User Manual for One-step GWAS

(Version 1)

2025.5.30

Jianjian Qi Laboratory

Inner Mongolia University

Support documents:

One-step GWAS is freely available to non-commercial users at JianjianQi-IMU/One-step-GWAS: Genome-wide association study.

The test data is sourced from academic papers (Zhao et al. *Nat. Commun*, 2011) and obtained from JianjianQi-IMU/One-step-GWAS: Genome-wide association study.

Citation:Multiple statistical methods are implemented in One-step GWAS. Citations of One-step GWAS very depending on methods used in the analysis:

Method	Method paper	Version 1
General Linear Model (GLM)	Xiang et al, 2012, *Nature Genetics*[1]	√
Mixed Linear Model(MLM)_GEMMA	Xiang et al, 2012, *Nature Genetics*[1]	√
Mixed Linear Model(MLM)_EMMAX	Hyun et al, 2010, *Nature Genetics*[2]	√
Structure	Jonathan et al, 2000, *Genetics*[3]	√
faststructure	Anil et al, 2014, *Genetics*[4]	√
Kinship	Xiang et al, 2012, *Nature Genetics*[1]	√

Note: These references are listed in section of Reference.

Table of Contents

1 INTRODUCTION

1.1 Why One-step GWAS?

One-step GWAS implemented a series of methods for Genome Wide Association (GWAS). The GWAS models include General Linear Model (GLM), Mixed Linear Model (MLM).

The One-step GWAS software integrates algorithms from various tools, including classic GWAS analysis software such as EMMAX and GEMMA, as well as population structure analysis tools like STRUCTURE and fastStructure. It primarily facilitates GWAS analysis and visualization functions on the Windows operating system, with interactive visualization capabilities tailored to meet the diverse needs of researchers. Additionally, it includes features such as principal component analysis, kinship analysis, population structure analysis. Moreover, it empowers researchers without a background in bioinformatics to perform analyses effortlessly, without the need to learn coding, simply by clicking the mouse.

1.2 Getting Started

1.2.1 Installtion

1. Windows x86 / x64

Download 'One-Step-GWAS.libopenblas.zip' from the website: https://github.com/JianjianQi-IMU/One-Step-GWAS/releases/tag/v1.0.0. After extracting the file, enter the folder—'One Step GWAS.exe' is the program executable.

1.2.2 Starting One-step GWAS

To run One-step GWAS, open the folder that holds the One-step GWAS files. Double click ‘One Step GWAS.exe’ will open the GUI of One-step GWAS as shown in Figure 1 below (Figures in the manual are produced in Windows 11):

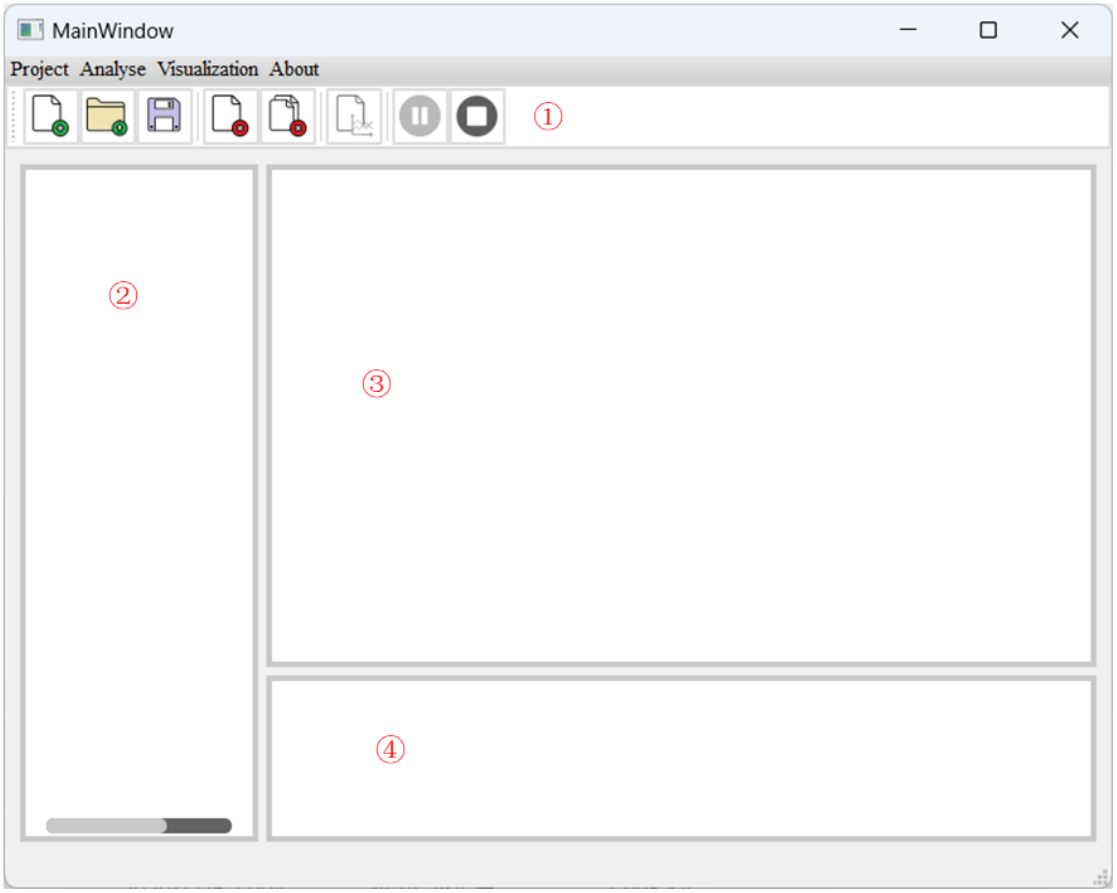

Figure 1 Graphical User Interface (GUI) of One-step GWAS.

The GUI shown above is scalable, i.e., the GUI size can be changed by dragging the borders of the screen, so that the GUI can use the full screen or part of the screen, allowing efficient and flexible use of the computer screen.

1. menu bar

The main menu in the menu bar includes functions for “Project”, “Analyse”, “Visualization” and “About”. The dropdown menu for “Project” includes “New project” and “Remove project”. Click “New project” to create a new project, Click “Remove project” to delete a project. The dropdown menu for “Analyse” includes GWAS analysis (GLM, GEMMA, EMMAX), PCA, population structure (STRUCTURE, FastStructure), Kinship analysis. The dropdown menu for “Visualization” includes the visualization function of GWAS, PCA, QQ plot, Structure and Kinship. The submenu includes “New project”, “load file”, “save result”, “remove”, “remove all project”, “result visualization”, “pause”, “stop running”, the commonly used function buttons.

2. project bar

When the data is imported, in this section, you can see the imported file names and paths.

3. running bar

This section is mainly used for analysis.

4. status bar

The status of project analysis is displayed in this area including running time, genotype and phenotype number etc.

1.3 How to use the One-step GWAS user manual

The manual is structured into seven sections. It begins with an introductory overview of the basic information of One-step GWAS in section 1. Sections 2 to 5 delve into the software's format, analysis functions, and visualization functions. Section 6 is dedicated to elucidating the output results, while Section 7 provides comprehensive guidance for testing the software using sample data. Lastly, the final section compiles all the references cited within the manual. Prior to delving into the intricacies of Sections 2 to 6, it is highly recommended to proceed directly to Section 7 for tutorial instructions.

2 INPUT DATA

2.1 Genotype Data

2.1.1 Diploid Format

1.The binary file format of PLINK (*.bed, *.bim) can be found on the PLINK official website (https://www.cog-genomics.org/plink/1.9/formats) .

2.VCF file format

2.1.2 Polyploid Format (with Tetraploid as an example)

Simplified Variants File(*.pped, *.bim)

1. pped file (Rows: SNP locus information; Columns: Sample information)

（The missing values represented as “NA”, The file is delimited by tab characters.）

0001	0000	0001	0001	0000
1111	0111	0011	0111	0011
1111	0111	0011	0111	0011
0111	0111	0111	0111	0111
1111	0111	0011	0111	0011
0111	0111	0111	0111	0111
0001	0000	0001	0011	0000
0111	1111	0111	0111	1111
1111	1111	1111	1111	1111
0000	0000	0000	0000	0000

Notice: If the file is in numeric format (representing the dosage of the minor alleles), then the correspondence is as follows:

0	0000
1	0001
2	0011
3	0111
4	1111

Notice: If the file is a bi-allelic SNP genotypes file, then the correspondence is as follows:

nulliplex(AAAA)	0000
simplex(AAAB)	0001
duplex(AABB)	0011
triplex(ABBB)	0111
quandriplex(BBBB)	1111

2. *.bim file (Columns: Chromosome number SNP identifier Genetic distance Physical position Allele 1 Allele 2)

1	c1_10000	0	1001	A	T
2	c1_10001	0	1221	A	T
3	c1_10010	0	1688	A	T
3	c1_10011	0	1754	A	T
4	c1_10012	0	1757	A	T
5	c1_10013	0	1822	A	T
6	c1_10014	0	1891	A	T
7	c1_10031	0	1896	A	T

Notice: If the Genetic distance is unknown, it can be indicated as 0;

2.2 Phenotype Data

Phenotype file, with only one column representing phenotypes. The order of phenotypes should match the order of samples in the genotype file, with missing values represented as “NA”.

0.926954733
1.078313253
1.016129032
0.810185185

2.3 Covariate Variables Data for GWAS

covariance file（The missing values represented as “NA”, The file is delimited by tab characters.）.

1	-0.0486	0.003	0.0752	-0.0076
1	0.0672	-0.0733	0.0094	-0.0005
1	0.0544	0.0681	-0.0062	-0.0369
1	-0.0073	0.0224	-0.0121	0.2602
1	0.0509	0.0655	-0.0058	-0.0378
1	-0.0293	-0.0027	-0.0677	-0.0085
1	-0.0333	-0.0039	-0.0675	-0.0008
1	-0.0399	0.0042	0.0435	0.0145

Notice: The first column must be 1. The other columns are: Q1, Q2, Q3…

2.4 PCA Data for Visualization

-0.0486	0.003	0.0752
0.0672	-0.0733	0.0094
0.0544	0.0681	-0.0062

Notice: The other columns are: Q1, Q2, Q3…, The file is delimited by tab characters.

2.5 Population Structure Data

0.000005	0.000005	0.999991
0.000006	0.99999	0.000005
0.878083	0.121912	0.000006
0.340348	0.000677	0.658975
0.880208	0.119786	0.000006
0.037514	0.04137	0.921116
0.000006	0.015215	0.984779

Notice: The other columns are: Q1, Q2, Q3…, The file is delimited by tab characters.

2.6 Kinship Data

kinship matrix file

-0.041562	0.028631	0.095294	-0.002573
0.015633	-0.035754	-0.030985	0.003589
0.028069	-0.041809	-0.033899	0.004297
0.022275	-0.041096	-0.033736	0.002053

Notice: The kinship matrix format is the same for GWAS analysis and visualization, The file is delimited by tab characters.

2.7 Chromosome Length File

Format: Chromosome number Chromosome length information

1	159038749
2	72943711
3	92503428
4	82398023
5	80503876
6	87043187
7	58785265
8	26434011
9	141712163
10	116138521
11	93888508

Notice: The order of chromosomes in this file determines the order of chromosomes in the visualization graph, The file is delimited by tab characters.

2.8 Genome Annotation File

1.GFF (General Feature Format) File

2.TSV (Tab-Separated Values) File Format (gene name Chromosome number Gene start position Gene end position Gene orientation information)

Soltu.DM.09G000160.1	9	61988	67895	-
Soltu.DM.09G000170.1	9	74474	75057	-
Soltu.DM.09G000180.1	9	76078	76779	+
Soltu.DM.09G000190.1	9	77149	78874	-
Soltu.DM.09G000200.1	9	85525	86595	-
Soltu.DM.09G000210.1	9	87346	88047	+
Soltu.DM.09G000220.1	9	108592	121632	-
Soltu.DM.09G000230.1	9	124664	125751	-
Soltu.DM.09G000240.1	9	135661	142242	+
Soltu.DM.09G000250.1	9	142199	142852	-
Soltu.DM.09G000260.1	9	145037	149809	+

2.9 Fam File Format

The binary file format of PLINK (*.fam) can be found on the PLINK official website (https://www.cog-genomics.org/plink/1.9/formats) .

2.10 Genome Sequence File

Fasta format

3 EDIT FILE

If you want to change a certain value after inputting the files, you can do it by right-clicking, as shown in
Figure 2. Then, you can edit the phenotype, covariate, and kinship relationship values.

Right click Edit Phenotype Data / Covariate Data / Kinship Data

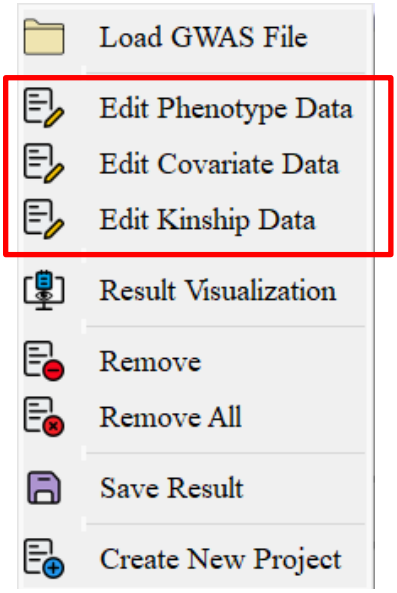

Figure 2 Edit Tab

4 ANALYSIS

4.1 PCA Analysis

The steps of Principal Component Analysis (PCA) are shown in Figure 3.

1. Create New Project (two methods):
   (1) Open the software > **Right click** > **Create New Project**.
   (2) Navigate to the "**Project**" function: **Project** > **New Project**.
2. Input Data:
   Right-click on the created project > **Load GWAS File** > Input **Genotype File** > Input **Phenotype File.**
3. Analysis:
   Navigate to the "**Analyse**" function: **Analyse** > **PCA**.
4. Save Result:
   When the analysis is completed, **Right-click** on the created project > **Save Result**.

Figure 3 shows the process of PCA analysis

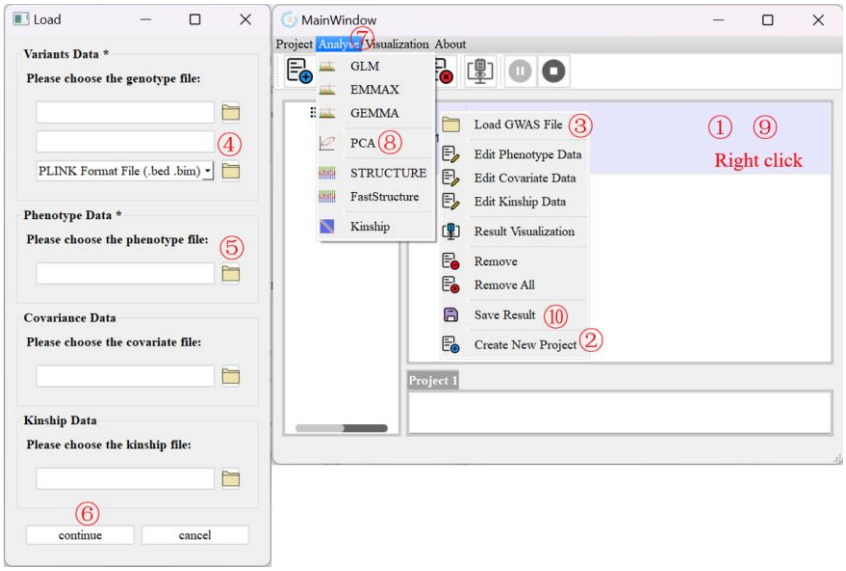

Figure 3 The process of PCA analysis

4.2 Structure Analysis

The steps for population structure analysis are shown in Figure 4.

1. Create New Project (two methods):

   (1) Open the software > **Right click** > **Create New Project**.

   (2) Navigate to the "**Project**" function: **Project > New Project.**

2. Input Data:

   Right-click on the created project > **Load GWAS File** > Input **Genotype File** > Input **Phenotype File.**

3. Analysis:

   Navigate to the "**Analyse**" function: **Analyse** > **STRUCTURE/ FastStructure**.

4. Save Result:

   When the analysis is completed, **Right-click** on the created project > **Save Result**.

Figure 4 shows the process of structure analysis

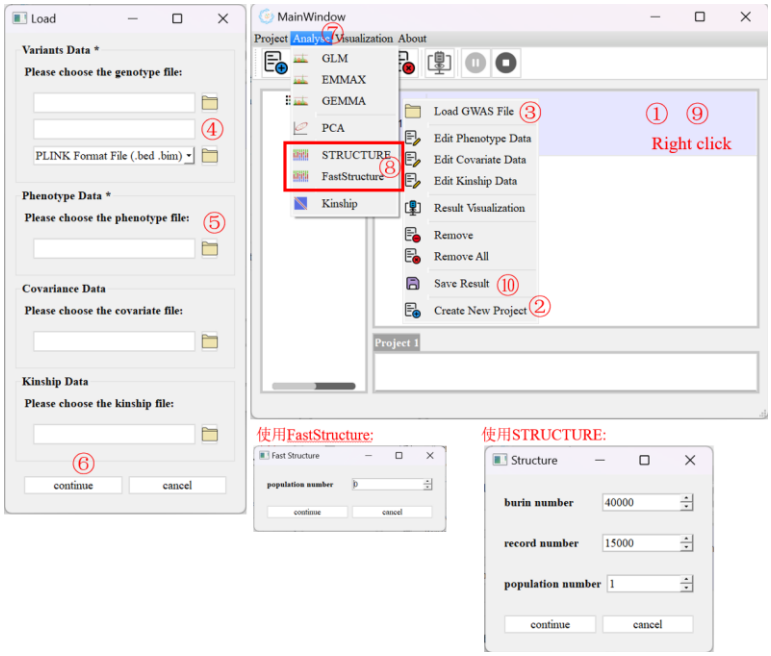

Figure 4 The process of Structure analysis

4.3 Kinship Analysis

The steps for the kinship analysis matrix are shown in Figure 5.

1. Create New Project (two methods):

   (1) Open the software > **Right click** > **Create New Project**.

   (2) Navigate to the "**Project**" function: **Project > New Project.**

2. Input Data:

   Right-click on the created project > **Load GWAS File** > Input **Genotype File** > Input **Phenotype File.**

3. Analysis:

   Navigate to the "**Analyse**" function: **Analyse** > **Kinship**.

4. Save Result:

   When the analysis is completed, **Right-click** on the created project > **Save Result**.

Figure 5 shows the process of kinship analysis

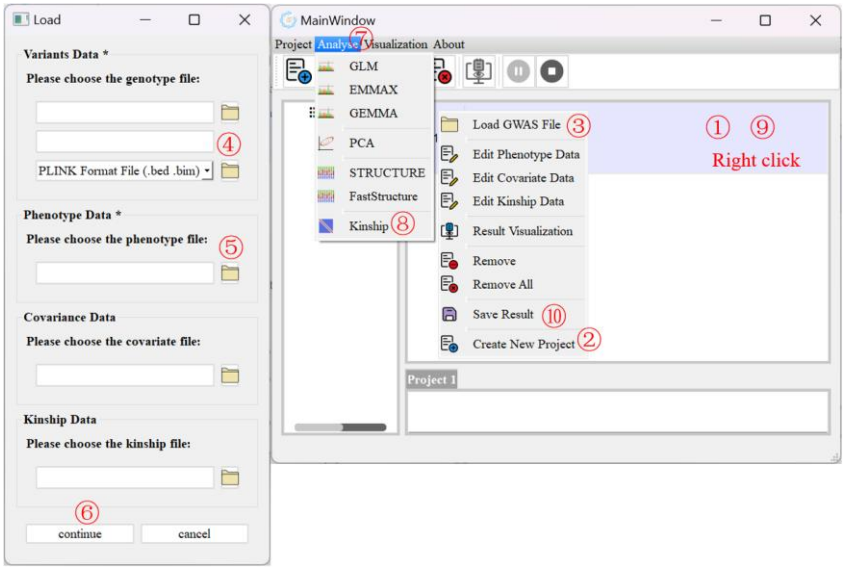

Figure 5 The process of kinship analysis

4.4 GWAS Analysis

In this software, available models for GWAS analysis include GLM, EMMAX, and GEMMA, with EMMAX and GEMMA being MLM models derived from classical GWAS analysis software. The steps for conducting GWAS analysis are shown in Figure 6.

1. Create New Project (two methods):
   (1) Open the software > **Right click** > **Create New Project**.
   (2) Navigate to the "**Project**" function: **Project > New Project.**
2. Input Data:
   (1) GLM:
   Right-click on the created project > **Load GWAS File** > Input **Genotype File** > Input **Phenotype File > (**Input **covariate File).**
   (2) GEMMA/EMMAX:
   Right-click on the created project > **Load GWAS File** > Input **Genotype File** > Input **Phenotype File > > Input covariate File >** Input **Kinship File.**
3. Analysis:
   (1) GLM:
   Navigate to the "**Analyse**" function: **Analyse** > **GLM**.
   (2) GEMMA/EMMAX:
   Navigate to the "**Analyse**" function: **Analyse** > **GEMMA/EMMAX**.

   Notice: Under polyploid analysis conditions, you can also select the desired model to use within GLM, EMMAX, or GEMMA.
4. Save Result:
   When the analysis is completed, **Right-click** on the created project > **Save Result**.

Figure 6 shows the process of GWAS analysis

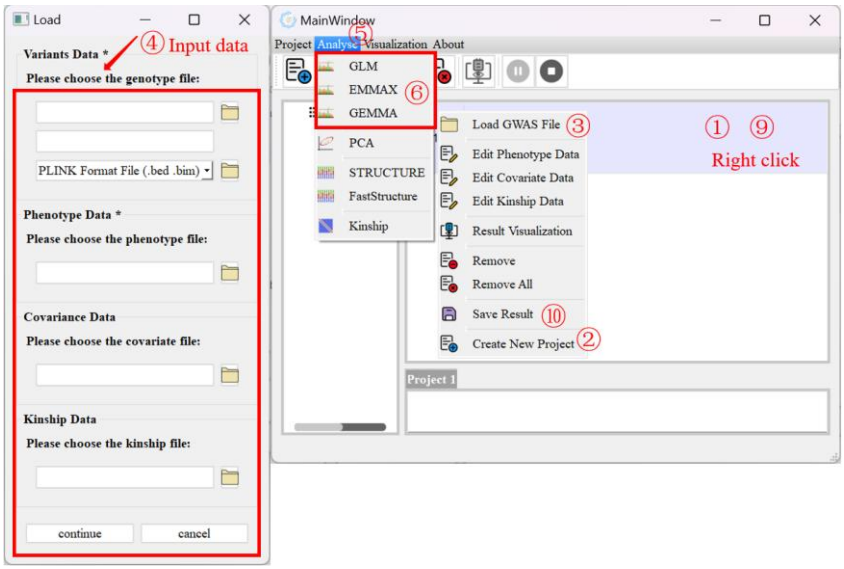

Figure 6 The process of structure analysis

5 VISUALIZATION

5.1 PCA Visualization

The PCA visualization process is shown in Figure 7 and Figure 8

There are two methods to visualize Principal Component Analysis results:

1. Navigate to the "**Visualization**" function: **Visualization** > **PCA Visualization** > Input ***.fam File** > Input **PCA Output File** > Choose **2D** or **3D**.

Figure 7 shows the process 1 of PCA visualization

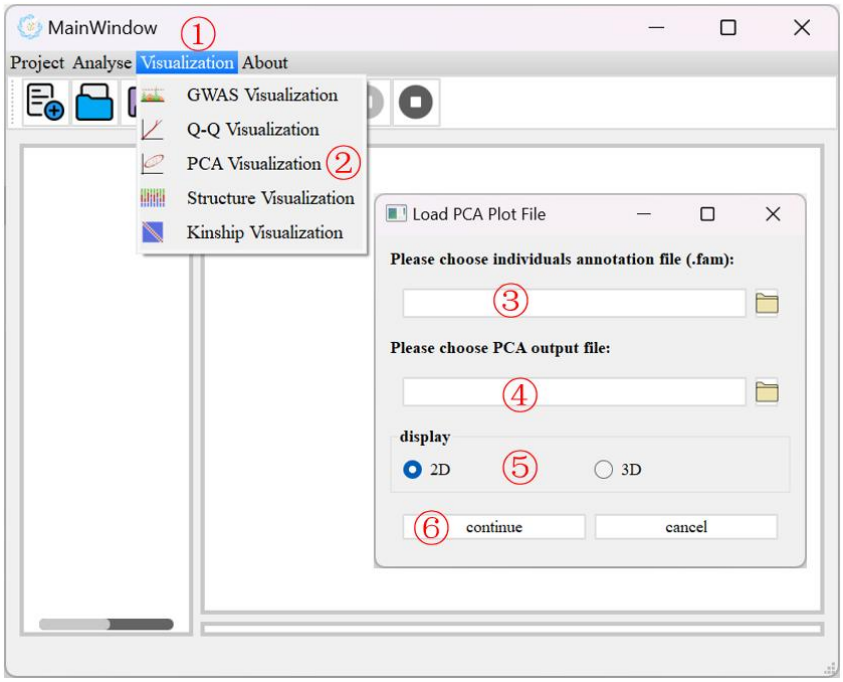

Figure 7 The process 1 of PCA visualization

2. After the PCA analysis is completed, you can **right-click** > select **Result Visualization** > Input ***.fam File** > choose **2D** or **3D**.

Figure 8 shows the process 2 of PCA visualization

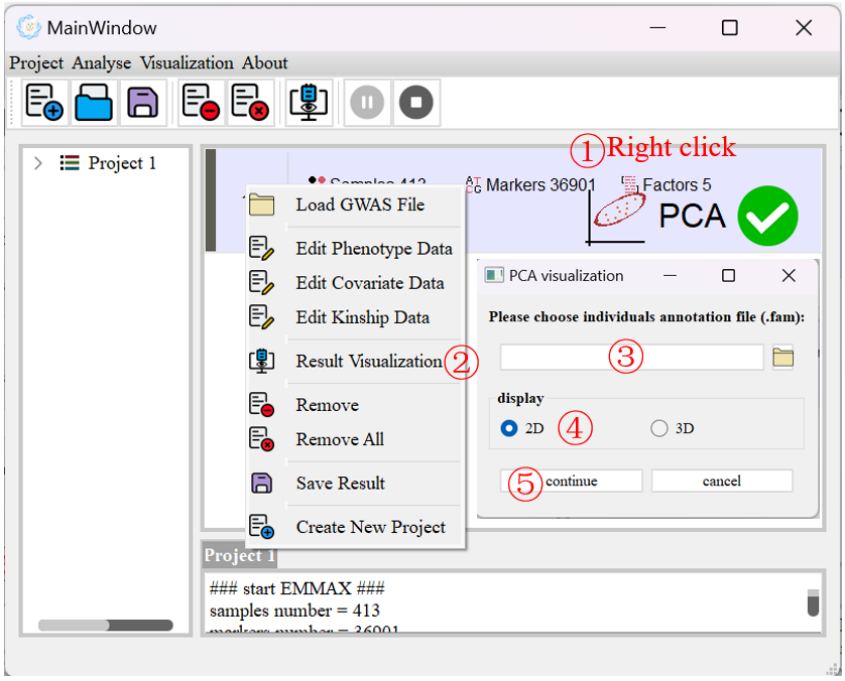

Figure 8 The process 2 of PCA visualization

5.2 Population Structure Visualization

The visualization steps for population structure are shown in Figure 9 and Figure 10.

There are two methods to visualize the results of population structure.

1. Navigate to the "**Visualization**" function: **Visualization** > **Structure Visualization** > Input ***.fam File** > Input **Structure Matrix File**.

Figure 9 shows the process 1 of structure visualization

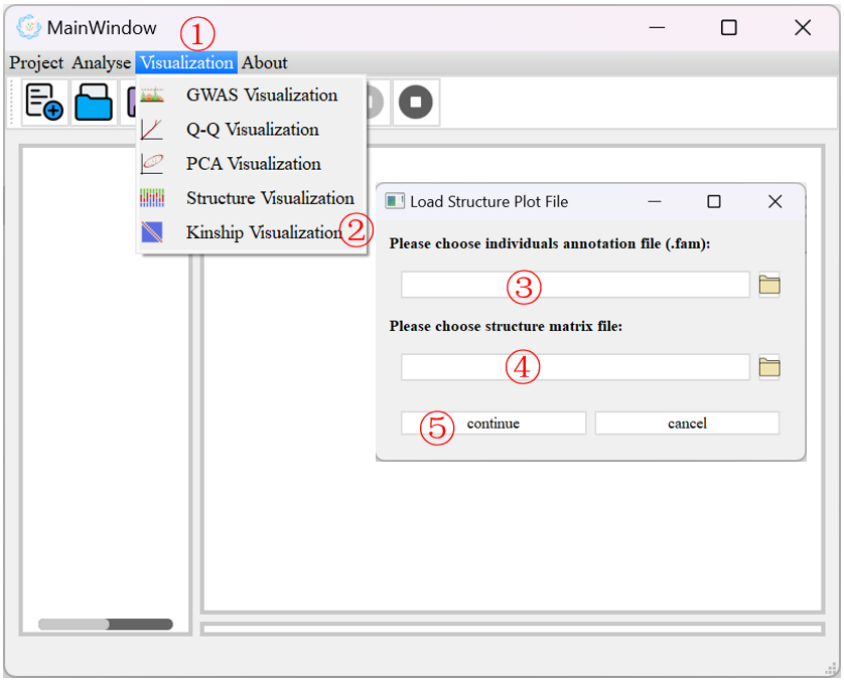

Figure 9 The process 1 of structure visualization

2. After the population structure analysis is completed, you can **right-click** > select **Result Visualization** > Input ***.fam File**.

Figure 10 shows the process 2 of PCA visualization

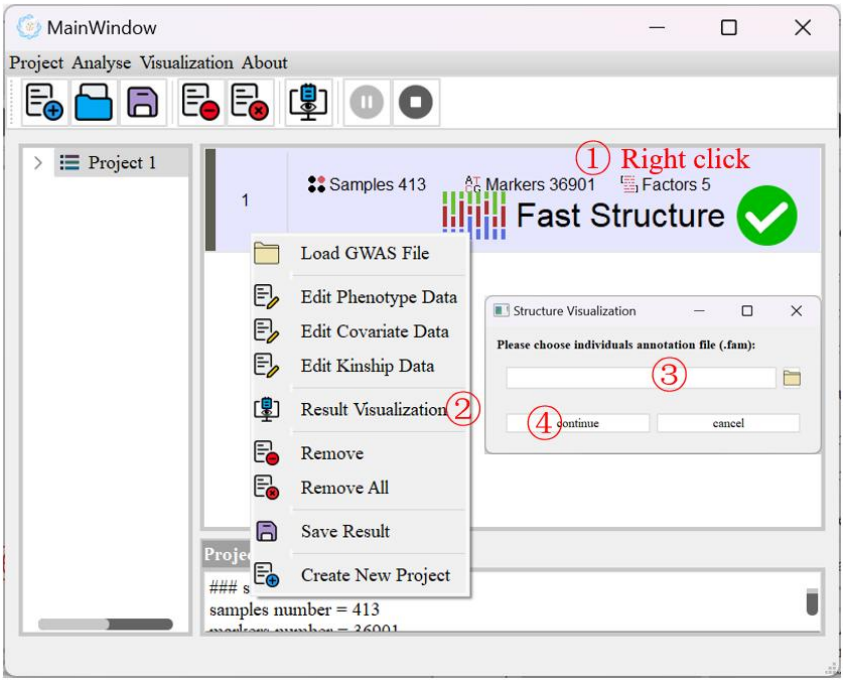

Figure 10 The process 2 of structure visualization

5.3 Kinship Visualization

The visualization steps for the kinship matrix are shown in Figure 11 and Figure 12.

There are two methods to visualize kinship relationships.

1. Navigate to the "**Visualization**" function: **Visualization** > **Kinship Visualization** > Input ***.fam File** > Input **Kinship Matrix File**.

Figure 11 shows the process 1 of kinship visualization

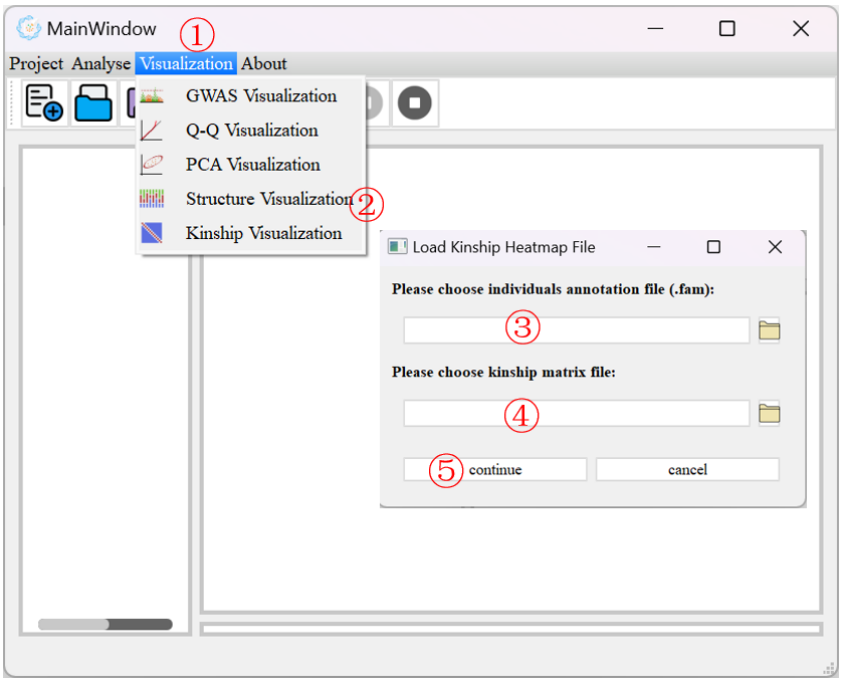

Figure 11 the process 1 of kinship visualization

2. After the Kinship analysis is completed, you can **right-click** > select **Result Visualization** > input the ***.fam file**.

Figure 12 shows the process 2 of kinship visualization

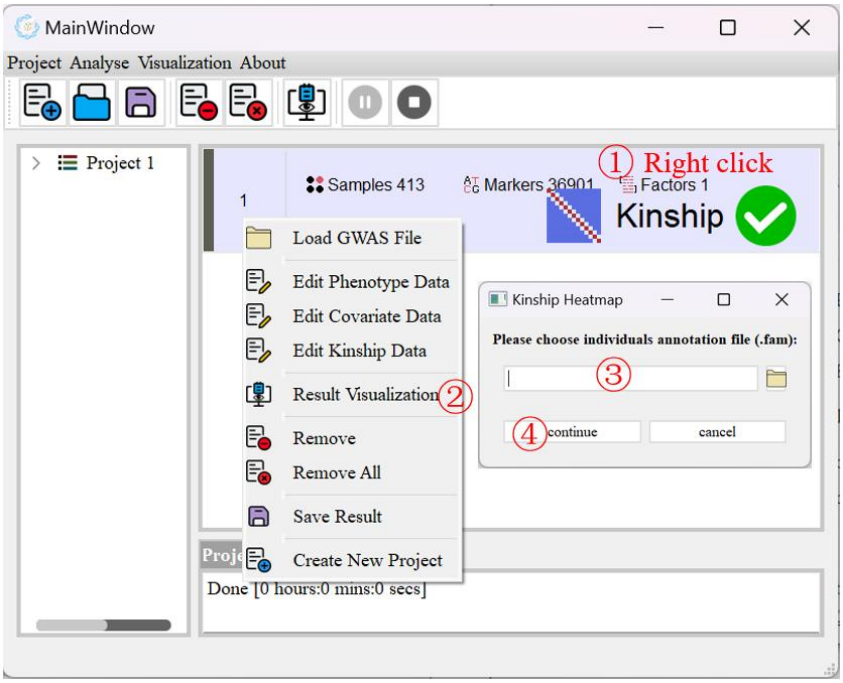

Figure 12 The process 2 of kinship visualization

5.4 Manhattan Plot

Display Manhattan plots after GLM, EMMAX, and GEMMA analyses.

The visualization steps for the Manhattan plot sre shown in Figure 13 and Figure 14.

There are two methods for visualization:

1. Navigate to the "**Visualization**" function: **Visualization** > **GWAS Visualization** > Input **Chromosome Length File** > Input **GWAS Output File** > Input ***.bim File** > choose **P value column** > Input **Genome Annotation File**.

Figure 13 shows the process 1 of Manhattan plot

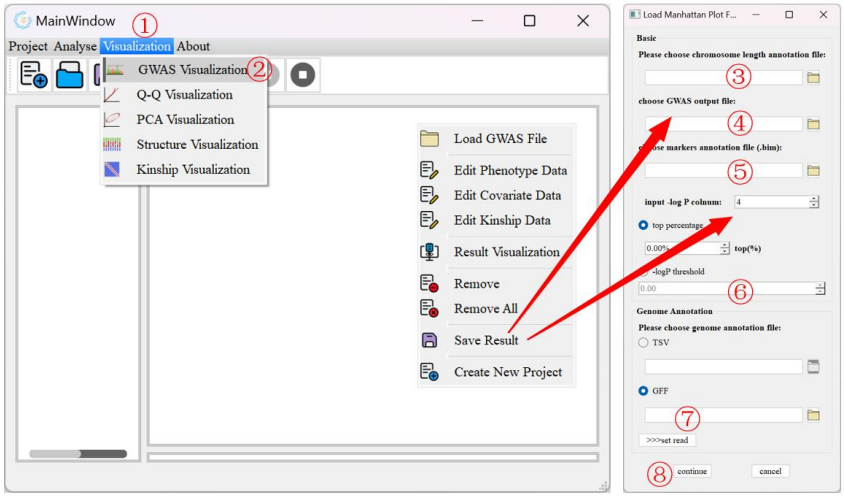

Figure 13 the process 1 of Manhattan plot

2. After the GWAS analysis is completed, you can **right-click** > select **Result Visualization** > input **the *.fam File** > Input **Genome Annotation File**.

Figure 14 shows the process 2 of Manhattan plot

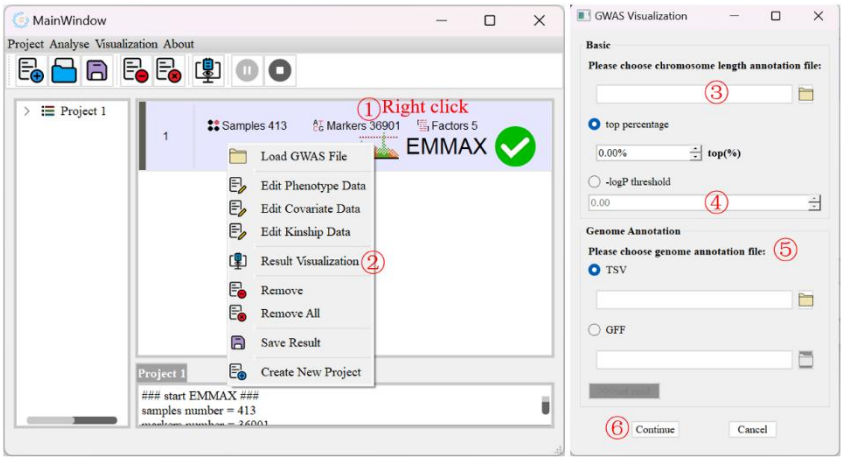

Figure 14 The process 2 of Manhattan plot

5.5 Q-Q Plot Visualization

The visualization process for the Q-Q plot is shown in Figure 15.

Navigate to the "**Visualization**" function: **Visualization > Q-Q Visualization >** Input **Chromosome Length File** > Input **GWAS Output File** > Input ***.bim file** > choose **P value column.**

Figure 15 shows the process of Q-Q plot

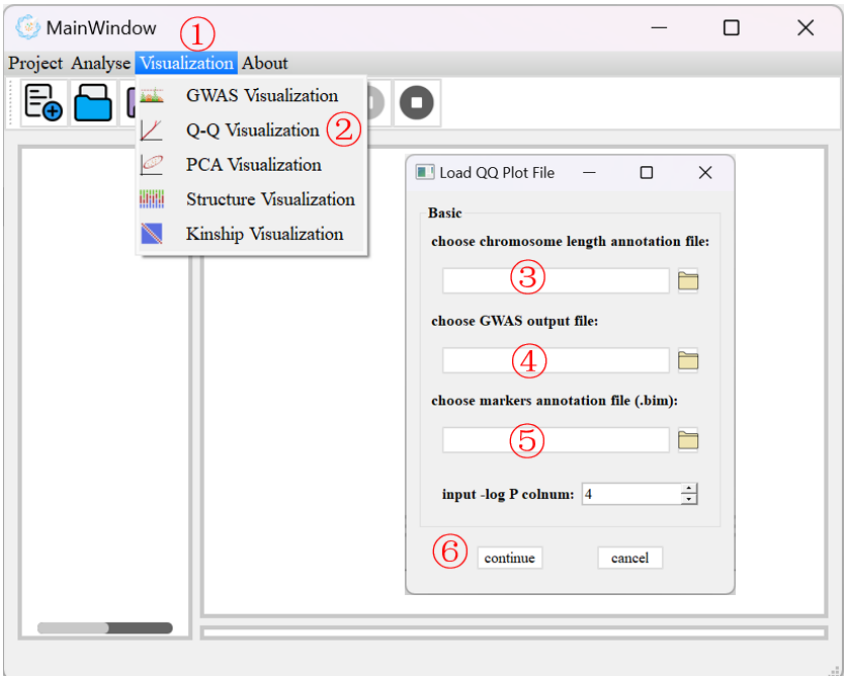

Figure 15 The process of Q-Q plot

6 OUTPUT RESULTS

6.1 Analysis Results

6.1.1 GWAS Output File

The results of the GWAS output file are shown in the following figure.

id1000001	-5.132093e-02	-2.685751e+00	7.588600e-03
id1000003	-4.892445e-02	-2.552283e+00	1.113478e-02
id1000005	-3.159476e-02	-1.088232e+00	2.772569e-01
id1000007	-4.016547e-02	-2.090504e+00	3.730832e-02
id1000008	-4.705157e-02	-2.478231e+00	1.368417e-02
id1000011	-2.025820e-02	-9.506943e-01	3.424290e-01
id1000013	-1.502854e-02	-5.066655e-01	6.127151e-01
id1000015	-4.821420e-02	-2.543575e+00	1.141081e-02
id1000016	-4.494457e-02	-2.446562e+00	1.492408e-02
id1000020	-3.619757e-02	-1.909484e+00	5.703369e-02
id1000024	-2.539926e-02	-1.117393e+00	2.646083e-01

SNP id

P value

Notice: If you have GWAS results files generated by other software, you only need to place the SNP identifiers in the first column. The file also includes p-values, which can be placed in any column.

6.1.2 Other Output File

PCA result file:

0.055	0.0987	0.0005	-0.0161
-0.0406	-0.0061	-0.0584	-0.0244
-0.0469	0.0014	0.0283	-0.0135
0.002	0.0349	-0.0216	0.0887
0.0621	0.1142	0.0031	-0.0277
0.0532	0.0958	0.0017	-0.0148
-0.0313	-0.0061	-0.0689	-0.0175
0.0683	-0.078	0.0099	-0.0084
0.0605	-0.0652	0.0107	-0.0076
-0.0343	-0.0042	-0.06	0.0085

Structure analysis output result and kinship matrix analysis output results format is identical to PCA visualization analysis)

6.2 Visualization Results

6.2.1 PCA plot

In the PCA visualization interface, "**File**" is used to save the generated PCA file. There are two saving options: SVG and PNG, and you can also choose the file size. When saving, it captures the displayed image on the page, meaning that if the image is zoomed in, the saved image will also be enlarged. "**Settings**" is used to adjust the image, including gridlines, axes, etc. "**Color**" is used to adjust colors. The PCA results are shown in Figure 16:

Figure 16 shows PCA plot

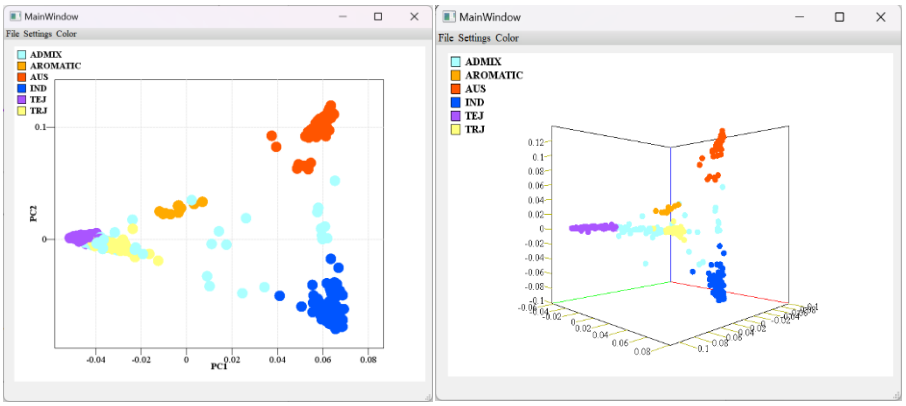

Figure 16 PCA plot

6.2.2 Structure plot

In the population structure visualization interface, "**File**" can be used to save the generated PCA results. There are two file formats available for saving: SVG and PNG, and you can select the desired file format to save. When saving, it captures the image displayed on the page, which means if the image is enlarged, the saved image will also be enlarged. "**Settings**" can be used to make some adjustments to the image, including sample sorting. "**Color**" allows for color modifications. In the left menu bar, you can make detailed adjustments to the image, including the spacing between samples, the width and height of the image, and so on. The population structure diagram is shown in Figure 17:

Figure 17 shows population structure plot

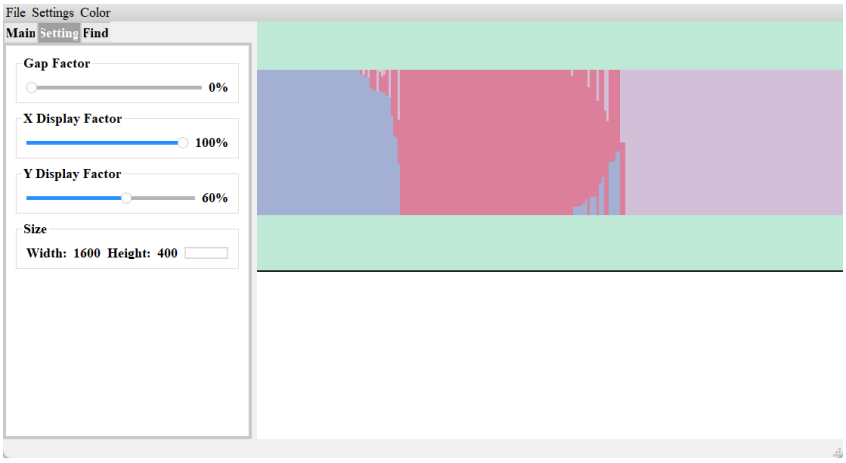

Figure 17 population structure plot

6.2.3 Kinship plot

In the kinship matrix visualization interface, "**File**" can be used to save the generated PCA results. There are two file formats available for saving: SVG and PNG, and you can choose the desired format to save. When saving, it captures the image displayed on the page, which means if the image is enlarged, the saved image will also be enlarged. "**Settings**" is used for clustering and modifying colors. The kinship matrix diagram is shown in Figure 18:

Figure 18 Kinship plot

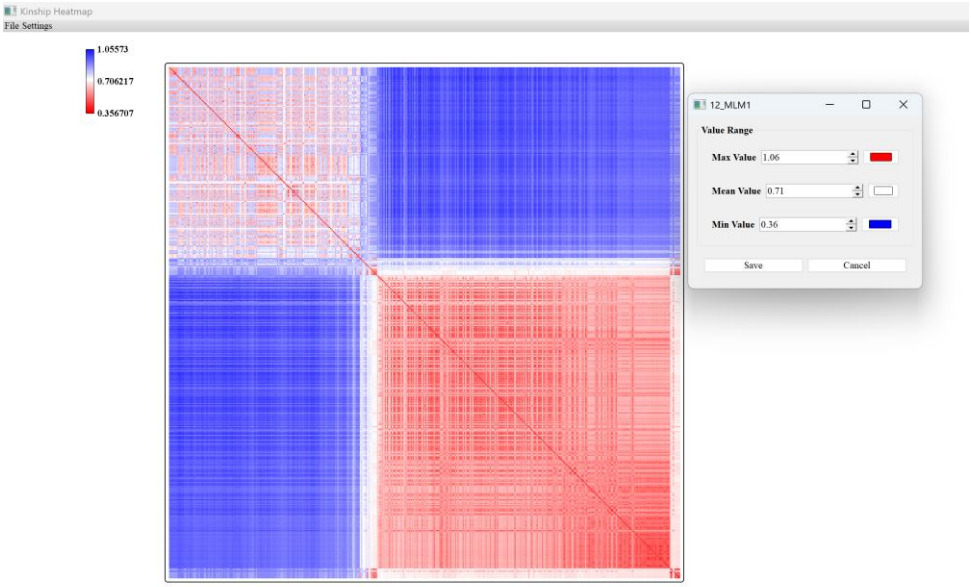

Figure 18 Kinship plot

6.2.4 Manhattan plot

In the Manhattan plot visualization interface, “**File**” is used to load genomic sequence file and save images. There are two file formats available for saving: SVG and PNG, and you can select the desired format to save. When saving, it captures the image displayed on the page, which means if the image is enlarged, the saved image will also be enlarged. “**Settings**” is used to adjust the image, including whether to display grid lines, axes, etc., as well as to set threshold lines. “**Color**” is used to adjust the colors. On the Manhattan plot, you can right-click on an SNP locus and select "Label" to annotate it. The Manhattan plot is shown in Figure 19:

Figure 19 shows the Manhattan plot

Figure 19 The Manhattan plot

6.2.5 Q-Q plot

In the Q-Q plot visualization interface, “**File**” is used to save images. There are two file formats available for saving: SVG and PNG, and you can select the desired format to save. When saving, it captures the image displayed on the page, which means if the image is enlarged, the saved image will also be enlarged. “**Settings**” is used to adjust the coordinate axes. “**Color**” is used to adjust the color and size of the points. The Q-Q plot is shown in Figure 20:

Figure 20 shows the Q-Q plot

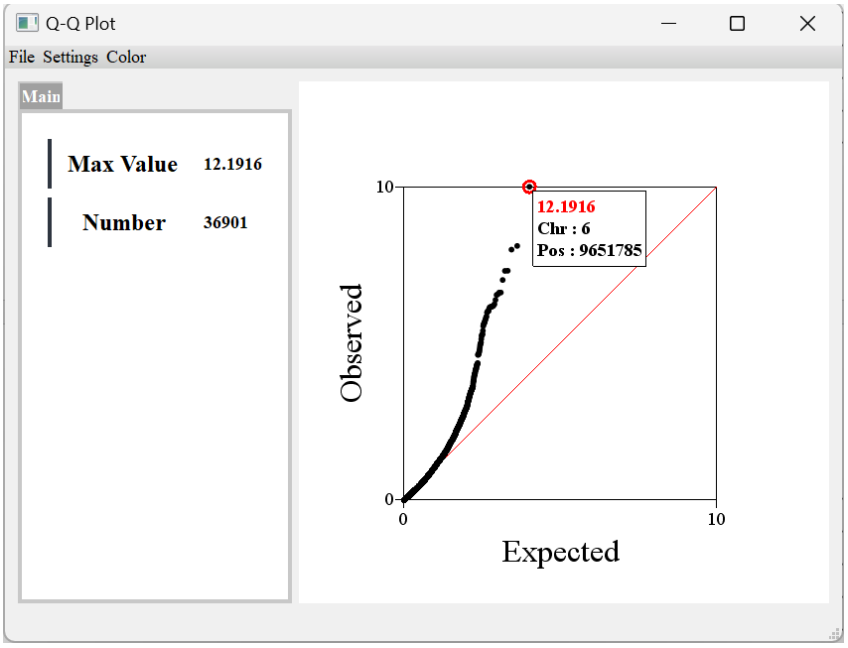

Figure 20 The Q-Q plot

7 TUTORIALS

The test data is sourced from academic papers (Keyan Zhao et al.nature communications,2011).

The One-step GWASrice demonstration data (described at DOI: 10.1038/ncomms1467) consisting of 413 diverse lines. The genotypic data consist of 36,901 SNPs distributed across the rice genome, and are available in plink binary format. The phenotype is FT ratio of Faridpur/Aberdeen. Covariant file was provided from paper. The kinship matrix was calculated using ONE STEP GWAS. The process is as shown in the lower figure in Figure 21.

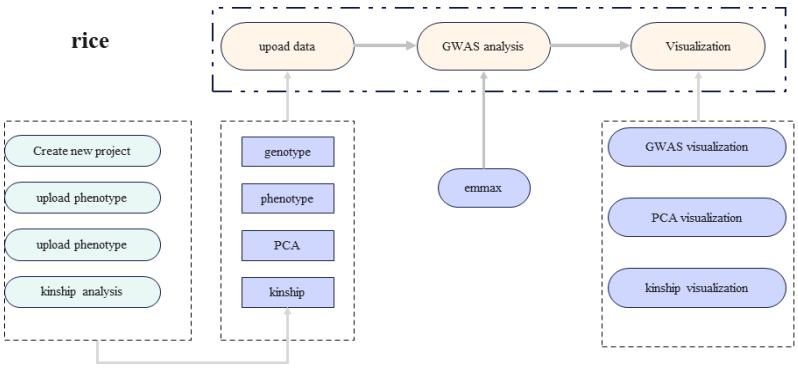

Figure 21 The process of rice GWAS analysis

7.1 Kinship Analysis

Genotype, phenotype, and PCA data were provided, and the kinship relation matrix data was generated using this software. The steps for generating the kinship matrix and the results outputs are shown in Figure 22.

1. Create New Project:

   Open the software > **Right click** > **Create New Project**.

2. Input Data:

   Right-click on the created project > **Load GWAS File** > Input **Genotype File** > Input **Phenotype File.**

3. Analysis:

   Navigate to the "**Analyse**" function: **Analyse** > **Kinship.**

4. Save Result:

   When the analysis is completed, **Right-click** on the created project > **Save Result.**

Figure 22 shows the process of rice kinship analysis and results.

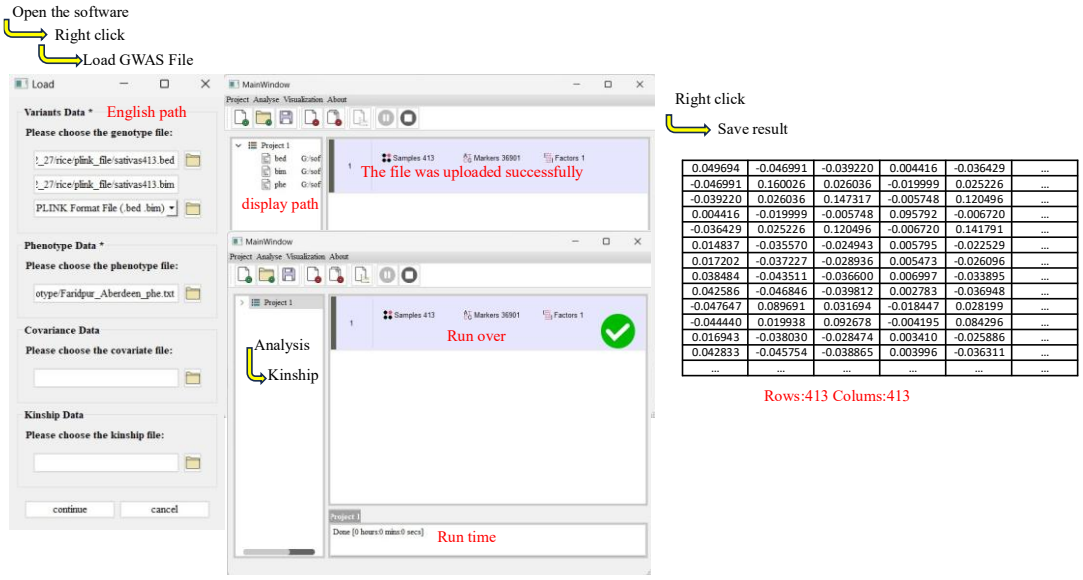

Figure 22 The process of rice kinship analysis and results

7.2 GWAS Analysis

Following the completion of the previous step, GWAS analysis using "**EMMAX**". The steps and the results outputs are shown in Figure 23.

1. Create New Project:

   Open the software > **Right click** > **Create New Project**.

2. Input Data:

   Right-click on the created project > **Load GWAS File** > Input **Genotype File** > Input **Phenotype File >** Input **covariate File >** Input **Kinship File.**

3. Analysis:

   Navigate to the "**Analyse**" function: **Analyse** > EMMAX**.**

4. Save Result:

   When the analysis is completed, **Right-click** on the created project > **Save Result.**

Figure 23 shows the process of rice GWAS analysis and results.

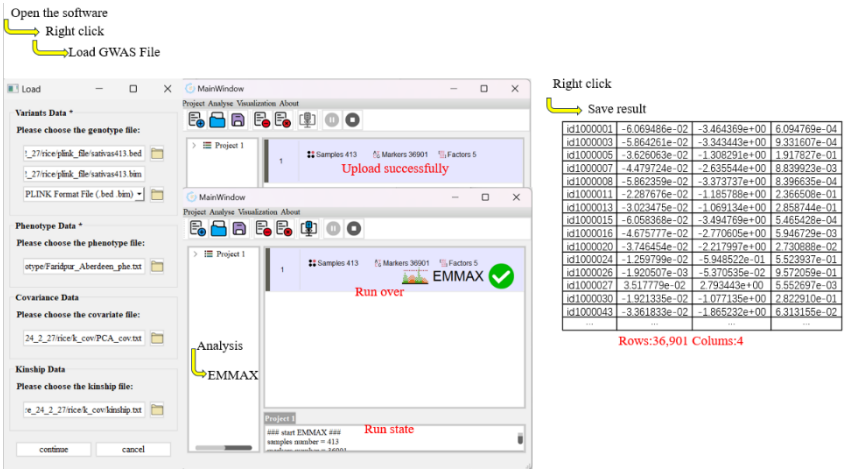

Figure 23 The process of rice GWAS analysis and results.

7.3 Manhattan plot

Navigate to the "**Visualization**" function: **Visualization** > **GWAS Visualization** > Input **Chromosome Length File** > Input **GWAS Output File** > Input ***.bim File** > choose **P value column** > Input **Genome Annotation File**. The steps and the results outputs are shown in Figure 24.

Figure 24 shows the process of rice Manhattan plot and visualization.

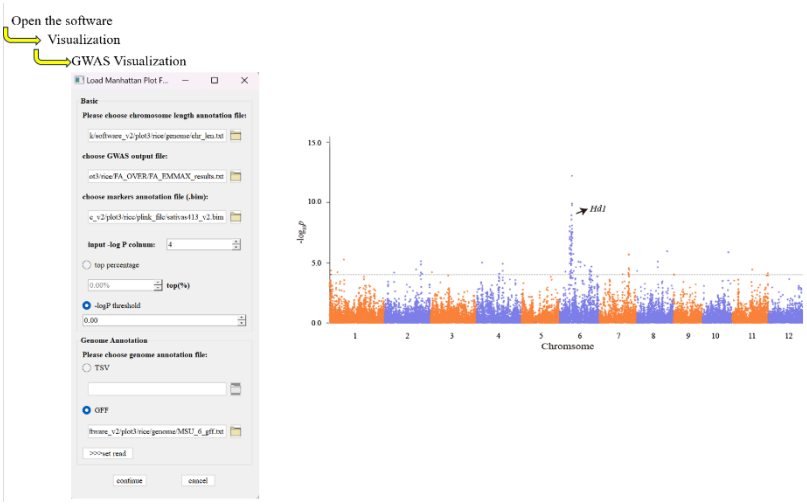

Figure 24 The process of rice Manhattan plot and visualization

7.4 Q-Q Plot

Navigate to the "**Visualization**" function: **Visualization > Q-Q Visualization >** Input **Chromosome Length File** > Input **GWAS Output File** > Input ***.bim file** > choose **P value column.** The steps and the results outputs are shown in Figure 25.

Figure 25 shows the process of rice Q-Q plot andvisualization.

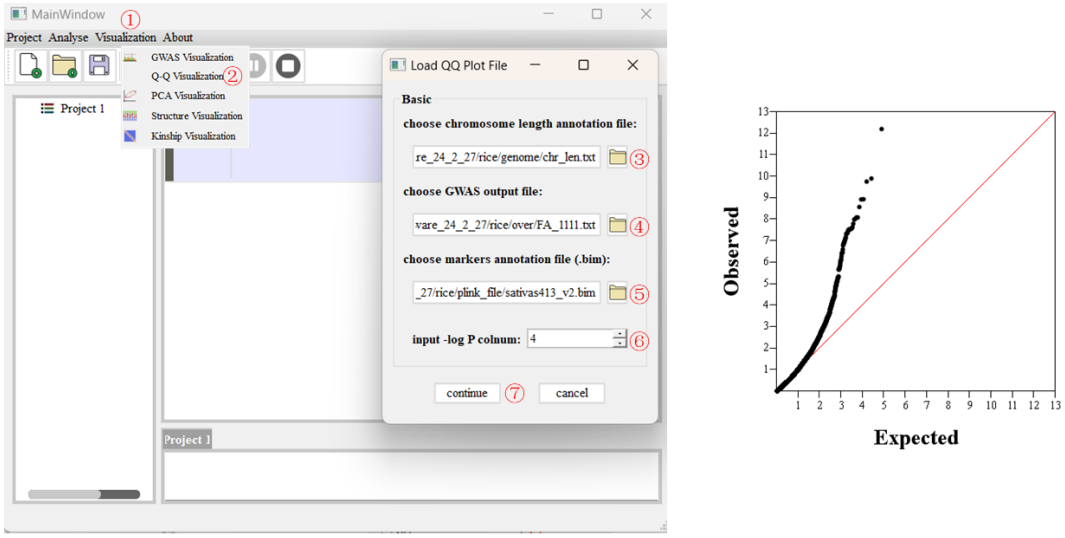

Figure 25 The process of rice Q-Q plot and visualization.

REFERENCES

[1] ZHOU X, STEPHENS M. Genome-wide efficient mixed-model analysis for association studies [J]. Nat Genet, 2012, 44(7): 821-4.

[2] KANG H M, SUL J H, SERVICE S K*, et al.* Variance component model to account for sample structure in genome-wide association studies [J]. Nat Genet, 2010, 42(4): 348-54.

[3] PRITCHARD J K, STEPHENS M, DONNELLY P. Inference of Population Structure Using Multilocus Genotype Data [J]. Genetics, 2000, 155(2): 945-59.

[4] RAJ A, STEPHENS M, PRITCHARD J K J G. fastSTRUCTURE: variational inference of population structure in large SNP data sets [J]. 2014, 197(2): 573-89.