Emhimad Abdalla
Emhi.abdalla@gmail.com

Constructing G matrix with missing SNPs:

Handling missing SNPs when calculating the **genomic relationship matrix** (**G;** VanRaden (2008)) is relatively straightforward. The standard procedure used for complete data can still be followed, with one key modification: when multiplying the **Z matrix** by its transpose, the missing genotype values should be treated as zero. This ensures that missing markers do not contribute to the calculation, effectively excluding them from the computation as if they were not present.

Example:
Suppose we have the following genotype matrix **M** for 3 animals and 2 SNPs, where SNP #1 for animal #2 is missing:

$$\mathbf{M} = \begin{bmatrix} 1 & 2 \\ \text{NA} & 0 \\ 2 & 0 \end{bmatrix}$$

If you are using **R**, missing values should be represented by `NA`. For instance, if missing values were previously coded as 5, you can convert them to `NA` using:

```
M[M==5]=NA
```

Now, the allele frequencies and the **P matrix** can be calculated in the usual way, without requiring any special treatment for the missing data:

```
n0 <- apply(M==0,2,sum,na.rm=T)
n1 <- apply(M==1,2,sum,na.rm=T)
n2 <- apply(M==2,2,sum,na.rm=T)
# Compute Z as described in Mrode 2015, page 180.
freq <- ((2*n2)+n1)/(2*(n0+n1+n2))
freq
#[1] 0.7500000 0.3333333

freq2 <- 2*(freq)
# create matrix P
P <- matrix(freq2, nrow=1)
P <- P[rep(1:nrow(P), times = nrow(M)), ]
P <- round(P, digits = 2) # This is just to round the outputs!
P
#     [,1] [,2]
#[1,]  1.5 0.67
#[2,]  1.5 0.67
#[3,]  1.5 0.67
```

Next, comput **Z** matrix:

Emhimad Abdalla
Emhi.abdalla@gmail.com

```
Z = M - P
Z
#      [,1]  [,2]
#[1,] -0.5  1.33
#[2,]   NA -0.67
#[3,]  0.5 -0.67
```

Since we need to multiply **Z** by its transpose, when replacing NAs by 0s, the missing SNPs do not contribute to the result:

The matrix Z above is:

$$\mathbf{Z} = \begin{bmatrix} -0.5 & 1.33 \\ 0 & -0.67 \\ 0.5 & -0.67 \end{bmatrix}$$

The missing SNP = NA as above

And its transpose is:

$$\mathbf{Z}' = \begin{bmatrix} -0.5 & 0 & 0.5 \\ 1.33 & -0.67 & -0.67 \end{bmatrix}$$

Now we can construct the **ZZ':**

$$\mathbf{ZZ}' = \begin{bmatrix} (-0.5 \times -0.5) + (1.33 \times 1.33) & (-0.5 \times 0.0) + (1.33 \times -0.67) & (-0.5 \times 0.5) + (1.33 \times -0.67) \\ (0.0 \times -0.5) + (-0.67 \times 1.33) & (0.0 \times 0.0) + (-0.67 \times -0.67) & (0.0 \times 0.5) + (-0.67 \times -0.67) \\ (0.5 \times -0.5) + (-0.67 \times 1.33) & (0.5 \times 0.0) + (-0.67 \times -0.67) & (0.5 \times 0.5) + (-0.67 \times -0.67) \end{bmatrix}$$

Note that the missing SNP did not contribute to the results (since they were multiplied by 0s).

Accordingly, the **ZZ'** is:

$$\mathbf{ZZ}' = \begin{bmatrix} 2.0189 & -0.8911 & -1.1411 \\ -0.8911 & 0.4489 & 0.4489 \\ -1.1411 & 0.4489 & 0.6989 \end{bmatrix}$$

Emhimad Abdalla
Emhi.abdalla@gmail.com

```
Z[is.na(Z)] = 0
ZZp <- Z%*%(t(Z))
ZZp
#         [,1]    [,2]    [,3]
#[1,]  2.0189 -0.8911 -1.1411
#[2,] -0.8911  0.4489  0.4489
#[3,] -1.1411  0.4489  0.6989
# OR
ZZP <- tcrossprod(Z)
ZZP
#         [,1]    [,2]    [,3]
#[1,]  2.0189 -0.8911 -1.1411
#[2,] -0.8911  0.4489  0.4489
#[3,] -1.1411  0.4489  0.6989
```

Another example:

Suppose we have this SNP data (M):

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ \text{NA} & 2 \\ 0 & 1 \end{bmatrix}$$

And suppose we calculated the allele frequencies and they were:

$$\mathbf{P} = \begin{bmatrix} -0.65 & 0.32 \\ -0.65 & 0.32 \\ -0.65 & 0.32 \end{bmatrix}$$

No, $\mathbf{Z} = \mathbf{M} - \mathbf{P}$

$$\mathbf{Z} = \begin{bmatrix} 1.65 & -0.32 \\ \text{NA} & 1.68 \\ 0.65 & 0.68 \end{bmatrix}$$

Replace NAs by 0s:

$$\mathbf{Z} = \begin{bmatrix} 1.65 & -0.32 \\ \boldsymbol{0} & 1.68 \\ 0.65 & 0.68 \end{bmatrix} \quad \text{and} \quad \mathbf{Z}' = \begin{bmatrix} 1.65 & \boldsymbol{0} & 0.68 \\ -0.32 & 1.68 & 0.68 \end{bmatrix}$$

Then **ZZ'** is:

Emhimad Abdalla
Emhi.abdalla@gmail.com

$$\mathbf{ZZ}^{'} = \begin{bmatrix} (1.65\times1.65)+(-0.32\times-0.32) & (-0.32\times1.68) & (1.65\times0.65)+(-0.32\times0.68) \\ (1.68\times-0.32) & (1.68\times1.68) & (1.68\times0.68) \\ (0.65\times1.65)+(0.68\times-0.32) & (0.68\times1.68) & (0.65\times0.65)+(0.68\times0.68) \end{bmatrix}$$

```
rm(list=ls(all=TRUE))
M <- read.table(textConnection("
                               1  0
                               NA 2
                               0  1
                               "), header=F)
M <- as.matrix(M)
closeAllConnections()
colnames(M) <- NULL

P <- read.table(textConnection("
                               -0.65 0.32
                               -0.65 0.32
                               -0.65 0.32
                               "), header=F)
P <- as.matrix(P)
closeAllConnections()
colnames(P) <- NULL

n0 <- apply(M==0,2,sum,na.rm=T)
n1 <- apply(M==1,2,sum,na.rm=T)
n2 <- apply(M==2,2,sum,na.rm=T)
# Compute Z as described in Mrode 2015, page 180.
freq <- ((2*n2)+n1)/(2*(n0+n1+n2))
freq
freq2 <- 2*(freq)
Z <- M-P
Z[is.na(Z)] = 0
Zp <- t(Z)
Z%*%Zp

D <- 2*(sum(freq*(1-freq)) )
G <- Z%*%Zp/D
# OR :
G <- tcrossprod(Z)/D
```

Emhimad Abdalla
Emhi.abdalla@gmail.com

References:

VanRaden, Paul M. "Efficient methods to compute genomic predictions." Journal of dairy science 91.11 (2008): 4414-4423.