illumina®

DNA Copy Number and Loss of Heterozygosity Analysis Algorithms

Detection of copy-number variants and chromosomal aberrations in GenomeStudio® software.

Introduction

Illumina has developed several algorithms for detecting copy number variants (CNVs) and other structural variants from microarray data (Table 1). These algorithms are available as individual software plug-ins for the GenomeStudio Genotyping Module and can be downloaded from the GenomeStudio Portal, from the download section of the Illumina Connect web page[1], or from MyIllumina (my.illumina.com). The plug-ins are used within the CNV Analysis workbench, and results can be visualized within the GenomeStudio Full Data Table, in the Illumina Genome Viewer (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.

CNV Region Report

CNV Region Report is a software plug-in for GenomeStudio that generates three separate CNV reports.

- **Standard Report**—Lists each CNV and loss of heterozygosity (LOH) region for each selected sample.
- **Allele-Specific Copy Number Report**—Estimates the allele-specific copy number for each probe entry (e.g., A- or AAB). In the output file, the CN_GTYPE column is calculated using the CNV Value (as determined by CNVPartition), the B Allele Frequency data, the GTYPE (genotype column), and the theoretical B Allele Frequency normal distributions for each copy number.
- **PLINK CNV Input Report**—Creates input files for some of the CNV features of the PLINK genome-wide association study (GWAS) and CNV analysis application[2].

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: the log R ratio (LRR) and B allele frequency (BAF). Because 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 resolution for detecting true copy number changes.

Table 1: GenomeStudio Copy Number Algorithms

Algorithm	Function
CNV Region Report	Generates three separate CNV reports
cnvPartition	Calculates copy numbers with confidence scores and generates CNV regions
Homozygosity Detector	Autobookmarks samples with extended tracts of homozygosity (single-sample analysis only)
LOH Score	Estimates the likelihood of a region exhibiting LOH

Copy Number Estimation

cnvPartition models LRRs and BAFs for each of 14 different copy number scenarios as simple bivariate Gaussian distributions (Table 2).

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 14 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 2 and the standard normal density are used:

$$L_{AAB} = \frac{1}{0.18\sqrt{2\pi}}\exp\left(-\frac{(LRR-0.3)^2}{2(0.18^2)}\right) + \frac{1}{0.03\sqrt{2\pi}}\exp\left(-\frac{(BAF-\frac{1}{3})^2}{2(0.03^2)}\right)$$

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:

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

Table 2: 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	0.75	0.03
BBBB	4	0.75	0.18	1	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)

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

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

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 indexes i and j such that the values $X_i \ldots X_j$ 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:

$$S_i = X_1 + \ldots + X_i, 1 \le i \le n$$

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

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[3,4]. 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. After the optimal window size value is found, the algorithm attempts to extend the window in both directions to maximize the value of $|Z_{ij}|$ further. 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. When 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 2, 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.

Additional Usage Notes

- Regions with copy number = 1 on the X or Y for males are filtered from the CNV results.
- Probes that are designated as Intensity Only are treated differently than normal probes. The B Allele Frequency is ignored for these probes.
- Y probes are not considered for samples designated as female.

Homozygous Region Detection

cnvPartition also includes a homozygosity detection algorithm that runs separately from the partitioning algorithm already described. This algorithm only runs on copy number 2 regions by default. Therefore, it is sometimes called a Copy Neutral LOH Detector. The logic is similar to that used in the homozygosity detector autobookmarking plug-in (see next section). However, additional logic has been added to simplify usage for the end user. Instead of adjusting the ChiSquare threshold as in the autobookmarking plug-in, the user can simply adjust the MinHomozygousRegionSize configuration parameter. By default, this is set to 10 Mb based on empirical testing.

Experimental Features

Several experimental features are included in cnvPartition v3.2.0. These features are all disabled by default but can be activated by adjusting the associated configuration files. Configuration details for each feature are available in the cnvPartition user document[5]. These features include:

- **LOH detection for the entire genome.** Previous versions only allowed LOH detection to be run on regions with a copy number of 2.
- **Log R Ratio adjustment for the Y chromosome.** Y SNPs are clustered using only males, so the log R Ratio indicates that the copy number = 2. The adjustment lowers all log R Ratios on Y. If Y chromosome SNP clusters are already adjusted, additional adjustment could provide inconsistent results.
- **Support for highly amplified genomes.** These are common with many cancer samples. Log R Ratios are adjusted upward based on calculation of average genomic ploidy.
- A GC wave adjustment based on linear regression of LRR vs. GC content in probes. GC waves can be present when using incorrectly quantified DNA in the Infinium Assay, or they might be present in regions of high or low GC content.

CNV Display in GenomeStudio

The copy number values are then used to create CNV regions and bookmarks in GenomeStudio 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 (1 Mb) was determined empirically to help prevent CNV regions from being falsely detected across centromeres and other large gaps.

Algorithms for Automated Bookmarking

GenomeStudio can use 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 to verify known regions of interest. Bookmarks can be edited whether they are created manually or generated using the autobookmarking tool. All bookmarks can be exported to share with other users. The Homozygosity Detector is an autobookmarking plug-in available for GenomeStudio.

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.

Homozygosity Detector Algorithm Process

1. The X^2 threshold is preset to 23.5, but users can change this value.
2. The minimum number of SNPs per region is preset to 50, but users can change this value.
3. The allele frequencies are calculated for each SNP before analysis. 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 predefined or user-supplied X^2 threshold value and has more SNPs than the predefined or user-supplied minimum number of SNPs is bookmarked.

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).

Figure 1: LOH Score Equation

$$P_{error} = 0.001$$

$$LOHScore = Log_{10} \frac{\prod_{i=1}^{N} P_i(gt_i|LOH)}{\prod_{i=1}^{N} P_i(gt_i|NoLOH)}$$

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

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

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

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 chromosomal position of the SNP. The equation used to determine the LOH Score is shown in Figure 1. The recommended window size depends upon the density of probes on the product in use, as defined by the following equation

$$\text{Minimum Window Size} = \frac{\text{Factor Number}}{\text{Number of Probes}}$$

where Factor Number = 88,000

The window size for a specific workspace may require optimization depending upon the type of aberration under examination and the quality of the data.

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 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 270 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 $(1\text{-}e)^N$. The likelihood of no LOH is $(1\text{-}h)^N$. Therefore, the log odds ratio is: $\mathbf{\log_{10}\{(1\text{-}e)^N/(1\text{-}h)N\}}$, which is the same as $\mathbf{N\{\log_{10}(1\text{-}e) - \log_{10}(1\text{-}h)\}}$, 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.

Conversely, if that stretch contains M heterozygote calls, the likelihood of LOH decreases and the equation is adjusted to $\mathbf{(1\text{-}e)^{(N-M)} \times e^M}$, because heterozygotes in a region with LOH occur only through genotyping errors. The likelihood of No LOH also changes and becomes $\mathbf{(1\text{-}h)^{(N-M)} \times h^M}$. 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 that 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.

Summary

GenomeStudio 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 GenomeStudio users to provide extended functionality.

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

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 identify interesting regions in large sample sets quickly or to analyze a more refined region further.

The open architecture of Illumina GenomeStudio 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 the GenomeStudio Portal or from the Illumina Connect web page at www.illumina.com/illuminaconnect. Illumina Connect is a collaborative program for facilitating the development of third-party tools and applications for DNA and RNA analysis of Illumina BeadArray™ products. Users should check the Illumina Connect web page regularly for new or updated plug-ins.

References

1. www.illumina.com/illuminaconnect
2. pngu.mgh.harvard.edu/~purcell/plink/
3. 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.
4. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23: 657–663.
5. my.illumina.com/

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 970-2007-008 Current as of 09 July 2014

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA