Sample Pair Mismatch

Author: Christophe Lambert and Greta Linse Peterson, Golden Helix, Inc.

Overview

This script compares genotype calls from NSP and STY files and calculates the correlation between the nearest markers in the two sets. If there is a high correlation, the NSP and STY markers correspond to the same person, otherwise there is a mismatch.

Recommended Directory Location

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

Note: The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between operating systems. The easiest way to locate this directory on your computer is to open SVS and go to **Tools >Open Folder > UserScripts Folder** and save the script in the **\SVS\Spreadsheet\Analysis** folder. If you save the script to the proper folder, it will be accessible from the spreadsheet **Analysis** menu.

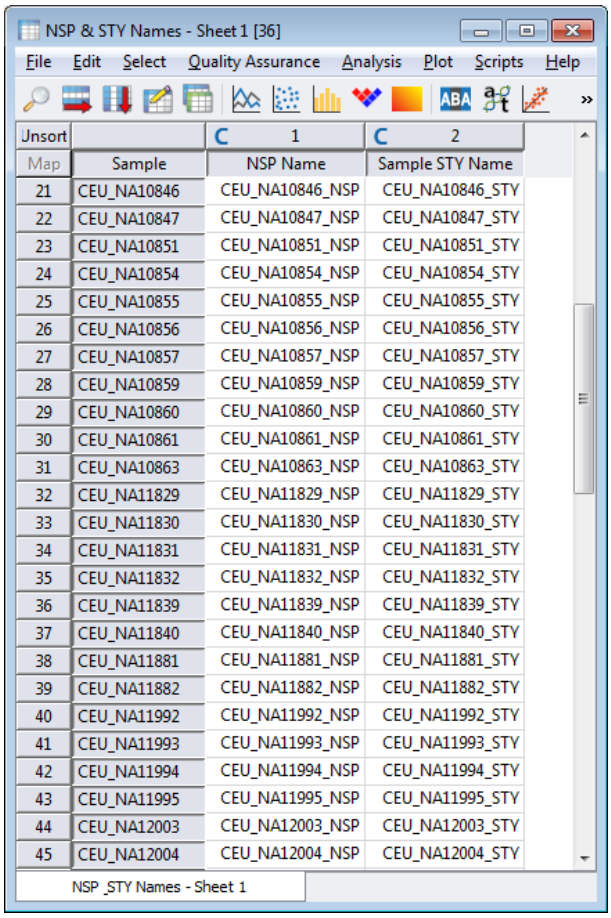
Obtaining the Required Datasets

In order to compare the NSP and STY genotypes you need the following items:

1. NSP genotypes for all samples with a marker map applied to the spreadsheet.

2. STY genotypes for all samples with a marker map applied to the spreadsheet.

3. A matching spreadsheet with a common sample name as the row labels and NSP file names in the first column and STY file names in the second column. See **Figure 1** for an example of a matching spreadsheet.



	Sample	NSP Name	Sample STY Name
21	CEU_NA10846	CEU_NA10846_NSP	CEU_NA10846_STY
22	CEU_NA10847	CEU_NA10847_NSP	CEU_NA10847_STY
23	CEU_NA10851	CEU_NA10851_NSP	CEU_NA10851_STY
24	CEU_NA10854	CEU_NA10854_NSP	CEU_NA10854_STY
25	CEU_NA10855	CEU_NA10855_NSP	CEU_NA10855_STY
26	CEU_NA10856	CEU_NA10856_NSP	CEU_NA10856_STY
27	CEU_NA10857	CEU_NA10857_NSP	CEU_NA10857_STY
28	CEU_NA10859	CEU_NA10859_NSP	CEU_NA10859_STY
29	CEU_NA10860	CEU_NA10860_NSP	CEU_NA10860_STY
30	CEU_NA10861	CEU_NA10861_NSP	CEU_NA10861_STY
31	CEU_NA10863	CEU_NA10863_NSP	CEU_NA10863_STY
32	CEU_NA11829	CEU_NA11829_NSP	CEU_NA11829_STY
33	CEU_NA11830	CEU_NA11830_NSP	CEU_NA11830_STY
34	CEU_NA11831	CEU_NA11831_NSP	CEU_NA11831_STY
35	CEU_NA11832	CEU_NA11832_NSP	CEU_NA11832_STY
36	CEU_NA11839	CEU_NA11839_NSP	CEU_NA11839_STY
37	CEU_NA11840	CEU_NA11840_NSP	CEU_NA11840_STY
38	CEU_NA11881	CEU_NA11881_NSP	CEU_NA11881_STY
39	CEU_NA11882	CEU_NA11882_NSP	CEU_NA11882_STY
40	CEU_NA11992	CEU_NA11992_NSP	CEU_NA11992_STY
41	CEU_NA11993	CEU_NA11993_NSP	CEU_NA11993_STY
42	CEU_NA11994	CEU_NA11994_NSP	CEU_NA11994_STY
43	CEU_NA11995	CEU_NA11995_NSP	CEU_NA11995_STY
44	CEU_NA12003	CEU_NA12003_NSP	CEU_NA12003_STY
45	CEU_NA12004	CEU_NA12004_NSP	CEU_NA12004_STY

Figure 1: NSP and STY matching spreadsheet.

Next, the genotypes need to be converted to integers using an additive model.

1) Open the NSP Genotype spreadsheet and convert to integers by going to **Edit > Recode Genotypes** and select **Encode genotypes numerically based on genetic model: Additive model: DD=2, Dd = 1, dd = 0**.

2) Open the STY Genotype spreadsheet and convert to integers by going to **Edit > Recode Genotypes** and select **Encode genotypes numerically based on genetic model: Additive model: DD=2, Dd = 1, dd = 0**.

Using the Script

1) To use the script, open the NSP and STY matching spreadsheet and go to **Analysis > Sample Pair Mismatch.**

2) Choose the columns that contain the NSP and STY CEL names, the numerically-coded NSP and STY spreadsheets, and the maximum distance in base pairs to determine

the markers to compare. The script finds the nearest marker in the STY set less than the threshold for every marker in the NSP set.

3) The NSP and STY correlation will be calculated and output in the NSP STY correlation spreadsheet as a child of the matching spreadsheet. See **Figure 2**.

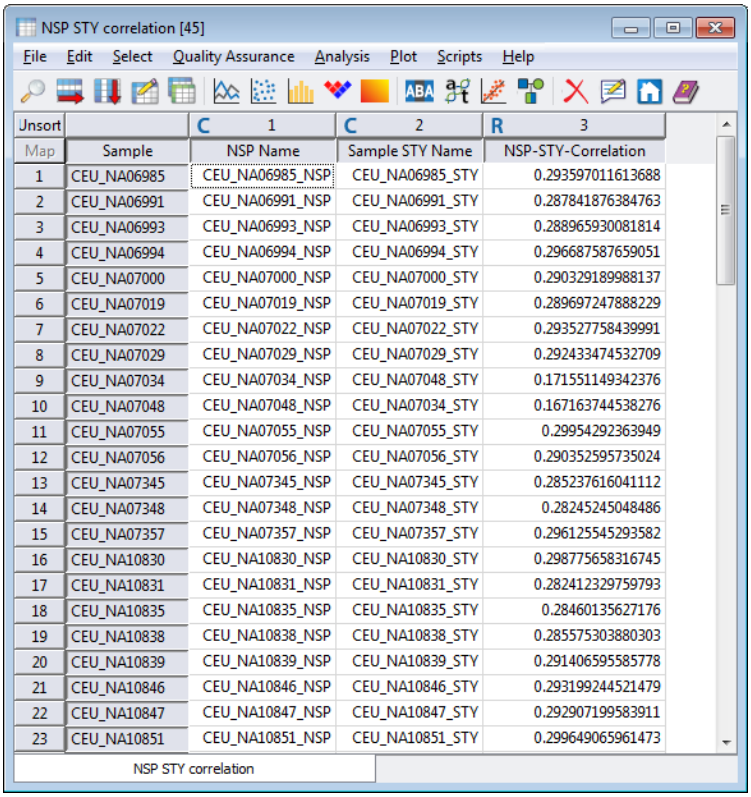
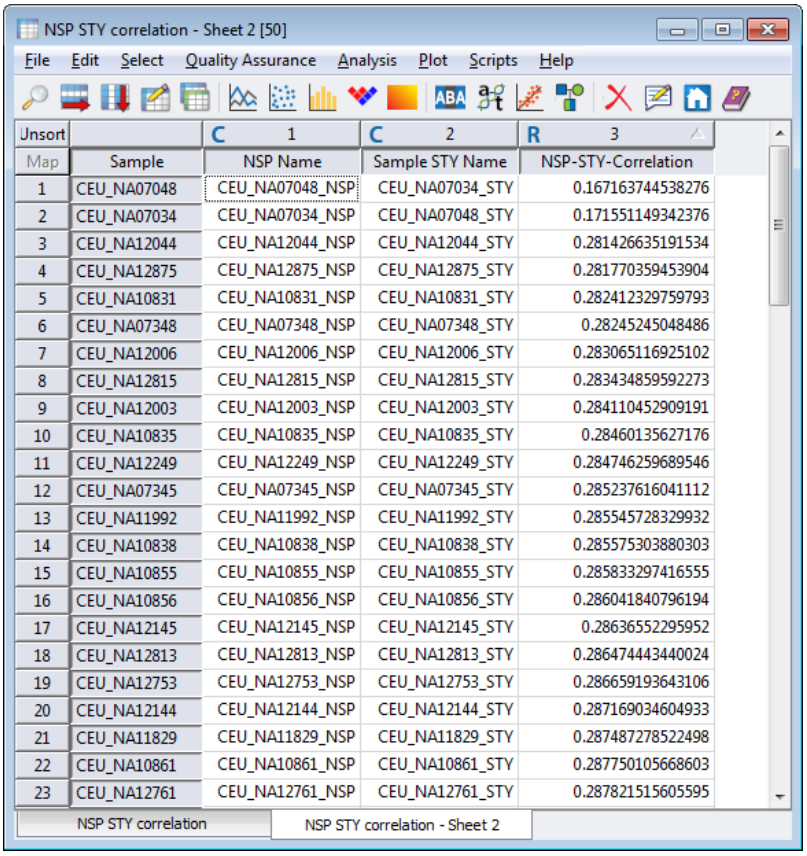


Figure 2: NSP STY correlation spreadsheet

4) Finally, the easiest way to identify pairs that are potential mismatches is to choose **Quality Assurance >Compare Columns** and add the first two columns. Samples that do not match are inactivated. See **Figure 3**. In this case the STY names for CEU_NA07048 and CEU_NA07034 were intentionally switched and you can see that they have the lowest correlation inconsistent with the rest of the correlation values.

5) A histogram is a good visualization tool, but it cannot identify samples that do not fit the distribution. See **Figure 4**.



Unsort		C 1	C 2	R 3
Map	Sample	NSP Name	Sample STY Name	NSP-STY-Correlation
1	CEU_NA07048	CEU_NA07048_NSP	CEU_NA07034_STY	0.167163744538276
2	CEU_NA07034	CEU_NA07034_NSP	CEU_NA07048_STY	0.171551149342376
3	CEU_NA12044	CEU_NA12044_NSP	CEU_NA12044_STY	0.281426635191534
4	CEU_NA12875	CEU_NA12875_NSP	CEU_NA12875_STY	0.281770359453904
5	CEU_NA10831	CEU_NA10831_NSP	CEU_NA10831_STY	0.282412329759793
6	CEU_NA07348	CEU_NA07348_NSP	CEU_NA07348_STY	0.28245245048486
7	CEU_NA12006	CEU_NA12006_NSP	CEU_NA12006_STY	0.283065116925102
8	CEU_NA12815	CEU_NA12815_NSP	CEU_NA12815_STY	0.283434859592273
9	CEU_NA12003	CEU_NA12003_NSP	CEU_NA12003_STY	0.284110452909191
10	CEU_NA10835	CEU_NA10835_NSP	CEU_NA10835_STY	0.28460135627176
11	CEU_NA12249	CEU_NA12249_NSP	CEU_NA12249_STY	0.284746259689546
12	CEU_NA07345	CEU_NA07345_NSP	CEU_NA07345_STY	0.285237616041112
13	CEU_NA11992	CEU_NA11992_NSP	CEU_NA11992_STY	0.285545728329932
14	CEU_NA10838	CEU_NA10838_NSP	CEU_NA10838_STY	0.285575303880303
15	CEU_NA10855	CEU_NA10855_NSP	CEU_NA10855_STY	0.285833297416555
16	CEU_NA10856	CEU_NA10856_NSP	CEU_NA10856_STY	0.286041840796194
17	CEU_NA12145	CEU_NA12145_NSP	CEU_NA12145_STY	0.28636552295952
18	CEU_NA12813	CEU_NA12813_NSP	CEU_NA12813_STY	0.286474443440024
19	CEU_NA12753	CEU_NA12753_NSP	CEU_NA12753_STY	0.286659193643106
20	CEU_NA12144	CEU_NA12144_NSP	CEU_NA12144_STY	0.287169034604933
21	CEU_NA11829	CEU_NA11829_NSP	CEU_NA11829_STY	0.287487278522498
22	CEU_NA10861	CEU_NA10861_NSP	CEU_NA10861_STY	0.287750105668603
23	CEU_NA12761	CEU_NA12761_NSP	CEU_NA12761_STY	0.287821515605595

Figure 3: NSP STY Correlation sorted by correlation values in ascending order.

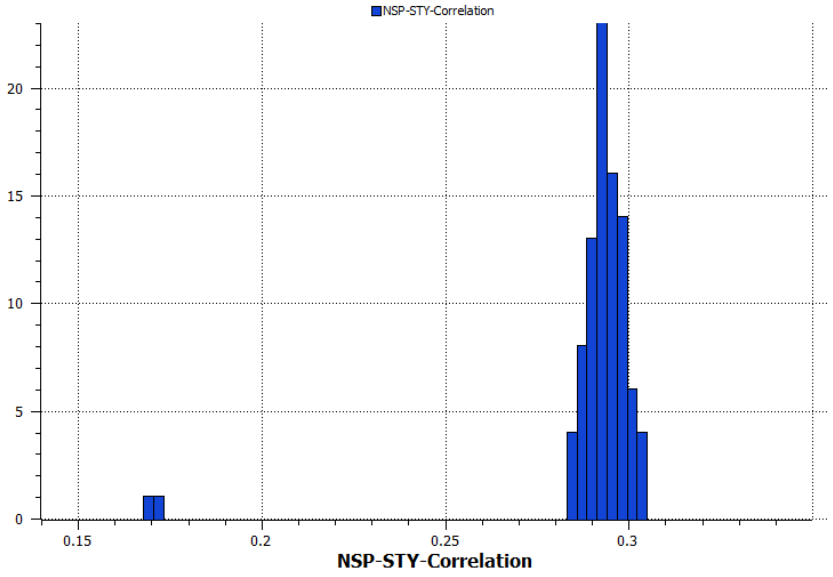


Figure 4: Histogram of NSP STY Correlation Values showing the two samples that are outliers.