Import Unsorted VCF Files

Author: Sam Gardner, James Grover, Gabe Rudy, Golden Helix, Inc.

Overview

This script will import 1000 Genomes .vcf file date into multiple spreadsheets and/or marker map fields. All resulting genotype information is unphased.

This has been tested successfully on well formatted VCF input from version 4.1, 4.0, 3.3, and 3.2 of the 1000genomes.org spec: http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

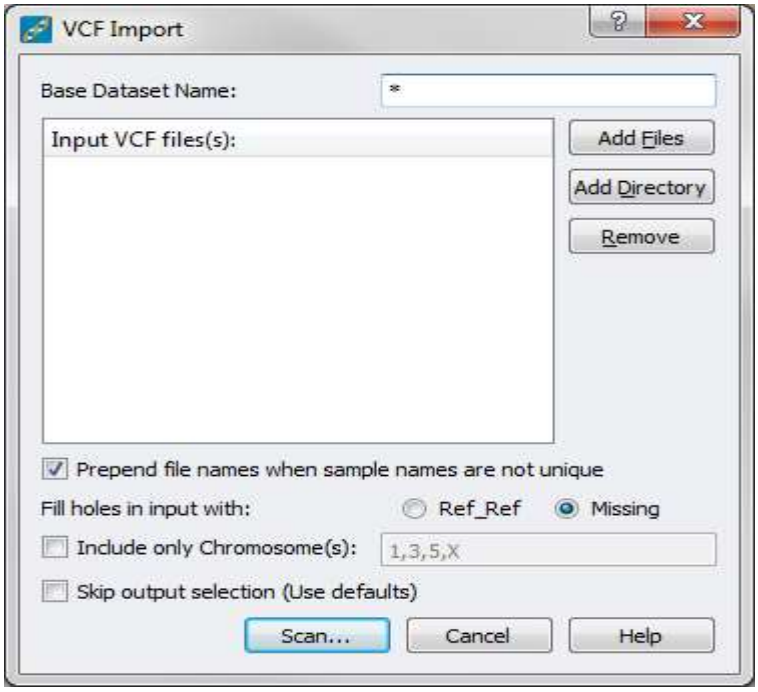
Recommended Directory Location

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

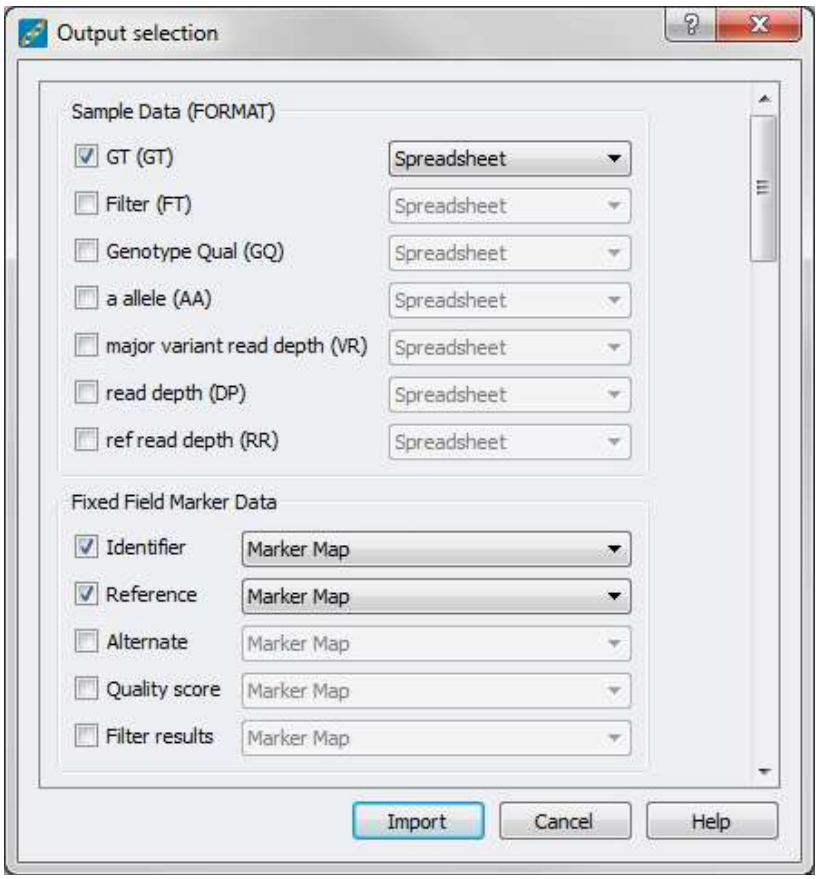
Note: The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between XP and Vista. The easiest way to locate this directory on your computer is to open SVS and select the **Tools > Open Folder > User Scripts Folder** menu option. If saved to the proper folder, this script will be accessible from the project navigator **Import** menu.

Using the Script

From an open project select **Import** > **Import Unsorted VCF Files**



Options for the import include the base dataset name. If specified this will define the naming prefix for each dataset created by the import. If the default value of "*" is used, the dataset will take the name of the first file in the input list. By default, sample names found to be duplicated in multiple files are disambiguated by pre-pending a part of the source file name. This option can be disabled by preference. The user may also specify whether holes in the input are filled with data indicating either homozygous reference or a missing genotype. Holes in the input occur when one of the input samples has data for a particular chromosome and position, yet another does not. The import may be limited to a subset of the chromosomes provided in the input files by activating the "Include only Chromosome(s)" option and typing in a comma delimited list of the chromosomes for which data is to be imported. The second dialog (output selection) may also be skipped at the user's option.



The second dialog will appear after scanning the input files for available data to be imported. Depending on the available data, fields in four categories may be presented for selection. The categories are:

- Sample Data (FORMAT) - Any format data fields provided in the VCF(s)
- Fixed Field Marker Data - Data from predefined VCF fields

- Computed Marker Data - Data which can be computed from other fields during import
- Marker Data (INFO) - Any info data fields provided in the VCF(s)

Some selections allow the user to choose which output to generate. Possible output types include:

- Spreadsheet - A spreadsheet containing data for a grid of samples vs chromosome and position
- Marker Map - A field added to the marker map which is applied to the columns of each spreadsheet
- Both - Both a spreadsheet and a marker map field

The “Reference” field available under “Fixed Field Marker Data” is selected as a marker map field by default and, for best utility in further analysis, should remain so.

The variant type information is appended to each column in the resulting spreadsheet(s). The following variant abbreviations are added to the column headers:

- *Ins*: Insertion
- *Del*: Deletion
- *Sub*: Substitution
- *SNV*: Single Nucleotide Variation