Technical Note

Associating Copy Number and SNP Variation with Human Disease

Abstract

The Genome-Wide Human SNP Array 6.0 is an affordable tool to examine the role of copy number variation in disease by combining high-powered SNP genotyping with highly accurate and sensitive detection of copy number state across the human genome. The combined density of SNPs and non-polymorphic probes on the SNP Array 6.0 provides high coverage of copy number variants—in particular those containing few SNPs that can be genotyped—and enables you to detect up to 10 times more copy number changes than competing platforms.

This technical note describes the design of the SNP Array 6.0 and explains how it can be used to examine the role of copy number variation in association studies of human diseases and traits.

Introduction

Genome-wide association studies seek to identify variation in the human genome that underlies a particular disease, drug response, diagnosis, or prognostic outcome. The cataloging of human variation and subsequent association analysis has traditionally focused on single nucleotide polymorphisms (SNPs). This assessment of common SNP variation in human disease has proven fruitful; more than 50 common variants have been found to be associated with disease such as type 2 diabetes, cardiac, and immunological disease[1,2].

However, recent work has demonstrated that other types of genomic variation—copy number polymorphisms (CNPs) and copy number variants (CNVs)—play a significant role in determining phenotype in common diseases, and are likely to be found at reasonably high frequencies in the population at large[3].

Recent data suggests that copy number variation could account for up to 4 megabases (Mb) of normal genetic differences, compared to roughly 2.5 Mb for SNP variation[4-8]. Many examples of diseases are known to be associated with copy number changes, including psoriasis[9], autism[10], lupus glomerulonephritis[11], as well as HIV infection and progression[12] (Table 1, adapted from Cohen, 2007[13]).

Until recently, however, cost-effective, genome-wide methods for analysis of CNVs in association analysis have been hampered by technical and practical limitations. In particular, SNPs in regions of copy number variation often fail Hardy-Weinberg and Mendelian inheritance checks, and are therefore not represented on most commercially available microarrays that principally interrogate SNPs[14].

Table 1: Examples of disease-associated CNVs (adapted from Cohen, 2007).

Disease	Gene	Phenotype
Psoriasis	ß-defensin	Red-scaling, elevated plaques
Autism	Segmental duplication	Neurobehavioral, includes social disability
HIV/AIDS	*CCL3L1*	Increased susceptibility to infection and disease
Lupus	*FCGR3B*	Increased susceptibility to kidney failure
Charcot-Marie-Tooth type 1A	*PMP22*	Demyelination, peripheral neuropathy
X-linked hypopituitarism	*SOX3*	In males, short stature, mild mental retardation
Autosomal dominant leukodystrophy	*LMNB1*	Demyelination, white brain matter abnormalities
Parkinson's	*SNCA*	Neuron degeneration, rigidity, tremor
Alzheimer's	*APP*	Amyloid beta precursor protein buildup
Altered drug metabolism	*CYP2D6*	Increased side effects, increased or decreased efficacy
Smith-Magenis syndrome	*RAI1*	Mental retardation
Pelizaeus-Merzbacher Smith-Magenis syndrome	*PLP1*	Demyelination, paralysis of legs, involuntary jerking of head
Spinal muscular atrophy	*SMN1*	Spinal deterioration, milder disease with later onset
Rett-like syndrome	*MECP2*	Mental retardation, spasticity, language/speech problems

Based on the need to expand the scope of association studies to include copy number variation, Affymetrix has developed microarrays designed specifically to interrogate CNVs alongside SNPs. In collaboration with the Broad Institute of Harvard and MIT, Affymetrix developed the Genome-Wide Human SNP Array 6.0, which allows you to analyze CNVs with 906,600 SNPs and more than 900,000 additional probes. This high-resolution array enables you to identify CNPs, CNVs, and/or SNPs that are statistically associated with a given phenotype. Large-scale whole-genome association studies can now be designed to assess true variation across the genome at an affordable cost.

Genotyping Console™ Software further enhanced the usefulness of the SNP Array 6.0 for association studies by integrating the algorithms for SNP genotyping (Birdseed v2), common CNP genotyping (Canary), as well as de novo/rare detection.

Performance and accuracy

The SNP Array 6.0 contains more than 900,000 nonpolymorphic probes and 906,600 SNP probes for copy number analysis. All probes on the array are designed to test sequences present on Nsp I or Sty I restriction enzyme fragments of approximately 200 to 1,100 base pairs (bp) that are amplified using the Genome-Wide Human SNP Nsp/Sty Assay Kit 5.0/6.0.

To obtain highly accurate and precise copy number intensity measurements, the linearity of response to copy number dosage was used in conjunction with probe spacing considerations to select a set of high-performing copy number probes that evenly span the genome and specifically target regions of known copy number variation. Targeted regions were derived from the Toronto Database of Genomic Variants (S. McCarroll, personal communication).

Figure 1: Dose response to copy number was assessed in a chromosome X dosage experiment consisting of five replicates each, of five cell lines with one to five copies of chromosome X. The plot presents a random selection of copy number probes on chromosome X, with each sub-plot corresponding to an individual copy number probe. For each sub-plot, the x axis is the log of the expected copy number ratio (using a copy number of 2 in the denominator) and the y axis is the scaled log ratio observed. The linearity and tightness of the relationship can be further improved (at the expense of genomic resolution) by averaging adjacent markers, but in this plot probes are assessed on an individual basis with no inter-probe averaging.

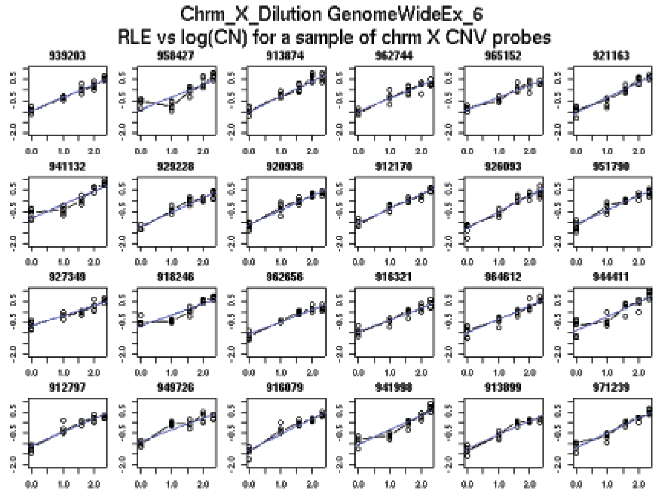
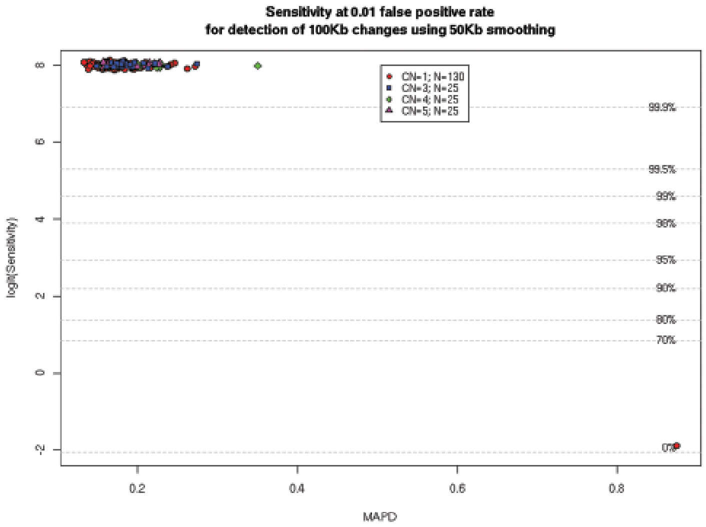


Figure 2: Sensitivity at 0.01 false positive rate for detection of 100 kb changes using 50 kb smoothing. One point was plotted for each chip; 205 total chips were analyzed. Of these, 130 were based on normal males with one copy of chromosome X. The rest consisted of 25 replicates of samples with three, four, and five copies of chromosome X. The replicates were processed at four different labs. The y axis presents sensitivity (proportion of the simulated variant correctly called) on a logit scale. The x axis presents MAPD, the copy number-specific QC metric used with the Genome-Wide SNP Array 6.0. Lower values of MAPD correspond to higher quality, and as the plot shows, the MAPD values are well corrected with actual copy number performance.



Probes for assessing copy number were empirically selected using a screen on a 13 million-probe array set designed against the fraction of the genome present on Nsp I fragments of a particular size. An Nsp/Sty mixture experiment was used to identify probes displaying strong linear response to change in copy number dosage. Final probe content of the SNP Array 6.0 was determined by the performance of the probes in these experiments, in addition to distribution in the genome. Efficacy of the copy number probes was explored in a chromosome X dosage experiment, using cell lines carrying from one to five copies (Figure 1).

In order to assess the sensitivity of detection for copy number changes with respect to the diploid state, single-sample detection of de novo variation was characterized in samples (N = 205) with randomly spliced 100 kilobase (kb) copy number variant regions. This analysis method assesses the physical marker distribution of the platform; if the selected region has few or no markers, sensitivity of detection will be lower. Hundreds of replicates were performed to relate sample quality control to typical sensitivity.

The SNP Array 6.0 provides extremely high sensitivity for the detection of real copy number changes at very stringent false positive rates. Figure 2 presents the sensitivity at a fixed false positive rate of 0.01. Sensitivity for the detection of a copy number of three (the most challenging case) is typically around 99 percent, and for the easier cases of one or greater than three copies, sensitivity is also around 99 percent. Moreover, the copy number-specific quality metric used on each chip is very predictive of functional performance.

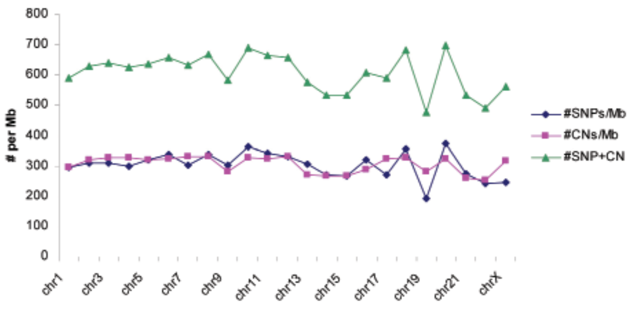
To further assess single-sample detection of de novo variation, 169 known verified regions in the commonly used standard fosmid (NA15510) and reference (NA10851) sample pairing[4] were analyzed. An HMM-CN algorithm was used to obtain a baseline across all 20 NA10851 replicates and then compared to the HMM-CN results from 20 replicates of NA15510. On average, 57 of the verified regions were detected.

An assessment of the number of copy number variants called on the same test sample (NA15510) using different experimental platforms and algorithms has been reported previously[3]. The SNP Array 6.0 detected the highest proportion of de novo variation than any microarray platform to date—approximately 10 times more than non-Affymetrix platforms. Table 2 summarizes these results.

Table 2: A summary of CNVs detected using various platforms in the test sample NA15510.

Platform	Analysis tool	CNVs detected
SNP Array 6.0	Genotyping Console™ Software	57
500K Array Set	GEMCA	24
	dCHIP	7
	CNAG	7
Illumina 650Y	QuantiSNP	9
	BeadStudio	5

Figure 3: Distribution of SNP probes and CNV probes is relatively even across the genome, resulting in a median marker distance of 680 bp.



High-resolution coverage of CNV

Most reported CNV locations in databases, such as the Database of Genomic Variants and the deCode database, actually correspond to the locations of CNV-*containing regions* as they were identified using low-resolution BAC clones, lower-resolution SNP microarray platforms or bioinformatic analysis. However, the SNP Array 6.0 has a median marker distance of 680 bp for CNV detection (Figure 3) and possesses fewer large gaps than other commercially available products (Figure 4).

The Broad Institute screened 270 HapMap samples using the SNP Array 6.0 to complete an initial high-resolution, genome-wide map of CNPs. These maps define more accurate CNP boundaries as well as allelic states for polymorphic CNPs (bi-allelic deletion loci have a diploid copy number of 0, 1, or 2, representing three possible genotypes; bi-allelic duplications have a diploid copy number of 2, 3, or 4).

Initial analysis of these maps indicates that the SNP Array 6.0 has more than twice the coverage of these better-defined CNPs than other commercial platforms (Table 3). Additionally, these maps confirm that the better-defined CNPs tend to exclude SNPs for the reasons previously discussed (i.e., they fail Hardy-Weinberg). For this reason, the high content of non-polymorphic copy number probes on the SNP Array 6.0 has proved to be of great importance to achieve extensive coverage of these refined common CNPs.

Figure 4: Intermarker distance ($\log_{10}$) across two commercially available platforms.

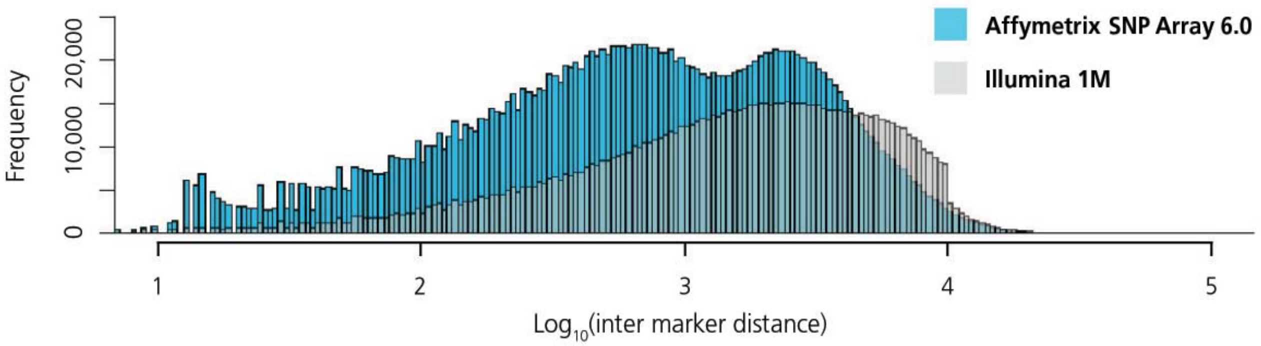


Figure 5a shows a polymorphic CNV on chromosome 1 that has been finely mapped by the SNP Array 6.0. The region is shown in Genotyping Console™ Software alongside a corresponding genomic variant loci from the Database of Genomic Variants.

Figure 5b shows this same region, as displayed in the Database of Genomic Variants browser. Unlike the SNP Array 6.0, lower-resolution microarray platforms that interrogate SNPs alone have low coverage of this CNV. Green represents the number of probes in non-overlapping 10 kb segments; pink represents the number of probes in nonoverlapping 10 kb segments on the Illumina 550 and 1M arrays.

Analysis tools

Current Affymetrix analysis tools, including Genotyping Console Software, enable several levels of CNV analysis for association studies. After completing array hybridization and data acquisition using GeneChip® Command Console® Software, you can easily open sample data (CEL) files in Genotyping Console Software, where you can perform copy number and LOH analysis using a reference model file. Available methods include smoothed and unsmoothed log ratios relative to a pre-defined reference, HMM estimates of copy number relative to a pre-defined reference, visualization of allelic differences, and estimation of regions of loss of heterozygosity (LOH).

Using the Segment Reporting Tool, you can generate a comprehensive summary of copy number aberrations across all samples. The segment summary table (Figure 6) details whether each copy number segment is a gain or loss, the copy number integer state, chromosomal location, cytoband location,

Table 3: SNP and non-polymorphic probe coverage of a high-resolution CNV map generated by the Broad Institute using the SNP Array 6.0 to screen 270 HapMap samples.

	SNP Array 6.0	ILMN 1M	ILMN 550K	ILMN 650Y
SNP	16,678	19,662	8,998	10,292
CN	30,360	841	0	0
SNP + CN	47,038	20,503	8,998	10,292

Figure 5a: A view in Genotyping Console Software of a copy number variant present across 34 of 270 HapMap samples. Fine-mapping with the SNP Array 6.0 across these samples demonstrates that only a small region of the annotated Toronto genomic variant loci (Toronto DGV) is the actual copy number variant.

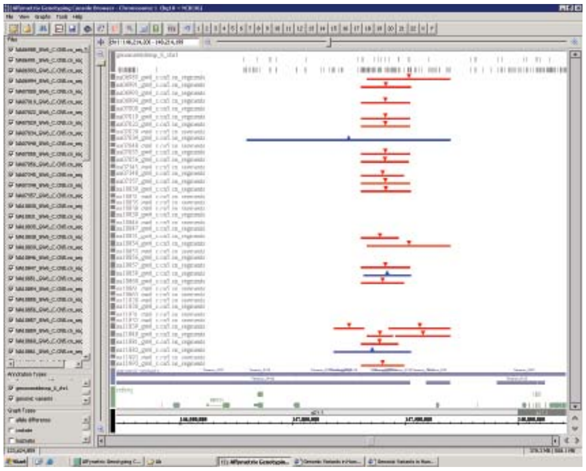
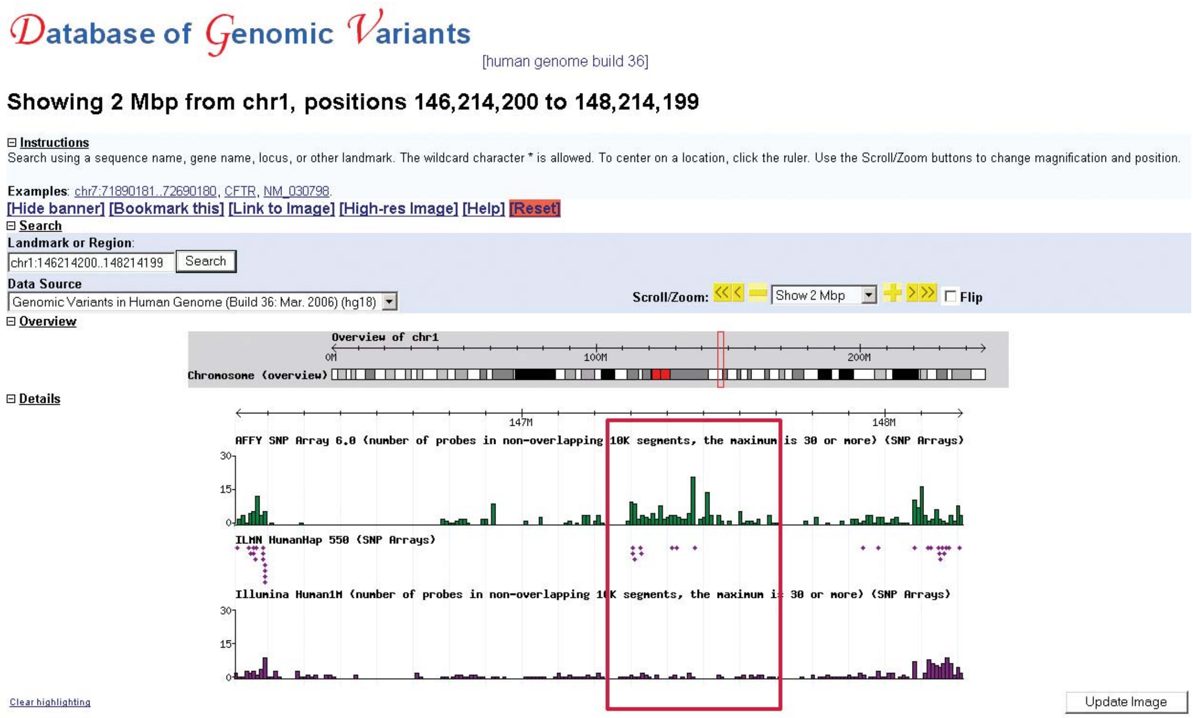


Figure 5b: Genotyping Console Software allows users to easily link out to the Database of Genomic Variants and other databases. This screenshot from the Database of Genomic Variants shows the same CNV (outlined in the red box). Unlike the SNP Array 6.0, microarray platforms that interrogate SNPs alone have low coverage of this CNV. (Green: number of probes in non-overlapping 10K segments; pink: number of probes in non-overlapping 10K segments on Illumina 550 and 1M).



Database of Genomic Variants

[human genome build 36]

Showing 2 Mbp from chr1, positions 146,214,200 to 148,214,199

⊟ Instructions

Search using a sequence name, gene name, locus, or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: chr7:71890181..72690180, CFTR, NM_030798.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help] [Reset]

⊟ Search

Landmark or Region: chr1:146214200..148214199 Search

Data Source: Genomic Variants in Human Genome (Build 36: Mar. 2006) (hg18)

Scroll/Zoom: << < − Show 2 Mbp + > >> □ Flip

⊟ Overview

⊟ Details

Clear highlighting

Update Image

Overview of chr1
0M
100M
200M
Chromosome (overview)
147M
148M
AFFY SNP Array 6.0 (number of probes in non-overlapping 10K segments, the maximum is 30 or more) (SNP Arrays)
30
15
0
ILMN HumanHap 550 (SNP Arrays)
Illumina Human1M (number of probes in non-overlapping 10K segments, the maximum is 30 or more) (SNP Arrays)
30
15
0

size of the copy number segment, number of markers in the segment, as well as overlap with known CNVs. The Segment Reporting Tool allows you to ask:

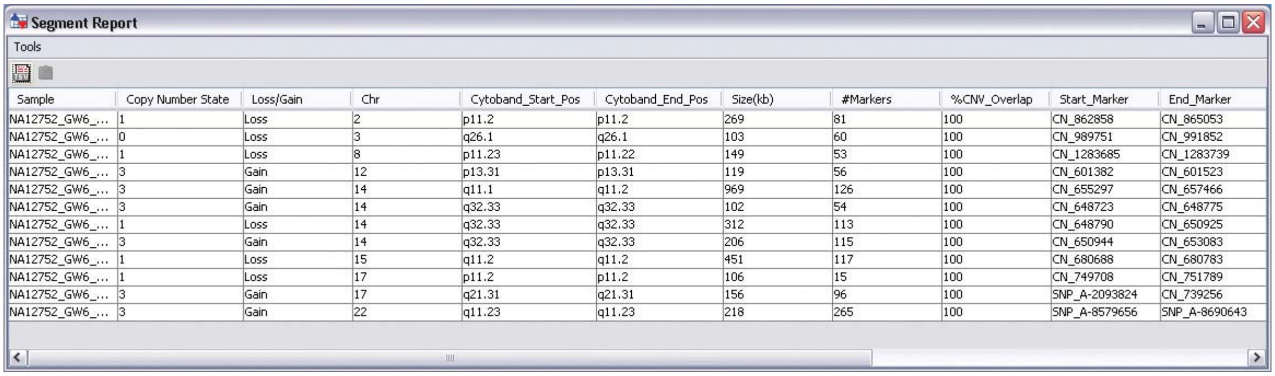
- Is the CNV a gain/loss?
- What is the precise integer level of genomic gain/loss?
- What are the breakpoints of the CNV?
- What is the precise size of the CNV?
- How reliable is the CNV call?
- Is the CNV novel or a common variant?
- How does the CNV compare across all of my samples?

CNVs of interest can be identified using the Segment Reporting Tool in combination with the Genotyping Console browser. The GTC browser includes a whole-genome karyoview, which provides a high-level view of copy number variation across the entire genome to correlate copy number changes with cytoband location. Select a chromosome to view CNVs of interest as they align with LOH results, as well as a number of annotation tracks including genes, CNVs in the Database of Genomic Variants, or any custom imported track (Figure 7). Multiple samples can be viewed simultaneously in the chromosome view to visually line up potential CNVs.

The CNV results computed in Genotyping Console™ Software can be used in subsequent association analysis using various statistical tools, including those incorporated into GeneChip-compatible™ software. Command-line tools for analyzing very large numbers of samples are also being developed for genotype and copy number analysis algorithms. These open-source implementations (upon which Genotyping Console Software is based) give you full visibility into the analysis details, enabling you to modify analysis settings and develop new applications for your unique studies.

In order to fully incorporate CNVs into whole-genome association analysis, a refined map of common copy number variation was developed[15]. This detailed map also has been used to better understand the LD structure between CNPs and SNPs as described by McCarroll, *et al.* The refined CNP map that has

Figure 6: Example of a segment report for a single sample in Genotyping Console™ Software. The segment report details whether a copy number change is a gain or loss, the copy number integer state, location, size, number of markers in the region, as well as overlap with known CNVs. The segment report and copy number predictions can be exported in CSV format. This data can be used to compute the frequency of the identified CNV across all samples, enabling subsequent association analysis using various statistical tools, including those incorporated into GeneChip-compatible™ software.

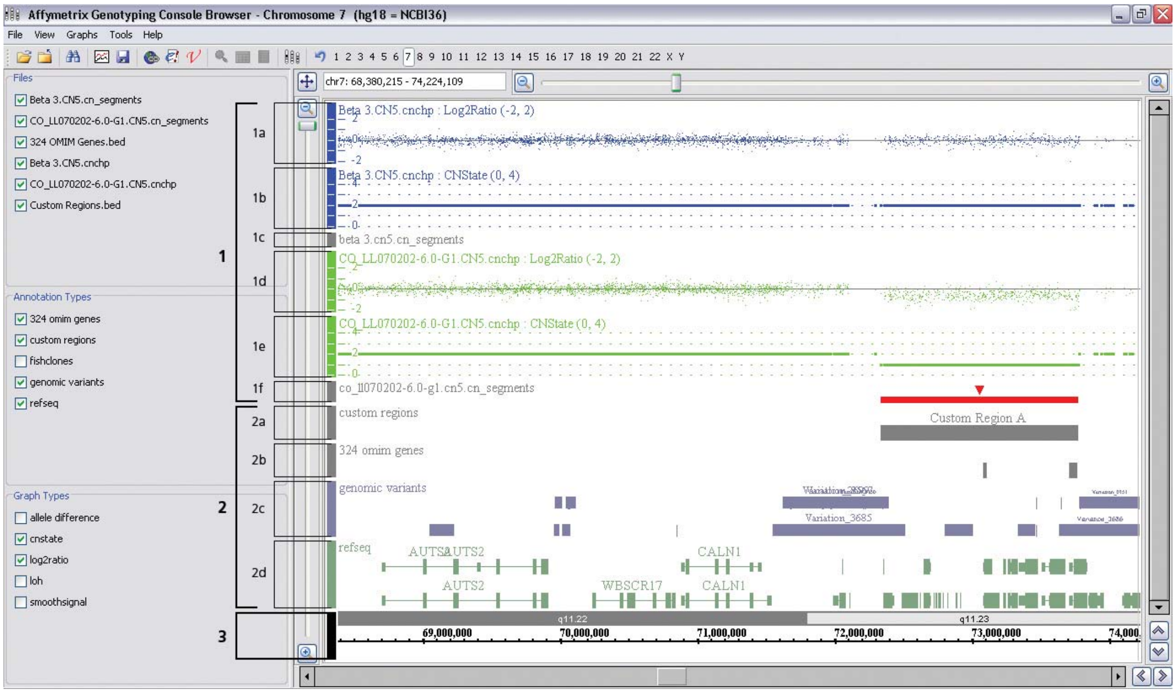


Segment Report

Tools

Sample	Copy Number State	Loss/Gain	Chr	Cytoband_Start_Pos	Cytoband_End_Pos	Size(kb)	#Markers	%CNV_Overlap	Start_Marker	End_Marker
NA12752_GW6_...	1	Loss	2	p11.2	p11.2	269	81	100	CN_862858	CN_865053
NA12752_GW6_...	0	Loss	3	q26.1	q26.1	103	60	100	CN_989751	CN_991852
NA12752_GW6_...	1	Loss	8	p11.23	p11.22	149	53	100	CN_1283685	CN_1283739
NA12752_GW6_...	3	Gain	12	p13.31	p13.31	119	56	100	CN_601382	CN_601523
NA12752_GW6_...	3	Gain	14	q11.1	q11.2	969	126	100	CN_655297	CN_657466
NA12752_GW6_...	3	Gain	14	q32.33	q32.33	102	54	100	CN_648723	CN_648775
NA12752_GW6_...	1	Loss	14	q32.33	q32.33	312	113	100	CN_648790	CN_650925
NA12752_GW6_...	3	Gain	14	q32.33	q32.33	206	115	100	CN_650944	CN_653083
NA12752_GW6_...	1	Loss	15	q11.2	q11.2	451	117	100	CN_680688	CN_680783
NA12752_GW6_...	1	Loss	17	p11.2	p11.2	106	15	100	CN_749708	CN_751789
NA12752_GW6_...	3	Gain	17	q21.31	q21.31	156	96	100	SNP_A-2093824	CN_739256
NA12752_GW6_...	3	Gain	22	q11.23	q11.23	218	265	100	SNP_A-8579656	SNP_A-8690643

Figure 7: Log$_2$ ratios, copy number states, and copy number segment results can be viewed simultaneously with multiple annotation tracks in the chromosome view of the Genotyping Console browser. Tracks can be removed, added, and rearranged, and custom tracks can also be added into the browser for aligning sample data with known annotation data. An example of a custom view displaying results in chromosome 7 from one sample without and one sample with a hemizygous deletion within the ELN gene, which is associated with Williams Syndrome. This view has been customized into three sections: results from the two samples (1), annotations (2), and the cytoband and linear genomic coordinates (3). The first section shows the log$_2$ ratios (tracks 1a, 1d), copy number states (tracks 1b, 1e), and copy number segment (tracks 1c, 1f) results from each of the two samples. Within the annotation section, tracks 2a and 2b are two custom regions, with one specifying the full region of interest (2a) and the other being the OMIM list (2b); track 2c indicates the annotated copy number variants published in the Toronto Database of Genomic Variants; and track 2d represents RefSeq genes.



Affymetrix Genotyping Console Browser - Chromosome 7 (hg18 = NCBI36)

chr7: 68,380,215 - 74,224,109

Files
☑ Beta 3.CN5.cn_segments
☑ CO_LL070202-6.0-G1.CN5.cn_segments
☑ 324 OMIM Genes.bed
☑ Beta 3.CN5.cnchp
☑ CO_LL070202-6.0-G1.CN5.cnchp
☑ Custom Regions.bed

Annotation Types
☑ 324 omim genes
☑ custom regions
□ fishclones
☑ genomic variants
☑ refseq

Graph Types
□ allele difference
☑ cnstate
☑ log2ratio
□ loh
□ smoothsignal

been generated using the SNP Array 6.0 has already led to the discovery of two novel associations with CNPs[16,17]. These findings highlight the importance of CNPs in understanding the full picture of genetic contribution to disease risk and variability.

The Broad Institute and Affymetrix also developed an algorithm that enables direct genotyping of CNPs. The Canary algorithm is based on a pre-defined set of more than 1,000 CNPs and provides discrete copy number states or genotypes to each CNP region. The Canary algorithm is fully integrated into Genotyping Console Software, thereby enabling the SNP Array 6.0 to genotype a set of common CNPs. The results of a genome-wide association study for myocardial infarction using the SNP Array 6.0 to look at SNPs, CNPs, and rare CNVs has been published in *Nature Genetics*[18]. This publication provides additional information on data handling and quality control metrics for CNPs and rare CNVs using the SNP Array 6.0.

Conclusion

The Genome-Wide Human SNP Array 6.0 is a powerful new tool for studying variation—both copy number and SNPs—associated with human genetic disease. You can now combine genotypes for more than 906,000 SNPs with accurate intensity measurements across more than 1.8 million markers on more samples than ever before, providing the most genetic power to detect CNV breakpoints and disease associated variation.

References

1. Bowcock A. M. Genomics: guilt by association. *Nature* **447**:645-646 (2007).
2. Altshuler D. and Daly, M. Guilt beyond a reasonable doubt. *Nature Genetics* **39**:813-815 (2007).
3. McCarroll S. A. and Altshuler, D. M. Copy-number variation and association studies of human disease. *Nature Genetics* **39**(7 Suppl):S37-42 (2007).
4. Redon, *et al.* Global variation in copy number in the human genome. *Nature* **444**(7118):444-54 (2006).
5. Tuzun E., *et al.* Fine-scale structural variation of the human genome. *Nature Genetics* **37**:727–732 (2005).
6. Altshuler D., *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**:513-516 (2000).
7. Wang D. G., *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**:1077-1082 (1998).
8. Sachidanandam R., *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**:928-933 (2001).
9. Hollox E. J., *et al.* Psoriasis is associated with increased ß-defensin genomic copy number. *Nature Genetics* **40**:23-25 (2007).
10. Weiss L. A., *et al.* Association between Microdeletion and Microduplication at 16p11.2 and Autism. *NEJM* (published online January 9, 2008).
11. Aitman T. J., *et al.* Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature* **439**:851-855 (2006).
12. Gonzalez E., *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**:1434-1440 (2005).
13. Cohen J. Genomics: DNA Duplications and Deletions Help Determine Health. *Science* **317**:1315-1317 (2007).
14. Wang K., *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**:1665-1674 (2007).
15. McCarroll S. A., *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**(10):1166-74 (2008).
16. Willer C. J., *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics* **41**:25-34 (2008).
17. McCarroll S. A., *et al.* Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nature Genetics* **40**:1107-1112 (2008).
18. Myocardial Infarction Genetics Consortium [Kathiresan S., *et al.*]. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature Genetics* **41**:334-341 (2009).

Affymetrix, Inc.
3420 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-DNA-CHIP (1-888-362-2447)
Fax: 1-408-731-5380
sales@affymetrix.com
support@affymetrix.com

Affymetrix UK Ltd.
Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 0HH
United Kingdom
Tel: +44 (0) 1628 552550
Fax: +44 (0) 1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

Affymetrix Japan K.K.
Mita NN Bldg., 16 F
4-1-23 Shiba, Minato-ku,
Tokyo 108-0014 Japan
Tel: +81-(0)3-5730-8200
Fax: +81-(0)3-5730-8201
salesjapan@affymetrix.com
supportjapan@affymetrix.com

www.affymetrix.com Please visit our website for international distributor contact information.
For research use only. Not for use in diagnostic procedures.

P/N 702626 Rev. 2

PRINTED ON RECYCLED PAPER