Genomic Selection Galaxy Analysis Pipeline

- Definition
- Field Descriptions
- Galaxy Analysis Pipeline Workflow
- Tutorial with Test Datasets

Definition

Genomic selection (GS): is a new approach for improving quantitative traits in large plant breeding populations that uses whole genome molecular markers (high density markers and high throughput genotyping). Genomic prediction combines marker data with phenotypic and pedigree data (when available) in an attempt to increase the accuracy of the prediction of breeding and genotypic values.

Training Population: also called candidate population or reference population that are both phenotyped and genotyped and can be used to predict the performance of related individuals that are only genotyped in related environment and management conditions.

Training Datasets: GS uses a training population of individuals with known phenotypes and marker data to build a model for the prediction of performance in a population of untested individuals based on marker data.

Prediction Populations: also called test or validation populations, which are genotyped but not phenotyped. Performance of individuals in the population are estimated based on marker effects in the training set.

Prediction Datasets: The predictor of an individual phenotype is the genomic estimated breeding value (GEBV) obtained as the sum of all corresponding marker effects of the individual.

Training model fitting: GS uses this marker data in two different ways: by using markers to model relationships between individuals or by estimating the effects of each marker on the trait of interest, some of which are presumed to be in linkage disequilibrium (LD) with relevant quantitative trait loci (QTL).

Prediction accuracies: are measured as correlations between the observed and predicted phenotypes using a cross validation method.

Cross validation: the cross validation method randomly partitions the genotypes into folds and isolates folds as target populations while omitting the target's phenotypic data. The remaining folds of genotypes with their phenotypic data intact are used as the training dataset; this process is repeated for each fold.

BLUP: best linear unbiased prediction

GEBV: genomic estimated breeding values, GEBVs are used to rank and select genotypes, without phenotypic data, for the next generation of breeding.

Methods of genomic prediction: Bayes B, Bayes C,Bayes Ridge Regression (BRR), Bayesian LASSO (BL), Bayesian Reproducing Kernel Hilbert Spaces, (RKHS) and Genomic BLUP (GBLUP).

The Bayesian methods address the problem of small number of observations (n) and a large number of parameters (p) to be estimated ($n<<p$) by restricting the size of the regression coefficients via shrinkage or regularization

COP: coefficient of parentage

BLUE: best linear unbiased estimator

PCA: principal component analysis

PEV: Prediction error variance

Field Descriptions

- Input file for phenotypes:
  the phenotype file contain two columns are:
  1. the sample names (alphanumeric values)
  2. the summarized phenotype data (real values)

Example: Test_Grain_PhenotypeforrAmpSeq.csv

- Input file for genotypes matrix, with sample names in the first row (unique and no duplicated sample names) and genotypic matrix in the table where marker names removed from the first column in the matrix:
  the genotype file which each column is a different sample name (alphanumeric values) with their genotypes data (real values) in its rows associated.

  Example: GenotypesForGS.csv

- Output files for predicted phenotypic values: output file contain three columns:

  1. the sample names (alphanumeric values)
  2. the observed phenotypic data (real values)
  3. the predicted phenotypic values

Example: Galaxy13-[Bayesian_Generalized_Linear_Regression_on_data_6_and_data_7].csv

Galaxy Analysis Pipeline Workflow

1. Prepare phenotype input files: 1. Prepare Phenotype Files
2. Prepare genotype input files: 2. Prepare Genotype Files
3. Log in Galaxy user interface
4. Import phenotype input files
5. Import genotype input files
6. Perform genotype data format encoding
7. Cross-reference check sample names and order in phenotype and genotype files: 3. Sample Matching
8. Select and define training datasets and parameters:
9. Define test datasets and parameters
10. Specify prediction models and parameters: 4. GEBV Calculators
11. Excute and run prediction analysis
12. Export prediction results
13. Visualize input and output results on visualization software
14. Make selection

The address of the CBSU/GOBII dockerized Galaxy is http://galaxy-demo.excellenceinbreeding.org/

Tutorial with Test Datasets

1. Download test datasets
2. Access to Galaxy through guest user log in
3. Download visualization software
4. Access to tutorials links to follow along