**Lab1 - Quality control of SNP data, VCE and genomic predictions with single-step using the BLUPF90 family**

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

The data for this lab was simulated using QMSim (Sargolzaei & Schenkel, 2009). A single trait animal model was simulated assuming heritability of 0.4. 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 0.40 and the residual variance was 0.60. The simulated phenotype was generated using the following model:

```
Phenotype = sex_effect + true_breeding_value + residual
```

Files are available in the folder day1. Copy the entire folder using the following command:
`curl http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=lab1_ufv.zip -o lab1.zip`

## Description of files

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

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

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

**mrkmap.txt**:
1: SNP ID
2: Chromosome
3: position
(Note: in this exercise, SNPs are sorted but this is not needed. The SNP ID has to match the order in the SNP file).

1. Modify an existent renumf90 parameter file (or create a new one), according to the data file, to fit the following model:

$$y = sex + animal + e$$

2. Run renumf90 program to renumber the data.

3. Check renf90.par, renf90.dat, and renaddxx.ped. From the renaddxx.ped file, identify genotyped animals, and check with wiki (http://nce.ads.uga.edu/wiki/doku.php?id=readme.renumf90) the content of each column. What is the content of **snp3.2k_XrefID**?

4. preGSf90 is a stand-alone program that encapsulates the genomic library including reading pedigree and markers, quality control and buildup of **G** and $\mathbf{A}_{22}$ and their inverses. Run preGSf90 including the option to save clean SNP file after quality control. In addition, do include `OPTION msg 100` (to have more output on the screen). Check the output. Which quality checks for both SNP and animals were done by default? Are there any duplicated genotypes? What is the correlation between **G** and $\mathbf{A}_{22}$? Check averages of **G** and $\mathbf{A}_{22}$.

5. Run blupf90+ with the VCE option to estimate variance components under AI-REML and EM-REML. Include the option to compute SE for heritability. Do it with and without genomic information. Compare CPU times across methods.

   Hint: use the following command to provide computing time while saving outputs to a log file:
   ```
   time blupf90+ renf90.par | tee blup1.log
   ```

6. Run blupf90+ to compute breeding values without SNP information (BLUP). Now run blupf90+ to compute breeding values using genomic information (ssGBLUP) and compare cpu time and solutions.
   Now run both BLUP and ssGBLUP with an option to compute and store reliabilities of breeding values. Save solutions with original ID.
   Hint: the current practice is to use variance components without gnomic information in ssGLBUP and BLUP. If the genotyped animals are a good representation of the pedigreed population, VCE should be similar.

7. Do a validation on young selection candidates or individuals from the 5th generation with genotypes and no phenotypes. Compare EBV and GEBV with true breeding value (TBV). Remember that correlation between (G)EBV and a benchmark (i.e., TBV) is a measure of accuracy. What happened with accuracy when genomic information was included? Check also intercept and regression coefficient from a regression of TBV on EBV and GEBV.

   Hint 1: remove the phenotypic information from the 5th generation and obtain solutions from a model with SNP information and with no SNP information.
   Hint 2: have renumf90 passing to the renumbered data a column containing generation number.
   Hint 3: if generation column is number 4, new data can be created using the AWK Linux tool:
   ```
   awk '$4!=5' renf90.dat > renf90.dat.reduced
   ```

8. Comparing EBV "before" and "After" – Assuming true breeding values are not known, as in real populations, do a validation based on the Method LR. Check correlations, intercept, and regression coefficient.

   Hint: The method compares EBV predictions with all data (whole, subindex "w") vs EBV with partial data (subindex "p"). Consider animals in the 5th generation as the "focal group", and compare their breeding values including their own records (whole data) or not (partial data). Comparing the EBVs of the same animals, those in the focal group, using either whole ($u_w$) or partial ($u_p$) yields statistics that are approximations to bias, slope, and ratios of accuracies:

   Bias: $\mu_{wp} = \overline{\hat{\boldsymbol{u}}_p} - \overline{\hat{\boldsymbol{u}}_w}$

   Slope or dispersion (also called $b_1$ and sometimes also called bias): $b_{w,p} = \frac{cov(\hat{\boldsymbol{u}}_w, \hat{\boldsymbol{u}}_p)}{var(\hat{\boldsymbol{u}}_p)}$

   Accuracies: $\rho_{p,w} = \frac{cov(\hat{\boldsymbol{u}}_p, \hat{\boldsymbol{u}}_w)}{\sqrt{var(\hat{\boldsymbol{u}}_w) var(\hat{\boldsymbol{u}}_p)}}$ is an estimator of the ratio of accuracies using the "partial" or the "whole" data $\frac{acc_p}{acc_w}$

   $b_{p,w} = \frac{cov(\hat{\boldsymbol{u}}_w, \hat{\boldsymbol{u}}_p)}{var(\hat{\boldsymbol{u}}_w)}$ is an estimator of the square of the ratio of accuracies $\left(\frac{acc_p}{acc_w}\right)^2$