Compute Centered and Scaled Genotype Matrix: W

Izel Fourie Sørensen, Pernille Merete Sarup, Palle Duun Rohde

April 6, 2018

Here we present two approaches for preparing the centered and scaled genotype matrix, $\boldsymbol{W}$. The genotype data prepared earlier $\boldsymbol{W}$ is computed based on the genotype data prepared earlier in the qgg user guide.

Approach 1

The additive genomic relationship matrix $\boldsymbol{G}$ (VanRaden PM. 2008. J Dairy Sci. 91:4414-4423) is constructed using all genetic markers as follows: $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{W}'/m$, where $\boldsymbol{W}$ is the centered and scaled genotype matrix, and m is the total number of markers. Each column vector of $\boldsymbol{W}$ was calculated as follows: $w_i = (m_i - 2p_i)/\sqrt{(2p_i(1-p_i)}$, where p_i is the minor allele frequency of the i^{th} genetic marker and m_i is the i^{th} column vector of the allele count matrix, $\boldsymbol{M}$, which contains the genotypes coded as 0, 1 or 2 counting the number of minor allele. (For the nearly homozygous DGRP lines the genotypes are coded as 0 or 2.)

Load the edited SNP genotype data file, `snpGE.Rdata`, created previously in the qgg user guide. The genotype data frame, `snpG`, is loaded directly into the object `W` by the `readRDS()` function.

```
W <- readRDS(file="./genotypes/snpGE.rds")
```

Count the number of minor alleleles (`nMinor`) and total number of alleles (`nAlleles`) for each SNP. Actually for `nAlleles` we count the number of genotypes that were measured per row/SNP (thus not "NA's"). This corresponds to counting one allele per genotype. Therefore the total number of alleles are 2 times `nAlleles`.

```
nMinor <- rowSums(W, na.rm=TRUE)
nAlleles <- rowSums(!is.na(W))
```

Compute minor allele frequencies:

```
p<-nMinor/(2*nAlleles)
  min(p)
  max(p)
```

Center and scale $\boldsymbol{W}$ using the observed allele frequencies:

```
for ( i in 1:205) {
  W[,i] <- (W[,i]-2*p)/sqrt(2*p*(1-p))
  isNA <- is.na(W[,i])
  W[isNA,i] <- 0
}
W <- t(W)
```

Save centered and scaled $\boldsymbol{W}$ calculated with approach 1 as `dgrp2_W1.Rdata`.

```
save(W, file="./genotypes/dgrp2_W1.Rdata")
```

Approach 2

In this approach the columns in W is scaled using the `scale()` function whose default method centers and/or scales the columns of a numeric matrix.

```
rm(list=ls(all=TRUE))
```

Load the edited SNP genotype data file:

```
W <- readRDS(file="./genotypes/snpGE.rds")
```

$\boldsymbol{W}$ is transposed because the `scale()` function by default scales the *columns* of a numeric matrix.

```
W <- t(W)
W <- scale(W)
dim(W)
```

Set missing values equal to 0.

```
for ( i in 1:205) {
  isNA <- is.na(W[i,])
  W[i, isNA] <- 0
}
```

Save centered and scaled $\boldsymbol{W}$ calculated with approach 2 as `dgrp2_W2.Rdata`.

```
save(W, file="./genotypes/dgrp2_W2.Rdata")
```