**Lab5 SNP effects and GWAS in ssGBLUP**

Prepared by D. Lourenco, A. Legarra, Z. Vitezica, and I. Aguilar

The data for this lab was simulated by D. Lourenco using QMSim (Sargolzaei & Schenkel, 2009). A single trait animal model was simulated assuming heritability of 0.40. All the genetic variance was explained by 500 QTL. Animals were genotyped for 45,000 SNP and the average LD was 0.18. The simulated additive genetic variance was 0.4 and the residual variance was 0.60. Phenotype was generated using the following model:

*Phenotype = sex_effect + true_breeding_value + residual*

**Description of files:**

**data3.txt:**
1: animal ID
2: generation
3: sex
4: phenotype
5: true breeding value (TBV)

**ped3.txt:**
1: animal ID
2: sire ID
3: dam ID

**snp3.2k:**
1: animal ID
2: SNP genotype

**mrkmap.txt**:
1: SNP ID
2: Chromosome
3: position

1. Files are available in the folder day5. Copy the entire folder using the following command:
   `cp -r /home/labs/lab5 .`

2. Run `renumf90` program using `renum.par` parameter file to renumber the data.

3. Calculating SNP effects in ssGBLUP: given that SNP effects are calculated based on GEBV, run `blupf90` to get solutions. Before that, check the options you will need to include in the parameter file, so `blupf90` can provide all the files needed for the calculation of SNP effects (in exercise 5). Assume that quality control was already done (i.e., turn off the intrinsic quality control).

4. Add an option to read a map file (mrkmap.txt) and run `postGSf90`. Check the output files and the content of each column (http://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90).

5. Indirect predictions for young individuals: `postGSf90` creates a file `snp_pred` with information about the random effects (number of traits + correlated effects), the gene frequencies and the solutions of SNP effects. This is the file used by `predf90` to provide indirect predictions for young genotyped animals as `Za`, where `Z` is a matrix of SNP content and `a` is a vector of SNP effects. Run `predf90` to get indirect predictions for young individuals that were not included

in the `blupf90` and `postGSf90` runs. Genotypes for young animals are in a file called `new_animals`

6. Single-step GWAS (ssGWAS): SNP effects computed based on GEBV are used to compute SNP weights or variance explained by SNP. The default way to calculate SNP weight (`w`) in postGSf90 is:

$$w_i = 2p_i(1-p_i)a_i^2$$

where `p` is the allele frequency and $a$ is SNP effect. A new method that has better convergence properties has been recently implemented in `postGSf90`. This method is called nonlinearA and is described in VanRaden (2008) as:

$$w_i = \mathrm{CT}^{\frac{|\widehat{a_i}|}{sd(\hat{a})}-2}$$

where CT is a constant set to 1.125, and $\frac{|\widehat{a_i}|}{sd(\hat{a})}$ is capped to 5 by default. To use this method, the following option should be used in `postGSf90`:

```
OPTION which_weight nonlinearA
```

Run postGSf90 including an option to calculate variance based on a window of 20 SNPs and an option to generate Manhattan plots. Use the default linear weight and the nonlinearA weight. Check the output files and compare results.

Hint 1: `postGSf90` reads a file with weights for each SNP if `OPTION weightedG file_name` is used. By default, all SNP have the same weight, so this file is actually a vector of dimension N x 1; where N is the number of SNP. If there are 50,000 SNP, the following `awk` command will create a vector of dimension 50,000 x 1: `awk 'BEGIN { for (i==1;i<50000;i++) print 1}' > w.txt`

Hint 2: `postGSf90` prints Manhattan plots on the screen and also creates files for printing in R (Sft1e2.R and Vft1e2.R) and in Gnuplot (Sft1e2.gnuplot and Vft1e2.gnuplot).

7. Computing p-values for SNP in ssGWAS: Although variance explained by SNP (SNP weight) is useful in the context of breeding and genetics, p-values can be computed to declare significance of SNP. This computation requires the inverse of the LHS of the MME (Aguilar et al., 2019 - https://doi.org/10.1186/s12711-019-0469-3). When p-values are computed, only 1 iteration should be used.

Run blupf90 and postGSf90 using the following option in both programs to obtain p-values for SNP:

```
OPTION snp_p_value
```

8. Iterative WssGBLUP: The weights computed in exercise 6 can be used to construct a weighted genomic relationship matrix $\mathbf{G_w} = \frac{\mathbf{ZDZ'}}{2\sum p_i(1-p_i)}$, and this matrix is used to compute new GEBV assuming SNP explain different proportions of variance. In this way, ssGBLUP becomes weighted ssGBLUP (WssGBLUP). The WssGBLUP is an iterative method, where weights are used in several rounds, until there is no more change between weights in iteration t and t-1. Usually, 3 to 5 rounds are enough to obtain convergence. Run one more round

of `blupf90` and `postGSf90` for both linear and nonlinearA methods. Compare Manhattan plots and maximum variance explained.

Hint: updated weights are in column 7 of snp_sol from exercise 6. Use the following Unix command to create the new weight file (dimension N x 1) to be used in exercise 8.

```
awk '{ if ($1==1) print $7}' snp_sol > W
```