A brief of pipeline for genomic selection using PCR and PLSR

This document mainly introduces how to implement genomic selection using principal components regression (PCR) and partial least squares regression (PLSR) to produce the results in our study. They include two procedures, one for simulation data and the other for the real data sets. Two strategies are used to evaluate the performance of prediction, cross validation and HAT method. A standard genomic selection method (BLUP) is also included in this program. Note this is just a pipeline rather than R package.

The procedure is following:

(1) Changing current working directory using **setwd.**

(2) Two procedures are included in the simulation data: "sim.hat.R" using HAT method to evaluate the performance of model and "sim.cv.R" using cross validation to evaluate the performance of model. When used, firstly change into your phenotype and genotype filename. At the same time, set "model" parameter into the selected method (such as "pcr" and "blup" for sim.hat.R, "pcr" and "plsr" for sim.cv.R. Then **source("sim.hat.R")** or **source("sim.cv.R")** in R Console.

(3) Three real data sets are included in the analysis, four agronomic traits, 1000 metabolites and 24,973 expressions as the target phenotype in an RIL rice population with 210 lines and 1619 genotypes). Two procedures are included, "rice.cv.R" and "rice.hat.R", respectively corresponding to the cross validation and HAT methods. When performing analysis, the settings are similar to the simulation analysis, just changing the input and output filenames and selecting the using model. Then **source("rice.cv.R")** or **source("rice.cv.R")** in R Console.

(4) Input files contain genotype and phenotype files. And both of the two files are data frame, which can be read using function **read.csv.** All the involved data sets are

available in the github.