Multilocus Risk Scores

Author: Alison Figueira, Golden Helix, Inc.

Overview

This script will output the risk scores for each sample for each score provided. This script uses the *numpy* package. For more information, please see: PLINK's scoring routine documentation.

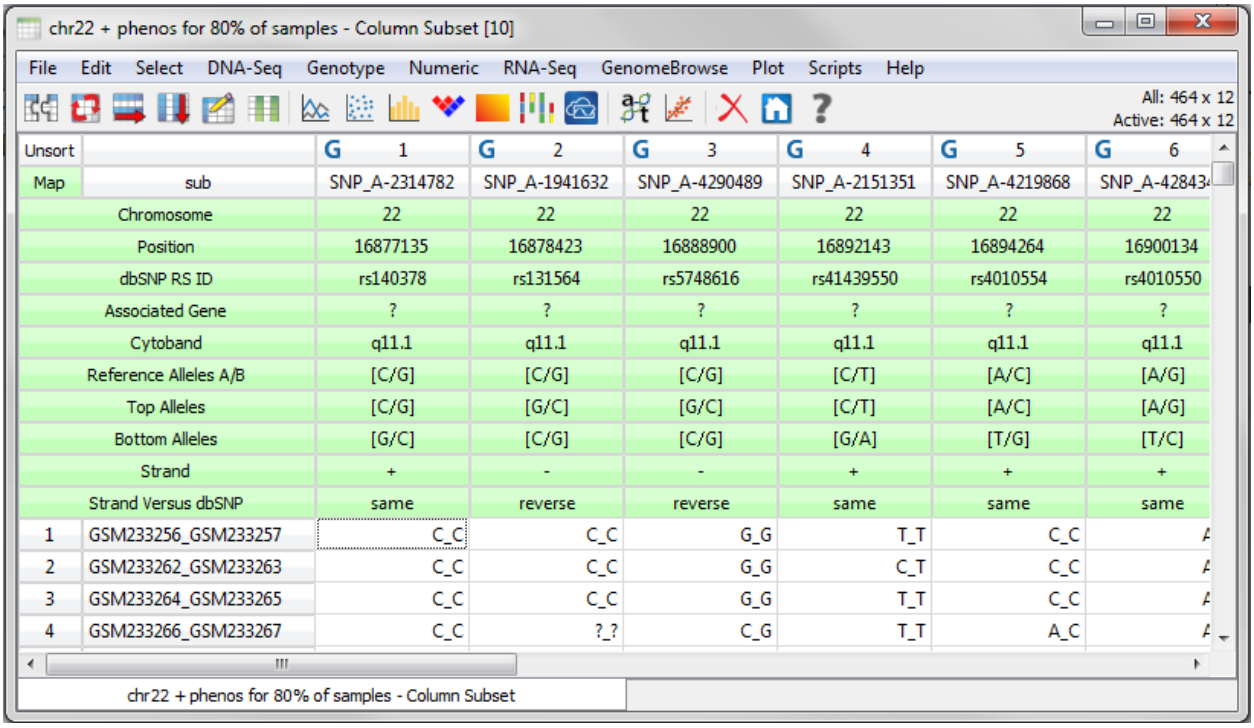
Recommended Directory Location

Save the script to the following directory:
***..\AppData\Local\Golden Helix SVS\UserScripts\Spreadsheet\Genotype**

Note: The **AppData (or Application Data)** folder is a hidden folder on Windows operating systems and its location varies between various versions. The easiest way to locate this directory on your computer is to open SVS and select the **Tools >Open Folder > UserScripts Folder** menu option and save the script in the **\Spreadsheet\Genotype** folder. If saved to the proper folder, this script will be accessible from the spreadsheet menu.

Using the Script

1. Open the spreadsheet containing the genotype data for each sample. The data should be genotypic or recoded numeric with a marker map applied to the columns, such as the example below.



Unsort		G 1	G 2	G 3	G 4	G 5	G 6
Map	sub	SNP_A-2314782	SNP_A-1941632	SNP_A-4290489	SNP_A-2151351	SNP_A-4219868	SNP_A-42843
	Chromosome	22	22	22	22	22	22
	Position	16877135	16878423	16888900	16892143	16894264	16900134
	dbSNP RS ID	rs140378	rs131564	rs5748616	rs41439550	rs4010554	rs4010550
	Associated Gene	?	?	?	?	?	?
	Cytoband	q11.1	q11.1	q11.1	q11.1	q11.1	q11.1
	Reference Alleles A/B	[C/G]	[C/G]	[C/G]	[C/T]	[A/C]	[A/G]
	Top Alleles	[C/G]	[G/C]	[G/C]	[C/T]	[A/C]	[A/G]
	Bottom Alleles	[G/C]	[C/G]	[C/G]	[G/A]	[T/G]	[T/C]
	Strand	+	-	-	+	+	+
	Strand Versus dbSNP	same	reverse	reverse	same	same	same
1	GSM233256_GSM233257	C_C	C_C	G_G	T_T	C_C	A
2	GSM233262_GSM233263	C_C	C_C	G_G	C_T	C_C	A
3	GSM233264_GSM233265	C_C	C_C	G_G	T_T	C_C	A
4	GSM233266_GSM233267	C_C	?_?	C_G	T_T	A_C	A

Figure 1: Example Spreadsheet

Make sure to inactivate (gray) any columns that you do not wish to include in your analysis.

2. While in the spreadsheet window, select **Genotype > Multilocus Risk Scores**.

3. In the first prompt window, select the spreadsheet with the scores associated with each SNP. This spreadsheet should have the scores by column. If your spreadsheet has the scores, by row, this script will transpose your spreadsheet and use that one for the calculation. The data type of the first column will be the type chosen for the transpose. Please see http://doc.goldenhelix.com/SVS/latest/svsmanual/spreadsheets.html#transposing in the documentation for how the transpose function works.

   Or, the spreadsheet could be manually transposed beforehand by selecting "Transpose Spreadsheet."

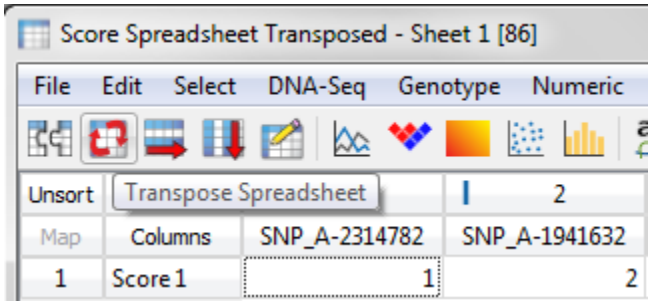


Figure 2: Transposing a spreadsheet

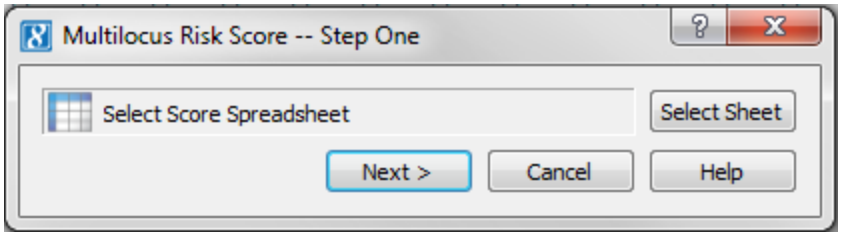


Figure 3: Selecting a score spreadsheet

(Continued on next page)

4. In the second prompt window, select with scores (columns) to use, whether missing data should be imputed or excluded, where the resultant spreadsheet should be placed, and if scores should be split by threshold values. For threshold values, the score column will be expanded into separate columns going from the lower threshold to the upper threshold by the increment value with only scores that are at the value or lower included, this is optional.

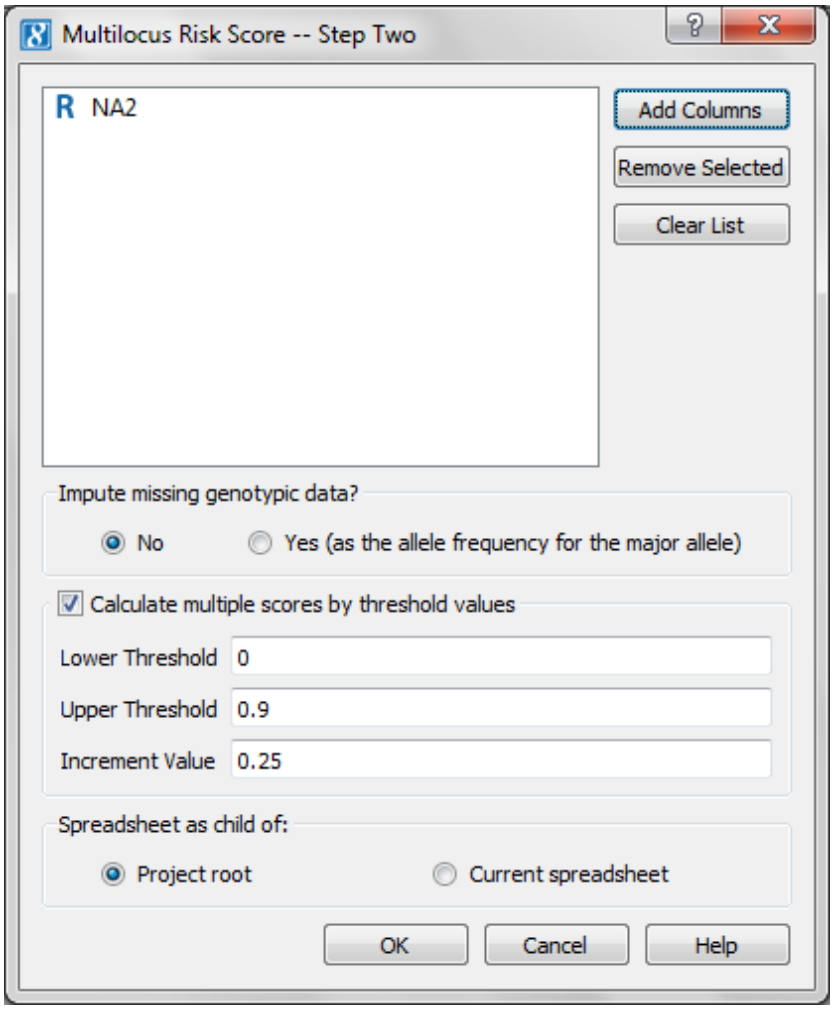
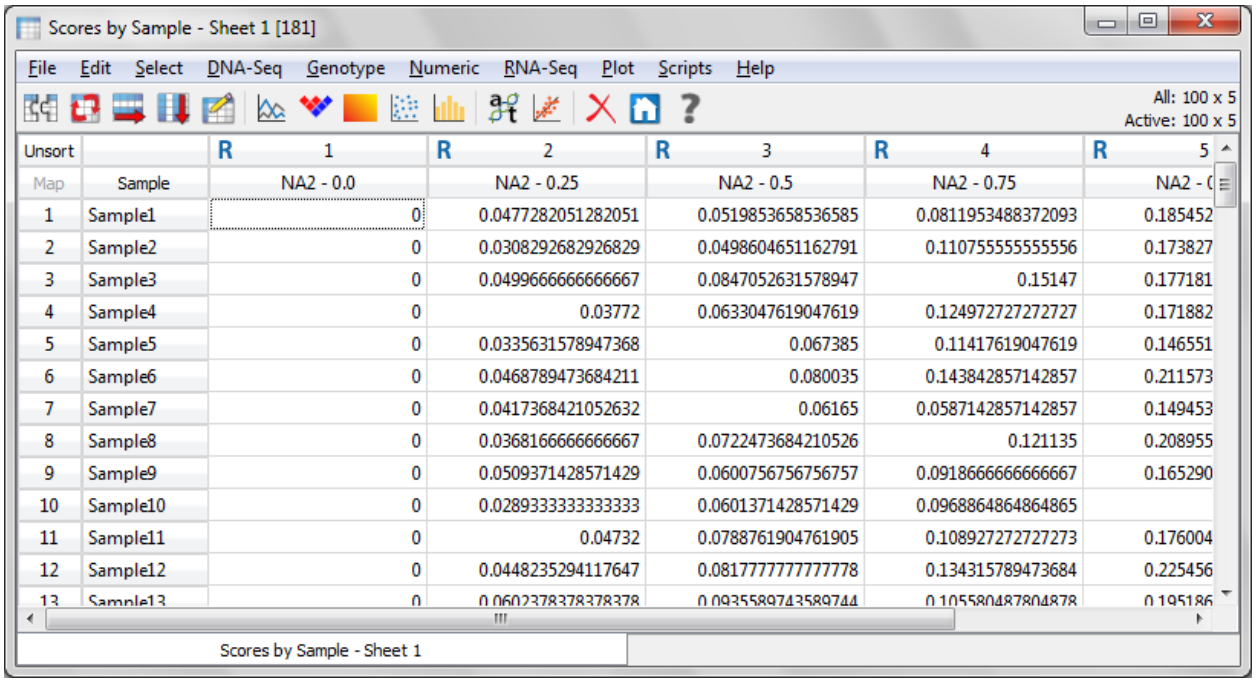


Figure 4: Second prompt, choose score columns, imputation, threshold values, and spreadsheet placement

5. Click **OK** to calculate the scores.

6. While calculating the scores, only SNPs that appear in both the genotype spreadsheet and the score spreadsheet will be included. Additionally, any SNPs that have a missing score will be excluded from the calculation for that particular score column.

7. When done, a results spreadsheet, called “Scores by Sample,” will be created which will have the sample names as the row labels and the scores by column. If separating the scores by threshold values was chosen, you’ll see separate columns for each score and threshold value pair. Another spreadsheet called “Per Sample Marker Counts” will be created that will display how many markers were included for each sample and score.



Scores by Sample - Sheet 1 [181]

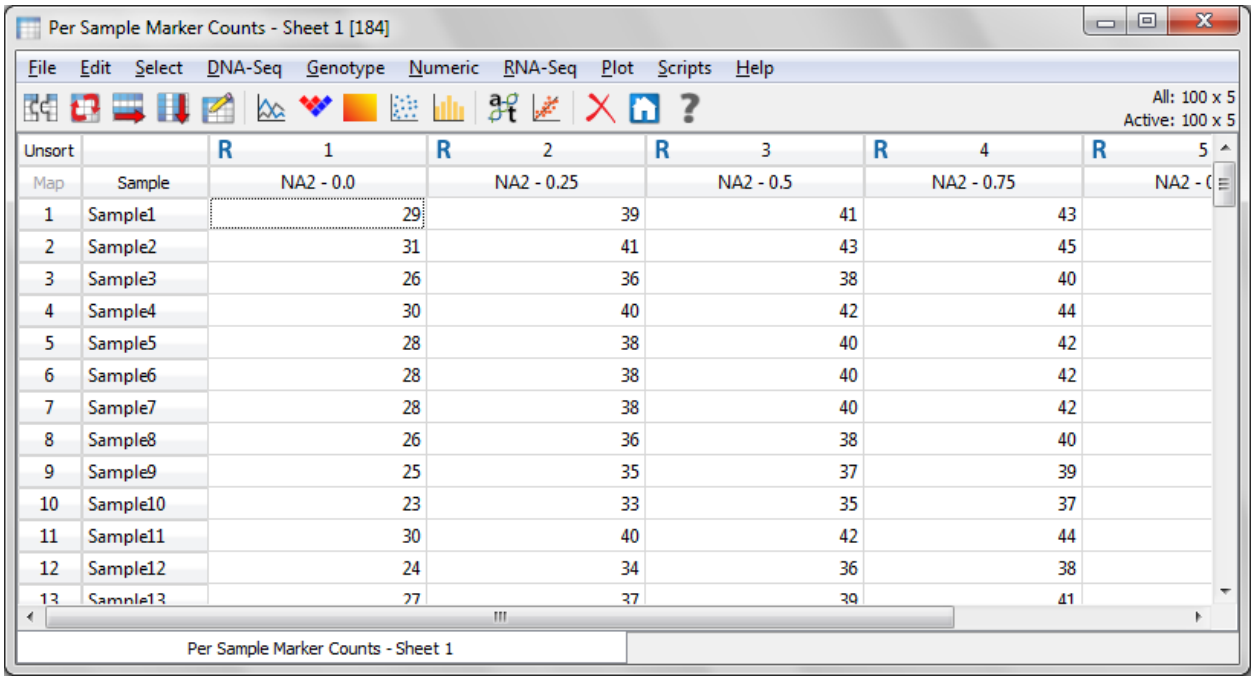
File Edit Select DNA-Seq Genotype Numeric RNA-Seq Plot Scripts Help

All: 100 x 5
Active: 100 x 5

Unsort		R 1	R 2	R 3	R 4	R 5
Map	Sample	NA2 - 0.0	NA2 - 0.25	NA2 - 0.5	NA2 - 0.75	NA2 - (
1	Sample1	0	0.0477282051282051	0.0519853658536585	0.0811953488372093	0.185452
2	Sample2	0	0.0308292682926829	0.0498604651162791	0.110755555555556	0.173827
3	Sample3	0	0.0499666666666667	0.0847052631578947	0.15147	0.177181
4	Sample4	0	0.03772	0.0633047619047619	0.124972727272727	0.171882
5	Sample5	0	0.0335631578947368	0.067385	0.11417619047619	0.146551
6	Sample6	0	0.0468789473684211	0.080035	0.143842857142857	0.211573
7	Sample7	0	0.0417368421052632	0.06165	0.0587142857142857	0.149453
8	Sample8	0	0.0368166666666667	0.0722473684210526	0.121135	0.208955
9	Sample9	0	0.0509371428571429	0.0600756756756757	0.0918666666666667	0.165290
10	Sample10	0	0.0289333333333333	0.0601371428571429	0.0968864864864865	
11	Sample11	0	0.04732	0.0788761904761905	0.108927272727273	0.176004
12	Sample12	0	0.0448235294117647	0.0817777777777778	0.134315789473684	0.225456
13	Sample13	0	0.0602378378378378	0.0935589743589744	0.105580487804878	0.195186

Scores by Sample - Sheet 1

Figure 5: Example Results, Scores by Sample



Per Sample Marker Counts - Sheet 1 [184]

File Edit Select DNA-Seq Genotype Numeric RNA-Seq Plot Scripts Help

All: 100 x 5
Active: 100 x 5

Unsort		R 1	R 2	R 3	R 4	R 5
Map	Sample	NA2 - 0.0	NA2 - 0.25	NA2 - 0.5	NA2 - 0.75	NA2 - (
1	Sample1	29	39	41	43	
2	Sample2	31	41	43	45	
3	Sample3	26	36	38	40	
4	Sample4	30	40	42	44	
5	Sample5	28	38	40	42	
6	Sample6	28	38	40	42	
7	Sample7	28	38	40	42	
8	Sample8	26	36	38	40	
9	Sample9	25	35	37	39	
10	Sample10	23	33	35	37	
11	Sample11	30	40	42	44	
12	Sample12	24	34	36	38	
13	Sample13	27	37	39	41	

Per Sample Marker Counts - Sheet 1

Figure 6: Example Results, Per Sample Marker Counts