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 the design of Infinium Assays and general guidelines for analyzing structural variation using Illumina whole-genome genotyping technology.

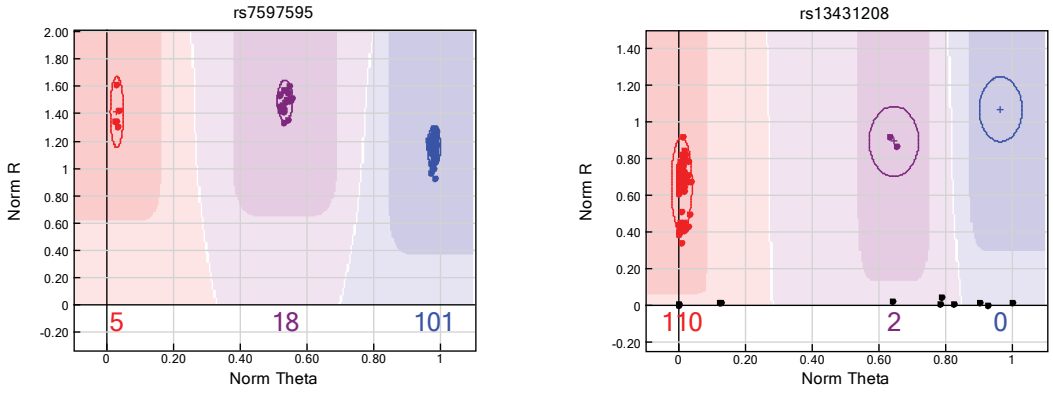
INTRODUCTION

Along with single nucleotide polymorphisms (SNPs), abnormalities in chromosomal structure are an important source of genetic variability with direct impacts on phenotypic variation and disease susceptibility. 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). 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

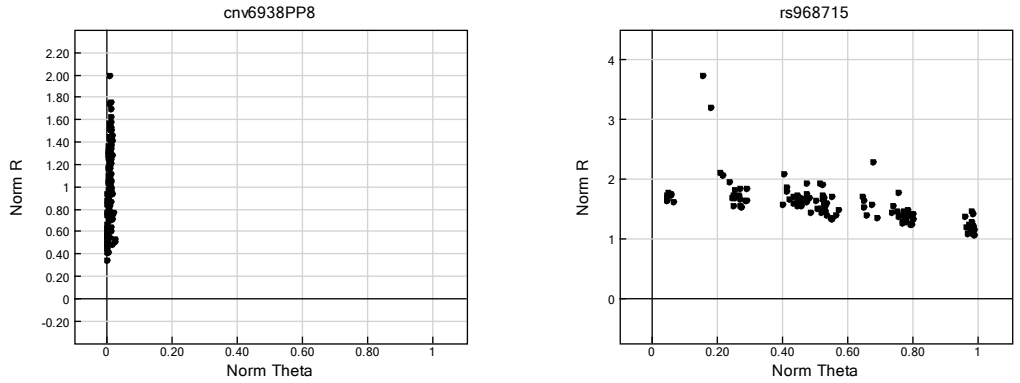
To create the best tools for cytogenetic analysis and copy number variation (CNV) identification, Illumina has taken advantage of the unconstrained marker design of the Infinium Assay and the high-density Infinium HD BeadChip platform. All whole-genome panels consist of a uniform distribution of SNP markers to create the fewest large gaps across the entire genome for high-resolution breakpoint mapping. To supplement this uniform coverage of the genome, Illumina worked closely with deCODE genetics to develop specific content to target the least stable 6% of the genome, which are the most likely regions to contain medically relevant copy number variation (CNV), such as segmental duplications and unSNPable regions lacking SNPs[1].

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 MARKERS



The genoplot in the left 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 right 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 and color coded regions on the plot.

The majority of markers on Infinium BeadChips are SNP genotyping markers, combined with intensity-only non-polymorphic probes filling in for regions underrepresented by SNPs or replacing SNPs that do not perform well. SNPs have higher signal-to-noise ratios and provide additional genetic information compared to intensity-only probes used on most DNA microarrays. Higher signal-to-noise ratios are a result of the essentially digital nature of genotypes (allele A or allele B).

Intensity-Only Probes

Infinium BeadChips use the signals generated from all markers (both polymorphic SNP and non-polymorphic intensity-only markers) for CNV detection. BeadStudio uses 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 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*, left 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*, right panel).

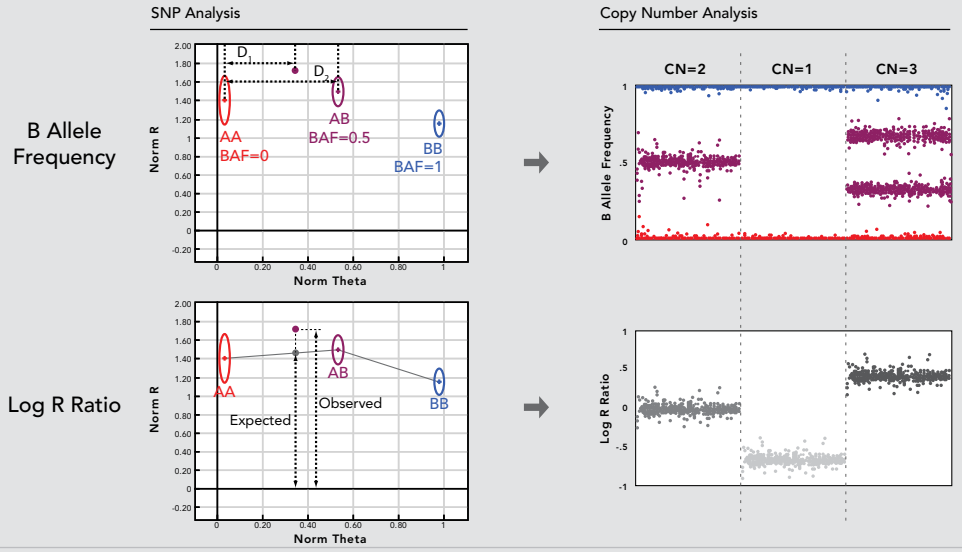
HOW DATA ARE GENERATED

For each SNP marker, the Infinium 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 BeadStudio 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*). BeadStudio 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 ($\log_2(R_{observed}/R_{expected})$), where $R_{expected}$ is interpolated from the observed allelic ratio with respect to the canonical genotype clusters)[2,3].

Standard Cluster File

As described, the standard canonical cluster positions used to compare against experimental data are essential to the computation of both log R ratio and B allele frequency. Thus, the use of an appropriate cluster file is

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 θvalue 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 (D_1/D_2 = .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 exhibit 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.

essential to 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. It is of note that Infinium HD BeadChip cluster position files are generated after excluding X chromosome SNPs[4]. 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[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 (four-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 likely 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 ratio values 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.

In general, log R ratio is used to diagnose physical aberrations, and B allele frequency monitors genetic aberrations. Copy number differences are readily apparent in plots of log R ratio as deflections in the y-dimension (*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.

FIGURE 4: TYPICAL WORKFLOWS FOR ILLUMINA CNV OR CYTOGENETIC ANALYSIS

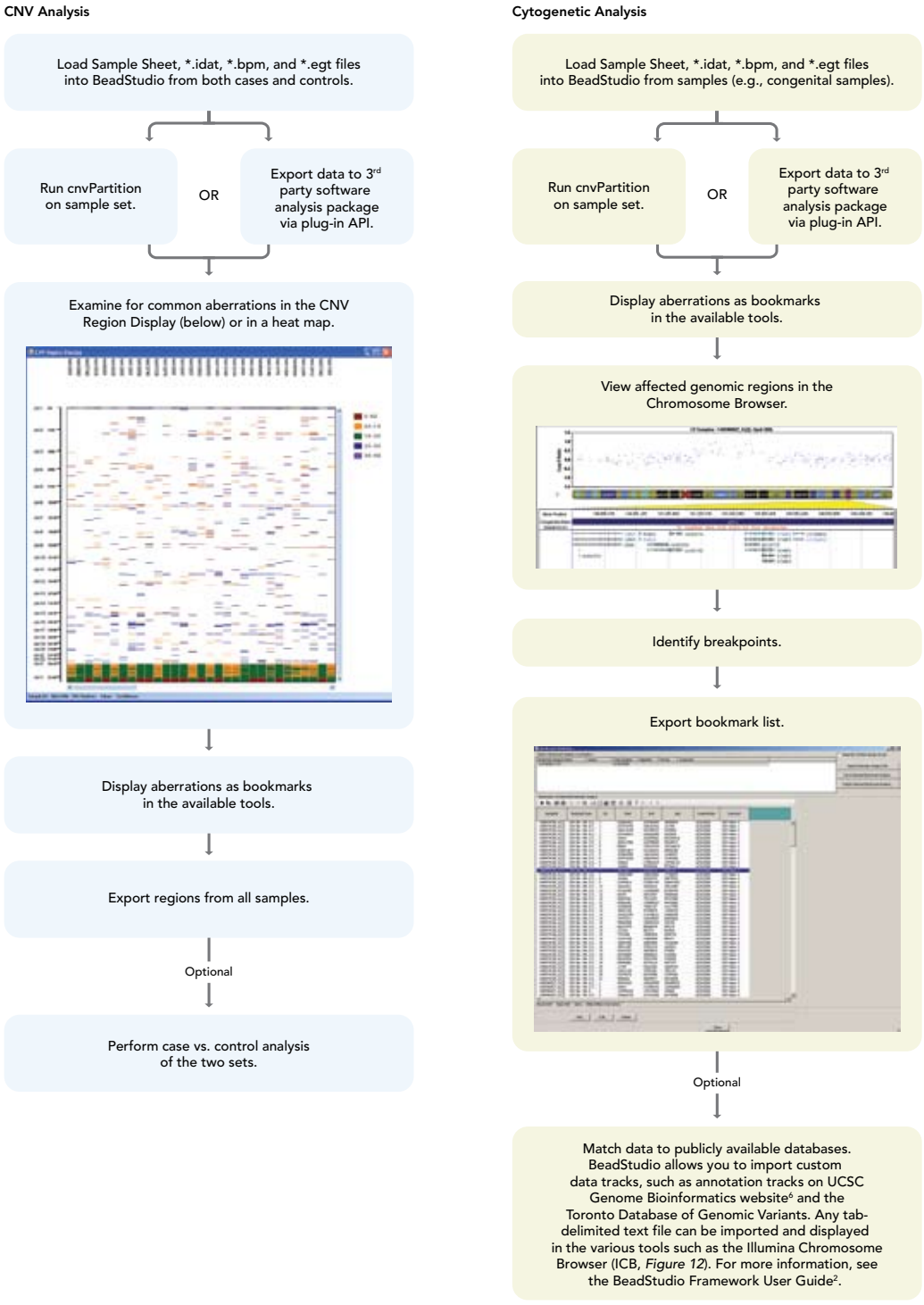


TABLE 2: BEADSTUDIO PLUG-INS FOR STRUCTURAL VARIATION ANALYSIS

ALGORITHM	SOURCE	FEATURES
cnvPartition	• Illumina* (algorithm & plug-in)	• Fast to analyze and report (~1 min / sample) • Detects CNV regions and estimates CNV values • Low CNV detection limit (< 100kb) • Computes confidence value for each locus
QuantiSNP	• Oxford University (algorithm) • Illumina* (plug-in)	• High data quality and accuracy • Finds small CNV regions (<100kB) • Provides confidence score for each locus • Optimized for Illumina BeadChips • Slow (~60min/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
dChip	• Harvard (algorithm) • Illumina* (report plug-in)	• Popular LOH analysis • Community supported
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 CGH	• 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 • User-friendly and 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 CN detection
Exemplar for CN	• Sapio (algorithm & plug-in)*	Not tested by Illumina
ArrayAssist	• Stratagene (algorithm & plug-in)*	Not tested by Illumina

*Download from www.illumina.com/illuminaConnect

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 2–3 heterozygous SNPs in a human genome with an average heterozygosity of 30%–40%. For example, the Human1M-Duo, which provides nearly 1.2 million markers at a median spacing of 1.5kb, has an effective resolution of ~7.5kb (*Table 1*).

ANALYSIS SOFTWARE

BeadStudio

BeadStudio is an integrated suite of software modules for the visualization and analysis of Illumina microarray and sequencing data. Analysis of Infinium BeadChip data is performed using the BeadStudio Genotyping (GT) Module, and the display tools of the BeadStudio framework. These

TABLE 1: EFFECTIVE RESOLUTION OF INFINIUM HD BEADCHIPS

	HUMANCNV370-QUAD	HUMAN610-QUAD	HUMAN1M-DUO
Number of Markers	373,397	620,901	1,199,187
Mean Spacing (kb)	7.8	4.7	2.4
Median Spacing (kb)	4.9	2.7	1.5
90th %ile Spacing (kb)	17.2	11.0	6.0
Effective Resolution (kb)	**24.5**	**13.5**	**7.5**

tools include automated algorithms that can scan for and characterize aberrations, the integrated Illumina Genome Viewer (IGV) that 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.

cnvPartition

Illumina's recommended analysis tool for CNV detection and characterization is cnvPartition, a plug-in algorithm for use with the CNV Analysis workbench in the BeadStudio GT Module. The cnvPartition algorithm is based on a recursive partition method that is described in the *DNA Copy Number Analysis Algorithms* technical note[7]. cnvPartition has been optimized for speed and accuracy to use log R intensity and B allele frequency for identification of chromosomal aberrations. During analysis, it estimates copy number values and calculates per-region confidence scores. CNV regions can then be converted into bookmarks in the Illumina Genome Viewer for a whole-genome graphical display.

The confidence score generated by cnvPartition is defined as the sum of all logged likelihoods for the assigned copy number for markers in the region, minus the sum of all logged likelihoods of copy number equal to two (normal) for markers in the region. Confidence scores provide a means to rank regions relative to their (dis)similarity to normal (copy number = 2), segments. Higher values represent higher confidence in the copy number designation of an aberration. Instructions for using cnvPartition and parameter definitions are contained in a document that can be downloaded from the plug-ins section of the BeadStudio Portal.

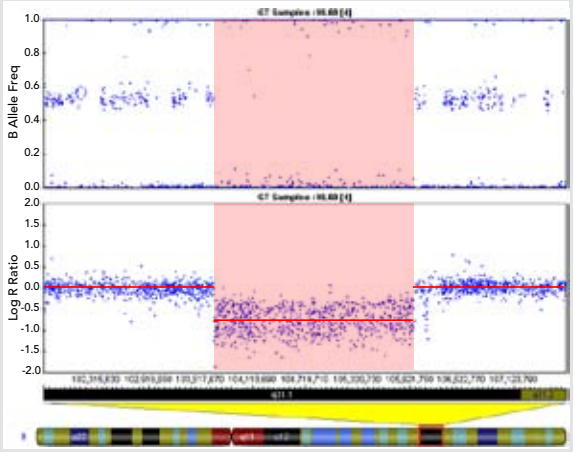
Third-Party BeadStudio Plug-ins

BeadStudio provides an open API for integrating third-party applications for downstream data analysis. In addition to cnvPartition, there are third-party algorithms that can be used for CNV and cytogenetics analysis, generally in conjunction with a BeadStudio plug-in. Some of these algorithms and customer report plug-ins to third-party platforms are described in Table 2.

EXAMPLES OF STRUCTURAL ABERRATIONS

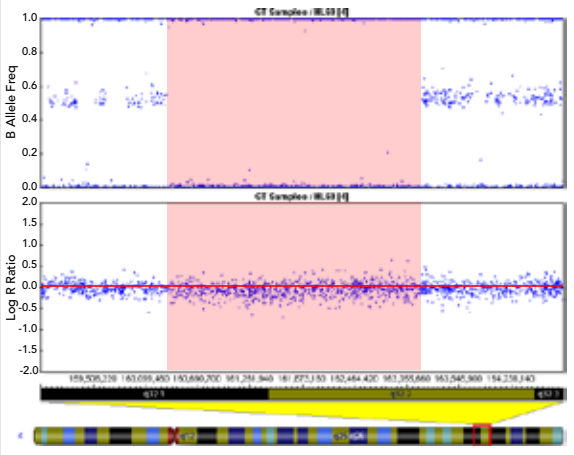
The following figures (*Figures* 5–13) are examples of typical Log R Ratio and B Allele Frequency plots that indicate the presence of various structural variants. For clarity, aberrant regions are annotated with shading and the average log R ratio over a region is indicated with a red line.

FIGURE 5: HEMIZYGOUS DELETION



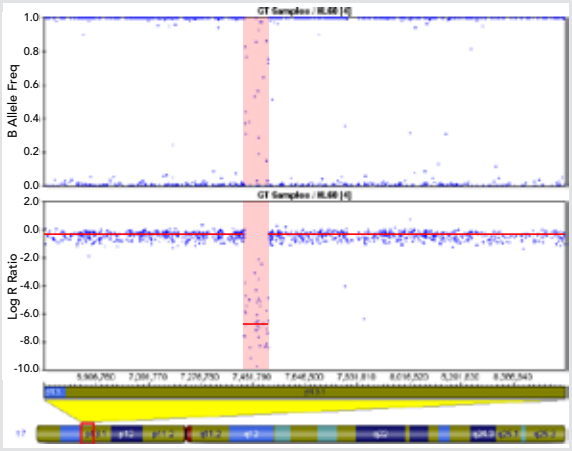
A hemizygous deletion (loss of one copy), shown in the ICB, is depicted as a loss of heterozygotes in the B Allele Freq plot (top) and a loss of signal intensity in the Log R Ratio plot (bottom). In the region of the deletion (shaded), the log R ratio is log2 of 1/2, or -1.

FIGURE 6: COPY-NEUTRAL LOH



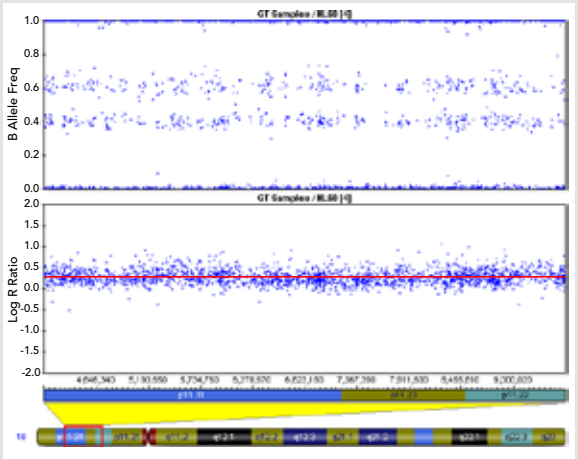
A region of copy-neutral LOH (shaded) is depicted by a loss of heterozygotes in the B allele frequency data but no change in the log R ratio (physical copy number).

FIGURE 7: HOMOZYGOUS DELETION



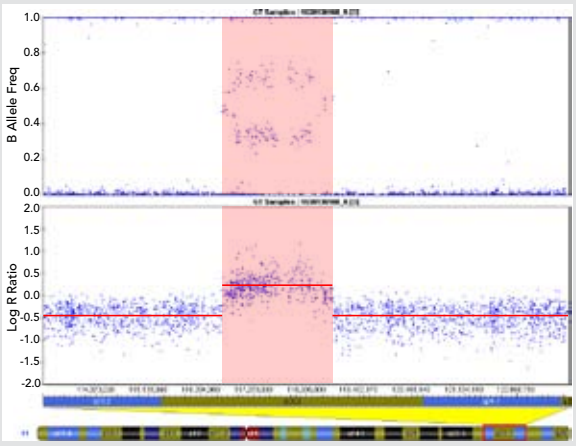
A region of homozygous deletion is where both copies of the chromosome have been lost (shaded). In this case, there are no SNPs present, so the genotyping data (B Allele Freq plot) appears like a "waterfall" as a result of noise in the absence of signal. The log R ratio in this region is the log2 of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

FIGURE 8: DUPLICATION



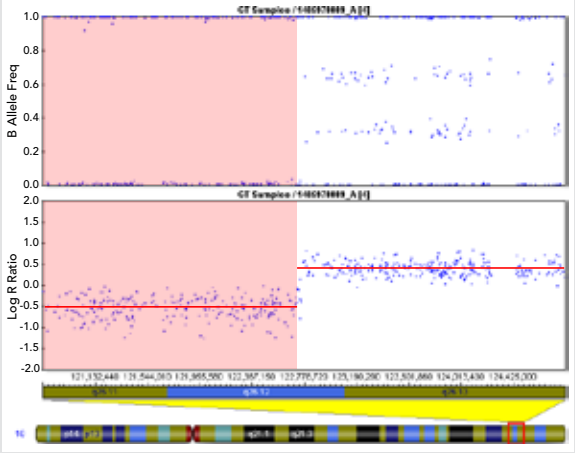
A duplicated region results in three total copies. This duplication is depicted by the B Allele Freq plot splitting into two new populations of data points representing the allelic ratios 1:2 and 2:1 (genotypes ABB and AAB). The duplication is also depicted by an increase in the log R ratio to ~0.4 (log2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

FIGURE 9: DUPLICATION NESTED WITHIN TWO FLANKING DELETIONS



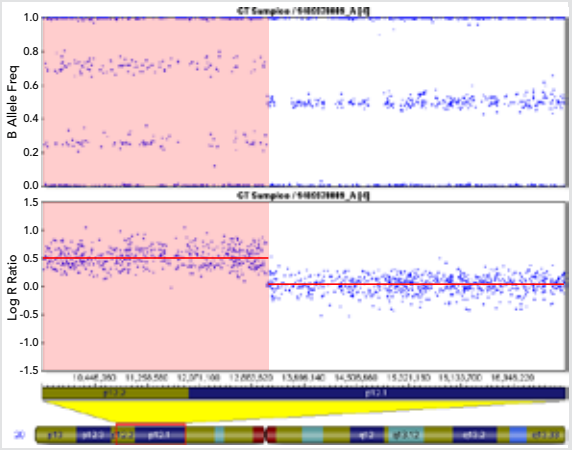
Regions of deletion (not shaded) are depicted by loss of signal intensity in the Log R Ratio plot to -0.5. An overlapping duplication (shaded) is depicted in the middle of the window by an increase in the Log R Ratio plot.

FIGURE 10: COMBINATION OF MULTIPLE ABERRATIONS IN A TUMOR SAMPLE



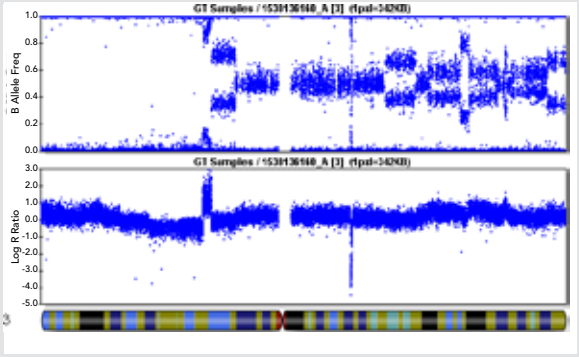
A hemizygous deletion (shaded) is depicted by the loss of heterozygotes in the B Allele Freq plot and a loss of intensities in the Log R Ratio plot. There is also a duplication (not shaded) indicated by the two clusters of data points in the B Allele Freq plot and an increase in the Log R Ratio plot.

FIGURE 11: DUPLICATION IN A TUMOR SAMPLE



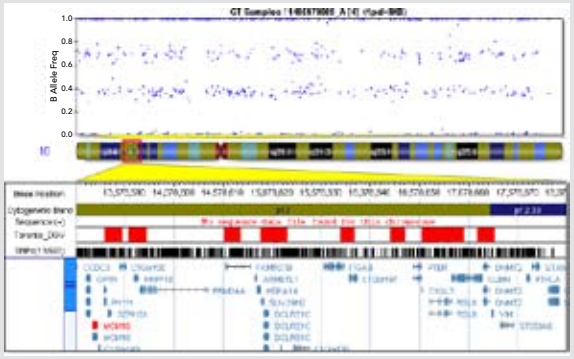
A duplication (shaded) is depicted by a splitting of heterozygotes in the B Allele Freq plot and an increase in intensities in the log R ratio to ~0.5.

FIGURE 12: DATA COMPLEXITY OF A TUMOR SAMPLE



Shown is 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). BeadStudio 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.

REFERENCES FOR FURTHER INFORMATION

For examples of successful CNV and cytogenetic analyses using Illumina DNA Analysis BeadChips, the following list of citations provides a good starting point. A more complete set of literature references is available at www.illumina.com/publications.

CNV Studies

- Blauw HM, Veldink JH, van Es MA, van Vught PW, Saris CG, et al. (2008) Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. Lancet Neurol 7(4): 319-326.
- Jones A, Mitter R, Springall R, Graham T, Winter E, et al. (2008) A comprehensive genetic profile of phyllodes tumours of the breast detects important mutations, intra-tumoral genetic heterogeneity and new genetic changes on recurrence. J Pathol 214(5): 533-544.
- Assie G, Laframboise T, Platzer P, Bertherat J, Stratakis CA, et al. (2008) SNP Arrays in Heterogeneous Tissue: Highly Accurate Collection of Both Germline and Somatic Genetic Information from Unpaired Single Tumor Samples. Am J Hum Genet 82(4): 903-15.
- Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, et al. (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82(3): 763-771.
- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181): 998-1003.
- Matarin M, Simon-Sanchez J, Fung HC, Scholz S, Gibbs JR, et al. (2008) Structural genomic variation in ischemic stroke. Neurogenetics 9(2):101-8.
- Ionita-Laza I, Perry GH, Raby BA, Klanderman B, Lee C, et al. (2008) On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. Genet Epidemiol 32(3):273-84.

Cytogenetic Studies

- Poot M, Eleveld MJ, van 't Slot R, van Genderen MM, Verrijn Stuart AA, et al. (2007) Proportional growth failure and oculocutaneous albinism in a girl with a 6.87 Mb deletion of region 15q26.2-->qter. European journal of medical genetics 50(6): 432-440.
- Brunetti-Pierri N, Grange D, Ou Z, Peiffer D, Peacock S, et al. (2007) Characterization of de novo microdeletions involving 17q11.2q12 identified through chromosomal comparative genomic hybridization. Clin Genet 72(5): 411-419.
- Lennon PA, Cooper ML, Peiffer DA, Gunderson KL, Patel A, et al. (2007) Deletion of 7q31.1 supports involvement of FOXP2 in language impairment: clinical report and review. Am J Med Genet A 143(8): 791-798.

Demonstration Data Sets

To help begin performing Illumina CNV data analysis in your own lab, demonstration BeadStudio Projects (*.bsc files) pre-loaded with HapMap samples are available to customers. Please contact Technical Support for information about getting access to 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. The large number of markers on Illumina BeadChips provide for 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) Expanding CNV Detection into the unSNPable Genome Technical Note. http://www.illumina.com/downloads/CNVdeCode_TechNote.pdf

(2) Beadstudio User Guides. Available from http://www.illumina.com/pagesnrn.ilmn?ID=275 or the BeadStudio 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/downloads/XChrClustering_TN.pdf

(5) Infinium Genotyping Data Analysis (PDF). http://www.illumina.com/downloads/GT-DataAnalysis_TechNote.pdf

(6) UCSC Genome Browser: Custom Annotation Tracks. http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html

(7) DNA Copy Number Analysis Algorithms Technical Note (PDF). http://www.illumina.com/downloads/CNValgorithms_TechNote.pdf

ADDITIONAL INFORMATION

For more information about Illumina DNA Analysis tools for CNV and cytogenetic analysis, please visit www.illumina.com/cyto or contact us at the address below.

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

FOR RESEARCH USE ONLY

Pub. No. 370-2008-014 Current as of 17 July 2008