Step by Step Guide to Importing Genetic Data into JMP Genomics

USING JMP GENOMICS 8

Introduction

Data for genetic analyses can exist in a variety of formats. Before this data can be analyzed it must imported into one or more properly formatted SAS data sets. JMP Genomics contains numerous processes for importing and formatting data from a variety of sources. In this document, we will explore two of these processes.

Typically the analyses will be performed on wide data sets where genotypes/alleles are in columns and samples are in rows. If available, it is recommended that additional sample annotation or phenotype information be added as additional columns to the wide data set.

JMP Genomics uses the variable/column/header name conventions of SAS. Typically variables/columns/headers should be no more than 32 characters and start with a letter. The following link provides more information:

http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=getting_started.01.16.html

For more information about the different data sets used by JMP Genomics, go to this link:

http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=getting_started.01.21.html#356832

Objectives

In this document, we explore the following import processes:

- Importing data from text-format files
- Importing VCF files

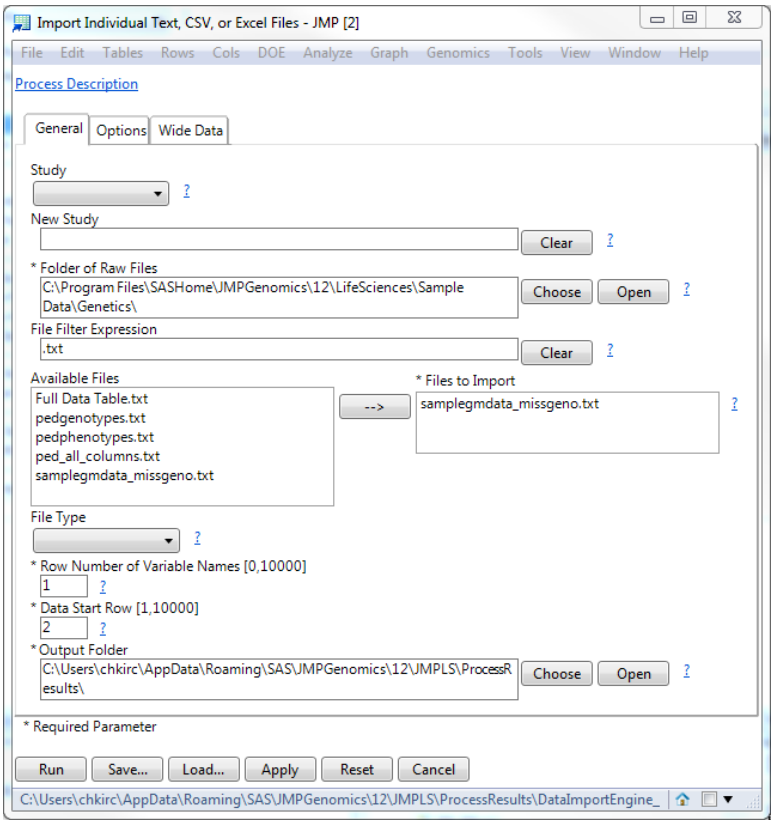
Importing Text Format Genotype Data

The JMP Genomics text file format import engine can be used to import large genotype data sets. For small data sets (fewer than 5,000 markers or so), the JMP text file import (**File > Open)** may be the best option. Regardless, the resulting data table is saved as a **sas7bdat** file. In this section, we will cover importing large data sets.

A typical data set will have samples as rows, phenotypes, traits and genotypes as columns. Phenotypes and traits can be categorical or numeric. Genotypes are typically in the form of (AA, AB, BB), (A/A, A/B, B/B) or (0, 1, 2)

*Note that if any text file is imported and a map file (genotype annotation file) is separately imported, you must run the **Subset and Reorder** genetic utility prior to any other process.*

1) Select **Import > Text > Import Individual Text, CSV, or Excel Files** from the Genomics starter menu.
    - We recommend that you import data from text or csv formatted files, as Excel formats often have formatting that is incompatible with the import process.
2) Complete the **General** tab as shown below. This file has all samples listed as rows and genotypes as columns. The location of the sample data is listed in the “Folder of Raw Files” field.



- The **File Filter Expression** box is optional.
- Multiple files can be imported, but all the files must be in the same format.
- **File Type** does not need to be specified if the file suffix is .txt or .csv. If the file has a different suffix but is in one of these formats, specify the type with this pull-down menu.
- The **Row Number of Variable Names** should have the appropriate value entered.
- The **Data Start Row** should have the appropriate value entered.
- The **Output Folder** is where you would like to save the resulting SAS data set.

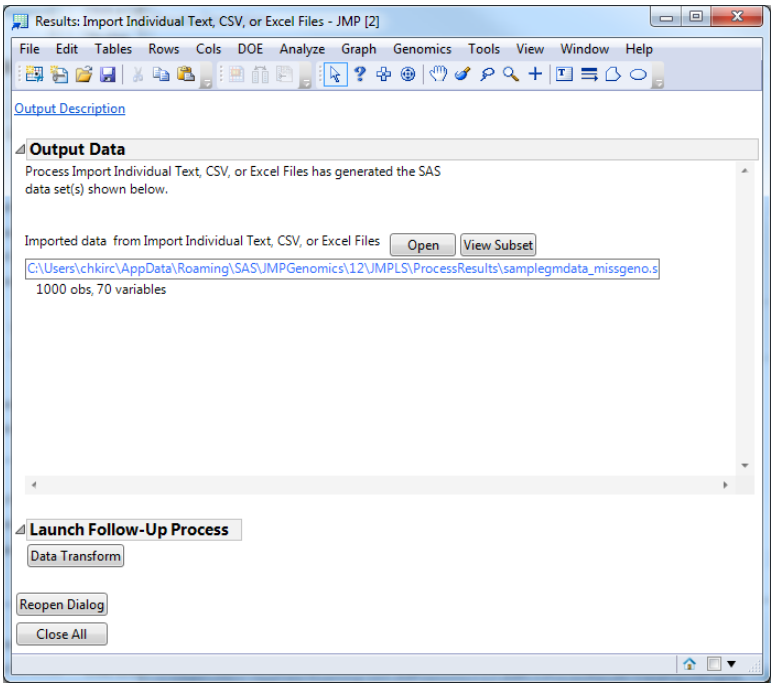
3) Select the **Options** tab.

- Here you can set the number of rows to scan.
- In this example, we make no changes to the default settings on the **Options** tab.

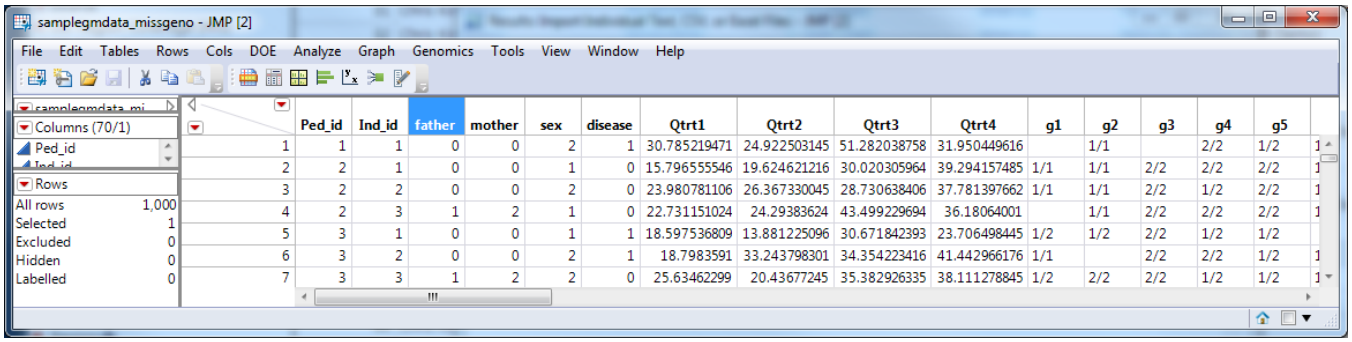
4) Select the **Wide Data** tab.

- Complete the **Minimum Number of Columns to Scan** box. Set this value to the column number in the data set past which the type of data (categorical vs. numeric) is constant from that column to the last column of the data table. In this case, you can leave it to its default of 0.
- When JMP Genomics imports a text or csv file, the software checks each column for the data type. Setting the minimum number of columns to scan will turn off this process for all columns beyond the selected value and significantly improve the import speed.
- All other fields can be left as the default.

5) Click **Run** to start the process.
6) The result is a window with a link to the created SAS data set:



7) The resulting SAS data set will look like this:



	Ped_id	Ind_id	father	mother	sex	disease	Qtrt1	Qtrt2	Qtrt3	Qtrt4	g1	g2	g3	g4	g5
1	1	1	0	0	2	1	30.785219471	24.922503145	51.282038758	31.950449616		1/1		2/2	1/2
2	2	1	0	0	1	0	15.796555546	19.624621216	30.020305964	39.294157485	1/1	1/1	2/2	2/2	2/2
3	2	2	0	0	2	0	23.980781106	26.367330045	28.730638406	37.781397662	1/1	1/1	2/2	1/2	2/2
4	2	3	1	2	1	0	22.731151024	24.29383624	43.499229694	36.18064001		1/1	2/2	2/2	2/2
5	3	1	0	0	1	1	18.597536809	13.881225096	30.671842393	23.706498445	1/2	1/2	2/2	1/2	1/2
6	3	2	0	0	2	1	18.7983591	33.243798301	34.354223416	41.442966176	1/1		2/2	2/2	1/2
7	3	3	1	2	2	0	25.63462299	20.43677245	35.382926335	38.111278845	1/2	2/2	2/2	1/2	1/2

8) Information can be found at this link regarding importing text files:

http://www.jmp.com/support/downloads/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=PR_G_IM_0028.html

Importing VCF files

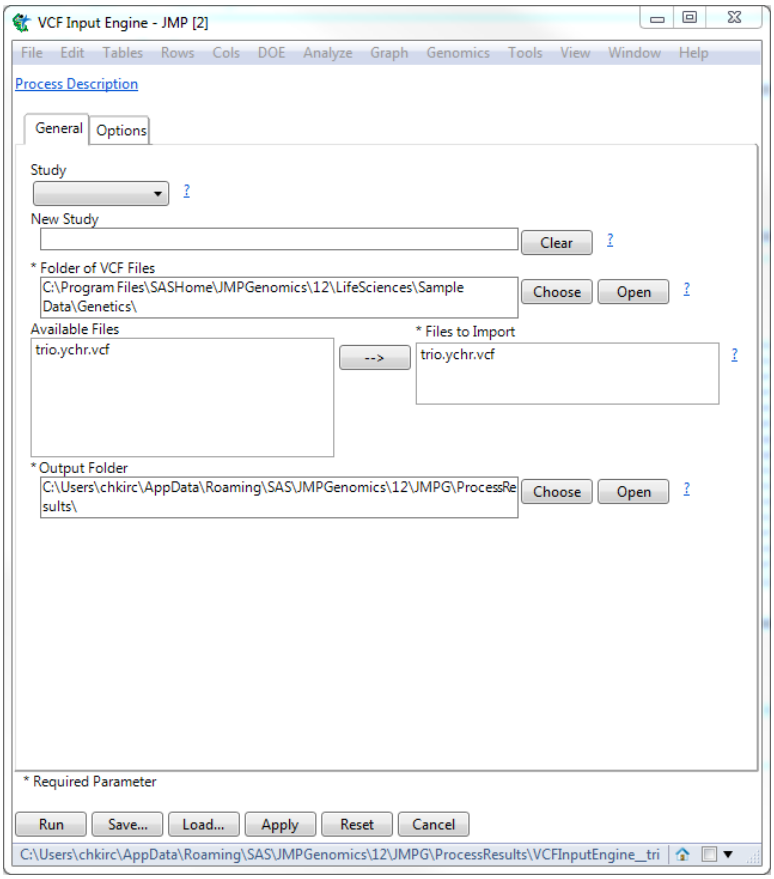
The JMP Genomics VCF file format import engine can be used to import large or small genotype data sets.

*Note that if any text file is imported and a map file (genotype annotation file) is separately imported, you must run the **Subset and Reorder** genetic utility prior to any other process.*

More information about VCF importing can be found here:

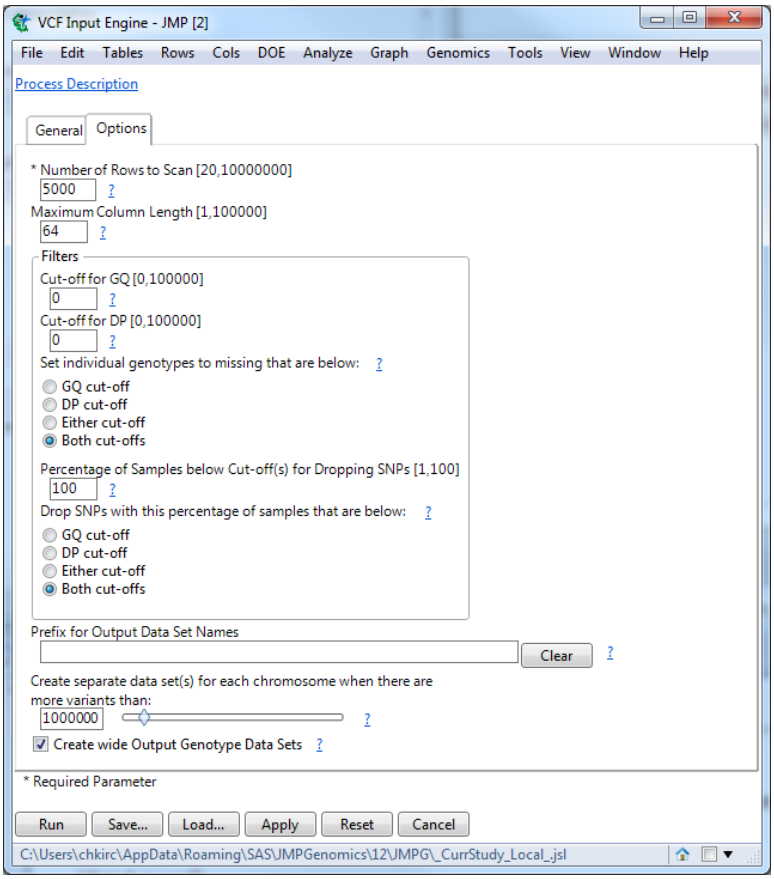
file:///C:/Program%20Files/SASHome/JMPGenomics/12/LifeSciences/Documentation/life_sciences_documentation/wwhelp/wwhimpl/js/html/wwhelp.htm#href=PR_G_IM_0053.html

1) Select **Import > Next-Gen Sequencing > Import VCF Files** from the Genomics starter menu.
   - You can have one or more VCF files. VCF version 4.0 or 4.1 is supported
2) Complete the **General** tab as shown below. This file has all samples listed as rows and genotypes as columns. The location of the sample data is listed in the "Folder of VCF Files" field.



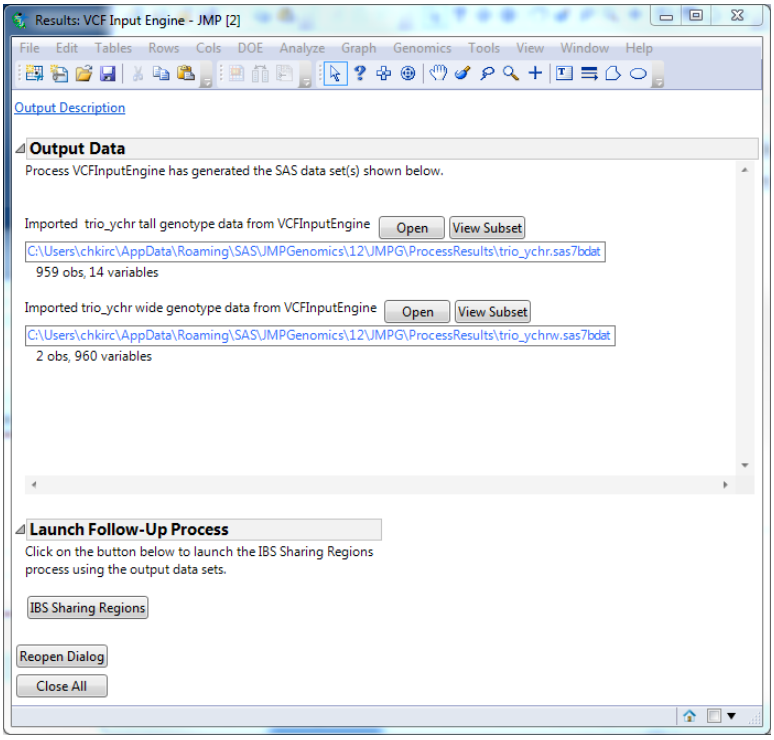
- Choose the directory of where the files are and which Files to Import.
- **Output Folder** is where you would like to save the resulting SAS data set.

3) Go to the **Options** tab:

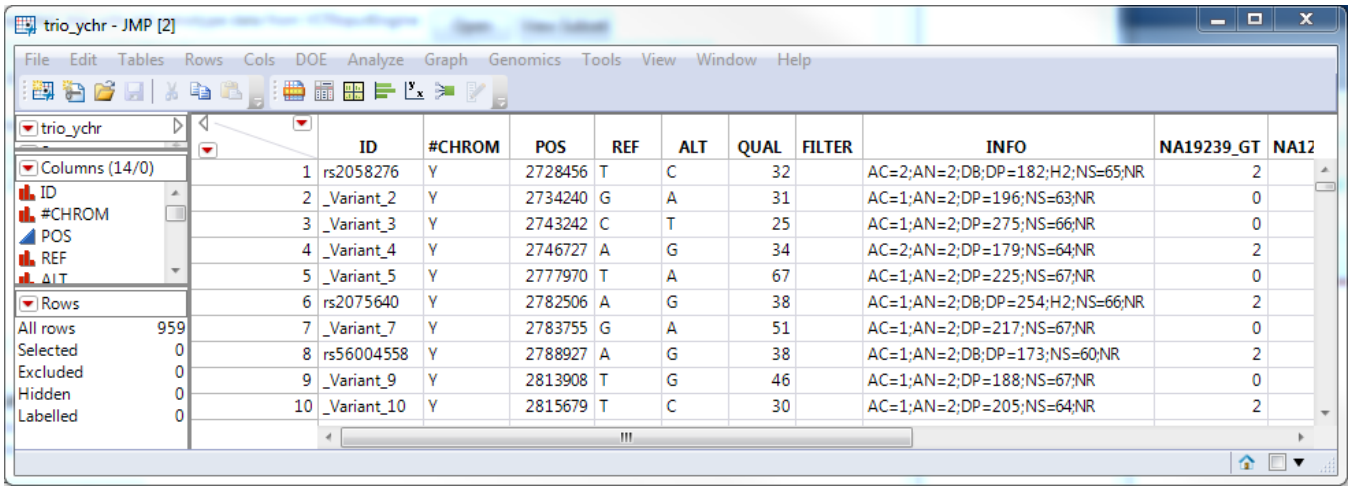


- Here you can set the **Number of Rows to Scan**, **Maximum Column Length** and filters for GQ and DP cut-offs
- You can also specify if that there are more than x number of variants, to split them up into separate files for each chromosome
- Make sure that the **Create wide Output Genotype Data Sets** option is checked

4) Click on the **Run** button to begin the process
5) The resulting window will look like this:



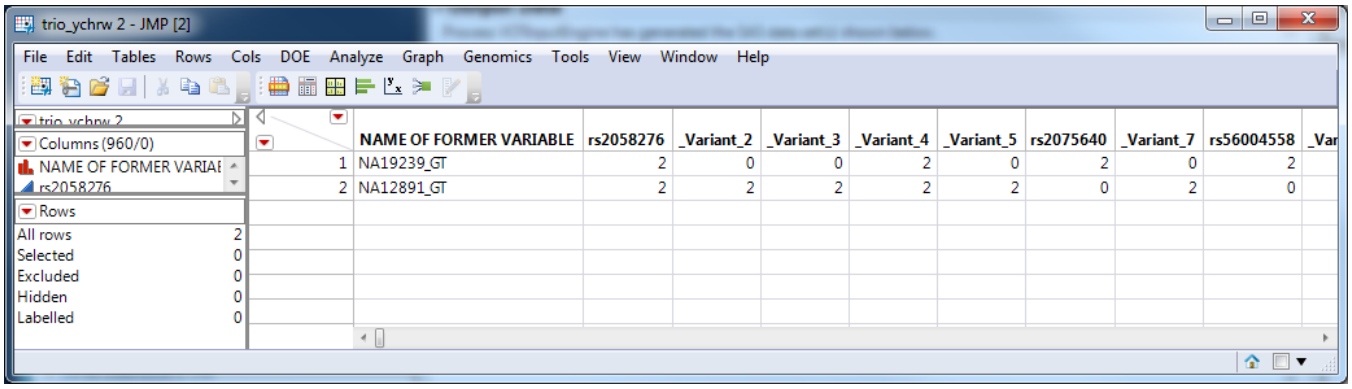
6) There will be a link to the tall data set that will have additional columns including the GT (genotype), DP and GQ columns for each sample. Click on **Open** button to see the file. Sample names are shown in the column header/name. Variant ID is in the ID column. See example below:



	ID	#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	NA19239_GT	NA12
1	rs2058276	Y	2728456	T	C	32		AC=2;AN=2;DB;DP=182;H2;NS=65;NR	2	
2	_Variant_2	Y	2734240	G	A	31		AC=1;AN=2;DP=196;NS=63;NR	0	
3	_Variant_3	Y	2743242	C	T	25		AC=1;AN=2;DP=275;NS=66;NR	0	
4	_Variant_4	Y	2746727	A	G	34		AC=2;AN=2;DP=179;NS=64;NR	2	
5	_Variant_5	Y	2777970	T	A	67		AC=1;AN=2;DP=225;NS=67;NR	0	
6	rs2075640	Y	2782506	A	G	38		AC=1;AN=2;DB;DP=254;H2;NS=66;NR	2	
7	_Variant_7	Y	2783755	G	A	51		AC=1;AN=2;DP=217;NS=67;NR	0	
8	rs56004558	Y	2788927	A	G	38		AC=1;AN=2;DB;DP=173;NS=60;NR	2	
9	_Variant_9	Y	2813908	T	G	46		AC=1;AN=2;DP=188;NS=67;NR	0	
10	_Variant_10	Y	2815679	T	C	30		AC=1;AN=2;DP=205;NS=64;NR	2	

- Notice that the genotype is the numerically coded genotype where 0 is major homozygous, 1 is heterozygous and 2 is minor homozygous. The call of the genotype is determined from within the VCF file and extracted from the GT portion of the column with the sample name. If called as 0 in VCF , then 0 in the SAS data set. 1 in the VCF refers to 2 in the SAS data set.

7) There will also be a link/path for the wide data set (you can view by clicking on the **Open** button). It is the data set you will use for the rest of the analysis. An example is below:



- Be sure to add phenotype and additional sample information for each sample as separate columns. You can join the information into this file by using the Name of Former Variable column as a unique ID. Use either the **Join** tool from JMP found under the Tables menu or the SAS Utility found **under SAS Data Set Utilities>Tables> Merge** from the Genomics Starter menu

This concludes the import of genetic data. Following import, you should next refer to the **Basic Genetic Analysis** document which reviews data cleaning and assessment as well as basic association and other statistical tests.