**Supplementary Note Data processing process of this research**

1. **Hyperspectral indices extraction**

The programs using for hyperspectral indices extraction of 533 *O. sativa* accessions

is same as that in the published article (<https://github.com/fenghuifh2006/Maize-RGB-CT-HSI-program>)

The acquired hyperspectral indices were organized and stored in Table S3

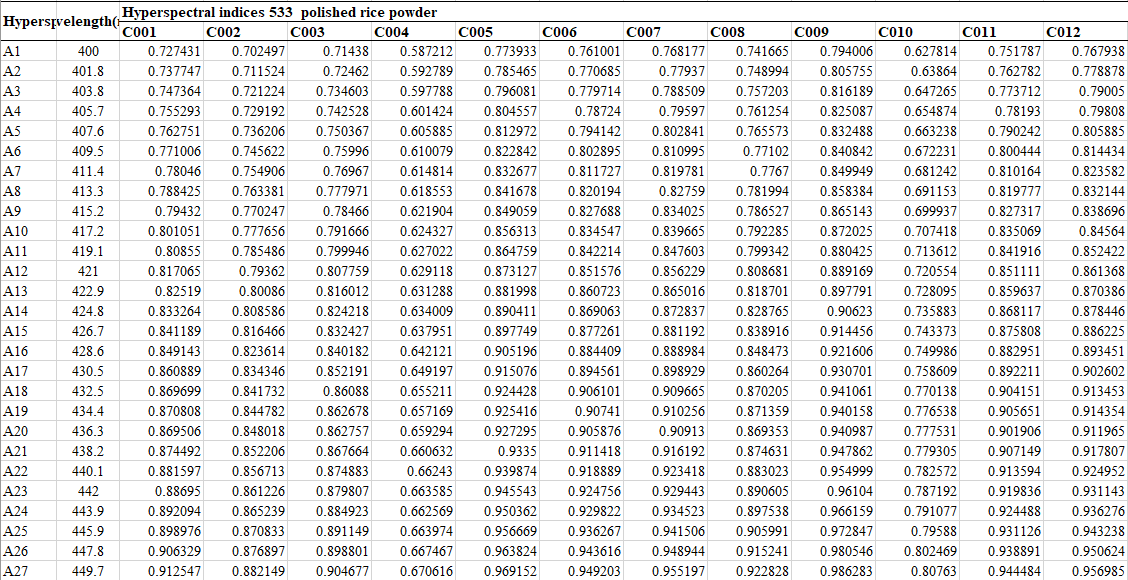


Figure 1 Hyperspectral indices of 533 rice accessions acquired in this study

1. **Metabolite levels extraction**

The metabolites data of 533 *O. sativa* accessions recorded were processed with LabSolutions 5.91 software. The data matrix was *log2* transformed.

The metabolites data were organized and stored in Table S5

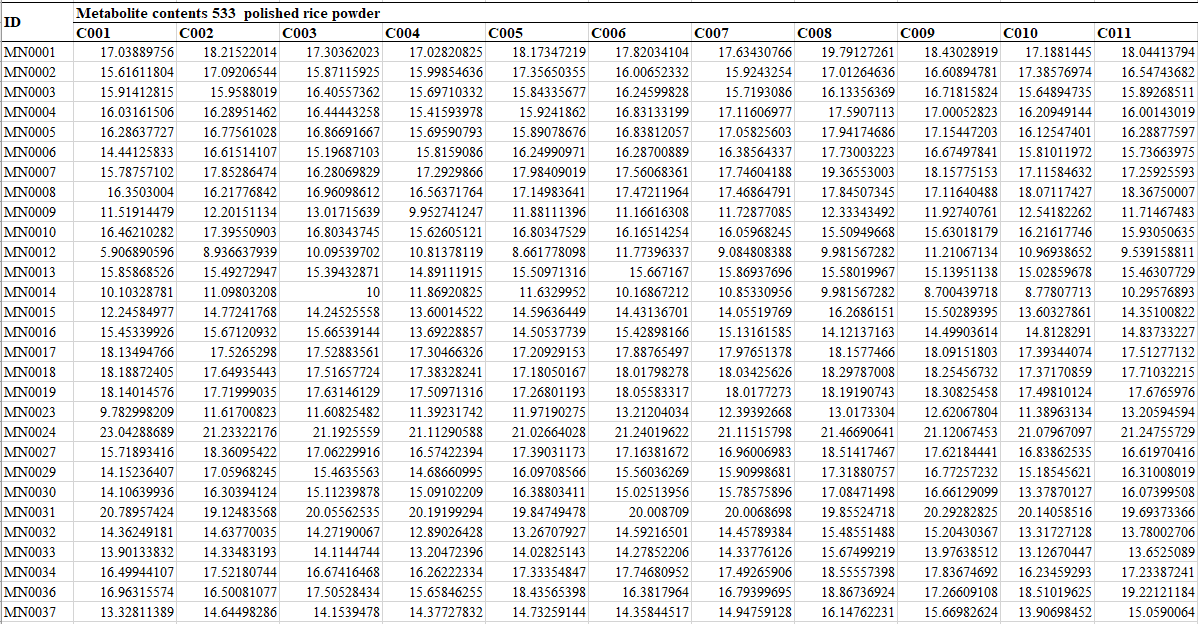


Figure 2 Metabolite levels of 533 rice accessions measured in the research

1. **Pearson correlation coefficient calculation between hyperspectral indices and metabolites**

The Pearson correlation coefficient calculation between hyperspectral indices and metabolites was calculated with the R script corr\_person.R using the built-in function ‘cor’ in the R4.2.1 environment . The method “pearson” was selected and the parameter “pairwise.complete.obs” was used for missing value processing.

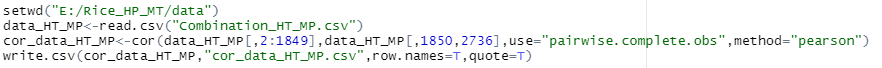


Figure 3 The code of calculating pearson correlation coefficient

The combination\_HT\_MP.csv is the combination of Table S3 and Table S5, The hyperspectral and metabolic data for same rice accession was listed in a row.

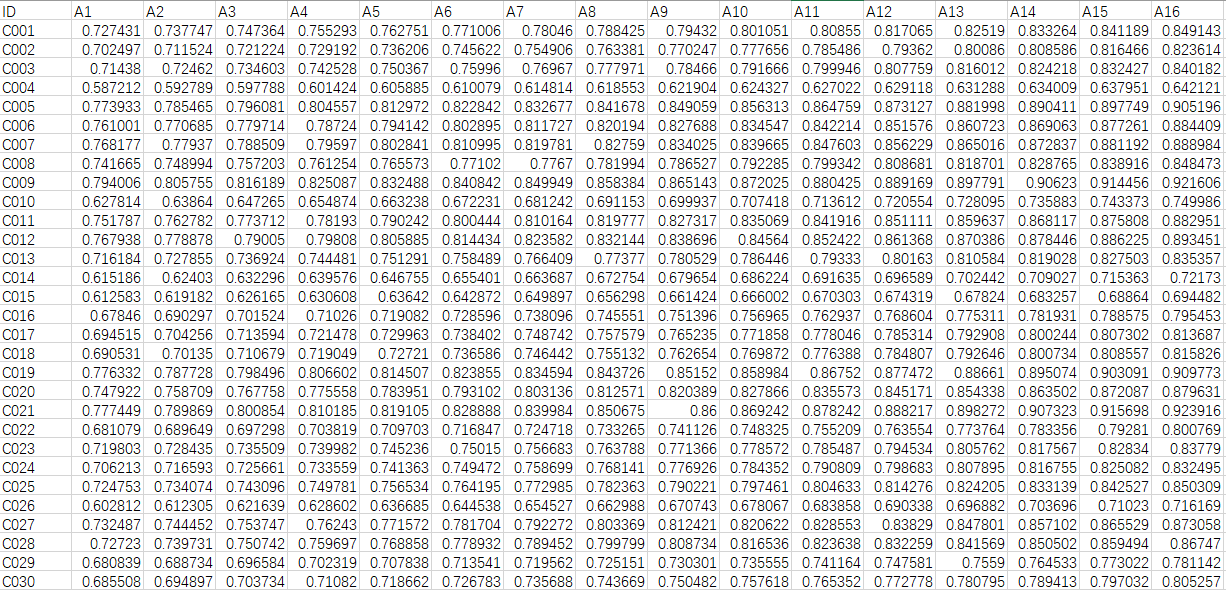


Figure 4 The content of combination\_HT\_MP.csv

A correlation matrix between hyperspectral indices and metabolites will be acquires after running the script corr\_person.R.

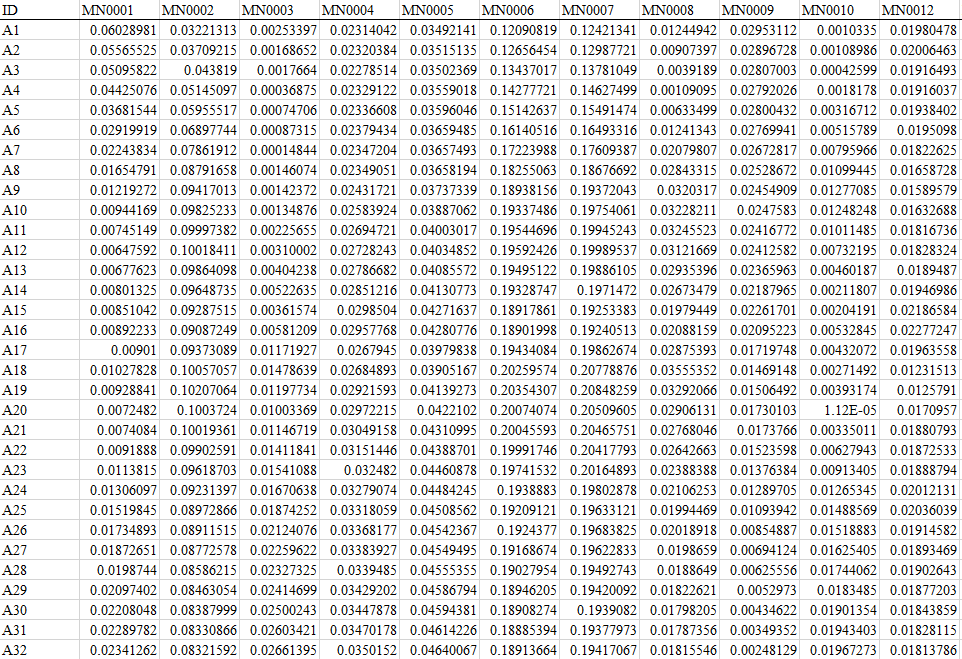


Figure 5 The correlation matrix between hyperspectral indices and metabolites.

1. **Eight machine learning models**

Eight machine methods including PLSR, LGBM, LASSO, RR, CNN, SVM, RF, SLR Regression were used for metabolites prediction. The codes of these methods were established in Python 3.8 environment based on sklearn module. The PLSR, LASSO, RR, SVM, RF, SLR regression models were integrated in the script ML\_regression.py. The LGBM regression models were calculated with the script LightGBM\_reg.py. The CNN regression models were calculated with the script CNNpytorch.py. The core codes for regression model construction were shown as follows

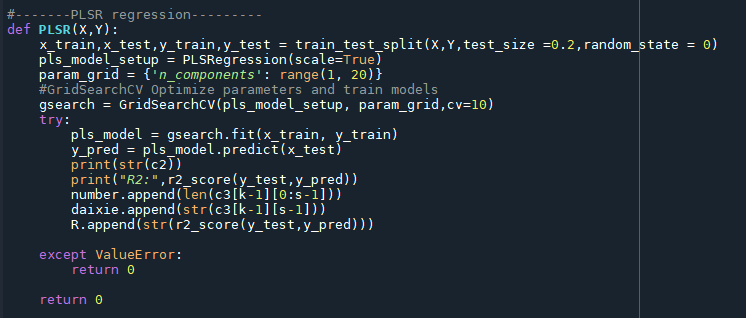


Figure 6 The function of PLSR regression

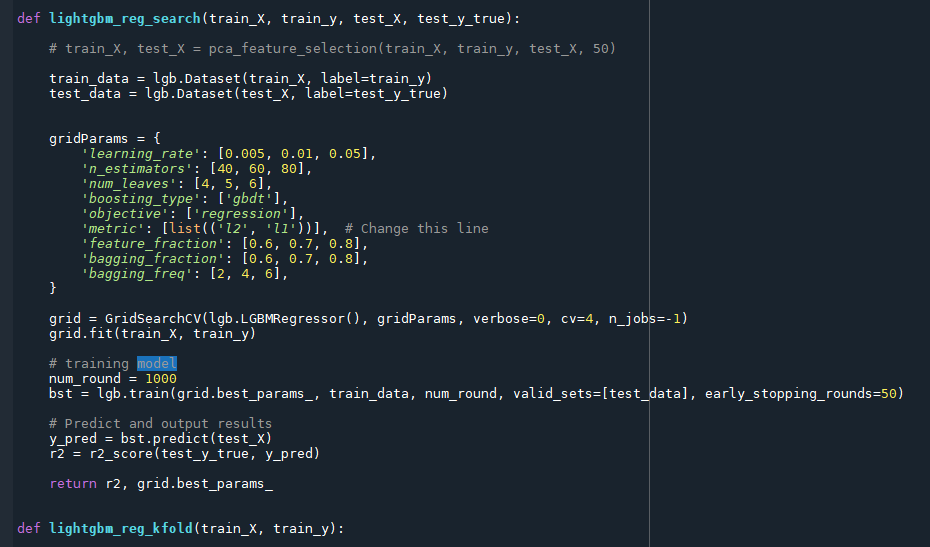


Figure 7 The function of LGBM regression

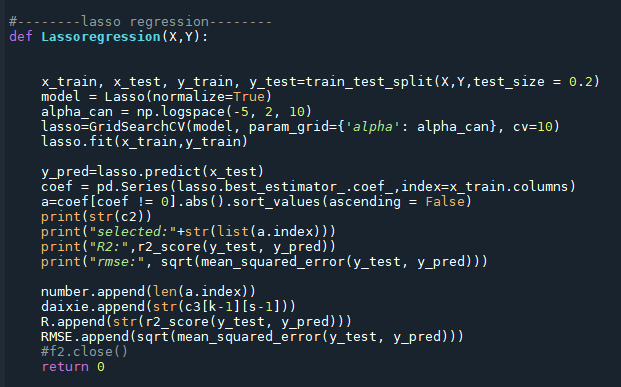


Figure 8 The function of LASSO regression

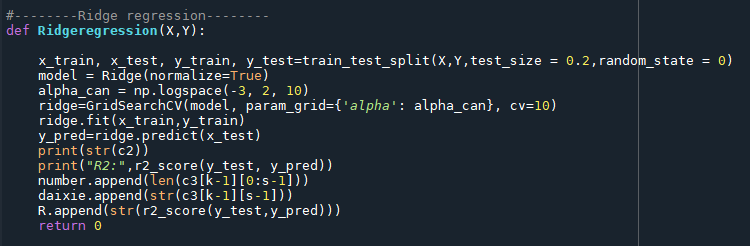


Figure 9 The function of RR regression

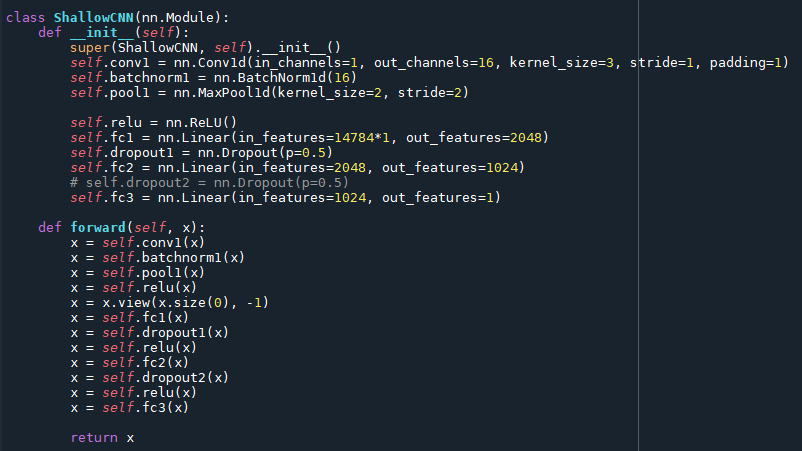


Figure 10 The function of CNN regression

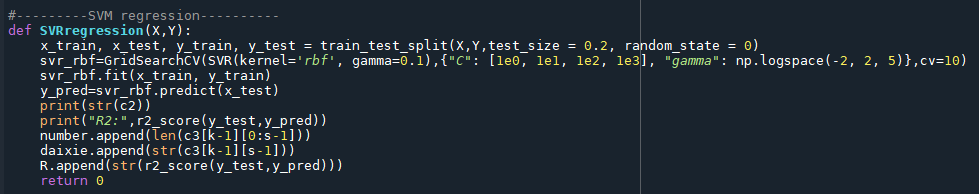


Figure 11 The function of SVM regression

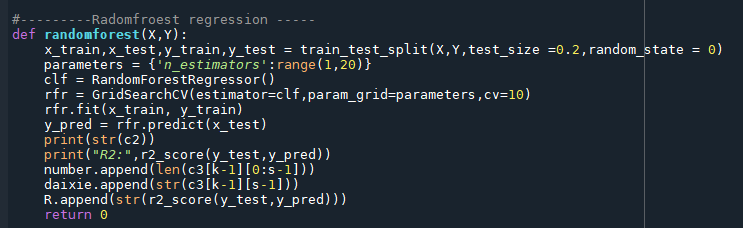


Figure 12 The function of RF regression

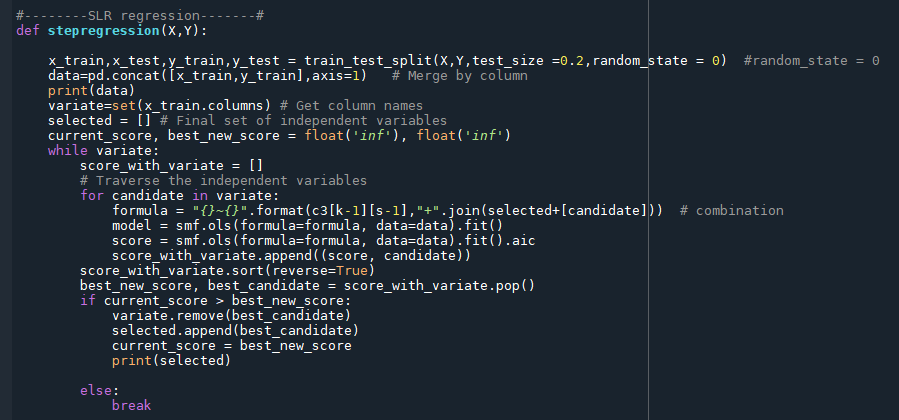


Figure 13 The function of SLR regression

The input data these scripts were the hyperspectral data listed in Table S3 served as X of the regression models and metabolite data listed in Table S5 served as Y of the regression models.



Figure 14 The input data of regression models

The training and test sets were derived randomly by dividing the data into a ratio of 1:4. The code to divide the training and test sets was shown as follows



Figure 15 Partition of the training and test sets

The R value and RMSE value of the regression models were reserved, all the details were listed in Table S7

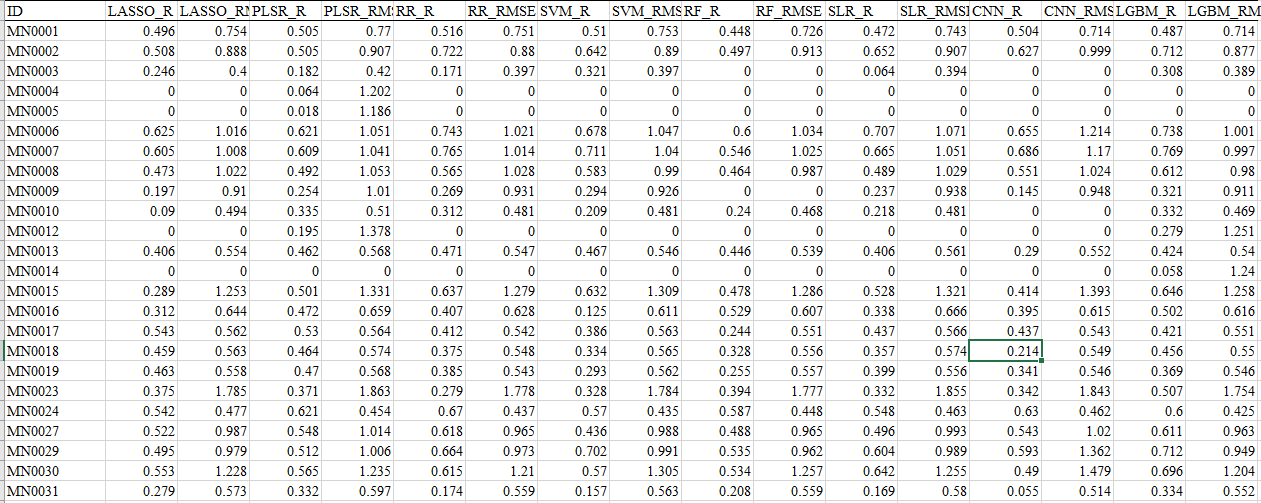


Figure 16 Model summary for eight machine learning methods.

1. **GWAS analysis**

The script for GWAS analysis and Manhattan maps drawing was listed in gwas\_script\_GB.txt shown as follows.

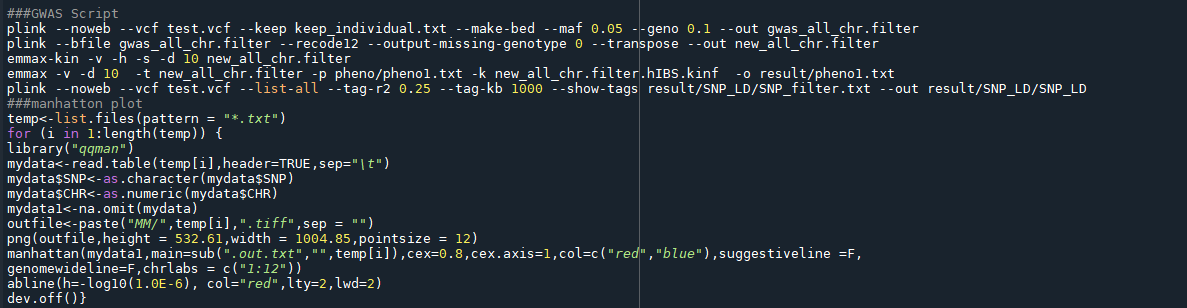


Figure 17 The script for GWAS analysis and Manhattan maps drawing

1. **Co-localization analysis**

Co-localization analysis was performed with the screened locus of hGWAS and mGWAS with the script co-localization.py according to the distance of the leading SNPs, while the threshold was 300kB within the chromosome.



Figure 18 The Python script of colocalization analysis

The co-localization results of hGWAS and mGWAS were reserved in Table S12

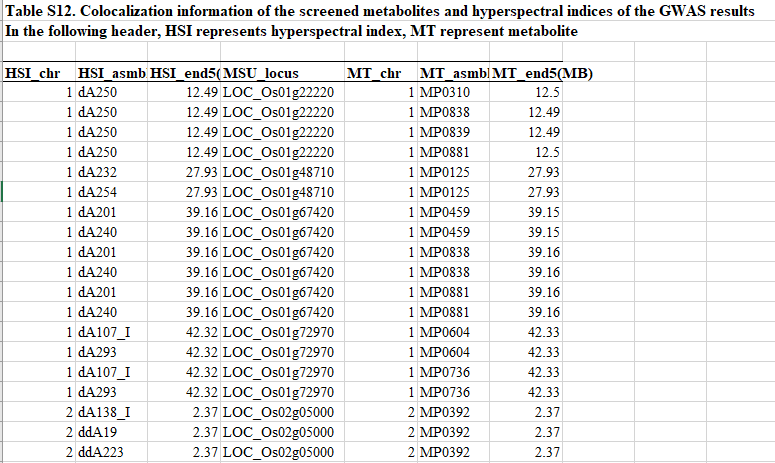


Figure 19 Colocalization information of the screened metabolites and hyperspectral indices of the GWAS results.

1. **Gene expression selection**

The co-localized locus with expression level higher than 200 in heading stage of rice panicles or 7-21 days after pollination of rice endosperms were reserved for further analysis. This process was handled by the script Gene\_expression\_selection.py shown as follows

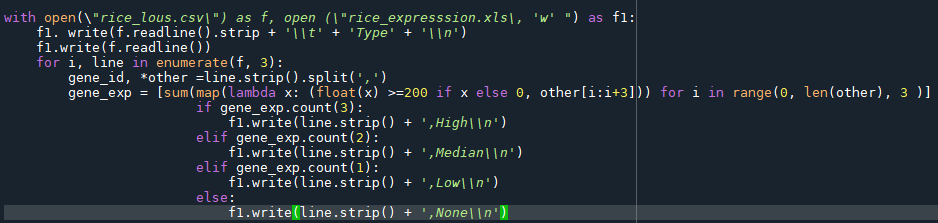


Figure 20 The Python script of Gene expression selection

1. **KEGG keywords mapping**

Several keywords were attached to each metabolite group, and the co-localized locus of corresponding metabolites having these keywords were screened. This process was handled by the script Kegg\_keywords.py shown as follows

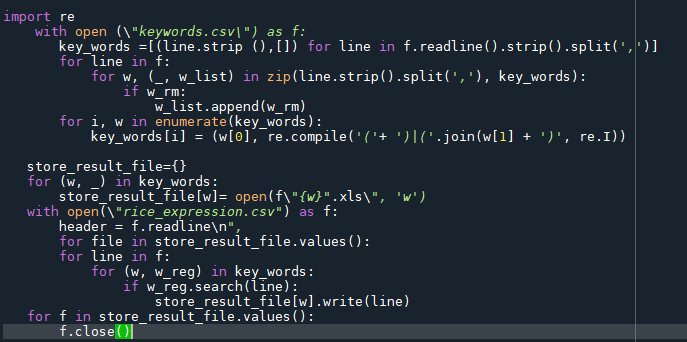


Figure 21 The Python script of KEGG keywords mapping

The keywords information of each metabolite group was shown in Table S13.

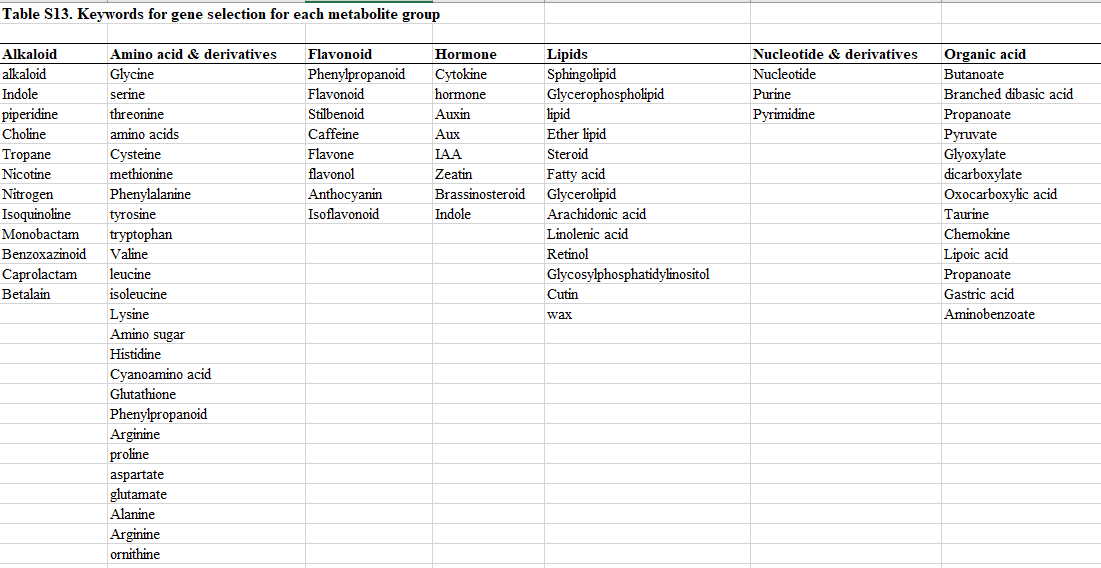


Figure 22 Keywords information for each metabolite group

Phenotypically and genetically related metabolites and hyperspectral indices could be screened after gene expression selection and KEGG keywords mapping, the details were listed in Table S15.

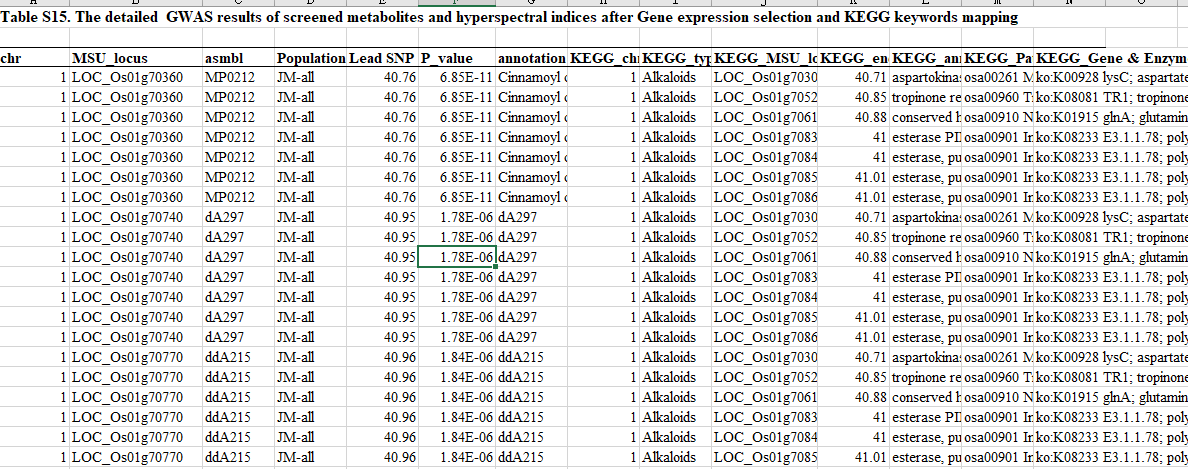


Figure 23 The detailed GWAS results of screened metabolites and hyperspectral indices after gene expression and KEGG keywords mapping.