to disect σ_A^2 and σ_D^2 to estimate the narrow sense heritability (h^2). When no special design is available we can still disect the genetic variance (σ_G^2) and estimate the broad sense heritability. In this example we will show the broad sense estimation which doesn't use covariance structures for random effects. For big models with no covariance structures, sommer's direct inversion is a bad idea to use but we will show anyways how to do it, for very sparse models we recommend using the lmer function from the lme4 package from Douglas Bates.

The dataset has 41 potato lines evaluated in 5 locations across 3 years in an RCBD design. We show how to fit the model and extract the variance components to calculate the h^2.

```
library(sommer)
data(h2)
head(h2)
```

```
##                Name     Env Loc Year     Block y
## 1           W8822-3 FL.2012  FL 2012 FL.2012.1 2
## 2           W8867-7 FL.2012  FL 2012 FL.2012.2 2
## 3         MSL007-B MO.2011  MO 2011 MO.2011.1 3
```

```
## 4         CO00270-7W FL.2012  FL 2012 FL.2012.2 3
## 5 Manistee(MSL292-A) FL.2013  FL 2013 FL.2013.2 3
## 6           MSM246-B FL.2012  FL 2012 FL.2012.2 3
```

```
ans1 <- mmer2(y~1, random=~Name + Env + Name:Env + Block,data=h2, silent = TRUE)
vc <- ans1$var.comp
V_E <- vc[2,1];V_GE <- vc[3,1];V_G <- vc[1,1];Ve <- vc[5,1]

n.env <- length(levels(h2$Env))
h2 <- V_G/(V_G + V_GE/n.env + Ve/(2*n.env)) #the 2 is a reference for block
h2
```

```
## [1] 0.8594805
```

Recently with markers becoming cheaper, thousand of markers can be run in the breeding materials. When markers are available, an special design is not neccesary to disect the additive genetic variance. The availability of the additive, dominance and epistatic relationship matrices allow us to estimate σ^2_A, σ^2_D and σ^2_I.

Assume you have a population, and a similar model like the one displayed previously has been fitted. Now we have BLUPs for the genotypes but in addition we have genetic markers.

```
data(CPdata)
CPpheno <- CPdata$pheno; CPpheno$idd <-CPpheno$id; CPpheno$ide <-CPpheno$id
CPgeno <- CPdata$geno
### look at the data
head(CPpheno)
```

```
##        id Row Col Year      color  Yield FruitAver Firmness  idd  ide
## P003 P003   3   1 2014 0.10075269 154.67     41.93  588.917 P003 P003
## P004 P004   4   1 2014 0.13891940 186.77     58.79  640.031 P004 P004
## P005 P005   5   1 2014 0.08681502  80.21     48.16  671.523 P005 P005
## P006 P006   6   1 2014 0.13408561 202.96     48.24  687.172 P006 P006
## P007 P007   7   1 2014 0.13519278 174.74     45.83  601.322 P007 P007
## P008 P008   8   1 2014 0.17406685 194.16     44.63  656.379 P008 P008
```

```
CPgeno[1:5,1:4]
```

```
##      scaffold_50439_2381 scaffold_39344_153 uneak_3436043 uneak_2632033
## P003                   0                  0            0            1
## P004                   0                  0            0            1
## P005                   0                 -1            0            1
## P006                  -1                 -1           -1            0
## P007                   0                  0            0            1
```

```
## fit a model including additive and dominance effects
A <- A.mat(CPgeno) # additive relationship matrix
D <- D.mat(CPgeno) # dominance relationship matrix
E <- E.mat(CPgeno) # epistatic relationship matrix

ans.ADE <- mmer2(color~1, random=~g(id) + g(idd) + g(ide),
                 G=list(id=A,idd=D,ide=E), silent = TRUE, data=CPpheno)
(H2 <- sum(ans.ADE$var.comp[1:3,1])/sum(ans.ADE$var.comp[,1]))
```

```
## [1] 0.7514288
```

```
(h2 <- sum(ans.ADE$var.comp[1,1])/sum(ans.ADE$var.comp[,1]))
```

```
## [1] 0.6214712
```

In the previous example we showed how to estimate the additive (σ^2_A), dominance (σ^2_D), and epistatic (σ^2_I) variance components based on markers and estimate broad (H^2) and narrow sense heritability (h^2).

2) Half and full diallel designs

When breeders are looking for the best single cross combinations, diallel designs have been by far the most used design in crops like maize. There are 4 types of diallel designs depending if reciprocate and self cross (omission of parents) are performed (full diallel with parents n^2; full diallel without parents n(n-1); half diallel with parents 1/2 * n(n+1); half diallel without parents 1/2 * n(n-1)). In this example we will show a full dialle design (reciprocate crosses are performed) and half diallel designs (only one of the directions is performed).

In the first data set we show a full diallel among 40 lines from 2 heterotic groups, 20 in each. Therefore 400 possible hybrids are possible. We have pehnotypic data for 100 of them across 4 locations. We use the data available to fit a model of the form:

$y = X\beta + Zu_1 + Zu_2 + Zu_S + \epsilon$

We estimate variance components for GCA_1, GCA_2 and SCA and use them to estimate heritability. Additionally BLUPs for GCA and SCA effects can be used to predict crosses.

```
data(cornHybrid)
hybrid2 <- cornHybrid$hybrid # extract cross data
head(hybrid2)
```

```
##   Location GCA1   GCA2         SCA Yield PlantHeight
## 1        1 A258 AS5707 A258:AS5707    NA          NA
## 2        1 A258     B2     A258:B2    NA          NA
## 3        1 A258    B99    A258:B99    NA          NA
## 4        1 A258   LH51   A258:LH51    NA          NA
## 5        1 A258   Mo44   A258:Mo44    NA          NA
## 6        1 A258  NC320  A258:NC320    NA          NA
```

```
modFD <- mmer2(Yield~Location, random=~GCA1+GCA2+SCA, data=hybrid2,silent = TRUE, draw=FALSE)
summary(modFD)
```

```
##
## Information contained in this fitted model:
## * Variance components, Residuals, Fitted values
## * BLUEs and BLUPs, Inverse phenotypic variance(V)
## * Variance-covariance matrix for fixed & random effects
## * Predicted error variance (PEV), LogLikelihood
## Use the '$' symbol to access such information
## ========================================================
## Linear mixed model fit by restricted maximum likelihood
## ******************  sommer 2.7  *********************
```

```
## ============================================================
## Method:[1] "NR"
##
## logLik    AIC    BIC
##  -1342   2691   2707
## ============================================================
## Random effects:
##                VarComp VarCompSE  Zratio
## Var(GCA1)        0.000     17.65  0.0000
## Var(GCA2)        7.205     19.60  0.3677
## Var(SCA)       187.736     44.59  4.2107
## Var(Residual) 221.142     17.78 12.4406
## Number of obs: 400  Groups: 20 20 400
## ============================================================
## Fixed effects:
##                    Value   Std.Error t.value
## (Intercept)  1.3793e+02  2.1193e+00 65.0845
## Location2   -1.5632e-13  2.1030e+00  0.0000
## Location3    7.8353e+00  2.1030e+00  3.7257
## Location4   -9.0975e+00  2.1030e+00 -4.3259
## ============================================================
## Use the '$' symbol to access all information
```

```
Vgca <- sum(modFD$var.comp[1:2,1])
Vsca <- modFD$var.comp[3,1]
Ve <- modFD$var.comp[4,1]
Va = 4*Vgca
Vd = 4*Vsca
Vg <- Va + Vd
(H2 <- Vg / (Vg + (Ve)) )
```

```
## [1] 0.7790583
```

```
(h2 <- Va / (Vg + (Ve)) )
```

```
## [1] 0.02879398
```

Don't worry too much about the small h2 value, the data was simulated to be mainly dominance variance, therefore the Va was simulated extremely small leading to such value of narrow sense h2.

In this second data set we show a small half diallel with 7 parents crossed in one direction. n(n-1)/2 crosses are possible $7(6)/2 = 21$ unique crosses. Parents appear as males or females indistictly. Each with two replications in a CRD. For a half diallel design a single GCA variance component can be estimated and an SCA as well ($\sigma^2_G CA$ and $\sigma^2_S CA$ respectively). And BLUPs for GCA and SCA of the parents can be extracted. We would create the design matrices in sommer using the `overlay` and `model.matrix` functions for the GCA and SCA matrices respectively.

$y = X\beta + Zu_g + Zu_s + \epsilon$

```
data(HDdata)
head(HDdata)
```

```
##   rep geno male female     sugar
```

```
## 1   1   12    1      2 13.950509
## 2   2   12    1      2  9.756918
## 3   1   13    1      3 13.906355
## 4   2   13    1      3  9.119455
## 5   1   14    1      4  5.174483
## 6   2   14    1      4  8.452221
```

```
HDdata$geno <- as.factor(HDdata$geno)
HDdata$male <- as.factor(HDdata$male)
HDdata$female <- as.factor(HDdata$female)
# Fit the model
modHD <- mmer2(sugar~1, random=~male + and(female) + geno,
               data=HDdata, silent = TRUE)
summary(modHD)
```

```
##
## Information contained in this fitted model:
## * Variance components, Residuals, Fitted values
## * BLUEs and BLUPs, Inverse phenotypic variance(V)
## * Variance-covariance matrix for fixed & random effects
## * Predicted error variance (PEV), LogLikelihood
## Use the '$' symbol to access such information
## ============================================================
## Linear mixed model fit by restricted maximum likelihood
## ********************  sommer 2.7  **********************
## ============================================================
## Method:[1] "NR"
##
## logLik    AIC    BIC
## -58.18 118.36 120.09
## ============================================================
## Random effects:
##                   VarComp VarCompSE Zratio
## Var(and(female))   5.508    3.5771  1.540
## Var(geno)          1.816    1.3637  1.332
## Var(Residual)      3.117    0.9619  3.241
## Number of obs: 42  Groups: 7 21
## ============================================================
## Fixed effects:
##             Value Std.Error t.value
## Intercept 10.3332    1.8183  5.6828
## ============================================================
## Use the '$' symbol to access all information
```

```
Vgca <- modHD$var.comp[1,1]
Vsca <- modHD$var.comp[2,1]
Ve <- modHD$var.comp[3,1]
Va = 4*Vgca
Vd = 4*Vsca
Vg <- Va + Vd
(H2 <- Vg / (Vg + (Ve/2)) ) # 2 technical reps
```

```
## [1] 0.9494872
```

```
(h2 <- Va / (Vg + (Ve/2)) )
```

```
## [1] 0.7140583
```

3) Genome wide association analysis (GWAS) in diploids and tetraploids

With the development of modern statistical machinery the detection of markers associated to phenotypic traits have become quite straight forward. The days of QTL mapping using biparental populations exclusively are in the past. In this section we will show how to perform QTL mapping for diploid and polyploid organisms with complex genetic relationships. In addition we will show QTL mapping in biparental populations to clarify that the fact that is not required anymore doesn't limit the capabilities of modern mixed model machinery.

First we will start doing the GWAS in a biparental population with 363 individuals genotyped with 2889 SNP markers. This is easily done by creating the variance covariance among individuals and using it in the random effect for genotypes. The markers are added in the `W` argument to fit the model of the form:

$y = X\beta + Zu + Wg + \epsilon$

In this case $X\beta$ is the fixed part only for the intercept, Zu is the random effect for genotypes with the additive relationship matrix (A) as the variance-covariance of the random effect, Wg is the marker matrix and the effects of each marker. This is done in this way:

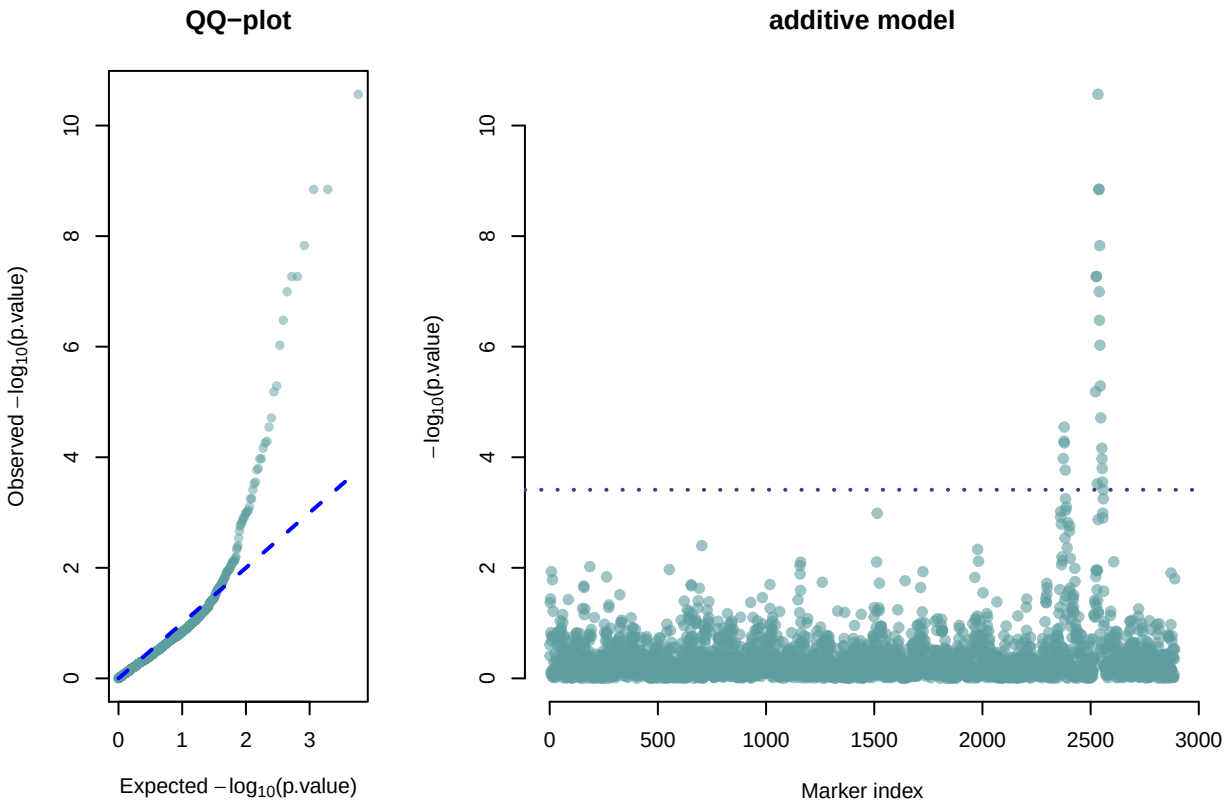
```
data(CPdata)
CPpheno <- CPdata$pheno
CPgeno <- CPdata$geno
### look at the data
head(CPpheno); CPgeno[1:5,1:4]
```

```
##        id Row Col Year      color  Yield FruitAver Firmness
## P003 P003   3   1 2014 0.10075269 154.67     41.93  588.917
## P004 P004   4   1 2014 0.13891940 186.77     58.79  640.031
## P005 P005   5   1 2014 0.08681502  80.21     48.16  671.523
## P006 P006   6   1 2014 0.13408561 202.96     48.24  687.172
## P007 P007   7   1 2014 0.13519278 174.74     45.83  601.322
## P008 P008   8   1 2014 0.17406685 194.16     44.63  656.379
```

```
##      scaffold_50439_2381 scaffold_39344_153 uneak_3436043 uneak_2632033
## P003                   0                  0             0             1
## P004                   0                  0             0             1
## P005                   0                 -1             0             1
## P006                  -1                 -1            -1             0
## P007                   0                  0             0             1
```

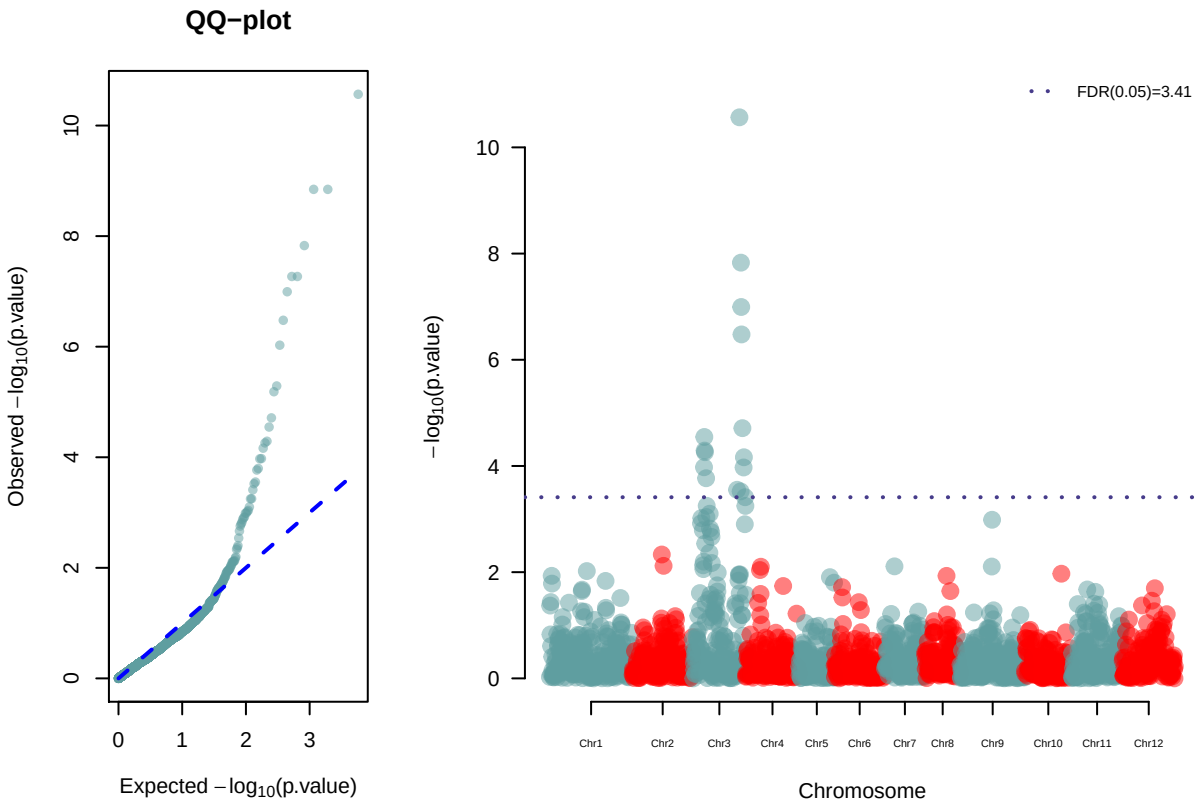
```
A <- A.mat(CPgeno) # additive relationship matrix
### fit the model
ans.A <- mmer2(color~1,random=~g(id), G=list(id=A),
               W=CPgeno, data=CPpheno, silent=TRUE) # fit the model
```

```
## Response is imputed for estimation of variance components in GWAS models.
```




```
### if you have a genetic map you can use it
my.map <- CPdata$map
ans.A <- mmer2(color~1,random=~g(id), G=list(id=A),
               W=CPgeno, data=CPpheno, silent=TRUE, map=my.map) # fit the model
```

```
## Response is imputed for estimation of variance components in GWAS models.
```



Now we will show how to do GWAS in a tetraploid using potato data. Is not very different from diploids. We only need to pay attention to the `ploidy` argument in the `atcg1234` and `A.mat` functions. In addition, when running the `mmer` model there is more models that can be implemented according to Rosyara et al. (2016).

```
data(PolyData)
genotypes <- PolyData$PGeno
phenotypes <- PolyData$PPheno
## convert markers to numeric format
numo <- atcg1234(data=genotypes, ploidy=4, silent = TRUE); numo[1:5,1:4]; dim(numo)
```

```
## Obtaining reference alleles
## Checking for markers with more than 2 alleles. If found will be removed.
## Converting to numeric format
## Calculating minor allele frequency (MAF)
## Imputing missing data with mode

##             c2_41437 c2_24258 c2_21332 c2_21320
## A96104-2           1        2        2        4
## A97066-42          2        3        2        4
## ACBrador           2        4        2        4
## ACLPI175395        0        4        0        4
## ADGPI195204        0        4        0        4

## [1]  221 3521
```

```r
# get only plants with both genotypes and phenotypes
common <- intersect(phenotypes$Name,rownames(numo))
marks <- numo[common,]; marks[1:5,1:4]
```

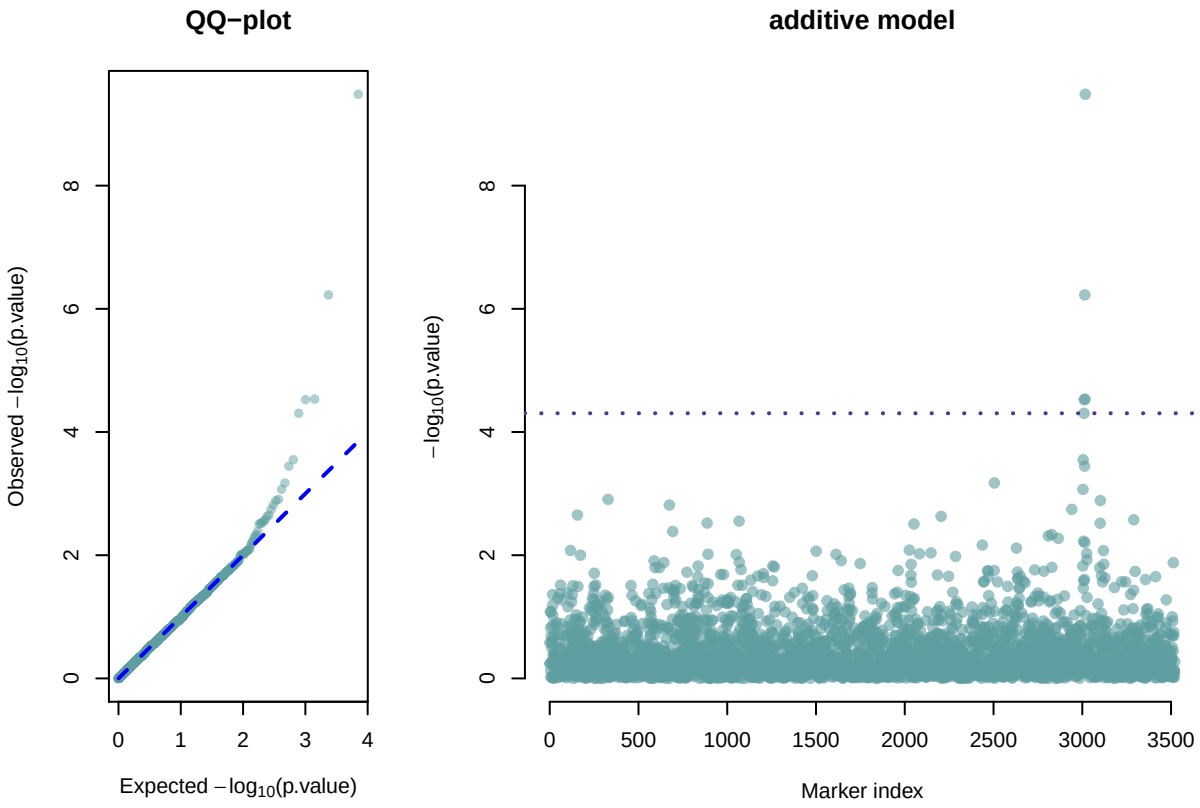
```
##                c2_41437 c2_24258 c2_21332 c2_21320
## A97066-42             2        3        2        4
## ACBrador              2        4        2        4
## AdirondackBlue        2        2        2        4
## AF2291-10             0        4        2        4
## AF2376-5              1        3        2        4
```

```r
phenotypes2 <- phenotypes[match(common,phenotypes$Name),];
phenotypes2[1:5,1:4]
```

```
##             Name total_yield chip_color tuber_eye_depth
## 1      A97066-42       13.10       2.35            3.03
## 2       ACBrador       15.56       2.63            4.37
## 3 AdirondackBlue       11.77       2.82            3.76
## 4      AF2291-10       13.43       1.50            4.50
## 5       AF2376-5       12.58       1.83            4.50
```

```r
# Additive relationship matrix, specify ploidy
K1 <- A.mat(marks, ploidy=4)
# run the model you want
models <- c("additive","1-dom-alt","1-dom-ref","2-dom-alt","2-dom-ref")
ans2 <- mmer2(tuber_shape~1, random=~g(Name), G=list(Name=K1), W=marks,
              method="EMMA", data=phenotypes2, silent = TRUE)
```

```
## Response is imputed for estimation of variance components in GWAS models.
```




```
summary(ans2)
```

```
##
## Information contained in this fitted model:
## * Variance components, Residuals, Fitted values
## * BLUEs and BLUPs, Inverse phenotypic variance(V)
## * Variance-covariance matrix for fixed & random effects
## * Predicted error variance (PEV), LogLikelihood
## Use the '$' symbol to access such information
## ============================================================
## Linear mixed model fit by restricted maximum likelihood
## ********************  sommer 2.7  **********************
## ============================================================
## Method:[1] "EMMA"
##
## logLik    AIC    BIC
## -192.5  387.0  390.3
## ============================================================
## Random effects:
##               VarComp
## Var(g(Name)) 0.60778
## Var(Error)   0.03807
## Number of obs: 187  Groups: 187
## ============================================================
## Fixed effects:
```

```
##              Value Std.Error t.value
## Intercept 3.307861  0.018302  180.74
## ========================================================
## Use the '$' symbol to access all information
```

4) Genomic selection

In this section we will use wheat data from CIMMYT to show how is genomic selection performed. This is the case of prediction of specific individuals within a population. It basically uses a similar model of the form:

$y = X\beta + Zu + \epsilon$

and takes advantage of the variance covariance matrix for the genotype effect known as the additive relationship matrix (A) and calculated using the `A.mat` function to establish connections among all individuals and predict the BLUPs for individuals that were not measured. The prediction accuracy depends on several factors such as the heritability (h^2), training population used (TP), size of TP, etc.

```
data(wheatLines);
X <- wheatLines$wheatGeno; X[1:5,1:4]; dim(X)
```

```
##      wPt.0538 wPt.8463 wPt.6348 wPt.9992
## [1,]       -1        1        1        1
## [2,]        1        1        1        1
## [3,]        1        1        1        1
## [4,]       -1        1        1        1
## [5,]       -1        1        1        1

## [1]  599 1279
```

```
Y <- data.frame(wheatLines$wheatPheno); Y$id <- rownames(Y); head(Y);
```

```
##              X1          X2          X4         X5   id
## 775   1.6716295 -1.72746986 -1.89028479  0.0509159  775
## 2166 -0.2527028  0.40952243  0.30938553 -1.7387588 2166
## 2167  0.3418151 -0.64862633 -0.79955921 -1.0535691 2167
## 2465  0.7854395  0.09394919  0.57046773  0.5517574 2465
## 3881  0.9983176 -0.28248062  1.61868192 -0.1142848 3881
## 3889  2.3360969  0.62647587  0.07353311  0.7195856 3889
```

```
rownames(X) <- rownames(Y)
# select environment 1
K <- A.mat(X) # additive relationship matrix
# GBLUP pedigree-based approach
set.seed(12345)
y.trn <- Y
vv <- sample(rownames(Y),round(dim(Y)[1]/5))
y.trn[vv,"X1"] <- NA
ans <- mmer2(X1~1,random=~g(id), G=list(id=K), method="EMMA",
             data=y.trn, silent = TRUE) # kinship based
cor(ans$u.hat$`g(id)`[vv,],Y[vv,"X1"])
```

```
## [1] 0.4885687
```

```r
## maximum prediction value that can be achieved
sqrt(ans$var.comp[1,1]/sum(ans$var.comp[,1]))
```

```
## Var(g(id))
##  0.5771923
```

5) Single cross prediction

When doing prediction of single cross performance the phenotype can be dissected in three main components, the general combining abilities (GCA) and specific combining abilities (SCA). This can be expressed with the same model analyzed in the diallel experiment mentioned before:

$y = X\beta + Zu_1 + Zu_2 + Zu_S + \epsilon$

with:

$u_1 \sim \mathrm{N}(0,\ K_1\sigma_u^2 1)$

$u_2 \sim \mathrm{N}(0,\ K_2\sigma_u^2 2)$

$u_s \sim \mathrm{N}(0,\ K_3\sigma_u^2 s)$

And we can specify the K matrices. The main difference between this model and the full and half diallel designs is the fact that this model will include variance covariance structures in each of the three random effects (GCA1, GCA2 and SCA) to be able to predict the crosses that have not ocurred yet. We will use the data published by Technow et al. (2015) to show how to do prediction of single crosses.

```r
data(Technow_data)

A.flint <- Technow_data$AF # Additive relationship matrix Flint
A.dent <- Technow_data$AD # Additive relationship matrix Dent

pheno <- Technow_data$pheno # phenotypes for 1254 single cross hybrids
head(pheno);dim(pheno)
```

```
##    hybrid dent flint     GY    GM      hy
## 1 518.298  518   298  -8.04 -0.85 518:298
## 2 518.305  518   305 -11.10  1.70 518:305
## 3 518.306  518   306 -16.85  2.24 518:306
## 4 518.316  518   316   2.08 -1.33 518:316
## 5 518.323  518   323   5.65 -2.71 518:323
## 6 518.327  518   327 -16.95 -0.52 518:327
```

```
## [1] 1254    6
```

```r
# CREATE A DATA FRAME WITH ALL POSSIBLE HYBRIDS
DD <- kronecker(A.dent,A.flint,make.dimnames=TRUE)
hybs <- data.frame(sca=rownames(DD),yield=NA,matter=NA,gcad=NA, gcaf=NA)
hybs$yield[match(pheno$hy, hybs$sca)] <- pheno$GY
hybs$matter[match(pheno$hy, hybs$sca)] <- pheno$GM
hybs$gcad <- as.factor(gsub(":.*","",hybs$sca))
hybs$gcaf <- as.factor(gsub(".*:","",hybs$sca))
head(hybs)
```

```
##       sca yield matter gcad gcaf
## 1 513:316 10.02  -2.05  513  316
## 2 513:323  6.97  -3.78  513  323
## 3 513:330    NA     NA  513  330
## 4 513:336    NA     NA  513  336
## 5 513:340    NA     NA  513  340
## 6 513:341    NA     NA  513  341
```

```
# RUN THE PREDICTION MODEL
y.trn <- hybs
vv1 <- which(!is.na(hybs$yield))
vv2 <- sample(vv1, 100)
y.trn[vv2,"yield"] <- NA
anss2 <- mmer2(yield~1, random=~g(gcad) + g(gcaf), G=list(gcad=A.dent, gcaf=A.flint),
               method="EM", silent=TRUE, data=y.trn)
```

```
## With var-cov structures (G) present you may want to try the AI or NR algorithm.
```

```
summary(anss2)
```

```
##
## Information contained in this fitted model:
## * Variance components, Residuals, Fitted values
## * BLUEs and BLUPs, Inverse phenotypic variance(V)
## * Variance-covariance matrix for fixed & random effects
## * Predicted error variance (PEV), LogLikelihood
## Use the '$' symbol to access such information
## ============================================================
## Linear mixed model fit by restricted maximum likelihood
## ********************  sommer 2.7  **********************
## ============================================================
## Method:[1] "EM"
##
## logLik    AIC    BIC
##  -5036  10073  10078
## ============================================================
## Random effects:
##               VarComp
## Var(g(gcad))    16.17
## Var(g(gcaf))    11.26
## Var(Residual)   17.66
## Number of obs: 1154  Groups: 123 86
## ============================================================
## Fixed effects:
##             Value Std.Error t.value
## Intercept 0.24555   0.20354  1.2064
## ============================================================
## Use the '$' symbol to access all information
```

```
cor(anss2$fitted.y[vv2], hybs$yield[vv2])
```

```
## [1] 0.8779213
```

In the previous model we only used the GCA effects (GCA1 and GCA2) for practicity, altough it's been shown that the SCA effect doesn't actually help that much in increasing prediction accuracy and increase a lot the computation intensity required since the variance covariance matrix for SCA is the kronecker product of the variance covariance matrices for the GCA effects, resulting in a 10578x10578 matrix that increases in a very intensive manner the computation required.

A model without covariance structures would show that the SCA variance component is insignificant compared to the GCA effects. This is why including the third random effect doesn't increase the prediction accuracy.

6) Multivariate genetic models and genetic correlations

Sometimes is important to estimate genetic variance-covariance among traits, multi-reponse models are very useful for such task. Currently sommer can deal with multivariate models for multiple random effects of the form:

$Y = X\beta + Zu + \epsilon$

with:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_t \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} X & ... & ... \\ ... & ... & ... \\ ... & ... & X \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} Z_1 G_1 Z_1' + ... + Z_1 R_1 Z_1' & ... & Z_1 H_1 Z_t' + ... + Z_1 S_1 Z_t' \\ ... & ... & ... \\ Z_t H_1 Z_1' + ... + Z_t S_1 Z_1' & ... & Z_t G_1 Z_t' + ... + Z_t R_1 Z_t' \end{bmatrix}$$

for 't' traits, where G are H are variance and covariance matrices among random effects for the "t" trait, and R and S are variance and covariance matrices among residuals. Here R=S=Iσ_ϵ , where I is an identity matrix. We can specify the covariance matrices. BLUPs will also be corrected for such covariances usually leading to more accurate predictions.

```
data(CPdata)
CPpheno <- CPdata$pheno
CPgeno <- CPdata$geno
### look at the data
head(CPpheno);CPgeno[1:5,1:4]
```

```
##        id Row Col Year      color  Yield FruitAver Firmness
## P003 P003   3   1 2014 0.10075269 154.67     41.93  588.917
## P004 P004   4   1 2014 0.13891940 186.77     58.79  640.031
## P005 P005   5   1 2014 0.08681502  80.21     48.16  671.523
## P006 P006   6   1 2014 0.13408561 202.96     48.24  687.172
## P007 P007   7   1 2014 0.13519278 174.74     45.83  601.322
## P008 P008   8   1 2014 0.17406685 194.16     44.63  656.379
```

```
##      scaffold_50439_2381 scaffold_39344_153 uneak_3436043 uneak_2632033
## P003                   0                  0             0             1
```

```
## P004                   0                  0             0             1
## P005                   0                 -1             0             1
## P006                  -1                 -1            -1             0
## P007                   0                  0             0             1
```

```
## fit a model including additive effects
A <- A.mat(CPgeno) # additive relationship matrix
####================####
#### ADDITIVE MODEL ####
####================####
ans.A <- mmer2(cbind(color,Yield,Firmness)~1, random=~g(id),G=list(id=A),
               MVM=TRUE, data=CPpheno, silent = TRUE)
summary(ans.A)
```

```
## Information contained in this structure:
## * Results for a multi response model
## Displayed:
## * Variance-covariance component summaries
## Use the '$' sign to access parameters
## ==========================================================
##     Multivariate Linear Mixed Model fit by REML
## ********************  sommer 2.7  **********************
## ==========================================================
## Method:[1] "MNR"
##        logLik      AIC      BIC
## MVM -435.9851 877.9703 892.9493
## ==========================================================
## Variance-Covariance components:
##
## Var-Covar(g(id))
##             color     Yield  Firmness
## color    0.005219    0.2984    0.5462
## Yield    0.298362  663.1160 -122.6967
## Firmness 0.546208 -122.6967 1264.8873
##
## Var-Covar(Residual)
##              color     Yield   Firmness
## color     0.002713    0.2224   -0.06052
## Yield     0.222351 4011.1719  189.58719
## Firmness -0.060518  189.5872 1200.92942
## ==========================================================
## Standard errors for variance components:
##                               VarComp VarCompSE  Zratio
## g(id).color-color           5.219e-03 1.047e-03  4.9844
## g(id).color-Yield           2.984e-01 4.334e-01  0.6885
## g(id).color-Firmness        5.462e-01 3.977e-01  1.3734
## g(id).Yield-Yield           6.631e+02 3.260e+02  2.0341
## g(id).Yield-Firmness       -1.227e+02 2.260e+02 -0.5430
## g(id).Firmness-Firmness     1.265e+03 2.969e+02  4.2602
## Residual.color-color        2.713e-03 2.979e-04  9.1068
## Residual.color-Yield        2.224e-01 2.261e-01  0.9833
## Residual.color-Firmness    -6.052e-02 1.320e-01 -0.4586
## Residual.Yield-Yield        4.011e+03 3.424e+02 11.7162
## Residual.Yield-Firmness     1.896e+02 1.433e+02  1.3230
```

```
## Residual.Firmness-Firmness  1.201e+03 1.194e+02 10.0548
## ============================================================
## Fixed effects:
##             color Yield Firmness
## (Intercept) 1.417 144.6      638
## ============================================================
## Groups and observations:
##       Observ Groups
## g(id)    363    363
## ============================================================
## Use the '$' sign to access parameters
```

Now you can extract the BLUPs using the 'randef' function or simple accesing with the '$' sign and pick 'u.hat'. Calculate genetic correlations and heritabilities easily.

```
## genetic variance covariance
gvc <- ans.A$var.comp$`g(id)`
## extract variances (diagonals) and get standard deviations
sd.gvc <- as.matrix(sqrt(diag(gvc)))
## get possible products sd(Vgi) * sd(Vgi')
prod.sd <- sd.gvc %*% t(sd.gvc)
## genetic correlations cov(gi,gi')/[sd(Vgi) * sd(Vgi')]
(gen.cor <- gvc/prod.sd)
```

```
##              color      Yield   Firmness
## color    1.0000000  0.1603880  0.2125959
## Yield    0.1603880  1.0000000 -0.1339714
## Firmness 0.2125959 -0.1339714  1.0000000
```

```
## heritabilities
(h2 <- diag(gvc) / diag(cov(CPpheno[,names(diag(gvc))], use = "complete.obs")))
```

```
##     color     Yield  Firmness
## 0.8389640 0.1457021 0.4936389
```

7) Multivariate GWAS

Following the same theory of multivariate methods, theorically the marker effects can take advantage of the information contained in the correlation among traits besides exploiting the correlations between individuals. We have extended the GWAS framework to multivariate GWAS. Here, we will show a multivariate GWAS in a biparental population using the information of 4 traits.

```
data(CPdata)
CPpheno <- CPdata$pheno
CPgeno <- CPdata$geno
### look at the data
head(CPpheno);CPgeno[1:5,1:4]
```

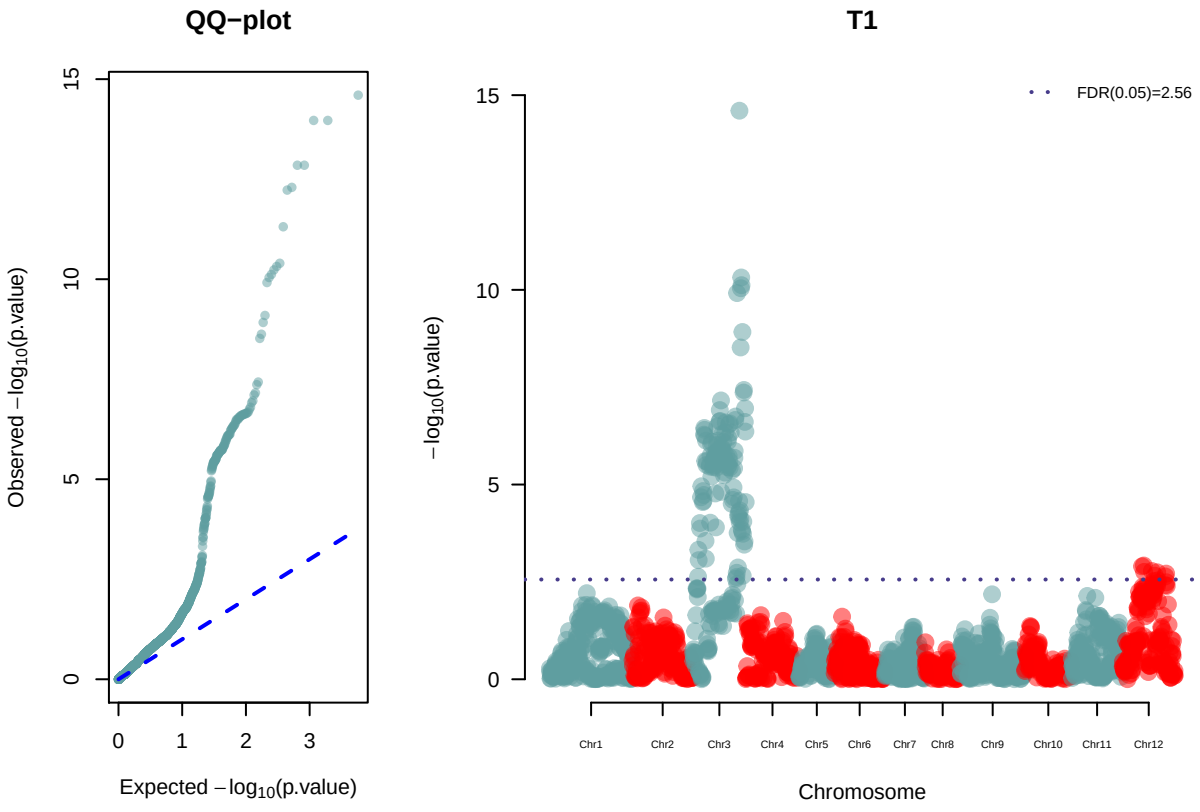
```
##        id Row Col Year      color  Yield FruitAver Firmness
## P003 P003   3   1 2014 0.10075269 154.67     41.93  588.917
## P004 P004   4   1 2014 0.13891940 186.77     58.79  640.031
```

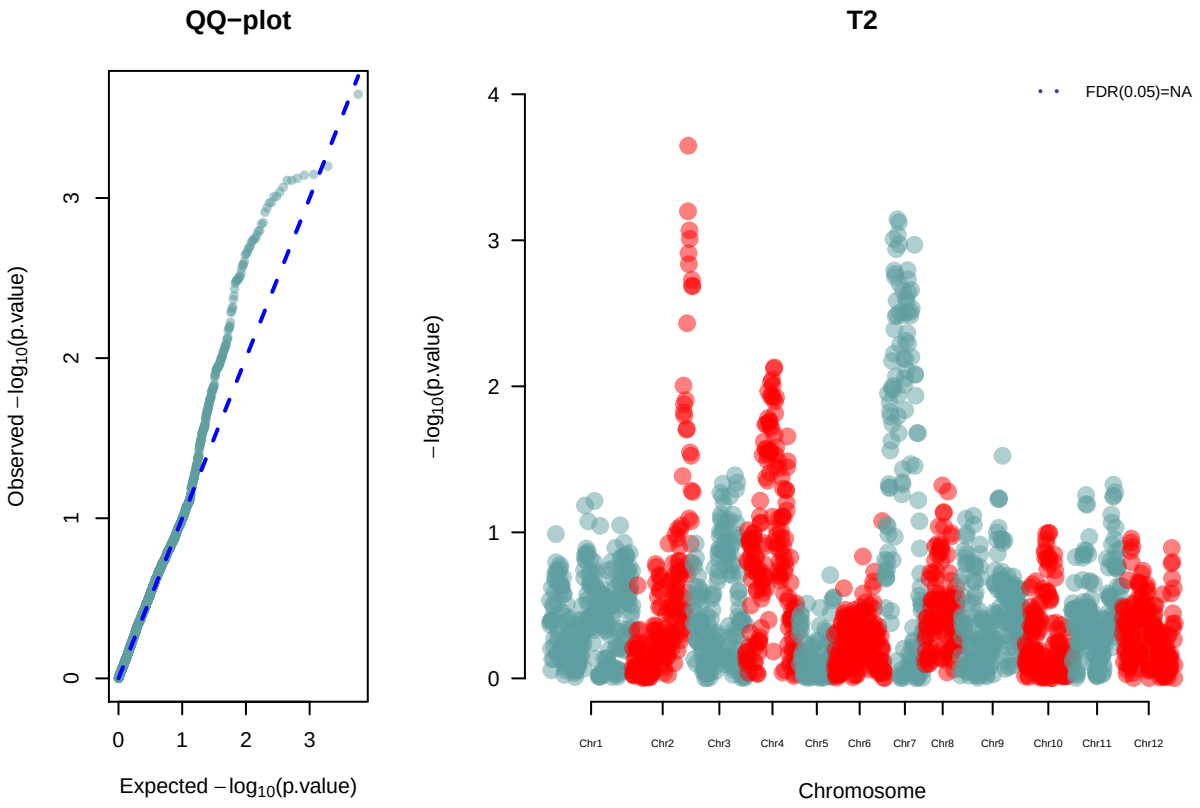
```
## P005 P005   5   1 2014 0.08681502  80.21     48.16  671.523
## P006 P006   6   1 2014 0.13408561 202.96     48.24  687.172
## P007 P007   7   1 2014 0.13519278 174.74     45.83  601.322
## P008 P008   8   1 2014 0.17406685 194.16     44.63  656.379
```

```
##      scaffold_50439_2381 scaffold_39344_153 uneak_3436043 uneak_2632033
## P003                   0                  0            0            1
## P004                   0                  0            0            1
## P005                   0                 -1            0            1
## P006                  -1                 -1           -1            0
## P007                   0                  0            0            1
```

```
## fit a model including additive effects
A <- A.mat(CPgeno) # additive relationship matrix
####================####
#### ADDITIVE MODEL ####
####================####
ans.A <- mmer2(cbind(color,Firmness)~1, random=~g(id),G=list(id=A),
               MVM=TRUE, data=CPpheno, silent = TRUE, W=CPgeno, IMP=TRUE,
               map=CPdata$map)
```

```
## Response is imputed for estimation of variance components in GWAS models.
```






```
summary(ans.A)
```

```
## Information contained in this structure:
## * Results for a multi response model
## Displayed:
## * Variance-covariance component summaries
## Use the '$' sign to access parameters
## ========================================================
##     Multivariate Linear Mixed Model fit by REML
## ********************  sommer 2.7  **********************
## ========================================================
## Method:[1] "MNR"
##        logLik      AIC     BIC
## MVM -261.3855 526.7709 535.946
## ========================================================
## Variance-Covariance components:
##
## Var-Covar(g(id))
##           color Firmness
## color    0.7695   0.1323
## Firmness 0.1323   0.4954
##
## Var-Covar(Residual)
##             color Firmness
## color     0.39991 -0.01485
```

```
## Firmness -0.01485  0.46815
## ========================================================
## Standard errors for variance components:
##                             VarComp VarCompSE  Zratio
## g(id).color-color           0.76953   0.15437  4.9848
## g(id).color-Firmness        0.13228   0.09560  1.3838
## g(id).Firmness-Firmness     0.49536   0.11634  4.2578
## Residual.color-color        0.39991   0.04392  9.1064
## Residual.color-Firmness    -0.01485   0.03164 -0.4693
## Residual.Firmness-Firmness  0.46815   0.04664 10.0385
## ========================================================
## Fixed effects:
##             color Firmness
## (Intercept) 0.023 -0.02618
## ========================================================
## Groups and observations:
##       Observ Groups
## g(id)    363    363
## ========================================================
## Use the '$' sign to access parameters
```

8) Specifying heterogeneous variances in univariate models

Very often in multi-environment trials, the assumption that genetic variance is the same across locations may be too naive. Because of that, specifying a general genetic component and a location specific genetic variance is the way to go. Although the function 'mmer' implemented in sommer can be used to do that, can be quite cumbersome and messy to create the incidence and variance covariance matrices for fitting those models. For that reason the function 'mmer2' was added to the package to make such models easier to fit.

We estimate variance components for GCA_2 and SCA specifying the variance structure.

```
data(cornHybrid)
hybrid2 <- cornHybrid$hybrid # extract cross data
head(hybrid2)
```

```
##   Location GCA1   GCA2         SCA Yield PlantHeight
## 1        1 A258 AS5707 A258:AS5707    NA          NA
## 2        1 A258     B2     A258:B2    NA          NA
## 3        1 A258    B99    A258:B99    NA          NA
## 4        1 A258   LH51   A258:LH51    NA          NA
## 5        1 A258   Mo44   A258:Mo44    NA          NA
## 6        1 A258  NC320  A258:NC320    NA          NA
```

```
### fit the model
modFD <- mmer2(Yield~1, random=~ GCA2 + at(Location,c("3","4")):GCA2,
               data=hybrid2, silent = TRUE)
summary(modFD)
```

```
##
## Information contained in this fitted model:
## * Variance components, Residuals, Fitted values
## * BLUEs and BLUPs, Inverse phenotypic variance(V)
```

```
## * Variance-covariance matrix for fixed & random effects
## * Predicted error variance (PEV), LogLikelihood
## Use the '$' symbol to access such information
## =========================================================
## Linear mixed model fit by restricted maximum likelihood
## ********************  sommer 2.7  ********************
## =========================================================
## Method:[1] "NR"
##
## logLik    AIC    BIC
##  -1404   2810   2814
## =========================================================
## Random effects:
##                                     VarComp VarCompSE Zratio
## Var(GCA2)                             42.04     22.52  1.867
## Var(at(Location,c("3","4"))3:GCA2)    47.36     45.02  1.052
## Var(at(Location,c("3","4"))4:GCA2)   124.69     70.36  1.772
## Var(Residual)                        372.97     28.29 13.182
## Number of obs: 400  Groups: 20 20 20
## =========================================================
## Fixed effects:
##              Value Std.Error t.value
## Intercept 138.8469    1.8695  74.269
## =========================================================
## Use the '$' symbol to access all information
```

In addition, other functions can be added on top to fit models with covariance structures:

```
data(cornHybrid)
hybrid2 <- cornHybrid$hybrid # extract cross data
## get the covariance structure for GCA2
A <- cornHybrid$K
## fit the model
modFD <- mmer2(Yield~1, random=~ g(GCA2) + at(Location):g(GCA2),
               data=hybrid2, G=list(GCA2=A),
              silent = TRUE, draw=FALSE)
summary(modFD)
```

```
##
## Information contained in this fitted model:
## * Variance components, Residuals, Fitted values
## * BLUEs and BLUPs, Inverse phenotypic variance(V)
## * Variance-covariance matrix for fixed & random effects
## * Predicted error variance (PEV), LogLikelihood
## Use the '$' symbol to access such information
## =========================================================
## Linear mixed model fit by restricted maximum likelihood
## ********************  sommer 2.7  ********************
## =========================================================
## Method:[1] "NR"
##
## logLik    AIC    BIC
##  -1407   2815   2819
```

```
## ========================================================
## Random effects:
##                              VarComp VarCompSE  Zratio
## Var(g(GCA2))                   20.25     14.42  1.4042
## Var(at(Location)1:g(GCA2))      0.00     14.73  0.0000
## Var(at(Location)2:g(GCA2))      0.00     14.73  0.0000
## Var(at(Location)3:g(GCA2))     16.36     21.80  0.7505
## Var(at(Location)4:g(GCA2))     75.49     39.27  1.9222
## Var(Residual)                 381.91     29.32 13.0256
## Number of obs: 400  Groups: 20 20 20 20 20
## ========================================================
## Fixed effects:
##              Value Std.Error t.value
## Intercept 138.5460    1.3708  101.07
## ========================================================
## Use the '$' symbol to access all information
```

Good luck with your analysis.

Literature

Covarrubias-Pazaran G. 2016. Genome assisted prediction of quantitative traits using the R package sommer. PLoS ONE 11(6):1-15.

Bernardo Rex. 2010. Breeding for quantitative traits in plants. Second edition. Stemma Press. 390 pp.

Gilmour et al. 1995. Average Information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51(4):1440-1450.

Henderson C.R. 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics vol. 31(2):423-447.

Kang et al. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723.

Lee et al. 2015. MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information. Cold Spring Harbor. doi: http://dx.doi.org/10.1101/027201.

Searle. 1993. Applying the EM algorithm to calculating ML and REML estimates of variance components. Paper invited for the 1993 American Statistical Association Meeting, San Francisco.

Yu et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Genetics 38:203-208.

Abdollahi Arpanahi R, Morota G, Valente BD, Kranis A, Rosa GJM, Gianola D. 2015. Assessment of bagging GBLUP for whole genome prediction of broiler chicken traits. Journal of Animal Breeding and Genetics 132:218-228.

Tunnicliffe W. 1989. On the use of marginal likelihood in time series model estimation. JRSS 51(1):15-27.