

DNA Copy Number Analysis Algorithms

Several algorithms are available for BeadStudio that enable detection of copy number variants and other chromosomal aberrations.

INTRODUCTION

Illumina has developed several algorithms for detecting copy number variants (CNVs) and other structural variants (Table 1). These algorithms are available as individual software plug-ins for the BeadStudio Genotyping Module and can be downloaded from the BeadStudio Portal, from the download section of the illumina•Connect webpage¹, or from iCom. Plug-ins are also available from Illumina Tech Support. The plug-ins are used within the CNV Analysis workbench, and results can be visualized within the BeadStudio Full Data Table, in the IGV, or in a CNV region display window. This technical note describes the function of these algorithms and how they can be employed to analyze a chromosomal region of interest.

LOH SCORE STATISTICAL ALGORITHM

The LOH Score column plug-in reports the likelihood of loss of heterozygosity (LOH) existing in a region of interest. The LOH Score algorithm scans data sets to determine and identify the presence of LOH. Variances in LOH score can be plotted in the Chromosome Heat Map or in the Illumina Genome Viewer (IGV).

TABLE 1: BEADSTUDIO COPY NUMBER ALGORITHMS

ALGORITHM	FUNCTION
LOH Score	Estimates the likelihood of a region exhibiting LOH
cnvPartition	Calculates copy numbers with confidence scores and generates CNV regions
Homozygosity Detector	Autobookmarks samples with extended tracts of homozygosity (single-sample analysis only)
ChromoZone	Autobookmarks chromosomal aberrations in single-sample mode

If a chromosomal region is lost and LOH is observed, the only expected genotypes are AA and BB. In this case, AB would be observed only as a result of genotyping error. If there is no LOH, all three genotypes are possible.

The LOH score is a measure of the likelihood that a SNP is exhibiting LOH around a window over all N SNPs, where N is the number of SNPs in a user-designated window size centered at the SNP's chromosomal position. The equation used to determine the LOH Score is shown in Figure 1. The recommended window size depends upon the density of SNPs on the product in use. The window size for a specific workspace may require optimization depending upon the type of aberration under examination and the quality of the data. Table 2 lists the suggested window sizes for selected SNP densities.

If the number of heterozygote SNPs matches the prediction in the specified window, the LOH score is 0. The LOH score increases if there is an unexpectedly low number of heterozygote SNPs in the window. Because of this design, the algorithm is based purely on genotype calls and heterozygote frequencies. Taking both of these into account, the LOH Score algorithm is a log odds ratio of the probability of a region exhibiting LOH versus

FIGURE 1: LOH SCORE EQUATION

$P_{\text{error}} = 0.001$

$$\text{LOH Score} = \log_{10} \frac{\prod_{i=1}^N P_i(gt_i|LOH)}{\prod_{i=1}^N P_i(gt_i|NoLOH)}$$

$$P_i(gt_i|LOH) = \begin{cases} P_{\text{error}}, & gt_i = AB \\ 1 - P_{\text{error}}, & gt_i = AA \text{ or } BB \end{cases}$$

$$P_i(gt_i|NoLOH) = \begin{cases} \text{hetfreq}, & gt_i = AB \\ 1 - \text{hetfreq}, & gt_i = AA \text{ or } BB \end{cases}$$

P_{error} = genotyping error rate
 hetfreq = mean heterozygote frequency
 gt_i = genotype of locus i

not exhibiting LOH. Because levels of LOH can vary from sample to sample and region to region, it is difficult to assign LOH score thresholds that always positively identify regions exhibiting LOH. However, the LOH score is a valuable calculation that can be used to detect chromosomal aberrations.

An odds ratio is defined as the ratio of the number of subjects in a group with an event to the number of subjects without an event. The log odds ratio for each of these hypotheses is computed using the cluster file to estimate heterozygote allele frequencies for every SNP, assuming that genotyping calls are independent.

The LOH Score algorithm does not incorporate haplotype structures and assumes that heterozygote frequencies in the training set are representative of frequencies in the population under study. Therefore, the LOH score is a generalization of what may be occurring in a region of interest. In single-sample mode, the reference is a cluster file and it is possible that copy-neutral LOH detected may be due to haplotype block structure in the data.

A diverse panel of 120 HapMap samples, including Caucasian, Han Chinese, Japanese, and Yoruba HapMap populations, is used to create the default cluster file and to calculate heterozygote frequencies. Heterozygosity

rates are estimated for the combined group. If the population under study is not represented well by this group, it is beneficial to create a new cluster file based on the data from the unique population. Independent of the platform used, false positives in the LOH score may be due to some SNPs being rare in the studied population but common in the diversity panel used to create the cluster file.

LOH Score Example

To illustrate the flow of this algorithm, consider this simplified example. For a window containing N SNPs, all resulting in homozygote calls with heterozygote frequencies of $h=0.1$, let the genotyping error be $e=0.001$. The likelihood of LOH occurring is $(1-e)^N$. The likelihood of no LOH is $(1-h)^N$. Therefore, the log odds ratio is $\text{Log}_{10}\{(1-e)^N/(1-h)^N\}$, which is the same as $N\{\log_{10}(1-e) - \log_{10}(1-h)\}$, which equals $N(h-e)/2.3$. The odds of LOH or No LOH grows in a roughly linear fashion with the number of consecutive homozygotes.

On the other hand, if that stretch contains M heterozygote calls, the likelihood of LOH decreases and the equation is adjusted to $(1-e)^{N-M} \times e^M$, since heterozygotes in a region with LOH occur only through genotyping errors. The likelihood of No LOH also changes and becomes $(1-h)^{N-M} \times h^M$. The log odds ratio now becomes

equal to $\{(N-M)(h-e)/2.3\} + M\{\log(e)-\log(h)\}$ which equals $\{(N-M)(h-e)/2.3\}-2M$. In this case, the odds have diminished. When roughly 1 in 10 SNPs receives a heterozygote call, the odds of both hypotheses are equal. If more heterozygote calls are produced, the log odds ratio becomes negative as it becomes less likely that these observations come from a region with LOH.

It is important to remember that there are usually unknown haplotype structures and population-dependent heterozygote frequencies which may play a role in the accuracy of the LOH score. However, this score is provided as a starting point to determine whether a particular stretch of homozygotes contains LOH.

ALGORITHMS FOR AUTOMATED BOOKMARKING

BeadStudio can make use of several automated bookmarking algorithms. These plug-in algorithms automatically scan data for the presence of structural aberrations and CNVs. Each algorithm employs a different strategy to search for aberrations and CNVs. They are intended to assist with visually categorizing various types of aberrations present in samples of interest. Automated bookmarking algorithms can be used as data-mining tools for the discovery of new regions, or for the verification of known regions of interest. Bookmarks can be edited whether they are created manually or generated using the autobookmarking tool to create a list of bookmarks. Both manually- and automatically-generated bookmarks can be exported to share with other users. This technical note describes three autobookmarking algorithms available for BeadStudio:

- cnvPartition
- Homozygosity Detector
- ChromoZone

CNVPARTITION

The goal of the cnvPartition algorithm is to identify regions of the genome that are aberrant in copy number using two Infinium® Assay outputs: log R ratio (LRR) and B allele frequency (BAF). Since LRR is the logged ratio of observed probe intensity to expected intensity, any deviations from zero in this metric are evidence for copy number change. BAF is the proportion of hybridized sample that carries the B allele as designated by the Infinium Assay. In a normal sample, discrete BAFs of 0.0, 0.5, and 1.0 are expected for each locus (representing AA, AB, and BB). Deviations from this expectation are

indicative of aberrant copy number. For example, if a locus has a BAF of 0.66, this might indicate that there are two copies of the B allele and one copy of the A allele present in the sample ($\frac{2}{2+1} \approx 0.66$). Analyzing both of these metrics provides stronger resolving power to detect true copy number changes.

Copy Number Estimation

cnvPartition models LRRs and BAFs for each of fourteen different copy number scenarios as simple bivariate Gaussian distributions (Table 3).

Modeling copy number in this way allows for computation of a preliminary copy number estimate for each assayed locus by comparing its observed LRR and BAF to values predicted from each of the fourteen models. Specifically, the likelihood of observing a given LRR and BAF under each of the fourteen models is calculated. For example, to compute the likelihood of a particular LRR and BAF given a genotype of AAB (L_{AAB}), the AAB parameters from Table 3 and the standard normal density are used:

TABLE 3: GENOTYPES MODELED BY CNVPARTITION

GENOTYPE	CN	LRR MEAN	LRR SD	BAF MEAN	BAF SD
DD	0	-5	2	NA	NA
A	1	-0.45	0.18	0	0.03
B	1	-0.45	0.18	1	0.03
AA	2	0	0.18	0	0.03
AB	2	0	0.18	0.5	0.03
BB	2	0	0.18	1	0.03
AAA	3	0.3	0.18	0	0.03
AAB	3	0.3	0.18	1/3	0.03
ABB	3	0.3	0.18	2/3	0.03
BBB	3	0.3	0.18	1	0.03
AAAA	4	0.75	0.18	0	0.03
AAAB	4	0.75	0.18	0.25	0.03
ABBB	4	0.75	0.18	0.5	0.03
BBBB	4	0.75	0.18	0.75	0.03

Parameters for each of the fourteen genotypes considered by cnvPartition are shown. BAFs are modeled as a uniform distribution between zero and one for homozygous deletions (DD). All other distributions are modeled with Gaussian distributions with the given parameters. The genotype AABB is not modeled since this would represent two independent duplication events and rarely occurs in nature. (CN = copy number, DD = double deletion, SD = standard deviation)

TABLE 2: RECOMMENDED WINDOW SIZES FOR LOH SCORE ALGORITHM

PRODUCT	POPULATION	MEAN AB FREQUENCY	MINIMUM RECOMMENDED WINDOW SIZE (MB)
Human1M	88 unrelated HapMap individuals ¹	0.28	0.09
	33 unrelated CEU individuals ²	0.26	0.08
HumanCNV370-Duo	88 unrelated HapMap individuals ¹	0.30	0.25
	33 unrelated CEU individuals ²	0.32	0.22
HumanHap650Y	88 unrelated HapMap individuals ¹	0.29	0.13
	33 unrelated CEU individuals ²	0.28	0.12
HumanHap550	88 unrelated HapMap individuals ¹	0.30	0.16
	33 unrelated CEU individuals ²	0.31	0.15
HumanHap300-Duo	83 unrelated HapMap individuals ¹	0.31	0.26
	31 unrelated CEU individuals ²	0.35	0.23
HumanHap240S ³	84 unrelated HapMap individuals ¹	0.27	0.37
	31 unrelated CEU individuals ²	0.28	0.36

¹HapMap individuals include those from four HapMap analysis panels: Caucasian, Han Chinese, Japanese, and Yoruba
²CEU includes only unrelated individuals from Caucasians of European Ancestry
³Excludes 180 mitochondrial SNPs

$$L_{AAB} = \frac{1}{0.18\sqrt{2\pi}} \exp\left(-\frac{(LRR - 0.3)^2}{2(0.18^2)}\right) + \frac{1}{0.03\sqrt{2\pi}} \exp\left(-\frac{(BAF - 0)^2}{2(0.03^2)}\right).$$

Likelihoods are also computed for other model genotypes listed in Table 1 with the exception of the homozygous deletion (DD). For homozygous deletions, a very low LRR is expected, but the BAF may be any value between zero and one. Therefore, the likelihood of a double deletion (LDD) is calculated by the equation:

$$L_{DD} = \frac{1}{2 \times \sqrt{2\pi}} \exp\left(-\frac{(LRR - (-5))^2}{2(2^2)}\right).$$

These likelihoods are then summarized by four composite copy number likelihoods:

$$\begin{aligned} L_0 &= L_{DD} \\ L_1 &= L_A + L_B \\ L_2 &= L_{AA} + L_{AB} + L_{BB} \\ L_3 &= L_{AAA} + L_{AAB} + L_{ABB} + L_{BBB} \\ L_4 &= L_{AAAA} + L_{AAAB} + L_{ABBB} + L_{BBBB} \end{aligned}$$

where L_k denotes the likelihood of copy number k for integer values of k and the likelihood of a genotype for non-numeric values of k . The preliminary copy number estimate (X) is defined as the average of the five modeled copy numbers, weighted by their respective likelihoods:

$$X = \frac{L_1 + 2L_2 + 3L_3 + 4L_4}{L_0 + L_1 + L_2 + L_3 + L_4}.$$

Breakpoint Identification

Preliminary copy number estimates are the inputs to the core partitioning algorithm. The goal of partitioning is to identify regions of the genome where the values of X are consistently higher or lower than 2, the expected value for a diploid sample. To find an aberrant region, the algorithm orders the X values by their position along a chromosome and searches for the indices i and j such that the values $X_i \dots X_j$ are maximally different than those outside this region. Thus, the algorithm seeks to maximize $|Z_{ij}|$ over all i and j with $i < j$, defined by the equations:

$$S_i = X_1 + \dots + X_i, 1 \leq i \leq n$$

$$Z_{ij} = \{1/(j-i) + 1/(n-j+i)\}^{-1/2} \times \{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)\},$$

where n is the number of loci assayed on the chromosome.

An exhaustive search through all pairs of i and j scales quadratically with n and is therefore an inefficient process for use with Illumina whole genome genotyping products. To simplify the calculations required, *cnvPartition* uses a sliding window strategy to maximize $|Z_{ij}|$, but where $j = i + w$, with w the defined window size. Once the optimal window size value is found, the algorithm attempts to extend the window in both directions to further maximize the value of $|Z_{ij}|$. As implemented, the algorithm repeats this procedure for $w = 4, 8, 16$, and 32 then reports the i and j corresponding to the maximal $|Z_{ij}|$ found. Once a maximally different segment is found, $|Z_{ij}|$ is compared to a pre-determined threshold (default is 6). If the threshold is exceeded, the boundaries are noted and the algorithm is applied recursively to the regions between 1 and i , $i+1$ and j , and $j+1$ through n . The threshold of 6 was chosen as a default because it minimizes false positives, particularly for short aberrations.

Copy Number Assignment to Partitioned Regions

The partitioning procedure results in a set of putative breakpoints scattered across the genome. The next step is to assign a copy number for each region lying between two consecutive breakpoints. To do this, L_0, L_1, L_2, L_3 and L_4 for each locus within the region are used. For each putative copy number (0–4), the logarithms of all L_k for each k are summed. The k with the highest sum is the copy number assigned to this region. For regions with copy numbers other than two, the algorithm also assigns a confidence score for the copy number that is called. The confidence score is defined as the sum of all logged likelihoods in the region for the assigned copy number minus the sum of all $\log L_2$ values for loci in the region.

CNV Display in BeadStudio

The copy number values are then used to create CNV regions and bookmarks in BeadStudio for visualization of aberrant regions. The *cnvPartition* algorithm incorporates two user-definable thresholds for optimization of CNV detection. The Confidence Threshold allows users to filter out CNV regions that have low confidence values. The default of 35 was determined empirically using normal HapMap samples on the Illumina Human1M BeadChip. The Probe Gap Size Threshold allows users to filter out CNV regions that are in large probe gaps, such as centromeres. The default of 1,000,000 (1Mb) was determined empirically to help prevent CNV regions from being falsely detected across centromeres and other large gaps.

HOMOZYGOSITY DETECTOR

The homozygosity detector algorithm can be used to autobookmark samples with extended tracts of homozygosity (single-sample analysis only). Homozygosity tracts may result from inbreeding, large scale gene conversion, uniparental disomy (UPD), or chromosomal deletions. Other factors such as population history or low recombination rates may also contribute to creating extended regions of LOH. This algorithm uses SNP frequencies to calculate the expectation that a single SNP is homozygous in a single sample. The algorithm then calculates the X^2 value of the observation of zero heterozygotes in N SNPs versus the expected number of heterozygotes given the frequencies of the SNPs. In this algorithm, there is a fixed cutoff significance ($X^2 = 23.5$), which corresponds to 50 contiguous SNPs each with a minor allele frequency (MAF) of 0.2.

This algorithm requires that each LOH region contains at least 50 homozygous SNPs by default. All LOH regions with more than 50 SNPs and $X^2 > 23.5$ are bookmarked. For example, the mean SNP spacing of 2.7 and 7.9kb in the Human1M and HumanCNV370-Duo DNA Analysis Bead-Chips results in an average sensitivity to detect regions of approximately 135kb and 395kb length, respectively.

Homozygosity Detector Algorithm Process

1. The X^2 threshold is preset to 23.5, but users have the ability to change this value.
2. The minimum number of SNPs per region is preset to 50, but users have the ability to change this value.
3. Prior to doing any analysis, the allele frequencies are calculated for each SNP. The allele frequencies are used to calculate the expectation that each SNP is heterozygous in a single sample assuming Hardy-Weinberg equilibrium.
4. Next, for each sample, all of the genotypes on each chromosome are scanned, and all of the contiguous regions without heterozygous genotypes are located.
5. For each of these regions, the expected number of heterozygotes is calculated by the equation:

$$E_{het} = \sum_{i=1}^N 2f_i(1-f_i)$$

where N is the number of homozygous SNPs and f_i is the frequency of either of the SNP alleles in the general population. The X^2 value is given as:

$$X^2 = \frac{(N_{hom} E_{hom})^2}{E_{hom}} + \frac{(N_{het} - E_{het})^2}{E_{het}}$$

where N_{hom} and N_{het} are the number of homozygous and heterozygous genotypes respectively, and E_{hom} and E_{het} are the expected number of homozygous and heterozygous genotypes. By definition, there are no heterozygotes, so the X^2 value can be simplified to:

$$X^2 = \frac{NE_{het}}{N - E_{het}}$$

where N is the number of SNPs with genotype calls.

6. Each segment that is more significant than the pre-defined or user-supplied X^2 threshold value and has more SNPs than the predefined or user-supplied minimum number of SNPs is bookmarked.

CHROMOZONE ALGORITHM

The ChromoZone algorithm can be used to autobookmark chromosomal aberrations in single-sample mode. The two types of phenomena that this algorithm identifies are *heterozygote sparse* regions, which may include deletions/LOH and high-level amplifications, and *heterozygote split* regions, which includes aberrations such as duplications. There is no option for adjusting the window size as the algorithm automatically calculates this parameter based on the SNP density of the BeadChip used for the study. This algorithm uses the following parameters: SNP position, allele A and B calls, and GenCall score. Intensity (R) information is not used in this version of the algorithm. Therefore, this version is unable to distinguish between different types of amplifications and deletions/LOH.

ChromoZone Algorithm Process

1. The GenTrain score threshold is set to 0.25.
2. The window size is estimated based upon the SNP density of the BeadChip of interest. A window is centered on a SNP and four additional windows are created; these include two windows of the original size, one window larger than the original window, and one smaller than the original window, to improve the robustness of the algorithm.
3. The reflective copy angle (RCA) is computed, and RCA less than 0.2 is removed. In this case, the copy angle refers to the same terminology as the B allele frequency. This process reflects all data downwards so only values between 0.0 and 0.5 are examined.
4. Densities are calculated along with several other parameters. The mean RCA is calculated. The heterozygote density (hd) is calculated as:

(heterozygote calls)/(heterozygote+homozygote calls).

The no-call density (nd) is calculated as:

#no calls/#SNPs.

5. Features are extracted using the logic functions as follows:
 - A heterozygote split is noted if the mean RCA is small or the no-call density is large.
 - A heterozygote split is not noted if the meanRCA is large or if the no-call density is small and the heterozygote density is small.
 - A heterozygote sparse region is noted if the heterozygote density is small and the no-call density is small.
 - Heterozygote sparse is not noted if heterozygote density is large or the no-call density is large.
6. The remaining gaps are then filled in. If a region is not marked, but is flanked by two regions that are marked, this unknown region is filled in with the same bookmark.

Other Important Information about ChromoZone

- The window size is calculated separately for each chromosome based on the SNPs that fall into the 99th percentile of the distances between SNPs. This ensures that the weakest distance is covered.
- The bin size is limited to a minimum of 100Kb and a maximum of 2.0Mb.
- There is a minimum of 20 SNPs per window and 15 calls per window.

SUMMARY

BeadStudio provides several methods to analyze SNP and probe intensity data to identify chromosomal regions with LOH and copy number variations. The software plug-ins described in this technical note are freely available to BeadStudio users to provide extended functionality.

The LOH Score algorithm provides statistical information about chromosomal aberrations of interest. This information includes the probability of LOH existing. This algorithm can be used to quickly identify interesting regions in large sample sets or to further analyze a more refined region.

Automated bookmarking algorithms save time by automatically scanning and categorizing samples. Researchers can use *cnvPartition* to find and calculate copy numbers, *Homozygosity Detector* to identify extended tracts of LOH, or *ChromoZone* to identify various chromosomal aberrations.

The open architecture of Illumina BeadStudio software allows for customized and advanced analysis tools for the downstream analysis of Illumina DNA Analysis BeadChip Genotyping data. The plug-ins described in this document can be downloaded from within the BeadStudio Portal. Alternatively, plug-ins can be downloaded from the *illumina•Connect* webpage at <http://www.illumina.com/illuminaconnect>. *illumina•Connect* is a new collaborative program for facilitating the development of 3rd party tools and applications for DNA and RNA analysis of Illumina BeadArray™ products. Plug-in development is on-going and the *illumina•Connect* webpage will continue to be updated with new releases of these plug-ins.

ADDITIONAL INFORMATION

Visit our website or contact us at the address below to learn more about Illumina DNA Analysis Products and Software solutions.

Illumina, Inc.
Customer Solutions
 9885 Towne Centre Drive
 San Diego, CA 92121-1975
 1.800.809.4566 (toll free)
 1.858.202.4566 (outside the U.S.)
techsupport@illumina.com
www.illumina.com

REFERENCES

- (1) <http://www.illumina.com/illuminaconnect>
- (2) Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5: 557-572.
- (3) Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23: 657-663.

FOR RESEARCH USE ONLY