Lab3 SNP effects and GWAS in ssGBLUP + Metafounders

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. The 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 on the website. Use curl to download it to your Linux or Mac device:

```
curl http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=lab3_UF.zip -o lab3.zip
```

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

2. 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 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 windows of 20 SNPs and an option to generate Manhattan plots (`OPTION Manhattan_plot`). Use

the default linear weight and the nonlinearA weight. Include an option to compute p-values (`OPTION snp_p_value`). Check the output files and compare results.

Hint 1: 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). To obtain p-values for SNP, both blupf90 and postGSf90 should include `OPTION snp_p_value`.

Hint 2: `postGSf90` prints Manhattan plots on the screen and also creates files for printing in R (Sft1e2.R, Vft1e2.R, Pft1e2.R) and in Gnuplot (Sft1e2.gnuplot, Vft1e2.gnuplot and Pft1e2.gnuplo).

Hint 3: Check all the options related to GWAS here:
http://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90

Hint 4: Although this exercise was based on ssGBLUP, you can use the tricks from yesterday to switch to GBLUP.

3. Iterative WssGBLUP: The weights computed in exercise 2 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 a 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, 2 to 5 rounds are enough to obtain convergence. Run two rounds of `blupf90` and `postGSf90` for both linear and nonlinearA methods. Compare breeding values and SNP effects from both runs. You will see that Manhattan plots can also be generated. Check the Manhattan plots and the maximum variance explained by SNP.

   Hint 1: `blupf90` and `postGSf90` read 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 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: updated weights are in column 7 of snp_sol (output from `postGSf90`). Use the following Unix command to create the new weight file (dimension N x 1) to be used in the second round of `blupf90` and `postGSf90`.

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

OPTIONAL

4. Using metafounders to set the base populations. Files are available in the folder day3_metafounders: The data for this lab was simulated by D. Lourenco (Lourenco et al., 2016) using QMSim (Sargolzaei & Schenkel, 2009). A single trait animal model was simulated assuming heritability of 0.30. All the genetic variance was explained by 400 QTL. Two lines (1 and 2) under 9 generations of selection were simulated. Pure and F1 (12) progeny were generated in generation 10. Animals were genotyped for 40,000 SNP. The simulated additive genetic variance was 0.3 and the residual variance was 0.70. The phenotype was generated using the following model:

   Phenotype = general_mean + true_breeding_value + residual

Files are available in the website. Use curl to download it to your Linux or Mac device:

```
curl http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=lab3mf_UF.zip -o
lab3mf.zip
```

Description of files:

newdata.txt:
1: animal ID
2: sire ID
3: dam ID
4: generation
5: phenotype
6: true breeding value (TBV)
7: line code
8: mean incidence

newped.txt:
1: animal ID
2: sire ID
3: dam ID

snp_file.txt:
1: animal ID
2: SNP genotype

a) Run renumf90 and ssGBLUP using blupf90test.

b) Replace missing parents with -1 in line 1 and -2 in line 2 (MF coding).

c) In a separate folder, run renumf90 with the modified pedigree that contains metafounders, estimate Gamma using gammaf90, and run ssGBLUP using blupf90test. Do not forget to replace to rename gamma.txt and change the random type in the parameter file (check the slides).

d) Compare solutions from a) and b). One way to compare the methods is to perform a validation for young individuals (e.g., accuracy, level bias – b_0, and dispersion bias – b_1). For that, remove phenotypes for individuals in generation 10. Run ssGBLUP with and without metafounders. Correlate solutions with TBV (column 6 in data3mf.txt). Do not forget that solutions are with renumbered IDs (different in ssGBLUP and ssGBLUP with MF) and TBV are with original IDs.