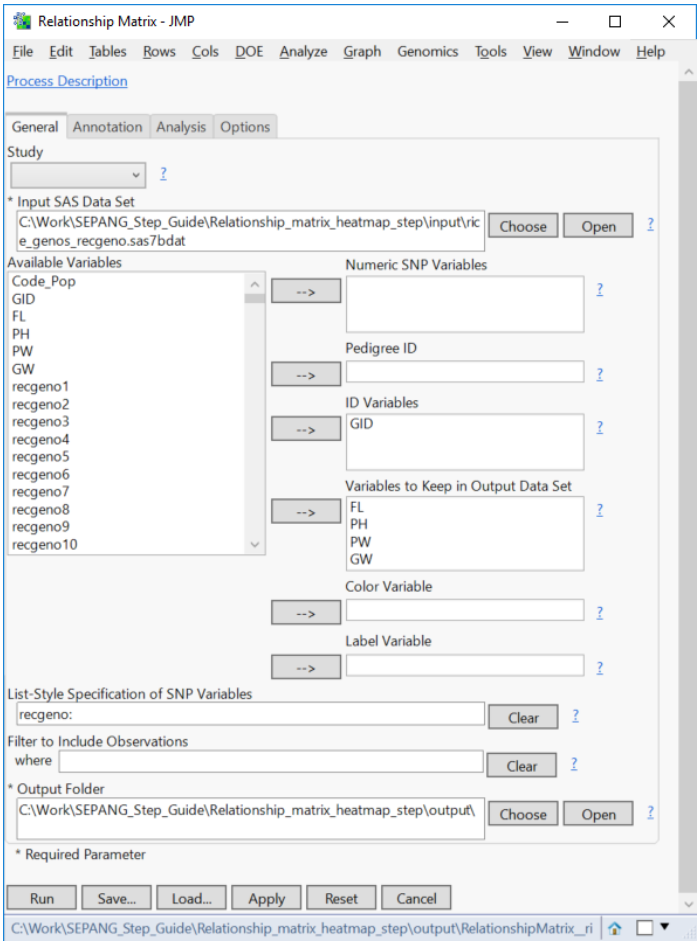
Relationship Matrix/Heatmap

In JMP Genomics there are two basic tools for computing and displaying relatedness among lines: **Relationship Matrix** and **Kinship Matrix**. The **Relationship Matrix** tool estimates the relationships among the lines using marker data, while the **Kinship Matrix** procedure takes pedigree information and computes the relationship measures directly. The **Kinship Matrix** process creates a matrix containing coancestry coefficients or covariance coefficients, while the **Relationship Matrix** computes one of three options: Identity-by-Descent, Identity-by-State, or Allele-Sharing-Similarity. Output from these two procedures can serve as the K matrix, representing familial relatedness, in Q-K association analysis. This step-guide will focus on the **Relationship Matrix** using a dataset containing 343 rice lines with 8,336 markers. Again, because this data does not have pedigree information, the **Relationship Matrix** process must be used. For a guide to the Kinship Matrix process, please refer to the *Kinship Matrix Step-Guide*.

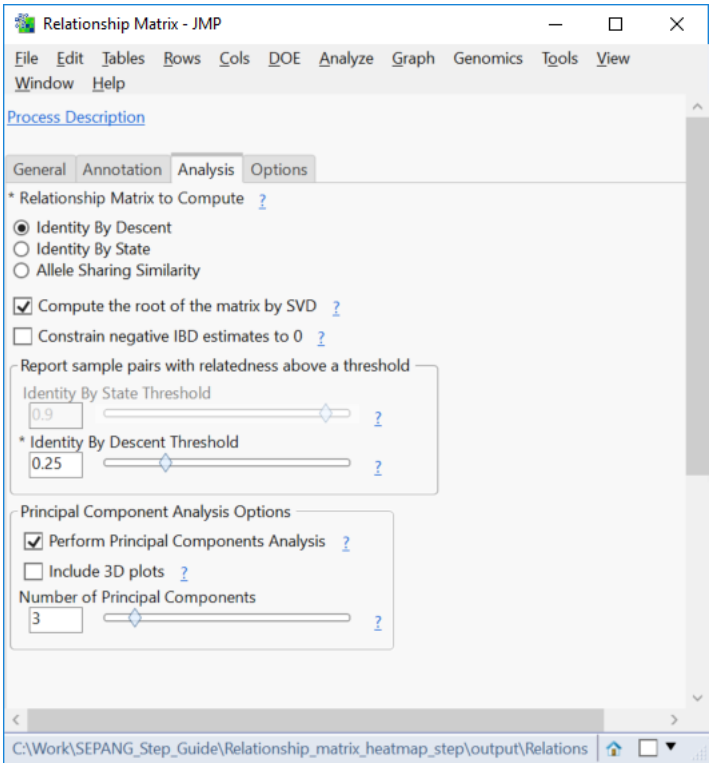
Relationship Matrix:

1. **Open** the **rice_genos.sas7bdat** dataset and inspect it in JMP. It has 343 rice lines in rows, 5 columns of annotation and traits, and 8,336 columns with marker data. These markers need to be coded as numeric genotypes. This numeric format is required for input to the **Relationship Matrix** procedure. Data in other formats should be converted to numeric genotypes using the **Recode Genotypes** procedure outlined in the *Recode Genotypes Step-Guide*.
2. From the **Genomics Starter** menu, choose **Genetics > Relatedness Measures > Relationship Matrix**.
3. Find the output dataset from the **Recode Genotypes** procedure, **rice_genos_recgeno.sas7bdat**, and **Choose** it as the **Input SAS Data Set**.
4. Select the **GID** variable from the **Available Variables** list, and place it into the **ID Variables** and **Label Variable** boxes.
5. Select all the non-marker variables, starting with **GID** and ending with **GW**, and place them in the box labeled **Variables to Keep in Output Data Set**.
    - These are columns containing phenotypic data.
6. In the box labeled **List-Style Specification of SNP Variables**, type “recgeno:” (without the quotes) to select all variables starting with the prefix “recgeno” as marker variables.
7. **Choose** an **Output Folder**.

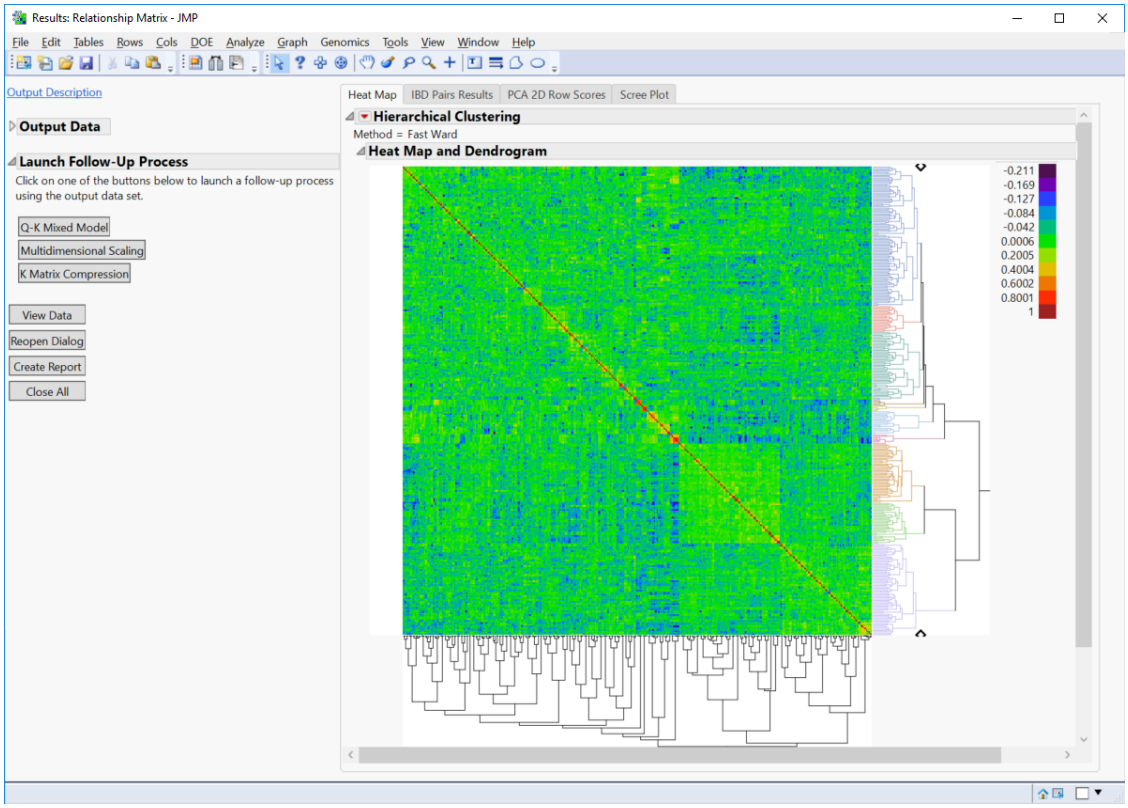


8. In the **Annotation** tab, select **rice_anno_recgeno.sas7bdat** as the **Annotation SAS Data Set**.
9. In the **Analysis** tab, leave the **Identity By Descent** option selected.
    - This will estimate the probability that individuals in the relationship matrix share an allele from a common ancestor at a specific locus. As noted above, options are available for Identity By State and Allele Sharing Similarity which use Gower's Similarity Metric to estimate the probability of two individuals sharing the same allele regardless of inheritance with and without a Range Standardization, respectively.
10. Check the **Compute the Root of the Matrix by SVD** box.
    - This option produces a file containing the square root of the relationship matrix, which can be used later in the QK association analysis.
11. The **Identity By Descent Threshold** slider can be changed to alter the threshold of IDB for pairs to be reported in an output dataset. The default setting is .25, meaning all pairs of rows with an IDB value greater than or equal to .25 will be included in the output dataset **rice_genos_recgeno_prs.sas7bdat.**

12. In the **Principal Component Analysis Options**, JMP gives options to perform PCA and set the number of Principal Components to include in the analysis.
    - Principal Component Analysis is a tool to combine input variables in a way that eliminates the facets of variables that do not explain variance in the data. The number of components will designate how many smaller factors will be used as new variables account for as much of the overall variance as possible without bloating or overfitting the model.

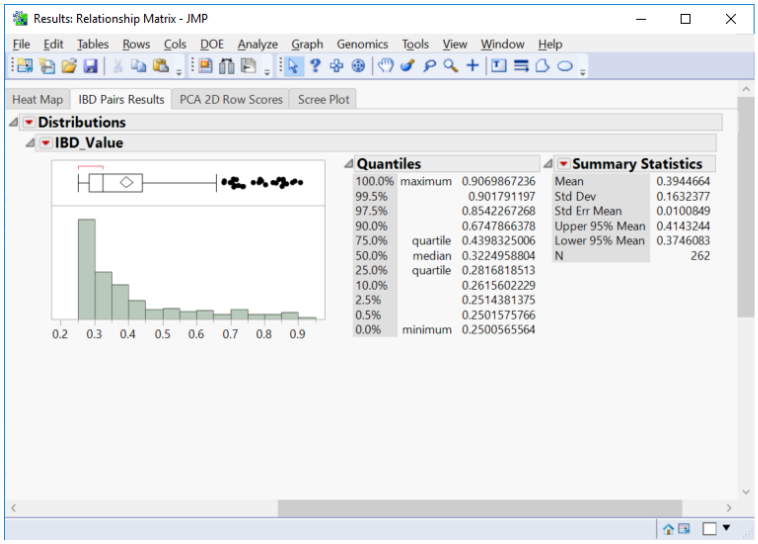


13. In the **Options** tab, check the box labeled **Plot Relationship Matrix Heat Map**. If you would like to append a prefix to the output variables it can be done in this tab as well.
14. Click **Run** to start the analysis. Examine the Heatmap Results in the first tab of the results dashboard:



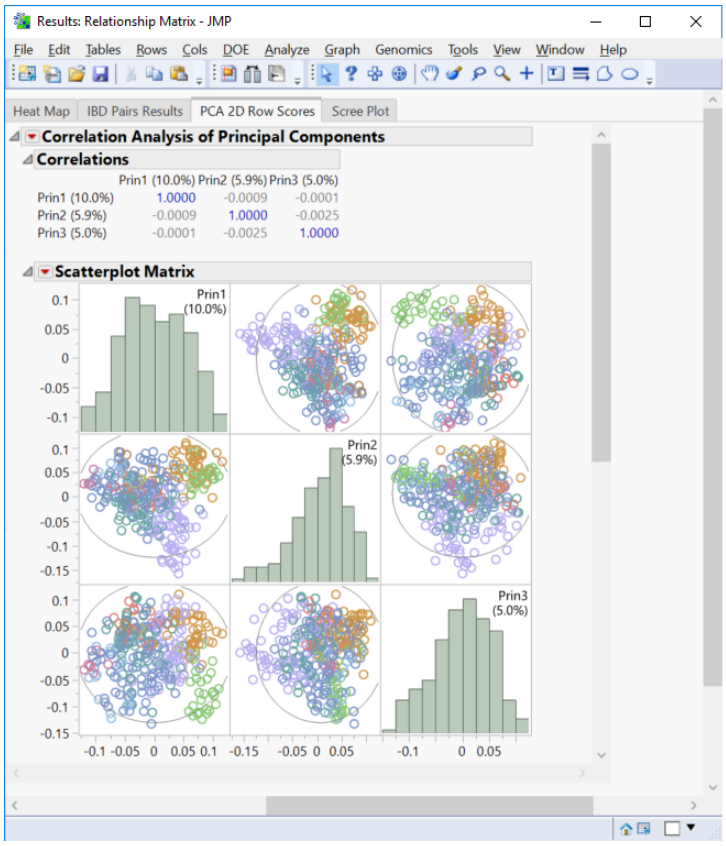
- The heatmap displays the relationships among the 343 lines. The red diagonal represents perfect relationship of each line with itself; the symmetric off-diagonal elements represent relationship measures (in this case IBD) for pairs of lines. The blocks of warmer colors on the diagonal show clusters of closely related lines.
- The dendrogram (tree diagram) on the right shows the results of a cluster analysis on the IBD matrix. Double click on any branch to zoom in and inspect the members. To revert to the top-level view, click on the **Hierarchical Clustering** hotspot and choose **Release zoom**.

15. Return to the results dashboard, and view the **IBD Pairs Results** tab.



- The histogram shows the distribution of IBD scores for the 262 pairs of lines with IBD values greater than 0.25. A dataset of these pairs has also been saved to the specified **Output Folder** titled **rice_genos_recgeno_prs.sas7bdat.** This table is also viewable by clicking the **View Data** button.

16. Look at the **PCA 2D Row Scores** tab.



- This scatterplot matrix shows the correlations between each of the 3 principal components. There is not evidence for strong population structure in these results since there isn't any stratification of points in these scatterplots.

17. Examining the **Scree Plot** tab shows the proportion of the variance accounted for by each Principal Component. In this case, the first two Principal Components account for most of the variation.
18. Find the section on the left of the dashboard labeled **Output Data**, and click on the triangle to display the contents of this section.
19. A file named **rice_genos_rc__rm.sas7bdat** is shown. This file contains the square root of the IBD matrix, to be used in QK association. This is different from the raw IBD values displayed in the heatmap.

Summary

This document served as a walkthrough for creating a **Relationship Matrix** from a data set containing 343 rice lines with 8,336 markers. This relationship matrix was composed of Identity By Descent values, but can be calculated for Identity By State and Allele Sharing Similarity as well. This process estimated the relationships among the lines using marker data since no pedigree information was available. Additionally, the input marker data was in numeric format, which is required for this analysis.

Follow-Up Processes

From the output window, Follow up processes are available for building a **Q-K Mixed Model**, **Multidimensional Scaling,** and **K-Matrix Compression.** The output data set, **rice_genos_rc__rm.sas7bdat**, from this analysis can be used as the **K matrix** in **Q-K association analysis**. For the next step in the **Q-K Association Analysis** pathway, view the *Principal Components Analysis Step-Guide.*