Expanding CNV Detection into the unSNPable Genome

In collaboration with deCODE Genetics, Illumina has created a panel of markers not available on any other SNP array that target copy number polymorphisms. This technical note describes the design strategy and preliminary validation results using this new content.

INTRODUCTION

Current genotyping products have allowed researchers to begin discovering the extent of CNV impacts on phenotypes and disease states. To create improved tools geared toward CNV identification and analysis, Illumina has worked closely with deCODE Genetics to develop specific content for several new DNA Analysis BeadChips. This content was designed to target the least stable 6% of the genome, which is the most likely to contain medically relevant copy number variant (CNV) regions such as segmental duplications and unSNPable regions lacking SNPs, while providing uniform coverage of the genome. This document describes how this content was selected and shows results demonstrating the large number of common CNVs that occur at greater than 5% frequency in the four HapMap populations and the Icelandic population. These important novel CNV regions are not targeted by other SNP genotyping arrays, but can be measured reliably using one of several new Illumina DNA Analysis BeadChips.

A NEW TOOL FOR COPY NUMBER ANALYSIS

It is now known that copy number variation is ubiquitous in the human genome. To improve the already fruitful results of whole-genome disease association studies, it is thought to be essential to incorporate the analysis of genomic structural variation[1].

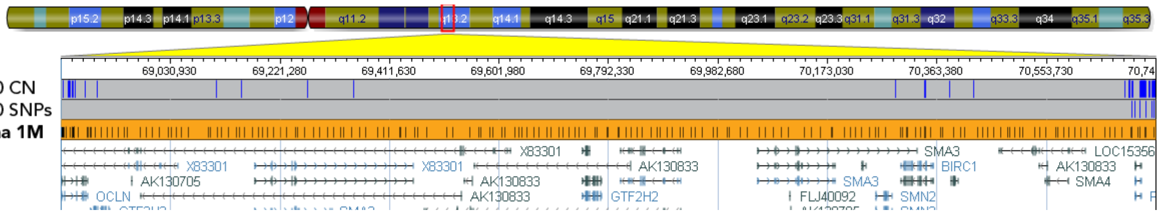
Although the influences of CNVs on disease are not entirely understood, there are numerous ways they can affect the human genome and disease susceptibility. A dosage-sensitive gene may be encompassed by a structural variant altering RNA levels. A hemizygous deletion in a gene region can unmask a recessive mutation on a homologous chromosome. Genes overlapping structural variants can be disrupted by deletion or copy-number variant breakpoints. There may also be positional effects whereby regulatory elements located within a CNV region can be amplified or deleted[1].

Although several studies have already shown the relevance of CNV regions to diseases such as breast cancer, SLE, and others[2-5], it has been hypothesized that a majority of the undiscovered CNVs lie within complex regions of the genome[6]. For that reason, we generated new content that expands current genotyping arrays to include coverage of the most unstable regions, comprising 6% of the genome. These new products are powerful tools for genome-wide SNP and CNV disease association studies.

CNVS IN THE DATABASE OF GENOMIC VARIANTS

The Toronto Database of Genomic Variants (DGV) is currently compiling CNV regions from various public studies. As of September 2007, there were several thousand regions deposited into this database that were

FIGURE 1: DENSE COVERAGE OF ILLUMINA/DECODE CNV CONTENT IN UNSNPABLE REGION



A region from Chromosome 5 with no coverage by either CN probes or SNPs from the Affymetrix6.0 (grey rows), is densely covered with the new CNV content from Illumina (orange row). Each vertical tick mark represents a single SNP or probe.

TABLE 1: NUMBER AND SIZE OF SEGMENTS IDENTIFIED AS LIKELY TO CONTAIN CNVS

SCREENING CRITERIA	SEGMENTS	TOTAL SIZE (MB)	UNIQUE SIZE (MB)
Megasatellites	167	6.9	1.7
Non-exact duplicons of 1,000 bases or greater	6,900	186	96
Exact duplicons of 100 bases and greater within 500,000 bases	15,800	63	10.4
UnSNPable genome			
15K gaps in HapMap SNPs	1,445	62.8	15.2
5–15K gaps and ≥ 2 SNPs with HW and inheritance problems	1,300	9.2	3.3
MHC region	1,500	3.4	2.5
Known rare and common CNVs (DGV)	1,769	77.5	60
Total	15,559 unique	263.3 Mb non-redundant	189.1 Mb

We identified over 15,000 non-redundant segments covering ~190Mb of total sequence. Approximately 14,000 of these regions were targeted with three or more markers at an average inter-marker spacing of 5kb.

largely identified with pre-existing microarray technologies[7]. Although some entries in the DGV represent commonly occurring CNVs, many of the entries occur at a low frequency (variant allele frequency < 5%)[8]. Additionally, most of the described regions are large since the methods used to discover them are less sensitive for detecting regions smaller than 50kb. Therefore, this database contains only a small percentage of the overall number of both common and rare CNVs[9]. Further work with new unbiased tools is needed to quantify how widespread and diverse CNVs truly are in the human genome.

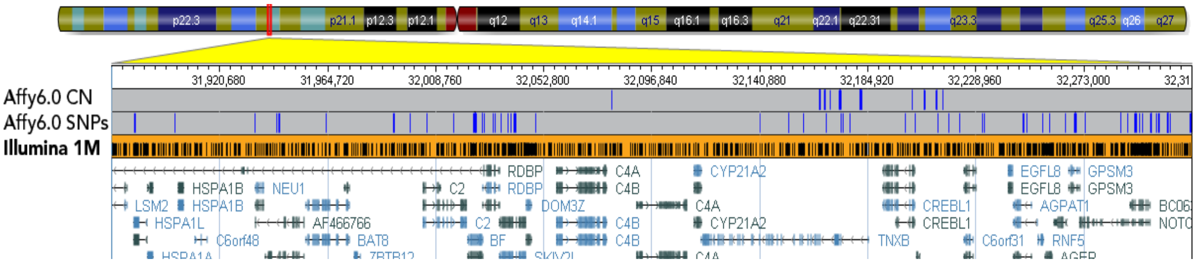
THE HUMAN GENOME ASSEMBLY DISGUISES CNV REGIONS

Both the Human Genome Assembly and the International HapMap Project have disguised regions of the genome most likely to represent copy number variants. For example, segmental duplications were collapsed as part of the Genome Project. However, such regions may be identified computationally. In addition, after completion of the HapMap project, we learned that there were regions of the genome that were unSNPable by existing methods. These regions either completely lacked SNPs or lacked SNPs that produced robust genotype results due to HWE and inheritance issues. Therefore, as we show here, these relatively complex and unstable regions yield a rich set of data for CNV studies.

GENERATING NEW CNV CONTENT

We started with dense coverage of DGV regions. For example, the Human1M BeadChip provides 206,000 SNPs in DGV regions with an average of 63 markers per region. Then, working in collaboration with deCODE Genetics, we specifically targeted additional regions that we expected to contain CNVs. Where possible, attempts were made to use SNPs to target a specific region for more robust CNV detection and measurement of LOH. The Infinium® Assay used for Illumina Beadchips is optimal for CNV detection because it provides a more representative amplification

FIGURE 2: UNIFORM COVERAGE OF ILLUMINA/DECODE CNV CONTENT IN UNSNPABLE MHC REGION



A region from within the MHC that has no coverage with either CN probes or SNPs from the Affymetrix6.0 (grey rows), but is densely covered with the new CNV content from Illumina on the HumanCNV370-Duo and Human1M BeadChips (orange row). Each vertical tick mark represents a single SNP or probe.

of the genome than methods based on PCR, which can incorporate an amplification bias into results. Importantly, the measurement at each locus is the average of up to 18 redundant bead types scattered randomly over the surface of a BeadChip, which increases the signal-to-noise ratio for precise CNV identification. Additionally, since SNPs and probes are developed using the same design strategy, both types of markers can be analyzed together.

This approach to identifying markers in unSNPable regions that are likely to interrogate CNVs has filled in areas left untouched by other SNP microarray design strategies (*Figures 1 and 2*). From the HumanCNV370-Duo to the Human1M BeadChips, Illumina has used a design strategy to ensure uniform distribution of markers across the genome, plus dense coverage of the DGV and other high-value regions likely to contain CNVs.

PRELIMINARY RESULTS

The Infinium Assay enables direct profiling of copy number changes by detecting changes in SNP genotypes and intensities of SNPs and probes. Figure 3 is a plot of intensity data from the two alleles of a selected SNP within a CNV region on Chromosome 11 (rs4132104) across a large sample set. The data points are distinctly grouped into clusters representing different copy number levels. The distinct separation of data point clusters facilitates monitoring single copy changes in unstable genomic regions.

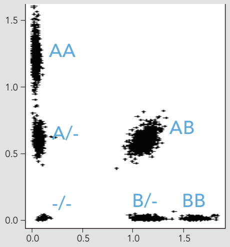
After creating a panel of markers that were expected to be a powerful tool for analyzing CNVs, we verified its performance with biological samples. A large set of samples (including HapMap and Icelandic samples) have been analyzed for the presence of CNVs using this new content on Illumina BeadChips. A group of CNV regions were arbitrarily selected for in-depth analysis to confirm findings using alternative methods, such as Southern blotting and TaqMan RT-PCR assays.

In the example shown in Figure 4, a novel yet common deletion was detected in RTDR1 (rhabdoid tumor deletion region gene 1). This CNV was validated using both Southern blotting (*Figure 4A*) and TaqMan RT-PCR assays (*Figure 4B*). TaqMan assays performed on several HapMap trios (parent-parent-child) show variable copy number levels in this region across a very large sample set. The Southern blot shows a fragment of DNA containing the SNP sequence from the original BeadChip. The expected fragment size is 4.4kb, but in some cases, a fragment of 1.8kb is seen in one or both chromosomes, indicating a loss of copy from two to one. Many patients in this study exhibited this copy number change.

NEW CONTENT FINDS COMMON CNVS

Preliminary results indicate that over 4,800 markers from the new Illumina CNV content show evidence for CNV at a frequency greater than 5%. Randomly chosen CNVs detected with this content have been confirmed with high success using alternative methods. These 4,800 markers represent over 2,100 segments of the nearly 10,000 autosomal segments selected in our original bioinformatics screen. These results demonstrate that the content avail-

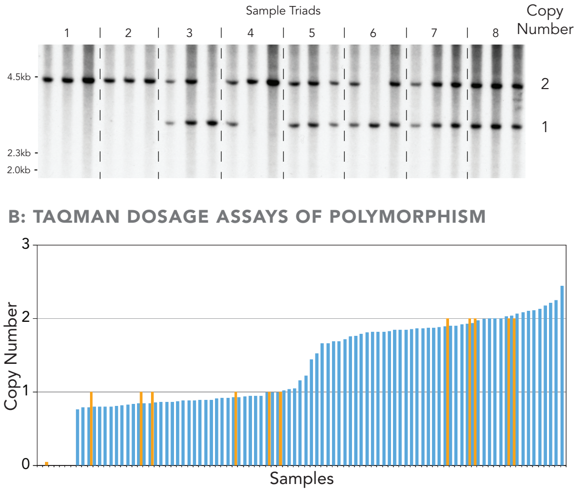
FIGURE 3: CNV DETECTED BY INFINIUM GENOTYPING



Genoplot of a representative SNP from Chromosome 11 located within PPFIBP2 (PTPRF interacting protein, binding protein 2) showing distinctly identifiable varying copy number levels, indicated by the genotype in blue.

FIGURE 4: VALIDATION OF CNV FOUND WITH ILLUMINA BEADCHIP

A: SOUTHERN BLOT OF TRIAD HAPMAP SAMPLES



A copy number variant detected in RTDR with the new CNV content that has been confirmed with both Southern blotting (A) and TaqMan assays in 95 samples (B). RT-PCR results of samples shown in Southern blot are indicated as orange bars.

TABLE 2: ILLUMINA BEADCHIPS TARGET COMMON POLYMORPHIC CNV REGIONS

CATEGORY	REGIONS SHOWING FREQUENCY > 5%	TOTAL AUTOSOMAL REGIONS	% POLYMORPHIC REGIONS
Megasatellites	56	109	51%
unSNPable segments >15kb	282	699	40%
unSNPable segments 5kb to 15kb	271	1,016	27%
Segments flanked by duplicons (100bp to 500bp)	894	3,442	26%
Segmental Duplications	705	4,205	17%
MHC	22	93	24%
Preliminary Total	**2,146**	**10,065**	**21%**

able on these Illumina BeadChips provides access to 10–20-fold more common CNVs than are defined in the SNPable genome or are available on other commercial SNP genotyping arrays.

The goal of this project was to provide researchers the most comprehensive CNV coverage of the human genome. To achieve this goal, Illumina and deCODE Genetics designed a set of markers (SNPs and probes) to target the "unSNPable genome," megasatellites, and segmental duplications. This new content has been added to several standard Illumina BeadChips which are all designed to minimize genome-wide gaps and provide dense coverage of DGV regions (*see inset*).

Intelligent marker design is essential for all disease association studies. The new CNV content designed by Illumina and deCODE Genetics contains highly polymorphic CNV regions missed by other SNP genotyping platforms. Illumina DNA Analysis products incorporate this new content for the most comprehensive genome-wide CNV detection and profiling.

ILLUMINA CNV ANALYSIS PRODUCTS

- **HumanCNV370-Duo DNA Analysis BeadChip:** Popular genome-wide tag SNP content plus CNV content
- **HumanExon510S-Duo DNA Analysis BeadChip:** Gene-centric content plus dense coverage of MHC, ADME, and CNV regions
- **Human1M DNA Analysis BeadChip:** Industry's only array with more than one million SNPs and CNV content. Uniformly spaced markers have the fewest large gaps of any array
- **HumanCNV-12 DNA Analysis BeadChip:** 12-sample BeadChip containing only the CNV content; can be combined with the HumanHap550-Duo BeadChip

REFERENCES

The majority of the data described in this document was presented at HGV 2007 (http://hgv2007.nci.nih.gov/home.cfm) and at ASHG 2007 (http://genetics.faseb.org/genetics/ashg/menu-annmeet.shtml) by both deCODE Genetics and Illumina.

(1) Feuk L, Marshall CR, Wintle RF, Scherer SW (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 15 Spec No 1: R57-66.

(2) Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L et al. (2007) FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39: 721-723.

(3) Frank B, Bermejo JL, Hemminki K, Sutter C, Wappenschmidt B et al. (2007) Copy number variant in the candidate tumor suppressor gene MTUS1 and familial breast cancer risk. Carcinogenesis 28: 1442-1445.

(4) McCarroll SA, Altshuler DM (2007) Copy-number variation and association studies of human disease. Nat Genet 39: S37-42.

(5) Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN et al. (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am J Hum Genet 80: 1037-1054.

(6) Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE et al. (2007) Challenges and standards in integrating surveys of structural variation. Nat Genet 39: S7-15.

(7) Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH et al. (2006) Global variation in copy number in the human genome. Nature 444: 444-454.

(8) Cooper GM, Nickerson DA, Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. Nat Genet 39: S22-29.

(9) Conrad DF, Hurles ME (2007) The population genetics of structural variation. Nat Genet 39: S30-36.

ADDITIONAL INFORMATION

Visit www.illumina.com or www.decode.com for more information. Contact us at the address below to learn more about CNV analysis using Illumina DNA Analysis BeadChips.

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

FOR RESEARCH USE ONLY

Pub. No. 370-2007-025 23Oct07