Genome-Wide Mapping of Copy Number Variations and Loss of Heterozygosity Using the Infinium® Human1M BeadChip

Contributed by Ku Chee-Seng, Sim Xueling, and Chia Kee-Seng, Centre for Molecular Epidemiology and Department of Community, Occupational, and Family Medicine, Yong Loo Lin School of Medicine, National University of Singapore

INTRODUCTION

Genetic variations within the human genome can take many forms, including single-nucleotide polymorphisms (SNPs), copy number variations (CNVs), and copy-neutral loss of heterozygosity (LOH). SNPs involve the change in a single nucleotide, while CNVs and LOH encompass larger segments of DNA. In this application note, we focus on methods for accurately mapping these structural variations and their potential involvement in disease manifestations.

CNVs, defined as additions or deletions in the number of copies of a particular segment of DNA (larger than 1kb in length) when compared to a reference genome sequence, provide further insight into the complexity and diversity of genetic variations. Since the initial discovery of hundreds of CNVs in the human genome reported in 2004[1,2], many more have been found[3]. In 2006, the largest and most comprehensive mapping of CNVs on International HapMap samples was completed, identifying nearly 1,500 CNV regions covering ~360 Mb, or ~12% of the nucleotide sequence in the human genome[4]. The significance of this discovery expands beyond the presence of CNVs themselves, and into the impact copy number changes have on complex diseases, as well as their importance in human evolution[5,6]. In fact, evidence is now available that links CNVs with complex diseases such as autoimmune disorders, HIV infection, cancers, schizophrenia, and autism[7–9].

Less information is currently available about LOH effects; however, their potential impact on complex diseases is enormous. Copy-neutral LOH is a continuous stretch of DNA sequence without heterozygosity. Although the biomedical relevance of regions of homozygosity to human complex diseases remains largely unexplored, some schizophrenia studies have shown significant differences in homozygous regions between cases and controls[10].

With only a preliminary understanding of the roles CNVs and LOH play in complex disease development, it is imperative that we generate a comprehensive catalog of structural variations in the human genome. This approach may provide the opportunity to unravel novel disease loci. To date, there has been little research into CNV information in Asian populations. Therefore, we have begun exploring the extent of CNVs in several South-East Asian populations (Singaporean Chinese, Malay, and Indian) with the goal of constructing a genome-wide map reflecting CNVs and copy-neutral LOH within these populations. In our study, we demonstrate the advantages of using high-density SNP arrays for this purpose.

MATERIALS AND METHODS

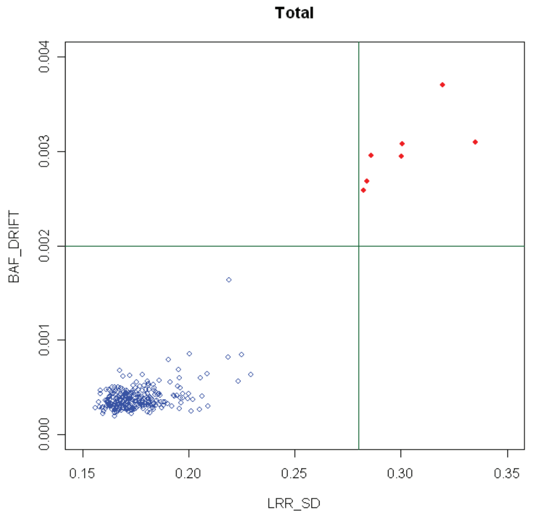
Samples

We genotyped 292 genomic DNA samples from unrelated healthy individuals without any known clinical disease. Genomic DNA samples were extracted from peripheral blood instead of lymphoblastoid cell lines, avoiding the introduction of artifacts (e.g., cell culture–induced chromosomal rearrangement) that may have incorrectly influenced our data. A stringent filtering criteria was applied to identify poor-quality samples. Samples with log R ratio standard deviation > 0.28 were removed from subsequent analyses to minimize the number of false-positive CNVs. Fewer than 3% of our samples failed these criteria (*Figure 1*).

Assay

Prepared DNA samples were run on the Infinium Human1M BeadChip from Illumina. We chose this platform for several reasons. The Human1M BeadChip offers significantly increased genomic coverage, resolu-

FIGURE 1: PENNCNV-GENERATED STANDARD DEVIATION OF LOG R RATIO AND B ALLELE FREQUENCY DRIFT VALUES FOR EACH SAMPLE AFTER CNV DETECTION



These are useful quality control parameters at the sample level. Large values indicate poor-quality samples. In our study, we set the thresholds of LRR_SR (log R ratio standard deviation) > 0.28 and BAF_Drift (B allele frequency drift) > 0.002 as the sample filtering criteria. Seven samples that failed the thresholds (red diamonds) were removed from further analyses.

tion, and probe uniformity across the human genome for unbiased, comprehensive detection of CNVs. Probes were specifically selected to cover genomic regions that potentially contain an excess number of CNVs, such as segmental duplications[11–12], for more accurate mapping of CNVs in these regions. In addition, the higher density array offers enhanced power for detecting smaller CNVs (< 50kb), which is especially critical for screening or discovery experiments where a large number of CNVs less than 10–50kb in length are yet to be uncovered[13]. Higher density arrays also increase the accuracy in mapping breakpoints of CNVs, providing a more accurate prediction or estimation of CNV size.

The Infinium Assay produced high-quality data in our study, achieving an average genotype call rate of > 99.5%. The simple workflow of the genotyping protocol involved only a few simple, straightforward steps, minimizing technical errors and ensuring a high genotyping success rate. This allowed our laboratory technician to complete the work within two weeks.

Analysis

The PennCNV algorithm, which employs a Hidden Markov Model, was used to detect both CNVs and copy-neutral LOH. CNV detection was mainly based on log R ratio (total signal intensity) and B allele frequency (allelic intensity ratio). In addition, this algorithm incorporates other sources of information, including population B allele frequency and distance between adjacent probes, to produce more reliable CNV calls. The PennCNV algorithm was developed for genome-wide detection of CNVs using Illumina SNP data[14], and is now available as a plug-in to Illumina's BeadStudio analysis software.

RESULTS

More Accurate CNV Mapping

We are excited to report that the majority of the CNVs detected were < 50kb in length. Figure 2 shows the distribution of deletions and duplications across chromosome 1 in our studied populations. These results contrast with preceding studies performed using lower resolution BAC and oligonucleotide array-based CGH or SNP arrays that are limited in their abilities to detect smaller sizes of CNVs. In fact, a recent study found that 88% of known CNV regions were smaller than the sizes reported in the Database of Genomic Variants and that more than a 50% reduction in size was reported for 76% of the CNVs[15]. This study was completed using a high-resolution, customized oligonucleotide CGH array with a 1kb resolution, emphasizing that use of lower-resolution arrays in most of the previous studies led to overestimation of CNV sizes.

Accurately estimating CNV sizes will have a significant impact when overlapping CNVs with known annotated genes to predict functional roles or mRNA expression studies, because gene function may be disrupted if part or all of the gene is deleted or duplicated. We believe that many of the genes that were found to overlap with CNVs were spurious findings resulting from overestimation of CNV sizes. With greater accuracy in estimating CNV breakpoints, the number of genes mapped to CNVs will likely be reduced.

DISCUSSION

Continuing Studies

In addition to unrelated individuals, we genotyped a number of families (father, mother, and one pair of monozygotic twins) using the Infinium platform. These samples were derived from the Singapore Twin Project.

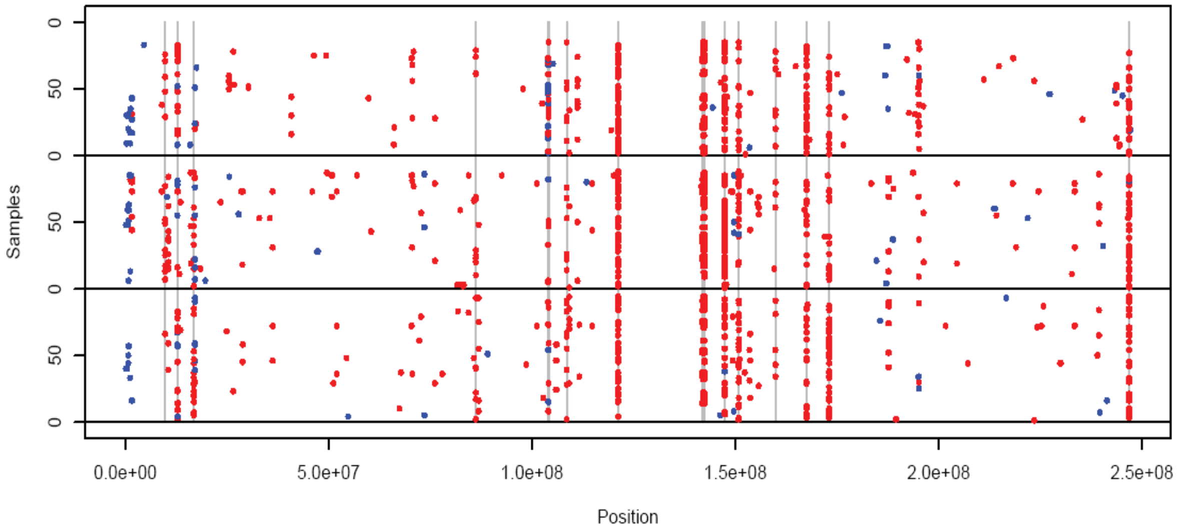
The goal of this study was to interrogate the *de novo* occurrence of CNV events compared to inherited germ line CNVs. CNVs detected in offspring but absent in their parents, or differences in the CNVs between a pair of monozygotic twins, are indications of putative *de novo* CNVs. Due to the noise inherent in any CNV detection method, it is important to validate these putative *de novo* CNVs using a second method. Comparing our CNVs data with other populations will provide further insight into the extent of similarities and differences in the CNV profile among populations of distinct ancestral backgrounds.

The Illumina iControlDB database made this comparison work possible. From iControlDB, we downloaded the data from 118 HapMap samples (49 Caucasians, 30 Asians, and 39 Africans) previously genotyped using the Human1M BeadChip. With this data set, we were able to detect CNVs using the same detection algorithm, applying the same quality control criteria to remove poor quality samples, filtering out likely false-positive CNVs, and analyzing CNV data in the same manner as our samples. This standardized analysis method allowed us to compare CNV profiles of our studied populations with those in the International HapMap populations.

Looking Ahead

Current data about the relative proportion of various types of structural variation within the human genome, and the genomic distribution and population frequencies of CNVs, are still rudimentary. More population-based studies are needed using various CNV detection methods in diverse populations. As the case for a link between CNVs and diseases grows stronger, it will be of paramount importance to build a near-complete, accurate map of CNVs and other structural variants representing populations worldwide. Currently, no single method is capable of detecting all the structural variations in a single experiment. With the rapid advances in sequencing platforms and technologies, it is now feasible to use sequencing paired-end mapping to characterize CNVs, inversions, insertions, translocations, and more complex chromosomal rearrangements such as genomic regions which are duplicated and inverted at the whole-genome scale. Unfortunately, this method is not yet sufficiently cost-effective for use in population-based studies that include hundreds of samples or genome-wide association studies (GWAS) of several thousand cases and controls. A comprehensive, accurate CNV database would enable more targeted and efficient platforms to genotype CNVs in thousands of samples. This database would be a valuable resource for future genetic studies of complex

FIGURE 2: DISTRIBUTION OF DELETIONS AND DUPLICATIONS IN CHROMOSOME 1



The X-axis is the physical chromosomal position and each line in the Y-axis represents one individual in all the three populations. Deletions are indicated with red points and duplications are indicated with blue points. Bottom panel: Chinese. Middle panel: Malay. Upper panel: Indian.

diseases and pharmacogenetic matters.

Over the last two years, GWAS have played a key role in uncovering novel genetic variations associated with complex human diseases. Future studies will need to explore CNV-structural variations, as well as gene-gene and gene-environment interactions. Adding environmental factors to experimental variables will require environmental data collection prior to disease onset. Large cohorts with repositories of biological samples will need to be developed. Several notable efforts, such as the UK Biobank and LifeGene Sweden, are moving in this direction. At the Centre for Molecular Epidemiology, we have set up the Singapore Consortium of Cohort Studies (SCCS) (http://www.med.nus.edu.sg/cof/cme.html) with the primary goal of understanding both genetic and environmental components in various complex diseases and quantitative traits such as metabolic and cardiovascular diseases. GWAS within the SCCS will be based on a nested case-control design and use next-generation genotyping and sequencing technologies to interrogate the genetic basis of complex diseases.

Currently, there are several ongoing studies at our Centre, including a GWAS of high-density lipoprotein cholesterol with well-annotated environmental exposure data. We are embarking on another GWAS on Type 2 diabetes where two thousand samples from our cohort will be genotyped using the Infinium HD Human610-Quad BeadChip. With the high-quality data from Illumina's BeadChips and our experience in CNVs and copy-neutral LOH detection, our future GWAS will not be restricted to SNP association analysis. Genome-wide CNV association analysis and whole-genome homozygosity mapping can be performed to discover other disease loci that may have eluded us when analysis was performed solely by SNP associations.

We are also undertaking a genetic diversity project—the Singapore Genome Variation Project—where 268 samples have been genotyped for ~1.4 million SNPs to characterize the extent of genetic variations in the Chinese, Malay, and Indian populations.

CONCLUSION

We are fortunate to live in an era where we may apply cutting-edge technologies to explore the human genome in unprecedented detail. We hope that research studies at our Centre will contribute even more to the current pool of knowledge of human genetic variations and improve our understanding of the environmental exposures and genetic basis underlying human complex diseases. The potential impact of genomics on medical sciences is tremendous, from identifying new molecular drug targets to developing new therapeutic interventions.

ACKNOWLEDGEMENTS

We would like to thank all of the staff at the Centre for Molecular Epidemiology and the Singapore Genome Variation Project team for their contributions.

REFERENCES

(1) 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.

(2) Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK et al. (2004) Detection of large-scale variation in the human genome. Nat. Genet. 36: 949-951.

(3) 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.

(4) 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.

(5) 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.

(6) Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nat. Genet. 39: 1256-1260.

(7) Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J et al. (2008) Psoriasis is associated with increased ß-defensin genomic copy number. Nat. Genet. 40: 23-25.

(8) Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R et al. (2008) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307: 1434-1440.

(9) Park J, Chen L, Ratnashinge L, Sellers TA, Tanner JP et al. (2006) Deletion polymorphism of UDP-glucuronosyltransferase 2B17 and risk of prostate cancer in African American and Caucasian men. Cancer Epidemiol. Biomarkers Prev. 15: 1473-1478.

(10) Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV et al. (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc. Natl. Acad. Sci. USA. 104: 19942-19947.

(11) Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA et al. (2005) Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. 77: 78-88.

(12) Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56-64.

(13) Estivill X and Armengol L (2007) Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet. 3: 1787-1799.

(14) Wang K, Li M, Hadley D, Liu R, Glessner J et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17: 1665-1674.

(15) Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L (2008). The fine-scale and complex architecture of human copy-number variation. Am. J. Hum. Genet. 82: 685-695.

ADDITIONAL INFORMATION

Visit www.illumina.com or contact us at the address below to learn more about Illumina DNA analysis products.

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. 370-2008-026 Current as of 15 December 2008