
GOLDEN HELIX SNP & VARIATION SUITE

USER GUIDE

SNP Genome-Wide Association Tutorial

Release 8.7.0

Golden Helix, Inc.

February 24, 2017

Contents

1. Sample QA - I: Basics	2
A. Open Project	2
B. Calculate Sample Statistics	2
B. Filtering Samples with Low Call Rate	3
C. Genotype Gender Check	5
D. Apply Filtering Results to Genotype Spreadsheet	7
2. Sample QA - II: Cryptic Relatedness	8
3. Sample QA - III: Population Stratification	13
4. SNP Quality Assurance	17
5. Genotype Association Analysis	20
A. Genotype Association Testing	20
B. Generating Q-Q Plots	20
C. Generating P-Value Plots	21
D. Creating a Manhattan Plot	27
E. Saving Plots as Images	27

Updated: February 24th, 2017

Level: Fundamentals

Version: 8.7.0 or higher

Product: SVS

The following tutorial is designed to systematically introduce you to a number of techniques for genome-wide association studies. It is not meant to replicate all the workflows you might use in a complete analysis, but instead touch on a sampling of the more typical scenarios you may come across in your own studies.

The genotype data included is a portion of a public GWAS dataset from the Gene Expression Omnibus database, as well as 270 HapMap samples. There is a population spreadsheet that identifies the HapMap subpopulation and the study data. Both studies were run on the Affymetrix 500K genotyping array. All phenotype data is simulated.

Requirements

To follow along you will need to download and unzip the following file:

Download

[SNP_GWAS_Tutorial.zip](#)

We hope you enjoy the experience and look forward to your feedback.

All open windows (except the Project Navigator) can be closed after each section's completion.

1. Sample QA - I: Basics

A. Open Project

For genetic association analysis in SVS 8 you will need, at minimum, three data sources: phenotype data, genotype data, and a genetic marker map file corresponding to the particular array used to produce the genotype data. Additional preparatory steps ensure that the analysis runs correctly for your study.

You need to have a project open before you can import data or perform analyses.

- From the Welcome Screen select **File >Open Project**. Locate the Golden Helix project file **SNP_GWAS_Tutorial.gbp** and select it, then click **Open**. This will open the Project Navigator.
- Select **Tools >Current Project's Options** and confirm that the correct default genome assembly is selected. For the data in this tutorial, the genome assembly should be set to *Homo sapiens(Human),GRCh37hg19(Feb 2009)*. This option primarily affects the genomic view and which data sources are available in the integrated GenomeBrowse plotting interface.

There should be four nodes visible in the project: **Phenotype, Population, 500K Geno Training Data**, and **500K Geno HapMap Data**.

- Double-click on **500K Geno Training Data - Sheet 1**. The spreadsheet contains the genotype information for the study sample data. There are total of 565 study samples with 498,784 SNPs as indicated in the top right portion of the window (rows x columns).
- Click on the green map button in the top-left corner of the spreadsheet to view the Marker Map information and other annotations (Figure 1-1).

The following sections lead you through quality assurance procedures performed in genome-wide association studies to identify samples of poor quality (call rates, heterozygosity, etc.) and those whose identity is of question (mismatched gender, ethnicity different than intended for the study, cryptically related, etc.). In some cases we'll automatically filter samples; in others we'll just identify samples for possible exclusion.

All of the sample statistics can be calculated with one tool. In the first step you will create a master sample statistics spreadsheet. Then you will use this spreadsheet to filter out samples based on a variety of undesirable characteristics.

B. Calculate Sample Statistics

Use the **Genotype Statistics by Sample** tool to calculate sample call rates and heterozygosity rates over the entire genome and over autosomes only.

- in the **500K Geno Training Data - Sheet 1** choose **Genotype >Genotype Statistics by Sample**.

Unsort		G 1	G 2	G 3	G 4	G 5	G 6	G 7	G 8
Map	NSP_STY	SNP_A-1909444	SNP_A-4303947	SNP_A-1886933	SNP_A-2236359	SNP_A-2205441	SNP_A-2116190	SNP_A-4291020	SNP_A-1902458
Chromosome		1	1	1	1	1	1	1	1
Position		752566	779322	785989	792480	799463	1003629	1097335	1130727
dbSNP RS ID		rs3094315	rs4040617	rs2980300	rs2905036	rs4245756	rs4075116	rs9442385	rs10907175
Associated Gene		?	?	?	?	?	?	?	TTL10
Cytoband		p36.33	p36.33	p36.33	p36.33	p36.33	p36.33	p36.33	p36.33
Reference Alleles A/B		[C/T]	[A/G]	[A/G]	[C/T]	[C/T]	[A/G]	[G/T]	[A/C]
Top Alleles		[G/A]	[A/G]	[T/C]	[C/T]	[C/T]	[T/C]	[G/T]	[A/C]
Bottom Alleles		[C/T]	[T/C]	[A/G]	[G/A]	[G/A]	[A/G]	[C/A]	[T/G]
Strand		-	+	-	+	+	-	+	+
Strand Versus dbSNP		same	same	reverse	reverse	same	same	same	same
1	GSM233256_GSM233257	T_T	A_A	G_G	T_T	C_C	A_A	G_G	A_A
2	GSM233258_GSM233259	C_T	A_A	G_G	T_T	C_C	A_A	G_G	A_A
3	GSM233260_GSM233261	C_C	?_?	A_A	T_T	C_C	A_A	G_G	A_A
4	GSM233262_GSM233263	C_T	A_G	A_G	T_T	C_C	A_G	G_T	A_A
5	GSM233264_GSM233265	T_T	A_A	G_G	T_T	C_C	A_A	G_G	A_C
6	GSM233266_GSM233267	T_T	A_A	G_G	T_T	C_C	G_G	G_G	A_A
7	GSM233268_GSM233269	C_T	A_G	A_G	T_T	C_C	A_G	G_G	A_C
8	GSM233270_GSM233271	T_T	A_A	G_G	T_T	C_C	A_A	G_T	A_A
9	GSM233272_GSM233273	T_T	A_A	G_G	T_T	C_C	A_A	G_T	A_A
10	GSM233274_GSM233275	T_T	A_A	G_G	T_T	C_C	A_G	G_G	A_C
11	GSM233276_GSM233277	T_T	A_A	G_G	T_T	C_C	A_A	G_T	A_A

Figure 1-1. Mapped genotype spreadsheet

- Check *Gender inference and X statistics* under X Chromosome Statistics choosing X from the dropdown and leaving default threshold value. Check *Output count and variant statistics for each autosomal chromosome* under Additional Outputs. The window should match Figure 1-2. Click **Run**.

Two output spreadsheets are created containing the statistics. The first spreadsheet, **Statistics by Sample**, contains the call rate and the X-chromosome heterozygosity rate. The second spreadsheet, **Autosome Statistics by Sample**, contains the statistics calculated separately for each autosome.

B. Filtering Samples with Low Call Rate

First filter out samples with low call rates. The sample call rate is defined as the fraction of called SNPs per sample over the total number of SNPs in the dataset. A standard quality threshold for excluding samples with a low call rate is 95%. It is important to not include the Y chromosome when calculating per sample call rates. It is up to the researcher whether or not the X chromosome should be included in call rate calculations. In this case, use the autosome-only call rate to determine which samples should be filtered.

- Open **Statistics by Sample** and right-click on the **Call Rate (Autosomes)** column header (9). Select **Activate by Threshold** and activate all samples with call rates ... **>= 0.95**. Click **OK**.

In the top-right corner, the number of active rows is listed at 468 (out of 565 total). This means 97 samples had an autosomal call rate < 95%.

- Create a row subset spreadsheet by going to **Select >Row >Row Subset Spreadsheet**. Rename this spreadsheet by right-clicking on the tab at the bottom of the spreadsheet window and selecting **Rename**. Change the name to: **Samples with Call Rate >= 0.95**.

Now use the renamed spreadsheet to compare heterozygosity rates.

Genotype Statistics by Sample

(No variable is set as dependent.)

Genotype Count Statistics

NOTE: Call rate and heterozygosity are always output.

☐ Number and fraction of genotypes with a minor allele (as determined from sample data)

Variant Statistics (Reference Field in Map: "Reference Alleles A/B")

☐ Number of variant genotypes (non reference)

☐ Number of singletons (variant genotype present only in given sample)

☐ Mean Ti/Tv of variant genotypes

Autosomal Statistics

☐ Hardy-Weinberg Thw P-Value (taken over all autosomal chromosomes and all samples)

Gender Chromosome Statistics

☒ Gender inference:

Select chromosome to use for gender inference: X

Threshold of heterozygosity for calling M/F 0.02

Additional Outputs (Verbose Output)

☒ Output count and variant statistics for each autosomal chromosome

Help Restore Options Save Options Run Cancel

Figure 1-2. Genotype Statistics by Sample dialog

C. Genotype Gender Check

Use the X chromosome heterozygosity rate to identify those samples whose inferred gender does not agree with their reported gender. Since the reported gender information is contained in the phenotype spreadsheet, you will first need to join the sample statistics spreadsheet with the phenotype information. Then create a histogram of the Heterozygosity rate and filter based on reported gender to detect discrepancies.

- From **Samples with Call Rate >= 0.95** choose **File >Join or Merge Spreadsheet**. Select the **Phenotype - Sheet 1** spreadsheet and click **OK**.
- Choose **Current Spreadsheet** under **Spreadsheet as Child of**. Leave all other parameters as the defaults, and click **OK**.

In the combined spreadsheet, inferred gender is located in the 22nd column (Inferred Gender) and reported gender is located in the 29th column (Gender).

- Right-click on the **Het Rate from All Columns (Chr. X)** column header (18) and choose **Plot Histogram**. There are two distinct distributions. The left one represents males and the right one females.
- To better visualize the distributions, in the plot viewer, click on the **Graph 1** node in the Graph Control Interface and set the **Bin Count** parameter to **128**.
- Next, click on the **Het Rate from All Columns (Chr. X)** node in the Graph Control Interface and select the **Color** tab. Select the **By Variable** radio button and click **Select variable...**
- Scroll to **Gender**, select it and click **OK**. Then click the blue box next to the **Gender == 'F'** option and change the color to Yellow, follow a similar process to change the **Gender == 'M'** to blue. With the plot colored on reported gender, the inconsistencies between inferred and reported gender are evident. More inconsistencies can be seen by changing the opacity.
- Click on the **Het Rate from All Columns (Chr. X)** node in the **Graph Control Interface** and under the **Item** tab move the opacity bar to the left, until it is in the center.

In this example, there are four samples who have a reported gender opposite of what is characteristic according to the heterozygosity rate. Compare the values of the two columns to identify mismatched samples.

- From the spreadsheet, **Samples with Call Rate >= 0.95 + Phenotype - Sheet 1**, select **Select >Compare and Activate by Column Agreement**. Click **Add Columns** and select the two previously mentioned categorical columns (Inferred Gender and Gender) in the menu. Click **OK**.

Note: Be sure that both columns have a C to the left of header. The categorical columns use the same naming conventions, 'M' and 'F', to identify gender.

- Check both **Rows with matching data values** and **Rows with differing data values** under **Create subset spreadsheet(s) of:**. This will allow you to examine samples that have consistent and inconsistent genders.
- Confirm that the options in the window match those in Figure 1-4. Click **OK**.

Two subset spreadsheets are created: **Rows with matching values in columns Inferred Gender and Gender** and **Rows with differing values in columns Inferred Gender and Gender**.

Note: In your own study, you may be able rectify the spreadsheet by verifying that the gender was simply a data entry error and not a genotyping anomaly.

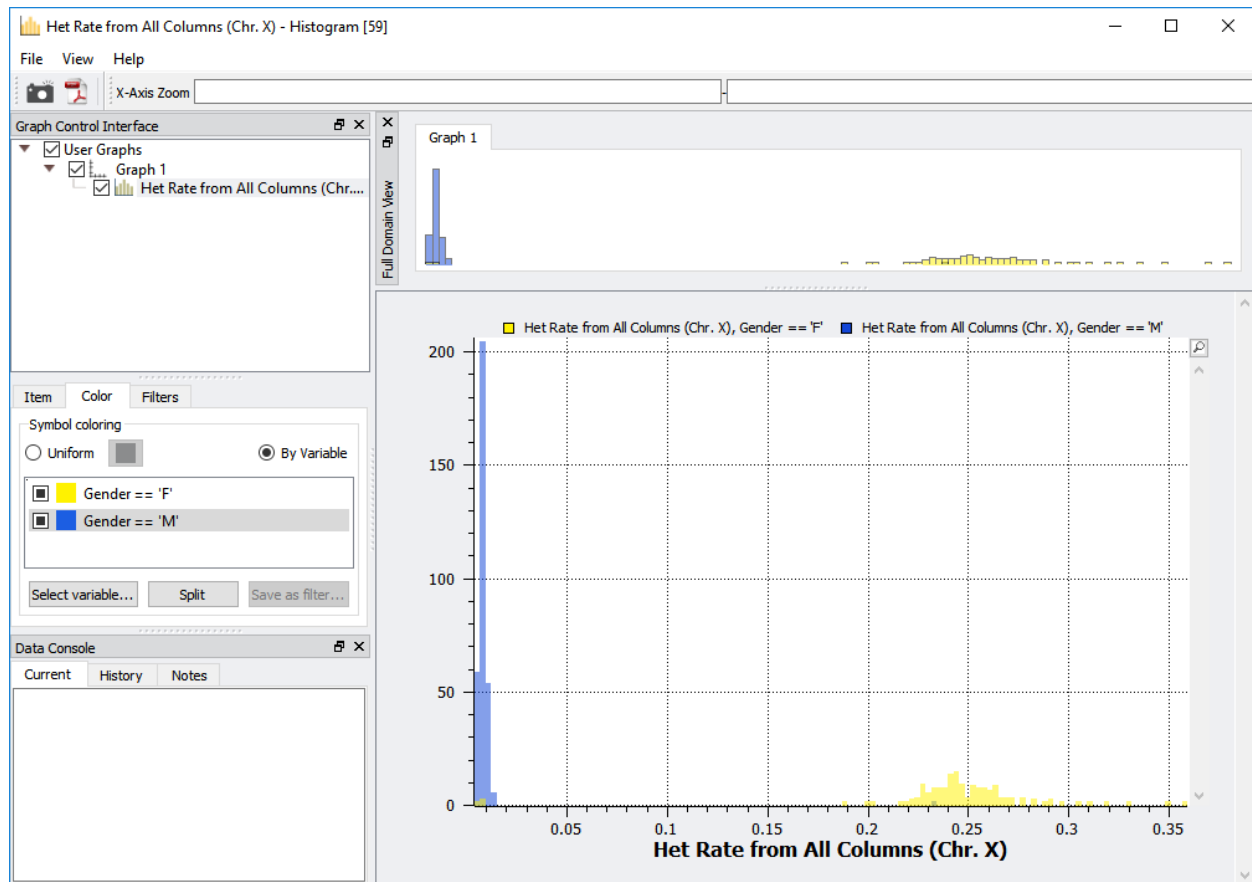


Figure 1-3. X Heterozygosity Histogram Colored by Reported Gender

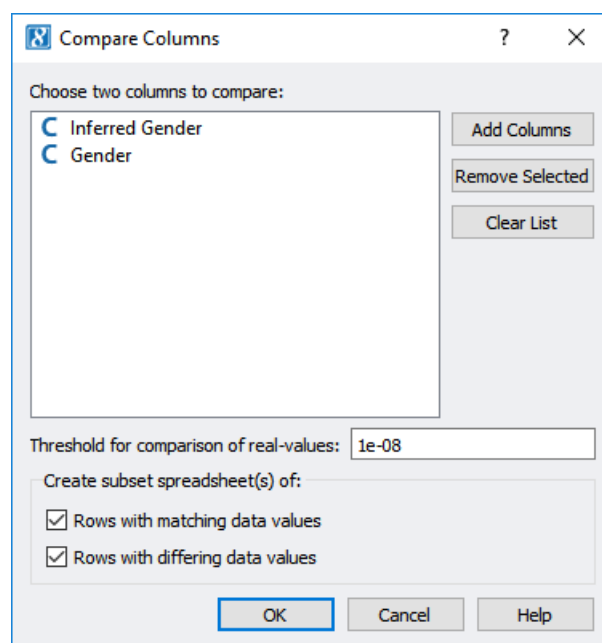


Figure 1-4. Compare Columns window

D. Apply Filtering Results to Genotype Spreadsheet

Now to exclude the filtered samples from the genotype dataset. To do this, use the rows in the **Rows with matching values in columns Inferred Gender and Gender** spreadsheet to activate their corresponding rows in the **500K Geno Training Data** spreadsheet.

- Open the **Rows with matching values in columns Inferred Gender and Gender** spreadsheet and choose **Select >Apply Current Selection to Second Spreadsheet**.
- Choose to *Apply filtered rows to the following spreadsheet* **500K Geno Training Data - Sheet 1**, then click **OK**.

You'll notice now in the upper-right portion of the window that there are only 464 rows active out of 565. Create a subset with these samples.

- Choose **Select >Row >Row Subset Spreadsheet**.
- Rename this spreadsheet in the Project Navigator (right click on the node and select **Rename Node**) to **Subset - Samples with Call Rate $\geq .95$ and Matched Gender**.

Note: You can close all open spreadsheets from the project navigator by going to **Window >Close All**

Note: Hiding child nodes in the project navigator that you will not be actively working with can help make the project easier to navigate. Clicking on the arrow next to **500K Geno Training Data - Sheet 1** results in the streamlined project navigator view below.

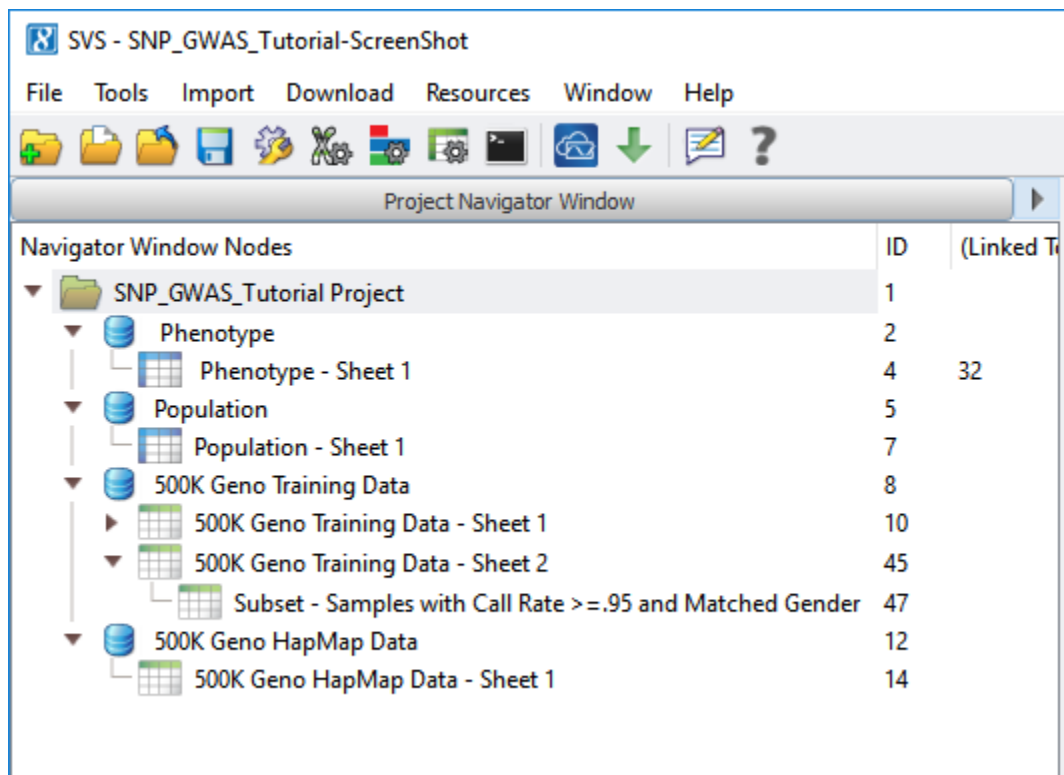


Figure 1-5. Streamlined Project Navigator.

2. Sample QA - II: Cryptic Relatedness

Next find and filter samples determined to be “related” to other samples. Relatedness is often defined as family-relatedness but identity by descent (IBD) estimation can also detect duplicate samples, duplicate samples from one of a pair of genotyping chips but not the other, or sample contamination. Before doing IBD, standard practice is to first prune the SNPs in LD with one another, reducing the number of association tests performed and thus the effect of multiple testing.

- Open **Subset - Samples with Call Rate $\geq .95$ and Matched Gender**.
- Choose **Genotype >Quality Assurance and Utilities >LD Pruning**.
- Input **100** for **Window Size**, **5** for **Window Increment**, and **0.5** for **LD r^2 Threshold**, **CHM** for **LD Computation Method**, and click **OK**.

This will take a few minutes. Upon finishing, 286,502 markers are inactivated as designated by the grayed out columns in the spreadsheet. You can also see in the upper-right portion of the window that only 212,282 columns are active out of 498,784. You will be using only the active columns (non-correlated markers) for autosomal chromosomes to perform IBD, so first create a column subset spreadsheet and then inactivate the X chromosome before continuing.

- Choose **Select >Column >Column Subset Spreadsheet**.
- Then from the subset spreadsheet, choose **Select >Activate by Chromosomes**, uncheck the **X** chromosome and click **OK**.

Note: In datasets with other non-autosomal chromosomes, inactivate these as well.

- In the Project Navigator rename this node to **Pruned SNP Subset**.

Now you are ready for IBD estimation.

- From the **Pruned SNP Subset** spreadsheet, choose **Genotype >Quality Assurance and Utilities >Identity by Descent Estimation**.
- For this exercise, uncheck **Output IBS distances ((IBS 2 + 0.5*IBS 1)/# non-missing markers)** and **Output untransformed estimates of $P(Z=0)$, $P(Z=1)$, and $P(Z=2)$** , make sure that **Output $PI = P(Z=1)/2 + P(Z=2)$** is checked, and check **Output all pairs where $PI > ___$** (enter **0**), and click **Run**.

This will take a few minutes. Upon completion, two spreadsheets are output: **IBD Estimate: Estimated PI** and **Pairwise IBD Estimates ($PI \geq 0$)**.

IBD Estimate: Estimated PI gives an $N \times N$ table where N is the number of samples in the dataset. By plotting a heatmap of this table we can detect patterns showing relatedness. **Pairwise IBD Estimates ($PI \geq 0$)** outputs various IBD statistics for all samples. These values can be sorted or plotted to find samples related to one another or detect sample contamination.

- From the **IBD Estimate: Estimated PI**, choose **Plot >Heat Map (Uniform)**.

You'll get the plot in Figure 2-1.

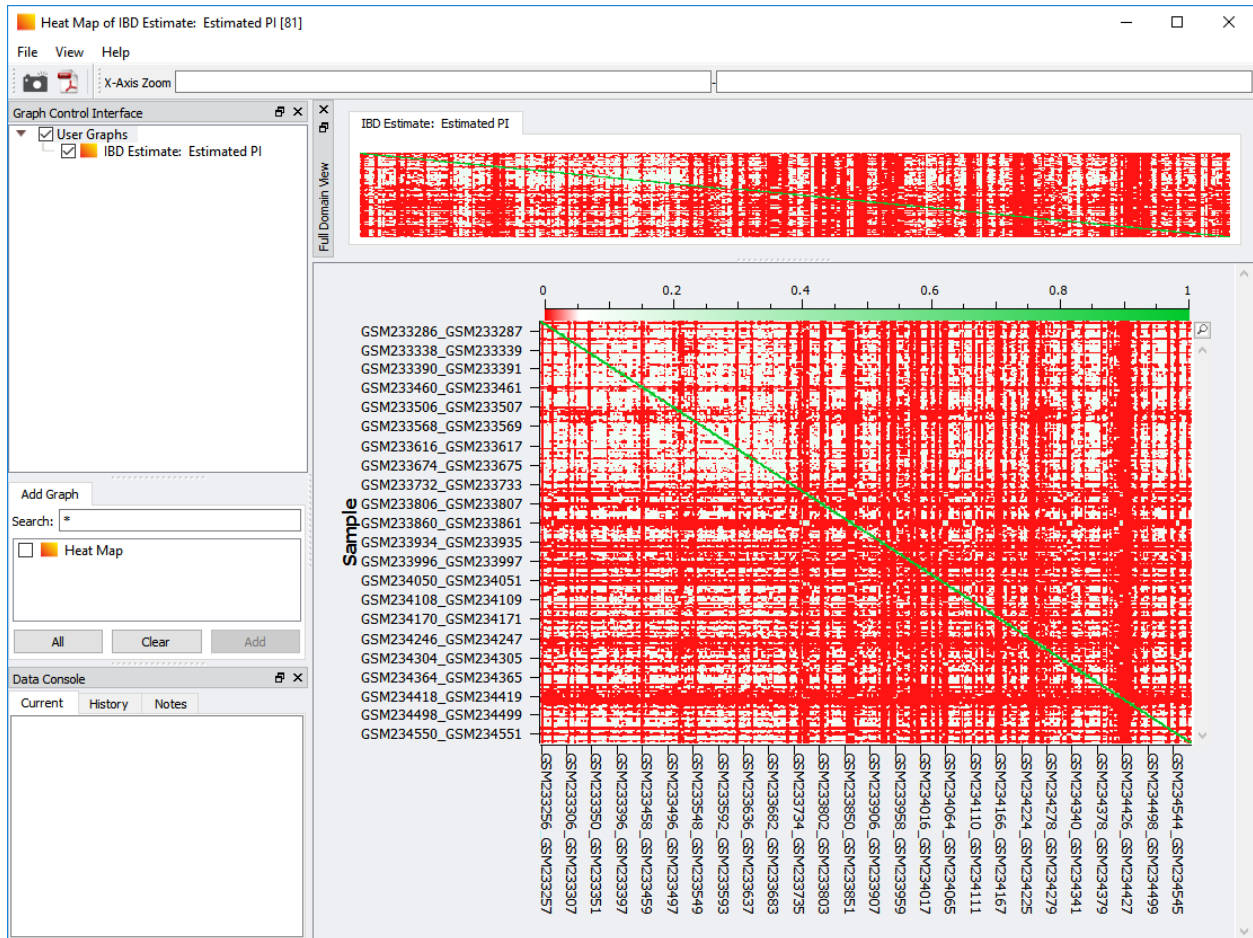


Figure 2-1. Default heat map of IBD PI estimates

By default, the heat map has a three color scheme calculated automatically. We want to define the color scheme manually based on a two color scheme where we look for sample pairs with a PI estimate of 0.25 or greater (PI of 0.25 = second degree relatives, 0.5 = first degree relatives, and 1 = identical twins (or duplicate samples)).

- Click on the **IBD Estimate: Estimated PI** node in the Graph Control Interface (upper-left portion of the window).
- On the **Color** tab choose **Manual** and then right-click the first option (0), and select **Delete**.
- Right click on the new first parameter, select **Edit**, and change it to **0.2**. Then click once on the color box next to the second parameter and select red; click **OK**. Your plot should look similar to Figure 2-2.

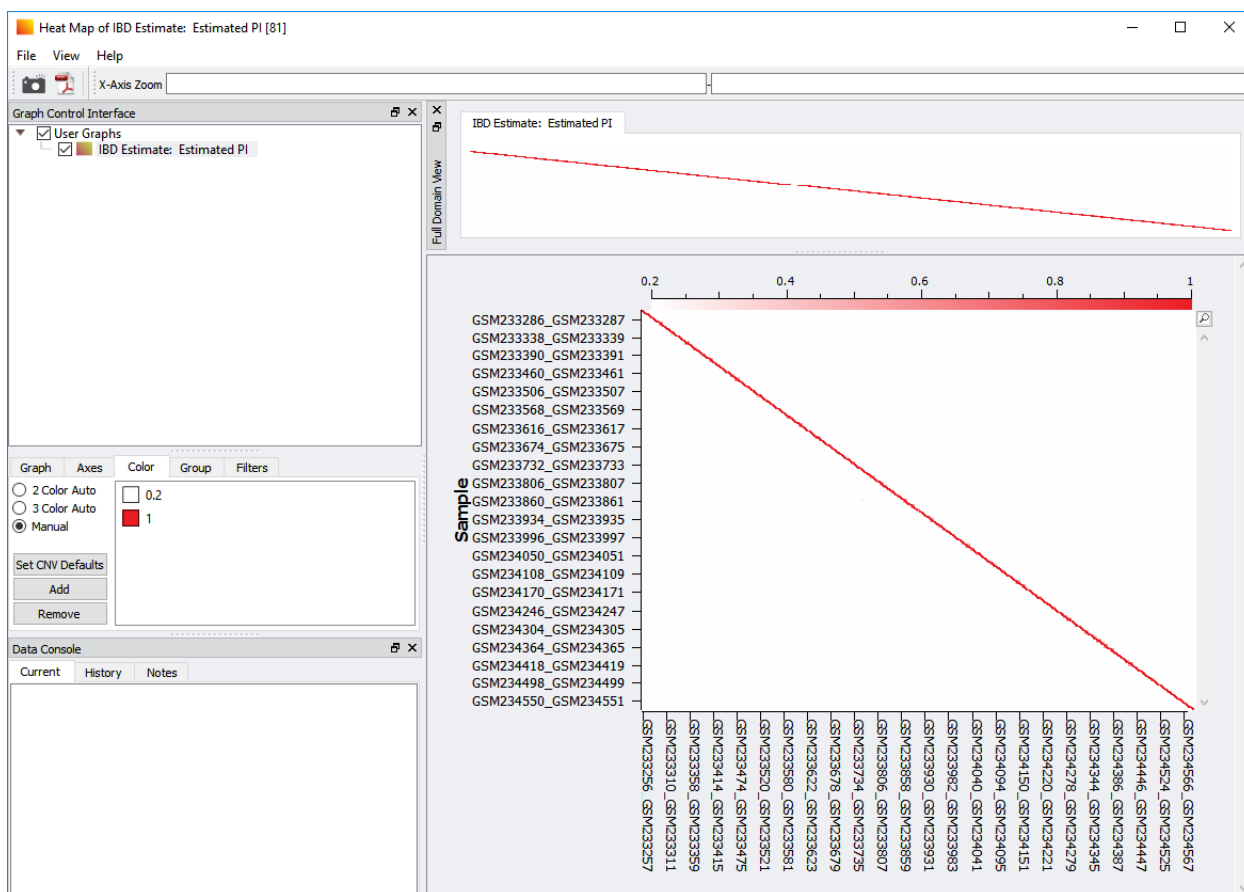


Figure 2-2. Heat map with two colors

You can begin to see samples along the diagonal line that appear to be related to one another. You can zoom in to see this more clearly by clicking and dragging a red box in the plot area (Figure 2-3).

In most population-based studies you'll typically find a couple sample pairs who are cryptically related in one way or another, which you can subsequently remove. In this particular study there are known to be several family trios and the cluster pattern above indicates this. For the purpose of this tutorial we won't exclude any sample pairs here, but if this were your own study, you would need to decide which member(s) of the trio to keep or discard from the study.

If you need to exclude samples from your study, the **Pairwise IBD Estimates (PI >= 0)** output can be used. Open the spreadsheet and right-click on the **Estimated PI** column (7) and select **Sort Descending**, those sample pairs that are

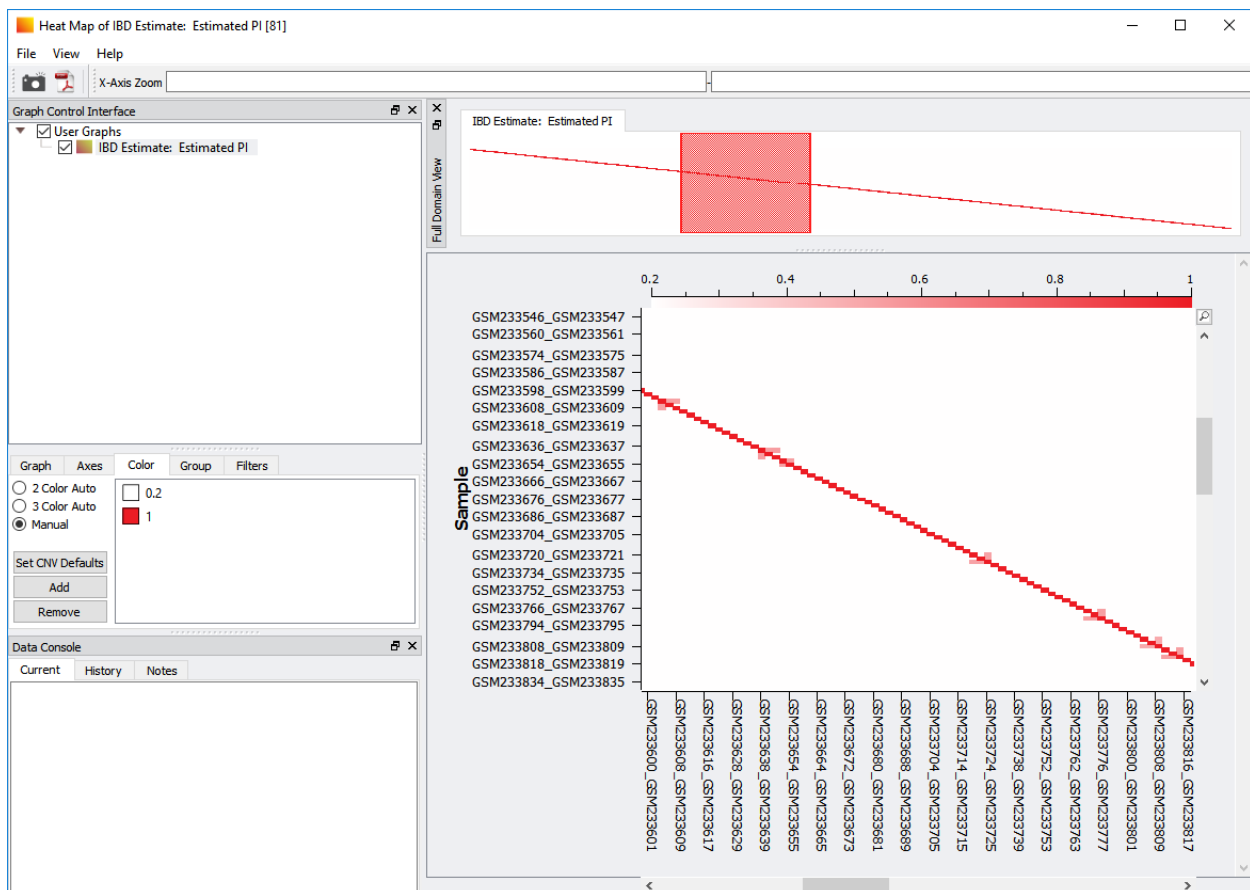


Figure 2-3. Zoomed in area around several related individuals

highly related will be listed at the top of the spreadsheet, you can then manual exclude either one or both samples of the pair from your dataset.

3. Sample QA - III: Population Stratification

The next step is to identify samples that depart from the expected homogenous ethnicity of your study. You can do this by performing principal component analysis on your data and comparing the first two principal components against reference samples of known ethnicities.

There are several ways to perform PCA. Some recommend using the pruned set of SNPs (as was done with IBD), some recommend using a filtered set of SNPs (on minor allele frequency and HWE for example), and some recommend using the entire SNP set. There are advantages to each. Here we'll use the pruned set of SNPs we created earlier. First we need to append the HapMap samples with our study samples.

- Open the **Pruned SNP Subset** spreadsheet and choose **File >Append Spreadsheets**.
- Choose the **500K Geno HapMap Data - Sheet 1** and click **OK**.
- In the **Append Spreadsheet** window enter **Pruned SNPs + HapMap** as the **New Dataset Name**, keep the rest of the parameters as the defaults and click **OK**.

The new spreadsheet should have 734 rows and 207,706 columns.

Note: You don't want to run PCA on non-autosomal chromosomes. The pruned SNP spreadsheet already had the X chromosome inactive so this chromosome was automatically dropped during the append process.

Now run PCA on this spreadsheet.

- From the spreadsheet, select **Genotype >Genotype Principal Component Analysis**.
- Under Principal Components, enter **5** for **Find up to top ____ components**.
- Leave the defaults for the rest of the options, and click **Run**.

Two spreadsheets, the **Principal Components (Additive Model)** spreadsheet and the **PC Eigenvalues (Additive Model)** spreadsheet, result from the analysis. To find out how many principal components are required to explain the majority of the population stratification, a PCA plot will be created and the eigenvalues will be visually inspected.

Look at the **PC Eigenvalues (Additive Model)** spreadsheet and notice that there is very little change between the third, fourth and fifth Eigenvalues, implying that three principal components explain the majority of stratification in the SNP data. This is consistent with there being three major populations in the data: CEPH (European), YRI (African), and CHB/JPT (Asian). You can visualize the population stratification by plotting the first few principal components against one another, juxtaposing the HapMap samples with the GEO study data. First you need to join the Population spreadsheet with the Principal Components spreadsheet.

- Open the **Population - Sheet 1** spreadsheet and select **File >Join or Merge Spreadsheets**.
- Select the **Principal Components (Additive Model)** spreadsheet.

- In the **Join or Merge Spreadsheet** window select the **Current spreadsheet** radio button under **Spreadsheet as Child of**, leave the rest of the parameters as defaults and then click **OK**. The combined spreadsheet should look like Figure 3-1.

Unsort		C 1	R 2	R 3	R 4	R 5
Map	NSP_STY	Population	EV = 35.8391	EV = 15.5366	EV = 2.04098	EV = 1.94041
1	GSM233256_GSM233257	Study	-0.00913848082717825	0.0152483756818808	0.00079288192320567	0.00116284181279436
2	GSM233258_GSM233259	Study	-0.0132832140660563	0.0160274568117261	-0.00267927413469164	-0.000919056416103216
3	GSM233262_GSM233263	Study	-0.0171541561898054	0.0144654467729679	-0.00151115502578382	-0.000264564463787557
4	GSM233264_GSM233265	Study	-0.0175477947358796	0.0179457088844222	-0.000371874804655187	-0.000242037050012731
5	GSM233266_GSM233267	Study	-0.0176869143656866	0.0192942175196264	-9.8021270587903e-006	-0.00101640789811311
6	GSM233270_GSM233271	Study	-0.0174722583523377	0.0185435996780324	-0.000304887286060429	-0.000720967511264825
7	GSM233276_GSM233277	Study	-0.0171744721818751	0.0193516580266033	0.000390985401467564	0.000138519300737938
8	GSM233278_GSM233279	Study	-0.0167299709588903	0.0188172288894126	0.000398072606222266	0.00105830110803665
9	GSM233280_GSM233281	Study	-0.00731550512550411	-0.0790740316310559	0.00125077230443915	0.000920141990854023
10	GSM233284_GSM233285	Study	-0.0176815540204872	0.0187010220735615	0.000158925142877242	0.000927627614467491
11	GSM233286_GSM233287	Study	-0.017094187251538	0.0187133319025732	-0.000186116671362646	-0.000485228562836639
12	GSM233288_GSM233289	Study	-0.016657779126341	0.0195007943419922	0.0015462503981008	0.000375433712215642
13	GSM233290_GSM233291	Study	-0.0178596992069173	0.0189509553764524	-0.000264614567949378	-0.000596937315595902
14	GSM233292_GSM233293	Study	-0.0177741336133809	0.0184774849731653	0.00135887390982303	-0.000977596419074744
15	GSM233298_GSM233299	Study	-0.0175636016189972	0.019912632628787	-0.00125602265906041	0.000600592945986911
16	GSM233300_GSM233301	Study	-0.0178470730247603	0.0192656058717757	0.00109939366893615	0.000175330604889279

Figure 3-1. Population added to the Principal Components spreadsheet

It is now possible to plot one component against the other and color-code each sample or data point according to its respective ethnicity.

- From the **Population + Principal Components (Additive Model) - Sheet 1** spreadsheet, select **Plot > XY Scatter Plots**.

The **XY Scatter Parameters** dialog appears with two list views. The list view on the left is for selecting the column (principal component) to represent the independent or X axis. The list view on the right is for selecting a single or multiple columns (principal components) to represent the dependent or Y axis.

- In the left list box select **EV = 35.8391**. In the right list box check **EV = 15.5366** and click **Plot**.

If we color each data point according to its respective ethnicity, the clusters become more obvious.

- In the **Graph Control Interface** in the upper-left pane of the Plot Viewer, select the Item **EV = 15.5366**.
- Select the **Color** tab and select the **By Variable** radio button
- Click **Select Variable** and select **Population** from the list. Click **OK**.

The four population groups are separated by color in the plot (Figure 4-2). We can see that our study consisted of mostly Caucasians (largest cluster), but also had Asians and some Asian Americans and African Americans. Similarly, four separate graph items are displayed in the **Graph Control Interface** making it easy to change the name, color and symbol for each.

- When finished, close the Plot Viewer and rename its associated node (under the **Population** spreadsheet) in the Project Navigator to **PCA Plot**.

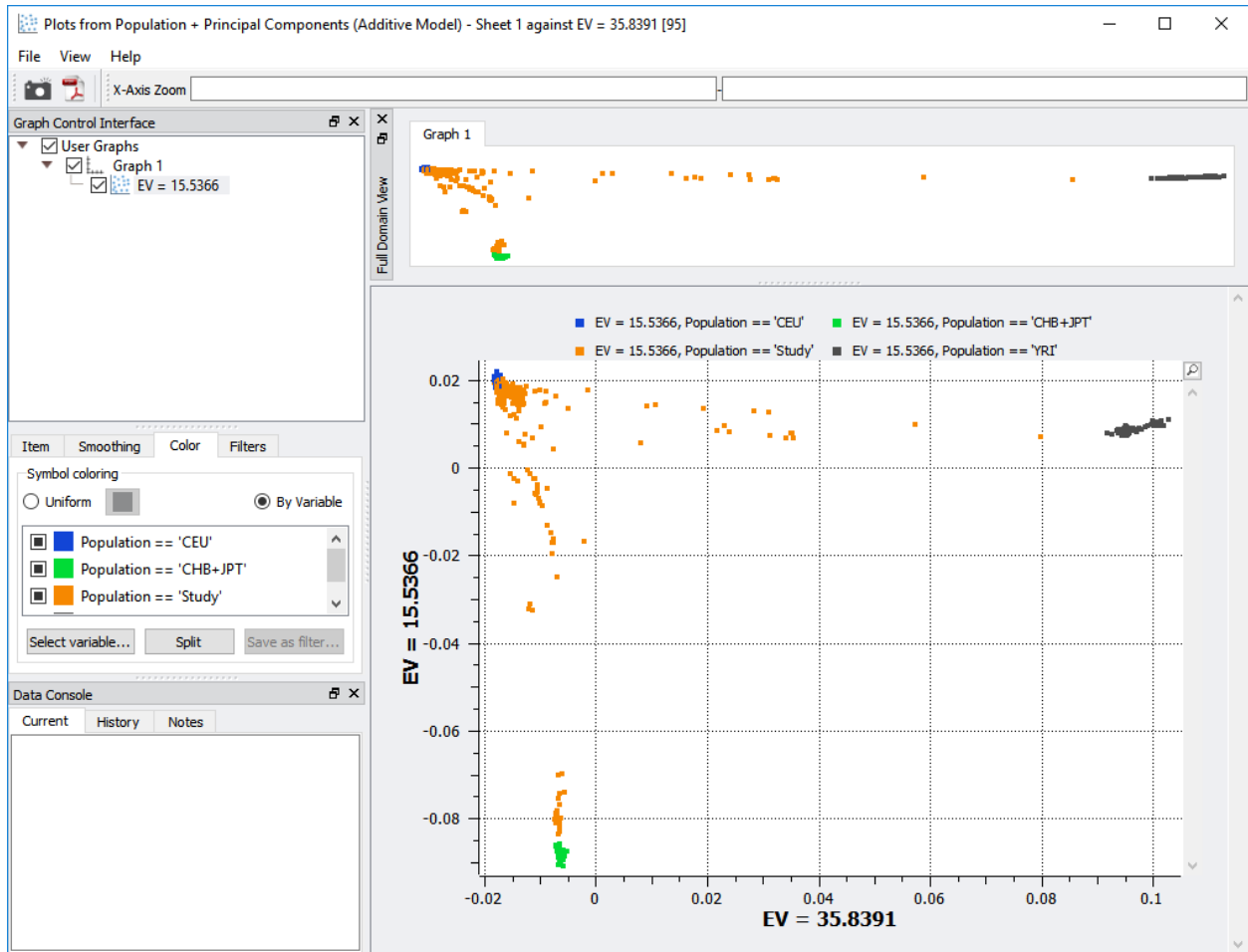


Figure 3-2. PCA plot of 3 HapMap populations and the study population

Note: Since the threshold between which samples should be excluded or kept can be ambiguous based on visual inspection of the PCA plots, we recommend calculating the inter-quartile range (IQR) distance around the centroid (median) of the study population cluster and excluding those that are 1.5 IQRs from the third quartile.

See [Multidimensional Outlier Detection](#) for examples of this process.

In summary, samples were filtered if they had a call rate below 95% and those whose reported and genotypically inferred genders did not match. We have also identified samples that could be excluded for cryptic relatedness and to remove population stratification in the dataset.

For the rest of this tutorial we will be working with those samples that remain after filtering by low call rate and filtering those samples where reported gender and genotypically inferred genders did not match, all other samples identified for possible exclusion will remain in the study.

Next, filter SNPs based on standard SNP quality assurance metrics.

4. SNP Quality Assurance

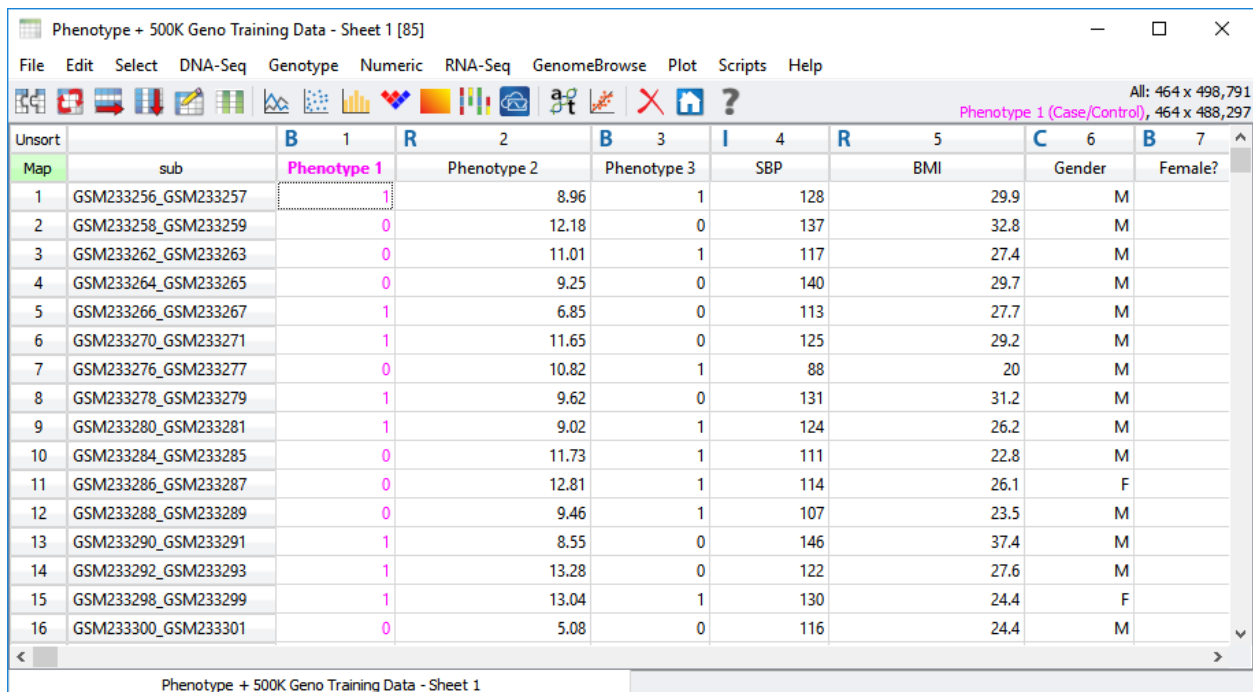
Before certain quality assurance measures can be performed on SNPs, the phenotype and genotype information needs to be merged.

- Open the **Phenotype - Sheet 1** spreadsheet and choose **File >Join or Merge Spreadsheets**.
- Select the **500K Geno Training Data - Sheet 2** spreadsheet and click **OK**.
- Leave all the parameters in the **Join or Merge Spreadsheet** window as the defaults and click **OK**.

Again, exclude the X chromosome from the following analysis. Also, specify case/control status because Hardy-Weinberg Equilibrium (HWE) will only be calculated based on control samples.

- Choose **Select >Activate by Chromosomes**, uncheck the **X** box, and click **OK**.
- Next left-click the **Phenotype 1** column label header. This will turn the column magenta denoting the column as the dependent variable.

The spreadsheet should look like Figure 4-1.



Unsort	sub	Phenotype 1	Phenotype 2	Phenotype 3	SBP	BMI	Gender	Female?
1	GSM233256_GSM233257	1	8.96	1	128	29.9	M	
2	GSM233258_GSM233259	0	12.18	0	137	32.8	M	
3	GSM233262_GSM233263	0	11.01	1	117	27.4	M	
4	GSM233264_GSM233265	0	9.25	0	140	29.7	M	
5	GSM233266_GSM233267	1	6.85	0	113	27.7	M	
6	GSM233270_GSM233271	1	11.65	0	125	29.2	M	
7	GSM233276_GSM233277	0	10.82	1	88	20	M	
8	GSM233278_GSM233279	1	9.62	0	131	31.2	M	
9	GSM233280_GSM233281	1	9.02	1	124	26.2	M	
10	GSM233284_GSM233285	0	11.73	1	111	22.8	M	
11	GSM233286_GSM233287	0	12.81	1	114	26.1	F	
12	GSM233288_GSM233289	0	9.46	1	107	23.5	M	
13	GSM233290_GSM233291	1	8.55	0	146	37.4	M	
14	GSM233292_GSM233293	1	13.28	0	122	27.6	M	
15	GSM233298_GSM233299	1	13.04	1	130	24.4	F	
16	GSM233300_GSM233301	0	5.08	0	116	24.4	M	

Figure 4-1. Joined spreadsheet with case/control status selected

- From the joined spreadsheet choose **Genotype >Genotype Filtering by Marker**.

The **Genotype Filtering** window lets you simultaneously choose thresholds for multiple statistics to filter SNPs failing to meet respective quality assurance measures.

- Check the following options and enter the following thresholds:
- **Drop if Call Rate** < 0.9
- **Drop if Minor Allele Frequency** < 0.01
- **Perform HWE filtering based on:** Controls
- **Drop if Fisher's exact test for HWE P-Value** < 1e-4
- Click **Run**.

Upon completion, SNPs in the **Phenotype + 500K Geno Training Data - Sheet 1** not meeting the specified thresholds are inactivated. A new spreadsheet, **Filtering Results**, will also be output with the various markers statistics for each SNP.

- To see how many SNPs were filtered, go to the Project Navigator and select the **Phenotype + 500K Geno Training Data - Sheet 1** spreadsheet. In the Node Change Log it will say how many SNPs were filtered (columns set to inactive).

Assuming all steps were followed correctly to this point, 104,035 SNPs should have been filtered. Though any further analyses only takes active columns and rows into consideration, it is often preferred to first create a subset spreadsheet of only those that are active.

- From the **Phenotype + 500K Geno Training Data - Sheet 1** spreadsheet, go to **Select >Subset Active Data**.

The new spreadsheet, **Phenotype + 500K Geno Training Data - Active Subset**, should have 464 rows and 384,263 columns.

- Rename this spreadsheet in the Project Navigator to **Filtered Data for Association Testing** and add a color tag to the node by right-clicking and selecting **Tag Node Green**. Colored tags are useful for easily locating specific spreadsheets.

You should now have a filtered set of samples and SNPs for association testing and your project should look similar to Figure 4-2.

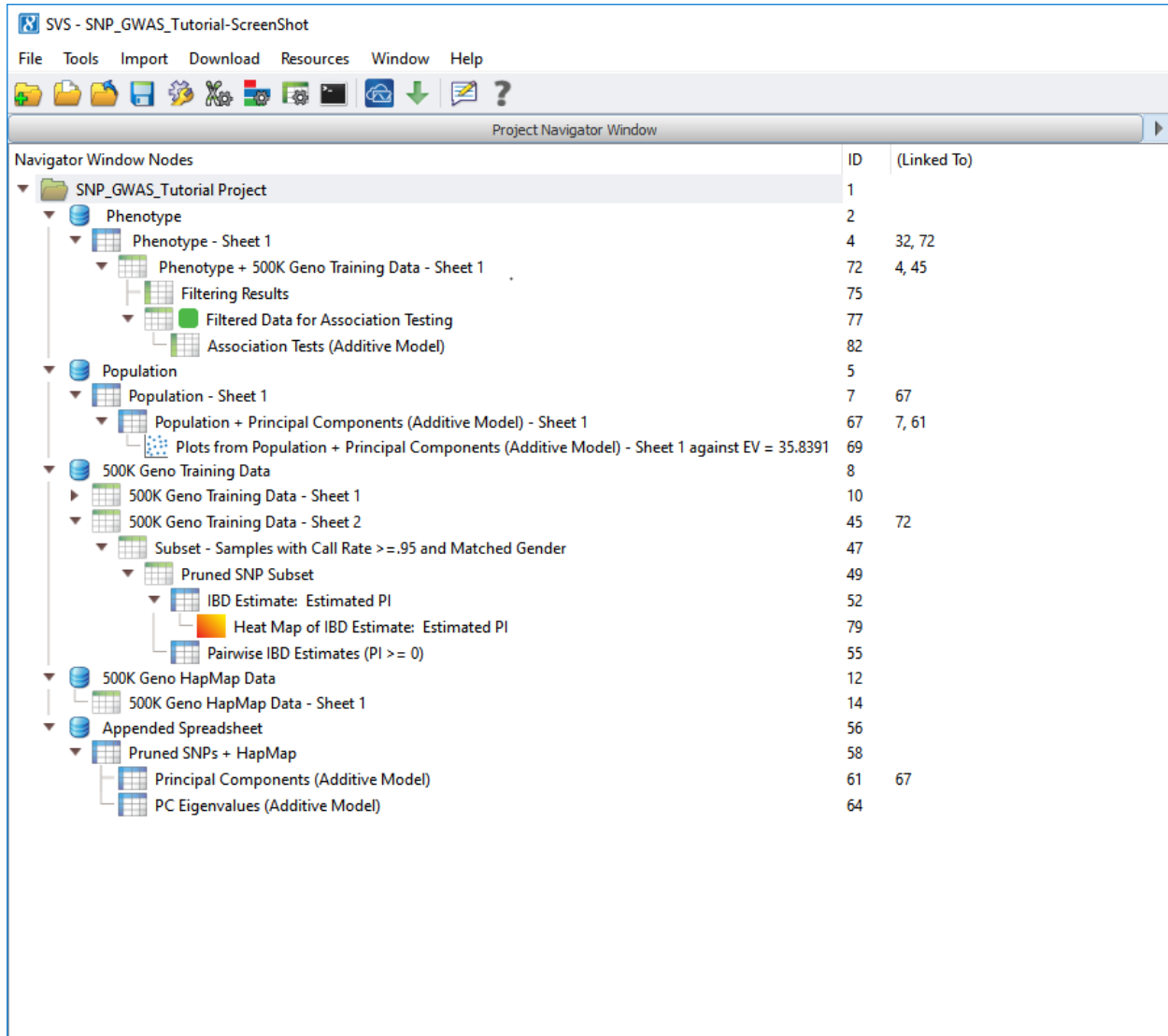


Figure 4-2. Project Navigator view

5. Genotype Association Analysis

After assuring the quality of the data, association testing can be performed.

A. Genotype Association Testing

- Open the **Filtered Data for Association Testing** spreadsheet and make sure the **Phenotype 1 Binary** column header is still set as the dependent variable.
- Choose **Genotype >Genotype Association Tests**.
- Make sure the **Additive Model: (dd) -> (Dd) -> (DD)** radio button is selected and check only **Correlation/Trend test** and **Exact form of Cochran-Armitage test** under **Test Statistic or Method**.
- Under **Multiple Testing Correction**, make sure **Bonferroni Adjustment** is checked and uncheck the other options.
- Check **Output data for P-P/Q-Q Plots**
- On the **Overall Marker Statistics** tab check **Genotype counts** under Count Tables and click **Run**.
- Upon completion a new spreadsheet is created, **Association Tests (Additive Model)**. This spreadsheet displays several association statistics for each SNP (Figure 5-1).

The results from the Exact Cochran-Armitage Test should be examined in the case when a SNP has a significant p-value but the counts in the contingency table of Case Status by Number of Minor Alleles has at least one count less than 5. In this case the assumptions of the Correlation/Trend test are violated.

- Right-click on the *Corr/Trend -log10 P* column and select **Sort Descending** to bring the most significant P-values to the top of the spreadsheet.

In this study the most significant marker: SNP_A-2070191 (Corr/Trend p-value = 2.499e-7) has the following contingency table:

	dd	Dd	DD	Total
Case	100	111	19	230
Control	147	69	5	221
Total	247	180	24	451

This results in an Exact Armitage P-Value of 1.869e-7. There is little difference between these results because all cell counts are 5 or higher.

B. Generating Q-Q Plots

Q-Q plots are generated by plotting the expected chi-squared values against the observed chi-squared values.

- From **Association Tests (Additive Model)**, select **Plot >XY Scatter Plots**. Two list views will appear.

Map	Marker	R	1	R	2	R	3	R	4	R	5
			Corr/Trend P		Corr/Trend -log10 P		Corr/Trend R		Corr/Trend X^2		Corr/Trend expected P
1	SNP_A-1909444		0.823610992556257		0.0842778655816102		0.0103591936230651		0.0496858692368305		0.824214388882383
2	SNP_A-4303947		0.607711955235808		0.216302219894464		-0.0241189244963382		0.263520301043608		0.609661552875044
3	SNP_A-1886933		0.419039164827852		0.377745384494962		-0.0379673710741759		0.65300913362673		0.418345890099023
4	SNP_A-2116190		0.390244651738654		0.408663040184332		-0.0399724317266052		0.738181427739809		0.389052842513435
5	SNP_A-4291020		0.803995332242444		0.0947464726287053		-0.0117123528649632		0.0615934651254015		0.805960885349573
6	SNP_A-1902458		0.204380733320902		0.68956004692674		-0.0591104809440866		1.6107565693804		0.202270627578041
7	SNP_A-2109914		0.726161942202304		0.138966515991351		0.0162768295070365		0.122664987784925		0.726196145788604
8	SNP_A-2291997		0.674846036031889		0.170795298767169		-0.019753807980215		0.175986031303362		0.677207843749593
9	SNP_A-4277872		0.867355283244052		0.0618029717894659		-0.00779577000275364		0.0278952797405475		0.866165697258331
10	SNP_A-4221087		0.229831562291807		0.638590330791662		0.056354897148033		1.44184699238468		0.228235156341492
11	SNP_A-1866065		0.733342611279004		0.134693079349566		0.0164483094862814		0.116064613646336		0.733420515022576
12	SNP_A-2288244		0.360313940856293		0.443318934325298		0.0425129440621263		0.83680324114005		0.358682385395115
13	SNP_A-1884606		0.132259429616758		0.878573354570407		-0.070259028449073		2.26577596508091		0.129587123134377
14	SNP_A-1783407		0.314539713220968		0.502324513513222		0.0471960009893667		1.01126797926229		0.312470885219451
15	SNP_A-2082515		0.73530936325985		0.133529903787277		0.0157115144262479		0.114292330417148		0.735320295116524
16	SNP_A-1910751		0.759533817513879		0.119452884825938		0.0145924136675373		0.093692956123593		0.761664779898765
17	SNP_A-2235839		0.679716693228732		0.167672064066117		-0.0193547127390384		0.170445231870866		0.681915654968706
18	SNP_A-2081399		0.830231187362503		0.0808009565211714		-0.0101184103144434		0.0459696200538541		0.830205202274531
19	SNP_A-1919019		0.894433258677915		0.0484520603166122		0.00617360918199544		0.0176084140533925		0.893561567188456

Figure 5-1. Association test results

- Select **Corr/Trend expected X^2** (seventh down) in the left list box and **Corr/Trend X^2** (fourth down) in the right list box.
- Click **Plot**.
- Select **Graph 1** in the Graph Control Tree.
- Under the **Add Item** tab select **f(x) = m(x) + b** and click **Add**.

This will generate a straight line with a slope of 1 and y-intercept of 0. You should have a Q-Q plot that looks like Figure 5-2.

- To change the weight and color of this line, select its associated graph item in the Graph Control Interface and choose the color and weight you like.
- When you've finished, close the Plot Viewer and rename its associated node in the Project Navigator to **Q-Q Plot**.

Similarly, you might also plot a P-P plot by using the expected -log10 P on the X axis and -log10 P on the Y.

C. Generating P-Value Plots

- From the **Association Tests (Additive Model)** spreadsheet, right-click on the **Corr/Trend -log10 P** column (2) and select **Plot Variable in GenomeBrowse**.

Notice the full-domain view now has chromosome bands and the X-axis is represented by chromosome and physical position (Figure 5-3).

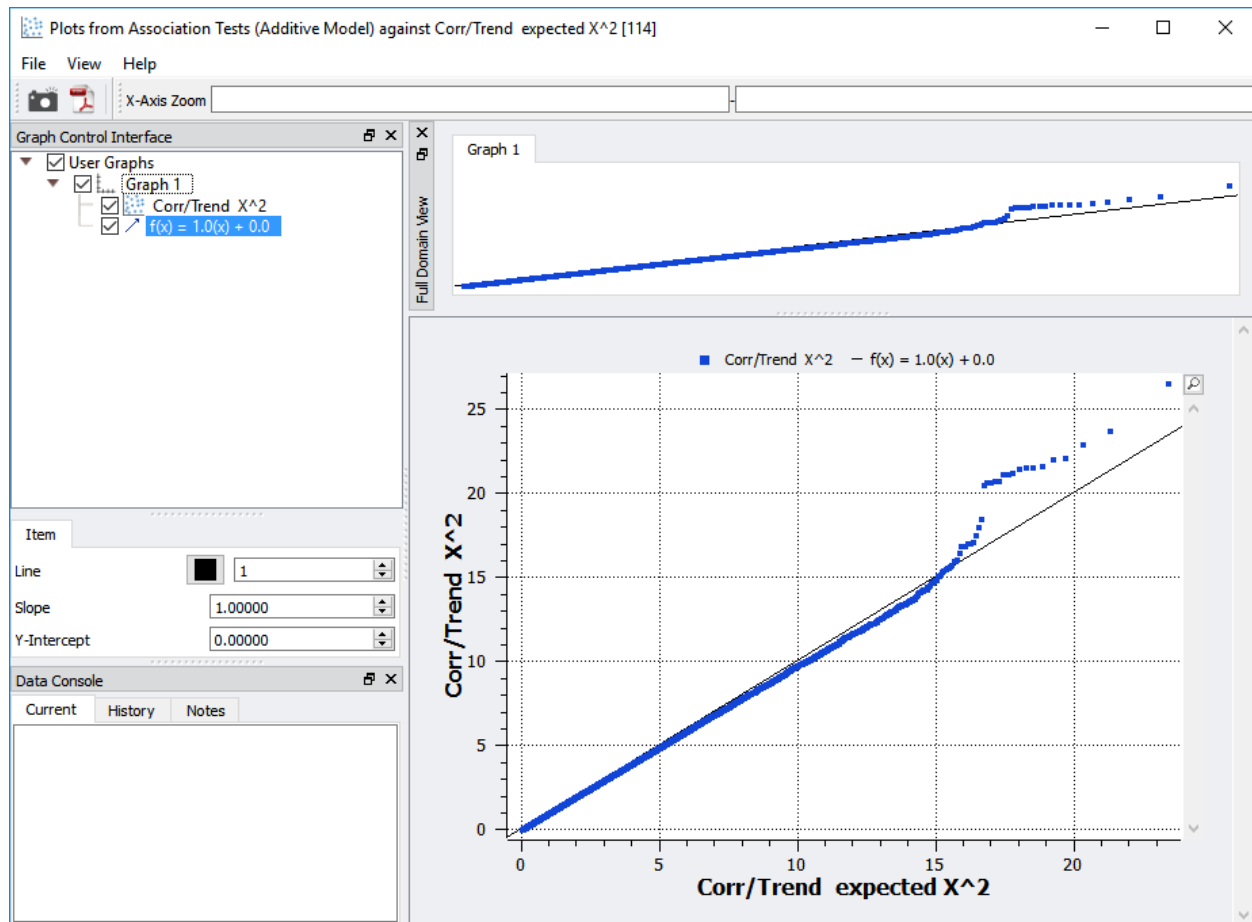


Figure 5-2. Q-Q plot of association results

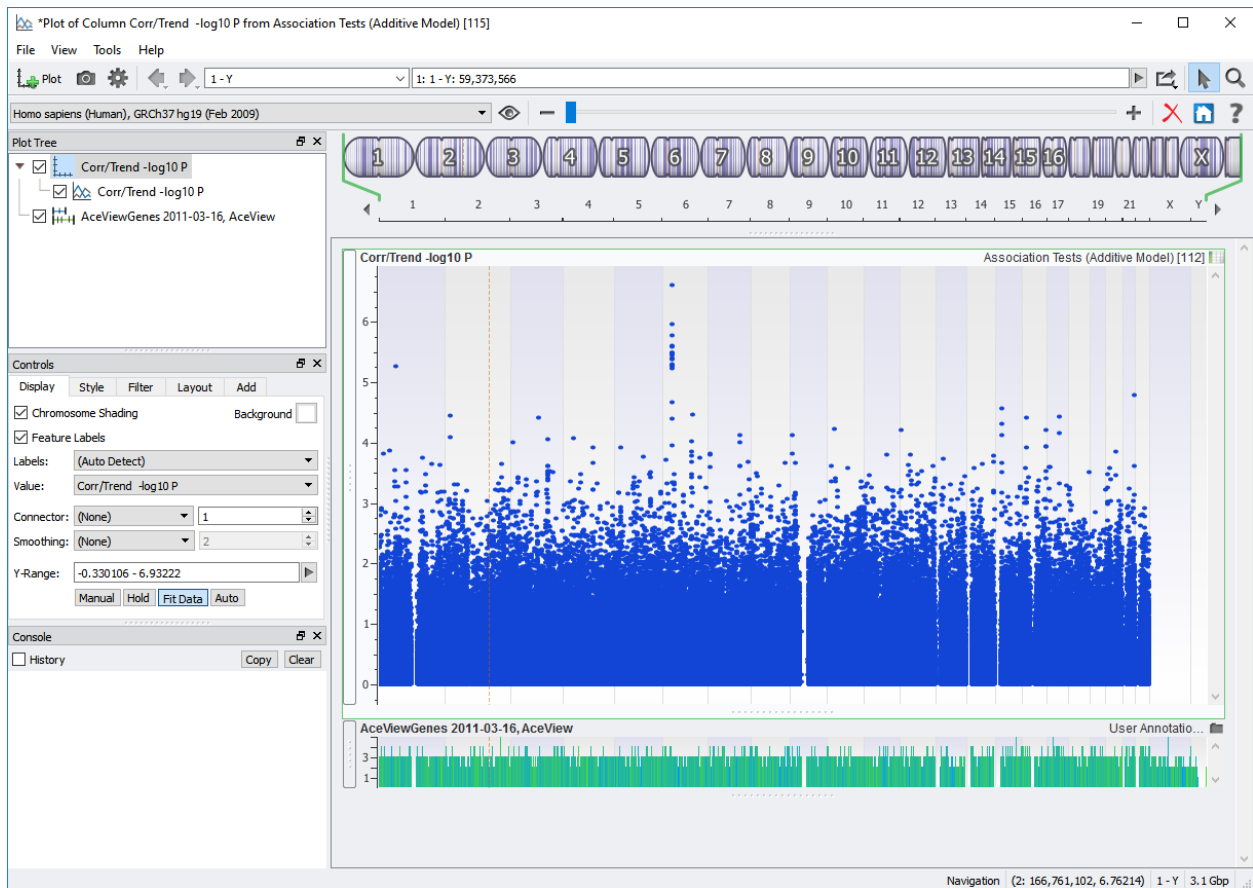


Figure 5-3. P-value plot in genome browser

There are many ways to zoom in the GenomeBrowse window: double-clicking a cytoband in the Full Domain View (top plot), manually selecting a chromosome and/or position in the Genomic Location Bar (at the top above the Full Domain Band), using the mouse scroll wheel, and using the zoom slider.

- Double-click the 6 in the **Full Domain View** since this is the location of the most significant p-values.

Zooming displays the karyogram view of chromosome 6. More information about SNPs are available with different annotation tracks.

- Zoom further into the peak (click and drag on the x-axis) and left-click on the top-most point in the plot (Figure 5-4).

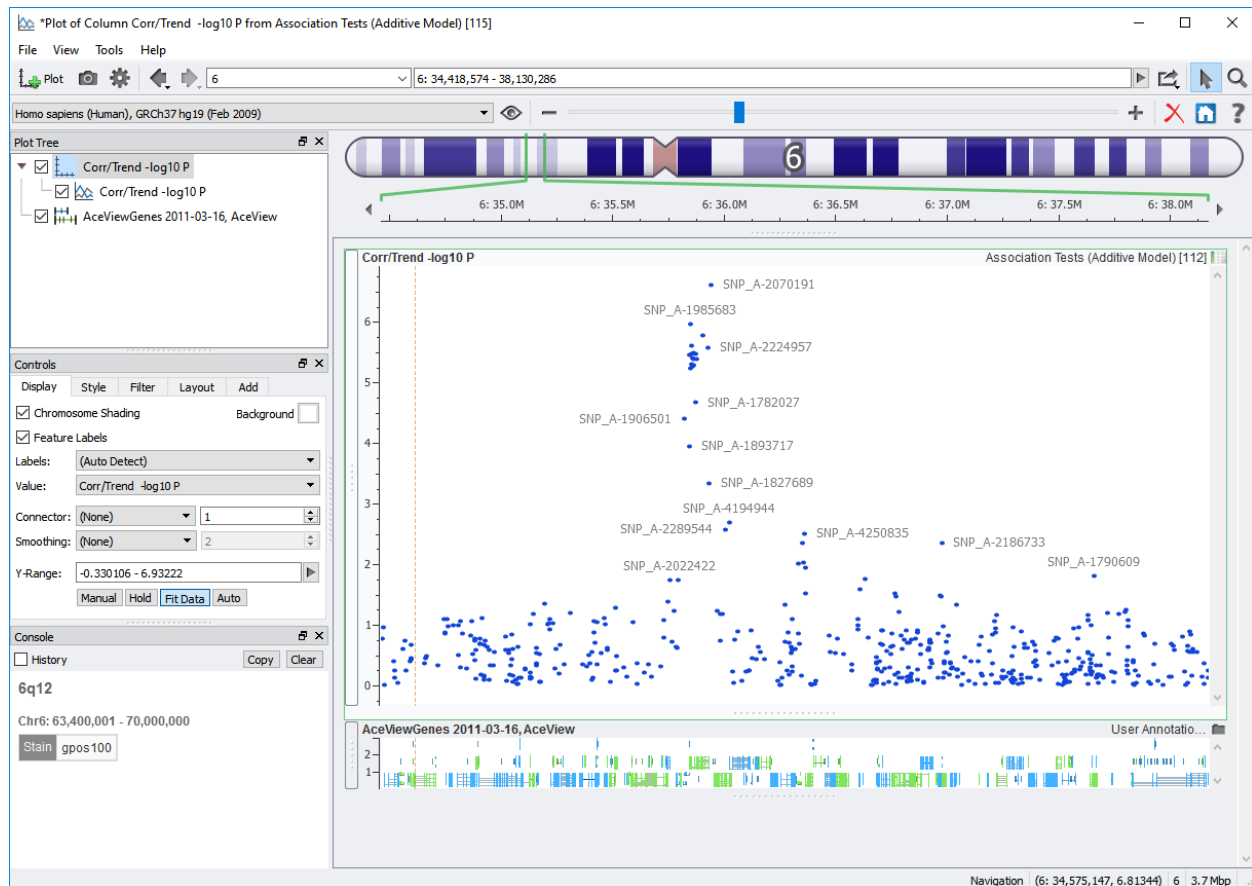


Figure 5-4. Zoomed in around the peak in the p-value plot

This displays the marker name, its p-value, chromosome, and position in the Console (bottom-left pane).

- You can see what gene(s) this SNP and others in the peak reside by looking at the annotation gene source that is loaded by default into the plot.

You can also find a listing of information available in each plot by activating the **Feature List** for that plot. Right-click anywhere on an active plot and select Feature List (Figure 5-6).

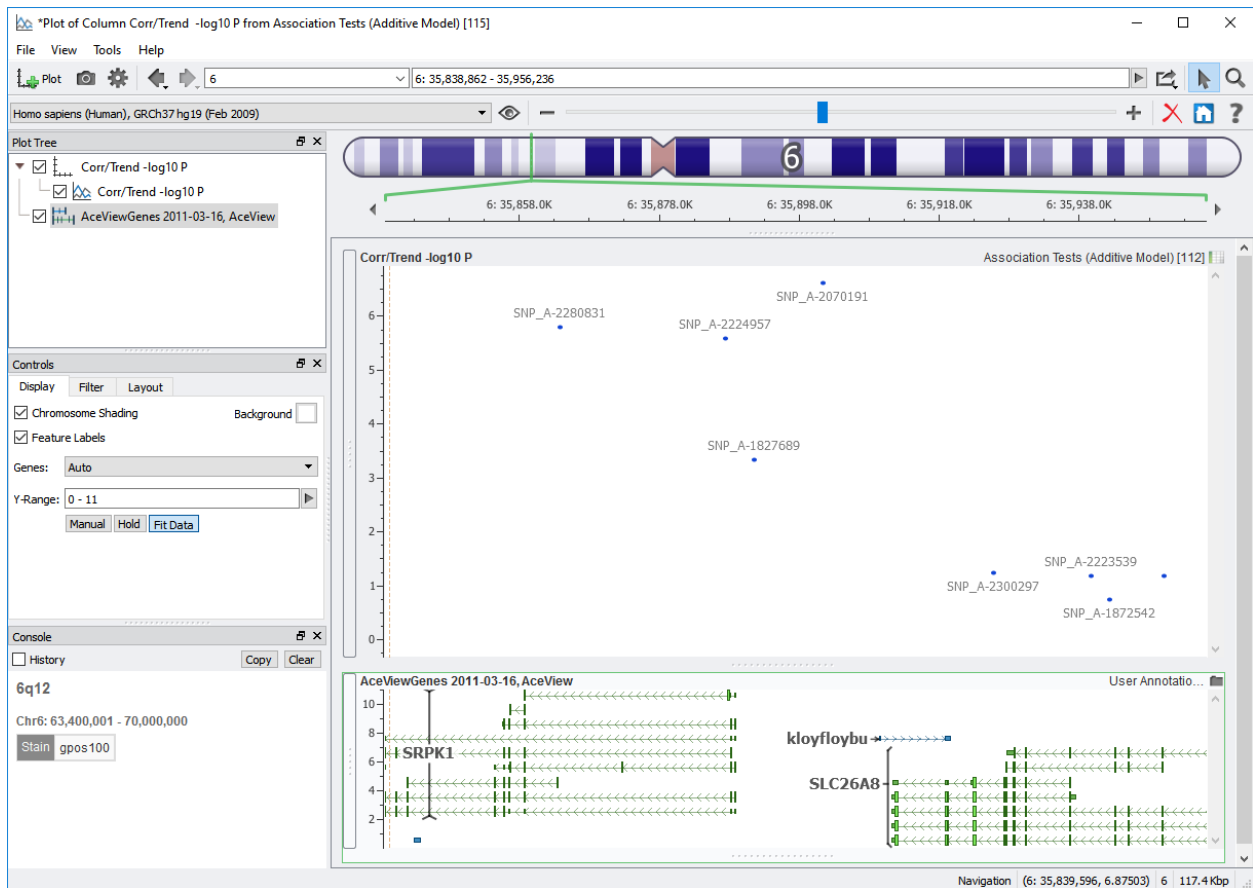


Figure 5-5. Observing the gene track

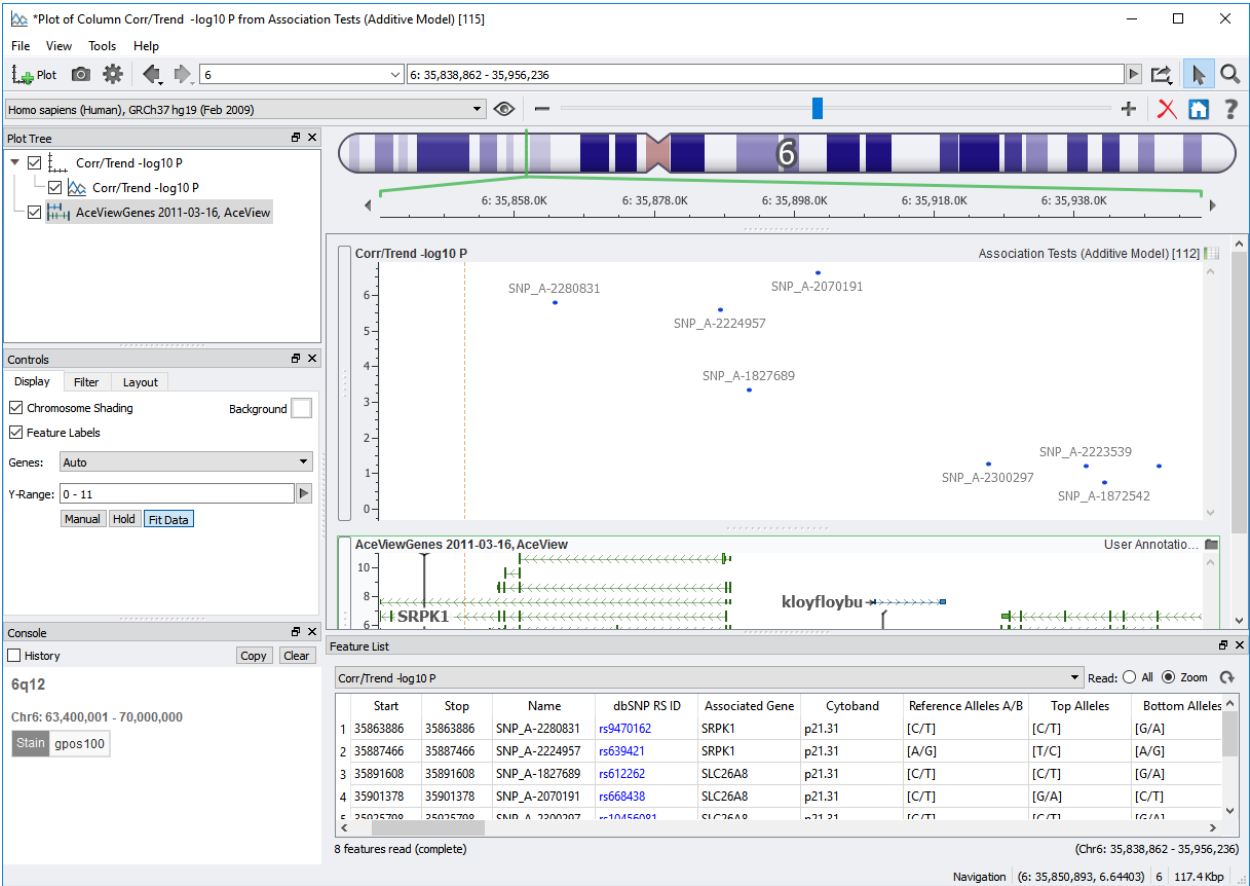


Figure 5-6. P-value Plot with Feature List

D. Creating a Manhattan Plot

Manhattan plots are popular images for publication purposes as they color-code by chromosome making it easy to see where significant markers reside.

- First, turn off the **Feature List** by clicking the X in the upper right corner.
- Then, in the Chromosome selection drop-down select **All** to rest the zoom.
- Select the **Corr/Trend -log₁₀ P** graph item in the Plot Tree.
- Under the **Style** tab, select **Chromosome** in the **Style By:** drop-down.

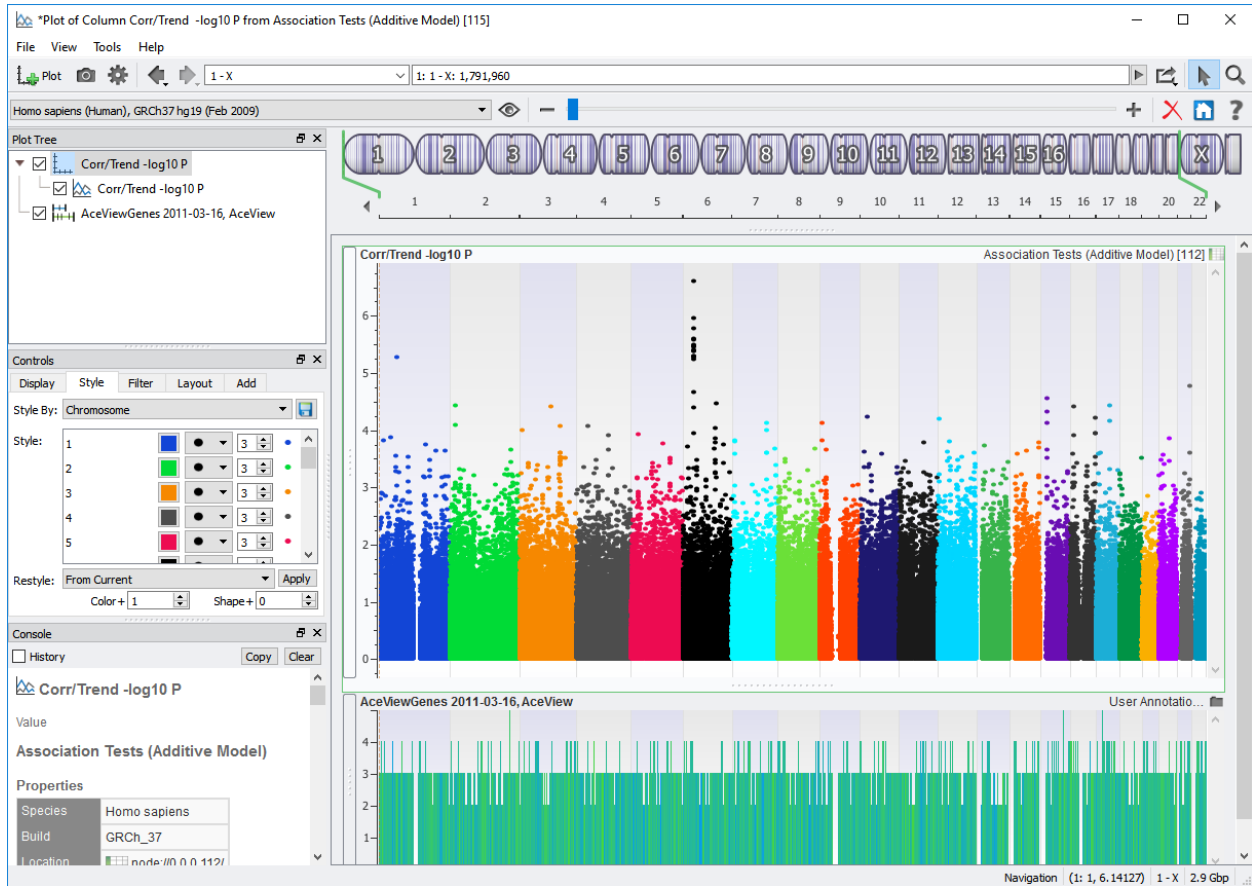


Figure 5-7. Manhattan plot

This will split the graph into 22 different colors, one for each chromosome (Figure 5-7). You can change the color, symbol choice, and size of each chromosome by selecting its respective option in the **Style** options box.

E. Saving Plots as Images

You can save all displayed plots in the plot view as an image.

- Choose **File > Save as Image** from the plot window.

This will bring up a preview window (Figure 5-8). Here you can manipulate various image parameters.

- Uncheck the **Domain View** option under the *Advanced* options.
- You can also change the margins of the image.
- Next, **Browse** to a folder where you want the image saved, give it the name **Manhattan Plot** and click **Save**.
- Click **Save** again at the bottom of the preview window to save the image.
- Once the image is saved close the Plot Viewer and rename its associated node in the Project Navigator to **Manhattan Plot**.

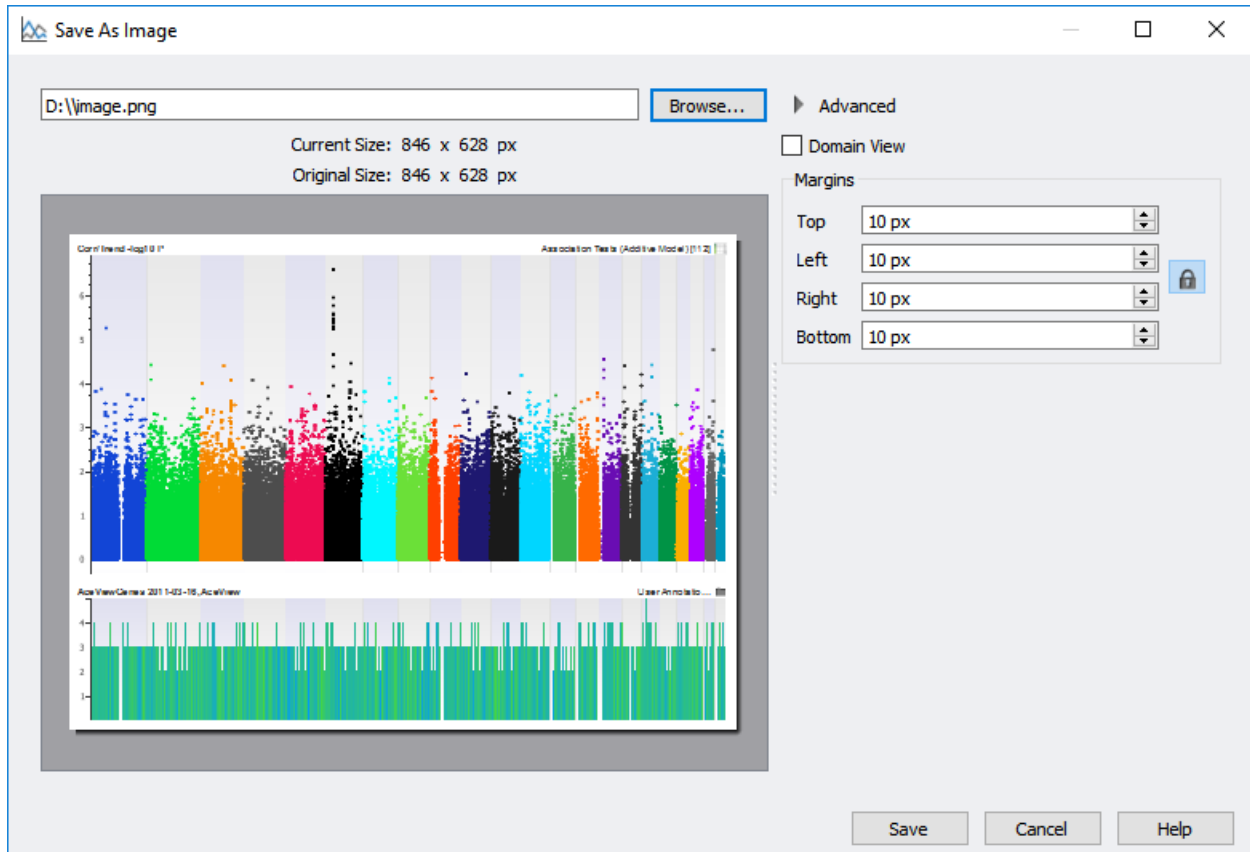


Figure 5-8. Save as Image preview window

You have now performed a cursory genome-wide association study on a case/control phenotype. For more challenging analyses, try running association tests and regression on the other phenotypes. If you click on the first node in the Project Navigator, SNP_GWAS_Tutorial, you will get more information in the User Notes text box on what can be found with each phenotype.