Affymetrix B Allele Frequency Calculation

Author: Greta Peterson, Golden Helix, Inc.

Overview

Using Affymetrix CEL files as its source, this script combines quantile normalized SNP A and B probe intensities for each marker into a theta value, then calculates B-Allele Frequencies for each marker.

For information on accessing B Allele Frequencies for Illumina data, see http://www.goldenhelix.com/Downloads/SVS/ExportingDSF-GenomeStudio_v4.doc

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 XP and Vista. The easiest way to locate this directory on your computer is to open SVS and select **Tools >Open Folder > User Scripts Folder**. If saved to the proper folder, this script will be accessible from the project navigator's Tools menu.

Preparing to use the Script

This script requires that CEL files be imported into SVS with intermediate spreadsheets. Follow the instructions below to obtain these intermediate files.

1. From an SVS project, a reference spreadsheet needs to be imported to use when importing CEL files unless all samples are used as reference, or one of the provided reference datasets are used.

   If a reference spreadsheet is used, then the content of the reference spreadsheet depends on the Affymetrix Array type.
   a. For combining 250K NSP and 250K STY arrays, matching columns are needed. The row labels do not matter, but will be used in the spreadsheets on import of the data. The second column is the CEL file name of the NSP array for the sample. The third column is the CEL file name of the STY array for the sample. Subsequent columns can contain phenotype information.
   b. For other Affymetrix arrays, the row labels need to match the CEL file names (with or without the .CEL extension), and one of the subsequent columns needs to be binary.

2. To import the CEL files, go to **Import > Affymetrix > CEL**. Select the CEL files. For combined Affymetrix 250K NSP and 250K STY arrays, select both sets of files. Click **Next**.

3. Choose the appropriate parameters in the Import CEL dialog, including selecting the **Quantile Normalized A/B intensities** under **Output Options**.

4. Click **Finish**. The CEL file import process might take a while, and when it is done the chosen spreadsheets are created. See **Figure 1** below. One spreadsheet should have "Quantile Normalized – SNP" in its name or for 500k data there should be two spreadsheets one with "Quantile Normalized – NSP" in its name, the other with "Quantile Normalized – STY" in its name. These are the spreadsheets to use for the B-Allele Frequency calculation.

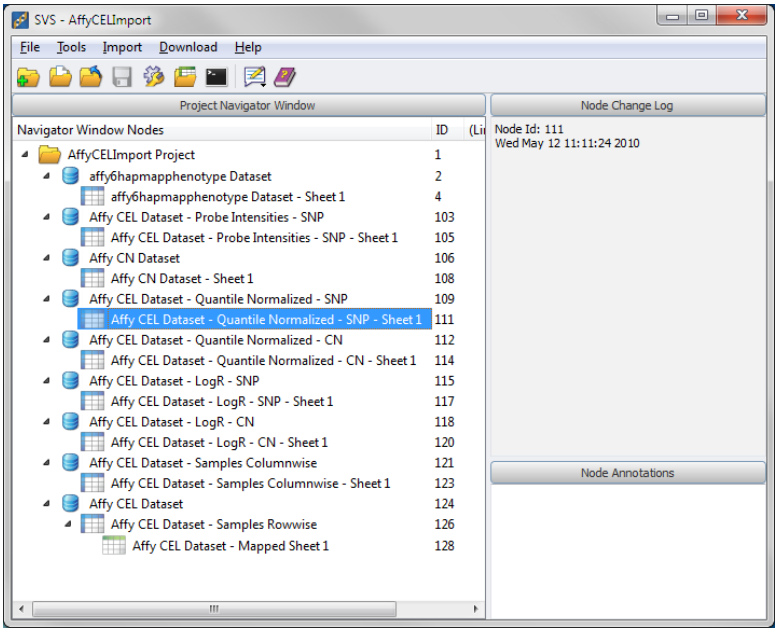

Figure 1: SVS Project Navigator after importing Affymetrix 6.0 CEL files with intermediate files

Using the Script

1. Go to **Tools > Affymetrix B Allele Frequency Calculation**.

2. You will be asked to select the spreadsheet containing quantile normalized A and B intensities and marker map. Select the appropriate spreadsheet and marker map file.

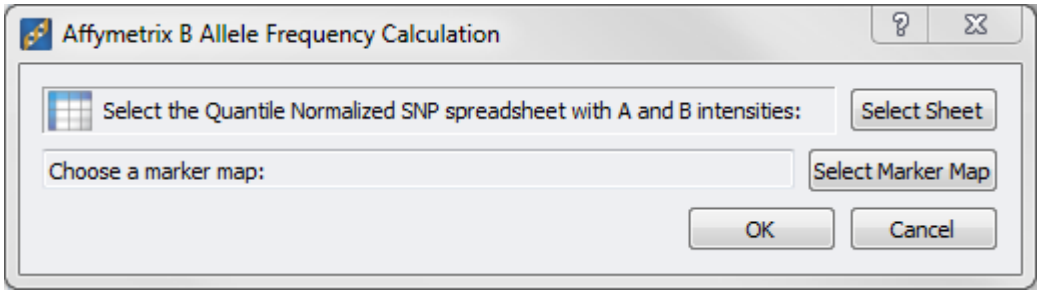

Figure 2: Prompt to select the spreadsheet with A & B intensities and Marker Map

3. Selecting a marker map is optional. If you do not select a marker map, then the final spreadsheets will not be marker mapped. Otherwise, select a marker map and press **OK**.

How the Script Works

The script will first use the A (X coordinate) and B (Y coordinate) intensities to calculate the theta coordinate in a polar coordinates transformation. The R coordinate is not needed for the B Allele Frequency computation. The theta coordinates are output in a spreadsheet for later reference. Plotting histograms of theta coordinates per SNP can yield important clustering information.

The formula (Wang, 2008) for the theta coordinate of the i^{th} sample and j^{th} marker is defined as:

$$\theta_{ij} = \frac{\arctan\left(Y_{ij}/X_{ij}\right)}{\pi/2}$$

Next, the script will use the theta values and the assumption that ideally the AA cluster will be near a theta value of 0.1, the AB cluster will be near a theta value of 0.5 and the BB cluster will be near a theta value of 0.9 to calculate approximate cluster means for each marker($\theta_{AA}, \theta_{AB}, \theta_{BB}$). These cluster means will be used to calculate the B-Allele Frequency with the following formula (Wang, 2008):

$$B_{ij} = \begin{cases} 0 & if\ \theta_{ij} < \theta_{AA} \\ \dfrac{0.5(\theta_{ij} - \theta_{AA})}{\theta_{AB} - \theta_{AA}} & if\ \theta_{AA} \le \theta_{ij} < \theta_{AB} \\ 0.5 + \dfrac{0.5(\theta_{ij} - \theta_{AB})}{\theta_{BB} - \theta_{AB}} & if\ \theta_{AB} \le \theta_{ij} < \theta_{BB} \\ 1 & if\ \theta_{BB} \le \theta_{ij} \end{cases}$$

The B-Allele Frequencies are output in a spreadsheet with markers in columns and samples in rows (see **Figure 3**). Then this spreadsheet is transposed to have samples in columns and markers in rows (see **Figure 4**) for plotting of the B-Allele Frequencies for particular samples (see **Figure 5**).

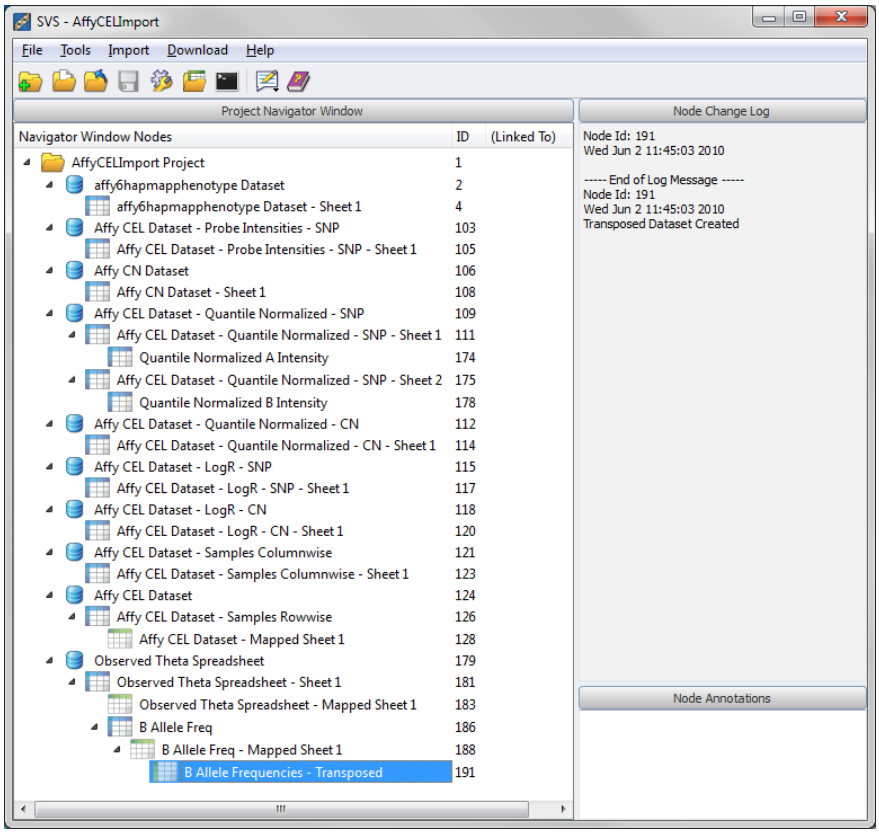

Figure 3: Project navigator after running script

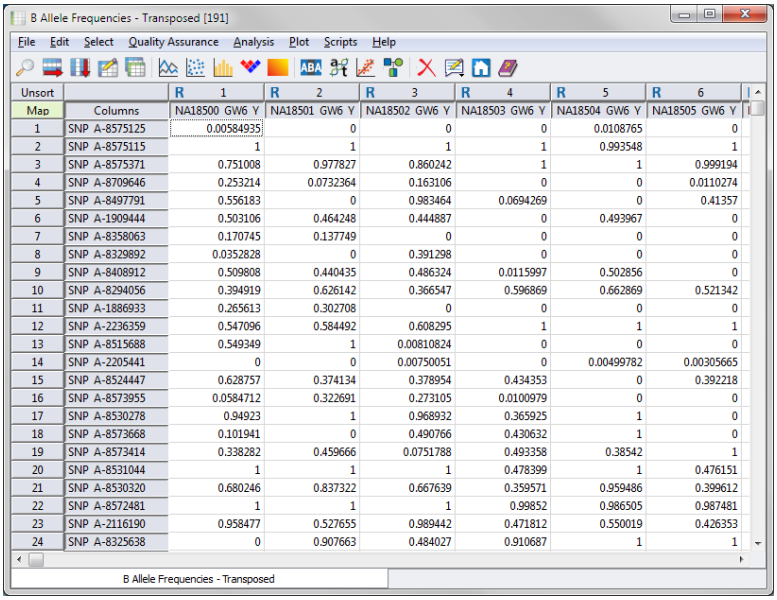

Figure 4: B Allele Frequency - Transposed spreadsheet, ready for plotting

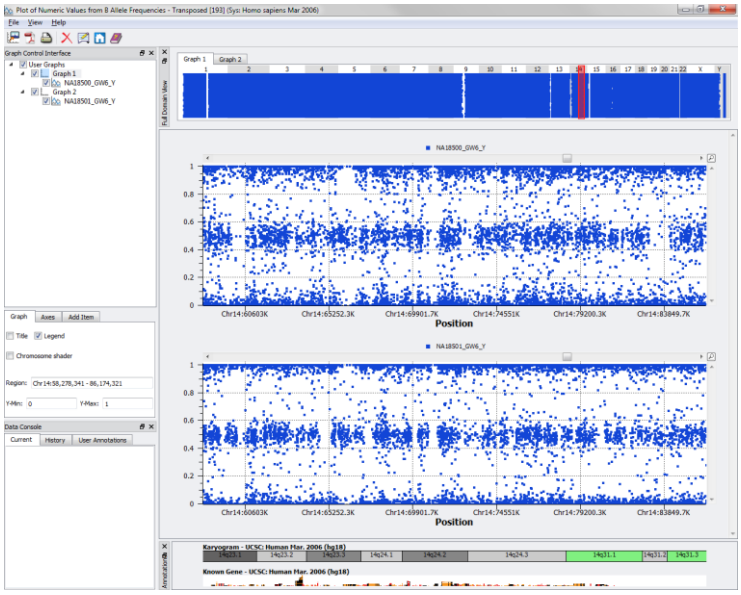

Figure 5: Two B-Allele Frequency plots

Works Cited

Wang, K. a. (2008). Copy Number Variation Detection via High-Density SNP Genotyping. *Cold Spring Harb. Protoc.*; doi:10.1101/pdb.top46.