Copy Number Analysis Module (CNAM)

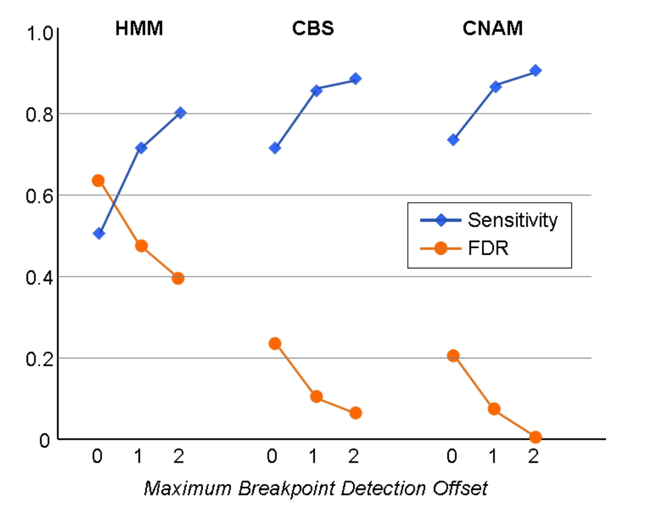
Software for Whole Genome Copy Number Association Studies

The Golden Helix Copy Number Analysis Module (CNAM), the newest product to join the SNP & Variation Suite, is the first software application capable of performing whole genome association studies on copy number variations.

Introduction

The Golden Helix Copy Number Analysis Module (CNAM) opens the door to the next breakthrough in genetic analysis. While there has been tremendous excitement in the genetic research community over the possibilities of whole genome copy number association, there simply have not been any tools capable of delivering on the promise. Until now. Using optimal methods for finding copy number segments, CNAM empowers researchers to find new value in their existing data, giving them the power to perform high resolution association studies on copy number variations.

Figure 1. False Discovery Rate vs. Sensitivity



This plot shows how the false-discovery rates and sensitivity compare among three different algorithms when tested on simulated data from Willenbrock,[6] illustrating that CNAM has better performance than CBS with a median false-discovery rate of 0.0 and a sensitivity of 0.90 versus 0.06 and 0.88 for CBS.

Problems with Existing Copy Number Analysis Methods

To date, most copy number analysis methods have had to rely on Hidden Markov Models (HMMs) or other methods that have proven to be sub-optimal due to high false discovery rates and low sensitivity. This hasn't been entirely by choice; there simply haven't been better, more effective methods available. Thus the bleeding-edge researchers that have undertaken copy number analysis have been handicapped with error-prone copy number calculations.

In recent years, two survey papers by Willenbrock et al.[6] and Lai et al.[4] have been published comparing various methods for finding copy number segments. The authors found that HMMs performed poorly, with high false discovery rates (~40-60%) and low sensitivity (~50-80%), even when applied to simulated data with known properties. The benefit of HMMs, however, is that they are relatively fast, which is the primary reason they are the most widely implemented methods today.

Additionally, intensity data, which is used for detecting and identifying copy number segments, is often "noisy," causing concern among researchers. If the data that feeds the challenged HMM algorithms is itself suspect, then the probability of the final result being reliable is even more questionable.

The end result is that scientists the world over have been spending endless cycles attempting to invent new methods based on existing paradigms, or hopelessly fine-tuning the above-mentioned, heuristic-based methods.

Segmentation

In the two studies previously mentioned, the authors concluded that the most effective method for finding copy number segments is Circular Binary Segmentation (CBS), a segmentation approach based on finding the change-points in data. Such an approach has been implemented in DNAcopy, for example. Though effective for finding segments and despite speed optimizations by Venkatraman et al.,[5] CBS is not computationally efficient for whole genome analysis. Case in point: analysis on Affymetrix 500K data has been shown to take over 20 minutes per sample, and roughly 45 minutes per sample on Illumina 550K data.

The challenge with segmentation methods is that in order to find the optimal segmentation, all possible change-points need to be evaluated, creating a combinatorial explosion. For example, if there are 10 copy number segments positioned across 2,000 probe intensities, then there are 2000^{10} places that these segments might lie (roughly a 1 followed by 33 zeros, or one decillion). Given a chromosome may have as many as a hundred change-points and today's whole genome arrays contain over 50,000 intensity values for a given chromosome, the search space can easily approach $50{,}000^{100}$ possible change-points.

This seemingly impossible series of calculations is what has led researchers to adopt various heuristics to make this process computationally viable, as was done with CBS, or by avoiding segmentation altogether.

Golden Helix Dynamic Optimal Segmentation Algorithm

In its initial research, Golden Helix found that the same powerful technology used for segmentation in its predictive analytics tools could be applied to this problem.

CNAM Highlights

- **Accurate:** Segmenting algorithm has low false-discovery rate and high sensitivity. Permutation testing ensures segments are statistically significant.
- **Fast:** Segmentation takes just over 5 minutes per sample with ~500K SNPs, and can be accelerated with multi-core systems.
- **Compatible:** The product is integrated with Illumina's BeadStudio software and provides direct support for Affymetrix CNT files.
- **Validated:** The algorithm has been validated on published simulated data as well as Affymetrix and Illumina HapMap data with known regions of copy number variation.

In fact, for nearly a decade, Golden Helix has been using an optimal segmenting algorithm that exhaustively searches through all possible change-points to find the optimal segmentation without succumbing to the combinatorial explosion. It is able to do so with speed that makes it efficient enough for whole genome analysis. The algorithm, based on published work by Dr. Douglas Hawkins [1,2,3] of the University of Minnesota's School of Statistics, utilizes dynamic programming to exhaustively look through all possible change-points to uncover those that optimize the sum of squared deviations from the mean within each segment. A creative implementation of this segmenting approach enabled Golden Helix to effectively solve the copy number computation problem.

Permutation Testing for Verifying Copy Number Segments

Solving the computational issue was not the final step in the process, however. The best results remained somewhat illusive due to the noise inherent in intensity data. There is also an added challenge in segmenting approaches: determining the correct number of copy number segments supported by the data.

A paired T-test comparing adjacent segments comes to mind, and is a reasonable approach. However, this approach does not account for the multiple testing involved in finding the pairwise optimal change-points. A Bonferroni adjustment can be used, but would be overly conservative because it would not account for correlations between adjacent points.

To overcome these last hurdles, Golden Helix supplemented its optimal segmentation algorithm with an efficient permutation testing approach, designed to cut through the noise to uncover true segments of copy number variation by determining and validating their statistical significance.

In short, each pair of adjacent segments must be statistically different. After the optimal segmentation process is completed, the data for adjacent segments is randomly shuffled. The shuffled data is optimally binary segmented, and the resulting sum of squared deviations from the mean is compared against the original segment pair. This process is repeated numerous times, according to a specified p-value threshold, until the original segment pair is determined to be significant versus the permuted ones. If not, the segment is discarded as insignificant. If all adjacent segments are statistically different, the data supports the segmentation. Otherwise, the next lower cardinality of optimal segmentation is evaluated.

Validation

Having successfully implemented this methodology for copy number analysis and developed the CNAM tool, a series of tests were conducted on the same simulated data used in the Willenbrock study, with better results than DNAcopy's Circular Binary Segmentation method (Figure 1). CNAM was able to achieve a median false-discovery rate of 0.0 with a sensitivity of 0.90 versus a median false discovery rate of 0.06 and sensitivity of 0.88 for DNAcopy. The computation time of the algorithm is also dramatically faster than CBS, taking roughly five minutes per sample to analyze Affymetrix 500K data and six minutes for Illumina 550K. To speed the process even further, the algorithm can be run using multi-core systems.

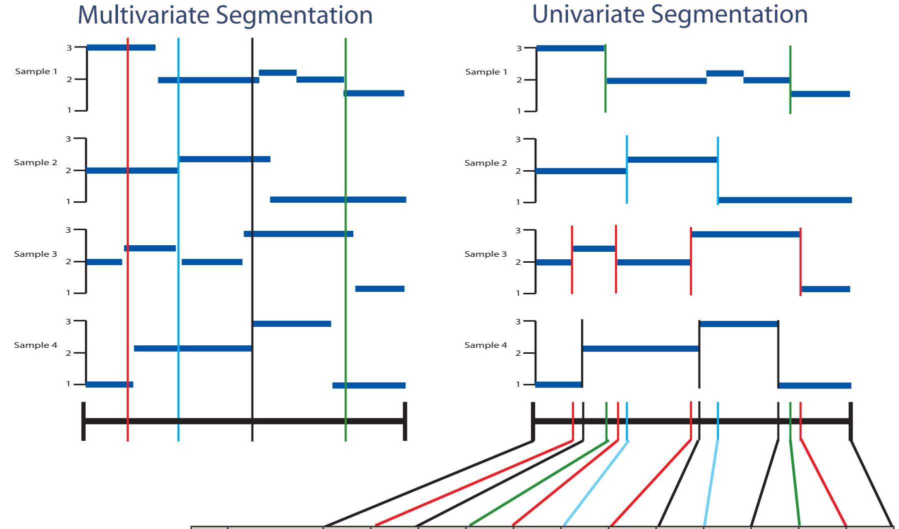
Additionally, the algorithm is robust to missing values and has been further validated on several simulated and real data sets.

Univariate vs. Multivariate Segmentation

CNAM offers two types of segmenting methods, univariate and multivariate (Figure 2). Both methods are based on the same algorithm, but use different criteria for determining segment change-points.

The univariate method segments each sample independently. Each time there is a change-point for a given sample, an additional covariate is created. The value of the covariate is the mean intensity within the original segment for that sample. The result is a spreadsheet showing all change-points found among all samples. The univariate approach is good for finding individual variations but may miss shorter copy number variations due to noise in the data. It is for this reason that a multivariate approach is also provided.

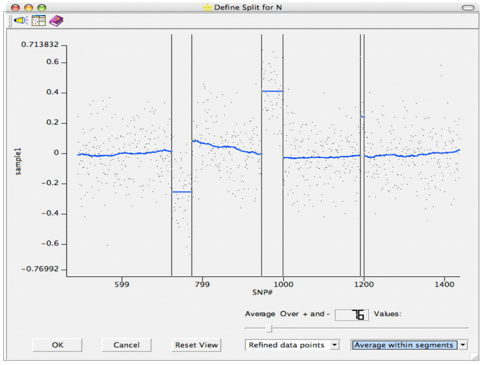
Figure 2. Univariate vs. Multivariate Segmentation



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		B-R	R-B	B-G	G-R	R-B	B-R	R-B	B-B	B-B	B-G	G-R	R-B
2	Sample 1	3	3	3	2.2	2.2	2.2	2.2	2.2	2.2	2.2	1.5	1.5
3	Sample 2	2	2	2	2	2	2.2	2.2	2.2	1	1	1	1
4	Sample 3	2	2.5	2.5	2.5	2	2	3	3	3	3	3	1

Above: The illustration above depicts how the multivariate and univariate segmentation methods differ. These methods are based on the same algorithm, but use different criteria for determining cut-points. The result of each is a spreadsheet with coavariates that correspond to the mean copy number intensities for each segment. Also shown is the resulting spreadsheet for the univariate approach.

Right: This is an actual screenshot of the univariate segmenting approach that shows the cut-points represent the beginning and end of segments. In this case, there are 7 distinct segments.

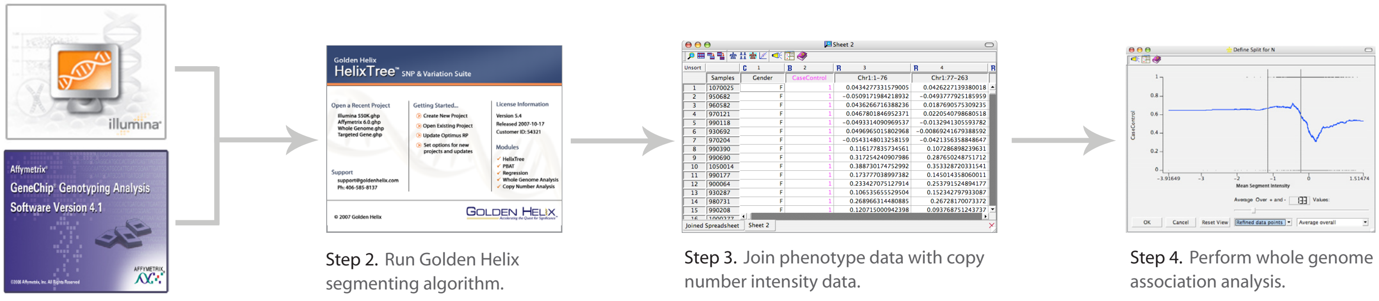


Average Over + and - 76 Values:

Refined data points

Average within segments

Figure 3. Whole Genome Copy Number Association Workflow


illumina
Affymetrix
GeneChip Genotyping Analysis Software Version 4.1
Golden Helix HelixTree SNP & Variation Suite
Sheet 2
Define Split for N

Step 1. Export copy number intensity data.

Step 2. Run Golden Helix segmenting algorithm.

Step 3. Join phenotype data with copy number intensity data.

Step 4. Perform whole genome association analysis.

The multivariate method segments all samples simultaneously, finding partial consensus copy number regions based on variation across many samples. In this case, every sample shares the same segment boundaries and the resulting covariates are the mean intensities for the given sample within the segment.

If there are consistent positions for copy number variation across multiple samples, the copy number segments will best be found using the multivariate method. It does make the fundamental assumption, however, that the boundaries for each sample are somewhat near one another. In reality there may not always be consistent copy number segments across multiple samples and the multivariate method may miss some segments where boundaries are irregular. Thus the multivariate method is preferable for finding very small copy number regions and for finding partially conserved regions that may be useful for association studies.

Univariate vs. Multivariate

Univariate Segmenting

- Analyzes each sample independently
- Good for finding individual variations
- Misses short CNVs due to noise

Multivariate Segmenting

- Analyzes all samples simultaneously
- If variations share boundaries will find "consensus" segments useful for association analysis
- Finds short CNVs that are replicated
- May miss CNV segments when boundaries are irregular

Whole Genome Association Workflow

Step 1. Export copy number intensity data.

CNAM offers direct support for Illumina and Affymetrix platforms and includes additional functionality for importing intensity data from other providers.

Illumina: A BeadStudio plug-in was developed in a joint effort with Illumina, providing the ability to export Illumina intensity data directly into Golden Helix's DSF format. All chromosomes can be exported at once or can be individually specified during the export process. If multiple DSF files are created, they can be concatenated during importation.

Affymetrix: Affymetrix provides tools for extracting data from cell intensity (CEL) files that can be used for copy number analysis. This data is saved in the Affymetrix CNT format. An option is available within CNAM to convert these CNT files into the required Golden Helix DSF format.

Step 2. Run segmenting algorithm.

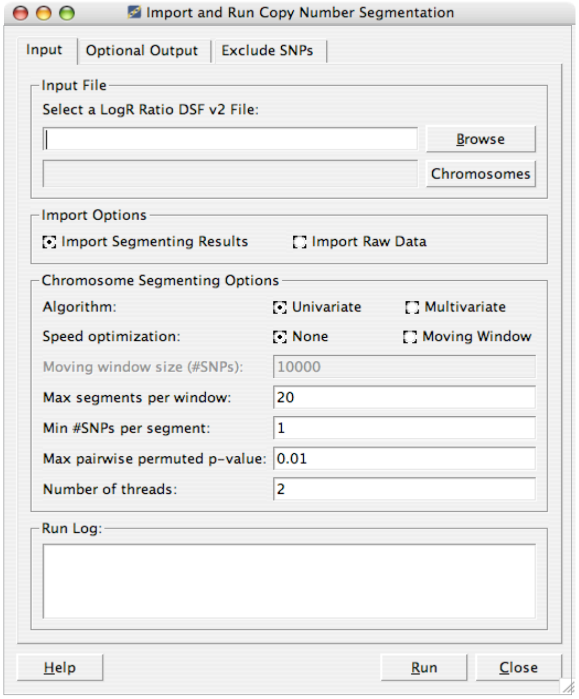
A simple dialog box guides the user through the importation and segmentation process. Certain parameters may be changed to optimize performance including choosing multivariate or univariate segmentation, enabling a moving window approach for performance optimization, and various other parameters that fine-tune the finding of copy number segment boundaries. An Illumina BeadStudio plug-in enables the viewing of copy number bookmarks using BeadStudio's Genome Browser.

Optionally, Wiggle files can be output for viewing with the UCSC Genome Browser.

CNAM works by scanning normalized probe intensity data (log R ratios) with the object of determining where copy number variations occur. During this process, covariates that describe the segments are created. The result is a covariate spreadsheet that can then be used for association analysis on the intensity data over these regions of variation, taking advantage of all the features and capabilities available within Golden Helix's HelixTree Software.

Depending on how stringently the parameters were set, both the univariate and multi-

Figure 4. Copy Number Segmentation Window



Input File

Select a LogR Ratio DSF v2 File:

Import Options

● Import Segmenting Results ○ Import Raw Data

Chromosome Segmenting Options

Algorithm: ● Univariate ○ Multivariate

Speed optimization: ● None ○ Moving Window

Moving window size (#SNPs): 10000

Max segments per window: 20

Min #SNPs per segment: 1

Max pairwise permuted p-value: 0.01

Number of threads: 2

Run Log:

Above is the Copy Number Segmentation window with default parameters set to run the Golden Helix optimal segmentation algorithm on copy number intensity data.

ate segmenting methods will create a fair of number of covariates to be used for association analysis. This number, however, will be orders of magnitude less than the total number of original intensity values, thereby reducing the multiple testing penalty.

Step 3. Join phenotype data with copy number intensity data.

The next step is joining the resulting segment mean covariate spreadsheet with additional phenotype data using HelixTree. The segment means covariate spreadsheet is treated the same as any other covariate spreadsheet within HelixTree: it can be joined with case/control status or other dependent variables (HelixTree supports binary, quantitative and categorical dependents) as well as additional phenotype variables.

Step 4. Perform whole genome association analysis.

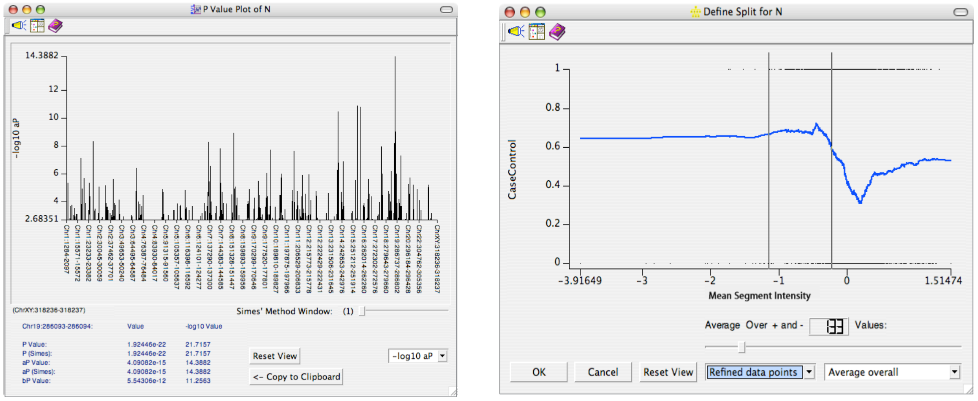
Once the data is joined, association analysis can be performed using tree-based approaches or regression within HelixTree.

The result of an association test is a window with various raw and adjusted p-values (Bonferroni, FDR, etc.) along with information regarding how the analysis was performed.

From here a p-value plot can be generated (Figure 5), which along with a table, gives a visual indication of which copy number segments are associated with the response variable. Additional tools include histograms and plots of intensity data versus case/control status (Figure 5). If a BeadStudio Bookmark or UCSC Wiggle file was created during segmentation, then copy number segments can also be visualized in these Genome Browsers.

Further, if the Regression Module is activated, linear or logistic regression on a quantitative or binary trait can also be performed.

Figure 5. Association Results



(ChrXY:318236-318237)

Simes' Method Window: (1)

Chr19:286093-286094:	Value	-log10 Value
P Value:	1.92446e-22	21.7157
P (Simes):	1.92446e-22	21.7157
aP Value:	4.09082e-15	14.3882
aP (Simes):	4.09082e-15	14.3882
bP Value:	5.54306e-12	11.2563

Average Over + and - 133 Values:

Refined data points

Average overall

A screenshot of the interactive p-value plot of the entire set of p-values for all found copy number segments across the genome. This plot gives a quick indication of which segments are associated with the response variable.

This plot shows how a particular segment is associated with case/control status. In this case, a lower copy number intensity resulted in a higher proportion of cases than normal copy number intensities (centered around zero) or greater.

References

1. Hawkins, D.M. (1972) 'On the choice of segments in piecewise approximation: Jour. Inst. Math. Applications, v. 9, no. 2, p. 250–256.
2. Hawkins, D.M., and Merriam, D. F. (1973) 'Optimal zonation of digitized sequential data': Jour. Math Geology, v. 5, no. 4, p. 389-395.
3. Hawkins, D. M., (2002). `Fitting multiple change-points to data', Computational Statistics and Data Analysis, 37, 323--341.
4. Lai, W., Johnson, M., Kucherlapati, R., and Park, P.: Comparative analysis of algorithms for identifying amplifactions and deletions in array CGH data. Bioinformatics 2005. v. 21(19):3763-70.
5. Venkatraman, E.S., and Olshen, A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23:657-663.
6. Willenbrock, H. and Fridlyand, A.: A comparison study: applying segmentation to array CGH data for downstream analyses. Bioinformatics. 2005 v. 21(22):4084-91.

Minimum System Requirements

Windows®
- Windows NT, 2000, XP
- Intel® Pentium® III processor or higher
- 512MB RAM

Linux®
- Intel Pentium III processor or higher
- Redhat Linux 7.3 - 9.0
- 512MB RAM

Macintosh®
- Mac OS X 10.4.x
- Power PC G4 & G5, Intel processors
- 500MB RAM

Recommended for WGA Analysis

- Intel Pentium 4 processor or higher
- 2GB RAM
- 80 GB Hard Drive

Join the growing Golden Helix community

- Hundreds of users worldwide
- Over 100 academic, government and industry organizations
- Nearly 100 peer reviewed publications

For more information visit:

www.goldenhelix.com

- Download a free fully-functional trial
- Sign up for a free webinar
- Explore the literature
- Discover how to integrate HelixTree into your research

Korea Biomics
Robinson Building #401, 542-4
Song-Dong, Gangdong-Gu
Seoul 134-030, Korea
Ph: +82 (0) 2-477-7470
Fax: +82 (0) 2-6081-6510
www.koreabiomics.com