Import Sorted VCF Files

Author: Gabe Rudy, Golden Helix, Inc.

Overview

This function imports 1000 Genomes .vcf file data into multiple spreadsheets. Special handling is provided for genotype data. The user can choose to import one VCF file or several VCF files simultaneously.

This import tool has been tested successfully on well-formed VCF input from versions 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

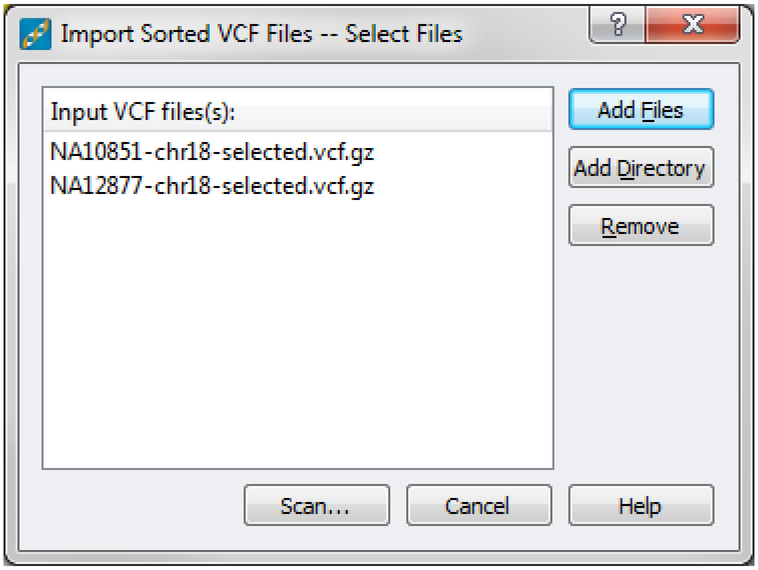
Recommended Directory Location

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 Sorted VCF Files.**



Import Sorted VCF Files -- Select Files

Input VCF files(s):

NA10851-chr18-selected.vcf.gz
NA12877-chr18-selected.vcf.gz

Select the VCF or VCF.GZ files to import. If both VCF and VCF.GZ files are available in the same directory, the bgzipped files will be preferred over the uncompressed files. All other files in the directory that are not *.vcf or *.vcf.gz files will be ignored.

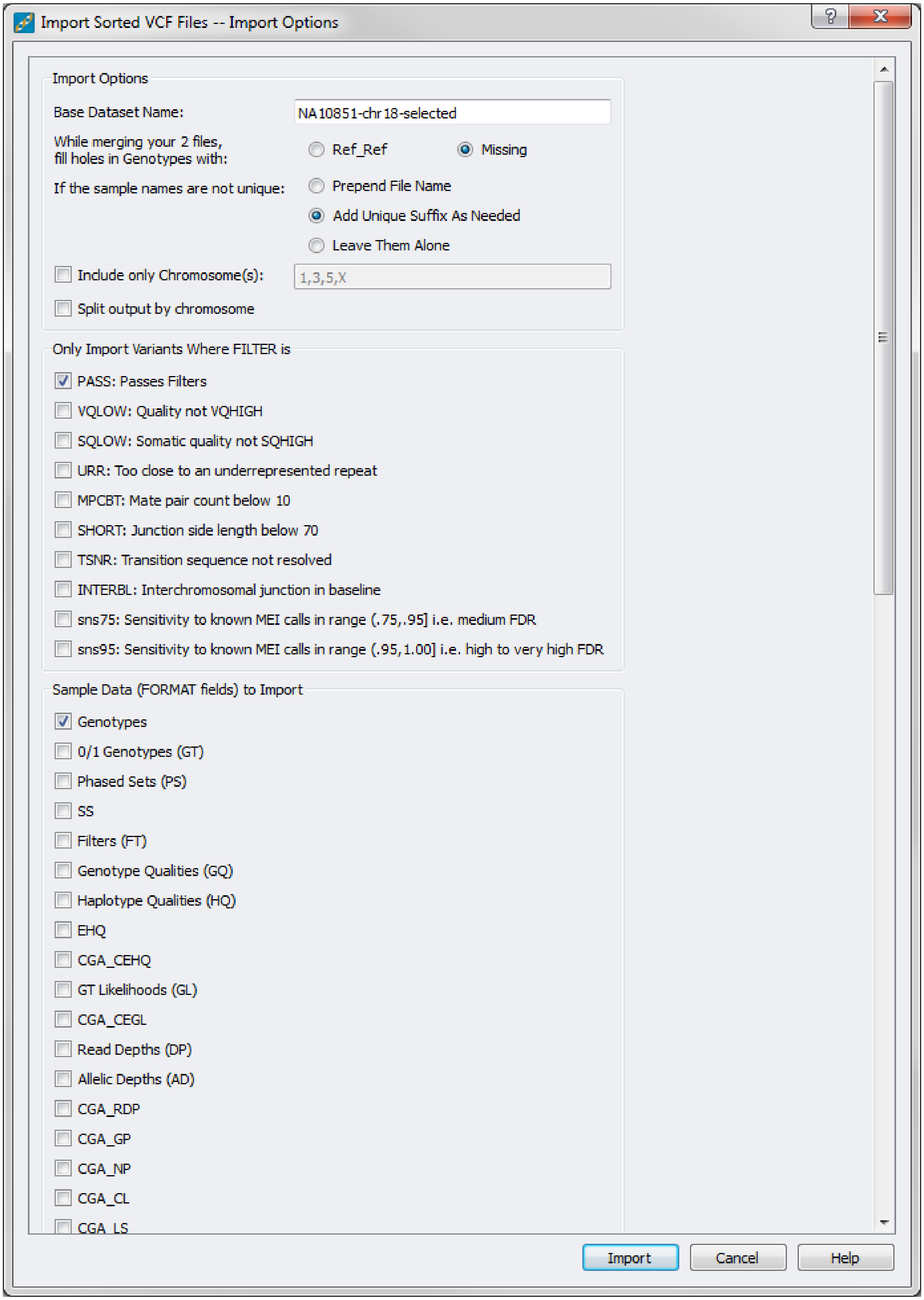
After the desired files have been added click **Scan...**

During the scan step the following happens:

- The directory is examined looking for *.vcf.gz files of the same name as selected *.vcf files.
- If compressed files are found they are examined to make sure they are compressed using bgzip.
- If no compressed files are found or if they are determined to be the wrong format the *.vcf files are compressed using bgzip. You are given the option on where to place the *.vcf.gz files. By default the directory is the same as where the first *.vcf file was found.
- After the files are compressed the directory is examined for indexed files as indicated by the extension TBI. If TBI files are not found, the VCF files are indexed with Tabix.

> **Note**: Tabix does require that the VCF files be sorted, if it appears that the files are not sorted it will present an error indicating as such. A third-party sorting tool will need to be used or you may use the script **Import Unsorted VCF Files** to import your data.

After the files have been scanned, compressed and indexed or the compressed and indexed files have been identified, a dialog opens with import parameters.



Import Sorted VCF Files -- Import Options

Import Options

Base Dataset Name: NA10851-chr18-selected

While merging your 2 files, fill holes in Genotypes with: ○ Ref_Ref ● Missing

If the sample names are not unique:
○ Prepend File Name
● Add Unique Suffix As Needed
○ Leave Them Alone

□ Include only Chromosome(s): 1,3,5,X

□ Split output by chromosome

Only Import Variants Where FILTER is

☑ PASS: Passes Filters
□ VQLOW: Quality not VQHIGH
□ SQLOW: Somatic quality not SQHIGH
□ URR: Too close to an underrepresented repeat
□ MPCBT: Mate pair count below 10
□ SHORT: Junction side length below 70
□ TSNR: Transition sequence not resolved
□ INTERBL: Interchromosomal junction in baseline
□ sns75: Sensitivity to known MEI calls in range (.75,.95] i.e. medium FDR
□ sns95: Sensitivity to known MEI calls in range (.95,1.00] i.e. high to very high FDR

Sample Data (FORMAT fields) to Import

☑ Genotypes
□ 0/1 Genotypes (GT)
□ Phased Sets (PS)
□ SS
□ Filters (FT)
□ Genotype Qualities (GQ)
□ Haplotype Qualities (HQ)
□ EHQ
□ CGA_CEHQ
□ GT Likelihoods (GL)
□ CGA_CEGL
□ Read Depths (DP)
□ Allelic Depths (AD)
□ CGA_RDP
□ CGA_GP
□ CGA_NP
□ CGA_CL
□ CGA_LS

VCF Import Window: Select Import Options (example of 2 Complete Genomics HapMap 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 is used, the dataset will take the name of the first file in the input list.

If two or more files are selected for import, the options of how to handle "holes" or locations where one file has a variant but the other file(s) do not include fill with Ref_Ref or fill with Missing. This is only applicable to the Genotype spreadsheet. All other spreadsheets are always filled with missing values.

If desired, only a subset of chromosomes can be imported. This can be specified by using the *Include only Chromosome(s)* flag and specifying the chromosomes in a comma separated list.

When importing whole genome data for numerous samples it can take a considerable amount of RAM to work with the spreadsheets if all of the chromosomes are included in the same spreadsheet. In this case it is recommended that you select *Split output by chromosome*.

If the VCF files have filter flags specified then it is possible to only import variants that meet a certain filter. If the filter information is missing for a variant it is always imported. To import all variants, check all of the filter boxes.

Sample level data is always imported into a spreadsheet. Any number of spreadsheets may be selected. To reduce the length of time it takes to import your data, only select the sample data fields necessary for your analysis. Adding more spreadsheets can cause the importer to take much longer and cause your project to be of considerable size.

Variant or Site level data is always imported into a marker map field.

> **Note**: When importing multiple VCF files, there is the possibility per-variant site INFO fields to differ between files. Depending on the data type of the field, we allow you to specify a merging function (Average, Min, Max, Unique, etc.).

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
* SNV: Single Nucleotide Variation
* MNP: Multi-Nucleotide Polymorphism
* REF: No alternate defined, calling reference
* MIX: A mixture of variation types