Tutorial of the OmniGS-R (v1.0)

Genomic Selection Pipeline Using R Packages

This tutorial provides step-by-step instructions for running three different use cases of OmniGS-R: Cross-Validation, prediction with test phenotype and prediction without test phenotype. It assumes you have installed the pipeline and its dependencies correctly.

Pipeline Setup and example Data

1. Ensure R and all required packages are installed as listed in the README.

2. Clone the repository and enter the example folder:

```
zhengc@biocomp-0-1.local:CZ_project>git clone https://github.com/ORDC-Crop-Bioinformatics/OmniGS-R.git
Cloning into 'OmniGS-R'...
remote: Enumerating objects: 178, done.
remote: Counting objects: 100% (178/178), done.
remote: Compressing objects: 100% (159/159), done.
remote: Total 178 (delta 62), reused 109 (delta 18), pack-reused 0 (from 0)
Receiving objects: 100% (178/178), 15.31 MiB | 11.86 MiB/s, done.
Resolving deltas: 100% (62/62), done.
zhengc@biocomp-0-1.local:CZ_project>cd OmniGS-R/example/
```

3. Review the example genotype, phenotype, and configuration files.

Cross-Validation

1. Input data: example/inputFile/train_genotype.vcf and example/inputFile/train_phenotype.txt

2. Configure Cross-Validation
   Edit the configuration file (e.g., gs_parameters_example_CV.config) to define cross-validation parameters. Specify the GS_Mode, number of replicates, input genotype and phenotype file, and output directory.

   As the first step, you can keep the same parameters as the file gs_example_CV.config. **But you MUST change the RScriptPath to point to your own Rscript in the configuration file.**

   # R path
   RScriptPath = your/Rscript

3. Run Cross-Validation
   Run the pipeline script using the example configuration:

   ./OmniGS-R_run_example_CV.sh

```
zhengc@biocomp-0-1.local:example>cat OmniGS-R_run_example_CV.sh
java -version
R --version

java -Xmx2000g -jar ../pipeline/OmniGS-R/OmniGS-R.jar  gs_parameters_example_CV.config
zhengc@biocomp-0-1.local:example>
zhengc@biocomp-0-1.local:example>./OmniGS-R_run_example_CV.sh
openjdk version "24.0.2-internal" 2025-07-15
OpenJDK Runtime Environment (build 24.0.2-internal-adhoc.conda.src)
OpenJDK 64-Bit Server VM (build 24.0.2-internal-adhoc.conda.src, mixed mode, sharing)
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under the terms of the
GNU General Public License versions 2 or 3.
For more information about these matters see
http://www.gnu.org/licenses/.


Check configuration...
√ All configuration parameters validated successfully

[Log] Writing console output to: outputFile_CV/gs_68687466975805498.log
```

4. Results and output Figures

The output folder contains detailed cross validation results (all_CV_results.txt) and summary statistics
(CV_summary_statistics.csv) for all the replications and folds.

```
zhengc@biocomp-0-1.local:example>ls -l outputFile_CV/
total 344
-rwxrwxr-x 1 zhengc domain users 21315 Oct 24 12:01 all_CV_results.txt
-rwxrwxr-x 1 zhengc domain users   479 Oct 24 12:01 CV_summary_statistics.csv
-rwxrwxr-x 1 zhengc domain users 14760 Oct 24 12:01 gs_68687466975805498.log
drwxrwxr-x 2 zhengc domain users   389 Oct 24 11:00 intermediate_data
drwxrwxr-x 2 zhengc domain users    94 Oct 24 11:00 pheno_data
drwxrwxr-x 2 zhengc domain users    48 Oct 24 12:01 plots
drwxrwxr-x 2 zhengc domain users     0 Oct 24 11:00 trait_predictions
zhengc@biocomp-0-1.local:example>
```
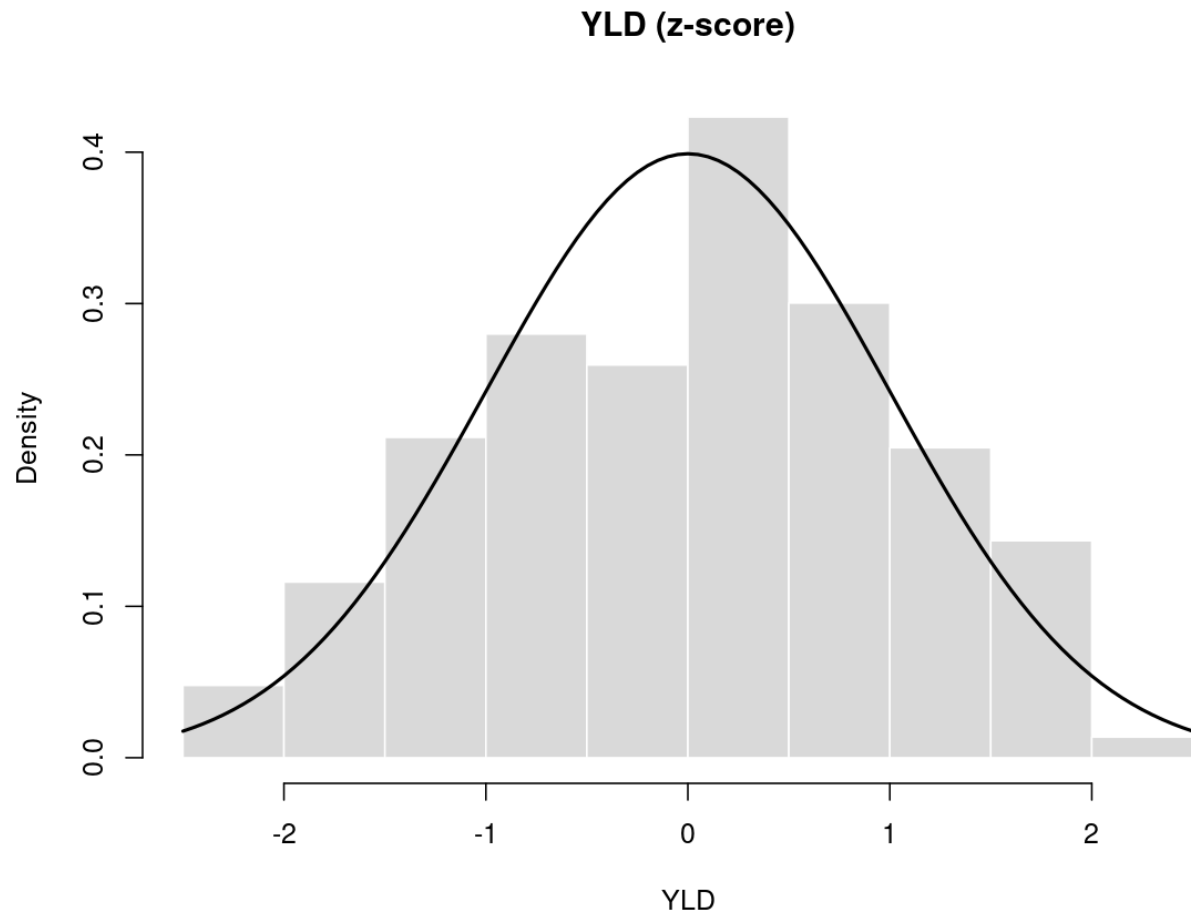
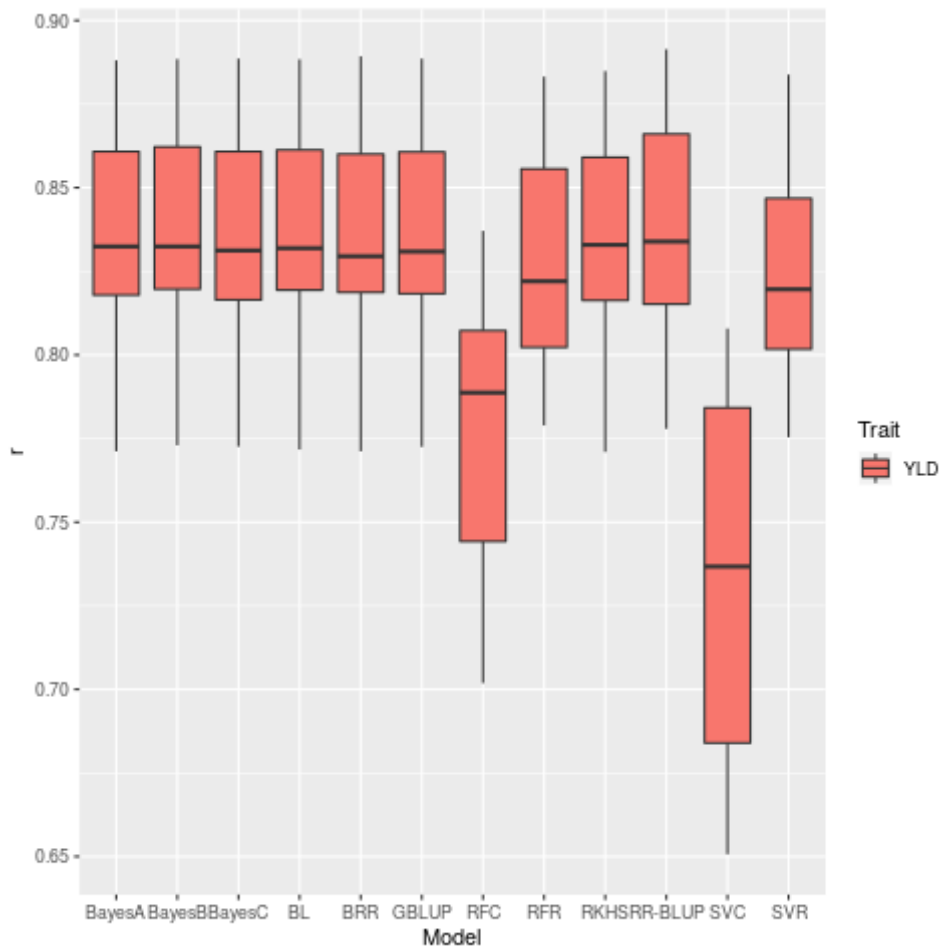Figure 1: Correlation between observed and predicted values across folds.

Figure 2: Correlation between observed and predicted values across folds.

Prediction with known test phenotype data

1. Input data:

    example/inputFile/train_genotype.vcf and example/inputFile/train_phenotype.txt

    example/inputFile/test_genotype.vcf and example/inputFile/test_phenotype.txt

2. Configure prediction with known test Phenotype data
    Edit the configuration file (e.g., gs_parameters_example_prediction1.config) to define cross-validation parameters. Specify the GS_Mode, number of replicates, input genotype and phenotype file, and output directory.

    As the first step, you can keep the same parameters as the file gs_parameters_example_prediction1.config. **But you MUST change the RScriptPath to point to your own Rscript.**

    # R path

3. Run the pipeline script using the example configuration:

```
zhengc@biocomp-0-1.local:example>cat OmniGS-R_run_example_prediction1.sh
java -version
R --version

java -Xmx2000g -jar ../pipeline/OmniGS-R/OmniGS-R.jar  gs_parameters_example_prediction1.conf
zhengc@biocomp-0-1.local:example>
zhengc@biocomp-0-1.local:example>./OmniGS-R_run_example_prediction1.sh
openjdk version "24.0.2-internal" 2025-07-15
OpenJDK Runtime Environment (build 24.0.2-internal-adhoc.conda.src)
OpenJDK 64-Bit Server VM (build 24.0.2-internal-adhoc.conda.src, mixed mode, sharing)
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under the terms of the
GNU General Public License versions 2 or 3.
For more information about these matters see
http://www.gnu.org/licenses/.


Check configuration...
√ All configuration parameters validated successfully

[Log] Writing console output to: outputFile_prediction_with_test_pheno/gs_68692401536040151.lo

================================================================
OmniGS-R (1.0): GENOMIC SELECTION PIPELINE USING R PACKAGES
================================================================
```

4. Results and output Figures.

```
zhengc@biocomp-0-1.local:example>ls -l outputFile_prediction_with_test_pheno/
total 248
-rwxrwxr-x 1 zhengc domain users 8242 Oct 24 12:45 gs_68692401536040151.log
drwxrwxr-x 2 zhengc domain users  827 Oct 24 12:41 intermediate_data
drwxrwxr-x 2 zhengc domain users   94 Oct 24 12:41 pheno_data
drwxrwxr-x 2 zhengc domain users 1528 Oct 24 12:45 plots
-rwxrwxr-x 1 zhengc domain users  306 Oct 24 12:45 prediction_detailed_results.txt
drwxrwxr-x 2 zhengc domain users  632 Oct 24 12:45 trait_predictions
```

Prediction with known test phenotype data

1. Input data:

   example/inputFile/train_genotype.vcf and example/inputFile/train_phenotype.txt

   example/inputFile/test_genotype.vcf and example/inputFile/test_phenotype.txt

2. Configure prediction with known test Phenotype data

Edit the configuration file (e.g., gs_parameters_example_prediction1.config) to define cross-validation parameters. Specify the GS_Mode, number of replicates, input genotype and phenotype file, and output directory.

As the first step, you can keep the same parameters as the file gs_parameters_example_prediction1.config. **But you MUST change the RScriptPath to point to your own Rscript.**

```
# R path
RScriptPath = your/Rscript
```

3. Run the pipeline script using the example configuration: