

Technical Note

## Strategies to Identify and Genotype Copy-Number Variants in the Human Genome

Improving the power of association studies with the Human1M DNA Analysis BeadChip.

by Gregory M. Cooper, Department of Genome Sciences, University of Washington, Seattle, WA 98195-5065

Resequencing of diverse humans has led to the discovery of millions of single-nucleotide polymorphisms (SNPs) in human populations, and large-scale genotyping projects have led to the characterization of these variants in terms of their frequency distributions and correlations with one another<sup>1</sup>. These efforts, coupled with advances in highly parallel and accurate genotyping technologies<sup>2</sup>, have facilitated considerable progress towards identifying the genetic basis for common diseases and complex traits in humans. Indeed, convincing genetic associations have now been reported for a wide variety of complex human traits, including diseases like diabetes and coronary artery disease<sup>3,4</sup>, drug metabolism and response<sup>5</sup>, and transcriptional regulation<sup>6,7</sup>, among many others.

Technological and practical concerns have dictated that the bulk of recent progress be explicitly centered upon the discovery and genotyping of SNPs. However, evidence is rapidly accumulating that genomic structural variants, including insertions, deletions, duplications, and inversions, are major contributors to the genetic diversity within human populations<sup>8,9</sup>. In particular, recent efforts have identified many common copy-number variants (CNVs, the subset of structural variants that results in a copy-number change) in human populations. In fact, while they occur less frequently than SNPs,

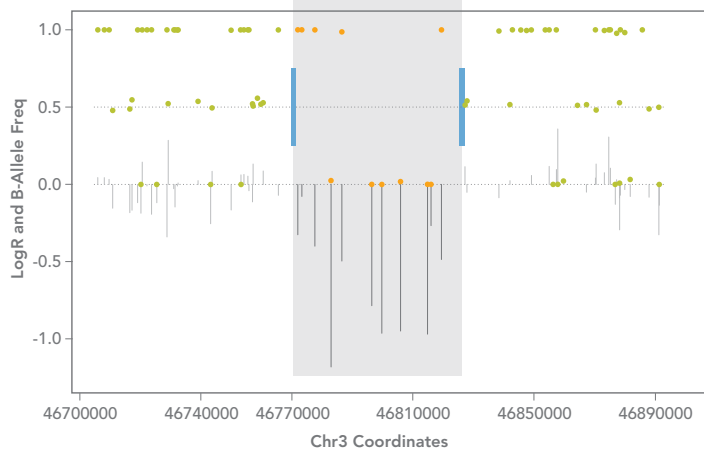
CNVs are likely to constitute a majority of the nucleotide differences between any two human beings since many of the events are quite large<sup>10</sup>. Furthermore, CNVs are likely to have major effects on human biology: they often result in gains or losses of whole genes or sets of genes and their associated regulatory elements, and can also generate hybrid or truncated gene products<sup>11,12</sup>. Accordingly, detailed studies on particular loci and their associated traits have identified a variety of human diseases that are at least partially attributable to CNVs, including genomic disorders resulting in severe early childhood diseases<sup>13,14</sup>, Parkinson's disease<sup>15</sup>, autism<sup>16</sup>, HIV susceptibility<sup>10</sup>, and other classical human genetic traits like red-green color blindness<sup>17</sup>. In fact, CNVs may also be causative variants for some of the genotype-phenotype correlations observed through SNP-based association studies, since in most cases the causal variants have only been localized to a genomic region rather than explicitly pinpointed.

As they are likely to contribute significantly to the genetic control of many human traits, an obvious goal is to systematically discover and precisely define the CNVs that are common in human populations, similar to the comprehensive analyses done on SNPs. While knowledge of an impressive number of CNVs has accumulated, a variety of technical challenges remain to be met before a comprehensive catalog of



Gregory Cooper is a post-doctoral fellow working at the University of Washington, Department of Genome Sciences in the labs of Dr. Evan E. Eichler and Dr. Deborah A. Nickerson. He received his Ph.D. in Genetics from Dr. Arend Sidow's lab at Stanford University. Dr. Cooper's current scientific interests focus on identifying genetic influences on complex human traits such as drug response and cardiovascular disease, with an emphasis on better understanding the role of copy-number variants in human biology.

**FIGURE 1: AN ESP-IDENTIFIED DELETION IS CLEARLY DETECTABLE ON THE ILLUMINA HUMAN1M BEADCHIP.**



The plot shows quantitative information obtained for the SNPs genotyped in this interval at the chromosome 3 coordinates indicated on the X-axis. The location of a sequence-resolved deletion in this sample is indicated in gray, and segmental duplications in the reference assembly are indicated in green. 'LogR' is shown as a vertical bar for each SNP, and corresponds to the total intensity observed for a given probe (global average is 0). 'B-allele Frequency', which indicates the proportion of the intensity attributed to the 'B' allele of the SNP, is indicated as a solid dot; 'AA' homozygotes have a 'B-allele Frequency' of 0, 'BB' are at 1, and 'AB' heterozygotes are near 0.5<sup>22</sup>. SNPs within the deletion are colored red (LogR) or blue (B-allele Frequency), while SNPs outside the interval are black. Note that, in contrast with the SNPs on the flanks, SNPs within the deletion are reduced in total intensity and also show a complete loss of heterozygosity, as expected for a hemizygous site. The presence of segmental duplications at the precise boundaries of this deletion is not likely to be coincidental, as this mutation was probably generated by a non-allelic homologous recombination event involving this pair of duplicated sequences.

“Direct CNV typing on SNP genotyping arrays would improve the power of future association studies, and also allow retrospective mining of previously generated SNP array data sets for CNV associations.”

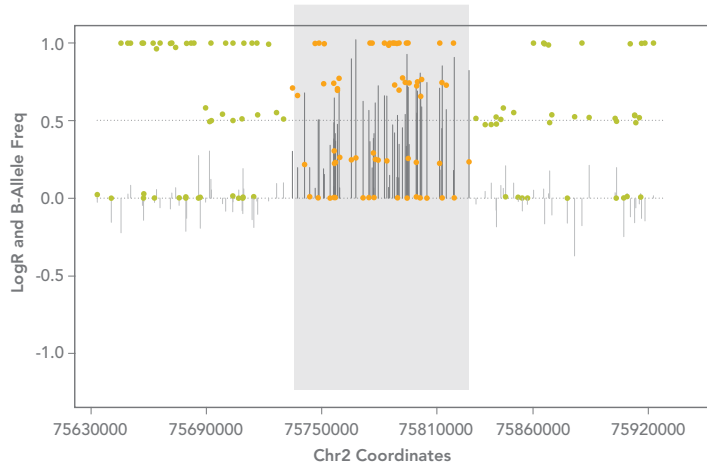
CNVs can be completed. In contrast with SNPs, CNVs are generally much larger than a typical sequencing read length and not amenable to discovery with a simple sequence and alignment approach.

Additionally, resulting from the kinds of mutational events that yield deletions and duplications of DNA (e.g. non-allelic homologous recombination<sup>13,14</sup>, Figure 1), CNVs often occur in and around regions of the genome that are enriched for blocks of duplicated, highly similar sequences<sup>18</sup>. This can confound genomic assembly of CNV-containing regions (CNVRs) from shotgun sequence data and inhibit or add noise to the discovery of CNVs through this approach.

Several alternatives to shotgun-sequence assembly and alignment exist for the detection of CNVs. The most popular and widely employed of these is array comparative genomic hybridization (CGH), wherein DNA from a 'sample' and 'reference' individual are differentially labeled and com-

petitively hybridized to an array containing genomic DNA. This genomic material can range in size from BACs to oligonucleotides, and can be tiled to either span the entire genome or densely target individual regions<sup>19</sup>. CGH experiments have proven to be an effective approach to identify CNVs spread across the genome within hundreds to thousands of individuals; in fact, recent experiments that include the HapMap samples have identified hundreds of distinct, common CNVs in the human genome<sup>10,13</sup>. However, in general, CGH experiments sacrifice either resolution in order to span the entire genome or genomic coverage in order to have high resolution. For example, the largest CNV studies were performed using CGH experiments with genome-wide BAC arrays; while these experiments are genome-wide, they suffer from poor resolution and collectively annotate ~20% of the human genome. It is likely that these annotations overestimate the true extent of copy-number variation by ~10-fold, and

**FIGURE 2: A DUPLICATION EVENT COMPUTATIONALLY INFERRED FROM THE HUMAN1M BEAD-CHIP THAT HAS BEEN VALIDATED WITH AN ESP LIBRARY RESOURCE FOR THIS INDIVIDUAL.**



Plot is similar to that seen in Figure 1. Note that, in contrast with the deletion event, the SNPs within the duplication exhibit elevated total intensity. Also, heterozygous SNPs in the duplication have B-allele Frequency values that deviate strongly from 0.5, since one of the alleles at each site is present multiple times (i.e. the true SNP genotypes are either 'AAB' or 'ABB').

thus it is critical to distinguish the experimental annotation of a CNVR from a precisely bounded CNV. Also, the results from CGH experiments can be difficult to validate since orthogonal platforms are often incapable of detecting all the identified variants, and systematic, targeted resequencing (a typical gold-standard for SNPs) is difficult or impossible for many of the events.

Another strategy to identify structural variants is clone-end sequence-pair mapping (ESP)<sup>20,21</sup>. In these experiments, a library of size-controlled clones is generated for a particular individual, and sequence is obtained from each end of these clones. These paired-ends are then computationally mapped to the reference human genome assembly. Regions of structural difference between the sample genome and the reference assembly manifest as *in silico* mapped clones with sizes that are implausibly large (implying a deletion in the sample relative to the reference), small (insertion relative to the reference), or have an inconsistent orientation (inversion relative to the reference<sup>20</sup>; also see Kidd et al. in preparation). The key benefits of the ESP approach include the ability to discover all classes of structural variation (including inversions,

which do not result in a copy-number change and are otherwise difficult to detect) and define them with comparatively high resolution. Note also that by retaining copies of the ESP clone library, the newly discovered sites of structural variation can be validated and defined at the nucleotide level through targeted resequencing. There are currently nine fosmid (~40 kb insert sizes) libraries that have been generated and analyzed, yielding a high-quality map of structural variation within the human genome. This includes many insertion and deletion polymorphisms that have been precisely defined at the nucleotide level (Kidd et al. in preparation). It should be noted that these experiments are expensive and not currently scalable as a genotyping technology, particularly in regards to the generation and maintenance of large-insert clone libraries. However, cheaper discovery projects and sequence validation of structurally variant clones may become available with the development of newer ultra-high-throughput sequencing technologies<sup>21</sup>.

Considering that large, genome-wide collections of CNVs have now been identified in many individuals with rough resolution and a few individuals with high resolution, there is hope that studies will soon be able

## REFERENCES

- (1) International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437: 1299–1320.
- (2) Gunderson KL, Steemers FJ, Lee G, Mendoza LG, and Chee, MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet* 37: 549–554.
- (3) The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
- (4) Saxena R, Voight BF, Lyssenko V, Burt NP, de Bakker PI, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316: 1331–1336.
- (5) Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 74: 106–120.
- (6) Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. *Nat Genet* 39: 1202–1207.
- (7) Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, et al. (2007) Gene-expression variation within and among human populations. *Am J Hum Genet* 80: 502–509.
- (8) Feuk L, Carson AR, and Scherer SW (2006) Structural variation in the human genome. *Nat Rev Genet* 7: 85–97.
- (9) Sharp AJ, Cheng Z, and Eichler EE (2006) Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 7: 407–442.
- (10) 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.
- (11) Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, et al. (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res* 65: 6071–6079.
- (12) Courseaux A, Richard F, Grosgeorge J, Ortolà C, Viale A, et al. (2003) Segmental duplications in euchromatic regions of human chromosome 5: a source of evolutionary instability and transcriptional innovation. *Genome Res* 13: 369–381.
- (13) Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, et al. (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am J Hum Genet* 79: 275–290.
- (14) Inoue K and Lupski JR (2002) Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet* 3: 199–242.

to genotype CNVs in the large population samples that have been assembled for SNP-based association studies<sup>3</sup>. ESP is impractical for more than dozens of samples, and CGH does not offer genome-wide coverage with high resolution. Therefore, neither is currently practical for large-scale genome-wide association studies, where low cost, high throughput, and high accuracy are necessary. As an alternative to these methods, SNP-based platforms have recently emerged as an appealing platform to genotype CNVs. Since quantitative intensity information is obtained for each of the genotyped SNPs and their respective alleles, inferences about the underlying copy-number of that site in the sampled genome can in principle be obtained. In fact, probes targeted to SNPs provide a distinct advantage over probes that do not capture polymorphic nucleotides; deletions manifest as regions of both reduced intensity and a loss of heterozygosity, while duplications are visible as regions of both elevated intensity and heterozygous allelic ratios that are not in a 1-to-1 balance<sup>10,22</sup>.

Direct CNV genotyping on SNP genotyping arrays would improve the power of future association studies, and also allow retrospective mining of previously generated SNP array data sets for CNV associations. It is with this optimism that many labs have set out to combine existing annotations of CNVs with whole-genome SNP array data and generate *de novo* CNV predictions using dense SNP genotype information. For example, in a collaboration between the Nickerson and Eichler labs at the University of Washington, we applied the fosmid ESP data resource collected on nine human samples to identify many common deletions that can be readily genotyped using the Illumina<sup>®</sup> Infinium<sup>®</sup> Human1M DNA Analysis BeadChip (Figure 1). We are also pursuing

the converse experiment, developing computational approaches to predict the existence of deletion and duplication events using SNP genotype data (Figure 2), and subsequently validating those predictions by targeted fosmid resequencing. Finally, we are beginning preliminary studies on the design of a custom genotyping chip, such as an Illumina's Infinium iSelect<sup>™</sup> Custom Genotyping BeadChip, to target validated, breakpoint-resolved CNVs for future large-scale CNV association study efforts.

While successes along these lines are clearly promising, much future work remains to be done. First, many 'known' CNVs still need to be validated and defined with higher precision; this is necessary to facilitate more robust CNV genotyping and future array designs. Second, even with current coverage levels, many identified CNVs lack a sufficient number of probes on any existing genome-wide array platform to permit effective genotyping. Finally, through both ESP and shotgun sequence projects on distinct human individuals<sup>23</sup>, many regions of human genomes have recently been discovered that are not present in the reference assembly. Until these intervals are fully sequenced and assembled, all current platforms that utilize a reference assembly (including all array-based experiments) will be unable to genotype these regions. Future array designs, both systematic (i.e. 'whole genome') and custom, should facilitate the analysis of these intervals and enable technology that can efficiently and rapidly type a more comprehensive spectrum of common human CNVs.

- (15) Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, et al. (2006) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum Mol Genet* 16: 1–14.
- (16) Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. (2004) Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528.
- (17) Deeb SS (2005) The molecular basis of variation in human color vision. *Clin Genet* 67: 369–377.
- (18) Cooper GM, Nickerson DA, and Eichler EE (2007) Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* 39: S22–S29.
- (19) Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, et al. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258: 818–821.
- (20) Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, et al. (2005) Fine-scale structural variation of the human genome. *Nat Genet* 37: 727–732.
- (21) Korb J, Urban AE, Affourtit JP, Godwin B, Grubert F, et al. (2007) Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science Epub*.
- (22) Peiffer DA, Le JM, Steemers FJ, Chang W, Jerniques T, et al. (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 16: 1136–1148.
- (23) Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The Diploid Genome Sequence of an Individual Human. *PLoS Biol* 5: e254.

#### ADDITIONAL INFORMATION

To learn more about Illumina's DNA Analysis BeadChips, visit our website at [www.illumina.com](http://www.illumina.com).

We are committed to providing you with the content you want as a member of the Illumina community. Please email us with comments and suggestions for topics at [icommunity@illumina.com](mailto:icommunity@illumina.com).

#### FOR RESEARCH USE ONLY

© 2007 Illumina, Inc. All rights reserved.

Illumina, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, and CSPro are registered trademarks or trademarks of Illumina. All other brands and names contained herein are the property of their respective owners. Pub. No. 370-2007-028 11Oct07

