Interpreting Infinium® Assay Data for Whole-Genome Structural Variation

Illumina offers a broad portfolio of DNA Analysis BeadChips for analyzing genotypes and structural variation. This document provides basic information about identifying copy number aberrations using Illumina whole-genome genotyping technology.

Introduction

Abnormalities in chromosomal structure are an important source of genetic variability with direct impacts on phenotypic variation and disease susceptibility[1]. Structural variation in the genome consists of several classes of variants. Illumina BeadChips based on the Infinium Assay can detect both copy number-variation (e.g., amplifications, duplications, deletions) and copy-neutral structural variants (e.g., copy-neutral LOH, uniparental disomy). The Infinium Assay delivers two principal types of data for assayed SNP loci: genotype and intensity. These two parameters are analyzed in combination for specific identification and precise breakpoint determination of variants.

Marker Design Strategy

To create the best tools for cytogenetic analysis and copy number variation (CNV) identification, Illumina has taken advantage of the unconstrained marker design of the Infinium Assay and the high-density Infinium HD BeadChip platform. All Illumina whole-genome panels consist of a uniform distribution of SNP markers to create the fewest large gaps across the entire genome for high-resolution aberration detection and breakpoint mapping. To supplement this uniform coverage of the genome, Illumina has leveraged the design flexibility of the Infinium Assay to target both common and rare regions likely to contain phenotypically relevant copy number variation (CNV), such as genes, regions commonly screened for cytogenetic abnormalities, and segmental duplications.

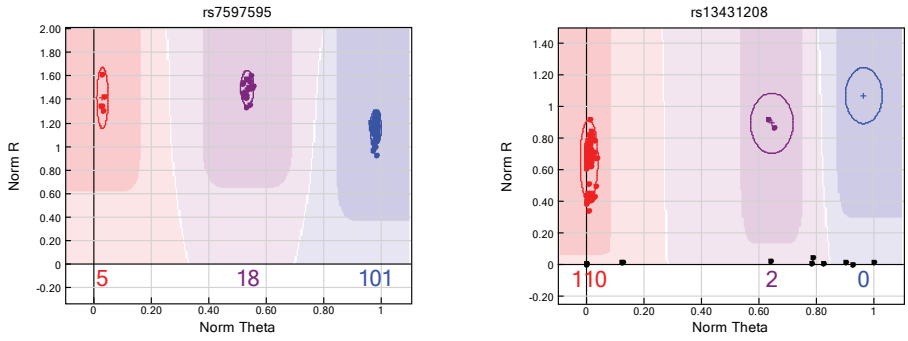
The majority of markers on Infinium BeadChips are SNPs, with intensity-only non-polymorphic probes used to fill in regions underrepresented by SNPs or to replace SNPs that do not perform well. SNPs are preferred for their higher signal-to-noise ratios and the additional genetic information they provide compared to the intensity-only probes used on most DNA microarrays. The essentially digital nature of genotypes (allele A or allele B) results in high signal-to-noise ratios, which provide a very robust measurement of copy number variation.

Intensity-Only Probes

GenomeStudio and KaryoStudio data analysis software use the signals generated from all Infinium BeadChip markers (both polymorphic SNP and non-polymorphic intensity-only markers) for CNV detection[2]. GenomeStudio and KaryoStudio use markers that are designated in stats for genotyping and copy number information (Figure 1). All in stats markers are SNPs (note, not all SNPs are in stats) and are therefore also used for calculating SNP statistics (e.g., call rate, heritability, and reproducibility). Markers designated out of stats are probes that are used only for intensity information in copy number calculations. Locus status (in stats or out of stats) is determined by Illumina and is preset in the bead pool manifest that is supplied with each product. For in stats markers, the Intensity Only column value in the manifest is 0; for out of stats markers, the Intensity Only column value is 1.

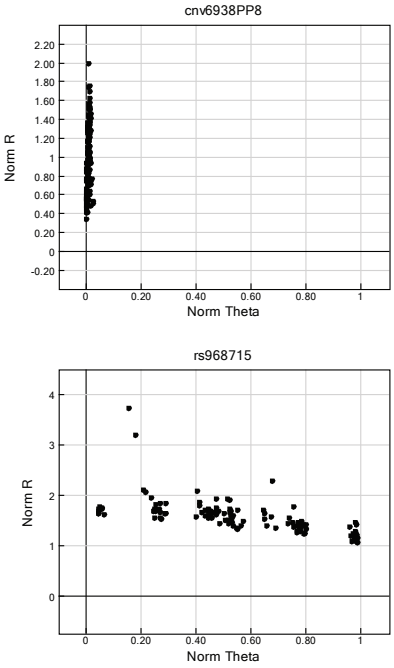
Illumina scientists analyze the genoplots during the development of each Infinium product to determine whether markers should be in

Figure 1: SNP Genoplot Examples



The left panel shows a normal genoplot with samples falling in each of three genotype clusters (red points are AA, purple points are AB, blue points are BB, expected cluster positions are indicated by ellipses). In the genoplot shown in the right panel, the samples represented by black points show a dramatic drop in intensity, which may signify a homozygous deletion in those samples.

Figure 2: Genoplots of out of Stats Markets



The genoplot in the top panel indicates that this intensity-only probe is assaying a non-polymorphic locus because all data points are along the θ = 0 axis and therefore monomorphic. Although the locus in the bottom panel is polymorphic, it does not exhibit typical clustering useful for genotyping; it is still used for CNV identification as an intensity-only probe. Since these genoplots are not of SNPs, there are no cluster positions or color-coded regions on the plot.

stats or out of stats. Reasons for a marker being out of stats include intentional design due to a lack of acceptable SNP locus in a region (Figure 2, top panel), or the finding that it is useful for CNV detection even though the SNP locus doesn't exhibit the typical three-genotype cluster pattern (Figure 2, bottom panel).

How Data are Generated

For each SNP marker, the Infinium HD Assay two-color readout results in intensity values measured in each of the two color channels (two alleles). Polar transformation of these data provide normalized intensity values (R) and allelic intensity ratios (θ). These parameters can be visualized in GenomeStudio as a genoplot (Figure 1). These values are used to calculate two metrics for each SNP marker in a sample—relative to those expected from a standard cluster position—which are used to determine SNP genotypes and copy number estimates (Figure 3). GenomeStudio software generates plots of all SNPs for B allele frequency (interpolated from known B allele frequencies of the three canonical clusters: 0, 0.5, and 1) and log R ratio ($\log2(R_{observed}/R_{expected})$), where $R_{expected}$ is interpolated from the observed allelic ratio with respect to the canonical genotype clusters[3,4,5]. KaryoStudio calculates these metrics, but does not display individual genoplots.

Standard Cluster File

A reference sample is not required to be run in parallel with each sample. Rather, standard canonical cluster position information is used to computate both log R ratio and B allele frequency. Therefore, the use of an appropriate cluster file is essential for accurate cytogenetic analysis. The standard cluster file (*.egt file) supplied by Illumina for Infinium HD BeadChips is generated by using a diverse set of more than 200 HapMap DNA samples, and should therefore be applicable to most general experimental cohorts. A custom-generated cluster file may provide improved analysis quality if experimental samples are from an isolated population and do not fit standard cluster positions well[4,5]

Because all calculations for log R ratio data points are made by comparing experimental data to canonical genotype clusters, it is imperative that the experimental conditions match the conditions used to determine canonical genotype clusters as closely as possible. These include precise quantification of DNA input with PicoGreen reagent. For all Infinium HD Quad and 12-sample BeadChips, 200 ng of DNA input is required. For all Infinium HD Duo (two-sample) BeadChips, 400 ng of DNA input is required. Deviations from these requirements typically expose GC-rich regions of the genome and can mask structural aberrations. In these cases, analysis algorithms may not be able to correctly identify aberrations. In cases where DNA input is accurately quantified, individual log R ratios in normal regions tend to be nearly zero, allowing accurate and precise identification of aberrations. Other important procedures for generating the highest log R ratio data quality are calibrating oven temperatures and following the Infinium Assay protocol exactly.

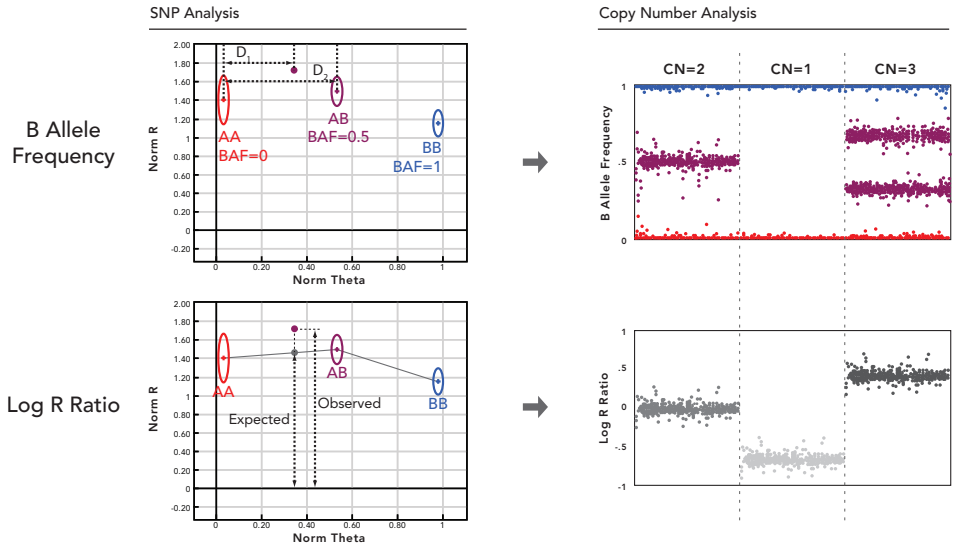
Copy Number Aberration Identification

Copy number aberrations are readily apparent in plots of log R ratio as deflections in the y-dimension and discrete patterns in the B allele frequency plot (Figure 3). Increases in log R ratio relative to the baseline result from increased signal intensity of a region, which represents increases in copy number (i.e., duplications or amplifications). Deletions show up in log R ratio plots as a decrease in signal intensity. For example, a log R ratio of approximately 1 ($\log_2$ of 50% signal decrease = -1) is expected theoretically from a hemizygyous deletion where there is only one copy of a region, rather than the normal two copies. Deviations from normal copy number result in greater or fewer numbers of potential genotypes, which are visible as more or less than three clusters of data points in the B allele frequency. Copy neutral aberrations are only visible with B allele frequency plots, since their copy number is two and looks normal in the log R ratio plot.

Infinium BeadChip Resolution

The effective resolution calculated for each BeadChip provides an estimate of the size of aberration that can be detected by analyzing data derived from that BeadChip. Effective resolution is defined as the median marker spacing of a BeadChip multiplied by an appropriate window size. For general purposes, we implement a window size of 5 because this encompasses two to three heterozygous SNPs in a human genome with an average heterozygosity of 30%–40%. For example, the HumanOmni1-Quad, which provides over 1.1 million markers at a median spacing of 1.2 kb, has an effective resolution of ~6.0 kb (Table 2).

Figure 3: Calculation of Log R Ratio and B Allele frequency



The allelic copy ratio in terms of B allele frequency (BAF) is calculated from the θ value of a sample and the expected cluster positions (ellipses, left top panel). The allele frequency is determined as a linear interpolation in the θ-dimension related to the allele frequency of each cluster (0.0, 0.5, and 1.0). In this example, a data point (purple dot) falling approximately 2/3 of the distance from the AA to the AB cluster (D1/D2 = .66) has an allele frequency of 0.33 (0.66 * 0.5).
The $\log_2$ R ratio is calculated as the ratio between observed normalized intensity of the experimental sample to the expected intensity (left bottom panel). The expected intensity is determined as a linear interpolation as a function of the sample θ (grey line) of the expected cluster positions (ellipses).
These two transformed parameters, B allele frequency and $\log_2$ R ratio, are then plotted along the entire genome for all SNPs on the array (right panel). The plot of these two parameters exhibits diagnostic signature profiles of copy number (example copy numbers 2, 1, and 3 shown) and specific classes of structural variation. Figure adapted from Peiffer, et. al, 2006.

Sample Types

The Infinium HD Assay supports a range of sample types that have been shown to yield high-quality data for CNV and cytogenetic analysis. These include blood, solid tumors, whole-genome amplified DNA, and fresh frozen samples.

Analysis Software

GenomeStudio Software

GenomeStudio is a feature-rich integrated suite of software modules for the exploratory visualization and analysis of Illumina microarray and sequencing data. Analysis of Infinium BeadChip data is performed using the GenomeStudio Genotyping (GT) Module, and the customizable display tools of the GenomeStudio framework. These tools include automated algorithms that can scan for and characterize aberrations, the integrated Illumina Genome Viewer (IGV) which is ideal for CNV and cytogenetic analysis, the Illumina Chromosome Browser (ICB) for closer examination of affected regions, and heat maps for the identification of common aberrations across large sample sets.

KaryoStudio Software

KaryoStudio is a streamlined software package for rapid and routine cytogenetic analysis applications using Infinium BeadChip data (Figure 4). Multiple samples can be loaded and analyzed together. Found regions in each sample are displayed automatically after data file input using the built-in algorithm and custom filtering settings.

An intuitive workspace with tables and chromosome browser allows for rapid visualization and verification of all results. Customizable cross-matching to external databases helps determine the potential relevance of found regions, and PDF report generation is provided for output of selected results.

cnvPartition

Illumina recommends the cnvPartition plug-in algorithm for CNV detection and characterization with KaryoStudio or GenomeStudio. The cnvPartition algorithm is based on a recursive partition method that is described in the DNA Copy Number Analysis Algorithms technical note[6]. cnvPartition has been optimized for speed and accuracy using log R intensity and B allele frequency in combination to identify chromosomal aberrations, estimate copy number values, and calculate per-region confidence scores.

Third-Party GenomeStudio Plug-ins

GenomeStudio software provides an open API for integrating third-party applications to support alternative downstream data analysis workflows for CNV or cytogenetic applications (Table 1).

Examples of Structural Aberrations

Figures 5–13 are examples of typical log R ratio and B allele frequency plots that indicate the presence of various structural variants. For clarity, aberrant regions are annotated with shading and the average log R ratio over a region is indicated with a red line. For examples of

Figure 4: Typical Workflows for Illumina CNV or Cytogenetic Analysis

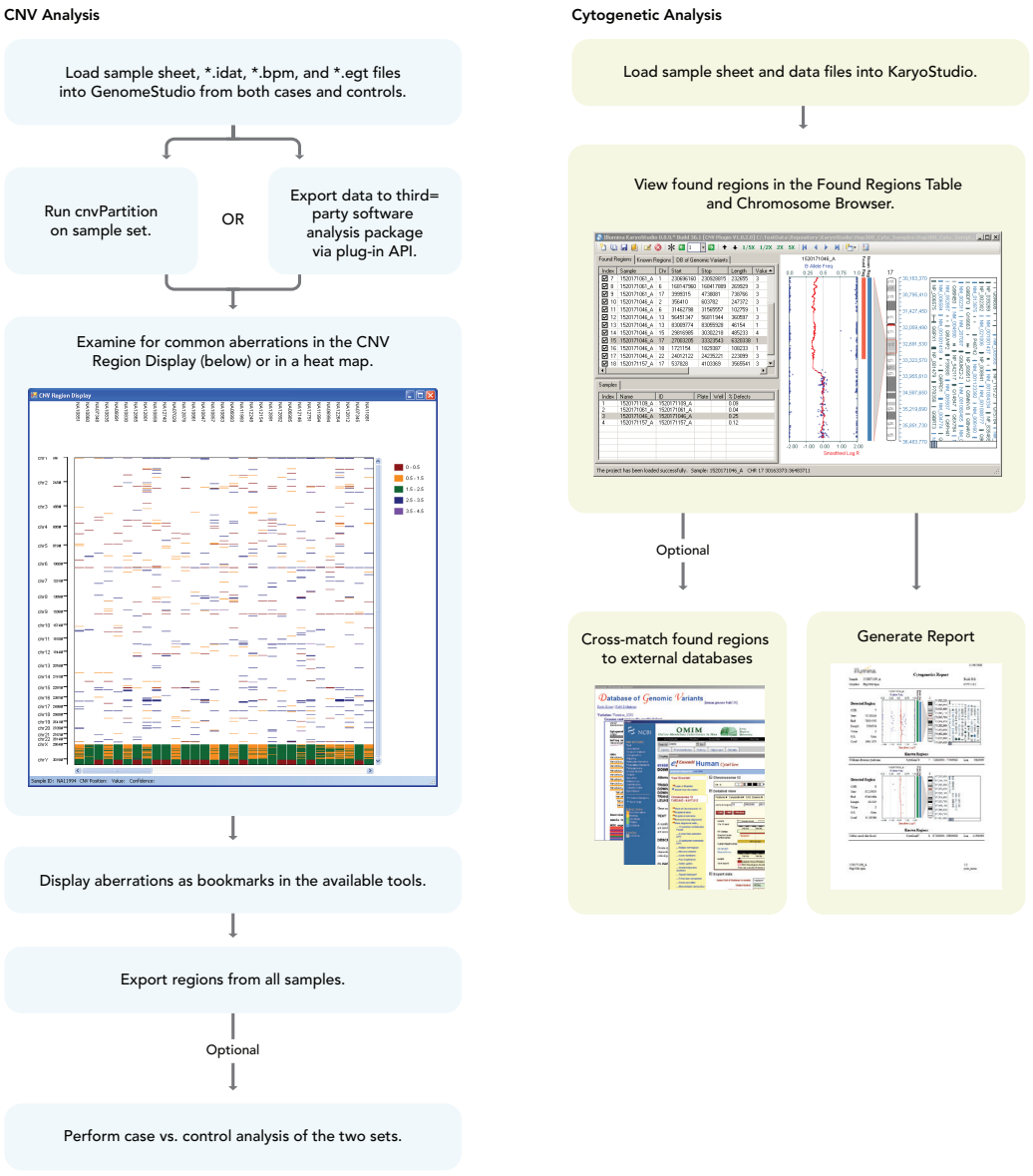


Table 1: Genomestudio Plug-ins for Structural Variation Analysis

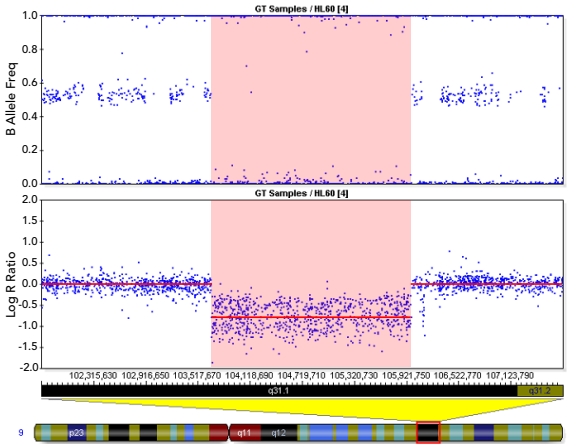
Algorithm	Source	Features
cnvPartition	• **Illumina* (algorithm & plug-in)**	• Fast to analyze and report (~1 min per sample) • Detects CNV regions and estimates CNV values • High data quality and accuracy • Finds small CNV regions (< 100 kb) • Computes confidence value for each locus
QuantiSNP	• **Oxford University (algorithm)** • **Illumina* (plug-in)**	• High data quality and accuracy • Finds small CNV regions (< 100 kb) • Provides confidence score for each locus • Optimized for Illumina BeadChips • Slow (~60 min per sample with Human1M BeadChip)
Partek GS v6.2	• **Partek (algorithm)** • **Partek plug-in* (custom report)**	• CNV detection and association with exon expression analysis • Can create models (PCA) for phenotypic association of CNVs • Leading bioinformatics software for genomic research
JMP Genomics v7.0	• **JMP/SAS (algorithm)**	• Applies 3D PCA for phenotypic association of CNVs
SNP & CN Variation Suite (CNAM)	• **Golden Helix (algorithm)** • **Golden Helix (plug-in)***	• Uses RP (segmentation) technique for CNV detection (fast detection) • Optimized to detect very small regions (high detection limit) • Includes permutation testing for accurate CNV detection • Suite enabled for WG CNV association with phenotype data
Nexus Copy Number	• **BioDiscovery (algorithm & plug-in)***	• Uses Hidden Markov Model (HMM) or Circular Binary Segmentation (CBS) approach to CNV detection • Enabled for WG CNV association with phenotypic data • Easy-to-use interface • Applies multiple permutations to improve accuracy of CNV detection • Used in both research and clinical cytogenetics labs
PennCNV	• **U Penn (algorithm)** • **Illumina* (plug-in)**	Published and available freely Uses Hidden Markov Model (HMM) approach for CNV detection
Exemplar for CN	• **Sapio (algorithm & plug-in)***	Not tested by Illumina
ArrayAssist	• **Stratagene (algorithm & plug-in)***	Not tested by Illumina

*Download from http://www.illumina.com/software/illumina_connect.ilmn

Table 2: Effective Resolution of Infinium HD Beadchips

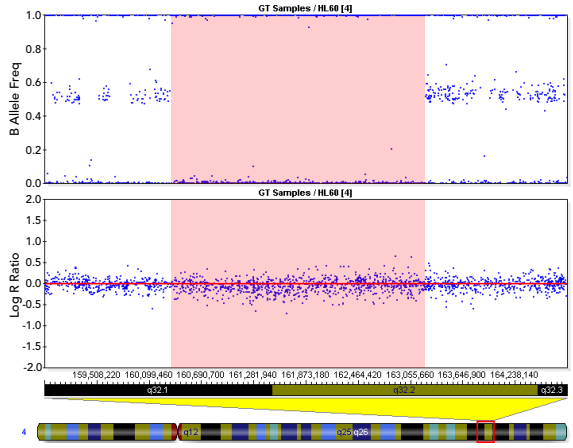
	HumanCytoSNP-12	Human610-Quad	Human660w-Quad	HumanOmni1-Quad	Human1M-Duo
Markers	299,671	620,901	657,366	1,140,419	1,199,187
Mean Spacing (kb)	9.6	4.7	4.4	2.5	2.4
Median Spacing (kb)	6.2	2.7	2.3	1.2	1.5
90th %ile Spacing (kb)	18.6	11.0	10.6	6.4	6.0
Effective Resolution (kb)	31	14	12	6	8

Figure 5: Hemizygous Deletion



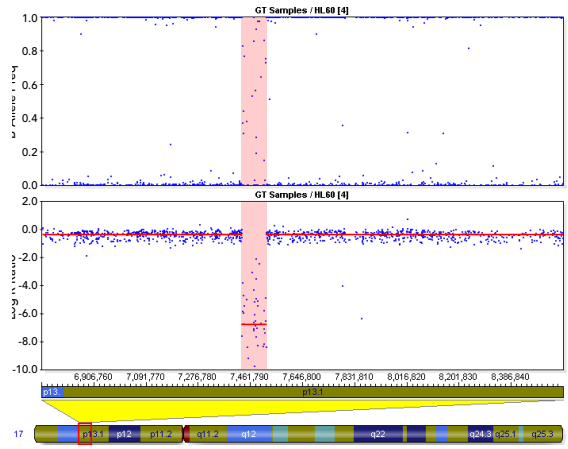
A hemizygous deletion (loss of one copy) is depicted in the ICB as a loss of heterozygotes in the B Allele Freq plot (top) and a loss of signal intensity in the Log R Ratio plot (bottom). In the region of the deletion (shaded), the log R ratio is log2 of 1/2, or -1.

Figure 6: Copy-Neutral LOH



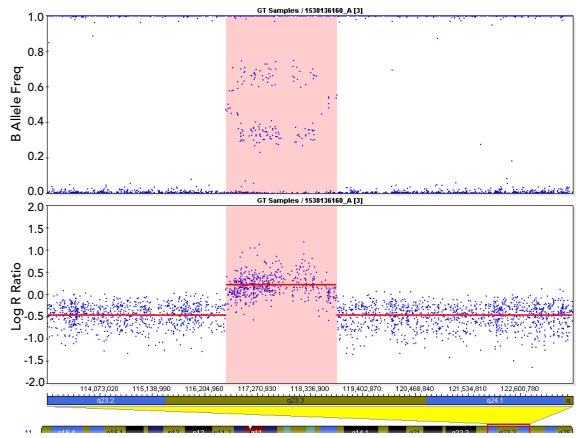
A region of copy-neutral LOH (shaded) is depicted by a loss of heterozygotes in the B allele frequency data but no change in the log R ratio (physical copy number).

Figure 7: Homozygous Deletion



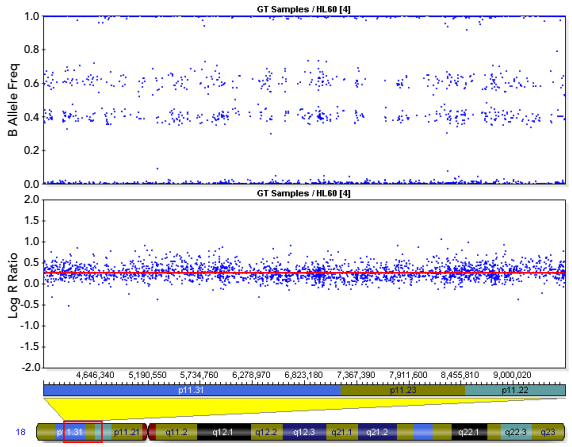
A region of homozygous deletion is where both copies of the chromosome have been lost (shaded). In this case, there are no SNPs present, so the genotyping data (B Allele Freq plot) appears like a “waterfall” as a result of noise in the absence of signal. The log R ratio in this region is the log2 of ~0/2, which is a highly negative value and is shown in the Log R Ratio plot as a large deflection downward.

Figure 9: Duplication Nested Within Two Flanking Deletions



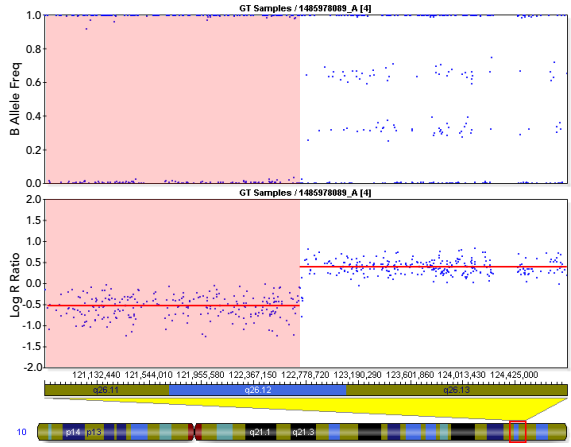
Regions of deletion (not shaded) are depicted by loss of signal intensity in the Log R Ratio plot to -0.5. An overlapping duplication (shaded) is depicted in the middle of the window by an increase in the Log R Ratio plot.

Figure 8: Duplication



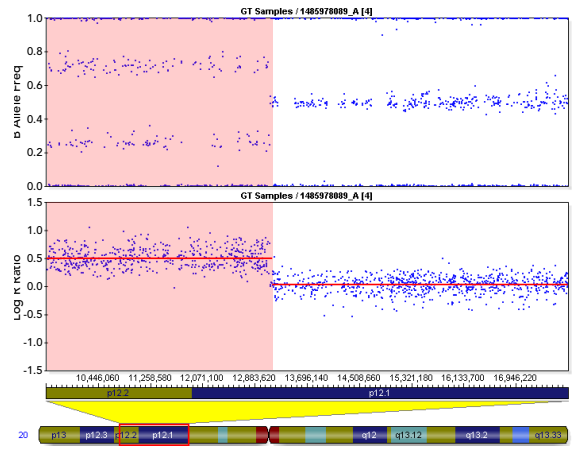
A duplicated region results in three total copies. This duplication is depicted by the B Allele Freq plot splitting into two new populations of data points representing the allelic ratios 1:2 and 2:1 (genotypes ABB and AAB). The duplication is also depicted by an increase in the log R ratio to ~0.4 (log2 of 3/2).

Figure 10: Combination of Multiple Aberations in a Tumor Sample



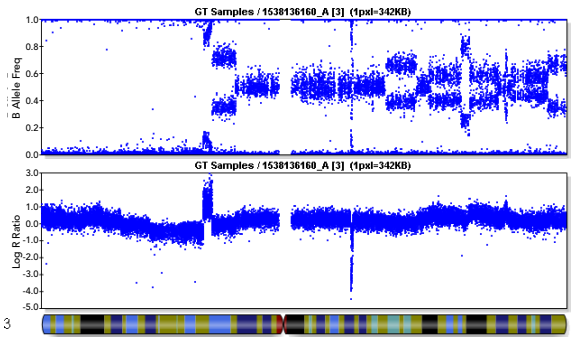
A hemizygous deletion (shaded) is depicted by the loss of heterozygotes in the B Allele Freq plot and a loss of intensities in the Log R Ratio plot. There is also a duplication (not shaded) indicated by the two clusters of data

Figure 11: Duplication in a Tumor Sample



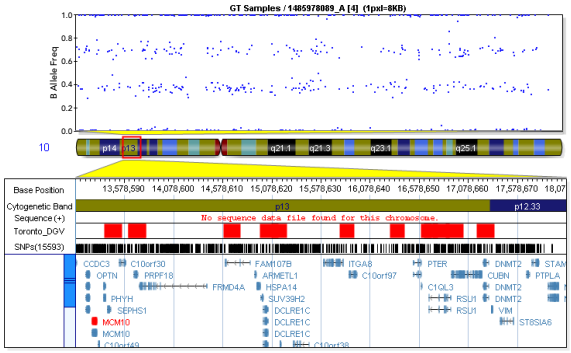
A duplication (shaded) is depicted by a splitting of heterozygotes in the B Allele Freq plot and an increase in intensities in the log R ratio to ~0.5.

Figure 12: Data Complexity of a Tumor Sample



The plot shows a profile of a breast tumor sample across the entirety of chromosome 3. The complexity of genomic aberrations coincident with tumor development is reflected in various and complex changes in the B allele frequency and log R ratio, including several duplications, deletions, and a homozygous deletion. Illumina recommends scanning such samples with cnvPartition. However, in some cases, due to the sample complexity and the majority of the genome not being diploid, the log R ratio may not accurately reflect the true copy number change.

Figure 13: View of Duplication in the Chromosome Browser



A duplication is depicted by the two populations of heterozygous data points in the B Allele Freq plot, representing the genotypes ABB and AAB (log R ratio is not shown). GenomeStudio allows you to import custom data tracks, such as those from the Toronto Database of Genomic Variants, shown as red regions below the B Allele Freq plot. accurately reflect the true copy number change.

successful CNV and cytogenetic analyses using Illumina DNA Analysis BeadChips, a list of citations is available at www.illumina.com/publications.

Structural Analysis in Mixed Tumor Samples

Many structural aberrations may not be present in every cell, but rather occur as mosaics with some cells having a normal karyotype and others being abnormal; this is especially true with tumor samples. The high-resolution Infinium HD Assay allows detection of low-level mosaicism in these sample types. The log R ratio and the B-allele frequency enable the quick determination of the level of mosaicism, as well as an indication of the mechanism behind the mosaicism7.

Demonstration Data Sets

To help begin performing Illumina CNV data analysis in your own lab, demonstration GenomeStudio Projects (*.gsc files) pre-loaded with HapMap samples are available. Please contact Technical Support for information about these demo projects.

Summary

Illumina DNA Analysis BeadChips are powerful tools for analyzing genome-wide structural variation. All of the features of Infinium DNA Analysis BeadChips—high density, broad coverage, and powerful markers, supported by comprehensive analysis software—provide a complete solution for CNV and cytogenetics applications. A high signal-to-noise ratio provides reliable and sensitive detection of changes in copy number. The large number of markers on Illumina BeadChips provides precise breakpoint mapping. By analyzing signal intensity and genotype, the two streams of data generated by the Infinium Assay, a wide variety of variants can be identified with high confidence.

References

1. http://genome.gov
2. GenomeStudio User Guides. Available from http://www.illumina.com/support/documentation.ilmn or the GenomeStudio Portal.
3. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, et al. (2006) High-Resolution Genomic Profiling of Chromosomal Aberrations Using Infinium Whole-Genome Genotyping. Genome Res 16: 1136-1148.
4. Updated Cluster Generation Protocol for X Chromosome SNPs (PDF). http://www.illumina.com/Documents/products/technotes/technote_x_chromosome_snp_cluster_generation.pdf
5. Infinium Genotyping Data Analysis (PDF). http://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf
6. DNA Copy Number Analysis Algorithms (PDF). http://www.illumina.com/Documents/products/technotes/technote_cnv_algorithms.pdf
7. Cytogenetics into Cytogenomics: SNP Arrays Expand the Screening Capabilities of Genetics Laboratories (PDF). http://www.illumina.com/Documents/products/appnotes/appnote_cytogenetics.pdf

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

illumina®