Activate ATCG SNPs to exclude SNPs

Author: Joost W. Morsink and Sander W. van der Laan, University Medical Center Utrecht, the Netherlands

Overview

This script can be used to identify SNPs that have ambivalent orientation by comparing a genotype dataset with a reference dataset, such as HapMap data. First the script will identify all A/T, T/A, C/G and G/C SNPs where the minor allele in your dataset differs from the minor allele in the reference dataset. Then it will further identify which of those SNPs have a minor allele frequency greater than or equal to a specified threshold in your dataset. These SNPs will be activated and an active subset will be created. The subset can be used to exclude the SNPs from the analysis, which had strand orientation that was hard to determine.

Recommended Directory Location

Save the script to the following directory:
*..**\Application Data\Golden Helix SVS\UserScripts\Spreadsheet\Quality Assurance\QA Scripts**

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 go to **Tools > Open Folder > User Scripts Folder**. If saved to the proper folder, this script will be accessible from the spreadsheet **Quality Assurance** menu.

Using the Script

To use this script you need genotype data, coded ACTG, as well as reference data such as that from HapMap samples. First you need to run Genotype Statistics by Marker on each spreadsheet then merge the two Marker Statistics spreadsheets together.

1. From a spreadsheet with genotype columns coded ACTG, choose **Quality Assurance > Genotype >Genotype Statistics by Marker**. Make sure that only the box preceding **Allele frequencies** is checked and click **Run**.

2. Repeat step 1 using the reference spreadsheet. Make sure that you will be able to distinguish between these spreadsheets when you merge them.

3. Open the **Marker Statistics** spreadsheet that corresponds to your data (from Step 1). Choose **File >Join or Merge Spreadsheets** and then select the **Marker Statistics** spreadsheet corresponding to the reference data. It is important that you do the

merging in this order because the first minor allele frequency columns will be used for filtering and you want to filter based on your sample rather than the reference samples.

4. Leave the default options for the merge and click **OK**. Now from your merged spreadsheet (**Marker Statistics + Marker Statistics** if you did not rename the spreadsheets), choose **Quality Assurance >QA Scripts >Activate ATCG SNPs to exclude SNPs**.

5. Choose an appropriate threshold, such that (MAF >= threshold) will activate the SNPs for exclusion and click **OK**.

A spreadsheet will be created that contains the subset of SNPs with high minor allele frequencies that have ambivalent orientation. You could now choose to exclude these SNPs from further analysis.