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\Numeric\CNV_QA**

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\Numeric\CNV_QA** folder. If you save the script to the proper folder, it will be accessible from the spreadsheet **Numeric > CNV QA** menu.

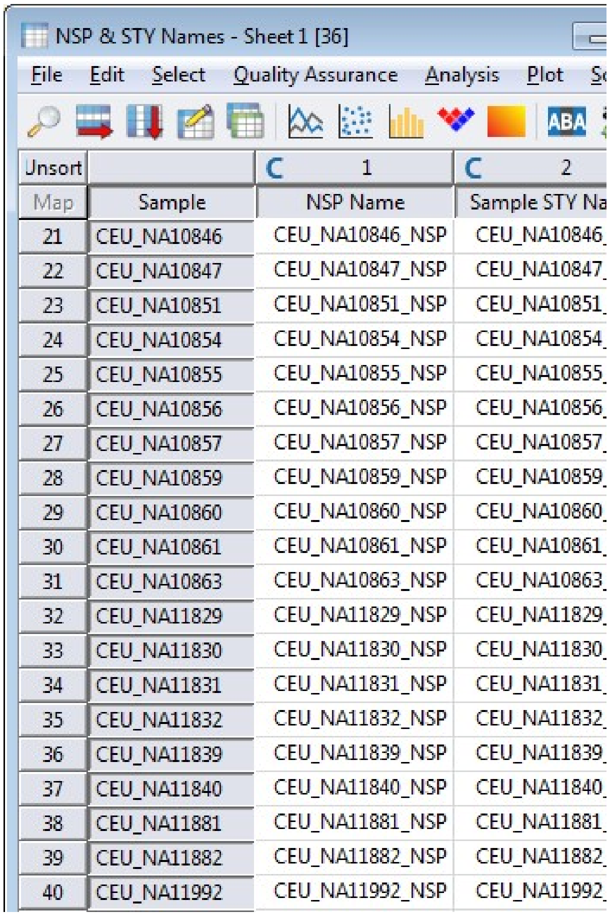
Obtaining the Required Datasets

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.



Unsort		C 1	C 2
Map	Sample	NSP Name	Sample STY Na
21	CEU_NA10846	CEU_NA10846_NSP	CEU_NA10846_
22	CEU_NA10847	CEU_NA10847_NSP	CEU_NA10847_
23	CEU_NA10851	CEU_NA10851_NSP	CEU_NA10851_
24	CEU_NA10854	CEU_NA10854_NSP	CEU_NA10854_
25	CEU_NA10855	CEU_NA10855_NSP	CEU_NA10855_
26	CEU_NA10856	CEU_NA10856_NSP	CEU_NA10856_
27	CEU_NA10857	CEU_NA10857_NSP	CEU_NA10857_
28	CEU_NA10859	CEU_NA10859_NSP	CEU_NA10859_
29	CEU_NA10860	CEU_NA10860_NSP	CEU_NA10860_
30	CEU_NA10861	CEU_NA10861_NSP	CEU_NA10861_
31	CEU_NA10863	CEU_NA10863_NSP	CEU_NA10863_
32	CEU_NA11829	CEU_NA11829_NSP	CEU_NA11829_
33	CEU_NA11830	CEU_NA11830_NSP	CEU_NA11830_
34	CEU_NA11831	CEU_NA11831_NSP	CEU_NA11831_
35	CEU_NA11832	CEU_NA11832_NSP	CEU_NA11832_
36	CEU_NA11839	CEU_NA11839_NSP	CEU_NA11839_
37	CEU_NA11840	CEU_NA11840_NSP	CEU_NA11840_
38	CEU_NA11881	CEU_NA11881_NSP	CEU_NA11881_
39	CEU_NA11882	CEU_NA11882_NSP	CEU_NA11882_
40	CEU_NA11992	CEU_NA11992_NSP	CEU_NA11992_

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 > 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 > Recode Genotypes** and select **Encode genotypes numerically based on genetic model: Additive model: DD=2, Dd = 1, dd = 0**.

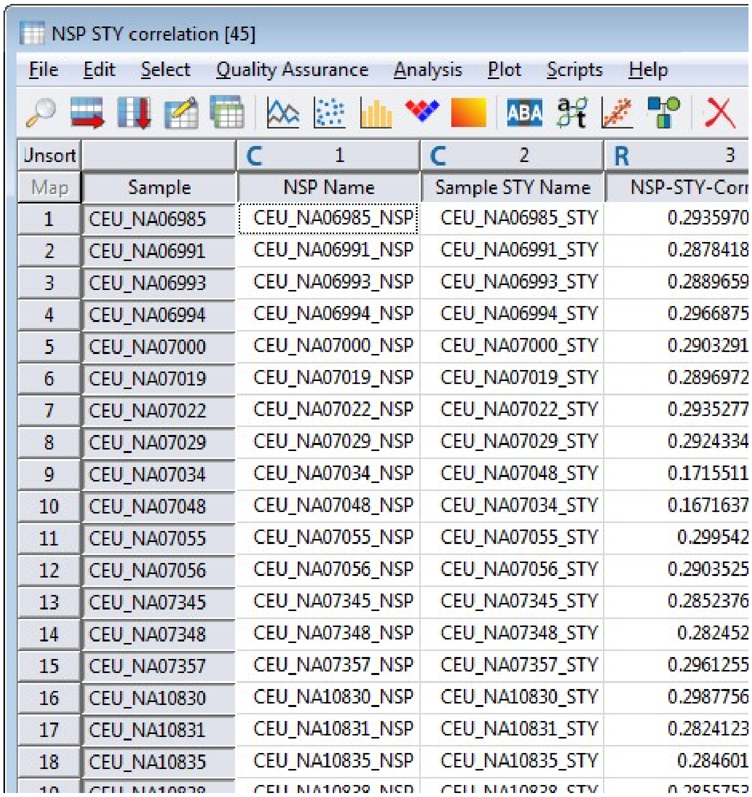
Using the Script

1) To use the script, open the NSP and STY matching spreadsheet and go to **Numeric > CNV QA > 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**.

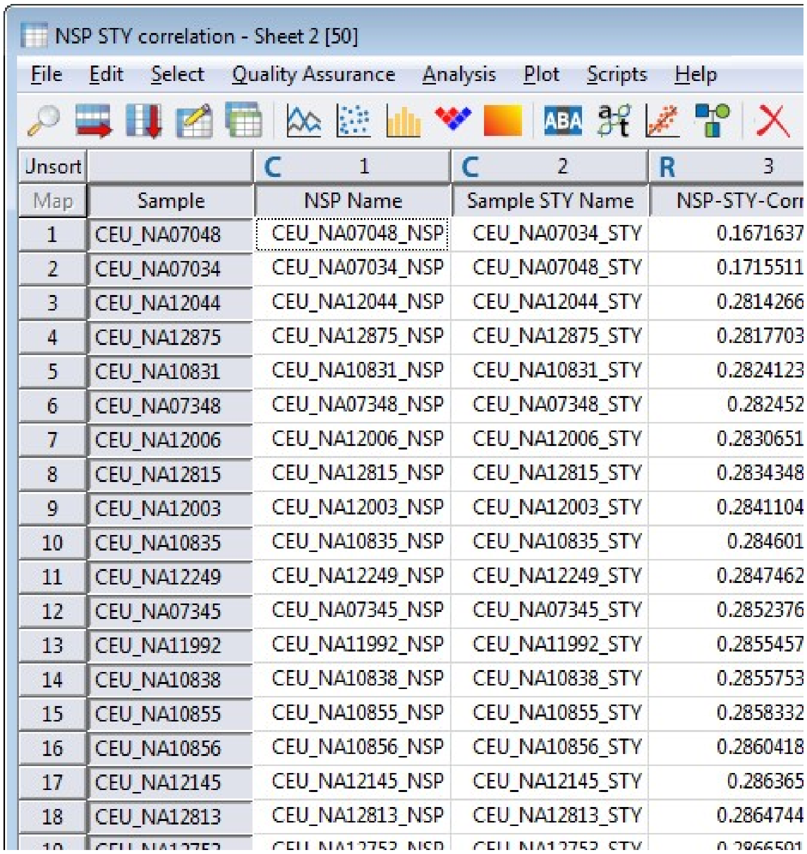


Unsort		C 1	C 2	R 3
Map	Sample	NSP Name	Sample STY Name	NSP-STY-Corr
1	CEU_NA06985	CEU_NA06985_NSP	CEU_NA06985_STY	0.2935970
2	CEU_NA06991	CEU_NA06991_NSP	CEU_NA06991_STY	0.2878418
3	CEU_NA06993	CEU_NA06993_NSP	CEU_NA06993_STY	0.2889659
4	CEU_NA06994	CEU_NA06994_NSP	CEU_NA06994_STY	0.2966875
5	CEU_NA07000	CEU_NA07000_NSP	CEU_NA07000_STY	0.2903291
6	CEU_NA07019	CEU_NA07019_NSP	CEU_NA07019_STY	0.2896972
7	CEU_NA07022	CEU_NA07022_NSP	CEU_NA07022_STY	0.2935277
8	CEU_NA07029	CEU_NA07029_NSP	CEU_NA07029_STY	0.2924334
9	CEU_NA07034	CEU_NA07034_NSP	CEU_NA07048_STY	0.1715511
10	CEU_NA07048	CEU_NA07048_NSP	CEU_NA07034_STY	0.1671637
11	CEU_NA07055	CEU_NA07055_NSP	CEU_NA07055_STY	0.299542
12	CEU_NA07056	CEU_NA07056_NSP	CEU_NA07056_STY	0.2903525
13	CEU_NA07345	CEU_NA07345_NSP	CEU_NA07345_STY	0.2852376
14	CEU_NA07348	CEU_NA07348_NSP	CEU_NA07348_STY	0.282452
15	CEU_NA07357	CEU_NA07357_NSP	CEU_NA07357_STY	0.2961255
16	CEU_NA10830	CEU_NA10830_NSP	CEU_NA10830_STY	0.2987756
17	CEU_NA10831	CEU_NA10831_NSP	CEU_NA10831_STY	0.2824123
18	CEU_NA10835	CEU_NA10835_NSP	CEU_NA10835_STY	0.284601
19	CEU_NA10838	CEU_NA10838_NSP	CEU_NA10838_STY	0.2855753

Figure 2: NSP STY correlation spreadsheet

4) Finally, the easiest way to identify pairs that are potential mismatches is to choose **Select > Compare and Activate by Column Agreement** 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-Corr
1	CEU_NA07048	CEU_NA07048_NSP	CEU_NA07034_STY	0.1671637
2	CEU_NA07034	CEU_NA07034_NSP	CEU_NA07048_STY	0.1715511
3	CEU_NA12044	CEU_NA12044_NSP	CEU_NA12044_STY	0.2814266
4	CEU_NA12875	CEU_NA12875_NSP	CEU_NA12875_STY	0.2817703
5	CEU_NA10831	CEU_NA10831_NSP	CEU_NA10831_STY	0.2824123
6	CEU_NA07348	CEU_NA07348_NSP	CEU_NA07348_STY	0.282452
7	CEU_NA12006	CEU_NA12006_NSP	CEU_NA12006_STY	0.2830651
8	CEU_NA12815	CEU_NA12815_NSP	CEU_NA12815_STY	0.2834348
9	CEU_NA12003	CEU_NA12003_NSP	CEU_NA12003_STY	0.2841104
10	CEU_NA10835	CEU_NA10835_NSP	CEU_NA10835_STY	0.284601
11	CEU_NA12249	CEU_NA12249_NSP	CEU_NA12249_STY	0.2847462
12	CEU_NA07345	CEU_NA07345_NSP	CEU_NA07345_STY	0.2852376
13	CEU_NA11992	CEU_NA11992_NSP	CEU_NA11992_STY	0.2855457
14	CEU_NA10838	CEU_NA10838_NSP	CEU_NA10838_STY	0.2855753
15	CEU_NA10855	CEU_NA10855_NSP	CEU_NA10855_STY	0.2858332
16	CEU_NA10856	CEU_NA10856_NSP	CEU_NA10856_STY	0.2860418
17	CEU_NA12145	CEU_NA12145_NSP	CEU_NA12145_STY	0.286365
18	CEU_NA12813	CEU_NA12813_NSP	CEU_NA12813_STY	0.2864744
19	CEU_NA12753	CEU_NA12753_NSP	CEU_NA12753_STY	0.2866501

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

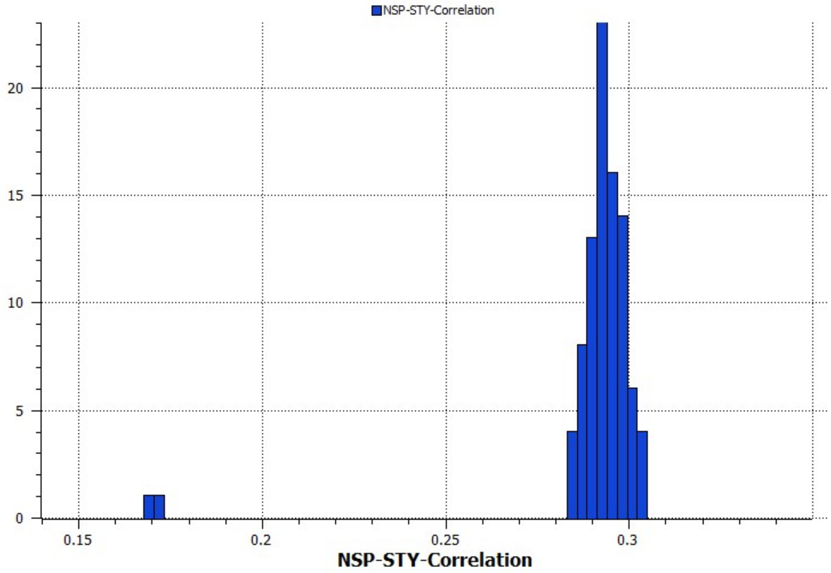


Figure 4: Histogram of NSP STY Correlation Values showing the two samples that are outliers.