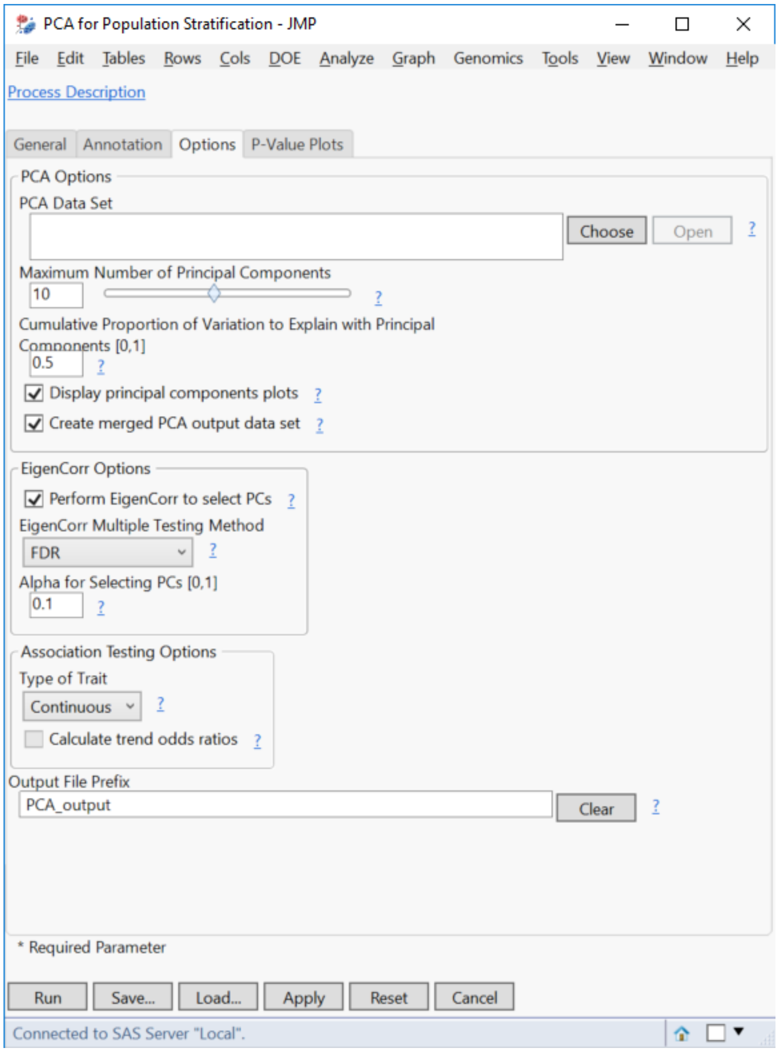
Population Structure: Principal Components Analysis

Population structure is genetic similarity across large groups of individuals or lines. In preparation for association mapping, population structure should be assessed. Like familial relatedness, population structure can be incorporated into an association mapping analysis. In QK association analysis, population structure is modeled with a Q matrix, and familial relatedness is modeled with a K matrix (See *K Matrix and Relatedness Measures Step-Guide*). There are two ways to construct a Q matrix in JMP Genomics: Principal Components Analysis (PCA) and Multidimensional Scaling (MDS).

Principal Components Analysis (PCA) is a data reduction technique similar to MDS, with some important differences. While MDS tries to reduce the data in such a way as to preserve the relationships among observations or lines, PCA is meant to describe the largest sources of variance in the data. As a result, PCA can be sensitive to smaller patterns found in a single maker or a handful of markers, patterns that would not be evident from MDS.

PCA can be used to control for population stratification in association testing in two ways in JMP Genomics. The first method, known as the **Eigenstrat** method, is found in the **PCA for Population Stratification** tool. This tool performs the PCA analysis first, saving the output, and then optionally performs Eigenstrat association analysis, if one or more trait variables are specified.

1. From the **Genomics Starter** menu, choose **Genetics > GWAS Testing > PCA for Population Stratification**.
2. Numeric genotypes are needed for this analysis. Reference the **Recode Genotypes** procedure outlined in the *Recode Genotypes Step-Guide*.
3. **Choose** the **Recode Genotypes** output, **rice_genos_recgeno.sas7bdat**, as the **Input SAS Data Set**.
4. Assign the four **Trait Variables** (**FL, PH, PW, GW**). Assign **GID** as the Label Variable.
5. Type “recgeno:” (without the quotation marks) in the box labeled **List-Style Specification of Marker Variables**.
6. **Choose** an **Output Folder**.
7. On the **Annotation** tab, **Choose rice_anno_recgeno.sas7bdat** as the **Annotation SAS Data Set**.
8. Fill out the **Annotation** tab with RS as the **Annotation** Label, chrom as the **Annotation Group Variable** and pos as the **Annotation Location Variable.**
9. Under the **Options** tab, check the box next to **Create merged PCA output data set.** This creates the file that will be used for Q-K analysis.
10. Select **Continuous** from the **Type of Trait** dropdown menu.
11. Type “PCA_output” in the **Output File Prefix** box.


PCA for Population Stratification - JMP
File Edit Tables Rows Cols DOE Analyze Graph Genomics Tools View Window Help
Process Description
General Annotation Options P-Value Plots
Choose Open
Clear
Run Save... Load... Apply Reset Cancel

PCA Options

PCA Data Set

Maximum Number of Principal Components
10

Cumulative Proportion of Variation to Explain with Principal Components [0,1]
0.5

☑ Display principal components plots

☑ Create merged PCA output data set

EigenCorr Options

☑ Perform EigenCorr to select PCs

EigenCorr Multiple Testing Method
FDR

Alpha for Selecting PCs [0,1]
0.1

Association Testing Options

Type of Trait
Continuous

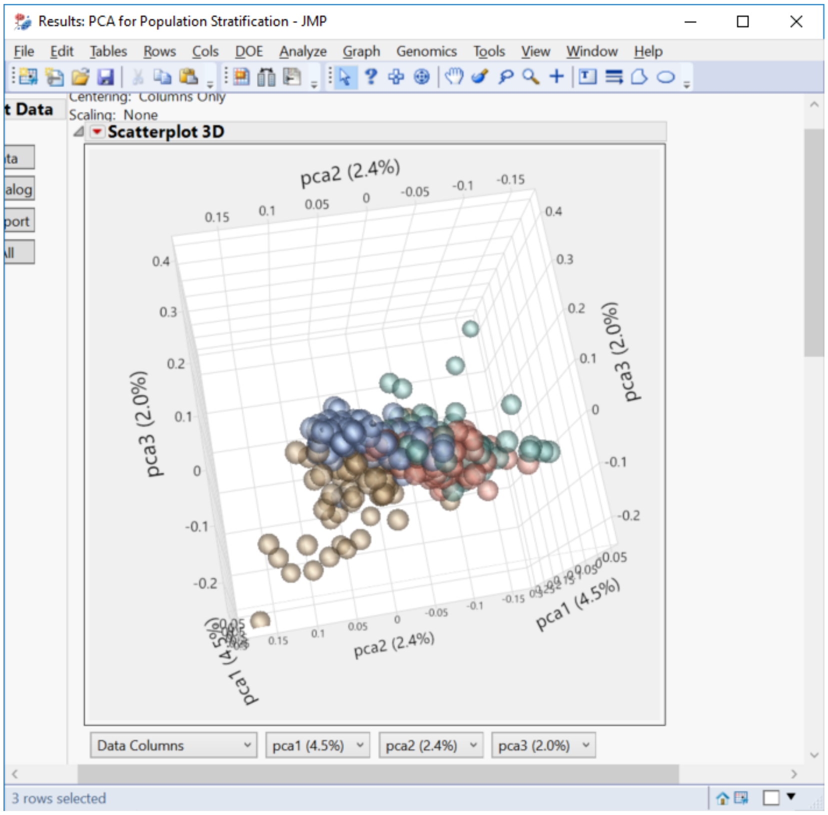
☐ Calculate trend odds ratios

Output File Prefix
PCA_output

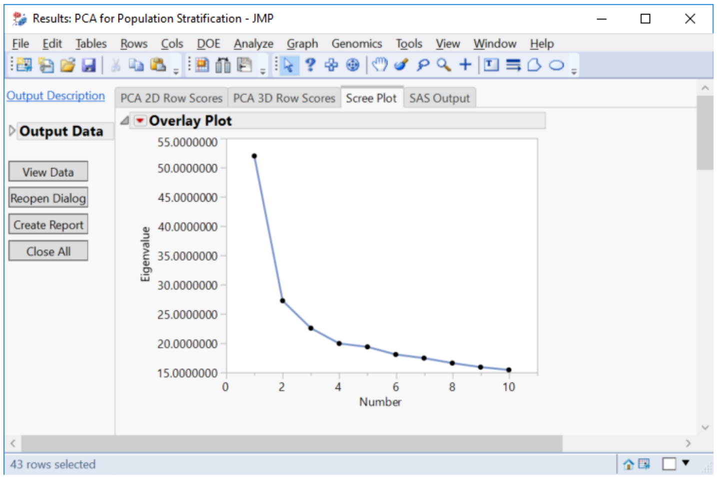
* Required Parameter

Connected to SAS Server "Local".

12. Click **Run** to start the analysis.
13. When the results dashboard appears, click on the **PCA 3D Row Scores** tab to view this output.



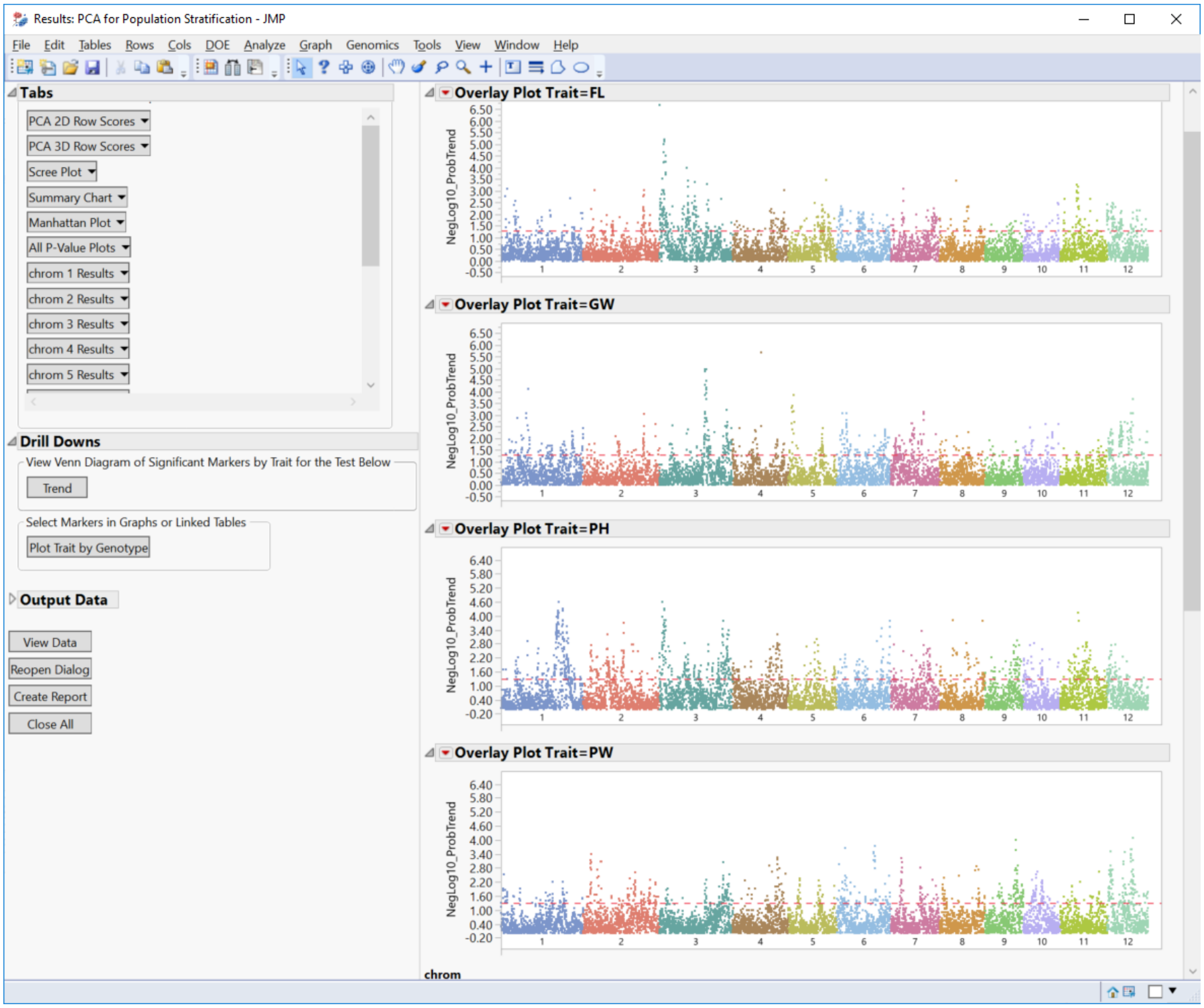
14. Explore the **Scree Plot**.
   - Look for the elbow in the plot to determine a sufficient number of dimensions for the analysis.



15. Click on the **Summary Chart** tab to view results from the Eigenstrat association analysis.

- The bar charts show statistical associations between the four traits and markers on multiple chromosomes.
- To interrogate a single chromosome at a time, find the button for that chromosome in the **Tabs** section and choose **View Tab**, or click the **All P-Value Plots** button. Or choose **View Data** from one of these buttons to look at the data table containing all the markers and significance results.

16. The Manhattan Plot tab shows significant markers for each of the four traits colored by chromosome. Points on the plot above the red line are considered significant markers.



17. The file **pca_output_pcm.sas7bdat** is now located in the **Output Folder** designated earlier and can be used in **Q-K Association Analysis**.

Follow-Up Processes

This guide covered the Q-matrix part of Q-K analysis. The Q matrix contains information about population structure, which can come from **Multidimensional Scaling, Principal Components Analysis**, or even manual assignment of the lines or individuals into groups curated by the user. The output data set, **pca_output_pcm.sas7bdat**, from this analysis can be used in **Q-K association analysis**. For the next step in the **Q-K Association Analysis** pathway, view the *Q-K Association Analysis Step-Guide.*