RRBLUP and **RRBLUP2**

<No-frills software for genomewwide prediction via ridge regression-best linear unbiased prediction>

Rex Bernardo
University of Minnesota

Background

RRBLUP and *RRBLUP2* are no-frills computer programs for (1) calculating genomewide marker effects, estimating the predictive ability by cross-validation, and (3) predicting the performance of a test population.

Cross-validation is done by a delete-one procedure. Suppose there are 200 individuals in the training population. The performance of individual 1 is predicted from information on individuals 2-200. The performance of individual 2 is predicted from information on individuals 1 and 3-200. The procedure is repeated until the performance of each individual is predicted from information on the remaining 199 individuals. In the end, the correlation between the predicted and observed performance of the 200 individuals is calculated as the predictive ability.

RRBLUP and *RRBLUP2* run as executable files in Windows (`RRBLUP` and `RRBLUP2`) and in macOS (`RRBLUPmac` and `RRBLUP2mac`); see last section of this document for details. *RRBLUP* requires prior estimates of the proportion of the variation that is due to genetic effects. *RRBLUP2* does not require such prior estimates but requires about 20 times the computation time as *RRBLUP*.

Files for the Training Population

Both software programs require three input files for the training population. Make sure all of the **files are closed** in Microsoft Excel or in Google Sheets prior to running the program.

1. **Marker names and chromosomes** (Sample: `Chrom.csv`)
   - This CSV file can be exported from Microsoft Excel or Google Sheets.
   - The marker names should have no spaces or commas, and not be more than 30 characters long.
   - The chromosome names should be numbers, with no letters.
   - The SNPs should be ordered by chromosome (lowest to highest), but the SNPs within a chromosome need not be in map order.

2. **SNP marker data** (Sample: `TrainSnp.csv`)
   - The rows correspond to the SNP markers and the columns correspond to the lines or individuals.
   - The order of SNPs in this file should correspond to the order of SNPs in the first file.
   - The SNP genotypes need to be coded as 1 for one homozygote, 0 for the heterozygote, and –1 for the other homozygote.
   - There cannot be any missing data. Imputed marker genotypes with a non-integer code (e.g., 0.5, –0.3) are acceptable.

3. **Phenotypic data** (Samples: `TrainPhen.csv` for *RRBLUP*, and `TrainPhen2.csv` for *RRBLUP2*)
   - The rows correspond to the lines or individuals and the columns correspond to the traits.
   - In *RRBLUP2*, the first row gives the name of each trait.

- In *RRBLUP*,
    - The **first row** gives the estimated proportion (between 0.01 and 0.99) of the trait variation that is due to genetic effects. This estimate can be the entry-mean heritability, or the proportion of the sums of squares that is due to genetic effects from an ANOVA.
    - The **second row** gives the name of each trait.
- The trait name cannot have any spaces.
- The order of lines in this file (down the rows) should match the order of lines in the SNP marker data file (across the columns).
- For each trait, only one record is allowed per line or individual. In other words, if a line was evaluated in, say, 3 locations, the mean of the line across the 3 locations needs to be calculated. The mean is then entered as the single phenotypic record for the line or individual.
- There cannot be any missing data.

Files for the Test Population

Both software programs allow—but do not require—the prediction of performance of individuals in a test population:

1. **SNP marker data** (Sample: `TestSnp.csv`)
    - Same format as in the corresponding SNP marker data file for the training population.

2. **Names of test individuals** (Sample: `TestNames`)
    - The rows correspond to the names of the individuals or lines whose performance is being predicted.
    - The names cannot have any spaces.
    - The order of lines in this file (down the rows) should match the order of lines in the SNP marker data file (across the columns) for the test population.

Parameter File (Samples: `ParmsRRBLUP.csv` and `ParmsRRBLUP2.csv`)

- The parameter file and the output file need to be in CSV format.
- Use the sample files as templates. Change only the information in the second column, and rename the file as needed.

#Markers	1022
SNP&ChromFile	Chrom.csv
#Traits	2
TRAINING_POPULATION	Y
#Individuals	242
MarkerFile	TrainSnp.csv
PhenotypeFile	TrainPhen.csv
MarkerEffOutputFile	SNPEffects.csv
TEST_POPULATION	Y
#Individuals	30
MarkerFile	TestSnp.csv
NamesFile	TestNames.csv
PredictionsFile	TestPred.csv

Change this to **N** if you do not want to predict the performance of a test population. The remaining lines of the parameter file will then be ignored.

Running the Software in Windows

1. Create a folder for the software and save the executable files (`RRBLUP.exe` and `RRBLUP2.exe`), training-population files, test-population files, and parameter file in this folder.
2. Double-click on an executable file to run it.

Running the Software in macOS

1. Create a `Desktop/RRBLUP` folder. Save the executable files (`RRBLUPmac` and `RRBLUP2mac`), training-population files, test-population files, and parameter file in this folder.
2. Open Terminal (e.g., `Command + Spacebar`, then search for `Terminal`).
3. Type `cd Desktop/RRBLUP`

Steps 3a and 3b apply only the first time you run the software:

3a. In the Terminal, type the following to allow access to the files:

```
chmod u+rwx ~/Desktop/RRBLUP/*
```

3b. Then enter each of the following to declare that each program is not a virus:

```
xattr -d com.apple.quarantine RRBLUPmac
xattr -d com.apple.quarantine RRBLUP2mac
```

4. Run the executable file by typing `./RRBLUPmac` or `./RRBLUP2mac` (don’t forget to include `./`)