Technical Article

by Jeanely Hunt

A Landmark CNV Study of Human Genetic Variation in Common Diseases

The human genome consists of approximately three billion bases. How we look, behave, metabolize food, and fight infections is determined by the sequence of codes that lie within the genome. The human genome and sequence of genes vary from one person to another. These genetic variations contribute to the diversity we see among the human population. Extreme variations in the genome, however, such as deletions of large sequences or the presence of excessive copies of a gene or multiple genes, are responsible for contributing to diseases and the predisposition of genetic disorders.

Genetic disorders can be classified into two broad categories: those that are caused by mutant genes, such as Huntington's disease, sickle-cell anemia, and muscular dystrophy, and those that run in families, such as diabetes, hypertension, and breast cancer. Even though humans have always been susceptible to genetic disorders, there have been considerable advances in scientific technologies and screening methods to help detect the major disorders caused by mutant genes and those that lie dormant, passed on by prior generations. Conventional genetic screening techniques are used to help identify mutant genes, but they are limited to searching for the presence of specific genes or gene combinations. Alternatively, scientists are currently researching copy number variations in human genetic studies.

Copy number variations (CNVs) are large DNA sequences between ten thousand and five million bases that undergo variation (i.e., insertions, deletions, and duplications) in portions of that sequence of DNA, many of which are found within segmental duplications and regions of the genome that are repetitive. Though the exact percentage of genetic diseases that are caused by CNV is somewhat still unknown, it is possible that the types of diseases that run in families may result from mutations in these large DNA sequences.

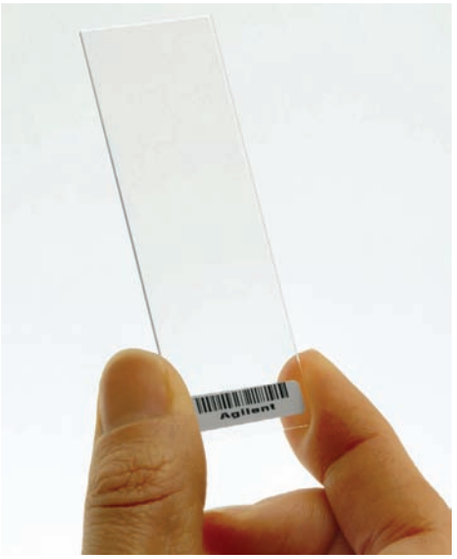

Figure 1 Human CNV Association 2 × 105 *array used in the study. For additional information about the microarray, please visit www.opengenomics.com.*

World's largest copy number variation study

Established in 2005, the Wellcome Trust Case Control Consortium (WTCCC) is an independent registered charitable trust comprised of 50 research groups across the U.K., including the Wellcome Trust Sanger Institute, University of Cambridge, and University of Oxford. The WTCCC is involved in ongoing research to increase the number of identified genes and new variants known to play a role in the susceptibility and cause of common diseases, including bipolar disorder, Crohn's disease, coronary artery disease, type 1 and type 2 diabetes, rheumatoid arthritis, breast cancer, and hypertension.

In August 2008, the WTCCC selected **Agilent Technologies** (Santa Clara, CA) to manufacture microarrays for a landmark human genome copy number variation study. The study employed the Agilent Human CNV Association 2 × 105 array (*Figure 1*), which uses two microarrays containing 105,000 probes on each array.

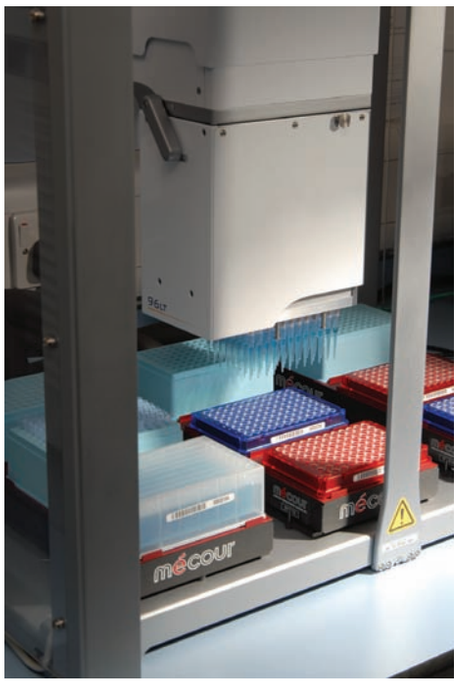

Figure 2 **OGT** *utilizes a Bravo Automated Liquid Handling Platform by* **Agilent** *Automation Solutions for sample preparation to facilitate the enormous throughput required in the study (shown with open inserts from* **Mécour***, Groveland, MA).*

Since **Oxford Gene Technology (OGT)** (Oxford, U.K.) is committed to providing high-quality data for a variety of high-throughput microarray applications, the WTCCC selected **OGT** as the microarray service provider for the study. Founded by Prof. Sir Edwin Southern, the pioneer of Southern blotting and microarray technologies in 1995, **OGT** is one of the largest **Agilent-**certified service providers today.

OGT offers a variety of products and services. In particular, the company provides high-quality microarray services, a service required by the

Figure 3 *Automated washing and drying eliminates degradation of fluorescent dyes and potential variability in data results. System from* ***SciGene*** *(Sunnyvale, CA).*

Figure 4 ***OGT*** *utilizes an ozone-free cabinet, which is important for detecting gene expression measurements since environmental ozone can degrade cyanine 5 (Cy5) in microarrays.*

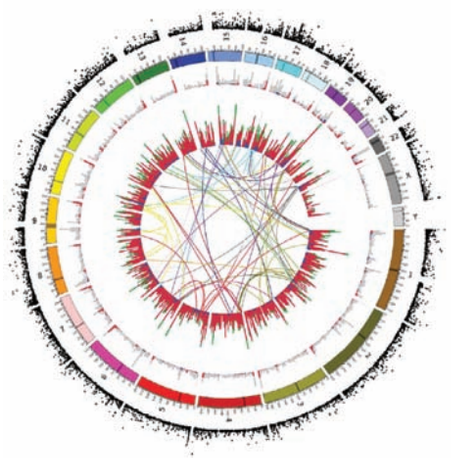

Figure 5 *Circular map showing the genomic distribution of different classes of CNVs and their population differentiation. (Figure 5 and Table 1 reproduced with permission from Ref. 1.)*

Table 1 **Trait-associated SNPs with possible causal CNVs[1]**

SNP	CNV	Location*	r^2†	Population‡	Data§	Reported gene	Trait	PMID
rs10492972	CNVR65.1	chr1: 10405137–10406094	0.92	CEU	Phased	*KIF1B*	Multiple sclerosis	18997785
rs11809207	CNVR118.1	chr1: 26332157–26337219	0.61	CEU	Phased	*CATSPER4*	Height	19343178
rs2815752	CNVR217.1	chr1: 72538870–72584557	0.96	CEU	Phased	*NEGR1*	Body mass index	19079261
rs7553864	CNVR240.1	chr1: 87385827–87386846	0.76	CEU	Intensities	*AK002179*	Smoking behaviour	19247474
rs4085613	CNVR358.1	chr1: 150822234–150856715	0.97	CEU	Phased	*LCE3D, LCE3A*	Psoriasis	19169255
rs11265260	CNVR381.1	chr1: 157915386–157916253	0.62	CHB+JPT	Phased	*CRP*	C-reactive protein	18439552
rs12029454	CNVR384.1	chr1: 160497369–160497846	0.57	CHB+JPT	Phased	*NOS1AP*	QT interval	19305408
rs6725887	CNVR1111.1	chr2: 203607766–203612122	1.00	CEU	Phased	*WDR12*	Myocardial infarction (early onset)	19198609
rs9311171	CNVR1355.1	chr3: 37953474–37961880	1.00	CHB+JPT	Phased	*CTDSPL*	Prostate cancer	17903305
rs3772255	CNVR1591.1	chr3: 157574746–157576258	0.90	CEU	Phased	*KCNAB1*	Ageing traits	17903295
rs9291683	CNVR1819.6	chr4: 9783252–9843664	0.51	YRI	Intensities	NR	Bone mineral density	17903296
rs9291683	CNVR1819.1	chr4: 9820419–9843664	0.51	YRI	Intensities	NR	Bone mineral density	17903296
rs401681	CNVR2293.1	chr5: 1386043–1386897	0.68	YRI	Intensities	*CLPTM1L*	Lung cancer	18978787
rs11747270	CNVR2646.1	chr5: 150157836–150161778	1.00	CEU	Phased	*IRGM*	Crohn's disease	18587394
rs11747270	CNVR2647_full	chr5: 150183562–150203623	1.00	CEU	Phased	*IRGM*	Crohn's disease	18587394
rs4704970	CNVR2659.1	chr5: 155409234–155427600	0.95	CEU	Phased	*SGCD*	Multiple sclerosis (age of onset)	19010793
rs12191877	CNVR2841.6	chr6: 31384505–31397416	0.79	CEU	Phased	*HLA-C*	Psoriasis	19169254
rs10484554	CNVR2841.6	chr6: 31384505–31397416	0.79	CEU	Phased	*HLA-C*	AIDS progression	19115949
rs3129934	CNVR2845.21	chr6: 32519885–32887814	0.87	CEU	Phased	*HLA-DRB1*	Multiple sclerosis	18941528
rs9277535	CNVR2846.3	chr6: 33156338–33162718	0.62	CEU	Intensities	*HLA-DPB1*	Hepatitis B	19349983
rs9277535	CNVR2846.5	chr6: 33159682–33163323	0.67	CEU	Intensities	*HLA-DPB1*	Hepatitis B	19349983
rs210138	CNVR2850.1	chr6: 33691917–33693857	0.55	CEU	Phased	*BAK1*	Testicular germ cell tumour	19483681
rs2301436	CNVR3164.1	chr6: 167408121–167409138	0.71	CEU	Intensities	*CCR6*	Crohn's disease	18587394
rs2705293	CNVR4074.1	chr8: 138980822–138981379	0.51	YRI	Intensities	*AK127771*	Neuroticism	18762592
rs1602565	CNVR5123.2	chr11: 29095953–29096982	0.64	CEU	Intensities	Intergenic	Schizophrenia	18677311
rs1602565	CNVR5123.1	chr11: 29096114–29096643	0.61	CEU	Intensities	Intergenic	Schizophrenia	18677311
rs7395662	CNVR5165.1	chr11: 48557432–48560877	1.00	CEU	Phased	*MADD, FOLH1*	HDL cholesterol	19060911
rs9300212	CNVR5492.1	chr12: 33606396–33608182	0.84	CEU	Phased	Intergenic	Cognitive test performance	17903297
rs1495377	CNVR5583.1	chr12:69818942–69819932	0.72	CEU	phased	NR	Type 2 diabetes	17554300
rs3118914	CNVR5871.1	chr13: 49967347–49973131	0.69	CEU	Phased	*DLEU7*	Height	19343178
rs763014	CNVR6576.1	chr16: 601068–603588	0.68	CEU	Intensities	*RAB40C*	Height	18391950
rs8049607	CNVR6636.1	chr16: 11591538–11592052	0.88	CHB+JPT	Phased	*LITAF*	QT interval	19305409
rs7188697	CNVR6746.1	chr16: 57231107–57233858	0.61	YRI	Phased	*NDRG4*	QT interval	19305409
rs1805007	CNVR6887.1	chr16: 88423599–88425903	0.87	CEU	Phased	*MC1R*	Skin sensitivity to sun	18488028

List of CNV correlations with trait-associated SNPs with $r^2 > 0.5$ (see main text for details). When a locus-trait association has been reported several times, only the results for the most recently published trait-associated SNPs are shown in this table. Some trait-associated SNPs are strongly correlated with more than one CNV in the same recombination hotspot interval. NR, no gene reported in original study; PMID, PubMed accession of the paper reporting the trait-associated SNP.
* Location of the CNV.
† Squared correlation coefficient.
‡ Population in which correlation observed; some SNP-CNV correlations are observed in several populations.
§ CNV data that correlates with the hit-SNP. Phased, phased SNP+CNV haplotypes; intensities, CNV intensity data and SNP genotypes. If present in phased and intensity data only phased data reported.

WTCCC in its study to identify genetic variants influencing disease susceptibility in a broad range of diseases. This service includes:

- Automated sample preparation, which can process in a 96-well format (*Figure 2*)
- Rigorous in-process quality control where samples undergo >30 QC analyses using a sensitive spectrophotometer before proceeding to the next stage
- Hybridization, completed on a 2 × 105K array in the study
- Automated washing and drying in an ozone-controlled environment (*Figure 3*)
- Data processing, which is also conducted in an ozone-controlled environment to avoid degradation and data variability (*Figure 4*).

Every stage in the process is directly monitored by a proprietary LIMS that allows for a 360° audit trail for every sample, which is very important for ensuring consistent results while meeting the project time line. In this study, the time factor was important because the WTCCC required 20,000 samples to be processed.

"In order to characterize genetic variants, reproducible performance and reliable processing of the high-resolution microarrays is essential. This project demanded high-quality data generated to tight deadlines, and we were very pleased with its rapid progress," said Dr. Matt Hurles of the Wellcome Trust Sanger Institute.

In early 2009, **OGT** reported that more than 20,000 samples generated by the WTCCC were successfully processed in 20 weeks. Processing over 1000 samples per week, this marked the world's largest copy number variation study.

"**OGT** is delighted to have successfully processed the huge number of samples, on time and to exacting QC standards, in this landmark CNV study," said Dr. John Anson, R&D Director at **OGT**.

CNV study

In October 2009, the scientists involved in this study published an article in *Nature*, entitled "Origins and Functional Impact of Copy Number Variation in the Human Genome," in which they described their comprehensive survey to detect common CNVs larger

than 1 kb in size in the human genome, and the development and application of experimental protocols to allow these CNVs to be assayed. *Figure 5* portrays a circular map showing the genomic distribution of different classes of CNVs and their population differentiation.[1]

The authors revealed that 80–90% of common CNVs larger than 1 kb (see note above re size ranges) in length have been discovered. Despite the fact that the CNVs most difficult to directly genotype were a result of duplications and contained multiallelic loci, they have been able to genotype approximately 40% of those discovered. *Table 1* shows trait-associated single-nucleotide polymorphisms (SNPs) with possible causal CNVs from the study.

The study indicated that CNVs are generated through mutational mechanisms that vary depending on the different sizes of the genomic alteration. Additionally, it appeared that the ability to regulate transcription also played a part in the formation of CNVs. This was suggested after an observation showed that the sequences enriched in CNV breakpoints happened to also be sequences in the promoter region with the inability to form B-DNA. Although additional research is to follow, this comprehensive population-based study concluded that CNVs may play a causative role in generating atypical variants of mutated genes in common diseases.

Reference

1. Conrad, D.F., et al. Origins and functional impact of copy number variation in the human genome. *Nature AOP* **2009**; doi:10.1038/nature08516.

Ms. Hunt is Editor, Integrated Media, for International Scientific Communications, Inc.; e-mail: jeanely_hunt@iscpubs.com.