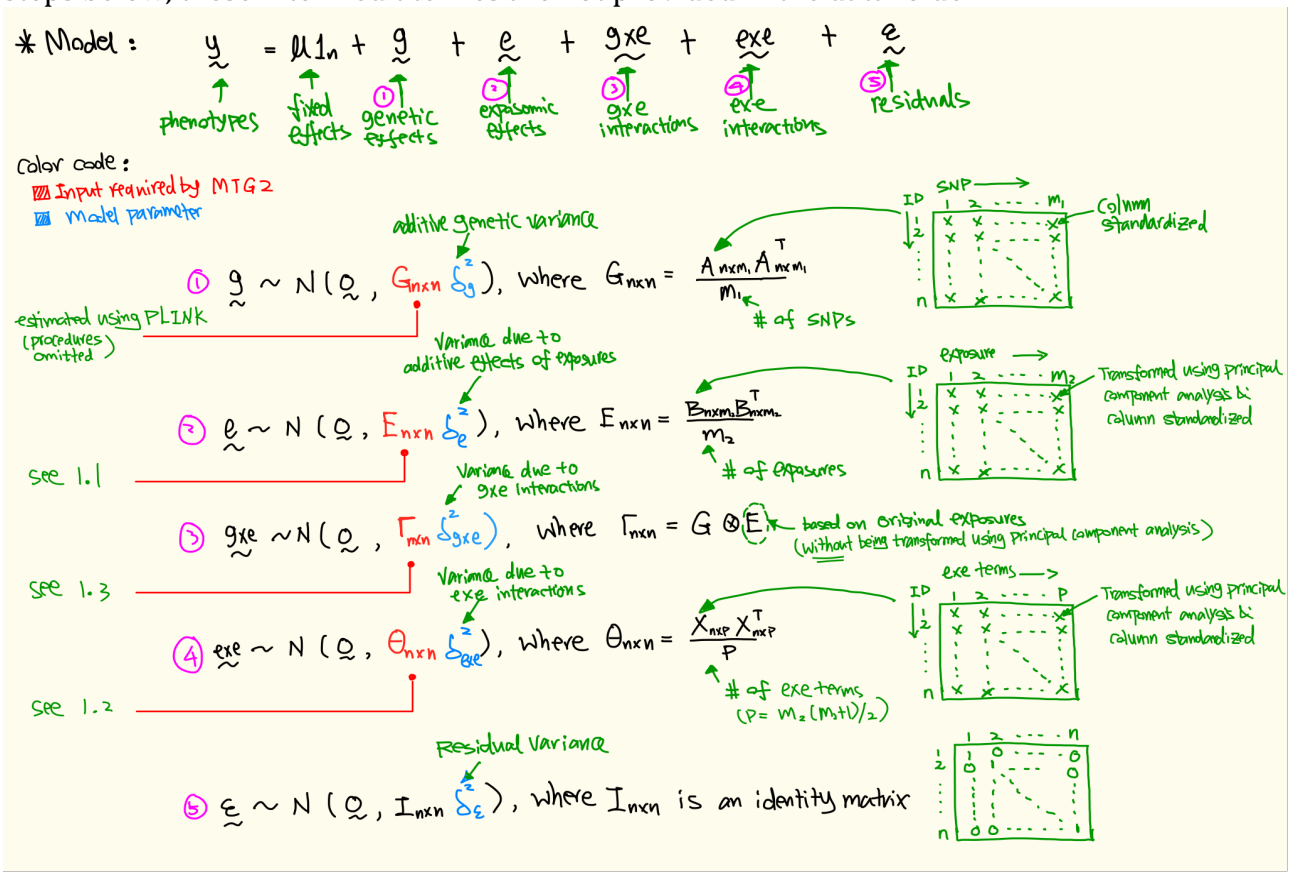
Integrative analysis of genomic and exposomic data using MTG2

0. Read me first

Here we use a toy example to illustrate how to use `MTG2` to perform integrative analyses of genomic and exposomic data (IGE). Suppose we have phenotypic data of a main trait for 1,000 participants, their genotype data and data of 10 exposomic variables or exposures. The interest is estimate phenotypic variance explained by additive genetic effecs, additive effects of the exposomic variables, interactions between exposomic variables (or exe interactions), and interactions between genotypes and exposomic variables (i.e., gxe interactions). The model to be fitted and steps involved are summarized in figures below.

NOTE: Since intermediate files used for fitting the model can be generated by following the steps below, these intermediate files are not provided in the data folder.



1. compute relationship matrices

1.1 E

```
exposures0=read.table("e.var", header=T)
# e.var is a 1000 x (10+1) matrix containing values of 10 exposures for 1000
participants, with id as the 1st column

# perform PCA
exposures1 = as.matrix(exposures0[,-1])
covmat=var(exposures1)
eig=eigen(covmat)
exposures2=exposures1%*%eig$vectors
exposures3=scale(exposures2)
#NOTE: exposure3 can be stored for GWEIS input (-snpbvr … (see section 3))
m=dim(exposures3)[2] # note: here m is the number of variables
# divide each element by sqrt(m)
exposures4=exposures3/sqrt(m)

# export data
write.table(cbind(exposures0$id,exposures0$id,exposures4),"e_pca_div_sqrtvarn
um", col.names=F, row.names=F,quote=F)

# create E
system(paste0("./mtg2.18 -p toy.fam -pdmx e_pca_div_sqrtvarnum -thread 80 -ou
t toy.e"), wait=F)
```

1.2 Θ

```
# load exposures
dat=read.table("e.var",header=T, stringsAsFactors = F)
e=scale(dat[,-1])

# exe interaction variables
for(i in 1:dim(e)[2]){
  col1=e[,i]
  for(j in i:dim(e)[2]){
    col2=e[,j]
    product=col1*col2
    if(i==1 & j==1){X=product}else{X=cbind(X,product)}
  }
}

# apply pca & standardization
covmat=var(X)
eig=eigen(covmat)
exposure1=X%*%eig$vectors
exposure2=scale(exposure1)
m=dim(exposure2)[2]
#NOTE: exposure2 can be stored for GWEIS input (-snpbvr … (see section 3))
```

```
# divide each element by sqrt(m)
exposure3=exposure2/sqrt(m)

# export data
write.table(cbind(dat$id,dat$id,exposure3),"exe_pca_div_sqrtvarnum",
col.names=F, row.names=F,quote=F)

# create Theta
system(paste0("./mtg2.18 -p toy.fam -pdmx exe_pca_div_sqrtvarnum -thread 80 -
out toy.exe"), wait=F)

# toy.fam file contains IID and FID of the 1,000 participants in the toy
example
```

1.3 Γ

First, we estimate the relationship matrix using original exposures without being transformed using a principal component analysis.

```
exposures0=read.table("e.var", header=T)
exposures1 = scale(exposures0[,-1])
m=dim(exposures1)[2]
# divide each element by sqrt(m)
exposures2=exposures1/sqrt(m)

# export data
write.table(cbind(exposures0$id,exposures0$id,exposures2),"e_div_sqrtvarnum",
col.names=F, row.names=F,quote=F)

# create E using exposures without PCA transformation
system(paste0("./mtg2.18 -p toy.fam -pdmx e_div_sqrtvarnum -thread 80 -out to
y.e_noPCA"), wait=F)
```

Second, we use the hadamard product of genomic relation matrix (i.e., G) and the exposomic relation matrix (i.e., Θ obtained using the above code) to obtain the relationship matrix for the estimation of variance due to gxe interactions (i.e., Γ).

```
paste toy.e_noPCA toy.grm | awk '{ print $1, $2, $3 * $7 }' > toy.gxe
```

Note: 'toy.grm' is the genomic relationship matrix (i.e., the G matrix in the model description under 'read me first'), and it can be obtained using PLINK (procedures omitted here; see PLINK for detail). 'toy.grm' contains three columns, the first two columns specify the position of off-diagonal entries of the relationship matrix (in the 3rd column). IMPORTANT: relationship matrices obtained using PLINK contain 4 columns instead; the extract column (i.e., 3rd) should be removed before using the syntax above. Alternatively, please adjust the above syntax to suit the file format of the relationship matrix.

1.4 Γ (using actual genotypic and E variables)

The same relationship matrix (toy.gxe) can be obtained from the Hadamard products of every pairs of genotypic and E variables, i.e.

```
# To get toy.gxe using column-standardized SNP coefficient matrix (e.g. using plink –recode A)


system(paste0("./plink1.9 --bfile toy --recode A --out toy"), wait=F)

plink=read.table("toy.raw",skip=1)
gen=scale(plink[,7:dim(plink)[2]])

# gxe interaction variables, i.e. pair-wise Hadamard products
for(i in 1:dim(gen)[2]){
  cat(i,"\n")
  col1=gen[,i]
  for(j in 1:dim(e)[2]){
    col2=e[,j]
    product=col1*col2
    if(i==1 & j==1){X=product}else{X=cbind(X,product)}
  }
}
#NOTE: X can be stored for GWEIS input (-snpbvr … (see section 3))

X=X/sqrt(dim(X)[2])
# export data

write.table(cbind(dat$id,dat$id,X),"gxe.variables",                col.names=F,
row.names=F,quote=F)

# create gxe kernel matrix

system(paste0("./mtg2.18  -p  toy.fam  -pdmx  gxe.variables  -thread  80  -out
toy.gxe"), wait=F)
```

This is useful when gxe.variables is required for GWEIS (SNP_BLUP) (see section 3 below, i.e. -snpbvr …). From toy.raw (first row) and GWEIS summary stats, allelic substation effect fro the reference allele for each SNP per environment can be found (see section 3).

2. model fitting

IGE provides estimated variance components via the following command:

```
./mtg2.18 -p toy.fam -mg toy.mat -d toy.dat -mod 1 -thread 100 -out toy.out > toy.log
```

where toy.mat has four relationship matrices (i.e. multiple random effects)

<toy.mat>

```
toy.grm
toy.e
toy.exe
toy.gxe
```

The output file <toy.out> contains variance estimates & SE

```
Ve        0.2231        0.0243 (estimated σ²ε & se)
Va        0.2665        0.0355 (estimated σ²g & se)
Va        0.2818        0.1279 (estimated σ²e & se)
Va        0.1212        0.0277 (estimated σ²exe & se)
Va        0.0866        0.0239 (estimated σ²gxe & se)
```

3. GWEIS / SNP BLUPs

Using the SNP BLUP method (see section 16 of MTG2 manual), IGE can output estimated beta coefficients (SE & p-value) of standardized SNP genotypes that are stored in the n x m1 A matrix (see the model description under 'read me first'). Importantly, this analysis can be applied to other variables of an IGE to output estimated beta coefficients of environmental exposures, exe interactions and gxe interactions.

Note the BLUP analysis is not REML; as such, estimated variance components from IGE should be provided as starting values (-sv toy.out). For example, to get BLUPs of the environmental exposures, the following syntax should be used:

```
./mtg2.18 -p toy.fam -d toy.dat -mg toy.mat -mod 1 -sv toy.out -thread 20 -
snpblup 2 -snpbvr e.wmat -wmatcol 10
```

where -snpblup 2 is to obtain beta coefficients for the 2nd component (i.e., effects of exposures; in the order of listed relationship matrix in toy.mat), -snpbvr e.wmat is the column-standardized matrix that contains exposures (i.e. exposure3 in section 1.3), -wmatcol 10 is the number of columns of e.wmat.

The output file will be named 'e.wmat.sumstat' (i.e., sumstat is attached after the name of the provided matrix, which is 'e.wmat' in this example)

<e.wmat.sumstat>

trait	SNP_EBVs	r2	SE	PEV	Chi	P
1	-0.17797	0.98482	0.16659	0.00043	1.141	0.28537E+00
1	0.18542	0.98496	0.16660	0.00042	1.239	0.26572E+00
1	-0.02133	0.98465	0.16658	0.00043	0.016	0.89810E+00
1	0.28916	0.98530	0.16663	0.00041	3.011	0.82683E-01
1	-0.24270	0.98477	0.16659	0.00043	2.123	0.14514E+00
1	0.25330	0.98547	0.16665	0.00041	2.310	0.12851E+00
1	-0.01004	0.98466	0.16658	0.00043	0.004	0.95194E+00
1	0.02040	0.98546	0.16664	0.00041	0.015	0.90259E+00
1	0.02250	0.98560	0.16666	0.00041	0.018	0.89259E+00
1	0.05828	0.98614	0.16670	0.00039	0.122	0.72664E+00

NOTE: the matrix that contains standardized SNP information (used to construct the genomic relation matrix 'toy.grm') can be obtained using PLINK (use `--recode A`; see PLINK manual for more detail). PLINK can convert PLINK files to a matrix with 0, 1 and 2 SNP coefficients which can be column-standardized for the estimation of SNP BLUPs using MTG2.

The above procedure can be used to obtain effect estimates for all other variance components (i.e., g, exe & gxe), when the order of the component and its associated column-standardized matrix are specified. Since the relationship matrix for gxe, i.e., Γ, is obtained using the hadamard product of G and and E (see the model description under 'read me first'), the matrix containing standardized gxe variables are not available. However, this matrix can be created using the same method as for exe interaction variables (see 1.3 above). Note that when the number of SNP is large, the process can be slowed down. It should be suggested to select a subset of SNPs (as discussed in Moore et al.).

For gxe component, the reference allelic substitution effect of each SNP (SNP BLUP) can be obtained for each environment (each of E variables), which can be derived from toy.raw (see section 1.4, i.e. reference allele information is in the first row) and summary stats results.

NOTE: Please also see section 16 (16-2) in MTG2 manual for more detail on SNP BLUP and example13 and 13-2.