Import Tall Skinny Format

Authors: Autumn Laughbaum, James Grover, Golden Helix, Inc.

Overview

This import script is designed to import genotypic data that is stored in a tall skinny format. The user will specify the sample name column, the variant name column, and the data column(s), as well as the data ordering.

Recommended Directory Location

Save the script to the following directory:
*..**\Application Data\Golden Helix SVS\UserScripts\SVS\Import**

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 > UserScripts Folder**. If saved to the proper folder, this script will be accessible from the project navigator's **Import** menu.

Using the Script

1. From the project navigator, choose **Import >Tall Skinny**. Select the appropriate file, choose a Base Dataset Name (that is, the name of the eventual spreadsheet node), the file delimiter and the missing value indicator.

2. If the file contains a header row (anywhere in the file, not necessarily the first line), check this option. If there are several sections in your file and the genotype data does not start immediately, enter a regular expression that matches the genotype section header. Otherwise uncheck this option. Click **OK**.

3. In the next dialog, you will specify the marker, sample and genotype (or allele) columns, as well as the data ordering. The drop-down lists contain the headers (if applicable, otherwise column numbers) and a few examples of the data found in that column.

    a) **Choose sample column:** Specify at least one sample name column. Optionally choose a second column and delimiter (to be inserted upon import) to identify unique sample names.
    b) **Choose marker column:** Specify at least one marker name column. Optionally choose a second column and a delimiter (to be inserted upon import) to identify unique marker names.
    c) **Data ordering:** Specify whether the data pertaining to an individual sample is grouped together or, on the other hand, whether the data pertaining to an individual marker is grouped together. **NOTE:** Sorting the data is one way to

group the data—for instance, sorting by sample is one way to group the data for each sample together—but sorting is not necessary for this application to work—only grouping the data, either a group for each sample or a group for each marker.

d) **Alleles in one column:** Check this option if the genotypes are found in one column with or without a delimiter. If there is no delimiter (AA, AB, etc.), each allele is expected to be one character long.

e) **Alleles in two columns:** Check this option if the alleles are found in separate columns.

4. Click **OK** to import the file.