

SNP Cluster Plots

Author: Greta Linse Peterson, Golden Helix, Inc.

Overview

This function creates scatter plots based on A and B allele intensities that can be split on SNP genotypes to create tri-colored cluster plots. The function will work for up to 100 SNPs at a time, although a separate scatterplot is created for each SNP.

You will need genotypes for up to 100 SNPs, and you must also have the Affymetrix CEL files available for import.

This document provides instructions on how to import the intensity data from the CEL files, how to use the function, and how to split the scatter plots based on genotypes.

Recommended Directory Location

Save the script to the following directory:

*..\Application Data\Golden Helix SVS\UserScripts\SVS\Tools\

Note: The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between operating systems. The easiest way to locate this directory on your computer is to open SVS and go to **Tools >Open Folder > UserScripts Folder** and save the script in the **\SVS\Tools** folder. If you save the script to the proper folder, it will be accessible from the project navigator **Tools** menu.

Obtaining the Required Datasets

In order to create cluster plots you need the following items:

1. Genotypes for all samples for at most 100 SNPs.
2. Affymetrix CEL files for the Allele A and B intensities

Assuming that genotype analysis has been performed to determine significant SNPs and that cluster analysis now needs to be performed, the genotypes should already be available in the project. See **Figure 1** for a sample project. The relevant nodes are number 25, the Association Test results where the top 25 markers are selected (see **Figure 2**, and number 26, where all other SNPs (and phenotypes) are inactivated except for these markers (see **Figure 3**).

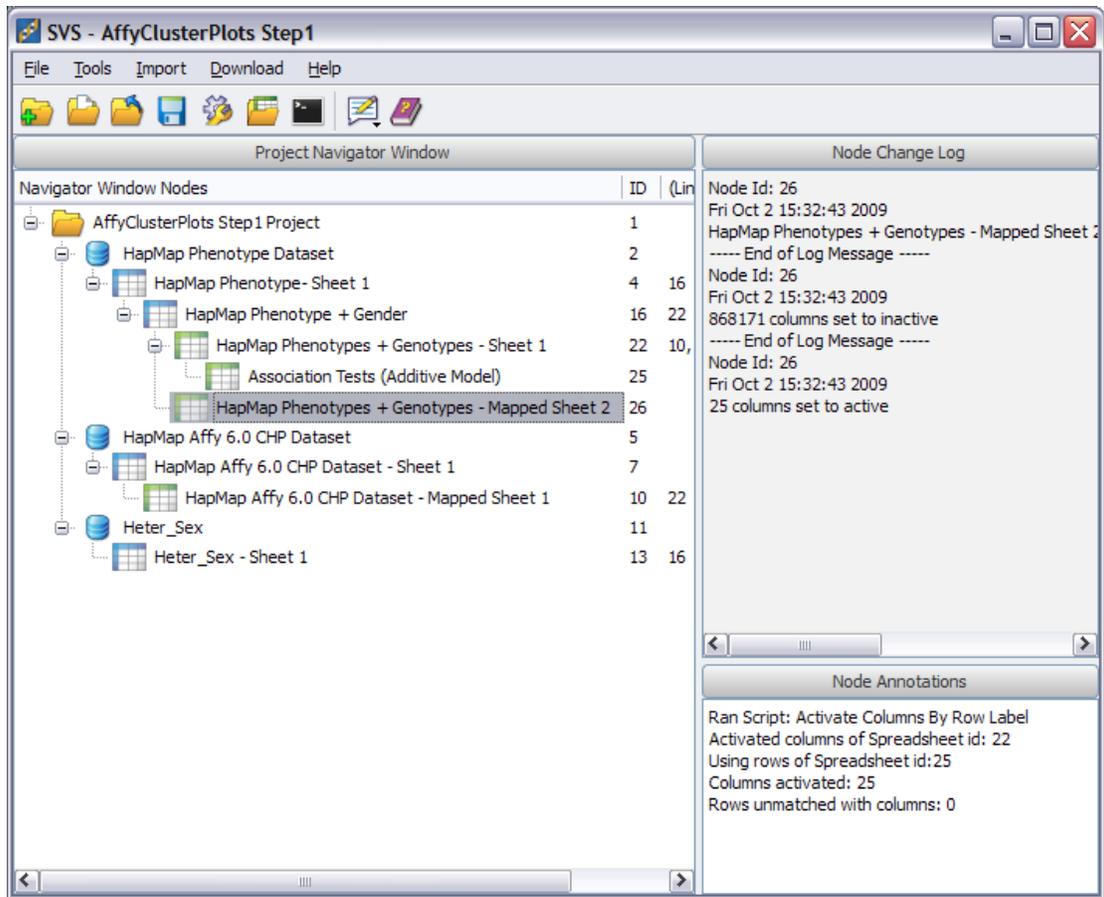


Figure 1: HapMap project with association results on gender

Map	Marker	R	1	R	2	R	3
			Corr/Trend P		Corr/Trend -log10 P		Corr/Trend R
1	SNP_A-8631489		5.7339302561275e-056		55.2415475938386		-0.977481152385415
2	SNP_A-8401046		9.74014364808917e-054		53.011434638079		-0.957146603234602
3	SNP_A-8547953		4.89436492634959e-052		51.3103036525585		-0.959984841489156
4	SNP_A-8325240		1.05829468427999e-051		50.9753933852905		-0.936401025876785
5	SNP_A-4304555		5.36177712933609e-006		5.27069124210829		0.277425128574724
6	SNP_A-2092035		5.75058465615366e-006		5.24028799878283		-0.276525660183003
7	SNP_A-8680119		5.93788983695091e-006		5.22636786393905		0.276112938232312
8	SNP_A-2206294		6.98720328716988e-006		5.15569662110636		0.274008768420886
9	SNP_A-8372209		9.01418728062345e-006		5.04507342317913		-0.270684911867888
10	SNP_A-8391943		1.52318114779921e-005		4.81724844412861		-0.263718751244533
11	SNP_A-2164086		1.52318114779921e-005		4.81724844412861		-0.263718751244533
12	SNP_A-2250877		1.53656966950338e-005		4.8134477436439		0.263601110234547
13	SNP_A-4208034		1.61066002958149e-005		4.79299611871052		0.262967256672239
14	SNP_A-2300919		1.72122872236835e-005		4.76416141540031		-0.26207120918048
15	SNP_A-8505180		1.93635238920692e-005		4.71301560427888		0.260960425316611
16	SNP_A-1815694		2.09236857440098e-005		4.67936181132991		0.259419675387037
17	SNP_A-2305874		2.44772246626159e-005		4.61123782598892		0.257271525080428
18	SNP_A-2090743		2.90132365097878e-005		4.53740382169926		-0.254924726337964
19	SNP_A-1915166		2.907143522042e-005		4.53653352712457		0.254896947044924
20	SNP_A-8638676		3.06242271435297e-005		4.51393486267755		0.254174644857863
21	SNP_A-8607724		3.16644000315045e-005		4.49942873645319		0.25371001313334
22	SNP_A-8454246		3.23825189901636e-005		4.4896893711763		-0.25386994330706
23	SNP_A-8331724		3.56744888978842e-005		4.44764224022925		0.252044947359688
24	SNP_A-8288234		3.89815099511941e-005		4.40914134246902		-0.250800548555345
25	SNP_A-2296771		4.18909851647735e-005		4.37787942606528		0.249785989402922
26	SNP_A-8573561		5.26812874119799e-005		4.27834362040185		-0.247452154886071
27	SNP_A-8460199		5.27510469029514e-005		4.27776891688973		0.246511632851483
28	SNP_A-4287427		5.63185524172114e-005		4.24934851660204		0.245574882558495
29	SNP_A-8435839		5.69832170722038e-005		4.244253035664		0.245406591476520

Figure 2: Association Test results, sorted ascending on p-value, top markers selected

Map	CELFILE	SNP_A-8479801	SNP_A-8631489	SNP_A-8475540
1	NAO6985_GW6_C	B_B	B_B	A_B
2	NAO6991_GW6_C	B_B	B_B	B_B
3	NAO6993_GW6_C	B_B	A_B	B_B
4	NAO6994_GW6_C	B_B	A_B	B_B
5	NAO7000_GW6_C	B_B	B_B	B_B
6	NAO7019_GW6_C	B_B	B_B	B_B
7	NAO7022_GW6_C	B_B	A_B	B_B
8	NAO7029_GW6_C	B_B	A_B	B_B
9	NAO7034_GW6_C	B_B	A_B	B_B
10	NAO7048_GW6_C	B_B	A_B	B_B
11	NAO7055_GW6_C	B_B	B_B	B_B
12	NAO7056_GW6_C	B_B	B_B	B_B
13	NAO7345_GW6_C	B_B	B_B	B_B

Figure 3: Only top markers are active in the Phenotype + Genotype spreadsheet

Next, the allele intensities need to be imported.

- 1) Open the Affymetrix CEL import window by going to **Import > Affymetrix > CEL**.
- 2) Select the CEL files using the File and Directory choosers. See **Figure 4** for an example of importing some HapMap data from Affymetrix 6.0 CEL files.

Import Options

Platform: GenomeWideSNP_6

Import Options

Reference Set

All samples

Defined subset from spreadsheet

0=Ref 1=Non-Ref Column:

Don't include reference samples in output spreadsheet

Precomputed populations

Marker map:

CDF File:

Specify temp directory

Datasets to Import

A/B intensities Quantile normalized A/B intensities

Unmerged LogRs in separate channels

LogR ratios - samples as columns LogR ratios - samples as rows

Dataset name base:

Figure 4: Affymetrix CEL file import window, settings for Affymetrix 6.0 CEL files

- 3) **IMPORTANT:** Notice that the option “**Quantile normalized A/B intensities**” is checked. This is a crucial step to import the allele A and B intensities. Otherwise only the log ratio data will be imported.

- After the CEL files are imported, the project should contain the CEL dataset and the normalized intensity data. See **Figure 5**. In the case of Affymetrix 500k datasets, there will be two Quantile Normalized SNP datasets, one for NSP and one for STY arrays. These two sheets can either be joined together or treated separately.

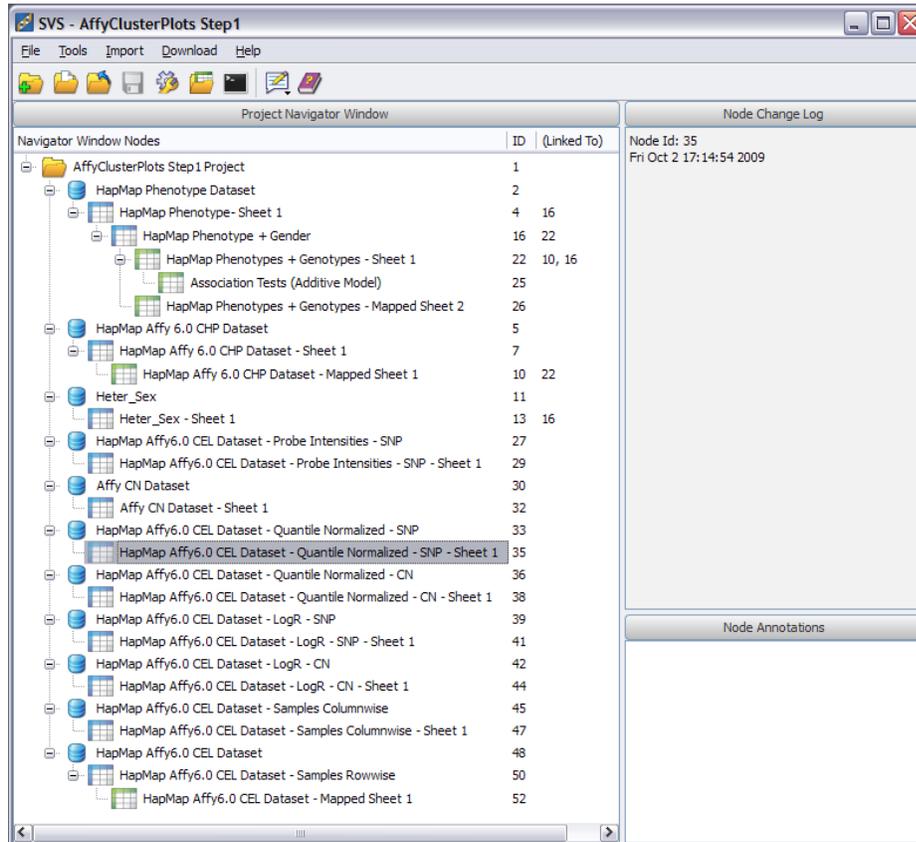


Figure 5: Project after importing CEL files including intermediate steps

Using the Function

- To use the function, go to **Tools > SNP Cluster Plots**.
- Select the genotype spreadsheet that has up to 100 markers active.
- Next, select the Quantile Normalized SNP spreadsheet.
- The Quantile Normalized SNP spreadsheet will be transformed from two columns per sample (first column for A allele intensity, second column for B allele intensity) to two columns per marker, (first column for A allele intensity, second column for B allele intensity) column headers labeled accordingly (see Error! Reference source not found.). XY scatter plots will also be generated for each marker selected. These plots can be split on genotypes to create up to three colored genotype plots.

- 5) To finish creating a genotype plot, follow the following steps.
- Select and open an XY scatter plot (**Figure 6**).

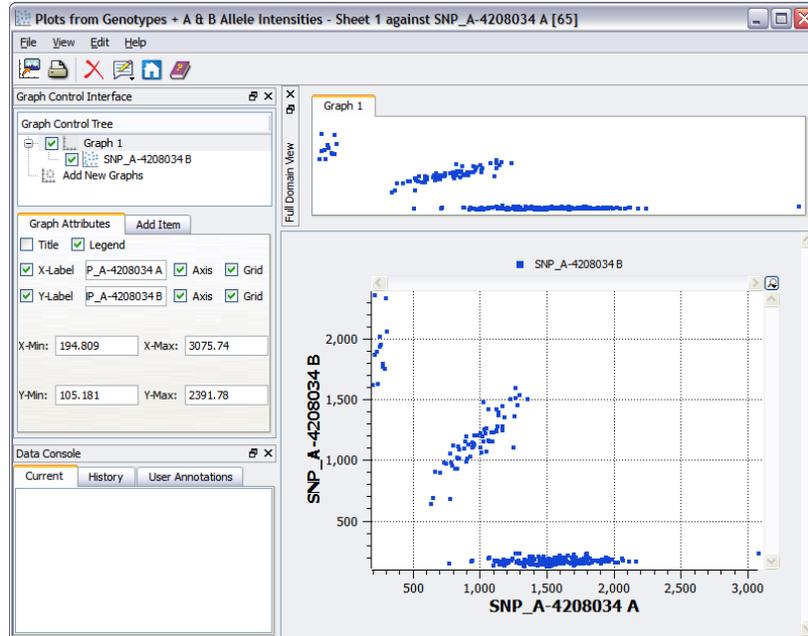


Figure 6: XY scatter plot of allele A & B intensities for SNP_A-4208034

- Select the graph item labeled SNP_A-4208034 B, and click on the **Filter** tab, **Figure 7**.

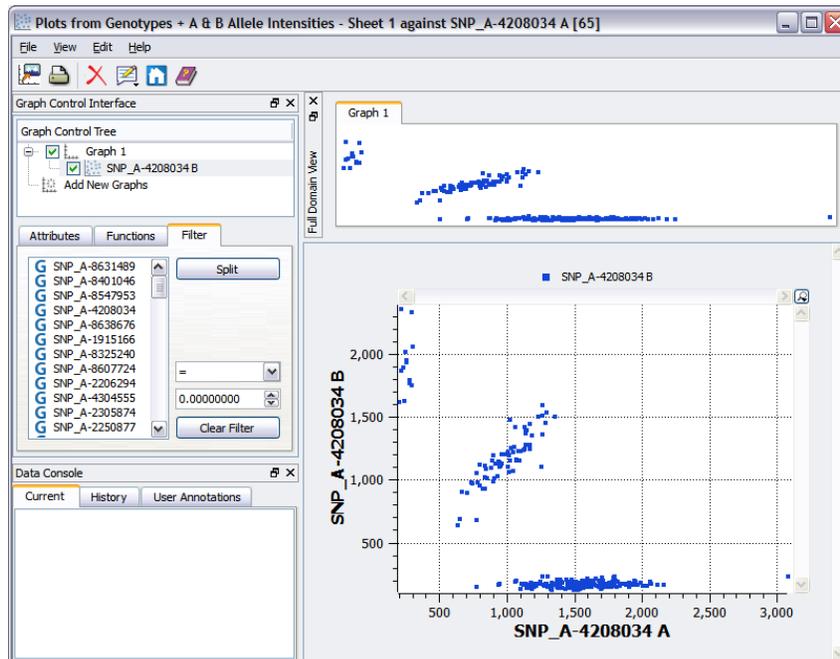


Figure 7: Graph item with Filter tab selected

- c. Next, select SNP_A-4208034 in the filter list, make sure that there is a blue G next to the item selected, and click **Split**, **Figure 8**.

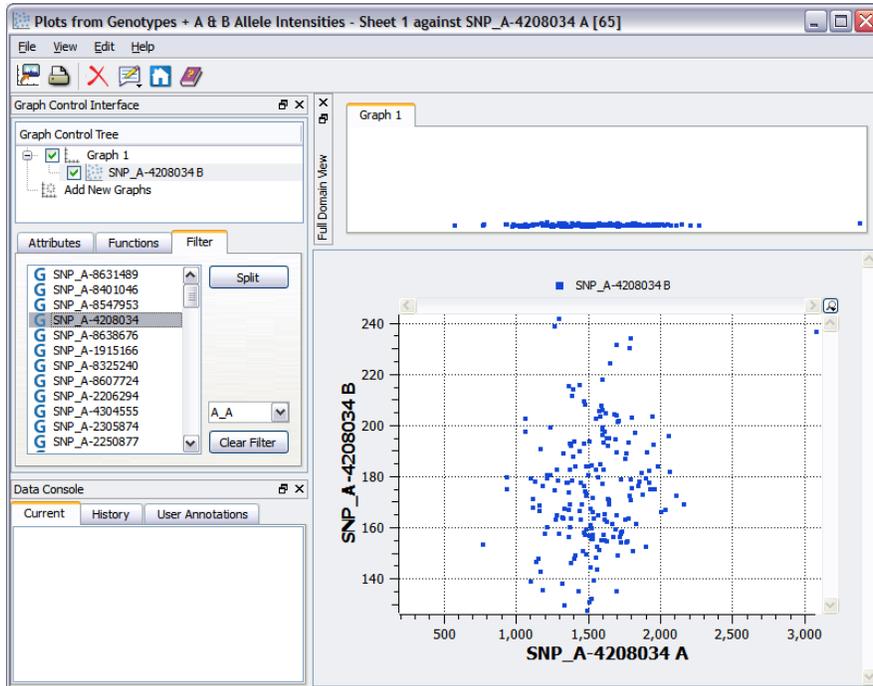


Figure 8: A & B Allele intensity scatter plot before clicking "Split" on the same marker genotypes

The resultant plot will have a different color for each genotype found for all samples for the selected marker. The plot can be edited to change the X and Y axis labels, the legend labels, and to add a title to the graph. The plot can be printed or saved by going to **File > Save as Image** or **Print**.

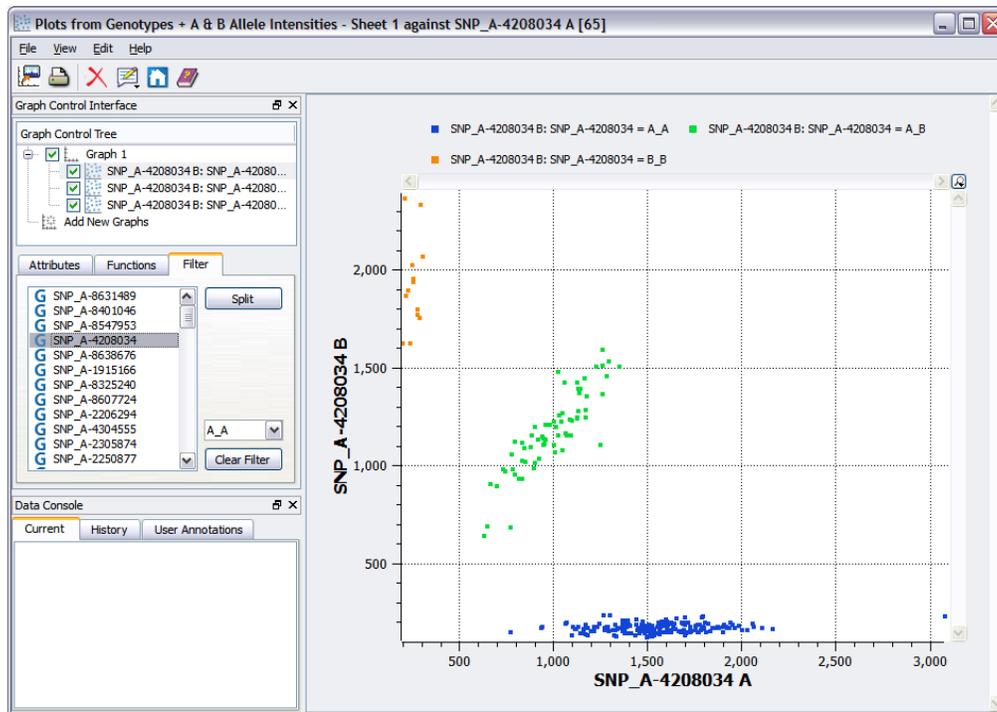


Figure 9: Genotype Cluster Plot for SNP_A-4208034