

Tutorial: Whole Genome Copy Number Association using CNAM and HelixTree

Overview

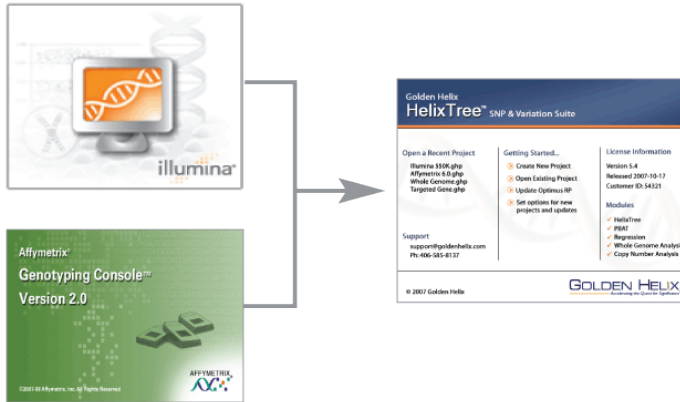
The Copy Number Analysis Module (CNAM), in conjunction with HelixTree, offers several methods for determining where regions of copy number variation occur and performing association analysis either on log₂ ratios directly or on found regions of copy number variation.

This tutorial breaks down the overall approach into seven steps, each consisting of one or more processes. They are:

1. Generate log ratios versus reference samples
2. Identify markers to exclude
3. Correct for batch effects and stratification
4. Perform whole genome log ratio association tests
5. Run segmenting algorithm to generate segmenting mean values and covariate table.
6. Perform association analysis on copy number segment covariates
7. Visualize segmenting results

Depending on your particular dataset and research goals, it may not always be optimal to perform each step sequentially. To guide you, recommendations for how to proceed are provided at the end of each section.

Step 1. Generate Log Ratios Versus Reference Samples.



Before you can perform copy number analysis, you first need a DSF file containing log₂ ratios (from now on referred to as LogRs) created by normalizing raw intensity data against a reference sample. CNAM offers direct support to create LogR DSF files from Illumina and Affymetrix platforms with additional functionality to create them from other providers. This tutorial will focus on preparing a LogR DSF file from Affymetrix CEL files. To learn how to create DSF files from Affymetrix CNT or CNCHP files, Illumina data or data from other providers, see section 4.4 of the HelixTree manual.

The workflow CNAM uses to generate normalized LogRs from Affymetrix 500K, SNP 5.0 and 6.0 CEL files is analogous to the methodology employed by Affymetrix. However, CNAM can perform quantile normalization without gender bias, scale to handle thousands of samples, and allows greater flexibility in choosing a reference set. To learn more about how CNAM processes Affymetrix CEL files, see section 25.9 of the HelixTree manual.

Preparing Files Needed to Process Affymetrix CEL files

Before CNAM can process CEL files, the following are needed:

- Spreadsheet matching NSP and STY (500K only)
- Spreadsheet with CEL file names and column indicating reference status
- Affymetrix marker maps
- Affymetrix library files

Spreadsheet matching NSP and STY (for 500K only)

To properly process both the NSP and STY CEL files, a spreadsheet matching the two needs to be imported into HelixTree. This spreadsheet will tell the CEL import tool how to join the NSP and STY samples together to create one sample per patient in the DSF file. *The matching spreadsheet must have a row label column and at least two data columns. The row labels must be the sample names. The first and second columns must be the NSP file names and the STY file names that are to be joined together, respectively.* Other columns in the data set are optional and may contain the reference status for the sample. If you include the reference status in this spreadsheet, you can also use this spreadsheet to indicate reference status (see below).

The easiest way to import this file into HelixTree is to create a CSV file with respective columns and then select **>File >Import Data >Import ASCII File** from the project navigator window. Make sure to enter the Row Label Column Number representing your sample names.

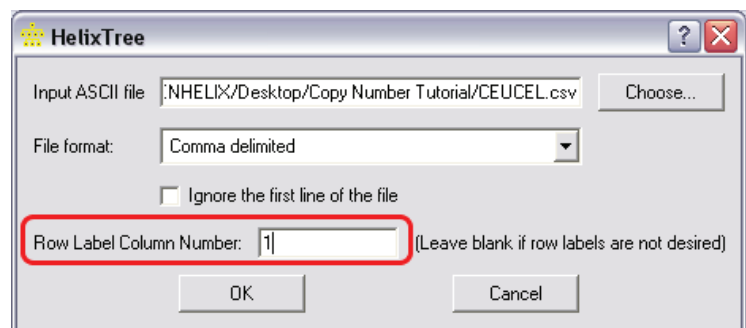


Figure 1. ASCII File Import option with Row Label Column Number set to 1.

The imported spreadsheet should resemble the image below.

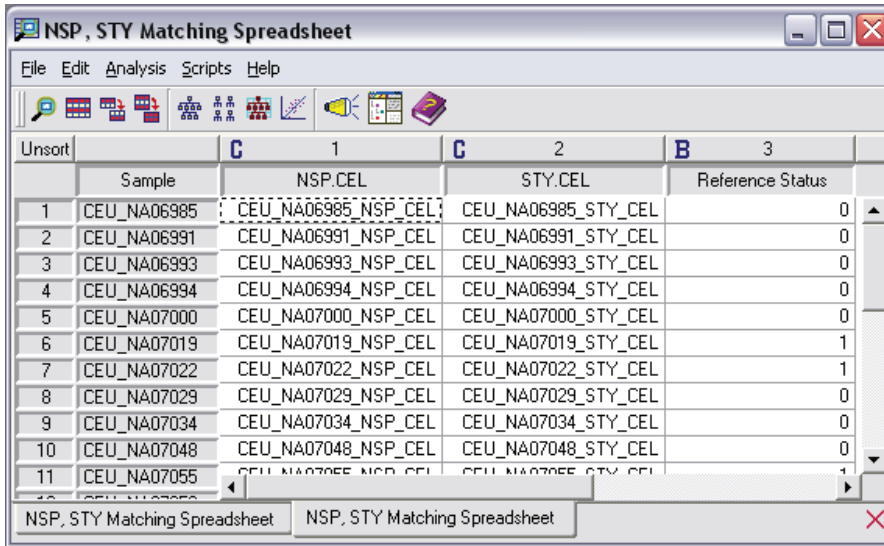


Figure 2. Spreadsheet for matching files from the NSP and STY arrays for 500K analysis.

Spreadsheet with CEL file names and column indicating reference status

This spreadsheet needs two columns, sample names (as row labels) and reference status. For the SNP 5.0 and SNP 6.0 Array, the row labels should be the file names of the CEL files with the “.CEL” extension removed. “0s” should denote samples to be used as references and “1s” should denote non-references.

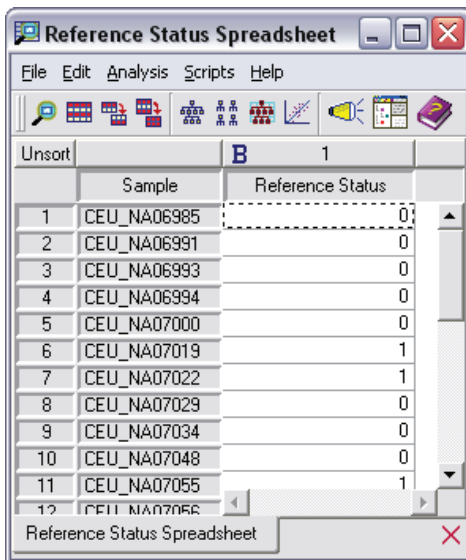


Figure 3. Reference status spreadsheet.

Note: It is up to the researcher to finalize a reference strategy. In CNAM you can use any external or internal samples as your reference set. Affymetrix recommends using at least 25 samples as references in un-paired copy number analysis. As discussed later, if using an external reference set, these samples can be dropped from the corresponding DSF file.

As with the NSP and STY matching spreadsheet above, the easiest way to import this file into HelixTree is to create a CSV file and then select **>File >Import Data >Import ASCII File** from the project navigator window. Make sure to enter the **Row Label Column Number** representing your sample names. The imported spreadsheet should resemble the image to the left (Figure 3).

Affymetrix marker maps (annotation files)

You will need an Affymetrix marker map corresponding to the CEL files you wish to import. Probes not contained in the marker map will not be included in the resulting LogR DSF file. For example, if the marker map does not contain copy number probes, those probes will not exist in the DSF file for copy number analysis.

The latest Affymetrix marker maps can be downloaded using the Affymetrix NetAffx service in HelixTree. To access this feature from the project navigator window, select **>File >Import Data >Download Affymetrix Marker Map**. You will be prompted for your Affymetrix NetAffx login information, which can be freely obtained by registering on Affymetrix’s website. After entering your NetAffx login information, the Download Annotations window will appear listing various Affymetrix annotation files (Figure 4).

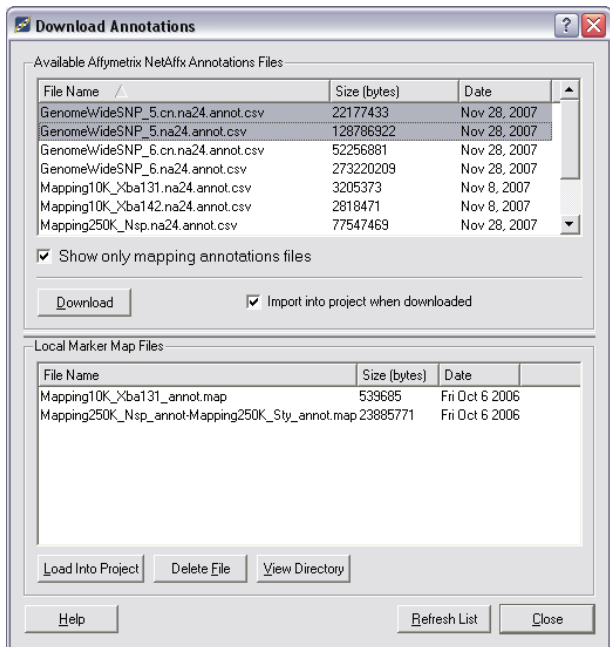


Figure 4. NetAffx marker map download window with both SNP 5.0 marker maps selected.

one or more files from the upper window, and click **Download**. The file(s) will automatically be downloaded to the `../HelixTree/AffyLibraryFiles` directory.

Converting Affymetrix CEL files to LogR DSF File

From the project navigator window, select **>CNAM >Import Affymetrix >Import CEL Files**.

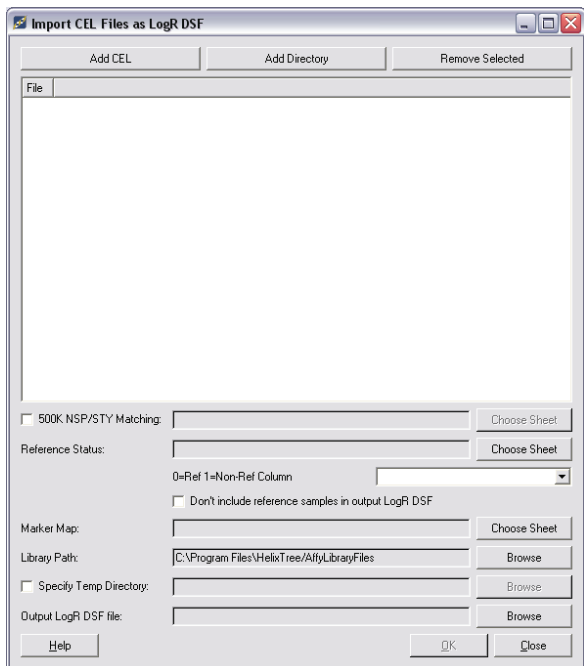


Figure 5. The Affymetrix CEL File Import Dialog in HelixTree.

It is important to note that there are actually two annotation files for each of the 500K, 5.0 and 6.0 files (500K = NSP + STY, SNP 5.0 and 6.0 = SNP + CN probes). For each file set, both corresponding annotation files need to be downloaded at the same time for HelixTree to properly merge them. To do this, highlight both annotation files (as seen in Figure 4), make sure the box **Import into project when downloaded** is checked and click **Download**. If you downloaded the annotation files previously, they should show up as a single merged file in the lower section of the window. If this is the case, you can just highlight that file and click **Load Into Project**.

Affymetrix library files

Similar to Affymetrix marker maps, Affymetrix library files can be downloaded using the NetAffx service in HelixTree. The library files should contain both SNPs and CN Probes when appropriate.

To download library files, select the **>Tools >Download Affymetrix Library File** menu option from the project navigator window. After entering your authentication information, HelixTree will load a list of library files available through the NetAffx service. To download library files, select

one or more files from the upper window, and click **Download**. The file(s) will automatically be downloaded to the `../HelixTree/AffyLibraryFiles` directory.

From the CEL file import dialog (Figure 5), first select the CEL files you want to include in the data set. For 500K data, you must select files from both the NSP and STY arrays for each sample. To select CEL files, click the **Add CEL** button, navigate to the appropriate folder and select multiple CEL files (you can hold down the **Shift** key to select multiple files at once). The CEL files you selected will appear in the CEL import dialog window (Figure 6). You may add all of the CEL files in a directory by using the **Add Directory** button. This is especially helpful if you store your NSP and STY files in separate directories. To remove CEL files from the window, select the unwanted samples and click **Remove Selected**. You may continue adding CEL files by clicking the **Add CEL** or the **Add Directory** buttons again.

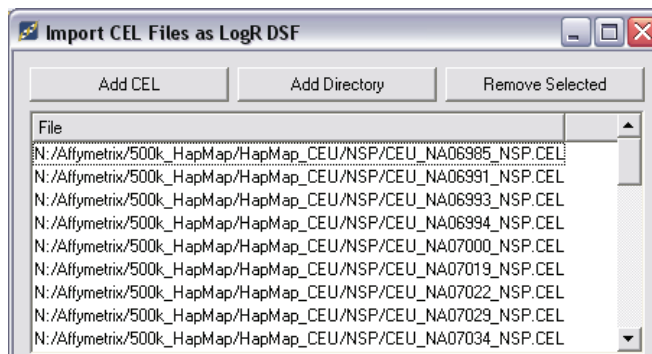


Figure 6. CEL files selected for conversion into a log2 ratio DSF file.

For 500K CEL file import, next check the 500K NSP/STY Matching check box and select the matching spreadsheet previously imported to be used. If you are importing 5.0 or 6.0 CEL files, leave this box unchecked.

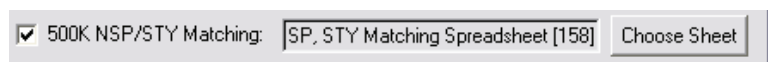


Figure 7. Selecting a spreadsheet to use for NSP and STY array matching.

Next, select a spreadsheet containing the **Reference Status** for the samples. When a spreadsheet is selected, the **0=Ref 1=Non-Ref Column** drop down box will fill with the different binary data column names in the selected spreadsheet. Select the name of the column to be used as the reference status.

Note: The gender of the reference samples should be considered for copy number analysis of the X and/or Y chromosomes.

Check the box **Don't include reference sample in output Log R DSF** if you are using external reference samples (e.g. HapMap data) and do not want them included in the resulting DSF file.

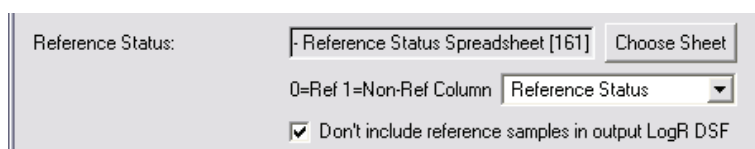


Figure 8. Spreadsheet and spreadsheet column selected for determining reference status.

Next, select the **Marker Map** previously imported to be used in the analysis and choose the **Library Path** where the CDF library files for the appropriate array can be found. These reside in the directory where you previously saved them.

Note: After using the CEL import tool for your given array (500K, 5.0, 6.0), an AffyLibraryFiles directory will be created in the HelixTree installation directory containing *.gcdf library files. These files can be used from that point on instead of CDF files.

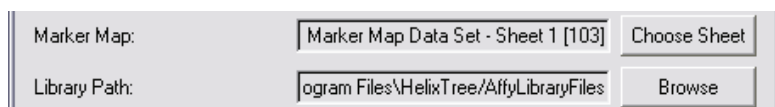


Figure 9. Marker map spreadsheet selected and library directory where CDF (or *.gcdf) library files are located.

You have the option to specify a **Temp Directory** (Figure 10) where intermediate DSF files will be stored. If your project is located on a shared network drive (not recommended), you should specify a **Temp Directory** on a local disk. Finally select the **Output LogR DSF** location and click **OK** to begin the import.

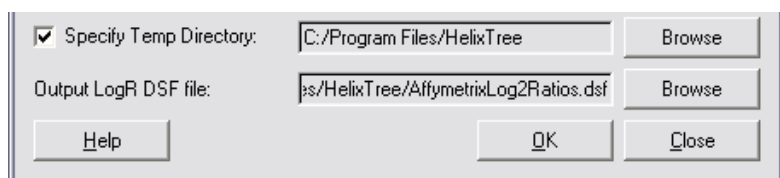


Figure 10. Optional temporary directory and output DSF file name selected.

The conversion will take several minutes per CEL file to complete. The DSF file created is ready to be imported and analyzed using either the Copy Number Segmentation tool or the LogR Association Tests and PCA window.

From here you can proceed to **Step 2:** Identify markers to exclude, **Step 3:** Correct for batch effects/stratification, **Step 4:** Perform whole genome log ratio association tests, or **Step 5:** Run segmenting algorithm to generate segment means and associated covariates.

Step 2. Identify Markers to Exclude.

It is sometimes desirable to filter out problematic markers before running segmentation, such as those with low call rates or gender-associated markers caused by effects of poorly randomized experiments. This tutorial will focus particularly on identifying gender-associated markers and creating a list of these markers to be excluded when performing LogR association tests and copy number segmenting (**Steps 4 and 5**). Other measures of quality control can be applied by following similar workflows.

To find gender-associated markers you need to perform an association test with gender as the dependent variable and LogRs as the independent variables. This can be done rather easily using the LogR Association Tests and PCA window in CNAM.

Select >CNAM >LogR Association Tests and PCA.

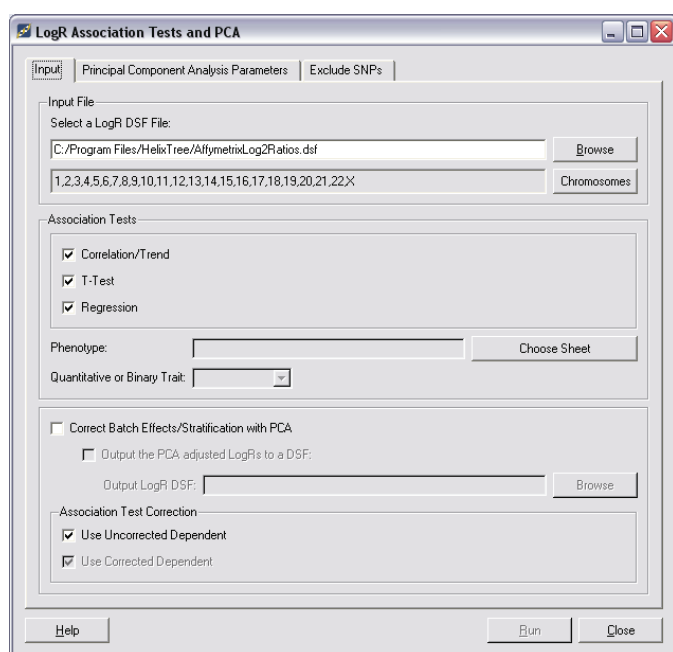


Figure 11. LogR Association Tests and PCA window.

From this window, **Browse** to the LogR DSF File created in **Step 1**. You can then choose which **Chromosomes** you want to perform association tests on. In this case, select all chromosomes – this will be the default if your LogR DSF contains information for all chromosomes. Then, **Check** the association tests you would like to perform.

Next, you need to select a spreadsheet within HelixTree containing your phenotype information - gender in this case. If you have not already imported your phenotype spreadsheet, from the project navigator window go to **>File >Import Data** and choose either **>Import Wizard** or **>Import ASCII File**. Make sure to indicate the column with your sample names as the row label column. This will ensure each sample's gender status is appropriately matched to its LogRs when performing the association test.

Once your phenotype spreadsheet is imported into HelixTree, from the LogR Association Tests and PCA window click **Choose Spreadsheet**, select the appropriate phenotype spreadsheet and click **OK**.

Having selected a phenotype spreadsheet, the **Quantitative** or **Binary Trait** box will be activated. Scroll down and select the gender column to make it the dependent variable.

Now you can perform association test(s) on gender. You may also simultaneously correct for batch effects or stratification using principle component analysis. This will be covered in **Step 3**, so for now **Uncheck** the **Correct for Batch Effects/Stratification with PCA** box. Click **Run**.

The result is a spreadsheet with markers as rows and p-values for each test statistic as columns (Figure 12).

Note: The p-values in this spreadsheet are not Bonferroni corrected.

Unsort	Label	R	7	R	8	R	9
			Corr/Trend R (Uncorrected)		T-Test P		Regression P
1	SNP_A-1909444		-0.00377131119582981		0.823508363199204		0.823492259119141
2	SNP_A-2237149		0.00297273861788001		0.860445447603698		0.860370113324788
3	SNP_A-4303947		0.0194605599286368		0.191918833784628		0.248892648619752
4	SNP_A-1886933		-0.00137270647033976		0.935297659738768		0.93527917183582
5	SNP_A-2236359		-0.00171814990235735		0.919065538794702		0.91903198312007
6	SNP_A-2205441		-0.002438045202298		0.895354075894425		0.885336284721447
7	SNP_A-2116190		-0.000358987451739524		0.983061865247547		0.98305571949784
8	SNP_A-4291020		0.0150409345276207		0.373698887332074		0.373183707330558
9	SNP_A-1902458		-0.0129743684393132		0.442884085341935		0.443055460974337
10	SNP_A-2131660		-0.0045808063208643		0.786461196029568		0.786410156719538
11	SNP_A-2109914		0.00161390723349787		0.923960583210927		0.92393626854908
12	SNP_A-2291997		0.0182299427396875		0.2053954472190104		0.280420071469763
13	SNP_A-4277872		0.00799803456117563		0.636207005526453		0.636393022730367
14	SNP_A-4221087						

Figure 12. P-value spreadsheet that results after performing LogR association tests.

You can plot each column to visually inspect the markers associated with gender by **Left Clicking** on the column number and selecting **Plot this Column**. You should see a plot similar to Figure 13 with significant peaks across the genome, signifying gender-associated markers.

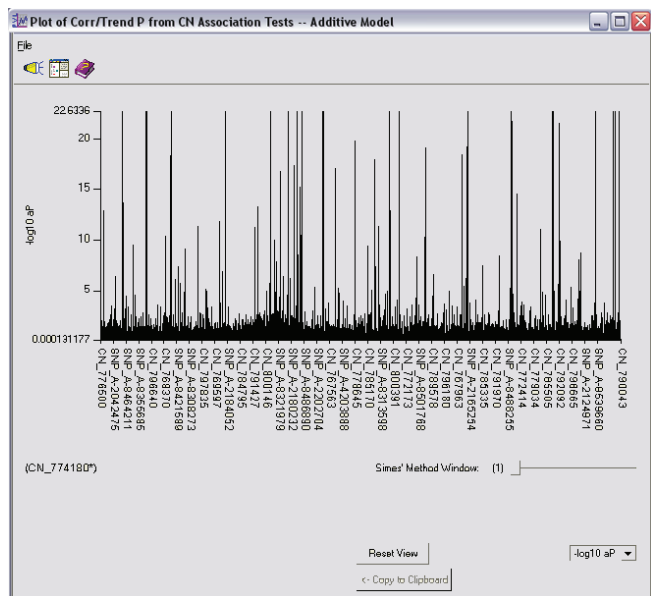


Figure 13. P-value plot representing gender-associated markers.

Now you need to create a row subset spreadsheet of only those markers whose p-values are less than a given cutoff (e.g. .01).

To do this, first go back to the p-value spreadsheet, **Left Click** on the column number again and this time select **Sort Ascending**. Scroll down to the first marker that exceeds your cutoff p-value and **Left Click** that marker's row label once.

This will turn the entire row grey, which means it is now inactive. To inactivate the rest of the markers above the cutoff, hold down the **Shift** key, scroll to the bottom of the spreadsheet and **Left Click** the last marker's row label. All markers greater than the p-value cutoff should now be inactivated, and all markers less than the cutoff should be activated (Figure 14).

Unsort	Label	T-Test P	Regression P
278	SNP_A-4270717	0.00935179304237022	0.00944028333258958
279	SNP_A-2272357	0.00940431732301606	0.00937219846291608
280	SNP_A-2003037	0.00949965135187087	0.00951882922146557
281	SNP_A-2189724	0.00954602505183484	0.00953102289084249
282	SNP_A-4286047	0.00959405048413044	0.0096017541064616
283	SNP_A-2247741	0.00965619176117508	0.00965967933700473
284	SNP_A-1983139	0.0097466983284939	0.00930675691598769
285	SNP_A-1911744	0.00975378186335047	0.00983452890012235
286	SNP_A-2248627	0.00999713404121669	0.00995728603223884
287	SNP_A-2303859	0.0100095384214313	0.00992814641650855
288	SNP_A-1788619	0.0100329865792353	0.0100677115256992
289	SNP_A-4249323	0.0101322339218938	0.00984037597610854
290	SNP_A-2238439	0.0101397180111156	0.0101989505372926
291	SNP_A-1888457	0.0101529102586662	0.0101953898671121
292	SNP_A-1935730	0.0101550758938977	0.00978139655487515
293	SNP_A-2019351	0.0102181721159165	0.010223332761201
294	SNP_A-2250476	0.0103219693139253	0.0102033313554147
295	SNP_A-4230709	0.0105683159938847	0.0105740435858816
296	SNP_A-2280591	0.0105745402228189	0.010398488219958
297	SNP_A-1820126	0.0106586752313221	0.0106903146481995
298	SNP_A-4247902	0.010820527583614	0.0109017592019541
299	SNP_A-4270572	0.0108287491883567	0.010543366460746
300	SNP_A-2158865		

Figure 14. Inactivating markers greater than p-value .01.

Next, from the spreadsheet menu, select **>Edit >Row >Subset Spreadsheet**. This will create a new spreadsheet with only the markers below the p-value cutoff (or those activated in the original spreadsheet). Finally, export this spreadsheet as a CSV file by selecting **>File >Save a Comma-Delimited Text File**. Browse to a path where you want the CSV file saved and click **Save**.

Next, you need to delete all the columns in the CSV file except for the first containing the marker names. This can be done easily within Excel or other spreadsheet editing programs. Once you have a CSV file with a single a column of marker names, it can be used in **Steps 4 and 5** to exclude markers when performing LogR association tests and copy number segmenting.

You can now proceed to **Step 3: Correct for batch effects/stratification** or skip to **Step 4: Perform whole genome LogR association tests** or **Step 5: Run segmenting algorithm to generate segment means and associated covariates**.

Step 3. Correct for Batch Effects and Stratification

Sometimes, finding an association can be confuted by population stratification. This is because a condition may be more prevalent in one group of samples than in a different group, and therefore there will be a spurious association between the condition or trait being tested for and any genetic characteristics varying between the two different groups.

While it is good practice for studies to be based upon the most homogenous test subjects as possible, it has been noted that even those who classify themselves as “Caucasian” have mild variation in genetic characteristics problematic enough to confound a study done over thousands of genetic markers.

Additionally, it has been our experience, especially with copy number analysis, that there is evidence of variations in test equipment confounding studies. These are referred to as batch effects, which are often the result of improperly randomizing the genotyping of cases and controls, males and females, etc. For example, perhaps all cases (or all females) were done at one site or day and controls (or males) at another.

This tutorial will lead you through correcting LogRs for batch effects and stratification using an Eigenstrat-based principal component analysis (PCA) method. The result will be a new DSF file with PCA corrected LogRs. This file can then be used to perform LogR association tests or copy number segmenting (**Steps 4** and **Step 5**).

Determining How Many Principle Components to Use

Before you apply PCA to your LogRs, you need to determine how many principal components to use in the analysis. Determining this is to some degree an open question. First off, if you choose as many components as there are markers, you will wind up subtracting out ALL effects, thus getting nothing from your tests.

A better option is to determine the components themselves with their corresponding Eigenvectors and then look at the pattern of the Eigenvalues.

To do this, from the project navigator window select >**CNAM** >**LogR association tests and PCA**.

Note: From this window you can also perform LogR association tests (**Step 4**) at the same time as correcting for batch effects or stratification. This tutorial will treat each step separately.

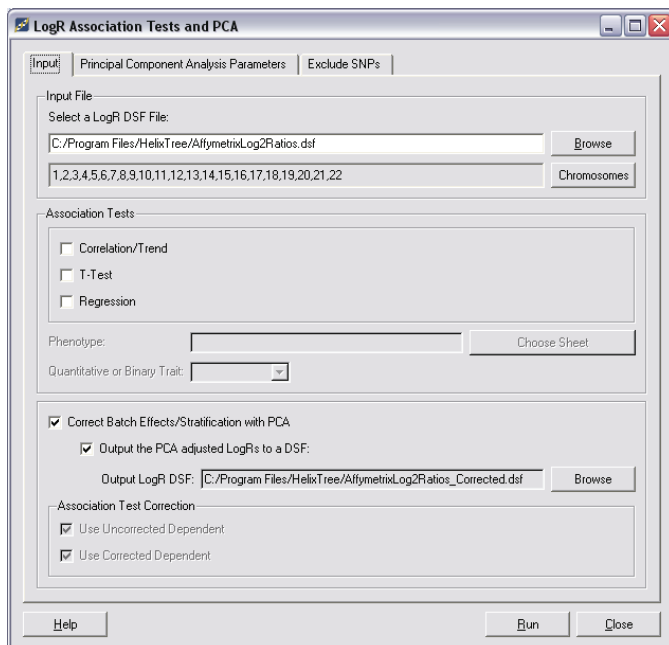


Figure 15. LogR Association Tests and PCA window ready to perform PCA analysis on LogR DSF file.

From this window, **Browse** to the LogR DSF file created in **Step 1** and click **Open**. Notice by default all the chromosomes will appear in the box below the input DSF file location. Only those chromosomes appearing here will be corrected.

Note: Including the X chromosomes may result in an erroneous systematic shift of the data, so we recommend excluding them from PCA analysis and adjusting the first 22 chromosomes. To do this click **Chromosomes**, uncheck 'X' and click **OK**.

Since you are not performing association tests during this step, **Uncheck** the three association test boxes. Also **Uncheck** the **Correct Batch Effects/Stratification with PCA** box since you first need to learn how many principal components to use before applying PCA.

Next, click on the **Principle Component Analysis Parameters** tab.

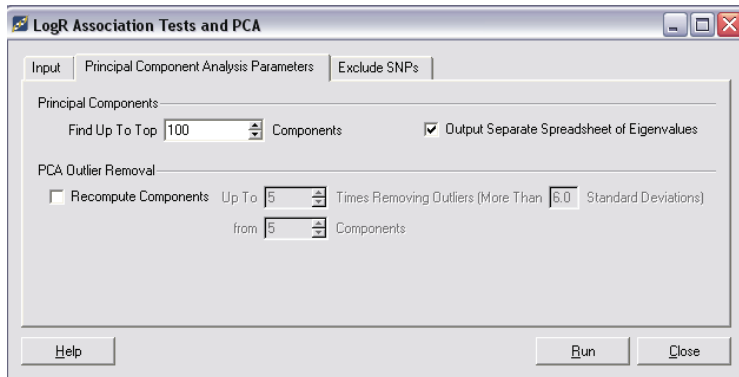


Figure 16. Principal Component Analysis Parameters tab set to find 100 principal components.

In this tab enter the number of principal components you want CNAM to find. For this first pass analysis, you should choose to find more than you believe are necessary. Beware that the more you enter, the longer the analysis will take. Next, check the **Output Separate Spreadsheet of Eigenvalues** box. For the purpose of this tutorial, do not worry about **PCA Outlier Removal**. Click **Run**.

When the analysis finishes, two spreadsheets are created, one containing Eigenvalues (Figure 17a) and the other principal components (Figure 17b). At this point, you are still deciding how many principal components to use, so you can delete the principal component spreadsheet.

Unsort	R	1
		Eigenvalue
1		63.2443901152054
2		18.7829048068888
3		10.7775568822379
4		8.3313612992495
5		5.60213501290017
6		4.85294501311866
7		3.89808921309776
8		3.4535059598795
9		2.76799234101659
10		2.54466596558315
11		2.26999033788234
12		2.13781868995749
13		

Figure 17a. Eigenvalues spreadsheet.

Unsort	R	1	R	2	R	3	R	4	
		EV = 63.2444			EV = 18.7829			EV = 10.7776	
		EV = 8.33136							
1	CEU_NA06985	-0.012626869		0.010334601		0.007304328		-0.021451419	
2	CEU_NA06991	-0.014542525		0.009906947		0.001879659		-0.023395637	
3	CEU_NA06993	-0.01100306		0.010912962		0.008529789		-0.02622865	
4	CEU_NA06994	-0.013021929		0.010529939		0.006428132		-0.019939854	
5	CEU_NA07000	-0.012433762		0.012604917		0.00305411		-0.029122289	
6	CEU_NA07019	-0.014046205		0.013648926		-0.003585103		-0.018744678	
7	CEU_NA07022	-0.012442558		0.010994228		0.002786595		-0.017174756	
8	CEU_NA07029	-0.013373914		0.012000247		0.002078799		-0.013594341	
9	CEU_NA07034	-0.013227088		0.012216104		0.005792213		-0.012735367	
10	CEU_NA07048								

Figure 17b. Principal component spreadsheet.

From the Eigenvalue spreadsheet, which contains one column of Eigenvalues, **Left Click** on the column number (1) and choose **Plot this Column**. You should get a plot similar to Figure 18.

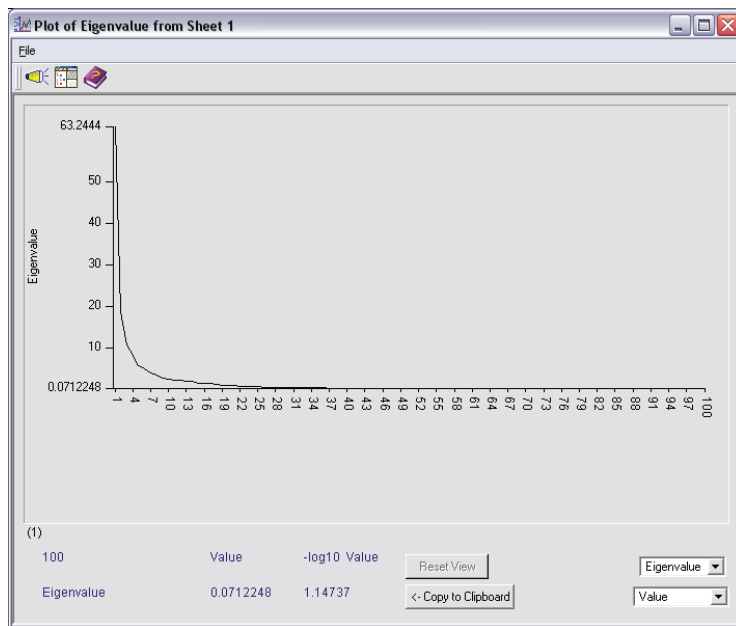


Figure 18. Plot of Eigenvalues.

Though there is currently no optimal way to decide how many principle components to use, if the first few Eigenvalues are very large compared with the remaining Eigenvalues, then a reasonable guideline is to use that many components in the second analysis when applying the PCA technique. In Figure 18, notice how the size of the Eigenvalues begins to smooth out around 10. Using more than nine may result in diminishing returns and may even adversely affect the results.

Note: For a more thorough discussion of stratification, principal components analysis, and Eigenvalues see:

Price, Alkes L., Patterson, Nick J. Plenge, Robert M. Weinblatt, Michael E. Shadick, Nancy A. Reich, David. (2006). Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. Nature Genetics 38, 904-909.

Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genetics 2(12): e190. doi:10.1371/journal.pgen.0020190.

Applying PCA Correction

Now that you have determined the number of principle components to use, you can go back into the LogR Association Tests and PCA window to correct for batch effects and stratification.

Again, **Choose** the DSF file you want to correct and **Uncheck** all the association test boxes.

Next, make sure the **Correct Batch Effects/Stratification with PCA** box is checked this time.

You also need to check **Output the PCA adjusted logRs to a DSF**. This allows you to import and segment the PCA corrected data using the Copy Number Segmentation window (**Step 5**), or use it as the input DSF in this window to run LogR association tests (**Step 4**) without having to re-run PCA correction each time. **Browse** to the path where you want to save the corrected DSF, name it and click **Save**.

Note: It is good to give the DSF file an intuitive name, such as one referring to the number of principal components used. The corrected LogRs will differ depending on how many principal components are used.

Click on the **Principal Component Analysis Parameters** tab.

Here you need to enter the number of principle components you determined previously. This time it is not necessary to output the Eigenvalue spreadsheet so you can **Uncheck** this box. Again, for this tutorial don't worry about PCA outlier removal.

Next, click on the **Exclude SNPs** tab.

From this tab, you can now exclude the markers you identified in **Step 2**. **Browse** to the CSV file created in **Step 2** and click **Open**. A list of markers from the CSV file should now be listed as in Figure 19.

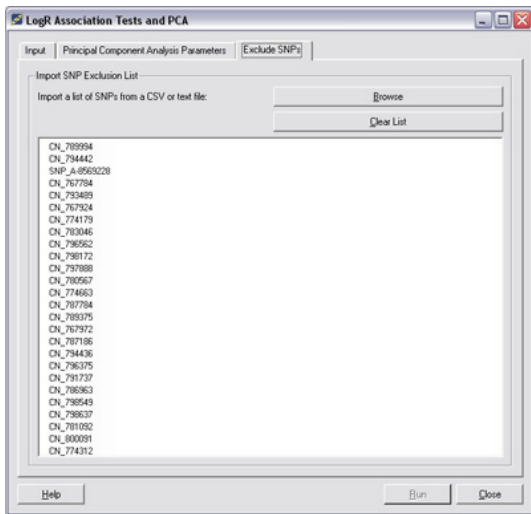


Figure 19. Exclude Markers tab.

When you are finished, click **Run**.

The result for this particular analysis is a new DSF file containing PCA corrected LogRs minus the excluded markers and a secondary principal component spreadsheet.

From here you can proceed to perform LogR association tests (**Step 4**) or perform copy number segmentation (**Step 5**) on the corrected DSF.

Step 4. Perform Whole Genome Log Ratio Association Tests.

It is advantageous to perform association directly on LogRs prior to running copy number segmentation (**Step 5**) because the computation time is much faster and it provides a good first look at the data. Please note however, the number of LogRs will be much greater than the number of copy number segment covariates resulting in a greater multiple testing penalty.

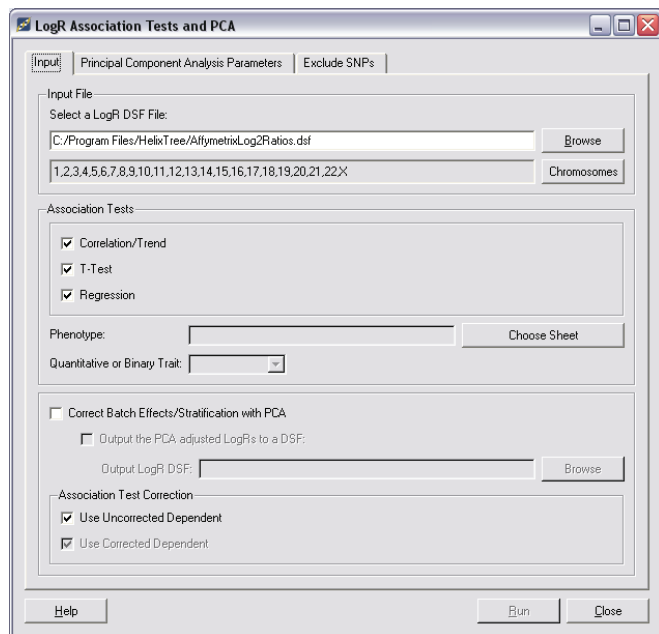


Figure 20. LogR Association Tests and PCA Window.

spreadsheet, from the project navigator window go to **>File >Import Data** and choose either **>Import Wizard** or **>Import ASCII File**. Make sure to indicate the column with your sample names as the row label column. This will ensure each sample's phenotype status is appropriately matched to its LogRs when performing the association test(s).

Once your phenotype spreadsheet is imported into HelixTree in the LogR Association Tests and PCA window, click **Choose Spreadsheet**, select the appropriate phenotype spreadsheet and click **OK**. A list of all quantitative and binary variables in the phenotype spreadsheet will now appear in the next box. **Select** the variable you want as your dependent.

Because the DSF file selected above already contains PCA corrected LogRs, uncheck the **Correct Batch Effects/Stratification with PCA** box. This will inactivate all PCA related options, including the parameters on the next tab.

If you wish to exclude other markers than those already excluded in **Step 3**, go to the **Exclude Markers** tab, **Browse** to the CSV file with the additional markers and click **Open**.

You are now ready to perform association. Click **Run**.

The result will be a p-value spreadsheet (not shown) with all the marker names in the first column and their corresponding p-values in each additional column.

You can plot each column by **Left Clicking** on the column number and selecting **Plot this Column**.

You can perform LogR association tests on any LogR DSF file. We recommend first correcting for batch effects and stratification and filtering out problematic markers as covered in **Steps 2** and **3**.

Note: LogR association tests can be performed simultaneously while correcting for batch effects and stratification and excluding problematic markers, all from the same window. This tutorial works through performing each process separately.

To get started, from the project navigator window select **>CNAM >LogR Association Tests and PCA**. In this window (Figure 20), **Browse** to the PCA corrected LogR DSF file created in **Step 3** and select the **Chromosomes** you want to test. **Check** the box next to the association test(s) you want to run.

Next, you need to select a spreadsheet within HelixTree containing your phenotype information, in this case, case-control status or quantitative trait. Your phenotype spreadsheet should have already been imported during **Step 2**. If you have not already imported your phenotype

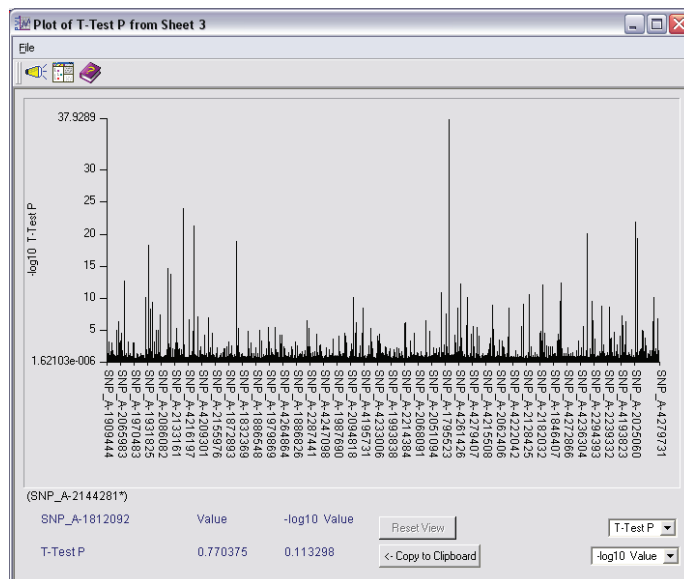


Figure 21. P-value plot from t-tests on PCA corrected LogRs.

In some cases, there may still be spurious associations across the genome (as seen in Figure 21). To reduce some of the noise you can apply a **Median Smooth** script to the p-values of each column. The Median Smooth script calculates a moving median centered about a given observation plus/minus a user-defined window size. This script is not provided with the software but can be downloaded from our online script repository.

http://www.goldenhelix.com/SNP_Variation/scripts/index.html

To use the Median Smooth script, **Download** the **mediansmooth.py** file from the webpage above and **Save** it in your **../HelixTree/scriptsht/spreadsheet/edit/** directory. Next, from the p-value spreadsheet select **>Edit >Median Smooth**. A dialogue window will appear asking for a moving window size (Figure 22).

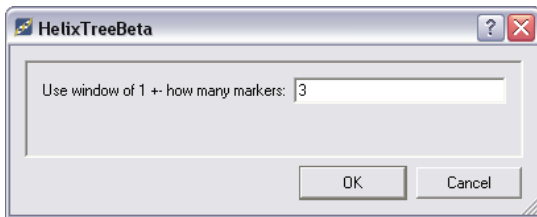


Figure 22. Median Smooth script with a 3 marker window size.

Enter an appropriate window size (i.e. 3) and click **OK**.

Note: A window size of 3 would include 7 observations

The result is a new p-value spreadsheet with the original p-value columns and new median smoothed p-valued columns. You can again plot each column by **Left Clicking** on the column header and selecting **Plot this Column**.

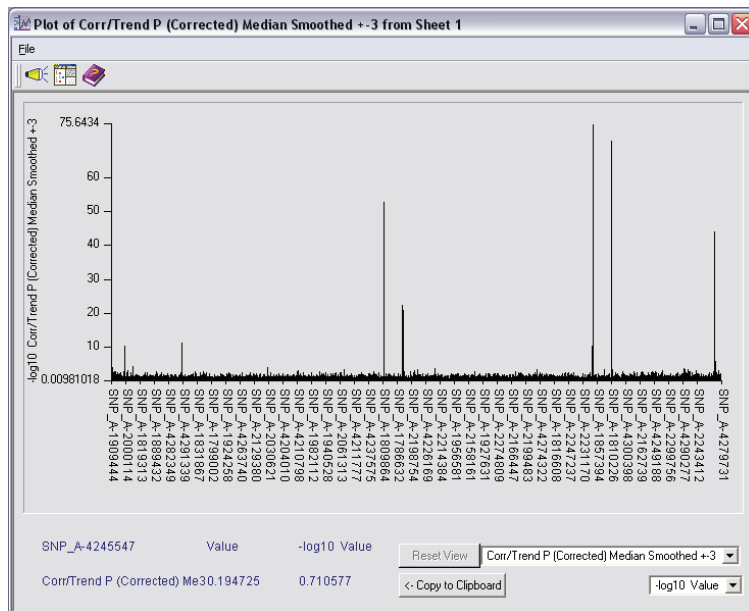


Figure 23. Plot of median smoothed p-values.

You should see far fewer spurious associations in a smoothed plot (Figure 23) versus an unsmoothed plot.

By **Left Clicking** and **Dragging** on the plot you can zoom into a region. Zooming in on a significant spike (Figure 24) should reveal a region spanning several markers with multiple significant p-values.

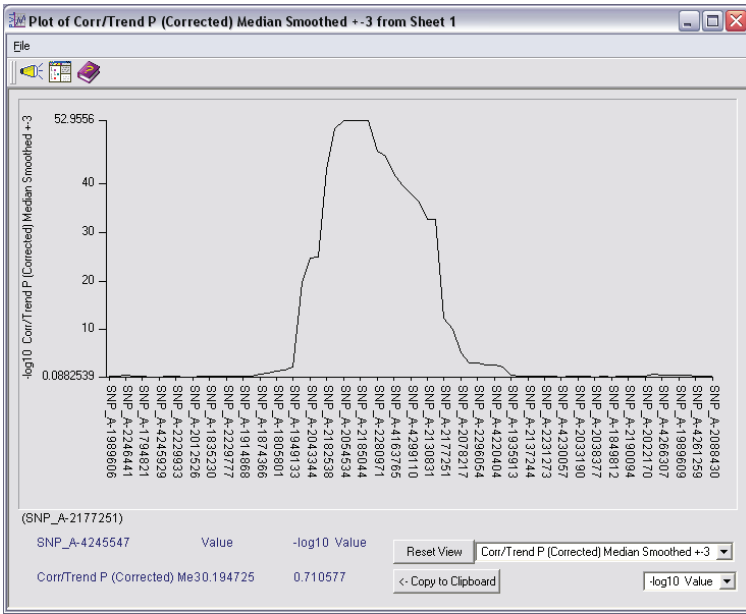


Figure 24. Zoomed p-value plot of ~10 marker significant peak.

A quick way to determine where significant markers reside on the genome and what genes they are associated with is to join the p-value spreadsheet with a marker map containing these markers. Such a marker map was imported in **Step 1** and should already be in the project.

To do this, go back to the Median Smoothed p-value spreadsheet and select **>File >Join Spreadsheets on Row Labels**. Select the appropriate marker map spreadsheet and click **OK**. The result is a spreadsheet with marker names, p-values and marker map information.

From here you can go to **Step 5** to run the Golden Helix segmenting algorithm on only those chromosomes revealing significant LogR association results or the entire PCA corrected LogR DSF.

Step 5. Run Segmenting Algorithm to Generate Segmenting Mean Values and Covariate Table.

CNAM employs an optimal segmenting algorithm that utilizes dynamic programming to exhaustively search through all possible change-points in LogR data to discover regions of markers in which LogRs vary significantly from region to region. These regions will generally be where there is copy number variation in your data.

The segmenting process is optimized by working at three levels:

1. If desired, the region of markers is subdivided into a moving window of sub-regions.
2. A unique segmenting algorithm is applied to find multiple segments wherever possible, and by implication, segment boundaries which we term “cut-points”.
3. A permutation algorithm is applied to validate the found cut-points.

CNAM offers two types of segmenting methods, univariate and multivariate. These methods are based on the same algorithm, but use different criteria for determining cut-points. The multivariate method segments all samples simultaneously, finding general copy number regions that may be similar across all samples. This method is preferable for finding very small copy number regions, and for finding conserved regions that may be useful for association studies. For a given sample, the covariate is the mean of the LogRs within each segment for that sample. If there are consistent positions for copy number variation across multiple samples, the copy number segments will be found. In reality, there may not always be consistent copy number segments across multiple samples. The univariate method segments each sample separately, finding the cut-points of each segment for each sample and outputting a spreadsheet showing all cut-points found among all samples.

This tutorial will focus on performing multivariate segmentation. The result will be two spreadsheets. The first is a copy number segment covariate spreadsheet which contains the mean LogR value for each sample within each segment of markers. The second is a segment means spreadsheet which contains columns for the chromosome number, segment start position, segment end position, segment mean, and the segment length in number of markers. Optionally, a Wiggle file may also be generated with found regions and can be viewed using a supported genome browser.

Performing Copy Number Segmentation

From the project navigator window, select **>CNAM >Copy Number Segmentation**. The window at (Figure 25) right will appear.

As with LogR association tests, you can perform segmentation on any LogR DSF file, though we recommend first correcting for batch effects and stratification and filtering out problematic markers as was done in **Steps 2 and 3**.

Browse to the LogR DSF file created in **Step 3** and select the Chromosomes you want to segment.

Note: Segmentation takes a fair amount of time to run. You may want to start with only those chromosomes that showed interesting regions from the LogR association tests in **Step 4**.

The next section, **Import Options**, can be somewhat confusing. There are two options, **Import Segmenting Results** or **Import Raw Data**. These are asking whether you want to import the results of the segmentation performed in this step into HelixTree or simply the LogRs of the DSF selected above. Notice if you select **Import Raw Data**, all the segmentation options below will be inactivated because segmentation will not be performed.

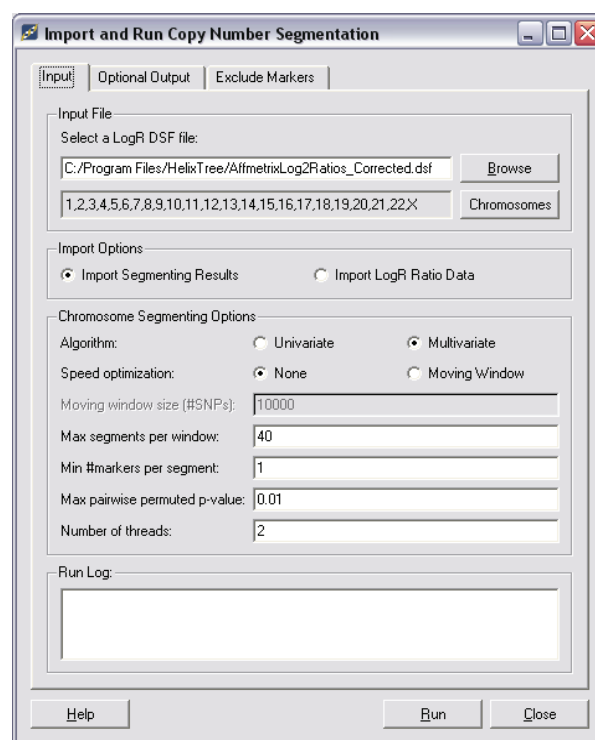


Figure 25. Copy Number Segmentation window.

Why import the raw data into HelixTree? Importing the raw data will enable you to perform tree-based association on the LogRs rather than the copy number segment means. With a few minor workarounds you can also visualize copy number segments directly in HelixTree rather than with an external genome browser.

Importing raw data into HelixTree is not covered in this tutorial. If you want to perform this operation, it is recommended that you either create a smaller DSF file based on a smaller sample size in **Step 1** or only import a single chromosome at a time. This is necessary because LogRs cannot be sparsely stored like SNP data, resulting in a spreadsheet that may exceed available computer RAM causing the program to crash. For this reason, we developed the **LogR association test and PCA** window as covered in **Step 4**.

Select **Import Segmenting Results**.

The next section, **Chromosome Segmenting Options**, allows you to choose which method you want to use for segmenting (multivariate or univariate) and to optimize the speed of the analysis. For the purpose of this tutorial, select the **Multivariate** algorithm and leave the rest of the parameters as the default values. For more information about these parameters, see section 25.3.4 of the HelixTree manual.

Next, click on the **Optional Output** tab. Check the **Optional Bookmark File Output** box if you want to export the segment means to a UCSC Wiggle Track (WIG) format file. WIG files allow you to view segment results in supported genome browsers such as UCSC's Genome Browser and Affymetrix's Integrated Genome Browser (IGB). Viewing segment results in UCSC's Genome Browser will be covered in **Step 7**.

Note: If you are considering using UCSC's Genome Browser, be aware there are limitations affected by the parameter settings in this step. Please see **Step 7** before proceeding.

Enter a file name and **Browse** to a path where you want to save the WIG file and click **Save**.

Note: If outputting WIG files while using the univariate segmenting algorithm, the browse button will have you select a directory location because a WIG file will be generated for each sample. These files will be named using the sample name.

If you wish to exclude additional markers other than those already excluded in **Step 3**, go to the **Exclude Markers** tab, **Browse** to the CSV file with the additional markers and click **Open**.

You are now ready to run segmentation. Click **Run**.

When segmentation has completed, two spreadsheets will pop up. The first is a copy number segment covariate spreadsheet (Figure 26), which contains the mean LogR value for each sample within each segment of markers. The second is a segment means spreadsheet (Figure 27), which contains columns for the chromosome name, segment start position, segment end position, segment mean, and the segment length in number of markers. If you selected to output a WIG file this will be stored at the path you chose above.

From here you can visualize found copy number segments in a genome browser (**Step 7**) or move on to performing association analysis on the copy number segment covariates (**Step 6**).

Sample	Chr1:1-6	Chr1:7-16	Chr1:17-24	Chr1:25-32	Chr1:33-44
sample001	-0.08144999916354822	-0.0539732991834171	-0.126536373990343	0.0266431259060038	0.036954001057893
sample002	0.14863459485783	-0.00611859988421202	0.0159128727391362	-0.0926297900358336	-0.0230768338466584
sample003	-0.0839106631040573	-0.0315207997326394	0.021117374683496	-0.126309749412071	-0.0447019367098537
sample004	0.0159898315787713	0.000723898629748233	0.137121260502637	-0.053303249525623	0.034237584866484
sample005	0.116444956323762	0.0279185997322202	-0.012622507096827	-0.0113341255346313	0.018159833193442
sample006	0.0669468343257904	-0.0243929994991049	0.0371860023587942	-0.0401262501254678	-0.0226741675287485
sample007	0.0866661680241426	0.00323359966278076	0.060426997921632	0.0375453747625518	-0.0539395020653804
sample008	0.024942168345054	-0.0181499993428988	0.0894268762709796	-0.0106605008477345	-0.0463773322699126
sample009	0.764146844546	-0.0434103889042342	0.0038424980561435	-0.01346879050487818	-0.085279168840498
sample010	0.12769150113066	0.0619117010384798	-0.00247012497857213	0.0379747496917844	0.0896872370511457
sample011	-1.30240116516749	-1.4400293012661	-1.4683173596859	-0.971435257233679	0.0209407490910962
sample012	0.0277521690974633	-0.0305076003074646	-0.0420574997951739	-0.0413959988373891	-0.00986516642539451
sample013	-0.0933095018068949	0.00101160164922476	0.133720874320716	0.0541891239117831	0.0284802503883839

Figure 26. Copy number segment covariate spreadsheet.

Sample Id	Chromosome Name	Base Start Position	Base End Position	Segment Mean	# SNPs
sample001	1	884	900	-0.052136528360493	17
sample002	1	884	900	0.0189965310903989	17
sample003	1	884	900	-0.014063411684485	17
sample004	1	884	900	0.0113641767050414	17
sample005	1	884	900	0.0405415303676444	17
sample006	1	884	900	0.0634152367574108	17
sample007	1	884	900	-0.00575309326887428	17
sample008	1	884	900	0.130309117629248	17
sample009	1	884	900	-0.0235232944216798	17
sample010	1	884	900	0.135334648411064	17
sample011	1	884	900	-0.0309441161944586	17
sample012	1	884	900	-0.0107459421475034	17
sample013	1	884	900	-0.0107459421475034	17

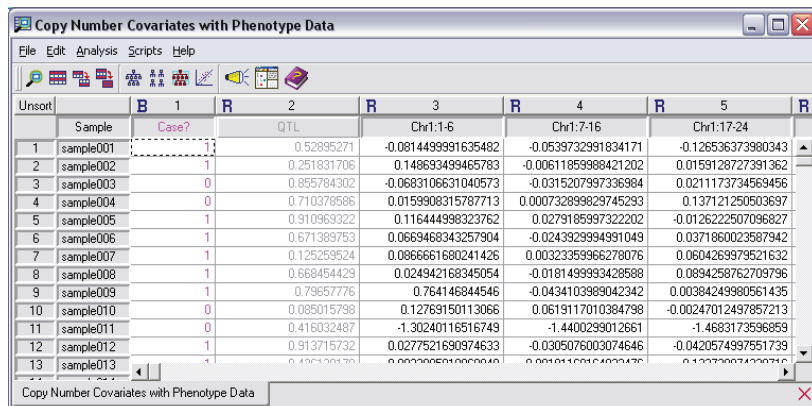
Figure 27. Segment means spreadsheet.

Step 6. Perform Association Analysis on Copy Number Segment Covariates

The next step is performing association analysis on the copy number segment covariates identified in **Step 5**. To do this, you first need to join the copy number segment covariate spreadsheet with your phenotype data.

Your phenotype spreadsheet should already be in HelixTree from **Step 2**. If you have not already done so, from the project navigator window go to **>File >Import Data** and choose either **>Import Wizard** or **>Import ASCII File**. Make sure to indicate the column with your sample names as the row label column. This will ensure each sample's phenotype status is appropriately matched to its copy number segment data when performing the association test(s).

Open your phenotype spreadsheet and select **>File >Join Spreadsheets on Row Labels**. Select your copy number covariate spreadsheet and click **OK**. A new spreadsheet will be created (Figure 28) with your sample names as row labels followed by each sample's phenotype information and then its mean LogR value for each segment.



Sample	Case?	QTL	R 3	R 4	R 5
sample001	1	0.52895271	-0.0814499991635482	-0.0539732991834171	-0.126536373980343
sample002	1	0.251831706	0.148693499465783	-0.00611859988421202	0.0159128727391362
sample003	0	0.855784302	-0.0683106631040573	-0.0315207997336984	0.0211173734569456
sample004	0	0.710378586	0.0159908315787713	0.000732899829745293	0.137121250503697
sample005	1	0.910969322	0.116444998323762	0.0279185997322202	-0.0126222507096827
sample006	1	0.671389753	0.0669468343257904	-0.0243929994991049	0.0371860023587942
sample007	1	0.125295924	0.0866661680241426	0.00323359966278076	0.0604269979521632
sample008	1	0.669454429	0.024942168345054	-0.0181499993428898	0.0894258762709796
sample009	1	0.79657776	0.764146844546	-0.0434103989042342	0.00384249980561436
sample010	0	0.085015799	0.12769150113066	0.0619117010384798	-0.00247012497857213
sample011	0	0.416032487	-1.30240116516749	-1.4400299012661	-1.4683173596859
sample012	1	0.913715732	0.0277521690974633	-0.0305076003074646	-0.0420574997951739
sample013	1	0.426120470	0.000000000000000	0.000000000000000	0.000000000000000

Figure 28. Simulated joined phenotype and copy number segment covariate spreadsheet.

Performing association analysis on this spreadsheet is done using the classic tree-based analysis approach in HelixTree. To do this, from your joined spreadsheet, first **Left Click** once on the column header of your dependent variable. This will turn the column magenta indicating its dependent variable status. You can also **Left Click** twice on the column header of each non-copy number segment column header to turn it grey. This will inactivate these columns so they won't be used during analysis. If you want to include them, leave them black.

Next select **>Analysis >Interactive Tree Analysis**.

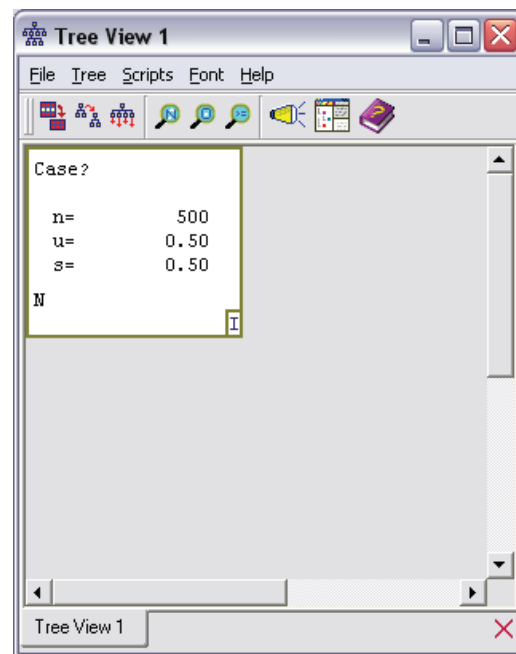
A window (Figure 29) with a single box (referred to as the root node) will appear and display the number of samples in your data set (n), the mean value of the response variable (u), and the response variables standard deviation (s). From this window you can either perform tree-based segmenting or regression association analysis.

Note: This tutorial only covers basic tree and regression analysis. To learn more about advanced capabilities of tree analysis see Chapter 7 in the HelixTree manual.

Tree-Based Segment Association

Similar to how the segmenting algorithm works in CNAM to identify copy number variations from LogRs (**Step 5**), HelixTree employs segmenting to find the optimal split(s) in a dataset whereby the mean(s) of the resulting subgroups differ based on a given variable (e.g. mean intensity values for a given marker). HelixTree then performs the appropriate association test to determine whether the differences in means among the subgroups are statistically significant.

HelixTree is unique in that can use this segmenting association approach to find multi-way splits supported by the data. This is especially powerful when considered in the context that a deletion or duplication is a



Case?	Value
n=	500
u=	0.50
s=	0.50

Figure 29. Interactive Tree Analysis root node: n = 500 samples, u = 62% cases, s = standard deviation of .48.

disruption, and a disruption may lead to higher incidence of disease. In cases where a basic test of association, such as regression and t-tests, may not find anything, a multi-way split may find that patients with a segment mean LogR near 0 (copy number two) may have a lower incidence of disease than patients with a higher or lower mean LogR (the tails).

To perform basic tree-based segment association, **Right Click** on the root node and select **Manual Split**. This will pop up the Manual Split window (Figure 30) with a list of p-values, the copy number segment covariates (referred to here as splitters), and the split rule used to perform the association.

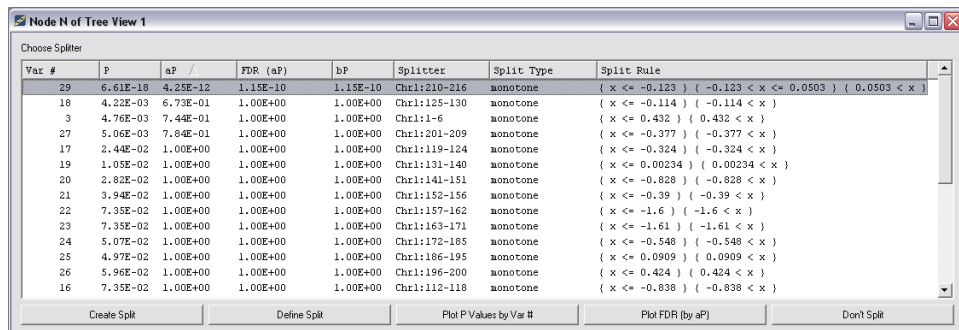


Figure 30. Manual Split window highlighting significant region Chr1:210-216 with F-Test Bonferroni corrected p-value (bP) of 1.15E-10.

Clicking on each row of the Manual Split window will change the tree view to reflect the various split rules. Clicking on **Plot P Values by Var #** will plot the p-values for each segment covariate, which will produce a p-value plot similar the ones created in **Step 4**.

You can visualize the association results for each splitter by clicking **Define Split** at the bottom of the Manual Split window. This will plot the segment mean LogR values against the response variable. For copy number segment association, you will typically see a plot with a binary (two-way) or three-way split (denoted by one or two vertical lines respectively).

Figure 31a, shown below, displays a binary split. Notice how the plot is centered around zero on the X-axis. This is because a LogR of zero is equivalent to copy number two. A rudimentary interpretation of this plot is that on average, a copy number equal to or less than two will result in a reduced incidence of disease (control), whereas a copy number variation greater than two will result in a greater incidence of disease (case).

Figure 31b, shown below, is a screenshot of a three-way split. The interpretation of this plot is that on average, a copy number less than or greater than two will result in a greater incidence of disease (case), whereas a copy number variation equal to two will result in a reduced incidence of disease (control).

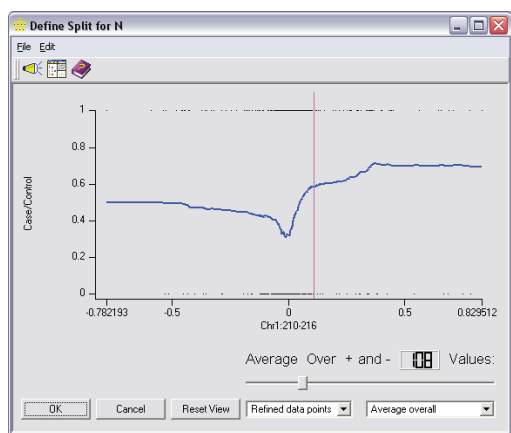


Figure 31a. Binary split.

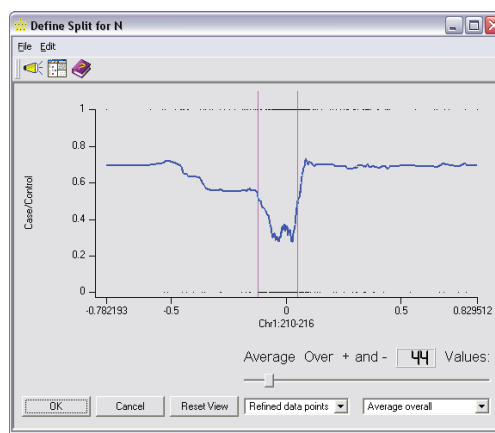


Figure 31b. Three-way split.

If you see a plot where the LogRs go from low to high or high to low, there is a good chance an additive affect is taking place. In this case, it would be best to perform regression association.

Tree-Based Regression Association

Regression association is also performed using interactive tree analysis with a modification to one of the tree options. To perform regression, go back to your joined phenotype/copy number segment covariate spreadsheet. Make sure your response variable is still magenta and select **>Analysis >Interactive Tree Analysis**. This will again pop up the tree view with a single root node. From this window select **>Tree >Options**. The following window will appear.

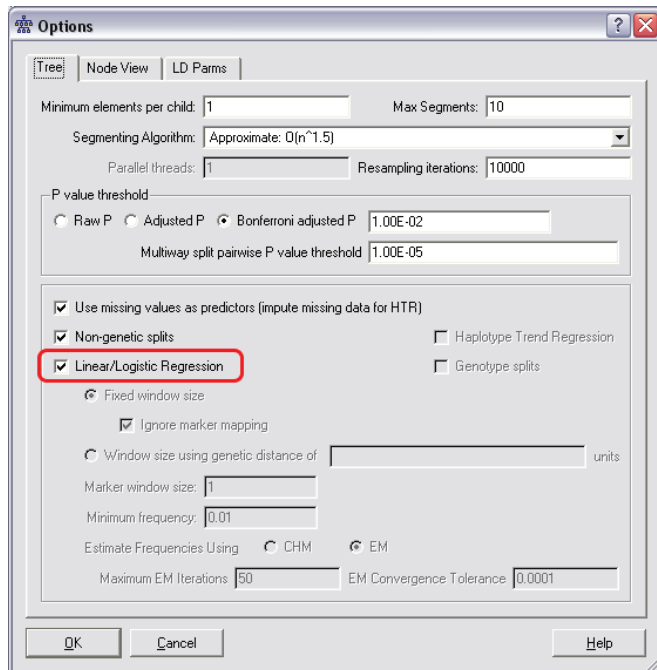


Figure 32. Tree options with Linear/Logistic Regression selected.

Toward the middle of the window there is a check box **Linear/Logistic Regression**. Check this box. This will enable regression to be performed from the Tree View.

Note: Based on your response variable, HelixTree will automatically perform a linear regression for quantitative dependent variables and logistic regression for binary dependent variables.

Click **OK**.

Now from the root node in the Tree View, **Right Click** and select **Manual Split**. A similar window to Figure 33 will appear.

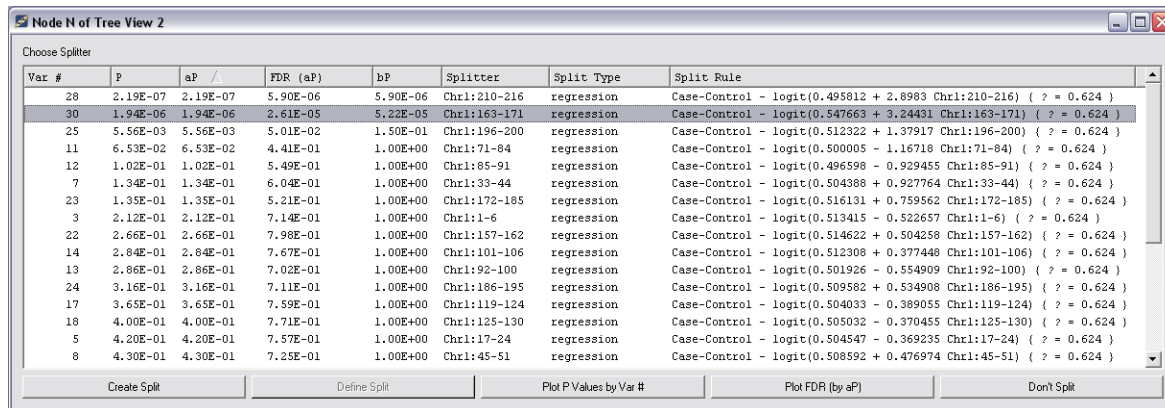


Figure 33. Manual Split window highlighting significant region (Chr1:163-171) with regression based Bonferroni p-value (bP) of 5.22E-05.

Notice now there are regression Split Type's and the split rule is equivalent to the regression equation. The p-values are also different because they are now based on a regression test.

Click Create Split. This will drop a residual node in the Tree View. (Figure 34). Notice the mean (u=) in the residual node is now zero. From here you can plot the association results by **Right Clicking** on the root node and selecting **>Visualize Split Data >Show Split Data**.

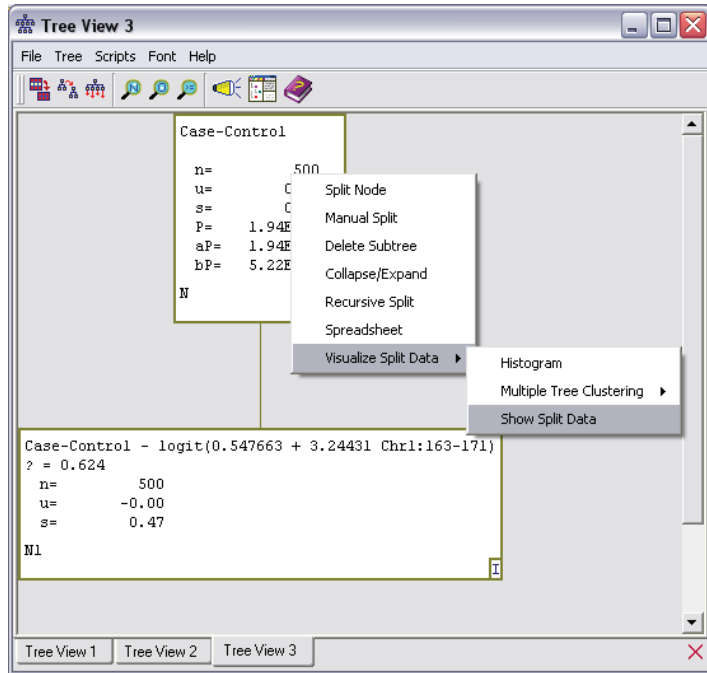


Figure 34. Residual node from regression “split” with option to view association results (split data).

A plot similar to that in the Define Split window will appear (Figure 35).

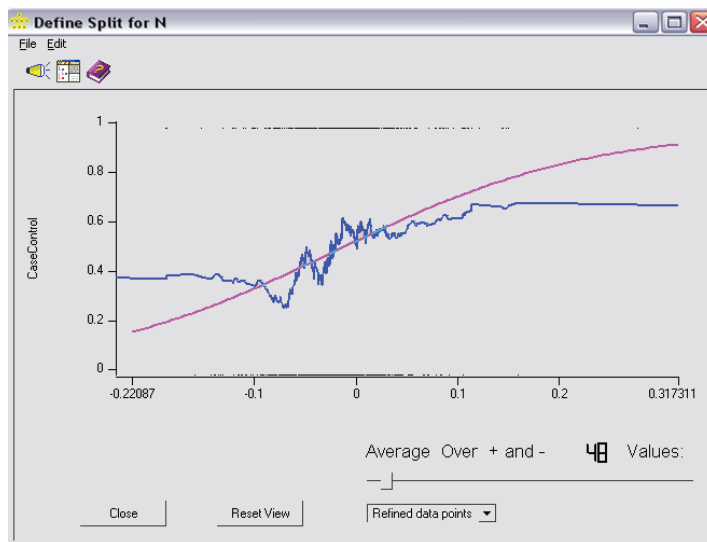


Figure 35. Regression association results.

From here you can repeat **Step 5** and segment additional chromosomes or visualize interesting regions in an external genome browser (**Step 7**).

Step 7. Visualize Segmenting Results

There are tools available outside HelixTree to visualize your copy number segmenting analysis results. Using the optional wiggle track file output in **Step 5** you can view your segmenting results using UCSC's Genome Browser or Affymetrix's Integrated Genome Browser (IGB). If you are analyzing Illumina data you can output the Segment Means spreadsheet created in **Step 5** as a bookmark CSV file to view your results in Illumina's BeadStudio Genome Browser. This tutorial will cover importing data into UCSC's Genome Browser. For more information on the other browsers see section 25.6 of the HelixTree manual.

The UCSC Genome Browser is an online tool created by the University of California, Santa Cruz (<http://genome.ucsc.edu/cgi-bin/hgGateway>), used to view genome data formatted in the Wiggle track format (WIG).

To import copy number analysis results in the UCSC Genome Browser, first go to the website address above. Select the **add custom tracks** button to open the **add custom tracks** page. From this page browse for the WIG file created in **Step 5** and select **Submit**. The next page will have a table to view the tracks found in the specified WIG file. To add more data tracks (samples), select add custom tracks. To view the tracks in the genome browser (Figure 36), select **go to genome browser**.

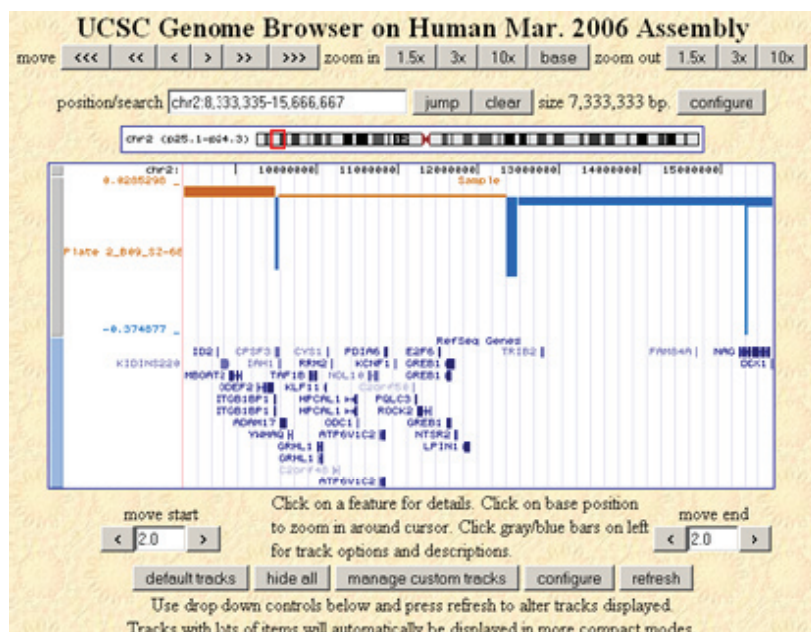


Figure 36. UCSC Genome Browser showing HelixTree segmenting results.

Note: There are some restrictions on the data that can be imported into the UCSC Genome Browser because it is an online tool. When performing multivariate segmenting analysis in CNAM, the results may contain segments that are only one marker in length. The UCSC Genome Browser does not accept data files containing segments with length equal to one. To avoid this conflict, in **Step 5** set the **Min #Markers** per segment parameter in the Copy Number Segmentation tool to a value greater than one.

The second conflict when using the UCSC Genome Browser to view CNAM segmenting results occurs when the difference between a segment's chromosome start position and chromosome end position is found to be larger than 10,000,000 bp. Currently the best way to handle this issue is to manually separate the large segments into multiple adjacent segments with the same segment mean. The WIG file can be opened and edited in a text editor. Following the format outputted by CNAM, insert rows to divide the long segment into multiple shorter segments.