A POPULATION-BASED STUDY OF COPY NUMBER VARIATIONS AND REGIONS OF HOMOZYGOSITY IN SINGAPORE AND SWEDISH POPULATIONS USING GENOME-WIDE SNP GENOTYPING ARRAYS

KU CHEE SENG

B. Sc. (Hons.), UM; M. Med. Sc., UM

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EPIDEMIOLOGY AND PUBLIC HEALTH

YONG LOO LIN SCHOOL OF MEDICINE

NATIONAL UNIVERSITY OF SINGAPORE

2011

ACKNOWLEDGEMENT

During the four years of my Ph.D. studies (August 2007 – August 2011), I'm grateful to the many people who in many and different ways have contributed to this work. Specifically, I would like to thank:

- Chia Kee Seng (main supervisor), Mark Seielstad (co-supervisor) and Yudi Pawitan (co-supervisor) for their guidance and encouragement throughout my Ph.D. studies, and for making all the publications possible
- Yudi Pawitan and Agus Salim for their guidance and discussion in data analysis
- Teo Shu Mei and Nasheen Naidoo, my course mates and colleagues, for helping in R package analysis (Shu Mei), critical reading and correcting the English of my manuscripts and thesis (Nasheen)
- All my colleagues and friends in the Center for Molecular Epidemiology and Department of Epidemiology and Public Health, National University of Singapore for their help and support

I would also like to acknowledge the funding agency. I was funded under the grant 'Singapore Consortium of Cohort Studies' from June 2007 – March 2011.

CONTENTS

SUMMARY

Population-based studies of copy number variations (CNVs) and regions of homozygosity (ROHs) have received considerable attention over the past few years. In addition, CNVs and ROHs were also found to be associated with various human complex diseases and traits such as schizophrenia, autism and height. Genome-wide mapping of CNVs and ROHs have been previously performed in European, East Asian and African populations using high-density SNP genotyping arrays. However, a comprehensive mapping study of CNVs and ROHs in the Singapore and Swedish populations has not been conducted previously. Therefore, the primary aim of this thesis was to detect and describe the characteristics of CNVs and ROHs in these two populations. A total of 292 samples from three Singaporean populations (99 Chinese, 98 Malay, and 95 Indian individuals) and 100 samples from the Swedish population were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 or/and Illumina Human1M BeadChip arrays. Subsequently, several hundred CNV loci and ROH loci were found in both populations. More interestingly, some of these CNV loci overlapped with known disease-associated or pharmacogenetic-related genes and showed substantial population frequency differences. Novel CNV loci that were not previously reported in public databases were also identified. Comparisons between these two populations and with the International HapMap III populations found substantial differences in their CNV and ROH profiles. Collectively, these results highlight the importance of characterizing CNVs and ROHs in individual populations. The studies in this thesis will establish a resource of CNVs and ROHs for future disease association studies in the Singapore and Swedish populations.

LIST OF PUBLICATIONS

1. **Ph.D. publications (see Appendices)**

(A) Research papers

1. Ku CS, Pawitan Y, Sim X, Ong RT, Seielstad M, Lee EJ, Teo YY, Chia KS, Salim A. Genomic copy number variations in three Southeast Asian populations. *Human Mutation* 31: 851-857 (2010).
2. Teo SM*, Ku CS*#, Naidoo N, Hall P, Chia KS, Salim A, Pawitan Y. A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals. *Journal of Human Genetics* 56: 524-533 (2011).
3. Ku CS#, Teo SM, Naidoo N, Sim X, Teo YY, Pawitan Y, Seielstad M, Chia KS, Salim A. Copy number polymorphisms in new HapMap III and Singapore populations. *Journal of Human Genetics* 56: 552-560 (2011).
4. Teo SM*, Ku CS*, Salim A, Naidoo N, Chia KS, Pawitan Y. Regions of homozygosity in three Southeast Asian populations. *Journal of Human Genetics* 57: 101-108 (2012).

* Joint first author

Corresponding author

(B) Review papers

1. Ku CS#, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics* 55:403-415 (2010).
2. Ku CS#, Naidoo N, Teo SM, Pawitan Y. Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* 129:1-15 (2011).

Corresponding author

(C) Encyclopedia/book chapters

1. Ku, Chee Seng; Naidoo, Nasheen; Teo, Shu Mei; and Pawitan, Yudi (February 2011) Characterising Structural Variation by Means of Next-Generation

Sequencing. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0023399

2. Chee-Seng, Ku; En Yun, Loy; Yudi, Pawitan; and Kee-Seng, Chia (April 2010) Next Generation Sequencing Technologies and Their Applications. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0022508
3. Chee-Seng, Ku; En Yun, Loy; Yudi, Pawitan; and Kee-Seng, Chia (April 2010) Whole Genome Resequencing and 1000 Genomes Project. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0022507
4. Chee Seng, Ku; Katherine, Kasiman; and, Kee Seng, Chia (September 2009) High-Throughput Single Nucleotide Polymorphisms Genotyping Technologies. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0021631

(D) Technical note

1. Ku Chee Seng, Sim Xueling, Chia Kee Seng. Genome-Wide Mapping of Copy Number Variations and Loss of Heterozygosity Using the InfiniumHuman1M BeadChip. Illumina Technical Note (2008).

2. Other publications during Ph.D. candidature (August 2007 – August 2011)

Publications	**Quantity**
Research paper	3
Review paper	7
Commentary	1
Encyclopedia/book chapters	6

COMPLETE LIST OF PUBLICATIONS (August 2007 – August 2011)

Research/Review papers

1. **Ku CS**, Pawitan Y, Sim X, Ong RT, Seielstad M, Lee EJ, Teo YY, Chia KS, Salim A. Genomic copy number variations in three Southeast Asian populations. *Human Mutation* 31: 851-857 (2010).
2. **Ku CS***, Teo SM, Naidoo N, Sim X, Teo YY, Pawitan Y, Seielstad M, Chia KS, Salim A. Copy number polymorphisms in new HapMap III and Singapore populations. *Journal of Human Genetics* 56: 552-560 (2011).
3. Teo SM, **Ku CS***, Naidoo N, Hall P, Chia KS, Salim A, Pawitan Y. A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals. *Journal of Human Genetics* 56: 524-533 (2011).
4. Teo SM, **Ku CS**, Salim A, Naidoo N, Chia KS, Pawitan Y. Regions of homozygosity in three Southeast Asian populations. *Journal of Human Genetics* 57: 101-108 (2012).
5. Mei TS, Salim A, Calza S, **Ku CS**, Chia KS, Pawitan Y. Identification of recurrent regions of Copy-Number Variants across multiple individuals. *BMC Bioinformatics* 11: 147 (2010).
6. Pawitan Y, **Ku CS**, Magnusson PK. How many genetic variants remain to be discovered? *PLoS One* 4: e7969 (2009).
7. Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, Small KS, **Ku CS**, Lee EJ, Seielstad M, Chia KS. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Research* 19: 2154-2162 (2009).
8. Naidoo N, Pawitan Y, Soong R, Cooper DN, **Ku CS***. Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Human Genomics* 5: 577-622 (2011).
9. **Ku CS***, Naidoo N, Wu M, Soong R. Studying the epigenome using next generation sequencing. *Journal of Medical Genetics* 48: 721-730.
10. **Ku CS***, Naidoo N, Teo SM, Pawitan Y. Regions of homozygosity and their impact on complex diseases and traits. *Human Genetics* 129:1-15 (2011).

11. **Ku CS***, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. *Human Genetics* 129:351-370 (2011).
12. **Ku CS***, Loy EY, Salim A, Pawitan Y, Chia KS. The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics* 55:403-415 (2010).
13. Hartman M, Loy EY, **Ku CS**, Chia KS. Molecular epidemiology and its current clinical use in cancer management. *Lancet of Oncology* 11: 383-390 (2010).
14. **Ku CS***, Loy EY, Pawitan Y, Chia KS. The pursuit of genome-wide association studies: where are we now? *Journal of Human Genetics* 55: 195-206 (2010).
15. **Ku CS***, Chia KS. The success of the genome-wide association approach: a brief story of a long struggle. *European Journal of Human Genetics* 16: 554-564 (2008).
16. **Ku CS**, Chia KS. Genome-wide association studies of type 2 diabetes. *Asia-Pacific Journal of Endocrinology* (2009).

***Corresponding author**

Commentary

1. Polychronakos C, **Ku CS**. Exome diagnostics: already a reality? *Journal of Medical Genetics* 48: 579.

Encyclopedia/book chapters (Encyclopedia of Life Sciences, Publisher: John Wiley & Sons)

1. **Ku Chee-Seng**, Loy En Yun, Pawitan Yudi, Chia Kee-Seng. Genome-wide Association Studies: The Success, Failure and Future. Published online: 15 December, 2009. **(*Keynote Article)**
2. **Chee Seng Ku**, Patrik K.E. Magnusson, Kee Seng Chia, Yudi Pawitan. Research on rare variants for complex diseases. Published online: 15 September, 2010. **(*Keynote Article)**
3. **Chee-Seng Ku**, Yudi Pawitan, Kee-Seng Chia. Genome-Wide Association Studies. Published online: 15 March, 2009.

4. **Ku Chee Seng**, Kasiman Katherine, Chia Kee Seng. High-Throughput Single Nucleotide Polymorphisms Genotyping Technologies. Published online: 15 September, 2009.
5. Jonathan T Tan, Kee Seng Chia, **Chee Seng Ku**. The Molecular Genetics of Type 2 Diabetes: Past, Present and Future. Published online: 15 September, 2009
6. **Ku Chee-Seng**, Loy En Yun, Pawitan Yudi, Chia Kee-Seng. **N**ext Generation Sequencing Technologies and Their Applications. Published online: 19 April, 2010.
7. **Ku Chee-Seng**, Loy En Yun, Pawitan Yudi, Chia Kee-Seng. Whole Genome Resequencing and 1000 Genomes Project. Published online: 19 April, 2010.
8. **Chee Seng Ku**, Nasheen Naidoo, Mikael Hartman, Yudi Pawitan. Genome wide association studies of cancers. Published online: 15 December 2010
9. **Chee Seng Ku**, Nasheen Naidoo, Mikael Hartman, Yudi Pawitan. Cancer genome sequencing. Published online: 15 December 2010
10. **Chee Seng Ku**, Nasheen Naidoo, Teo Shu Mei, Yudi Pawitan. Characterizing structural variation by means of next-generation sequencing. Published online: 15 February 2011

LIST OF TABLES

Chapter 2 - Background

Chapter 4 – Materials and methods

Chapter 5 - Results

LIST OF FIGURES

Chapter 2 - Background

Figure 6 – This figure illustrates the difference between ‘sequence coverage’ and ‘physical coverage'. At the specific nucleotide locus or position (red arrow), it is covered by two sequence reads highlighted by red circles (sequence coverage = 2), however, there are four paired-end sequence reads spanning the locus (physical coverage = 4) (adapted from Meyerson et al. (2010) Nat. Rev. Genet. 11:685-696).

Figure 7 – This figure illustrates that changes in sequencing depth (abundance of sequence reads) are used to identify copy number changes such as homozygous and hemizygous deletions and duplications.

Figure 8 – Plots of the differences in the LRR and BAF patterns for the ROH (left panels) and one-copy deletion (right panels) generated from a sample derived from our previous study (Ku et al. 2010) and genotyped by the Illumina 1M Beadchip (adpated from Ku et al. (2011) Hum. Genet. 129:1-15).

Chapter 5 - Results

Figure 9 – Number of CNVs per genome and their frequency in each of the three Singapore populations (adapted from Ku et al. (2010) Hum. Mutat. 31:851-857).

Figure 10 – Number of loci replicated by the Affymetrix platform and novel loci not found in the DGV.

Figure 11 - PCA comparing the Swedish and HapMap III populations.

Figure 12 - PCA results based on the common ROH loci for three Singapore populations.

LIST OF ABBREVIATIONS

ABI - Applied Biosystems

ADAMTSL3 - ADAMTS-like 3

ASW - people of African ancestry in the southwestern USA

BAC – bacterial artificial chromosome

BAF - B allele frequency

BMI – body mass index

Bp - basepair

CCDC60 - coiled-coil domain containing 60

CCL3L1 - chemokine (C-C motif) ligand 3-like 1

CEPH - Centre d'Etude du Polymorphisme Humain

CFH - complement factor H

CFHR1 - complement factor H-related 1

CFHR3 - complement factor H-related 3

CGH - comparative genomic hybridization

CHD - the Chinese community in Metropolitan Denver, Colorado, USA

Chr - chromosome

CNP – copy number polymorphism

CN – copy number

CNV – copy number variation

CTDSPL - CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like

CYP2A6 – cytochrome P450, family 2, subfamily A, polypeptide 6

CYP2A7 - cytochrome P450, family 2, subfamily A, polypeptide 7

DGV – database of genomic variants

DNA – deoxyribonucleic acid

DOC - depth-of-coverage

ERBB4 - v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)

FCGR3A - Fc fragment of IgG, low affinity IIIa, receptor

FCGR3B – Fc fragment of IgG, low affinity IIIb, receptor

FCGR2B - Fc fragment of IgG, low affinity IIb, receptor

FCGR2C - Fc fragment of IgG, low affinity IIc, receptor

FDR - false discovery rate

FISH – fluorescent in situ hybridization

GA - genome analyzer

GIH - Gujarati Indians in Houston, Texas, USA

GLG1 - golgi glycoprotein 1

GS FLX - genome sequencer FLX

GSTM1 - glutathione S-transferase mu 1

GSTM2 - glutathione S-transferase mu 1

GSTT1 - glutathione S-transferase theta 1

GSTT2 - glutathione S-transferase theta 2

GSTT2B - glutathione S-transferase theta 2B

GSTTP1 - glutathione S-transferase theta pseudogene 1

GWAS – genome-wide association studies

HapMap – haplotype map

HIV- human immunodeficiency virus

HLA – human leukocyte antigen

HLA-DRB1 - major histocompatibility complex, class II, DR beta 1

Indels – insertions and deletions

IRGM – immunity-related GTPase family, M

Kb - kilobase

LCE3B - late cornified envelope 3B

LCE3C - late cornified envelope 3C

LD – linkage disequilibrium

LRR - log R ratio

LWK - the Luhya inWebuye, Kenya

Mb - megabase

MEX - people of Mexican ancestry in Los Angeles, California, USA

MHC - major histocompatibility complex

MKK- the Maasai in Kinyawa, Kenya

mRNA – messenger ribonucleic acid

NEGR1 - neuronal growth regulator 1

NGS - next-generation sequencing

NHGRI - National Human Genome Research Institute

NUS-IRB - National University of Singapore-Institutional Review Board

PARK2 - parkinson protein 2, E3 ubiquitin protein ligase (parkin)

PC - principal component

PCA - principal component analysis

PCR – polymerase chain reaction

PEM - paired-end mapping

qPCR – quantitative polymerase chain reaction

RFLP – restriction fragment length polymorphism

ROH – region of homozygosity

ROMA - representational oligonucleotide microarray analysis

SGCD - sarcoglycan, delta

SMRT – single molecule real time

SNP - single nucleotide polymorphism

SNR - signal-to-noise ratio

SOLiD - supported oligonucleotide ligation detection

STR – short tandem repeat

SWED - Swedish

TLR7 - toll-like receptor 7

TGS – third generation sequencing

TMEM57 - transmembrane protein 57

TP63 - tumor protein p63

TSI - the Tuscans in Italy

UCSC – University of California, Santa Cruz

UGT2B17 - UDP glucuronosyltransferase 2 family, polypeptide B17

VNTR – variable number tandem repeat

WDR12 - WD repeat domain 12

WTCCC – Wellcome Trust Case Control Consortium

WWOX - WW domain containing oxidoreductase

YRI - Yoruba Ibadan Nigerian

ZNP510 - zinc finger protein 510

CHAPTER 1 – INTRODUCTION

A new era of copy number variations (CNVs) discovery began when two separate studies, published concurrently in 2004, identified several hundred deletions and duplications in the human genome[1, 2]. The comprehensive detection and characterization of CNVs has begun to lay the foundation to improve our understanding of human genetic variation and for deciphering the role of CNVs in the risk of complex diseases. Subsequently, recent evidence has linked CNVs to various complex diseases such as cancers, autoimmune diseases, schizophrenia and autism[3-8].

Over the past several years, most of the CNV data were generated by microarrays[9, 10]. However, a paradigm shift in the discovery of CNVs and copy-neutral variations was attributed to the development of a sequencing-based method known as paired-end mapping (PEM). This method was first demonstrated to be powerful in detecting structural variations (CNVs and copy-neutral variations) using next-generation sequencing (NGS) technologies in 2007[11]. Further studies also made use of the ability of NGS to generate several hundred million short sequence reads where CNV detection was based on the abundance or density of the sequence reads aligned to a reference genome. This approach is known as depth-of-coverage (DOC)[12].

However, at the time when our CNV project was started in 2007 as part of the Singapore Genome Variation Project[13], the sequencing-based methods to detect CNVs were still developing and were not well-established. The Singapore Genome Variation Project aimed to characterize the extent of common single nucleotide polymorphisms (SNPs) and

the patterns of linkage disequilibrium (LD) and haplotype in the human genome of DNA samples from each of the three populations in Singapore, i.e., Chinese, Malays and Indians (http://www.nus-cme.org.sg/SGVP/). Therefore, two high-density SNP genotyping arrays were chosen for the project. These arrays were the Affymetrix Genome-Wide Human SNP Array 6.0 and the Illumina Human1M BeadChip. As a result, the signal intensity data of these two genotyping arrays were also used for this CNV detection project. In addition, in collaboration with the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden, DNA samples from the Swedish population were also genotyped by the Affymetrix Genome-Wide Human SNP Array 6.0 for the project.

My thesis is divided into four studies (Study I – IV), each with a specific aim. The primary aim was to identify CNVs and study their population characteristics using high-density SNP genotyping arrays in the Singapore population (Study I) and the Swedish population (Study II). The motivation for these studies was that CNV data in the Singapore and Swedish populations is limited.

Besides our SNP dataset, the CEL-files of the Affymetrix SNP Array 6.0 for the seven populations in the International HapMap III project were downloaded from the International HapMap ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). This allowed us to investigate population differences of CNV profiles between the HapMap III and Singapore populations (Study III). It is important to study population differences,

particularly for those CNVs that overlap with known disease-associated genes, pharmacogenetics genes or other medically importance genes which could have different impacts in different populations[4, 14, 15]. Currently, the amount of data documenting the differences of CNVs in various populations is limited.

In addition to CNVs, regions of homozygosity (ROHs) can be also detected using high-density SNP genotyping arrays. ROHs are more abundant in the human genome of outbred populations than previously thought[16]. In addition, studies have identified ROHs to be associated with complex phenotypes such as schizophrenia, late-onset of Alzheimer's disease and height[17-19]. This suggests that studying ROHs may be useful for identifying genetic susceptibility loci harboring recessive variants for complex diseases and traits. Therefore, the secondary aim of this thesis was to identify and study ROH distribution patterns using the same set of SNP data (the Affymetrix SNP Array 6.0 and Illumina 1M datasets) in the Singapore population (Study IV). However, for the Swedish population, the ROH analysis was included in Study II.

In summary, the four studies in my thesis are:

Study I – Genomic copy number variations in three Southeast Asian populations

Study II – A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals

Study III – Copy number polymorphisms in new HapMap III and Singapore populations

Study IV - Regions of homozygosity in three Southeast Asian populations

CHAPTER 2 - BACKGROUND

2.1. Human genetic variations

Human genetic variations are the differences in the DNA sequence within the genome of individuals in populations and can take many forms, including single nucleotide changes or substitutions, tandem repeats, insertions and deletions (indels), additions or deletions that change the copies number of a larger segment of DNA sequence (i.e. CNVs), other chromosomal rearrangements such as inversions and translocations (also known as copy-neutral variations), and ROHs (Figure 1 and Table 1). These genetic variations span a spectrum of sizes from a single nucleotide to megabases. Single nucleotide substitutions or alterations involve a change in a single nucleotide at a particular locus in the DNA sequence, such as restriction fragment length polymorphisms (RFLPs), single nucleotide polymorphisms (SNPs) and single nucleotide indels. On the other extreme, CNVs, inversions, translocations and ROHs encompass larger segments of DNA sequences that range from kilobases to megabases (>1kb), whereas tandem repeats and indels fall between these extremes (>1bp to 1kb)[20, 21].

Table 1 – Categories of human genetic variations

Category	Genetic variation	Size
Single nucleotide changes	RFLP, SNP, single nucleotide indel	Single nucleotide
Tandem repeats	STR VNTR	2 – 8bp >8bp
Indels	Small indel Intermediate indel	2 – 100bp >100bp - <1kb
Structural variations	Deletion, duplication, inversion, translocation	>1kb
Copy-neutral loss of heterozygosity	ROH	>1Mb

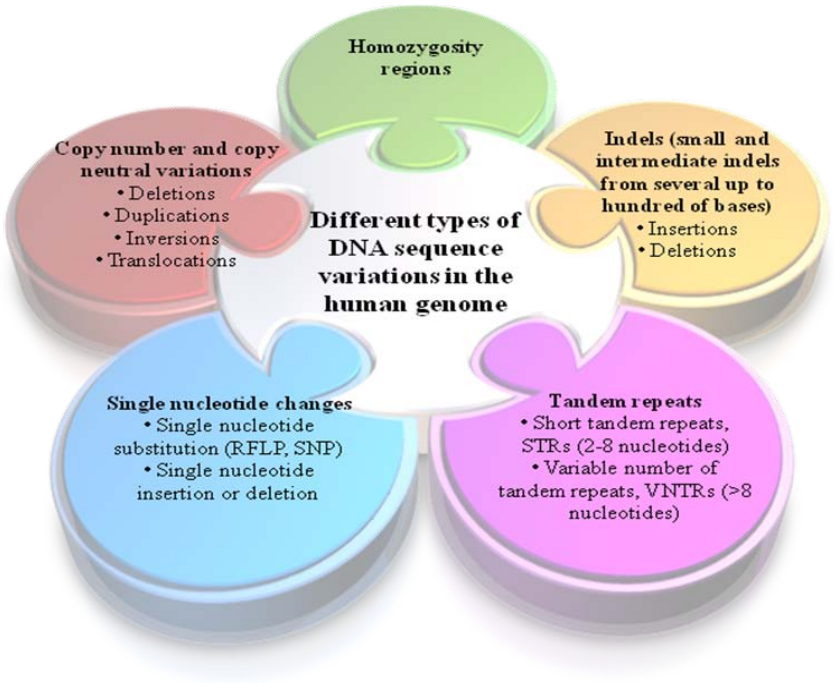


Figure 1 – Types of DNA sequence or genetic variations in the human genome. The genetic variations can be broadly divided into 5 categories (a) single nucleotide changes, (b) tandem repeats, (c) indels, (d) structural variations (CNVs and copy-neutral variations) and (e) ROHs.

In general, these genetic variations occur spontaneously in the human genome, and are the footprints of alterations that occur in DNA replication during cell division. External agents, such as viruses and chemical mutagens, can also induce changes in the DNA sequence. The occurrence of each type of genetic variation is mediated by different molecular mechanisms, although most of these are currently unclear. For example, several mechanisms have been proposed to explain the widespread occurrence of CNVs in the human genome, such as non-allelic homologous recombination and non-homologous end joining[22]. For ROHs, the homozygosity could have resulted from uniparental isodisomy and autozygosity[16]. Regardless of the molecular mechanisms that generated these genetic variations, they can be broadly classified as either somatic or germline variations depending on whether they arose during mitosis or meiosis, respectively.

The understanding of human genetic variations has advanced considerably over the past 30 years. Before the new millennium, the physical mapping of genetic variations such as RFLPs (in the 1980s)[23] and tandem repeats (in the 1990s)[24] was accomplished. By contrast, other genetic variations such as SNPs[25], indels[26, 27], CNVs[28-30] and ROHs[16] were identified after the turn of the new millennium. In addition to physical mapping, their biological functional roles, for example, their effects on or associations with mRNA expression levels, alternative splicing processes and other molecular and regulatory processes are now better understood[31-34]. Furthermore, these genetic variations were also found to be associated with various human diseases, including monogenic and complex diseases[4, 17, 34-37]. Presently, research in genetic variation is drawing much attention and

effort from the genetics community, as is evident from the initiation of the 1000 Genomes Project. A major aim of this project is to construct the most detailed map of genetic variations in the human genome. The pilot phase of the project was completed in 2010 (see section 2.10)[38].

2.2. Categories of genetic variations

There is still no clear consensus on how to define and categorize genetic variations. For example, SNPs are defined as single nucleotide substitutions; occasionally single nucleotide insertions or deletions also fall under this category (Figure 2a). Point mutations include both single nucleotide substitutions and single nucleotide indels with population frequencies of less than 1%. This is different from polymorphisms, when the population frequency is higher than the arbitrary cutoff of 1%.

Figure 2a – Single nucleotide changes (adapted from Ku et al. (2010) J. Hum. Genet. 55:403-415)[21]

Tandem repeats can be broadly divided into two classes: short and variable number tandem repeats (STRs and VNTR). STRs usually refer to tandem repeats in which the sequence length is arbitrarily set at eight nucleotides or less, and VNTRs are longer

tandem repeats (Figure 2b). They are also known as microsatellites and minisatellites respectively. The most common types of microsatellites are di-, tri- and tetra-nucleotide repeats. However, repeats of identical nucleotides of several bases or longer in the length are known as homopolymer sequences, for example, GGGGG or AAAAA. Although the sequence in the tandem repeats is simple compared with other more complex DNA sequence changes or rearrangements, these simple sequences can be repeated up to hundreds of times, thus creating very high heterozygosity or allelic diversity[20, 21, 39, 40].

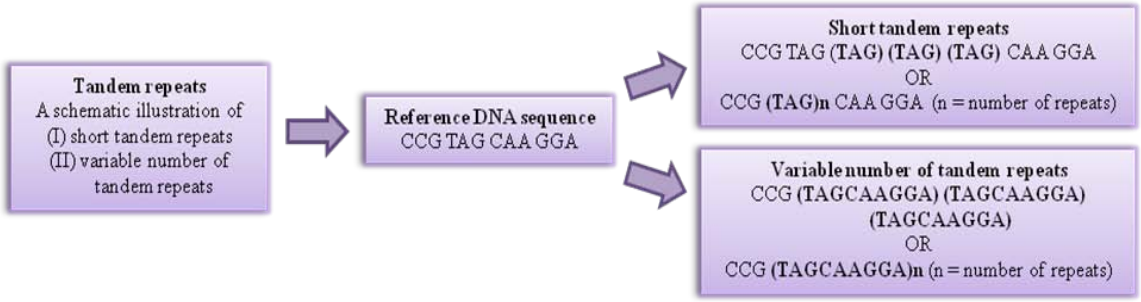


Figure 2b – Tandem repeats (adapted from Ku et al. (2010) J. Hum. Genet. 55:403-415)[21]

The boundary or distinction between CNVs and indels is even more unclear. In the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation/), deletions and duplications/insertions larger than 1kb are classified as 'CNVs', whereas those between 100bp to 1kb are grouped as 'InDels'. Table 2 summarizes the number of indels, CNVs and inversions cataloged in the DGV. As such, the remaining several hundred thousands of indels in the range of several nucleotides to tens of nucleotides, which were identified in the recent whole-genome resequencing studies, currently do not have their own category[41-47]. For example, Wang et al. (2008)[43] found approximately 140,000 indels

within 1-3bp in the Han Chinese Yan Huang (YH) genome, and approximately 400,000 indels defined from 1 to 16bp were also detected in the African NA18507 genome by Bentley et al. (2008)[44]. Thus, perhaps a new category such as 'short indels' (<100bp) is needed (Figures 2c and 2d). Similar to SNPs, common CNVs with population frequencies of 1% or higher are known as copy number polymorphisms (CNPs)[29].

Figure 2c – Indels (adapted from Ku et al. (2010) J. Hum. Genet. 55:403-415)[21]

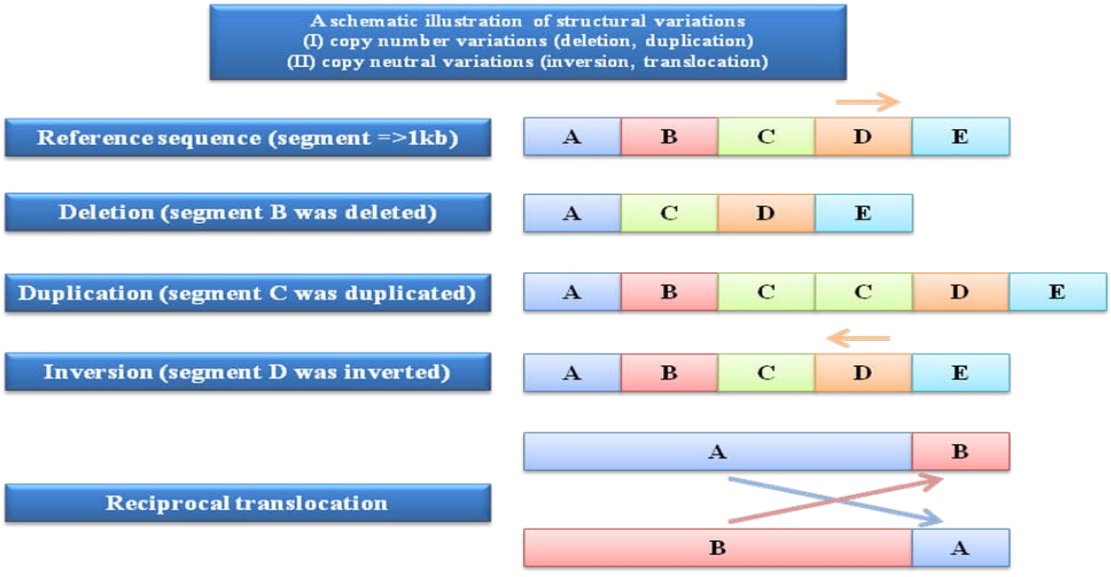


Figure 2d - Structural variations (adapted from Ku et al. (2010) J. Hum. Genet. 55:403-415)[21]

Table 2 – Summary statistics of the DGV

DGV entries	Number
Total entries	101923
CNVs	66741
Inversions	953
InDels (100bp – 1kb)	34229
Total CNV loci	15963

*Articles cited: 42 **Last updated: Nov 02, 2010

However, apart from single nucleotide changes, such as RFLPs and SNPs, all other genetic variations can be broadly grouped under the umbrella of structural variations[48]. It is important to note that these classifications are based primarily on patterns of changes in DNA sequence and an arbitrary definition of size. There is no consideration to the underlying biological mechanisms or their downstream functions that mediated their occurrences.

2.3. The evolution of genetic markers in disease gene mapping

Genetic variations in the human genome are useful as genetic markers for many different applications. These include:

(a) forensic investigations (for example, genetic or DNA fingerprinting)[49]

(b) routine clinical tests (for example, human leucocyte antigen typing for hematopoietic stem cell or organ transplantation)[50]

(c) prediction of drug responses or the tailoring of prescription doses (for example, genotyping tests for the SNPs in the thiopurine methyltransferase gene to predict patient responses to 6-mercaptopurine)[51]

(d) population genetics studies (for example, studies of human migration patterns)[52]

(e) genetic markers in disease gene mapping, such as family linkage and genetic association studies to identify the susceptibility loci or genes for monogenic and complex diseases.

Different genetic variations demonstrate different characteristics. Tandem repeats such as minisatellites and microsatellites are highly variable (polymorphic) in human populations. Therefore, they have higher allelic states and are more informative than the biallelic genetic markers, such as SNPs. Unlike SNPs in which a single nucleotide substitution will only give rise to two alleles, each repeat in minisatellites and microsatellites is considered as one allelic state. The genetic variations that occur in more than two allelic states are known as multiallelic markers. Tandem repeats have been widely used in genetic fingerprinting and as the genetic markers in linkage studies to locate the chromosomal regions harboring the mutations or genes for monogenic or familial disorders, complex diseases and quantitative traits[53-56]. Although tandem repeats are more informative than SNPs at the individual marker level, they are fewer in number than the several million SNPs in the human genome. Thus, tandem repeats are not ideal genetic markers for applications that require high marker density or resolution, such as genome-wide association studies (GWASs). In GWAS, a large number of genetic markers spanning the whole genome are required to achieve comprehensive coverage and adequate statistical power to detect unknown disease variants through LD[57, 58].

The rapid advances of high-throughput SNP genotyping technologies have enabled the genotyping of up to one million SNPs to be done efficiently on thousands of samples in GWAS. In contrast, no high-throughput method has been developed to assay microsatellites on a genome-wide scale[59-61]. This technological development, together with their abundance in the human genome, has resulted in SNPs becoming the primary genetic markers used in more than 500 GWAS (A Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/26525384). Almost all the GWAS have used the commercially available whole-genome SNP genotyping arrays from Illumina and Affymetrix.

Although SNPs have been studied in detail over the past decade, progress in the studies of other genetic variations, such as indels, CNVs and ROHs has been slow. CNVs started gaining more attention from the genetics community when several hundreds of deletions and duplications were first reported in 2004[1, 2]. Similarly, no large-scale attempt was made to identify indels until 2006, where a study by Mills *et al*. found several hundred thousand indels in the human genome[26]. The high frequency of ROHs in the genomes of outbred populations was also underappreciated until the first report in 2006[16]. Finally, the richness of genetic variations in the human genome has recently been further corroborated by several whole-genome resequencing studies, revealing a high frequency of new SNPs, indels, CNVs and other structural variations (Figure 3a and 3b). NGS technologies have facilitated and accelerated the process of identifying genetic variations through whole-genome resequencing and making the 1000 Genomes Project technically

feasible[62-65]. Several methods to detect structural variations based on NGS data were also developed (these methods will be discussed in sections 2.6. and 2.7).

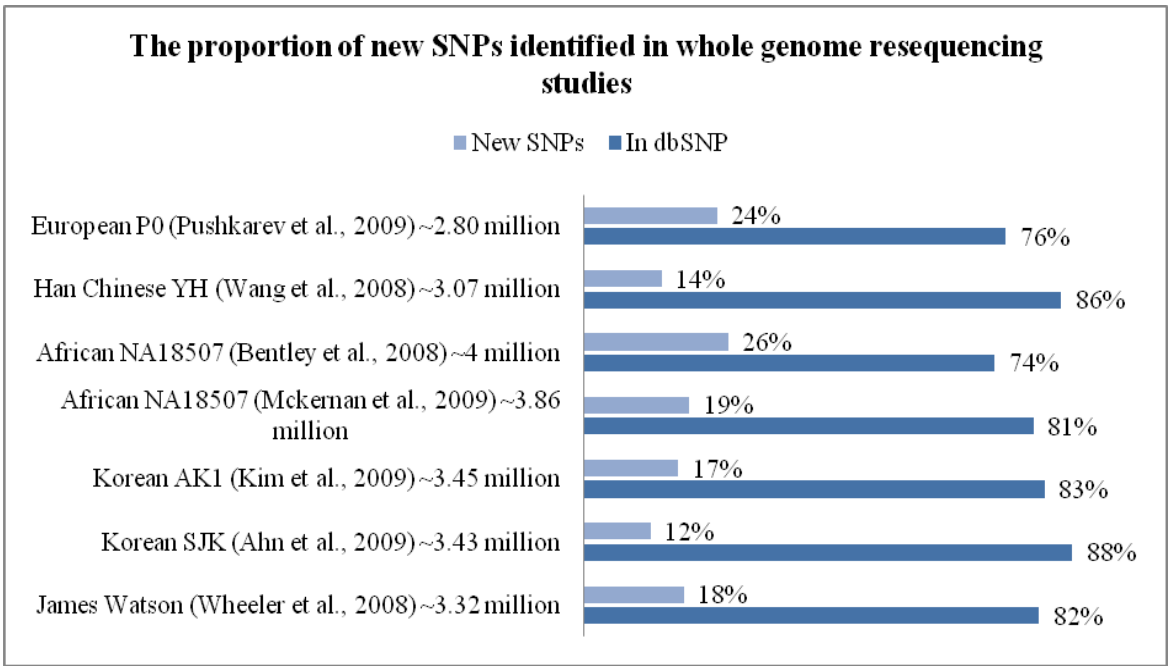


Figure 3a – The proportion of new SNPs identified in whole-genome resequencing studies (adapted from Ku et al. (2010) J. Hum. Genet. 55:403-415)[21]

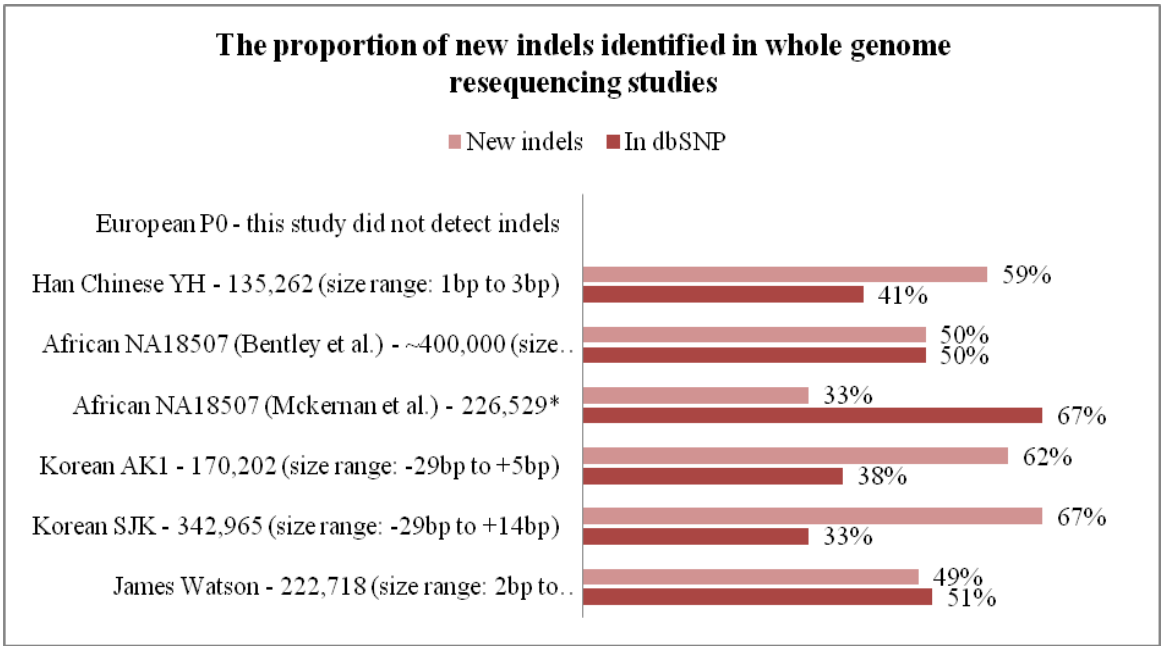


Figure 3b – The proportion of new indels identified in whole-genome resequencing studies (adapted from Ku et al. (2010) J. Hum. Genet. 55:403-415)[21]

** 89,679 insertions up to 3bp, 124,024 deletions up to 11bp, 12,826 larger indels. 67% of small indels in dbSNP (i.e., insertions up to 3bp and deletions up to 11bp)*

*** Approximately 0.4 million indels were identified and it was reported that approximately half of the indels are corroborated by entries in dbSNP*

In recent years, many studies have directly examined the associations of CNVs with complex diseases using SNP genotyping or comparative genomic hybridization (CGH) arrays. These studies have yielded exciting results for several diseases, such as schizophrenia and autism[66-68]. This further supports the use of CNVs as genetic markers to uncover new susceptibility loci for future disease association studies. Interestingly, genome-wide homozygosity mapping approaches have also been applied to dissect the genetic basis of complex diseases and have successfully identified a number of susceptibility loci for schizophrenia[17]. Conversely, short indels have not been directly interrogated in GWAS, but how much they can be tagged indirectly through LD by the

SNPs in genotyping arrays is unclear. Unlike CNVs and ROHs, which can be studied by SNP genotyping arrays, no high-throughput method has been developed to investigate short indels on a genome-wide scale. Direct detection and interrogation of short indels requires sequencing-based methods, as demonstrated in the whole-genome resequencing studies. As a result they cannot be effectively used as genetic markers in GWAS at the present time.

2.4. A new era of CNVs discovery through microarrays

A new era of CNVs discovery began when two separate studies, published concurrently in 2004, identified several hundred deletions and duplications in the human genome. Historically, large deletions and duplications were documented decades ago in clinical cytogenetics studies and found to cause various genomic or cytogenetic disorders[69]. The distinguishing feature of the recent studies was that these CNVs were more prevalent in the human genome than previously expected. These changes in copies number also did not result in any apparent phenotype or disorder and these regions of variable copies were found in the genomes of phenotypically normal individuals[1, 2]. As these submicroscopic (<5Mb) deletions and duplications are beyond the detection limit of traditional cytogenetics tools, such as molecular fluorescence in situ hybridisation (FISH), these recent discoveries can be credited to the use of whole-genome microarray technologies[10]. The term CNV was first introduced in 2006, and it is generally defined as additions or deletions in the number of copies of a particular segment of DNA (larger than 1kb in length) when compared with a reference genome sequence[70].

Although the early whole-genome microarray studies discovered several hundred CNVs, it was widely believed that the number of CNVs detected is likely to be under-estimated. For example, Sebat et al. (2004) detected a total of 221 CNVs in 20 individuals with an average CNV length of 465kb. These studies used 'low-resolution' microarrays such as ROMA (representational oligonucleotide microarray analysis) containing 85,000 probes with a resolution of approximately one probe for every 35kb[1], and the BAC-CGH array with a resolution of approximately one probe for every 1Mb[2]. Furthermore, these studies investigated a small sample size of only tens of individuals, which limit the detection of less common CNVs. CNVs smaller than 50-100kb will also not be detected as their size is below the resolution limits of these microarrays.

A later study by Tuzun et al. (2005) showed that approximately 85% of the 297 identified structural variations (139 insertions, 102 deletions and 56 inversions) were not detected by the two earlier studies. However, this study used a sequencing-based method, where the fosmid paired-end sequences were sequenced. Many of the structural variations that are being identified using this sequencing-based method are beyond the resolution limit of ROMA and the BAC-CGH microarrays. Inversions are also undetected by microarrays[1, 2, 71]. The discovery of many novel structural variations is due to the difference between the resolution of sequencing-based and microarray-based methods in detecting structural variations

However, the contribution of CNVs as a significant source of genetic variation in human populations has since been appreciated despite the limitations using microarrays. This is

evident from the enormous amount of interest and efforts generated towards mapping CNVs in different populations[28, 72, 73]. The first comprehensive mapping of CNVs in the 270 samples from the International HapMap I Project was completed in 2006. DNA samples from the HapMap individuals were screened for CNVs using two complementary microarray platforms, i.e., SNP genotyping and clone-based CGH arrays. A total of 1,447 copy number variable regions covering 360Mb (12% of the genome) were identified in these populations. More interestingly, these regions contained hundreds of genes, disease loci, functional elements and segmental duplications[28]. 'Human Genetic Variation' was then recognized as the 'Breakthrough of The Year' in 2007 by the journal *Science*. This was partly accomplished due to the significant progress made in the research of CNVs[74] (see Appendix: Table 1 - Summary of population-based CNV studies in different populations using SNP genotyping microarrays).

The limitations of ROMA and the BAC-CGH arrays have been overcome in later studies by using higher resolution microarrays and larger sample sizes of several hundred samples[29, 30, 75-78]. For example, Conrad et al. (2010) designed and custom-made a set of 20 tiling oligonucleotide-CGH microarrays comprising of 42 million probes with a median spacing of 56bp which were used for mapping CNVs in 40 HapMap samples. This study generated a comprehensive map of 11,700 CNVs greater than 443bp, of which 8,599 have been subsequently validated independently[30]. Other studies have also used the highest resolution SNP genotyping arrays that are commercially available, such as the Affymetrix SNP Array 6.0 and the Illumina Human 1M BeadChip[29, 78]. The 270 HapMap samples were rescreened with a higher resolution SNP genotyping array (i.e., the

Affymetrix SNP Array 6.0) and identified 1,320 common CNVs or copy number polymorphisms (CNPs) that segregate at an allele frequency >1%[29]. By contrast, Yim et al. (2010) screened CNVs in a much larger sample size (3,578 healthy Korean individuals) but used a lower density SNP array, i.e., the Affymetrix SNP Array 5.0 (an earlier version of the Affymetrix SNP Array 6.0)[77].

Over the past few years, most of the CNV data were generated using CGH and SNP microarrays, where fluorescence signal intensity information is used to detect deletions and duplications. These microarrays are highly accessible and affordable for population-based studies. Additionally, the methods of analysis and tools for detecting CNVs using microarray data, such as PennCNV and Birdsuite, have also been well-developed[79-81]. This has enabled studies of the characteristics of CNVs in various populations[29, 75, 77, 78]. However, due to the reliance on the relative or difference in signal intensity compared to a reference in inferring regions with copy number changes, this has hindered microarrays from detecting copy-neutral variations[10]. Furthermore, due to the limitations in marker density or resolution of microarrays used in the previous studies, these methods have poor sensitivity to detecting smaller CNVs (<50kb)[28]. However, the ability to detect smaller CNVs is critical as they are more numerous than the larger CNVs. The accuracy in determining the sizes or breakpoints of CNVs is highly dependent on the resolution of the microarrays as the sizes of CNVs found by previous studies were frequently over-estimated. It is notable that 88% of 1,153 CNV loci were smaller than sizes reported in the DGV, and that a reduction of >50% in size was observed for 76% of the CNV loci[82].

The latest developments in SNP genotyping arrays, such as an increase in marker density and uniformity of distribution in the genome and copy number probes to cover regions with sparse SNPs, have improved the sensitivity of microarrays. Nonetheless, these SNP microarrays still lack the sensitivity to detect CNVs smaller than 5-10kb, even with the use of the highest resolution microarrays such as the Illumina 1M and the Affymetrix SNP Array 6.0[29, 83]. While designing a set of high-resolution CGH microarrays comprising tens of millions of probes offers an unprecedented resolution, this method is more costly for several hundred samples[30], although, these improvements in microarrays are still unable to detect copy-neutral variations. Thus, developments of other methods that can overcome the limitations of microarrays and simultaneously detect both CNVs and copy-neutral variations are needed. Figure 4 illustrates the different signal intensity patterns of CNVs for oligonucleotide CGH and SNP genotyping arrays. Two types of signal intensity data were produced by SNP genotyping arrays, i.e., log ratio (total signal intensity) and B allele frequency (BAF, allelic intensity ratio). By contrast, the CGH array generated only a log ratio. As a result, ROHs can only be detected by a SNP genotyping array (see section 2.13).

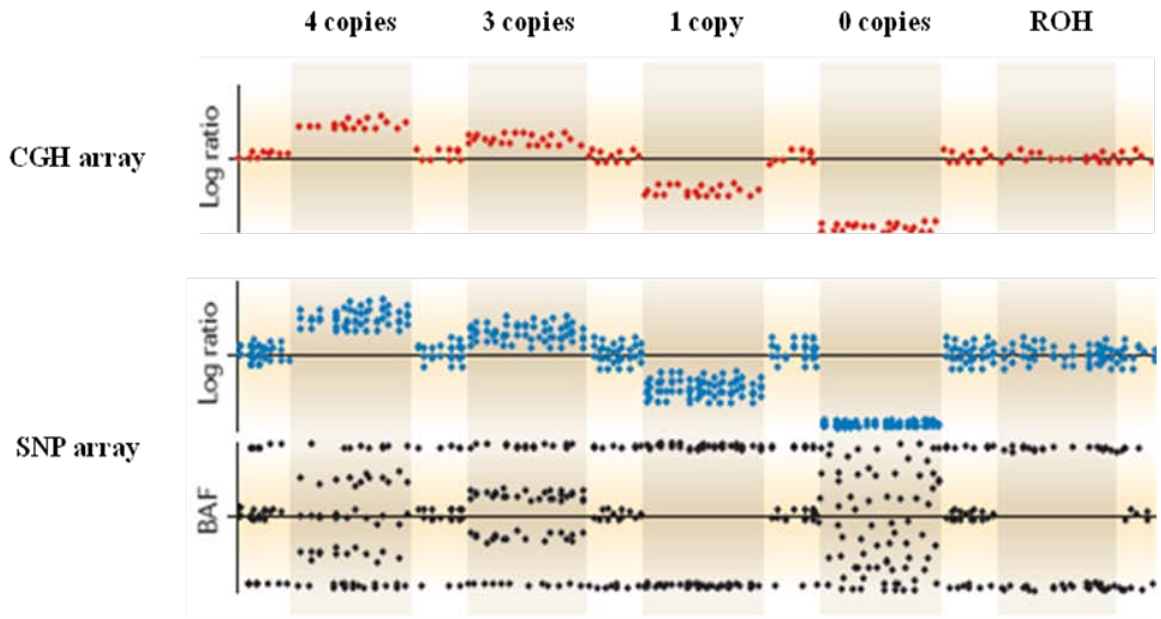


Figure 4 – Different patterns of signal intensity of CNVs for oligonucleotide CGH and SNP genotyping arrays (adapted from Alkan et al. (2011) Nat. Rev. Genet. 12:363-376)[84].

In array CGH (Figure 4, top panel), the signal ratio between a test and reference sample is normalized and converted to a $\log_2$ ratio, which acts as a proxy for copy number. An increased $\log_2$ ratio represents a gain in copy number in the test compared with the reference; conversely, a decrease indicates a loss in copy number. SNP arrays generate a similar metric by comparing the signal intensities of the sample being analysed to a collection of reference hybridizations, or the rest of the population being analysed. The log ratio metric for SNP arrays demonstrates a lower per-probe signal-to-noise ratio (SNR) than array CGH (compare log ratio for CGH and SNP arrays); however, SNP arrays offer an additional metric that enables a more comprehensive assignment of copy number than does array CGH (Figure 4, bottom panel). This metric, termed B allele frequency (BAF), can be calculated as the proportion of the total allele signal (A + B)

explained by a single allele (A). The BAF has a significantly higher per-probe SNR than the log ratio data and can be interpreted as follows: a BAF of 0 represents the genotype (A/A or A/–), whereas 0.5 represents (A/B) and 1 represents (B/B or B/–). Different BAF values occur for AAB and ABB genotypes or more complex genotypes (for example, AAAB, AABB and BBBA). Homozygous deletions result in a failure of the BAF to cluster. Thus, the BAF may be used to accurately assign copy numbers from 0 to 4 in diploid regions of the genome. The BAF also allows detection of copy-neutral events such as ROHs (also known as copy-neutral loss of heterozygosity) resulting from segmental uniparental isodisomy and identity by descent (see section 2.13).

2.5. Copy neutral variations - inversions and translocations

The discovery of CNVs in the human genome of healthy individuals from different populations has advanced rapidly over the last few years. However, similar progress is not seen in the detection of copy-neutral variations. This is due to the lack of a more powerful and efficient method for a genome-wide discovery of inversions and translocations. Unlike CNVs that can be studied by microarrays, the detection of copy-neutral variations usually requires sequencing-based methods. In addition, inversions and translocations are technically more difficult to detect. Relatively slower progress in the studies of copy-neutral variations is evident from the data entries recorded in the DGV (http://projects.tcag.ca/variation/), in which 66,741 CNVs and 34,229 indels have been reported in the database, whereas only 953 inversions have been found, and no data is available for translocations in the DGV presently (DGV last updated on 02 November 2010). However, one should be cautious with this interpretation as these are not

proportions. As the total number of CNVs, indels and inversions in the human genome is still unknown, the proportions of these genetic variations that have been discovered will remain unknown. The data in the DGV have been derived from the results of 42 studies using microarray-based, sequencing-based detection methods and other approaches. There are many more studies but their results have not been cataloged in the DGV. It is apparent that the entries in the database are still far from complete.

Most of the CNV data, available to date, were generated by microarray-based methods in which differences in signal intensities were used to detect deletions and duplications (Figure 4). As a result, these methods are unable to detect inversions and translocations (also known as balanced chromosomal rearrangements) because they do not lead to a gain or loss of chromosomal or DNA segments. Rather, several different strategies and approaches have been taken to try to identify inversions in the human genome. For example, Feuk et al. (2005) discovered regions that are inverted between the chimpanzee and human genomes by performing comparative analysis of their DNA sequence assemblies. In the study, they identified about 1,600 putative regions of inverted orientation in the genomes that covered >150Mb of DNA sequence. The inverted regions are distributed throughout the genomes and span sizes from 23bp to 62Mb in length. A number of inverted regions were also selected to be validated using PCR and FISH, and out of the 23 experimentally validated inversion regions, three were found to be polymorphic (>1%) in a panel of human samples, and were known as inversion polymorphisms[85].

A statistical method has also been developed to identify large inversion polymorphisms using high-density SNP genotyping data with unusual LD patterns. This method was developed to detect chromosomal regions that are inverted in a majority of the chromosomes in a population with respect to the reference human genome sequence. Although this method has worked using the International HapMap Project data to detect inversion polymorphisms, it has not been widely used by other studies. This study identified 176 inversions ranging from 200kb to several Mb in length using the HapMap Phase I data. However, their results were not cataloged in the DGV[86]. This, together with the study by Feuk et al. (2005)[85], also provided some evidence that a considerable portion of their detected inversions were flanked by highly homologous repeats or segmental duplications. This suggests that segmental duplications could be the favored spots mediating the chromosomal rearrangements that generate inversions.

The remarkable discovery of inversions was credited to the development of a sequencing-based method known as PEM, and the concurrent advances in NGS technologies. The PEM method also contributed greatly to the mapping of CNVs in the human genome. The power of this method to detect inversions was first demonstrated in a study by Tuzun et al. (2005) by sequencing the fosmid paired-end sequences. Their study successfully identified 56 inversion breakpoints. Kidd et al (2008) used the same strategy of fosmid clone sequencing to detect structural variations in eight individual genomes, and a total of 224 inversions were also identified[71, 87].

2.6. Sequencing-based detection methods - PEM

In the PEM method, a library of DNA fragments with a fixed insert size is prepared and both ends of the DNA fragments are sequenced to generate 'paired-end sequences' (the sequences at both ends of the DNA fragments). This sequence information is then aligned against the reference genome. The underlying principle of PEM is to detect the discrepancy or discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome to infer 'simple' deletion, insertion and inversion (Figure 5). The use of the term 'simple' is to distinguish from other more complex structural variations such as 'everted duplication', 'linked insertion' and 'hanging insertion'[11, 71].

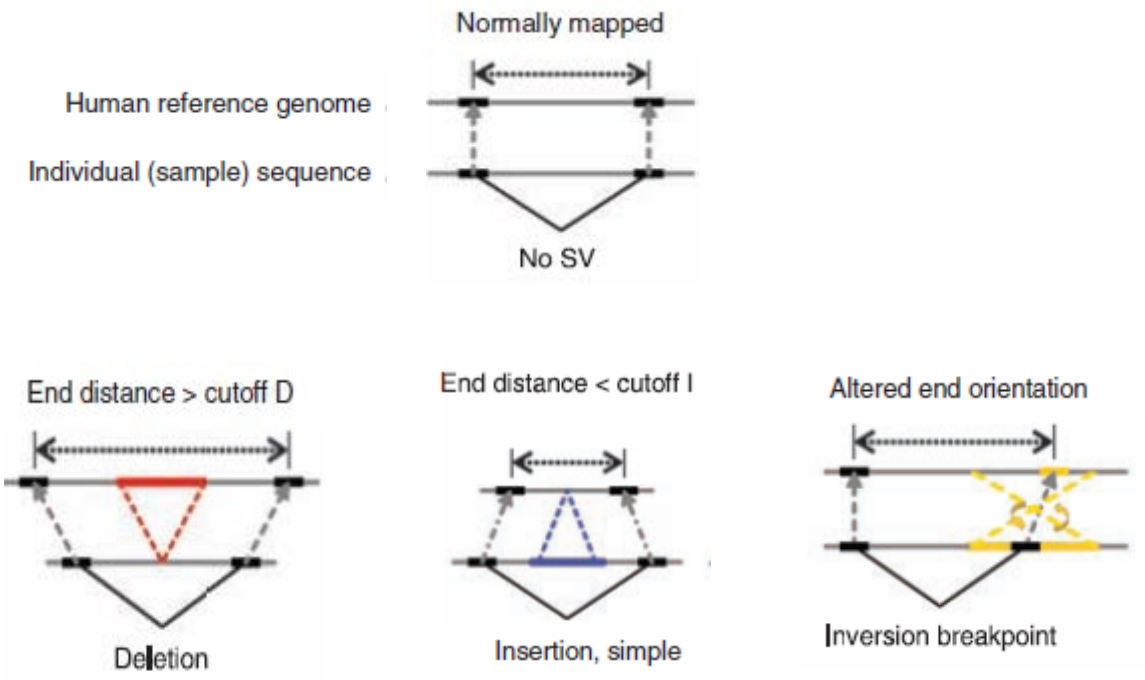


Figure 5 – Top panel: No discrepancy or discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome. Bottom panel: (a) Simple deletions were predicted from paired-end sequences span larger than a specified cutoff 'D' (red region indicates region deleted from sample genome); (b) simple insertions had

a span smaller than a specified cutoff 'I' (blue region; indicates region inserted in sample genome) and (c) inversions are seen when ends map to the genome at different relative orientations (yellow region indicates region inverted in sample genome)(adapted from Korbel et al. (2007) Science 318:420-426)[11].

When paired-end sequences that are being aligned to the reference sequence display discordance from the expected insert size or distance, this is an indication of deletion and insertion, whereas discordance in orientation suggests the presence of inversion (i.e., paired-end sequences are incorrectly oriented compared to the reference genome). Since the insert size of the DNA fragment library is known, the presence of insertion is indicated when paired-end sequences that align to the reference sequence are substantially shorter than expected. Conversely, a longer than the expected insert size suggests the presence of deletion, while other more complicated patterns of discordance when aligning the paired-end sequences provide hints at more complex rearrangements[11, 71, 88].

As such, the paired-end sequences are usually classified as 'concordant pairs' or 'discordant pairs', with only the discordant pairs informative for inferring structural variations. The presence of both concordant and discordant pairs spanning a locus suggests a heterozygote state with respect to the structural variation, e.g. a deletion occurs only in one homologous chromosome. In addition, multiple paired-end sequences are usually needed to reliably infer whether a locus harbors a structural variation. The requirement of multiple paired-end sequences spanning a locus to detect structural variations will reduce the number of false-positive signals. It will also minimize the false-

negative rate, for example, a heterozygous deletion will be missed by the presence of one concordant pair. However, with multiple paired-end sequences, it is more likely that both the concordant pair and discordant pair will be observed to detect the heterozygous deletion. As a result, a sufficient amount of sequencing is needed to ensure that there are multiple paired-end sequences spanning across the genome. This also means that a substantial amount of sequencing is needed for the PEM method and thus this method will be more costly when using Sanger sequencing compared to NGS technologies[11, 71, 88].

'Physical coverage' is important for detecting structural variations using PEM. Physical coverage measures the number of fragments spanning a site and this affects the ability to detect structural variations. It is different from 'sequence coverage' which measures the number of sequence reads that cover a site and this sequence coverage affects the ability to detect single nucleotide variants or point mutations (Figure 6). Thus, physical coverage can be increased by creating a library of larger insert sizes. When preparing a 'shotgun library' using standard methods, the sizes of DNA fragments or insert sizes are usually several hundred bases and approximately tens of bases on both ends of the DNA fragments which are sequenced using NGS technologies[89].

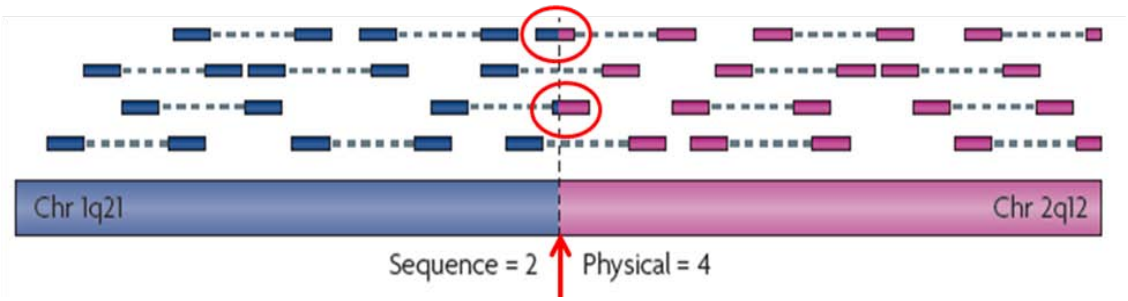


Figure 6 – This figure illustrates the difference between 'sequence coverage' and 'physical coverage'. At the specific nucleotide locus (red arrow), it is covered by two sequence reads highlighted by red circles (sequence coverage = 2), however, there are four paired-end sequence reads spanning the locus (physical coverage = 4) (adapted from Meyerson et al. (2010) Nat. Rev. Genet. 11:685-696)[89].

However, the insert size can be increased to several kilobases by creating a 'jumping library' or a 'mate-pair library'. Additional steps are involved in preparing a mate-pair library in comparison to a paired-end library, where both ends of the DNA fragments of several kilobases, e.g. 3kb in the Korbel et al. (2007) study were first ligated with biotinylated hairpin adapters. The DNA fragments were then circularized and randomly sheared. The fragments attached to biotinylated hairpin adapters were isolated to form a mate-pair library and then followed by sequencing[11]. Mate-pair library construction enables sequencing at both ends of longer DNA fragments of several kilobases. The mate-pair library with a larger insert size will increase the physical coverage of the genome. For example, by sequencing 50 bases from both ends of the DNA fragments from a library with a 3kb insert size, the physical coverage of the genome is 10-fold higher than from a library with a 300bp insert size. However, the sequence coverage is similar between both libraries as only 50 bases of paired-end sequences were generated with regards to the library insert size[89].

Thus the paired-end and mate-pair libraries differ only in the steps of constructing these libraries, as the sequencing and aligning of the paired-end sequences to the reference sequence to detect structural variations follow the same principle. Although creating a mate-pair library increases physical coverage, a larger insert size is less sensitive to detecting smaller structural variations because of the difficulty in tightly controlling the sizes of the DNA fragments in the library. Therefore, depending on the 'tightness' or 'narrowness' of the distribution pattern (standard deviation) of the insert sizes in the library, it can be difficult to distinguish a true PEM signature caused by a small indel (i.e., an indel of several or tens of bases) because of the variance in insert sizes in the library. This is due to the fact that it is not technically possible to generate an exactly similar size for each of the DNA fragments when preparing a library[89].

In comparison to microarray-based methods, PEM has a higher sensitivity to detecting smaller CNVs in addition to identifying copy-neutral variations, and it also has greater precision in determining the breakpoints or boundaries of structural variations. For example, the PEM method has been applied in a number of whole-genome resequencing studies where several thousand structural variations were detected[43, 46].

Nonetheless, this method could be biased against detection of duplications or insertions. This has been clearly shown in the YH genome, where most of the identified CNVs are deletions, namely 2441 deletions compared to 33 duplications. This is because PEM is unable to detect insertions larger than the insert size of the reference library. This also reveals the major limitation of PEM with fixed insert size in detecting insertions[43].

Deletions are easier to be detected because they are identified by a longer than expected insert size when aligned to the reference, whereas the detection of insertions is restricted by the insert size. This means that insertions larger than the insert size are beyond the detection range. Therefore, several paired-end and mate-pair libraries with short and long insert sizes will be needed to capture structural variations of varying sizes. This will also nevertheless increase the sequencing costs several fold depending on the number of libraries. As the bias against detection of insertions is partly due to the small insert size, therefore, larger insert sizes of several kilobases should improve the ability to detect more insertions[11].

2.7. Sequencing-based detection methods - DOC

Depth-of-coverage (DOC) is another method using the NGS data for CNVs detection. This method is based on the depth of coverage of the sequence reads to infer deletions and duplications. The DOC method is enabled by the production of several hundred million short sequence reads per instrument run by NGS technologies. The principle underlying the DOC approach is based on the assumptions that the sequencing process is uniform so that the number of sequence reads mapping to a region follows a Poisson distribution. As such, the number of sequence reads should be proportional to the number of times that a particular region appears in the genome (Figure 7). Therefore, it is expected that a duplicated region will have more reads aligning to it, and the converse is true for deletions[12, 88]. However the assumption that the sequencing process is uniform may not be valid. This is because of the sequencing bias of the NGS technologies which

leads to certain regions of the genome being over or under-sampled resulting in spurious signals[90].

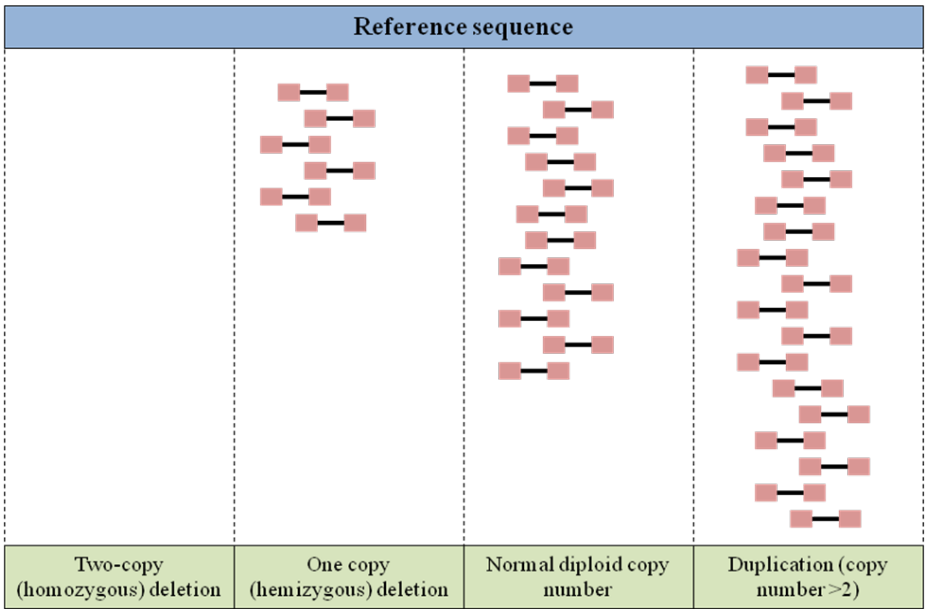


Figure 7 – This figure illustrates that changes in sequencing depth (abundance of sequence reads) are used to identify copy number changes such as homozygous and hemizygous deletions and duplications.

Based on the principle of the DOC method, the strength of a DOC signature (i.e., 'gain' or lose') is thus directly related to the coverage of the sequencing data (the number of sequence reads) and also to the size of the CNVs. This means that the DOC signatures will be stronger for larger CNVs, and is thus more powerful for detecting larger CNVs compared to PEM. In contrast, unlike PEM, the DOC method cannot detect copy-neutral variations. Moreover, the DOC method may not be powerful enough to identify smaller CNVs (related to the strength of DOC signatures) and it is also limited in defining breakpoints[88]. In comparison to microarrays, copies number can only be inferred to four

(CN=4) as the upper boundary for SNP microarray or copy number changes will be denoted as 'gain' or lose' for CGH microarray[29, 79]. The DOC method is also more robust and accurate at determining higher copies number.

2.8. Choosing a sequencing platform for PEM and DOC

The applications of high-throughput sequencing technologies that are commercially available and accessible to end-users or researchers for PEM and DOC will be discussed. The sequencing technologies that are currently available can be broadly grouped into NGS technologies such as the Roche 454 Genome Sequencer FLX (GS FLX) System, Illumina Genome Analyzer (GA) and HiSeq, and Applied Biosystems (ABI) Supported Oligonucleotide Ligation Detection System (SOLiD) and third generation sequencing (TGS) technologies such as the HeliScope Single Molecule Sequencer which is now commercially marketed by Helicos Biosciences. The features of NGS technologies are summarized in Table 3. It is noteworthy that the development of numerous other sequencing technologies are on the horizon, such as single molecule real time (SMRT) sequencing technology by Pacific Biosciences, which is now commercially marketed[91]. Others technologies such as nanopore sequencing may take several years to become a mature technology[92]. In comparison, Complete Genomics provides a sequencing service rather than selling their sequencing machines to end-users[93].

Table 3 - Summary of the features of NGS technologies

Feature	Roche® 454 GS FLX	Illumina® GAII/HiSeq	ABI® SOLiD
Commercially marketed	2005	2006	2007
Current generation of the sequencer	Roche® 454 GS FLX Titanium	Illumina® HiSeq	ABI® SOLiD 4.0
Massively parallel sequencing (number of DNA fragments)	Several hundred thousand to one million	Several hundred million	Several hundred million
Sequencing throughput per instrument run	Several hundred megabases per run in 10 hours	Several hundred gigabases per run in a few days	Several hundred gigabases per run in a few days
In vitro amplification method	Emulsion PCR	Bridge amplification on solid surface	Emulsion PCR
Sequencing approach	Sequencing by synthesis mediated by polymerase - pyrosequencing	Sequencing by synthesis mediated by polymerase - sequencing by reversible terminator chemistry	Sequencing by ligation of di-nucleotide probes mediated by ligase
Sequencing reagent	4 types of dNTPs	4 types of ddNTPs labeled by 4 different fluorescent colors	16 types of di-nucleotide probes labeled by 4 different fluorescent colors
Detection method of the incorporated nucleotides	Emission of chemiluminescent light	Fluorescent colors	Fluorescent colors
Sequence read length	400-500 bases	75-125 bases	50 bases
Read base or base calling error rate	0.5-1.5%	0.2-2%	<0.1%
Error type	Insertion or deletion of nucleotides in homopolymer sequences	Substitution of nucleotides	Substitution of nucleotides

(This table was adapted from Ku et al. (April 2010) Next Generation Sequencing Technologies and Their Applications. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0022508)

Although Roche 454 GS FLX, Illumina GA/HiSeq and ABI SOLiD are classified as NGS technologies, several features differ substantially between them. They are characterised by the ability of parallel sequencing of a very large number of sequence reads. However, the Roche 454 GS FLX can only generate approximately one million sequence reads per instrument run, in comparison to the Illumina GA/HiSeq and ABI SOLiD where several hundred million sequence reads are produced. Similarly, the HeliScope Single Molecule Sequencer can also produce several hundred million sequence reads[62-64, 94]. One of the major distinctions between NGS and TGS is that TGS does not require whole-genome amplification steps such as emulsion PCR and bridge amplification compared to NGS. Therefore, TGS has the potential to further increase the number of sequence reads or throughput per instrument run than their current capacity. Therefore, the Illumina GA/HiSeq, ABI SOLiD and HeliScope Single Molecule Sequencer provide an advantage for the DOC method that requires high density of sequence reads to infer CNVs. The specificity of DOC to detect CNVs and the precision to map the breakpoints can be improved by increasing the density or coverage of sequence reads[12, 88]. However, the length of sequence reads produced by the Roche 454 GS FLX is on average 400-500bp, which is substantially longer than that of the other three sequencing technologies which range from 32bp to 125bp[94]. Although PEM and DOC methods are targeting larger structural variations, the sequence read lengths produced by the Roche 454 GS FLX is better for detecting small indels of up to tens of bases. Moreover, the longer sequence read lengths produced by the Roche 454 GS FLX may also be more suitable for *de novo* genome assembly before read lengths of several kilobases are generated by future sequencing technologies.

The PEM method, when applied alone rather than integrated with DOC data, must ensure that the paired-end sequences be uniquely aligned to the reference genome to infer structural variations, compared to ambiguous paired-end sequences which align to multiple locations. As such, shorter sequence read lengths may be less specific in aligning against the reference genome, especially in repetitive regions such as segmental duplications. Moreover, the number of paired-end sequences is also important as multiple discordant pairs are usually needed to reliably detect a structural variation. In terms of preparing the PEM libraries for sequencing, all three NGS technologies are able to generate both paired-end and mate-pair libraries, thus allowing for sequencing of short and longer insert sizes[95, 96]. Each of the sequencing technologies has its own strengths and weaknesses, and a combination of these technologies in an experiment may be the ideal approach to detecting new structural variations and to also address the systematic biases in sequencing[90]. Table 4 summarizes the comparison between microarrays and sequencing-based methods for detecting structural variations.

Table 4 – Comparison between microarrays and sequencing-based methods for detecting structural variations

	Microarrays*	**PEM****	**DOC**
Principle	Based on the relative or difference in florescence signal intensity compared to a reference (one sample or a set of samples) to infer CNVs	Based on the discrepancy or discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome to infer 'simple' deletion, insertion and inversion	Based on the density of sequence reads being aligned to the reference genome to infer CNVs
Ability to detect CNVs	Yes	Yes	Yes
Ability to detect copy neutral variations	No	Yes	No
Reliably detecting CNVs	Up to tens of probes	Multiple discordant pairs	A high density of sequence reads
Application to population-based studies	Commonly applied to up to thousand samples	Has not yet been applied	Has not yet been applied
Sensitivity to detect smaller CNVs e.g. <10kb	Generally poor, but dependent on the resolution of the microarrays, e.g. a set of oligonucleotide CGH arrays containing 42 million probes has provided an unprecedented resolution	Yes, preparation of several libraries of different insert sizes are able to detect insertions and deletions of varying sizes, but the detection of insertions is limited by the insert sizes	It may not be powerful enough to detect smaller CNVs (related to the strength of DOC signatures and the coverage of the sequencing data or the number of sequence reads)
Sensitivity to detect larger CNVs	Yes, even low resolution BAC clone CGH arrays (with a resolution of approximately one probe for every 1Mb) have been used to detect CNVs of several hundred kilobases to megabases	Yes, however the detection of insertions is limited by the insert sizes, thus preparation of fosmid or BAC clone libraries with larger insert sizes are needed for detecting larger insertions	Yes, the DOC signatures will be stronger for larger CNVs
Precision in mapping breakpoints	Generally poor, however, it can be improved by increasing the	Good, theoretically the breakpoints can be mapped to a single	The precision to map the breakpoints can be improved by

	resolution of microarrays	nucleotide resolution	increasing the density or coverage of sequence reads
Role in 'discovery' and 'genotyping'	Can be used as an effective method to genotype newly discovered and known CNVs in population-based studies	Powerful for discovery of new structural variants	Discovery of CNVs especially in regions such as segmental duplications where PEM is less effective
Weakness as a result of technology limitations	Generally have poor signal-to-noise ratios for oligonucleotide-CGH and SNP microarrays compared to BAC clone CGH arrays	Short sequence reads are less specific in aligning uniquely to the reference genome especially in segmental duplications	Sequencing biases may lead to certain regions of the genome being over or under-sampled resulting in spurious DOC signatures
Scalability of sample throughput by technology	High sample throughput, for example, several hundred samples per week can be genotyped by SNP arrays, as evident in genome-wide association studies	Tens of gigabases of sequencing data can be produced per instrument run in several days by NGS technologies, and the sample throughput can be scaled up by 'barcoding', i.e., labeling the samples by barcodes	Tens of gigabases of sequencing data can be produced per instrument run in several days by NGS technologies, and the sample throughput can be scaled up by 'barcoding', i.e., labeling the samples by barcodes
Level of analytical and computational challenges	Lesser, analytical methods for detecting CNVs using microarray data are well-developed	Greater, an emerging and maturing method leveraging on the large amount of NGS data	Greater, an emerging and maturing method leveraging on the large amount of NGS data
Difficulty in sample preparation	Easier in processing the samples for hybridization on the microarrays	More challenging in preparing sequencing libraries especially clone-based libraries	More challenging in preparing sequencing libraries

*Whole-genome oligonucleotide-CGH and SNP microarrays **Paired-end and mate-pair libraries and clone-based libraries (such as fosmid and BAC clones) for PEM (This table was adapted from Ku et al. (February 2011) Characterising Structural Variation by Means of Next-Generation Sequencing. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester. DOI: 10.1002/9780470015902.a0023399)

2.9. International effort to characterize structural variations using PEM

The PEM method to detect structural variations was first demonstrated by Tuzun et al. in 2005 by mapping paired-end sequencing data from a human fosmid DNA genomic library. The average insert size of a fosmid library is approximately 40kb. This study identified 139 insertions, 102 deletions and 56 inversions[71]. However, sequencing of fosmid clones by means of Sanger sequencing is laborious and costly. These limitations have been overcome by NGS technologies which directly sequence the paired-end or mate-pair libraries without the need for cloning steps[11]. Both of the studies applied the PEM approach to investigate structural variations in the same sample (NA15510) from the International HapMap Project. However, their library insert sizes differed and this has enabled a comparison of the sensitivity between these studies. Korbel et al. (2007)[11] were able to confirm 41% of all deletion and inversion events detected by fosmid paired-end sequencing. Additionally, they identified an additional 407 structural variations in NA15510 that had not been previously detected by fosmid paired-end sequencing. This further suggests that several libraries with different insert sizes are needed to increase the sensitivity of PEM.

In addition to individual studies, a large-scale effort is currently underway by the Human Genome Structural Variation Working Group to comprehensively map structural variations in phenotypically normal individuals using the PEM approach[97]. More specifically, the objective is to characterize the pattern of human structural variations at the nucleotide sequence level from a collection of 48 individuals of European, Asian and African ancestry. This project plans to make fosmid clone libraries of approximately

40kb insert size from the genomic DNA of 48 unrelated females. These samples were already genotyped by the HapMap Project. A larger insert size of approximately 150kb prepared from BAC clone libraries will also be constructed from 14 unrelated HapMap males. This will aim to provide sequence information on structural variations that are too large to be included in the fosmid libraries, such as those associated with segmental duplications. As such, both the fosmid and BAC libraries will ensure the comprehensive capture of structural variations of varying sizes across the human genome.

Structural variation is biased toward complex duplicated and repetitive regions. Hence, developing clone libraries for a modest number of human genomes should serve as a valuable resource for characterizing complex and difficult-to-assay regions of genome structural variation. Since the underlying clones can be retrieved, the complete sequence context of the discovered structural variation can also be obtained[97]. This is crucial for precise breakpoint delineation of structural variation, which is then important for understanding the mutational mechanisms responsible for human genome structural variation. A total of 1,695 structural variations were discovered with fosmid libraries derived from nine individuals. This study also showed that 50% were seen in more than one individual and that nearly half lay outside regions of the genome previously described as structurally variant thereby indicating novel discoveries. More importantly, 525 new insertion sequences (that are not present in the human reference genome) were discovered and many of these were found to be variable in copy number between individuals[87]. This is important because it suggests that structural variations or CNVs

could have gone undetected as part of the 'missing sequences' in the human reference genome.

The complete sequencing of 261 structural variations provided insights into the different mutational processes that have shaped the human genome. Thus, the study by Kidd et al (2008) provided the first high-resolution sequence map of human structural variation[87]. A subsequent study then expanded the Human Genome Structural Variation clone resource by including capillary end sequencing of 4.1 million additional fosmid clones from eight additional human genomes. The combined set included 13.8 million clones derived from the genomes of six Yoruba Nigerians, five CEPH Europeans, three Japanese, two Han Chinese, and one individual of unknown ancestry[98]. This study characterized the complete sequence of 1,054 large structural variations and analyzed their breakpoint junctions to infer their potential mechanisms of origin. Three mechanisms were found to account for the bulk of germline structural variation: microhomology-mediated processes involving short (2–20bp) stretches of sequence (28%), nonallelic homologous recombination (22%), and L1 retrotransposition (19%).

2.10. The 1000 Genomes Project

The 1000 Genomes Project was initiated in 2008 with the aim of sequencing the genomes of at least 1000 individuals from different populations around the world (http://www.1000genomes.org/). The main aim of this international collaborative project was to provide a comprehensive map of human genetic variation for future disease

association studies and population genetics. As with the HapMap project, the data from this project will also be made available publicly.

Owing to the ease of high-throughput genotyping technologies, SNPs have been widely used as genetic markers in GWAS to search for disease variants. However, evidence has been accumulating to suggest that (common) SNPs alone are unlikely to account for all the heritable risk of complex diseases[99, 100]. Concurrently, the amount of data supporting associations of CNVs with complex diseases has grown[4, 7]. Similarly, the importance of rare variants in complex diseases is also increasingly being recognized[101, 102]. This indicates that future disease association studies need to interrogate non-SNP and rare genetic variants requiring a comprehensive catalogue of human genetic variations. Common SNPs have been well documented in the dbSNP, but rarer (or lower frequency) SNPs are still under-represented in the database and information on indels and structural variations is still incomplete.

The completion of the pilot phase of the 1000 Genomes Project identified approximately 15 million SNPs, 1 million short indels and 20,000 structural variations, most of which were previously unreported[38]. In addition, the location, allele frequency and local haplotype structure of these genetic variants were described. The sequencing data also enabled characterization of CNVs within heavily duplicated and near-identical regions[103]. Recently, a map of CNVs was constructed based on whole-genome sequencing data from 185 human genomes in the pilot phase of the project; this encompassed 22,025 deletions and 6,000 additional structural variations, including insertions and tandem duplications.

More importantly, approximately half of the structural variations were mapped to single nucleotide resolution, thereby facilitating the analysis of their origin and functional impact[104]. Precision in terms of the breakpoint delineation of structural variations is a prerequisite to obtaining insights into their underlying mutational mechanisms[98]. However, the nucleotide resolution analysis of the breakpoints is hampered by the low resolution of the microarrays used in previous studies.

A recent study also identified approximately two million small insertions and deletions (indels) ranging from 1bp to 10,000bp in length in the genomes of 79 humans. Interestingly, approximately half of these variants (i.e. 819,363 small indels) mapped to human genes. These small indels were frequently found in the coding exons of these genes, and several lines of evidence indicate that such variation is a major determinant of human biological diversity[27, 104]. This study also found that many of the small indels had high levels of LD with both HapMap-SNPs and GWAS-SNPs, suggesting that a proportion of these small indels have already been interrogated indirectly for their associations with complex phenotypes in GWAS through LD with the SNPs as surrogate markers. This also indicates that, in addition to SNPs and larger CNVs, small indel variation is likely to be a key factor underlying the genetics of human complex diseases and traits.

In comparison to whole-genome resequencing, which relies on a reference genome for aligning the sequence reads, *de novo* genome assembly will enable a more thorough and comprehensive detection of various genetic variations in the human genome ranging from

single nucleotide variants, small indels to large structural variations. Currently, *de novo* genome assembly is challenging and less practical because of the short sequence reads generated by NGS technologies, especially the Illumina and Life Technologies sequencing platforms. Recent studies have attempted to perform *de novo* human genome assembly using short sequence reads with limited success[105-108]. One such study showed that *de novo* assemblies were 16.2% shorter than the reference genome with thousands of coding exons being completely absent[84]. *De novo* genome assembly and haplotype phasing will eventually become more feasible with longer sequence read lengths of up to several kilobases being generated by future sequencing technologies[65, 91].

2.11. Associations of CNVs with complex diseases and traits

Although >4,000 SNPs have been reported to be associated with various human complex diseases and traits (http://www.genome.gov/26525384), these SNPs are more likely to be surrogate markers which are in strong LD with disease variants. The disease variants in most GWAS-implicated loci remain uncovered and the surrogate markers may not necessarily be tagging for SNPs, as the disease variants could also be in the form of indels and CNVs. This was well demonstrated in the discovery of a 20kb deletion located immediately upstream of the IRGM gene for Crohn's disease, and the finding of a 45kb deletion that was in perfect LD with BMI (body mass index)-associated SNPs in the NEGR1 gene[109, 110]. In addition, studies have also found evidence of correlations of CNVs with GWAS-SNPs at r^2 >0.5 suggesting the possibility of associations of CNVs with various human complex diseases and traits[30].

The genome-wide study performed by the Wellcome Trust Case Control Consortium (WTCCC) investigating the association between ~3,400 CNVs and 8 common diseases in 19,000 samples, however, did not find novel discoveries[111]. This is noteworthy in that WTCCC interrogated only a fraction of the total CNVs found in a large-scale discovery study using a high-resolution oligonucleotide array[30]. Although the genome-wide studies of the associations of common CNVs and complex phenotypes did not yield exciting novel discoveries at this stage[112-114], discoveries of rare CNVs with various complex phenotypes have been seen in schizophrenia[66, 115, 116], epilepsy[117] and severe early-onset obesity[118]. For example, these studies have found that rare structural variations that disrupt multiple genes in neurodevelopmental pathways are over-represented in schizophrenia cases than controls[66, 115].

In addition, CNVs that overlap with several genes such as FCGR3B and beta defensin genes (autoimmune and inflammatory diseases), CCL3L1 (HIV infection and rheumatoid arthritis), UGT2B17 (prostate cancer and graft versus host diseases), leptin receptor (type-2 diabetes) and TLR7 (childhood-onset systemic lupus erythematosus) have been found to be associated with various phenotypes from targeted approaches[4, 7, 119-121]. However, these associations warrant further validation as candidate-based association studies in small sample sizes have frequently been confounded by false-positive signals. The amount of evidence is expected to increase in the near future when we have a better understanding of the characteristics of CNVs and a more comprehensive map is constructed upon completion of the 1000 Genomes Project, and when more efficient and accurate methods are available to detect the CNVs for disease-association studies.

2.12. Regions of homozygosity (ROHs)

A ROH defines a continuous or uninterrupted stretch of a DNA sequence without heterozygosity in the diploid state, that is, in the presence of both copies of the homologous DNA segment. Thus, all the genetic variations, such as SNPs (biallelic marker) or microsatellites (multiallelic markers) within the homologous DNA segments have two identical alleles that create homozygosity[16]. The ROH is different from one-copy deletion (or hemizygous deletion), which could also lead to the homozygosity in genome-wide SNP genotyping data. However this is considered as a 'spurious homozygosity' because only one allele of the SNPs is present in the deleted region for one-copy deletions. Thus, the DNA fragments with only the single allele are hybridized on the genotyping array. As a result, the signal intensity of only one allele is measured and subsequently used in genotype calling, and hence it would be incorrectly labeled a homozygote genotype. Therefore, the result of 'homozygosity' is due to the absence of the other allele, instead of 'true homozygosity' where two identical alleles are present[122]. The distinction between 'true homozygosity' as opposed to 'spurious homozygosity' due to one-copy deletion is difficult to determine just by inspection of the genotype data alone. The allelic signal intensity ratio (the relative ratio of the fluorescent signals between two probes/alleles at each SNP) is needed to differentiate between the two types of homozygosity[79, 122]. Therefore, for studies that used only SNPs genotype data to identify the ROHs, i.e., to screen regions with a minimum consecutive homozygote SNPs, the possibility that some regions are caused by one-copy deletion cannot be firmly excluded, because deletions are also widespread in the human genome.

Cytogenetic abnormalities such as uniparental isodisomy can also result in homozygosity where two copies of a single parental homologous DNA segment are inherited from one parent. As such, it cannot be distinguished from homozygosity resulting from other factors such as parental consanguinity and autozygosity using the allelic signal intensity ratio as in the case of one-copy deletion. Thus, for studies that involved unrelated samples where checking the Mendelian transmission errors in the ROHs is not possible, the possibility of uniparental isodisomy leading to homozygosity cannot be definitively ruled out. Assessing the transmission errors requires data from trios or families. However, the likelihood that a considerable fraction of ROHs will be accounted for by uniparental isodisomy is low given that this cytogenetic abnormality is rare[123].

Currently, there is no consensus or standardized criteria used to define the ROH. However, previous studies have focused on regions ≥1Mb, and thus the true extent of homozygosity in the human genome could be underestimated[16, 124, 125]. More recent studies have defined a ROH at a minimum length of 500kb[19] with the intention of avoiding underestimation of the numbers of regions in the human genome. This is because shorter ROHs are now also thought to be associated with complex phenotypes. However, setting a shorter length for definition will increase the number of false positive signals, i.e., increase the sensitivity at the expense of specificity. Therefore, in discovery studies, balancing both the sensitivity and specificity when setting the criteria to identify ROHs is critical.

By focusing only on regions ≥500kb or 1Mb, the 'noise' introduced by one -copy deletions is likely to be minimal, thereby reducing the potential to cause spurious homozygosity. This is because large deletions of ≥500kb are relatively rare in the human genome, as supported by data from genome-wide mapping of CNVs studies[29, 30, 76-78]. Therefore, a critical issue to be addressed in future homozygosity mapping studies is determining the optimal cutoff of the length of the ROH to be adopted, as this will avoid over-estimating the homozygosity when the length is set too low and which can then be easily confounded by one-copy deletion of hundreds of kilobases or smaller. Although some studies have reduced the cutoff length to 500kb[19], it is still uncertain whether this new cutoff can readily reflect the true extent of homozygosity in the human genome.

It was not previously expected that the genomes of outbred populations contain ROHs of several megabases until the first few early reports in 2006 and 2007[16, 124, 125]. One study found ROHs of >5Mb in 26 of the 272 unrelated samples assessed[125]. Similarly, another study performed in Han Chinese also observed the high frequency of ROHs, where 34 out of the 515 unrelated individuals contained ROHs ranging from 2.94 to 26.27 Mb[124]. Gibson et al. (2006) studied the samples from the International HapMap Projects and identified 1,393 ROHs exceeding 1Mb in 209 unrelated HapMap individuals. Several hundred ROHs were found in each of the HapMap populations, where the average number of ROHs (>1Mb) per individual was found to be lowest in the Yoruba Ibadan Nigerian (YRI) population compared to other populations within the HapMap Phase I Project[16]. In addition to demonstrating that ROHs are common, even in the unrelated individuals from the apparently outbred populations, Gibson et al. (2006)[16] also

demonstrated the value of including diverse populations to examine the differences in ROHs. The YRI population has the least number of ROHs per individual. This finding is expected, as the populations of African ancestry are older in human history and hence have more generations and a higher number of recombination events than other populations (recombination occurs during meiosis in each generation). Recombination is an important process to interrupt the long continuous ROHs into smaller segments over generations. Population differences in ROHs have also been well documented in other studies[126].

Each of the previous studies identified a different number of ROHs per individual[124, 126-129]. These differences are likely to be reflective of technical and methodological variations such as using different genotyping platforms or SNPs data, different defining criteria and different analytical techniques. Both the genotyping platform and defining criteria can significantly influence the profile of ROHs by way of number, size, cumulative length and genomic distributions. Minor alterations in defining criteria can substantially affect the number of ROHs detected making comparisons between studies difficult. Therefore, it is critical to develop a set of standardized criteria for identifying ROHs, and to establish a database to catalog these and other regions in the human genome from published studies, similar to other databases developed for SNPs and structural variations (CNVs) such as the dbSNP and DGV respectively[2, 130]. This database will enable researchers to quickly compare their results with published data. Consensus on defining ROHs and the construction of a database to serve as a reference will help in expediting research in ROHs.

2.13. Methods of detecting ROHs

Several targeted and genome-wide molecular methods are available to detect structural variations. However, unlike structural variations, ROHs cannot be detected with technologies used in molecular genetics such as FSIH and BAC clone or CGH arrays[9, 10]. Furthermore, several new sequencing-based approaches for detecting structural variations such as PEM and DOC are also unsuitable for detecting ROHs[11, 12, 87].

The genome-wide mapping of ROHs can only be done using SNP genotyping arrays or direct sequencing. The whole-genome resequencing or *de novo* genome assembly using NGS or TGS will offer an almost complete solution to detecting most of the genetic variations, including ROHs within the human genome. However, these high-throughput sequencing technologies were not readily available until recently, and the cost is still prohibitively expensive to sequence the whole human genome in a population-based study[64, 65]. As a result, SNP genotyping arrays are the main tools for ROH mapping. The SNPs data can be used in two different ways to detect the ROHs. The first approach is to screen the whole genome using a sliding window analysis for consecutive SNPs showing homozygotes over a certain length, such as 1Mb, as implemented in PLINK[131]. Since this approach uses genotype data only, it is unable to distinguish between true homozygosity and the spurious homozygosity caused by one-copy deletion without further investigation of CNVs in the samples.

This limitation has been overcome by the second approach which relies on the signal intensity data. Two types of signal intensity data are generated by the SNPs genotyping

array: (a) the total signal intensity or log R ratio (LRR) and (b) the allelic intensity ratio or BAF. The combination of LRR and BAF can be used to determine several different states of copy numbers such as homozygous and hemizygous deletions, one-copy and two-copy duplications, and ROHs as implemented in the PennCNV algorithm. The BAF is needed to differentiate between ROHs from normal diploid copies and one-copy deletion[79]. Figure 8 illustrates the difference in LRR and BAF patterns between ROH and one-copy deletion. For the one-copy deletion, there is a decrease in the LRR in addition to the absence of heterozygosity as shown in the BAF panel. Conversely, no reduction in the LRR will be seen for ROH, but the absence of heterozygosity is observed. Most of the genome-wide studies of ROHs have used SNP genotyping arrays. In comparison, the commonly used CGH arrays for detecting CNVs produced only total signal intensity data. This renders them unusable for identifying ROHs.

The ROH and one-copy deletion were detected using the LRR and BAF information by the PennCNV algorithm (LRR: total fluorescent intensity signals from both sets of probes/alleles at each SNP, BAF: the relative ratio of the fluorescent signals between two probes/alleles at each SNP)[79]. The size of the ROH is approximately 1.06Mb (1,064,933 bases) spanning from 125374832 to 126439764 in chromsome 2. This region contains 246 markers. The size of the one-copy deletion is approximately 250kb (250,186 bases) spanning from 23994408 to 24244593 in Chromosome 22. This region contains 101 markers. The regions affected by the ROH and one-copy deletion were shaded and the blue dots represent markers in the genotyping array (Figure 8).

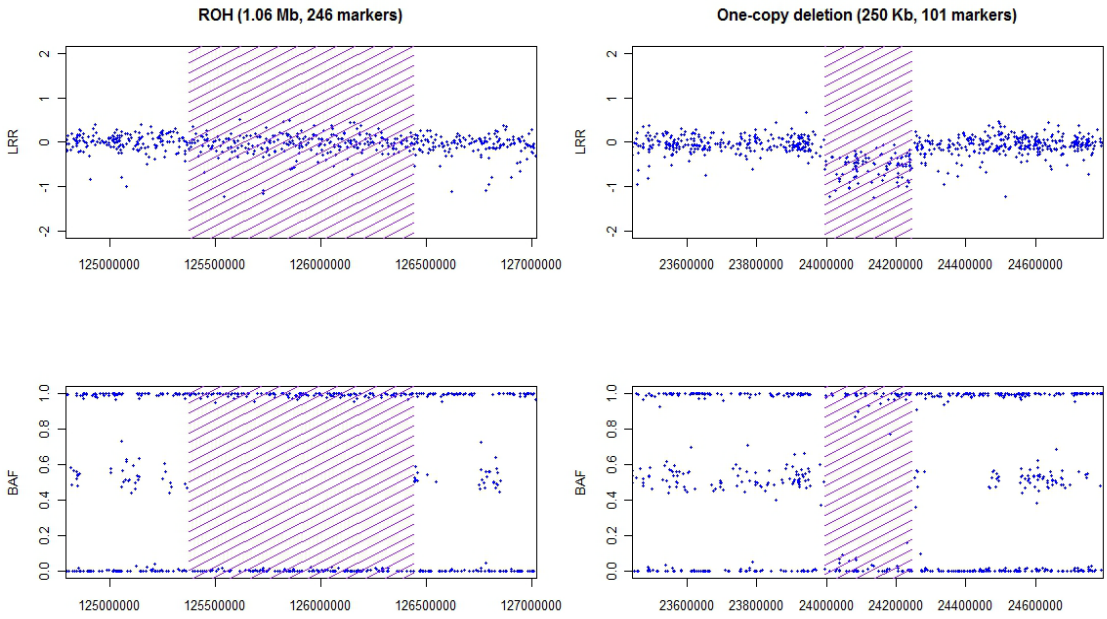


Figure 8 – Plots of the differences in the LRR and BAF patterns for the ROH (left panels) and one-copy deletion (right panels) generated from a sample derived from our previous study (Ku et al. 2010) and genotyped by the Illumina 1M Beadchip (adpated from Ku et al. (2011) Hum. Genet. 129:1-15)[132].

2.14. Associations of ROHs with complex diseases and traits

Although information regarding the extent of ROHs in the human genome is still limited compared to SNPs, indels and CNVs, the potential impact of ROHs on complex diseases and traits could also be as significant as other genetic variations. The importance of ROHs to complex phenotypes remains largely unexplored; however, several studies have shown significant differences in ROHs between cases and controls in a genome-wide investigation for schizophrenia[17], late-onset Alzheimer's disease[18] and height[19]. The idea underlying the homozygosity association approach is to uncover recessive variants contributing to complex phenotypes. The success of this approach has been demonstrated

in several studies. Nine common ROHs significantly differentiated schizophrenia cases from controls. More interestingly, four of the regions contained or were located near to the genes known to be associated with schizophrenia[17]. This proof-of-principle study has demonstrated the applications of the whole-genome homozygosity association approach to identifying genetic risk loci for complex phenotypes and represents an alternative to SNPs analysis.

Similarly, in a large-scale association study involving 837 late-onset Alzheimer's disease cases and 550 controls, a single ROH on chromosome 8 was identified, and three of the genes in the region are biologically plausible candidates[18]. Success was also achieved for complex quantitative traits such as height[19], where strong statistical evidence showing the association of one ROH with height was obtained in a total sample size of >10,000 in both the genome-wide discovery and replication studies. The height of individuals with this ROH was significantly higher (increased by 3.5 cm) than individuals without the ROH. Nonetheless, other studies produced negative results, as no evidence of homozygosity was found for bipolar disorder[133].

Many reasons can be considered for the inconsistencies with which associations of ROHs were only found in some diseases or studies but not others. This could also indicate that the effects of homozygosity on the risk of complex phenotypes may be disease or trait-dependent, for example, some quantitative traits have shown significant variance due to recessive alleles such as systolic blood pressure, total cholesterol and low-density lipoprotein cholesterol. This implies that the effects of homozygosity may be greater in

influencing the variation of these traits than others[134]. On the other hand, it could also be population-dependent since differences in homozygosity between populations have been documented. Although a number of genome-wide homozygosity association studies have been performed, the optimum study design or methods of analysis for assessing the associations of ROHs on disease risk have not been well established. This is vital before breakthrough discoveries can be made in this research area.

The aim of using the homozygosity association approach to dissect the genetics of complex phenotypes is to reveal the recessive loci that only express their effects (or increase the risk of complex diseases) in the presence of two deleterious recessive alleles, in a recessive disease model. In addition to autosomal recessive disorders, complex diseases can also be affected by recessive variants. The conventional single-SNP analysis approach applied in GWAS may not be statistically powerful enough to identify recessive alleles with small effect sizes and moreover, the recessive model is not usually tested. Until the effect of homozygosity on complex phenotypes is better understood, it is premature to make any conclusions, as the field is still in its infancy compared to association studies between SNPs and CNVs for complex diseases and traits. However, the published studies have collectively demonstrated the feasibility of using the homozygosity association approach to identify susceptibility loci for complex phenotypes and have subsequently produced encouraging results. This also underscores the need to further investigate and catalog the extent of ROHs in different populations. Similar to the other genetic variations, ROHs have the potential to become genetic markers in GWAS.

In fact, homozygosity mapping has been commonly used to identify the loci for recessive diseases in consanguineous families.

2.15. History and origin of the Singapore and Swedish populations

Singapore populations

The earliest known settlement in Singapore was documented as far back as the second century AD and was an outpost of the Sumatran Srivijaya empire, named Temasek. Between the 16th and early 19th centuries, it was part of the Sultanate of Johor. In 1613, Portuguese raiders burnt down the settlement and the island sank into obscurity for the next two centuries. Singapore had been a part of various local empires since it was first inhabited. Modern Singapore was founded as a trading post of the East India Company by Sir Stamford Raffles in 1819.

Singapore is a relatively young country with a migratory history predominantly consisting of immigrants with Chinese, Malay, and Indian genetic ancestries from neighboring countries such as China, India, Indonesia, and Malaysia (The population of Singapore, 2nd edition, Institute of South East Asian Studies, Singapore: 2007). The Chinese community mainly consists of descendents of Han Chinese settlers from the southern provinces of China, such as Fujian and Guangdong, and currently represents the dominant racial population in Singapore, accounting for 76.7% of the resident population from the Singapore Census conducted in 2000. While Han Chinese represents the largest ethnic group amongst the Chinese globally, there are a considerable number of sub-ethnicities within the Han classification with a diverse range of dialects and cultural

diversity, with established genetic heterogeneity following a geographical north–south cline. The majority of the early Chinese immigrants to Singapore belonged mainlyto the dialect groups of Hokkien, Teochew, Cantonese, Hakka, and Hainanese found predominantly in Southern China. While Malays formed the dominant race in Singapore prior to the colonization by British settlers, the proportion of indigenous Malays has been surpassed by migrant Malays from Peninsula Malaysia, as well as Javanese and Boyanese people from Indonesia. Cultural and religious similarities have resulted in intermarriages between the immigrant and local Malays, whose descendents are now collectively known as Malays and account for 13.9% of the Singapore population. The British colonization of Singapore also brought Indian migrants from the Indian subcontinent, with the majority consisting of Telagus and Tamils from southeastern India and a minority of Sikhs and Pathans from north India. The origins of the Indians in Singapore comprises of people with paternal ancestries tracing back to the Indian subcontinent, and as a race, Indians represent 7.9% of the Singapore population.

Swedish population

The first inhabitants to the area of present-day Sweden travelled from Central Europe after the ice age. For millennia, the country was sparsely inhabited by hunter-gatherer populations until the slow adoption of agriculture and ceramics that began around 4000 BC in southern Sweden. While the southern parts of the country developed strong contacts with the Germanic culture, the north associated to Finland and Karelia with a common culture covering the entire northern Fennoscandia. This culture has sometimes been suggested to be ancestral to the indigenous Sami population still inhabiting the area.

Sweden was not united under one ruler until the 11th century, and the traditional division to the southern Götaland, central Svealand, and northern Norrland is still widely known despite lacking any official status. There have been long-standing contacts with the neighboring populations, with Norwegian influence in western Sweden, Danish in the south, and Finnish in the north. The population density has been highest in Southern and Central Sweden, while in Norrland the population is centered on the eastern coast and in river valleys whereas the mountaineous regions in the northwest are largely uninhabited.

Genetically the Swedes have appeared relatively similar to their neighboring populations - for example the Norwegians, Danish, Germans, Dutch and British - both in a classical study based on a small number of autosomal markers and in the recent genome-wide studies. In contrast, the Finns seem to be an exception to this rule: they do not appear genetically very close to the Swedes although they are geographically nearby. However, the Finns tend to show inflated genetic distances relative to the European populations in general, not only relative to the Swedes.

CHAPTER 3 – AIMS

My thesis was motivated by several key concepts highlighted in Chapter 2. The first is that the roles of CNVs and ROHs in human complex diseases and traits are increasingly being recognized. The second is that limited CNV and ROH data is available for healthy individuals in Singapore and Swedish populations, therefore, the population characteristic of CNVs and ROHs in these populations are largely unknown. Thirdly is that comparisons among different populations are challenging because different studies have applied different analytical approaches. As a result, comparison of the final results from different studies was plausible.

This thesis is divided into four studies with a specific primary aim for each study:

Study I: To identify CNVs and to study their population characteristics in Singapore populations

Study II: To identify CNVs and ROHs, and to study their population characteristics in the Swedish population

Study III: To study population differences of CNVs and CNPs between HapMap III and Singapore individuals

Study IV: To identify ROHs and to study their population characteristics in Singapore populations

CHAPTER 4 – MATERIALS AND METHODS

The chapter summarizes the materials and methods sections of the four studies (Study I – IV). More details of the materials and methods of these four studies can be found in the full research publications attached in the appendix.

4.1. Study I (Genomic copy number variations in three Southeast Asian populations)

4.1.1. Samples

The genomic DNA samples used in this study were extracted from peripheral blood samples of individuals recruited under a previous project approved by the National University of Singapore-Institutional Review Board (NUS-IRB) (Reference Code: 07-199E). The project recruited 600 unrelated and apparently healthy individuals (without clinical disease) from the three populations in Singapore (Chinese, Malay, and Asian Indian) for identification and characterization of novel genetic variants in drug transporter and ion channel genes.

The DNA samples for this study (n=292) were chosen using stratified random sampling from the pool of samples to ensure approximate equal representation of each population and gender. The selected samples were genotyped using the Illumina Human 1M Beadchip and Affymetrix Genome-Wide Human SNP Array 6.0. The Illumina array was used to detect CNV loci in the study population, whereas the Affymetrix array was used to characterize loci that were independently replicated. The samples were anonymized, but basic demographic data such as gender, age, and self-reported ethnicity were retained. There were a total of 99 Chinese, 98 Malay, and 95 Indian individuals in the final

genotyped samples. Population membership was ascertained on the basis that all four grandparents belonged to the same population group.

Genotyping was performed using Illumina 1M for all DNA samples as per the manufacturer's protocol. The Singapore Genome Variation Project applied several filtering criteria to identify and remove unsuitable samples that we similarly adopted[13]. A total of 273 samples were used in the subsequent CNV calling after removing samples on the basis of a high SNP missingness rate (≥2%), excessive heterozygosity or cryptic relatedness by excessive identity-by-states, and based on samples displaying either evidence of admixture or clear evidence of discordance between self-reported and genetically inferred population membership.

4.1.2. CNV detection using PennCNV

We used the PennCNV algorithm to identify both deletions and duplications in the 22 autosomes and the X chromosome[79, 135]. We applied a set of filtering criteria, as recommended by the algorithm, to exclude poor quality samples, which resulted in a further seven samples being excluded because their intensity data failed to conform to these criteria. Our final set for analysis included 266 samples consisting of 93 Chinese, 88 Malay, and 85 Indian individuals. For each sample, PennCNV returned a list of regions with an abnormal copy number with their associated confidence scores. The score is a log Bayes Factor that measures the likelihood that the region harbors an abnormal copy number. A confidence score of 10 or larger has been suggested as a threshold to classify reliable CNV calls. In our case, we retained all CNVs called with confidence

scores higher than the median confidence score. This median score was calculated based on the confidence scores of CNVs detected in all individuals, and its value was approximately 12.

4.1.3. Analyses

Construction of CNV loci - As CNV regions called by PennCNV tend to overlap, we merged these regions into discrete, non-overlapping loci with the boundaries of each locus determined by the union of all CNV regions that belong to that particular locus[28]. If both deletions and duplications were observed in a particular locus, two separate loci were identified for each form of CNV.

Replicated CNV loci - To validate the CNV loci identified using the Illumina 1M platform, we genotyped the same 266 samples using the Affymetrix SNP Array 6.0. The signal intensity data were analyzed using PennCNV with the same parameters as used for the Illumina samples. A CNV locus found using the Illumina platform was considered to be replicated if there was at least one overlapping CNV locus found using the Affymetrix platform. A CNV locus detected using Illumina was considered replicated if it shared at least 50% of its length with a CNV locus detected using the Affymetrix platform.

Novel CNV loci - To identify CNV loci that are novel, we compared our results with the CNV loci published in the DGV. We classified a particular CNV locus identified using Illumina and subsequently replicated using the Affymetrix platform as novel if it did not share at least 50% of its length with any established CNV loci in the DGV database.

Population differentiation of CNV loci - We used a Vst statistic[28] to describe the overall population differentiation due to CNVs. We also compared the distribution of integer copy numbers across the three ethnic groups using the Fisher's exact test. A p-value<0.001 from this test was used to identify loci that segregated at different frequencies across the three populations.

Mapping against annotated genes - We used the UCSC gene annotation (http://genome.ucsc.edu/) to identify genes that are located within or partially overlap with CNV loci.

4.2. Study II (A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals)

4.2.1. Samples

One hundred randomly selected healthy Swedish individuals who volunteered as controls in case controls studies were studied. Peripheral blood samples of the participants were drawn for genomic DNA extraction and stored at the Karolinska Biobank. Identities of the participants were kept anonymous and no personal identifiers were used. All 100 samples were genotyped using the Affymetrix SNP Array 6.0 as per the manufacturer's protocol. Two samples were removed from further analysis because their genotype call rates were below 98%. The remaining 98 samples were used for CNV detection.

4.2.2. CNV-detection algorithms and analyses

CNV calling using PennCNV - We used two CNV-detection algorithms, namely PennCNV and Birdsuite for both comparison and validation[79, 80]. This study focused only

on the CNVs in the 22 autosomes. After PennCNV detection, we applied a set of filtering criteria, as recommended by the algorithm, to exclude samples with poor quality signal intensity data. This resulted in the exclusion of 11 samples with the final set for analysis consisting of 87 samples. For each sample, PennCNV generated a list of CNVs with their confidence scores. A confidence score of 10 or larger has been recommended as the threshold to classify reliable CNVs. Therefore, we retained all CNVs called with confidence scores ≥10 for subsequent analyses.

Construction of CNV loci using PennCNV output - The CNVs called by PennCNV were shown to overlap across samples. Thus, we merged or grouped these individual CNV calls into discrete, non-overlapping loci with the boundaries of each locus determined by the union of all CNVs that belonged to that particular locus. This was performed using the methods that we have previously developed[136]. We classified the status of these CNV loci into three categories, 'del' (loci containing deletions), 'dup' (loci containing duplications) and 'del/dup' (loci containing both deletions and duplications).

CNP calling using Canary (Birdsuite) - Birdsuite software was also used to analyze the Affymetrix SNP Array 6.0 data. There are two components in the software for detecting copy number changes, namely Canary and Birdseye. Canary is used to determine the integer copy number at each of the predefined 1316 CNPs. These 1316 CNPs are distributed in all the autosomes and sex (X and Y) chromosomes. However, 25 CNPs located in the sex chromosomes were removed because the CNP calling in these chromosomes was less accurate. Thus, the results reported in this study were comprised of only 1291 CNPs in the 22 autosomes. Confidence statistics generated for the CNPs

were also used to identify poor quality calls and only integer copy numbers detected with high confidence, as recommended by the software (confidence score >0.1), were used for subsequent analyses.

Correlation analysis of CNPs **-** We performed a correlation analysis of CNPs and the nearby SNPs. For each of the 1291 CNPs, SNPs within a 200kb window from the start- and end-position of the CNP were considered. We used the squared Pearson's correlation (r^2) for correlation analysis. The genotype calling of the Affymetrix SNP Array 6.0 was done using Birdsuite. In addition, to investigate the potential associations of CNPs with human diseases and traits, the same methods of r^2 calculations for the 1291 autosomal CNPs and the SNPs that were identified by GWAS was adopted. The list of GWAS-SNPs was downloaded from the National Human Genome Research Institute (NHGRI) website (http://www.genome.gov/gwastudies/) on 26 October 2010.

CNV calling using Birdseye (Birdsuite) **-** In addition to PennCNV we also used another algorithm, Birdseye, to analyse the same set of data as different algorithms tend to have different sensitivities and specificities for detection of CNVs in different regions throughput the genome. As such, CNV loci detected by PennCNV and Birdseye can be cross-validated. We used the Birdseye component in Birdsuite to detect additional CNVs throughput the genome, which was not restricted to the 1316 predefined CNPs. Similarly, only CNVs in autosomal chromosomes were used because of the inaccuracy of Birdseye in the sex chromosomes. CNVs with low confidence, as recommended by the software (confidence score $\leq$5), were removed from subsequent analysis.

Construction of CNV loci using Birdseye output - We also constructed CNV loci based on the Birdseye output using similar methods as were applied to the PennCNV output.

Comparison of CNV loci detected by PennCNV and Birdsuite - The CNV loci identified by PennCNV and Birdseye were compared as a 'validation' step. We used a 'reciprocal 50% overlapping' method to compare the CNV loci detected by these two algorithms and considered a CNV locus 'found' by both algorithms when this locus was detected in both PennCNV and Birdseye with an overlap of ≥50% of their lengths.

Novel CNV loci - To identify novel CNV loci, we compared the CNV loci detected by PennCNV and Birdseye with the data from the DGV. A CNV locus identified by PennCNV and Birdseye was considered novel if it did not share at least 50% of its length with any CNV loci cataloged in the DGV.

4.2.3. Comparison with HapMap Phase III populations

The CEL-files of the Affymetrix SNP Array 6.0 for the seven populations in the HapMap Phase III project were downloaded from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). The HapMap phase III populations studied are people of African ancestry in the southwestern USA (ASW), the Chinese community in Metropolitan Denver, Colorado, USA (CHD), Gujarati Indians in Houston, Texas, USA (GIH), the Luhya in Webuye, Kenya (LWK), people of Mexican ancestry in Los Angeles, California, USA (MEX), the Maasai in Kinyawa, Kenya (MKK) and the Tuscans in Italy (TSI). All the samples were analysed using Canary similar to the analysis of the Swedish population. Only unrelated samples were included in our study,

i.e., family-related samples were removed using the 'relationships' file provided by the International HapMap Project. After the sample exclusion step, a total of 594 unrelated samples from the seven HapMap III populations were analysed: ASW (n=52), CHD (n=89), GIH (n=89), LWK (n=90), MEX (n=53), MKK (n=132) and TSI (n=89). We performed principal component analysis (PCA) to compare the Swedish population with the HapMap Phase III populations using the CNP output generated by Canary.

4.2.4. ROH-detection algorithms and analyses

In addition to CNVs, we also detected ROHs using PennCNV in the 22 autosomes of the 87 Swedish individuals. However, we only focused on ROHs≥500Kb. For each of these we confirm that they are ROHs by determining the genotypes of the SNPs that fell within each region. We then calculated the percentage of heterozygosity (number of heterozygotes/total number of heterozygotes and homozygotes). We also calculated the percentage of missingness genotypes (number of missingness/total number of SNPs in each ROH). We excluded ROHs with >2.5% heterozygosity and >1% missingness. For the remaining ROHs, we also ensured a density of one SNP per 10kb.

4.3. Study III (Copy number polymorphisms in new HapMap III and Singapore populations)

4.3.1. Samples

In this study, a total number of 265 Singapore samples (93 Chinese, 88 Malays and 84 Indians) genotyped using the Affymetrix SNP Array 6.0 were available for analysis (please see section 4.1.1. for Singapore samples). We also studied the HapMap III

populations for comparisons (see section 4.2.3. for HapMap III populations). All the samples were analysed by Birdsuite.

4.3.2. CNP calling using Canary (Birdsuite)

All the Singapore and HapMap III samples were analysed by Canary. (Please see section 4.2.2. for more descriptions of CNP calling using Canary).

4.3.3. Correlation analysis

The correlation analysis of CNPs performed in this study differed slightly from Study II, because we were restricted to biallelic CNPs with a MAF $\geq$5%. All the correlation analyses of CNPs and nearby SNPs were done separately for each of the 10 populations. (See section 4.2.2. for further descriptions of correlation analysis of CNPs.)

4.3.4. Copy number loci calling using Birdseye (Birdsuite) and validation

The Birdseye component in Birdsuite was used to detect additional copy number loci located outside the 1316 CNPs in the 10 populations. (See section 4.2.2. for more descriptions of CNV calling using Birdseye and construction of novel copy number loci using Birdseye output.)

4.4. Study IV (Regions of homozygosity in three Southeast Asian populations)

4.4.1. Samples

The identical set of samples (268 samples) studied in the Singapore Genome Variation Project[13] was used in this study. (Please see section 4.1.1. for further description of these Singapore samples.)

4.4.2. ROH detection

The signal intensity data from the Illumina and Affymetrix arrays of these 268 samples were analyzed to identify ROHs in the 22 autosomes by the PennCNV algorithm. As such, ROHs detected by the Illumina and Affymetrix arrays could be cross-validated. We focused only on ROHs≥500Kb and confirmed each of the ROHs by determining the genotypes of the SNPs that fell within each region. The genotype data for each of the three populations was obtained from the Singapore Genome Variation Project (http://www.nus-cme.org.sg/sgvp/). There are approximately 1.58 million genotypes or SNPs in the '*QC+Mono file*' for each population. This quality control exercise was done separately for each of the populations using their own specific genotype file.

The percentage of heterozygosity (number of heterozygotes/total number of heterozygotes and homozygotes) was then calculated. We also calculated the percentage of missingness genotypes (number of missingness/total number of SNPs in each ROH). We removed ROHs with an arbitrary cutoff of >2% heterozygosity and >1% missingness. We also ensured a density of one SNP per 10Kb.

4.4.3. Validation

Similar to our previous study on CNVs (Study I), we used the Illumina array as the detection platform for ROHs in this study. The results were then validated using the Affymetrix array at the 'sample-level', i.e., ROHs detected by the Illumina and Affymetrix arrays were cross-validated sample-by-sample. This cross-validation was performed by overlapping the ROHs detected by the Illumina array against the ROHs identified by the Affymetrix array at a 50% overlap cutoff point.

4.4.4. Construction of 'common ROH loci'

We also clustered the individual ROHs into discrete common loci termed 'common ROH loci', which were identified using statistical methods developed by our group[136].

4.4.5. Population comparisons

For each of the common ROH loci, we tested the difference between the three populations (using the Fisher's exact test with p-values corrected using the false discovery rate (FDR)). We also generated a PCA plot using the common ROH loci. Since the Illumina array was used as the detection platform, these analyses (i.e., identification of common ROH loci and the PCA plot) were restricted to ROHs detected by the Illumina array.

4.5. Summary for Study I – IV

Table 5 – Summary of samples, genotyping platforms, detection algorithms and data used and generated by Study I - IV

Study	Final sample size	Genotyping platform	Detection algorithm	Data	Remark
I	266 Singapore samples	Affymetrix SNP Array 6.0 and Illumina 1M	PennCNV	• CNV data generated by PennCNV for both Affymetrix and Illumina arrays	• Illumina CNV data was used for detection and Affymetrix CNV data was used for validation
II	87 Swedish samples and 594 HapMap III samples	Affymetrix SNP Array 6.0	PennCNV and Birdsuite (Canary and Birdseye)	• CNV data generated by PennCNV and Birdseye • Canary determined the integer copy number of the 1316 predefined CNPs • ROH data generated by PennCNV	• The Swedish CNV data generated by PennCNV and Birdseye were cross-validated • The HapMap III samples were only analysed by Canary and used for PCA to provide some preliminary insights to the population differences (because this study focused on a detailed study of CNVs and ROHs in the Swedish population)
III	265 Singapore samples and 594 HapMap III samples	Affymetrix SNP Array 6.0	Birdsuite (Canary and Birdseye)	• CNV data generated by Birdseye • Canary determined the integer copy number of the 1316 predefined CNPs	• The HapMap III samples were analyzed in detail in this study for comparisons with Singapore populations
IV	268 Singapore samples	Affymetrix SNP Array 6.0 and Illumina 1M	PennCNV	• ROH data generated by PennCNV	• Illumina ROH data was used for detection and Affymetrix ROH data was used for validation

CHAPTER 5 – RESULTS

This chapter summarizes the major results of each of the four studies (Study I – IV). The complete set of the results of these four studies are reported in the full research papers attached in the appendix.

5.1. Study I (Genomic copy number variations in three Southeast Asian populations)

We discovered approximately 45 CNVs per individual with a ratio of deletions to duplications of approximately 4:1. The majority of individuals had 20–60 CNVs in their genome (Figure 9).

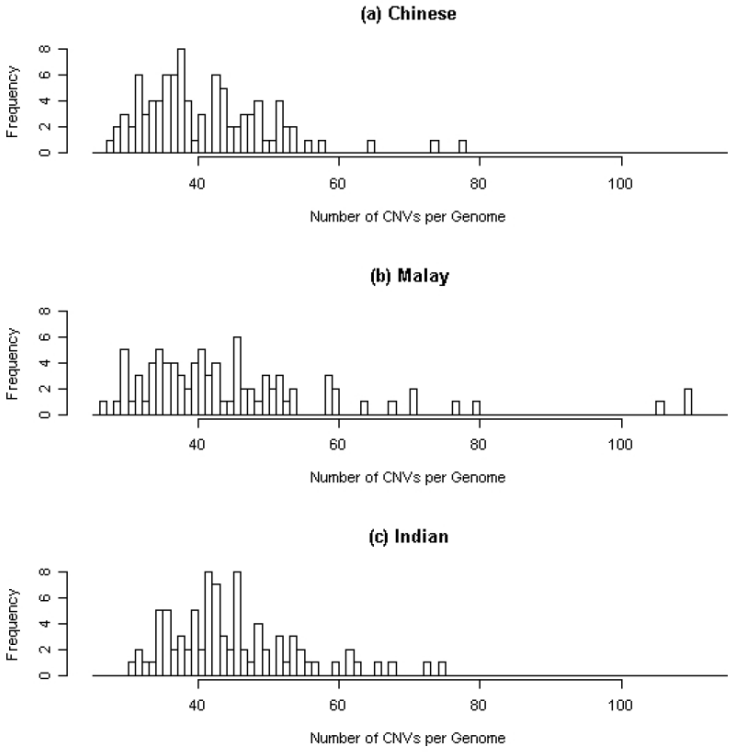


Figure 9 – Number of CNVs per genome and their frequency in each of the three Singapore populations (adapted from Ku et al. (2010) Hum. Mutat. 31:851-857)[78].

We merged overlapping individual regions and identified 1,841 deletion and 732 duplication loci in these populations (Figure 10). Less than 10% of deletion loci and less than 5% of duplication loci could be considered common (population frequency >5%) across the three populations. Using the Affymetrix 6.0 platform we identified 1,514 deletions and 560 duplications loci.

Comparing the CNV loci we found 752 (40.8%) deletion and 422 (57.8%) duplication loci identified using the Illumina platform that were replicated by the Affymetrix platform. Singletons constituted the majority of CNV loci that were not replicated, with 64.8% of non-replicated deletion loci and 71.6% of non-replicated duplication loci which were singletons.

We discovered that approximately 40% of the replicated deletion loci were novel, as only 467 out of 752 loci were found in the DGV. Similar to the deletion loci, approximately 37% of duplication loci were found to be novel (156 out of 422 loci).

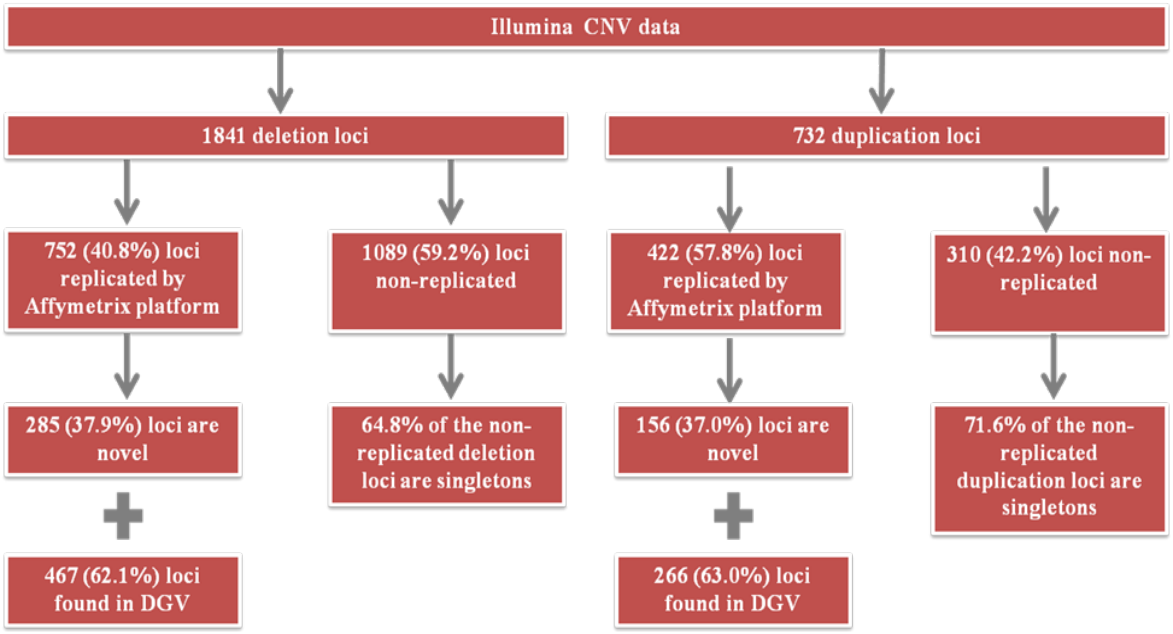


Figure 10 – Number of loci replicated by the Affymetrix platform and novel loci not found in the DGV.

The median Vst statistic between Chinese and Malay populations, computed across 1,174 replicated loci was 0.016 and lower than the corresponding comparisons for Chinese and Indian populations (median = 0.035) and Indian and Malay populations (median= 0.028). Over the whole genome, the Indian population was more differentiated from Chinese and Malay populations. We identified 27 deletion loci that segregated at significantly different frequencies across the three populations.

Compared to duplication loci, we found an appreciably lower percentage of deletion loci that overlap with known genes or uncharacterized transcripts in the UCSC database (p-value <0.001). Most of the 367 deletion loci that overlap with UCSC genes are rare (66.8%); however, there are 66 (18.0%) deletion loci with an intermediate frequency between 1% and 5%, and 56 (15.3%) common deletion loci that also overlap UCSC

genes. Likewise, we found that 229 (72.7%) out of 315 duplication loci overlapping with the UCSC genes are rare, 23.5% are intermediate, and 3.8% are common (Table 6). In addition, we also found several complex disease-associated genes overlapping with the common CNV loci (population frequency >5%), including FCGR3B, beta-defensin genes, UGT2B17, CCL3L1 and a number of drug-related genes such as CYP2A6 and CYP2A7.

Table 6 – The proportion of deletion and duplication loci overlapping with the UCSC database with varying population frequencies

Population frequency	Deletion loci overlap UCSC genes	Duplication loci overlap UCSC genes
Rare (<1%)	66.8%	72.7%
Intermediate (1-5%)	18.0%	23.5%
Common (>5%)	15.3%	3.8%

5.2. Study II (A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals)

In this study, an average of approximately 36 CNVs per individual with a ratio of deletions to duplications of approximately 2.6:1 was discovered for the Swedish individuals. The number of CNVs per individual ranged from 22 to 65.

We merged overlapping CNVs to construct CNV loci and identified 623 loci where 476 loci contained deletions (‘del-loci’), 102 loci contained duplications (‘dup-loci’) and 45 loci contained both deletions and duplications (‘del/dup-loci’) (Table 7). Of the 623 CNV loci, 268 loci were detected in≥2 individuals. Among the high frequency CNV loci (i.e. ,

the loci that were detected in multiple individuals) several overlapped with disease-related genes. For example, a deletion locus overlapping with WWOX (a tumor suppressor gene) was detected in 24 of the 87 individuals (27.6%), and a deletion locus encompassing GSTT1 was deleted at a population frequency of 13.8%. Additionally, the proportion of del-loci encompassing the UCSC genes (28.36%) was much lower than dup-loci (50.00%) overall.

In an effort to validate the 623 CNV loci constructed from the PennCNV output, we compared them with the CNV loci detected by Birdseye. We found 196 loci (31.46%) with ≥50% reciprocal overlap with the Birdseye data and the status of 'del', 'dup' and 'del/dup' of the 196 loci were consistent with the Birdseye data. For the remaining 427 CNV loci that were not confirmed by Birdseye data, we found 247 loci that were cataloged in the DGV. Therefore, by applying two different methods of validation, 443 (71.1%) of the 623 CNV loci detected by PennCNV were considered reliable in this study (Table 7).

Table 7 – Summary statistics of CNV loci constructed from PennCNV output

Summary statistics of CNV loci (PennCNV output)	Total	Del	Dup
Number of CNV loci	623	476 (76.40%) *	102 (16.37%) *
Number of CNV loci detected in ≥2 individuals	268 (43.02%) **	194 (40.76%) **	29 (28.43%) **
Sum of the length of loci	61.52Mb	19.83Mb	25.80Mb
Average length per locus	98.75Kb	41.66Kb	252.93Kb
Average number of markers per locus	58	34	141
Size distribution			
<10Kb	141 (22.63%)	132 (27.73%)	6 (5.88%)
≥10Kb to <50Kb	265 (42.54%)	236 (49.58%)	17 (16.67%)
≥50Kb to <100Kb	79 (12.68%)	54 (11.34%)	21 (20.59%)
≥100Kb to <500Kb	110 (17.66%)	52 (10.92%)	43 (42.16%)
≥500Kb	28 (4.49%)	2 (0.42%)	15 (14.71%)
Overlapping with DGV			
CNV loci that overlap	388 (62.28%)	298 (62.61%)	54 (52.94%)
CNV loci that did not overlap	235 (37.72%)	178 (37.39%)	48 (47.06%)
Overlapping with UCSC genes			
CNV loci that overlap	202 (32.42%)	135 (28.36%)	51 (50.00%)
CNV loci that did not overlap	421 (67.58%)	341 (71.64%)	51 (50.00%)
Overlapping with CNV loci from Birdseye data and consistent in CNV status, i.e., del/dup/del+dup			
CNV loci that overlap	196 (31.46%)	160 (33.61%)	30 (29.41%)
CNV loci that did not overlap	427 (68.54%)	316 (66.39%)	72 (70.59%)

*The percentage was calculated by dividing 623 loci

**The percentage was calculated by dividing 623, 476 and 102 loci respectively

Note: as there are only 45 CNV loci (7.22%) with status del+dup, the summary statistics of these loci were not shown in the table

(This table was adapted from Teo et al. (2011) J. Hum. Genet. 56:524-533)[137]

Approximately 49.81% of the 1,291 autosomal CNPs were non-polymorphic in the Swedish population. Numerous CNPs were found to overlap with important known disease- or pharmacogenetics-related genes (Table 8). To investigate the potential role of CNPs in the etiology of complex diseases or traits, we computed the r^2 between CNPs and the SNPs from the NHGRI GWAS Catalog. Of the >3,000 GWAS-SNPs that were found to be associated with various complex diseases and traits, only 8 GWAS-SNPs were found to be in strong correlation with 6 CNPs (Table 9).

Table 8 – CNPs that overlap with important and known disease- and pharmacogenetic-related genes

CNP ID	CN=0	CN=1	CN=2	CN=3	CN=4	Frequency	Chr.	Start	End	Length	UCSC gene (disease/trait)
118	0	1	70	0	1	2.78	1	159778034	159906183	128149	FCGR3A,FCGR2B,FCGR2C,FCGR3B (autoimmune or inflammatory diseases)
11164	0	1	83	2	0	3.49	6	162658558	162660430	1872	PARK2,parkin (Parkinson's disease)
530	1	10	71	0	0	13.41	3	190846372	190847332	960	TP63 (cancers)
147	3	31	53	0	0	39.08	1	194997658	195068695	71037	CFHR3,CFHR1 (age-related macular degeneration)
603	8	33	46	0	0	47.13	4	69043083	69168574	125491	UGT2B17 (prostate cancer, graft-versus-host disease)
2560	15	36	34	0	0	60.00	22	22680529	22726814	46285	GSTT1 (phase II metabolizing enzyme)
2203	20	46	17	1	0	79.76	16	76929941	76942266	12325	WWOX (cancers)
109	33	39	15	0	0	82.76	1	150822330	150853218	30888	LCE3C,LCE3B (psoriasis)
2559	32	41	1	0	0	98.65	22	22613016	22670785	57769	GSTT2,GSTT2B, GSTTP1 (phase II metabolizing enzyme)
88	46	1	0	0	0	100.00	1	110025907	110044476	18569	GSTM2,GSTM1 (phase II metabolizing enzyme)

(This table was adapted from Teo et al. (2011) J. Hum. Genet. 56:524-533)[137]

Table 9 – Correlation between CNPs and GWAS-SNPs at r^2>0.5

CNP	Chr.	Start position	End position	Length	GWAS-SNP	r2 value	Gene	Complex disease/trait
60	1	72541504	72583736	42232	rs2815752	1	NEGR1	BMI
147	1	194997658	195068695	71037	rs6428370	0.647399825	Intergenic	Acute lymphoblastic leukemia (childhood)
333	2	203608045	203610291	2246	rs6725887	0.84632626	WDR12	Myocardial infarction (early onset)
874	5	150185693	150198797	13104	rs13361189	0.927251567	IRGM	Crohn's disease
874	5	150185693	150198797	13104	rs1000113	0.927251567	IRGM	Crohn's disease
874	5	150185693	150198797	13104	rs11747270	0.927251567	IRGM	Crohn's disease
877	5	155409350	155415307	5957	rs4704970	1	SGCD	Multiple sclerosis
933	6	32539530	32681749	142219	rs3129934	0.664781909	HLA-DRB1	Multiple sclerosis

(This table was adapted from Teo et al. (2011) J. Hum. Genet. 56:524-533)[137]

We performed a PCA to compare the Swedish population with the HapMap Phase III populations using the CNP output generated by Canary. The PCA showed distinct clusters for populations with different ancestries. The first two principal components (PC1 and PC2) separated the African (ASW, MKK and LWK) and non-African (CHD, GIH, MEX, SWED and TSI) populations (Figure 11, top left). This suggested that the CNP profiles of the African populations were substantially different from the non-African populations. From the second and fourth principal component (PC2 and PC4), three distinct clusters were observed (Figure 11, top right). The three African populations remained as a distinct cluster; however, CHD was separated from the European populations (MEX, SWED and TSI) and the Gujarati Indians (GIH). This indicated that the CNP profile of Gujarati Indians in Houston (Texas, USA) resembles the European populations. PCA was also performed by restricting only the 'European cluster' populations (GIH, MEX, SWED and TSI) in PC2 versus PC4. More interestingly, we also found that the CNP profile of the Swedish population was substantially different from the other populations such as GIH and MEX, but it was also appreciably different from TSI (Figure 11, bottom left). These differences further justify the need to detect and characterize the CNV/CNP profile of the Swedish population.

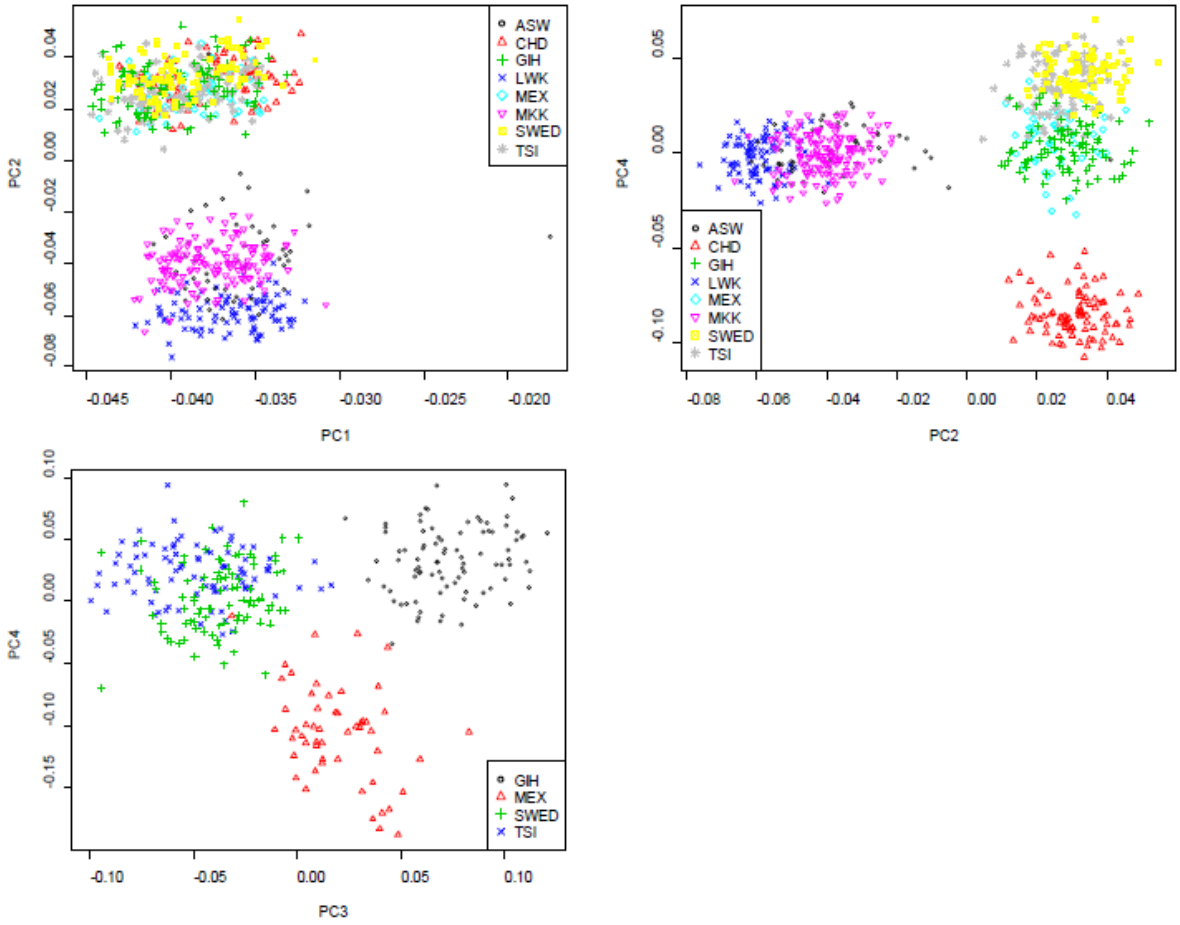


Figure 11 - PCA comparing the Swedish and HapMap III populations
(This figure was adapted from Teo et al. (2011) J. Hum. Genet. 56:524-533)[137]

By restricting ROHs to≥500Kb, a total of 14 ,815 regions were found in the 87 Swedish individuals with an average of 170 ROHs. The number of ROHs ranged from 105 to 220. The majority of these ROHs were less than 1Mb in length. However, by restricting ROHs to ≥1Mb, 2 ,814 ROHs with an average of 32 ROHs per individual were found. The median size of the ROHs was approximately 686Kb, with the largest ROH spanning a length of approximately 25Mb in chromosome 11. The sum of the length of ROHs in each individual (i.e., the total length of all the ROHs in one individual) was then

computed. It ranged from approximately 87Mb to 179Mb with a median and mean of approximately 141Mb. This finding suggests that on average, 141Mb or 4.92% of the human genome (2,867Mb) is homozygous in these Swedish individuals.

5.3. Study III (Copy number polymorphisms in new HapMap III and Singapore populations)

This study focused on comparing the Singapore and HapMap III populations. The Singapore populations were similar to the HapMap III populations of non-African descent in terms of the proportion of non-polymorphic loci and loci with varying population frequencies. More than half of the CNPs were non-polymorphic in the Singapore and HapMap III populations of non-African descent (i.e., CHD, GIH, MEX and TSI). This was in contrast to the populations of African descent (i.e., ASW, LWK and MKK), where only 26.41% to 37.72% of the CNPs were not polymorphic.

We identified 698 CNPs (FDR <0.01) that differ between populations with several loci encompassing known disease- or traits-associated or pharmacogenetic-related genes. There was a large inter-population difference in the frequencies of some of the CNPs overlapping these genes (Table 10).

The numbers of CNPs that showed significant differences (FDR <0.01) in pairwise comparisons of the 10 populations are shown in Table 11. Only 19 CNPs showed significant differences between Sing-Chinese and CHD, and 12 CNPs between Sing-

Indian and GIH suggesting that the CNPs profile in the two Chinese and two Indian populations are very similar.

To investigate the potential role of CNPs in the etiology of complex diseases or traits, we computed the r^2 between CNPs and the SNPs in the NHGRI GWAS Catalog. Out of the >2,500 GWAS-SNPs that have been found to be associated with various complex diseases and traits, only 17 GWAS-SNPs were found to be in strong correlation with 12 CNPs (Table 12).

Table 10 – CNPs (FDR <0.01) that overlap with known disease-associated or pharmacogenetic-related genes

CNP	Gene	Sing-Chinese	Sing-Malay	Sing-Indian	ASW	CHD	GIH	LWK	MEX	MKK	TSI
CNP2203	WWOX	2.38*	7.32	51.81	66.67	0.00	48.86	40.00	67.31	28.35	68.18
CNP340	ERBB4	0.00	2.33	12.05	7.69	0.00	17.24	0.00	0.00	0.00	4.49
CNP530	TP63	64.84	48.24	27.38	30.77	68.54	31.82	31.82	9.62	32.06	6.90
CNP2118	ADAMTSL3	67.05	46.84	11.54	38.46	51.19	4.49	49.40	24.32	48.80	19.51
CNP147	CFHR3,CFHR1	11.83	12.64	53.57	59.62	15.73	58.43	59.09	18.87	42.42	43.82
CNP2560	GSTT1	96.77	85.06	56.63	72.00	92.13	70.79	75.56	71.70	80.15	67.06
CNP603	UGT2B17	100.00	95.40	82.14	48.08	98.88	86.42	63.33	58.49	67.18	58.43
CNP2415	CYP2A6	18.89	36.25	5.13	6.00	23.86	11.49	8.05	2.04	8.80	4.60

*Population frequency (%) = deletion frequency + duplication frequency

(This table was adapted from Ku et al. (2011) J. Hum. Genet. 56:552-560)[138].

Table 11 - The number of CNPs that showed significant differences (FDR <0.01) in the pairwise comparisons among the 10 populations

Population	Sing-Chinese	Sing-Malay	Sing-Indian	ASW	CHD	GIH	LWK	MEX	MKK	TSI
Sing-Chinese	-	6	84	137	19	106	209	81	199	141
Sing-Malay	-	-	46	125	26	72	197	59	180	126
Sing-Indian	-	-	-	93	88	12	186	32	147	54
ASW	-	-	-	-	132	95	13	69	18	90
CHD	-	-	-	-	-	113	196	77	192	130
GIH	-	-	-	-	-	-	170	35	155	52
LWK	-	-	-	-	-	-	-	123	33	176
MEX	-	-	-	-	-	-	-	-	97	27
MKK	-	-	-	-	-	-	-	-	-	146
TSI	-	-	-	-	-	-	-	-	-	-

(This table was adapted from Ku et al. (2011) J. Hum. Genet. 56:552-560)[138].

Table 12 – Correlation between CNPs and GWAS-SNPs at r^2 >0.5 in 10 populations

CNP	Chr.	Start/End position	GWAS-SNPs	GWAS-SNPs position	Population	Gene	Disease/Trait
60	1	72541504 72583736	rs2815752	72585028	Sing-Chinese, Sing-Malay, Sing-Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	NEGR1	BMI
874	5	150185693 150198797	rs13361189	150203580	Sing-Chinese, Sing-Malay, Sing-Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	IRGM	Crohn's disease
874	5	150185693 150198797	rs1000113	150220269	Sing-Chinese, Sing-Malay, Sing-Indian, CHD, MEX, MKK, TSI	IRGM	Crohn's disease
874	5	150185693 150198797	rs11747270	150239060	Sing-Chinese, Sing-Malay, Sing-Indian, ASW, CHD, GIH, MEX, MKK, TSI	IRGM	Crohn's disease
877	5	155409350 155415307	rs4704970	155433570	Sing-Malay, Sing-Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	SGCD	Multiple sclerosis
333	2	203608045 203610291	rs6725887	203454130	Sing-Chinese, CHD, LWK, MEX, MKK, TSI	WDR12	Myocardial infarction (early onset)
399	3	37957108 37961932	rs9311171	37971481	Sing-Chinese, Sing-Malay, CHD, MEX, TSI	CTDSPL	Prostate cancer
28	1	25465715 25534592	rs10903129	25641524	Sing-Indian, GIH	TMEM57	Total cholesterol
147	1	194997658 195068695	rs6428370	195111216	Sing-Indian, ASW, GIH, MEX, TSI	Intergenic	Acute lymphoblastic leukemia (childhood)
147	1	194997658 195068695	rs10737680	194946078	GIH	CFH	Age-related macular degeneration
1491	9	98700200 98729161	rs10816533	98578959	CHD	ZNP510	Height

109	1	150822330 150853218	rs10888501	150804578	Sing-Malay, Sing-Indian	Intergenic	Response to antipsychotic treatment
12035	12	118473270 118475144	rs11064768	118302892	Sing-Chinese	CCDC60	Schizophrenia
2197	16	72953795 73009537	rs10871290	73030197	Sing-Indian	GLG1	Breast cancer
933	6	32539530 32681749	rs3135338	32509195	Sing-Malay, Sing-Indian	HLA	Multiple sclerosis
933	6	32539530 32681749	rs615672	32682149	Sing-Malay	HLA-DRB1	Rheumatoid arthritis
933	6	32539530 32681749	rs9272346	32712350	Sing-Malay	MHC	Type 1 diabetes

(This table was adapted from Ku et al. (2011) J. Hum. Genet. 56:552-560)[138].

The second component of the Birdsuite software, Birdseye, was used to identify novel copy number loci in the 10 populations. We subsequently found 5,947 copy number loci, of which 933 loci were excluded due to overlap with the 1,291 autosomal CNPs identified by McCarroll et al. (2008)[29]. As a result, only 5,014 were novel copy number loci, i.e., had not been previously found by McCarroll et al. (2008). Of these, 1,448 loci were detected in two or more individuals in the 10 populations. Using a more stringent definition of 'common' novel copy number loci (population frequency ≥1%), there were only 170 loci and of these, 42 loci had a population frequency ≥5%.

5.4. Study IV (Regions of homozygosity in three Southeast Asian populations)

By restricting ROHs to ≥500Kb in length, several thousand ROHs were found in each of the three populations (Table 13). On average, the Indian population had the lowest number of ROHs compared to Chinese and Malay populations. This result was consistent between the Illumina and Affymetrix arrays. When restricting to ROHs ≥1Mb in length, the average number of ROHs was still lower in the Indian (n=10) than in the Chinese (n=16) and Malay (n=15) populations when using the Illumina array. The Indian population had the lowest number of ROHs and smallest total length of ROHs per individual on average compared to the Chinese and Malay populations. These results also show that, on average, approximately 2-3% of the human genome is homozygous in these Chinese (2.9%), Malay (2.6%) and Indian (2.0%) individuals. Approximately 58% (14,414 ROHs) of the total number of 24,730 ROHs detected by the Illumina array qualified for validation by the Affymetrix array.

Table 13 – Characteristics of ROHs in three Singapore populations

	Illumina			Affymetrix		
Characteristics	**Chinese (n=96)**	**Malay (n=89)**	**Indian (n=83)**	**Chinese (n=96)**	**Malay (n=89)**	**Indian (n=83)**
Total number of ROHs	10,470	8,560	5,700	6,564	5,578	3,793
Number of ROHs per individual						
Mean (average)	109	96	69	68	63	46
Median	109	95	69	70	64	45
Minimum	87	54	42	48	38	27
Maximum	127	134	91	88	89	64
Length of ROHs (in kb)						
Mean (average)	761.6	771.1	836.0	826.5	837.1	933.2
Median	658.8	660.0	653.3	691.9	695.5	679.7
Maximum	5,078	13,620	44,840	5,088	36,270	44,810
Total length of ROHs per individual (in Mb)						
Mean (average)	83.06	74.16	57.41	56.51	52.46	42.65
Median	82.92	74.50	52.94	56.46	51.31	36.93
Minimum	65.02	37.64	30.31	40.45	29.71	20.15
Maximum	105.70	103.50	219.60	81.10	102.00	210.60
Size distribution of ROHs (proportion, %)						
500kb - 1Mb	84.9%	84.6%	85.5%	79.2%	79.0%	79.6%
≥1Mb	15.1%	15.4%	14.5%	20.8%	21.0%	20.4%
ROHs validated by Affymetrix at ≥50% overlap						
Number	5,918	5,070	3,426	-	-	-
Percentage	56.5%	59.2%	60.1%	-	-	-

We subsequently identified a total of 821 common ROH loci in all three populations, of which only 20 loci differed significantly (FDR <0.01) between the populations. Our PCA plot (Figure 12) showed a reasonably good separation of Chinese and Malay from Indian, and to the best of our knowledge, we demonstrated the utility of ROHs as a genetic marker in population structure analysis for the first time.

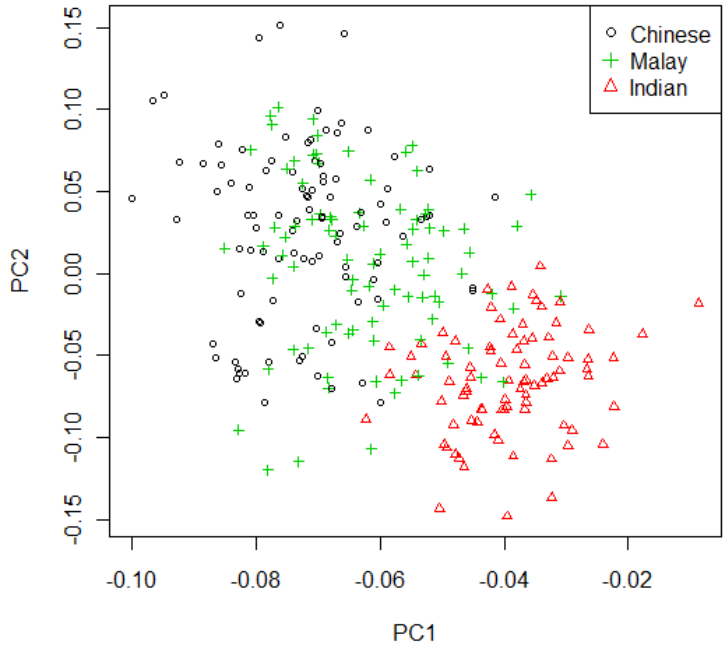


Figure 12 - PCA results based on the common ROH loci for three Singapore populations

CHAPTER 6 - DISCUSSION

6.1. CNV and ROH maps for each population

The four studies which comprise my Ph.D. thesis have investigated the population characteristics of CNVs (CNPs) and ROHs in the Singapore and Swedish populations. In addition, we have also investigated the population differences by comparison with the HapMap III populations. This has enabled us to discover substantial differences between these populations. For the three Singapore populations, our CNV results showed that the Indian population is more differentiated from Chinese and Malay populations (i.e., the differences between Indian and Chinese or Malay are greater than between Chinese and Malay) (Study I). This is in strong agreement with the findings based on SNP data from the Singapore Genome Variation Project[13], which also found that the Indian population is more differentiated from Chinese and Malay populations. This was further supported by the PCA based on the ROH data of Singapore populations, of which the plot showed a reasonably good separation of Chinese and Malays from Indians (Study IV). Although a number CNV studies have been performed in European populations, our results showed that the CNP profile of the Swedish population differed considerably from the HapMap III European populations (Study II). Finally, further comparison to the HapMap III populations using the CNP data generated by Canary, revealed that pairwise comparisons between (a) Singapore Chinese and HapMap CHD and (b) Singapore Indians and HapMap GIH are comparable. In addition, the CNP profile of the Indian population is closer to European populations than to Singapore Chinese and Malay populations and the HapMap III CHD population (Study III). Therefore, we have documented the substantial differences between the populations in the studies. This also suggests that the data

generated for the HapMap populations might not be applicable to other populations worldwide, supporting the notion of generating a genome-wide map of CNVs and ROHs for individual populations. Although my thesis focuses on CNVs and ROHs, a greater emphasis was placed on CNVs as it has, to date, received greater research attention..

Furthermore, some of the CNV loci or CNPs that showed differences in frequencies between populations overlapped with known disease-related genes and pharmacogenetics-associated genes. These differences may account for phenotypic differences between populations. In addition, the importance of CNPs was further demonstrated by their strong LD with several of the GWAS-SNPs. Taken together, these preliminary data warrant further studies to directly investigate the associations of CNVs/CNPs with various complex diseases and traits.

6.2. Major criticisms from reviewers

There were several major criticisms from reviewers for four submitted papers that are worth further discussion.

(a) Validation using quantitative PCR

One of the major criticisms from the reviewers for Study I and Study II was that quantitative PCR (qPCR) was not used to validate the CNVs identified from microarrays. It is argued that qPCR validation is necessary and important because the detection of CNVs using microarrays is prone to a high false-positive rate. We have performed an *in silico* validation in our studies.

In addition to applying a series of stringent quality control criteria at the 'sample-level' (i.e., removing samples with poor quality signal intensity data) and 'CNV-level' (i.e., removing CNV calls generated by algorithms using the accompanied statistic score for each CNV call), we have performed validation using a second genotyping platform, i.e., Illumina and Affymetrix genotyping arrays were used in Study I and *in silico* validation using two different algorithms, i.e., PennCNV and Birdsuite in Study II. In Study I, only CNV loci detected by the Illumina platform and subsequently replicated by the Affymetrix platform were discussed in detail. Furthermore, for loci that were replicated, the distribution of integer copy-numbers as detected by Illumina and Affymetrix were compared. We then tested the population differentiation of the CNV loci across the three populations, only if the integer copy-number frequency was estimated consistently across the two platforms. We found that for loci that were replicated, the vast majority (89.9% for deletion loci and 91.2% for duplication loci) had copy-number frequencies that were consistently estimated across the two platforms. On the other hand, validation through algorithm comparison method is a common and standard *in silico* approach used to validate CNV data or results[139]. Additionally, these results were also compared to the DGV (Study II). Approximately 70% of the CNV loci were considered validated by applying these two methods of validation (Table 7).

In contrast, there are several problems of applying qPCR to validate the results or estimate the FDR. First, multiple primer sets are needed to probe/validate one CNV locus as the breakpoints of CNVs are inaccurately determined using microarrays and also due to the large sizes of CNV loci. Furthermore, the boundaries of a CNV locus were also

determined through statistical methods by clustering individual CNVs into discrete loci (as described in the Material and Methods 'Construction of CNV loci' in Study I and Study II). In addition, qPCR should be performed for at least tens of CNV loci randomly selected from hundreds of CNV loci in order to provide a more accurate estimation of FDR. Although qPCR is a highly accurate method, it is not scalable to validate all the CNV loci, with the result that the FDR will be estimated from only a small number of randomly selected CNVRs. In contrast, validation through comparison of the datasets produced by two different genotyping platforms (Study I) or two different algorithms (Study II) allowed us to validate all the CNV loci at the 'whole-genome' level. The other commonly used *in silico* validation method is through comparison with the data from the DGV. Thus, we have used different *in silico* validation methods in our studies. However, one major limitation in our validation approaches (i.e., using two different genotyping platforms or software) is that a '50% overlap' was used to define the validation or replication of one locus by another genotyping platform or software. However, this is an arbitrary cutoff, and there is currently no consensus on the most appropriate cutoff to be used. Similarly, novel CNV loci were identified through comparison with entries from the DGV that did not have a >50% overlap.

(b) Small sample sizes

In addition to the validation limitation, the sample size for each population was also deemed to be inadequate for studying population characteristics by the reviewers. However, the sample size for each population in these studies is of similar size to other studies - most notably, the International HapMap Project. We have used a larger sample

size than other studies, for example, to date only one previous study has investigated CNVs in the Swedish population[140]. This study used a low resolution BAC-clone microarray with a sample size of approximately 30 samples. In our study, we overcame the limitations of the previous study by using a larger sample size (n=87) and using the highest density SNP genotyping array currently available, i.e., the Affymetrix SNP Array 6.0 (Study II).

(c) Comparisons with other populations

Finally, it was also strongly recommended that Study I and Study II could benefit substantially from a comparison to CNV data that was already available (mostly from European populations). However, it is important to emphasize that such a comparison (i.e., comparing the final results across different studies) would be inappropriate. The reason we did not compare the results of our studies (Study I and Study II) with existing data from published studies is due to the methodological inconsistencies in CNV and ROH detection in the different studies. Since different studies have used different platforms, quality control criteria and methods to construct CNV loci and to detect ROHs, this would make direct comparisons of the final results in different studies challenging, even if the same genotyping platform and algorithm were used. As a result, we would need to analyse the data from different populations with the same analytical procedure. Therefore, we have downloaded the Affymetrix SNP Array 6.0 dataset for the International HapMap III populations for comparisons with the Singapore (Study III) and Swedish (Study II) populations. As the HapMap III data was analysed in a similar way

and filtered by the same quality control criteria, this made the population comparisons more valid.

6.3. Technological limitations

One major limitation of this project was the use of SNP genotyping arrays to identify CNVs. Using the highest resolution SNP genotyping arrays in our studies has allowed us to detect smaller sizes of CNVs, as evident by our results showing that the majority of the detected CNVs were less than 50kb. Furthermore, these arrays were also supplemented with 'copy number' or 'non-polymorphic' probes in addition to the probes designed for SNPs genotyping. These copy-number probes have increased the marker or probe coverage and density in regions lacking SNPs (i.e., SNPs sparse regions) and repetitive regions such as segmental duplications, of which there are difficulties in assessing these regions. Taken together, these can be considered important improvements in CNV detection studies compared to previous studies that used earlier versions of arrays which were not purposely designed for CNVs detection. A further advantage over studies using CGH arrays is that SNP genotyping arrays generate two types of signal intensity data, thus enabling detection of ROHs. However, there are still a number of limitations associated with these high resolution SNP genotyping arrays. These limitations include (a) a limited sensitivity to detect CNVs smaller than 10kb compared to sequencing-based approaches, (b) inability to detect copy-neutral variations such as inversions and translocations and (c) imprecise mapping of the breakpoints of the CNVs.

The arrival of NGS technologies has provided alternative sequencing approaches to studying structural variations (both CNVs and copy-neutral variations) in the form of PEM and DOC. Although the sequencing-based approaches have overcome the major limitations of microarrays in detecting structural variations, it is still prohibitively expensive to be applied in population-based studies of up to several hundred samples. Thus far, sequencing-based methods have not been applied in population-based studies and published research has only shown the proof-of-concept in a few samples[11, 87]. Due to these limitations, the applications of these sequencing-based methods in population-based studies currently requires a large international effort such as the 1000 Genomes Project[38, 104], despite the sample size for each individual population not being more than several hundred in the 1000 Genomes Project. In addition, several libraries of different insert sizes, such as 3kb, 5kb or 10kb, will be needed to ensure a comprehensive detection of structural variations of varying sizes. Furthermore, the analysis of PEM and DOC is bioinformatically more challenging than microarrays, where well-developed algorithms such as PennCNV and Birdsuite are available to analyse microarray data[81]. As a result, this renders sequencing-based methods unsuitable for studying the population characteristics of CNVs and disease association studies where a large sample size would be needed.

6.4. Clinical and public health significance

Our studies have found substantial differences in the CNV/CNP profiles between Singapore, Swedish and HapMap III populations. Interestingly, many of these copy number loci overlap with known disease-associated genes and pharmacogenetic-related

genes (Study III). More specifically, we found a markedly lower deletion frequency of a CNP locus which overlapped with the WWOX gene (a tumour suppressor gene affected in multiple cancers) in Sing–Chinese and Sing–Malay compared with other populations. Another CNP of interest is a 46-kb deletion that overlaps with GSTT1 (an important detoxification enzyme and has a key role in the metabolism of carcinogenic compounds). The total deletion frequency of this CNP was high in all the 10 populations compared, ranging from 56.63% to 96.77%. However, Sing–Indians had a considerably lower total deletion frequency (56.63%) than Sing–Malays (85.06%) and Sing–Chinese (96.77%). This difference is attributable to two-copy deletion, as the difference in two-copy deletion frequency was 15.66% in Sing–Indian, 32.18% in Sing–Malay and 46.24% in Sing–Chinese.

Further, a 125-kb CNP deletion that overlapped with UGT2B17 also showed substantial differences in the deletion frequency between Asian and non-Asian populations. Asian populations (Sing–Chinese, Sing–Malay, Sing–Indian, CHD and GIH) had higher frequencies, which ranged from 82.14% to 100%, when compared with populations of European and African ancestry (48.08%–67.18%). The differences were even more apparent for two-copy deletions with the highest frequencies in CHD (70.79%), Sing–Chinese (65.59%) and Sing–Malay (52.87%), followed by the two Indian populations, GIH (37.04%) and Sing–Indian (30.95%), whereas the European and African populations were in the lower end of the spectrum with frequencies less than20%. Deletion of the UGT2B17 gene was also been found to be associated with an increased risk of prostate cancer. The functional role of the UGT2B17 enzyme is clear in prostate cancer, as it is

involved in steroid hormone (androgen) metabolism. In addition, the mismatch of UGT2B17 copy numbers in donors and recipients of stem cell transplantation were also associated with an increased risk of graft-versus-host disease.

Although a direct association between the CNPs and phenotypic differences has not been established in our studies, collectively our results suggest that CNP distributions are substantially different between populations and thus, may account for phenotypic or disease differences between them. As such, the potential implication of CNVs in clinical and public health practice is promising, however further studies are needed to establish their significance. For example, in the context of the Singapore populations, if the copy number changes of UGT2B17 were also found to be associated with an increased risk of graft-versus-host disease in the populations, then genetic screening could be implemented in the clinical setting prior to the transplantation. Similarly, a population screening program could be implemented in a high-risk group harboring multiple cancer predisposing CNVs of large effect sizes for early detection and treatment. Genetic information of CNVs overlapping with pharmacogenetic-related genes could also be beneficial for clinical drug trials, where it could be used to identify the population most likely to respond favorably to the drug.

CHAPTER 7 – FUTURE DIRECTIONS AND PERSPECTIVES

7.1. Technological developments

Microarrays have been widely used in the discovery of CNVs over the past several years. However, with the development of PEM and DOC, this raises the question of whether these sequencing-based methods will eventually replace microarrays in structural variation research. The answer is likely to be a resounding 'yes', but at present microarray and sequencing-based methods are proving to be more valuable by being complementary to each other in population studies of structural variations. The role of microarrays will likely need to be switched from that of 'discovery' to 'genotyping'. Although sequencing-based methods are more powerful in the discovery of new structural variations, these methods are costly for up to several thousand samples, especially when several libraries of different insert sizes are needed for PEM. This would limit the number of future studies of population characteristics and disease association. However, the newly discovered and the currently known structural variations can be characterised in population-based studies or investigating their associations with diseases using custom-designed oligonucleotide microarrays. However, this is limited to CNVs which are believed to be in the majority in structural variations. Thus other high-throughput methods to assay newly discovered and known copy neutral variations need to be developed.

Although the PEM and DOC methods have overcome the major shortcomings of microarrays in detecting structural variations, these methods have their own weaknesses. Nevertheless, these emerging sequencing-based methods will continue to play a role in

the discovery of structural variations until *de novo* genome assembly is more feasible[108]. *De novo* genome assembly will be more practical with the promise of TGS technologies to increase the sequence read length to several kilobases so that a full human genome can be assembled[91]. The developments of NGS and TGS technologies are occurring at a rapid pace, and more importantly, the cost of sequencing has been decreasing over the years making whole-genome sequencing attainable at a cost ranging from several thousand US dollars (e.g. the sequencing service provided by Complete Genomics) to approximately USD30,000 (e.g. whole-genome sequencing using the Illumina HiSeq2000 system). Therefore, sequencing-based approaches could be applied to Singapore populations in the near future to generate a more comprehensive map of genetic variations including SNPs, indels, CNVs and other structural variations, and ROHs. This will complement the international effort of the 1000 Genomes Project of various European, Asian and African populations. This will also allow a thorough comparison of the population characteristics of genetic variations between Singapore and other populations worldwide.

7.2. A perspective on a detailed genetic variation map for each population

Recent studies have increasingly documented the population differences of CNVs within and between populations with distinct ancestral backgrounds (i.e., African, European and Asian)[29, 30, 38, 141-143]. This was further supported by our studies documenting a wide range of differences of CNVs and CNPs between the Singapore, Swedish and International HapMap III populations[137, 138] (Study II and Study III). This further supports the concept that each geographically distinct population, despite common ancestral backgrounds e.g.,

northern Han Chinese in China or the HapMap CHB panel versus Singapore Chinese (are mostly Southern Chinese), is genetically varied to some extent.

This genetic diversity is due to differences in CNVs and other genetic variations such as SNPs, and small indels. Therefore, this may also suggest that public genetic databases, such as the International HapMap Project and the 1000 Genomes Project, are not completely representative of every geographical population worldwide. This has been shown by the inclusion of an additional seven populations of African, European and Asian ancestries in the International HapMap Phase III Project[144]. Similarly, the 1000 Genomes Project will eventually include at least 28 populations which will comprise of populations of (a) European, (b) East Asian, (c) South Asian, (d) North and South American and (e) West African ancestries upon completion (http://www.1000genomes.org/about#ProjectSamples). In addition to these international consortia, individual country efforts aiming to study population genetic differences, such as the Singapore Genome Variation Project, were also conceived to characterize the genetic profile (e.g. SNPs, haplotype and CNVs) of Chinese, Malay and Asian Indian populations in Singapore[13, 78] (Study IV). Taken together, this raises an important question of whether a comprehensive genetic variation map is needed for each geographical population worldwide.

Most of these population genetic differences had been identified by microarray data and limited to CNVs and SNPs, whilst other structural variations such as inversions and translocations were not interrogated. These differences are anticipated to be larger when

newer sequencing approaches are adopted to characterize small indels, structural variations, and rarer and population-specific variants (both SNPs and non-SNP variants). Structural variations account for a greater fraction of the diversity between individuals than SNPs[104, 108].

The detailed characterization of genetic variations in each geographical population worldwide will facilitate studies of human evolution and migration history, as well as the genetic basis underlying the phenotypic variability between populations. However, it also has negative ethical and societal implications[145, 146], for example, discrimination may occur if one population is found to harbor common deleterious risk variants that predispose to certain serious illness such as mental retardation and depression. Similarly, in the pharmacogenetics realm, if some variants are responsible for the non-responsiveness to certain drugs or lead to the requirement of a larger dosage to achieve the optimal therapeutic effect, this might label the population as an insurance risk. With the current pace of improvements in sequencing technologies, technical ability and financial feasibility, the goal of generating a detailed genetic variation map for each population is within reach. However, the negative ethical and societal impacts must be carefully monitored and minimized through the implementation and enforcement of regulations and policies by the respective parties. These ethical and societal impacts are now becoming more relevant as cheaper sequencing is becoming more available.

So will a detailed genetic variation map eventually be needed for each geographical population? The answer is a resounding ‘yes’. However, this genetic information must be

used ethically to benefit all populations and to minimize the potential harmful effects. Despite having the sequencing technologies, generating a detailed genetic variation map from population-based studies was not possible a few years ago due to the cost. However, cost may no longer be the most significant factor to consider as the sequencing costs are widely anticipated to eventually become more affordable for population-based studies. By contrast, the balance of beneficial and harmful impacts on ethical and societal aspects is the key consideration in the future generation of a detailed population genetic variation map.

REFERENCES

1. Sebat J., Lakshmi B., Troge J., Alexander J., Young J., Lundin P., et al. Large-scale copy number polymorphism in the human genome. Science. **305**, 525-528 (2004).
2. Iafrate A.J., Feuk L., Rivera M.N., Listewnik M.L., Donahoe P.K., Qi Y., et al. Detection of large-scale variation in the human genome. Nat Genet. **36**, 949-951 (2004).
3. Shlien A., Malkin D. Copy number variations and cancer susceptibility. Curr Opin Oncol. **22**, 55-63 (2010).
4. Wain L.V., Armour J.A., Tobin M.D. Genomic copy number variation, human health, and disease. Lancet. **374**, 340-350 (2009).
5. Zhang F., Gu W., Hurles M.E., Lupski J.R. Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet. **10**, 451-481 (2009).
6. Stankiewicz P., Lupski J.R. Structural variation in the human genome and its role in disease. Annu Rev Med. **61**, 437-455 (2010).
7. Girirajan S., Campbell C.D., Eichler E.E. Human Copy Number Variation and Complex Genetic Disease. Annu Rev Genet. (2010).
8. Girirajan S., Eichler E.E. Phenotypic variability and genetic susceptibility to genomic disorders. Hum Mol Genet. **19**, R176-187 (2010).
9. Carson A.R., Feuk L., Mohammed M., Scherer S.W. Strategies for the detection of copy number and other structural variants in the human genome. Hum Genomics. **2**, 403-414 (2006).
10. Carter N.P. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. **39**, S16-21 (2007).
11. Korbel J.O., Urban A.E., Affourtit J.P., Godwin B., Grubert F., Simons J.F., et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. **318**, 420-426 (2007).

12. Yoon S., Xuan Z., Makarov V., Ye K., Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. **19**, 1586-1592 (2009).
13. Teo Y.Y., Sim X., Ong R.T., Tan A.K., Chen J., Tantoso E., et al. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. **19**, 2154-2162 (2009).
14. He Y., Hoskins J.M., McLeod H.L. Copy number variants in pharmacogenetic genes. Trends Mol Med. **17**, 244-251 (2011).
15. Gamazon E.R., Huang R.S., Dolan M.E., Cox N.J. Copy number polymorphisms and anticancer pharmacogenomics. Genome Biol. **12**, R46 (2011).
16. Gibson J., Morton N.E., Collins A. Extended tracts of homozygosity in outbred human populations. Hum Mol Genet. **15**, 789-795 (2006).
17. Lencz T., Lambert C., DeRosse P., Burdick K.E., Morgan T.V., Kane J.M., et al. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci U S A. **104**, 19942-19947 (2007).
18. Nalls M.A., Guerreiro R.J., Simon-Sanchez J., Bras J.T., Traynor B.J., Gibbs J.R., et al. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics. **10**, 183-190 (2009).
19. Yang T.L., Guo Y., Zhang L.S., Tian Q., Yan H., Papasian C.J., et al. Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. J Clin Endocrinol Metab. **95**, 3777-3782 (2010).
20. Nakamura Y. DNA variations in human and medical genetics: 25 years of my experience. J Hum Genet. **54**, 1-8 (2009).
21. Ku C.S., Loy E.Y., Salim A., Pawitan Y., Chia K.S. The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet. **55**, 403-415 (2010).
22. Hastings P.J., Lupski J.R., Rosenberg S.M., Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. **10**, 551-564 (2009).

23. Botstein D., White R.L., Skolnick M., Davis R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. **32**, 314-331 (1980).
24. Weissenbach J., Gyapay G., Dib C., Vignal A., Morissette J., Millasseau P., et al. A second-generation linkage map of the human genome. Nature. **359**, 794-801 (1992).
25. Sachidanandam R., Weissman D., Schmidt S.C., Kakol J.M., Stein L.D., Marth G., et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. **409**, 928-933 (2001).
26. Mills R.E., Luttig C.T., Larkins C.E., Beauchamp A., Tsui C., Pittard W.S., et al. An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. **16**, 1182-1190 (2006).
27. Mills R.E., Pittard W.S., Mullaney J.M., Farooq U., Creasy T.H., Mahurkar A.A., et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. **21**, 830-839 (2011).
28. Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., et al. Global variation in copy number in the human genome. Nature. **444**, 444-454 (2006).
29. McCarroll S.A., Kuruvilla F.G., Korn J.M., Cawley S., Nemesh J., Wysoker A., et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. **40**, 1166-1174 (2008).
30. Conrad D.F., Pinto D., Redon R., Feuk L., Gokcumen O., Zhang Y., et al. Origins and functional impact of copy number variation in the human genome. Nature. **464**, 704-712 (2010).
31. Stranger B.E., Forrest M.S., Dunning M., Ingle C.E., Beazley C., Thorne N., et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. **315**, 848-853 (2007).
32. Gilad Y., Rifkin S.A., Pritchard J.K. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. **24**, 408-415 (2008).

33. Fraser H.B., Xie X. Common polymorphic transcript variation in human disease. Genome Res. **19**, 567-575 (2009).
34. Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. **18**, 1011-1019 (2008).
35. Haberman Y., Amariglio N., Rechavi G., Eisenberg E. Trinucleotide repeats are prevalent among cancer-related genes. Trends Genet. **24**, 14-18 (2008).
36. Hindorff L.A., Sethupathy P., Junkins H.A., Ramos E.M., Mehta J.P., Collins F.S., et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. **106**, 9362-9367 (2009).
37. Ng S.B., Nickerson D.A., Bamshad M.J., Shendure J. Massively parallel sequencing and rare disease. Hum Mol Genet. **19**, R119-124 (2010).
38. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. **467**, 1061-1073 (2010).
39. Schlotterer C. The evolution of molecular markers--just a matter of fashion? Nat Rev Genet. **5**, 63-69 (2004).
40. Frazer K.A., Murray S.S., Schork N.J., Topol E.J. Human genetic variation and its contribution to complex traits. Nat Rev Genet. **10**, 241-251 (2009).
41. Levy S., Sutton G., Ng P.C., Feuk L., Halpern A.L., Walenz B.P., et al. The diploid genome sequence of an individual human. PLoS Biol. **5**, e254 (2007).
42. Wheeler D.A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A., et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. **452**, 872-876 (2008).
43. Wang J., Wang W., Li R., Li Y., Tian G., Goodman L., et al. The diploid genome sequence of an Asian individual. Nature. **456**, 60-65 (2008).
44. Bentley D.R., Balasubramanian S., Swerdlow H.P., Smith G.P., Milton J., Brown C.G., et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. **456**, 53-59 (2008).
45. Kim J.I., Ju Y.S., Park H., Kim S., Lee S., Yi J.H., et al. A highly annotated whole-genome sequence of a Korean individual. Nature. **460**, 1011-1015 (2009).

46. Ahn S.M., Kim T.H., Lee S., Kim D., Ghang H., Kim D.S., et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. **19**, 1622-1629 (2009).
47. Pushkarev D., Neff N.F., Quake S.R. Single-molecule sequencing of an individual human genome. Nat Biotechnol. **27**, 847-850 (2009).
48. Feuk L., Carson A.R., Scherer S.W. Structural variation in the human genome. Nat Rev Genet. **7**, 85-97 (2006).
49. Tamaki K., Jeffreys A.J. Human tandem repeat sequences in forensic DNA typing. Leg Med (Tokyo). **7**, 244-250 (2005).
50. Petersdorf E.W. HLA matching in allogeneic stem cell transplantation. Curr Opin Hematol. **11**, 386-391 (2004).
51. Karas-Kuzelicki N., Mlinaric-Rascan I. Individualization of thiopurine therapy: thiopurine S-methyltransferase and beyond. Pharmacogenomics. **10**, 1309-1322 (2009).
52. Abdulla M.A., Ahmed I., Assawamakin A., Bhak J., Brahmachari S.K., Calacal G.C., et al. Mapping human genetic diversity in Asia. Science. **326**, 1541-1545 (2009).
53. Feng B.J., Huang W., Shugart Y.Y., Lee M.K., Zhang F., Xia J.C., et al. Genome-wide scan for familial nasopharyngeal carcinoma reveals evidence of linkage to chromosome 4. Nat Genet. **31**, 395-399 (2002).
54. Bakker S.C., van der Meulen E.M., Buitelaar J.K., Sandkuijl L.A., Pauls D.L., Monsuur A.J., et al. A whole-genome scan in 164 Dutch sib pairs with attention-deficit/hyperactivity disorder: suggestive evidence for linkage on chromosomes 7p and 15q. Am J Hum Genet. **72**, 1251-1260 (2003).
55. Garner C.P., Ding Y.C., Steele L., Book L., Leiferman K., Zone J.J., et al. Genome-wide linkage analysis of 160 North American families with celiac disease. Genes Immun. **8**, 108-114 (2007).
56. Lopez S., Buil A., Ordonez J., Souto J.C., Almasy L., Lathrop M., et al. Genome-wide linkage analysis for identifying quantitative trait loci involved in the regulation of lipoprotein a (Lpa) levels. Eur J Hum Genet. **16**, 1372-1379 (2008).

57. Wang W.Y., Barratt B.J., Clayton D.G., Todd J.A. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet. **6**, 109-118 (2005).
58. Hirschhorn J.N., Daly M.J. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. **6**, 95-108 (2005).
59. Matsuzaki H., Dong S., Loi H., Di X., Liu G., Hubbell E., et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods. **1**, 109-111 (2004).
60. Steemers F.J., Chang W., Lee G., Barker D.L., Shen R., Gunderson K.L. Whole-genome genotyping with the single-base extension assay. Nat Methods. **3**, 31-33 (2006).
61. Ragoussis J. Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet. **10**, 117-133 (2009).
62. Shendure J., Ji H. Next-generation DNA sequencing. Nat Biotechnol. **26**, 1135-1145 (2008).
63. Mardis E.R. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. **9**, 387-402 (2008).
64. Metzker M.L. Sequencing technologies - the next generation. Nat Rev Genet. **11**, 31-46 (2010).
65. Mardis E.R. A decade's perspective on DNA sequencing technology. Nature. **470**, 198-203 (2011).
66. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature. **455**, 237-241 (2008).
67. Glessner J.T., Wang K., Cai G., Korvatska O., Kim C.E., Wood S., et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. **459**, 569-573 (2009).
68. Cook E.H., Jr., Scherer S.W. Copy-number variations associated with neuropsychiatric conditions. Nature. **455**, 919-923 (2008).
69. Lee C., Iafrate A.J., Brothman A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet. **39**, S48-54 (2007).

70. Freeman J.L., Perry G.H., Feuk L., Redon R., McCarroll S.A., Altshuler D.M., et al. Copy number variation: new insights in genome diversity. Genome Res. **16**, 949-961 (2006).
71. Tuzun E., Sharp A.J., Bailey J.A., Kaul R., Morrison V.A., Pertz L.M., et al. Fine-scale structural variation of the human genome. Nat Genet. **37**, 727-732 (2005).
72. Zogopoulos G., Ha K.C., Naqib F., Moore S., Kim H., Montpetit A., et al. Germ-line DNA copy number variation frequencies in a large North American population. Hum Genet. **122**, 345-353 (2007).
73. Wong K.K., deLeeuw R.J., Dosanjh N.S., Kimm L.R., Cheng Z., Horsman D.E., et al. A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet. **80**, 91-104 (2007).
74. Pennisi E. Breakthrough of the year. Human genetic variation. Science. **318**, 1842-1843 (2007).
75. Matsuzaki H., Wang P.H., Hu J., Rava R., Fu G.K. High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. Genome Biol. **10**, R125 (2009).
76. Park H., Kim J.I., Ju Y.S., Gokcumen O., Mills R.E., Kim S., et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet. **42**, 400-405 (2010).
77. Yim S.H., Kim T.M., Hu H.J., Kim J.H., Kim B.J., Lee J.Y., et al. Copy number variations in East-Asian population and their evolutionary and functional implications. Hum Mol Genet. **19**, 1001-1008 (2010).
78. Ku C.S., Pawitan Y., Sim X., Ong R.T., Seielstad M., Lee E.J., et al. Genomic copy number variations in three Southeast Asian populations. Hum Mutat. **31**, 851-857 (2010).
79. Wang K., Li M., Hadley D., Liu R., Glessner J., Grant S.F., et al. 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 (2007).

80. Korn J.M., Kuruvilla F.G., McCarroll S.A., Wysoker A., Nemesh J., Cawley S., et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. **40**, 1253-1260 (2008).
81. Dellinger A.E., Saw S.M., Goh L.K., Seielstad M., Young T.L., Li Y.J. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res. **38**, e105 (2010).
82. Perry G.H., Ben-Dor A., Tsalenko A., Sampas N., Rodriguez-Revenga L., Tran C.W., et al. The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet. **82**, 685-695 (2008).
83. Cooper G.M., Zerr T., Kidd J.M., Eichler E.E., Nickerson D.A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet. **40**, 1199-1203 (2008).
84. Alkan C., Coe B.P., Eichler E.E. Genome structural variation discovery and genotyping. Nat Rev Genet. **12**, 363-376 (2011).
85. Feuk L., MacDonald J.R., Tang T., Carson A.R., Li M., Rao G., et al. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genet. **1**, e56 (2005).
86. Bansal V., Bashir A., Bafna V. Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Res. **17**, 219-230 (2007).
87. Kidd J.M., Cooper G.M., Donahue W.F., Hayden H.S., Sampas N., Graves T., et al. Mapping and sequencing of structural variation from eight human genomes. Nature. **453**, 56-64 (2008).
88. Medvedev P., Stanciu M., Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. **6**, S13-20 (2009).
89. Meyerson M., Gabriel S., Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. **11**, 685-696 (2010).
90. Harismendy O., Ng P.C., Strausberg R.L., Wang X., Stockwell T.B., Beeson K.Y., et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. **10**, R32 (2009).

91. Schadt E.E., Turner S., Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. **19**, R227-240 (2010).
92. Branton D., Deamer D.W., Marziali A., Bayley H., Benner S.A., Butler T., et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. **26**, 1146-1153 (2008).
93. Drmanac R., Sparks A.B., Callow M.J., Halpern A.L., Burns N.L., Kermani B.G., et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. **327**, 78-81 (2010).
94. Li Y., Wang J. Faster human genome sequencing. Nat Biotechnol. **27**, 820-821 (2009).
95. Koboldt D.C., Ding L., Mardis E.R., Wilson R.K. Challenges of sequencing human genomes. Brief Bioinform. **11**, 484-498 (2010).
96. Robison K. Application of second-generation sequencing to cancer genomics. Brief Bioinform. **11**, 524-534 (2010).
97. Eichler E.E., Nickerson D.A., Altshuler D., Bowcock A.M., Brooks L.D., Carter N.P., et al. Completing the map of human genetic variation. Nature. **447**, 161-165 (2007).
98. Kidd J.M., Graves T., Newman T.L., Fulton R., Hayden H.S., Malig M., et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell. **143**, 837-847 (2010).
99. Manolio T.A., Collins F.S., Cox N.J., Goldstein D.B., Hindorff L.A., Hunter D.J., et al. Finding the missing heritability of complex diseases. Nature. **461**, 747-753 (2009).
100. Eichler E.E., Flint J., Gibson G., Kong A., Leal S.M., Moore J.H., et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. **11**, 446-450 (2010).
101. Bodmer W., Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. **40**, 695-701 (2008).
102. Bodmer W., Tomlinson I. Rare genetic variants and the risk of cancer. Curr Opin Genet Dev. **20**, 262-267 (2010).

103. Sudmant P.H., Kitzman J.O., Antonacci F., Alkan C., Malig M., Tsalenko A., et al. Diversity of human copy number variation and multicopy genes. Science. **330**, 641-646 (2010).
104. Mills R.E., Walter K., Stewart C., Handsaker R.E., Chen K., Alkan C., et al. Mapping copy number variation by population-scale genome sequencing. Nature. **470**, 59-65 (2011).
105. Li R., Zhu H., Ruan J., Qian W., Fang X., Shi Z., et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. **20**, 265-272 (2010).
106. Li Y., Hu Y., Bolund L., Wang J. State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics. **4**, 271-277 (2010).
107. Paszkiewicz K., Studholme D.J. De novo assembly of short sequence reads. Brief Bioinform. **11**, 457-472 (2010).
108. Li Y., Zheng H., Luo R., Wu H., Zhu H., Li R., et al. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat Biotechnol. **29**, 723-730 (2011).
109. McCarroll S.A., Huett A., Kuballa P., Chilewski S.D., Landry A., Goyette P., et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet. **40**, 1107-1112 (2008).
110. Willer C.J., Speliotes E.K., Loos R.J., Li S., Lindgren C.M., Heid I.M., et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet. **41**, 25-34 (2009).
111. Craddock N., Hurles M.E., Cardin N., Pearson R.D., Plagnol V., Robson S., et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. **464**, 713-720 (2010).
112. Cronin S., Blauw H.M., Veldink J.H., van Es M.A., Ophoff R.A., Bradley D.G., et al. Analysis of genome-wide copy number variation in Irish and Dutch ALS populations. Hum Mol Genet. **17**, 3392-3398 (2008).

113. Blauw H.M., Veldink J.H., van Es M.A., van Vught P.W., Saris C.G., van der Zwaag B., et al. Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. Lancet Neurol. **7**, 319-326 (2008).
114. Bae J.S., Cheong H.S., Park B.L., Kim L.H., Park T.J., Kim J.Y., et al. Genome-wide association analysis of copy number variations in subarachnoid aneurysmal hemorrhage. J Hum Genet. **55**, 726-730 (2010).
115. Walsh T., McClellan J.M., McCarthy S.E., Addington A.M., Pierce S.B., Cooper G.M., et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science. **320**, 539-543 (2008).
116. Mulle J.G., Dodd A.F., McGrath J.A., Wolyniec P.S., Mitchell A.A., Shetty A.C., et al. Microdeletions of 3q29 confer high risk for schizophrenia. Am J Hum Genet. **87**, 229-236 (2010).
117. Mefford H.C., Muhle H., Ostertag P., von Spiczak S., Buysse K., Baker C., et al. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. PLoS Genet. **6**, e1000962 (2010).
118. Bochukova E.G., Huang N., Keogh J., Henning E., Purmann C., Blaszczyk K., et al. Large, rare chromosomal deletions associated with severe early-onset obesity. Nature. **463**, 666-670 (2010).
119. McCarroll S.A., Bradner J.E., Turpeinen H., Volin L., Martin P.J., Chilewski S.D., et al. Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. Nat Genet. **41**, 1341-1344 (2009).
120. Jeon J.P., Shim S.M., Nam H.Y., Ryu G.M., Hong E.J., Kim H.L., et al. Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus. BMC Genomics. **11**, 426 (2010).
121. Garcia-Ortiz H., Velazquez-Cruz R., Espinosa-Rosales F., Jimenez-Morales S., Baca V., Orozco L. Association of TLR7 copy number variation with susceptibility to childhood-onset systemic lupus erythematosus in Mexican population. Ann Rheum Dis. **69**, 1861-1865 (2010).

122. Peiffer D.A., Le J.M., Steemers F.J., Chang W., Jenniges T., Garcia F., et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. **16**, 1136-1148 (2006).
123. Curtis D. Extended homozygosity is not usually due to cytogenetic abnormality. BMC Genet. **8**, 67 (2007).
124. Li L.H., Ho S.F., Chen C.H., Wei C.Y., Wong W.C., Li L.Y., et al. Long contiguous stretches of homozygosity in the human genome. Hum Mutat. **27**, 1115-1121 (2006).
125. Simon-Sanchez J., Scholz S., Fung H.C., Matarin M., Hernandez D., Gibbs J.R., et al. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet. **16**, 1-14 (2007).
126. Nothnagel M., Lu T.T., Kayser M., Krawczak M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet. **19**, 2927-2935 (2010).
127. McQuillan R., Leutenegger A.L., Abdel-Rahman R., Franklin C.S., Pericic M., Barac-Lauc L., et al. Runs of homozygosity in European populations. Am J Hum Genet. **83**, 359-372 (2008).
128. Nalls M.A., Simon-Sanchez J., Gibbs J.R., Paisan-Ruiz C., Bras J.T., Tanaka T., et al. Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet. **5**, e1000415 (2009).
129. Curtis D., Vine A.E., Knight J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet. **72**, 261-278 (2008).
130. Day I.N. dbSNP in the detail and copy number complexities. Hum Mutat. **31**, 2-4 (2010).
131. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. **81**, 559-575 (2007).

132. Ku C.S., Naidoo N., Teo S.M., Pawitan Y. Regions of homozygosity and their impact on complex diseases and traits. Hum Genet. **129**, 1-15 (2011).
133. Vine A.E., McQuillin A., Bass N.J., Pereira A., Kandaswamy R., Robinson M., et al. No evidence for excess runs of homozygosity in bipolar disorder. Psychiatr Genet. **19**, 165-170 (2009).
134. Campbell H., Rudan I., Bittles A.H., Wright A.F. Human population structure, genome autozygosity and human health. Genome Med. **1**, 91 (2009).
135. Eckel-Passow J.E., Atkinson E.J., Maharjan S., Kardia S.L., de Andrade M. Software comparison for evaluating genomic copy number variation for Affymetrix 6.0 SNP array platform. BMC Bioinformatics. **12**, 220 (2011).
136. Mei T.S., Salim A., Calza S., Seng K.C., Seng C.K., Pawitan Y. Identification of recurrent regions of Copy-Number Variants across multiple individuals. BMC Bioinformatics. **11**, 147 (2010).
137. Teo S.M., Ku C.S., Naidoo N., Hall P., Chia K.S., Salim A., et al. A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals. J Hum Genet. **56**, 524-533 (2011).
138. Ku C.S., Teo S.M., Naidoo N., Sim X., Teo Y.Y., Pawitan Y., et al. Copy number polymorphisms in new HapMap III and Singapore populations. J Hum Genet. **56**, 552-560 (2011).
139. Pinto D., Marshall C., Feuk L., Scherer S.W. Copy-number variation in control population cohorts. Hum Mol Genet. **16 Spec No. 2**, R168-173 (2007).
140. Diaz de Stahl T., Sandgren J., Piotrowski A., Nord H., Andersson R., Menzel U., et al. Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array. Hum Mutat. **29**, 398-408 (2008).
141. Gautam P., Jha P., Kumar D., Tyagi S., Varma B., Dash D., et al. Spectrum of large copy number variations in 26 diverse Indian populations: potential involvement in phenotypic diversity. Hum Genet. (2011).

142. Wineinger N.E., Pajewski N.M., Kennedy R.E., Wojczynski M.K., Vaughan L.K., Hunt S.C., et al. Characterization of autosomal copy-number variation in African Americans: the HyperGEN Study. Eur J Hum Genet. (2011).
143. Chen W., Hayward C., Wright A.F., Hicks A.A., Vitart V., Knott S., et al. Copy Number Variation across European Populations. PLoS One. **6**, e23087 (2011).
144. Altshuler D.M., Gibbs R.A., Peltonen L., Dermitzakis E., Schaffner S.F., Yu F., et al. Integrating common and rare genetic variation in diverse human populations. Nature. **467**, 52-58 (2010).
145. McGuire A.L., Caulfield T., Cho M.K. Research ethics and the challenge of whole-genome sequencing. Nat Rev Genet. **9**, 152-156 (2008).
146. Guttmacher A.E., McGuire A.L., Ponder B., Stefansson K. Personalized genomic information: preparing for the future of genetic medicine. Nat Rev Genet. **11**, 161-165 (2010).

APPENDICES
(Ph.D. publications)

Appendix Table 1 - Summary of population-based CNV studies in different populations using SNP genotyping microarrays

Study and year	Microarray platform and algorithm	Population, sample size	Major findings
Redon et al. 2006	•Affymetrix 500K early access and CGH with a Whole Genome TilePath (WGTP) array	270 individuals from four HapMap II populations	•The average number of CNVs detected per experiment was 70 and 24 for the WGTP and 500K platforms respectively. •Identified 913 CNVRs on the WGTP platform and 980 CNVRs on the 500K platform. •Approximately half of these CNVRs were called in more than one individual and 43% of all CNVs identified on one platform were replicated on the other. •Combining the data resulted in a total of 1,447 discrete CNVRs, covering 12% (~360 Mb) of the human genome.
Zogopoulos et al. 2007	•Affymetrix 100K and 500 K •CNAG	1,190 North Americans	•Assembled a genomic map consisting of 578 CNVRs covering approximately 220 Mb (7.3%) of the human genome. •Copy number changes in the majority of these CNVRs are rare (>93% CNVRs occurring at <1% frequency). •Population frequencies of 1–5% and >5% were estimated for CNVs present in approximately 6 and 1% of CNVRs, respectively.
Kang et al. 2008	•Affymetrix 250K Nsp •CNAG, dChip and GEMCA	116 Korean individuals	•There were significant differences in the numbers and positions of CNVs identified by the three methods. •The dChip algorithm identified more CNVs than CNAG and GEMCA. In total, 772, 403 and 302 CNVs were found by the dChip, CNAG and GEMCA algorithms. •A total of 141 CNVs was identified (selecting CNVs represented by more than two algorithms) and defined 65 CNVRs from the 141 CNVs by merging overlapping CNVs from different individuals, among which 10 CNVRs (15.4%) were novel and not present in the DGV. •Most CNVs (75%) from the Korean population were rare (<1%), occurring just once among the 116 individuals
McElroy et al. 2009	Affymetrix 500K CNAT	385African Americans (from 28 US States) and 435 individuals of European descent (from Australia, East Europe, North Africa, North America,	•In the African Americans, a total of 1362 copy number events were identified, with a mean of 3.5 CNVs per individual. •A total of 1972 copy number events were identified in Whites, resulting in a mean of 4.8 CNVs per individual. •1068 CNV regions were identified across all individuals (412 were unique to African Americans, 580 were unique to Whites, and 76 were common between the two populations)

		North Europe, South America, South Europe, and West Europe)	
Lin et al. 2009	•Illumina HumanMap550K •PennCNV	813 Han Chinese residing in Taiwan	•4452 reliable CN-altered events (1025 non-redundant genomic regions which are defined as any overlap of CNVs) were found in the 813 individuals. •Only 365 of the 1025 non-redundant genomic regions were found to be CN variable in at least two individuals, and were regarded as CNVRs in this study. •The majority of the CNVRs (298; 81.6%) had been reported in the DGV. •Only 64 of 365 CNVRs had a CNV allele frequency greater than 1%.
Li et al. 2009	•Affymetrix GeneChip Mapping 500K Array •CNAT	985 US Caucasian and 692 Han Chinese individuals	•2,381 autosomal CNVs were identified in the 1,677 subjects. •Among the 2,381 autosomal CNVs, 15.4% were detected in both populations, 41.4% only in Caucasian, and 43.2% only in Han Chinese •1135 CNVRs covering approximately 439 Mb (14.3%) of the human genome were identified. •Compared with the DGV, 69% (680) of 985 autosomal CNVRs overlapped with previously published CNVs. The remaining 305 CNVRs were novel and covered 2.5% (72.1 Mb) of the 22 autosomes.
Yim et al. 2010	•Affymetrix Genome-Wide Human SNP array 5.0 •SW-ARRAY algorithm	3578 Korean individuals	•Identified 144207 CNVs •4003 CNVRs were defined that encompass 241.9 Mb accounting for ~8% of the human genome (a total of 3076 CNVs called in a single individual were excluded from defining CNVRs) •16% of the CNVRs (656/4003) were observed in ≥1% of 3578 study subjects. Among the CNVRs with an allele frequency≥1%, 130 CNVRs (3.2% of total CNVRs) were observed in ≥5% of study subjects •By comparing to DGV, 1926 CNVRs (48.1%) were known ones, and remaining 2077 CNVRs (51.9%) were potentially novel.
Wineinger et al. 2011	•Affymetrix SNP Array 6.0 •Birdsuite and PennCNV	446 African-American subjects	•Identified 11 070 CNVs that were called by both algorithms, including 8385 deletions and 2685 duplications. •1541 unique CNVRs identified by both Birdsuite and PennCNV, of which 309 were novel (did not overlap with any of the CNVs included in the DGV database) •Among the CNVRs identified by both algorithms, 655 were present in more than one individual •The majority of CNVRs that were called by both Birdsuite and PennCNV were rare (77.7%), occurring in <1% of the study population (≤4 individuals)

			•Many CNVRs were singleton (57.6%), only occurring in one individual
Chen et al. 2011	•Illumina Infinium HumanHap 300 •QuantiSNP and cnvPartition	2789 individuals from the island of Vis, Croatia (n=965), the Orkney Isles, Scotland (n=691) and South Tyrol, Italy (n=1133)	•Identified 4016 autosomal CNVs in 1964 individuals, out of the total 2789 samples •An average number of 2.05 detectable CNVs per sample •Fewer CNVs were detected on average in Orcadians (0.91 CNV per person) than in South Tyroleans (1.77 per person) or Vis islanders (1.43 per person). •The 4016 CNVs were clustered into 743 non redundant CNVRs which covered a total of 187.95 Mb (6.6%) of the 22 autosomes •Different patterns of CNV frequency were observed in different populations; 588 CNVRs (79.1%) were specific to just one of the three population isolates: 244 of them were detected only in Dalmatians, 112 only in Orcadians and 239 only in South Tyroleans.
Xu et al. 2010	•Illumina HumanHap 610 Quad and 1M •PennCNV	1917 Chinese, 2399 Malays, and 2217 Indians residing in Singapore	•Identified about 16 CNVs per individual •Over half of the CNVs in each population are of low frequency (population frequency < 10%), and more than one-third are rare (population frequency < 1%). •Over 70% of these rare CNVs are replicated in the same population (non-singletons), and the majority is shared by at least two populations. •In each population, about 20% of the CNVs are not found in the DGV and thus considered novel. Over 85% of the novel CNVs detected in the present study are rare (population frequency < 1%)

References

1. Redon R., Ishikawa S., Fitch K.R., Feuk L., Perry G.H., Andrews T.D., et al. Global variation in copy number in the human genome. *Nature.* 444, 444-454 (2006).
2. Zogopoulos G., Ha K.C., Naqib F., Moore S., Kim H., Montpetit A., et al. Germ-line DNA copy number variation frequencies in a large North American population. *Hum Genet.* 122, 345-353 (2007).
3. Kang T.W., Jeon Y.J., Jang E., Kim H.J., Kim J.H., Park J.L., et al. Copy number variations (CNVs) identified in Korean individuals. *BMC Genomics.* 9, 492 (2008).
4. McElroy J.P., Nelson M.R., Caillier S.J., Oksenberg J.R. Copy number variation in African Americans. *BMC Genet.* 10, 15 (2009).
5. Lin C.H., Lin Y.C., Wu J.Y., Pan W.H., Chen Y.T., Fann C.S. A genome-wide survey of copy number variations in Han Chinese residing in Taiwan. *Genomics.* 94, 241-246 (2009).
6. Li J., Yang T., Wang L., Yan H., Zhang Y., Guo Y., et al. Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. *PLoS One.* 4, e7958 (2009).

7. Yim S.H., Kim T.M., Hu H.J., Kim J.H., Kim B.J., Lee J.Y., et al. Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum Mol Genet.* 19, 1001-1008 (2010).
8. Wineinger N.E., Pajewski N.M., Kennedy R.E., Wojczynski M.K., Vaughan L.K., Hunt S.C., et al. Characterization of autosomal copy-number variation in African Americans: the HyperGEN Study. *Eur J Hum Genet.* 19, 1271-1275 (2011).
9. Chen W., Hayward C., Wright A.F., Hicks A.A., Vitart V., Knott S., et al. Copy number variation across European populations. *PLoS One.* 6, e23087 (2011).
10. Xu H., Poh W.T., Sim X., Ong R.T., Suo C., Tay W.T., et al. SgD-CNV, a database for common and rare copy number variants in three Asian populations. *Hum Mutat.* 32, 1341-1349 (2011).

RESEARCH ARTICLE

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Genomic Copy Number Variations in Three Southeast Asian Populations

Chee-Seng Ku,[1] Yudi Pawitan,[2] Xueling Sim,[1] Rick T.H. Ong,[1,3] Mark Seielstad,[3,4] Edmund J.D. Lee,[5] Yik-Ying Teo,[1,6–8] Kee-Seng Chia,[1,2,8] and Agus Salim[1,8]*

[1]*Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore;* [2]*Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden;* [3]*Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore;* [4]*Institute for Human Genetics and Department of Laboratory Medicine, University of California, San Francisco, California;* [5]*Department of Pharmacology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore;* [6]*Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom;* [7]*Department of Statistics and Applied Probability, National University of Singapore, Singapore;* [8]*Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Communicated by Pui-Yan Kwok

Received 24 September 2009; accepted revised manuscript 29 April 2010.
Published online 17 May 2010 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21287

ABSTRACT: Research on the role of copy number variations (CNVs) in the genetic risk of diseases in Asian populations has been hampered by a relative lack of reference CNV maps for Asian populations outside the East Asians. In this article, we report the population characteristics of CNVs in Chinese, Malay, and Asian Indian populations in Singapore. Using the Illumina Human 1M Beadchip array, we identify 1,174 CNV loci in these populations that corroborated with findings when the same samples were typed on the Affymetrix 6.0 platform. We identify 441 novel loci not previously reported in the Database of Genomic Variations (DGV). We observe a considerable number of loci that span all three populations and were previously unreported, as well as population-specific loci that are quite common in the respective populations. From this we observe the distribution of CNVs in the Asian Indian population to be considerably different from the Chinese and Malay populations. About half of the deletion loci and three-quarters of duplication loci overlap UCSC genes. Tens of loci show population differentiation and overlap with genes previously known to be associated with genetic risk of diseases. One of these loci is the CYP2A6 deletion, previously linked to reduced susceptibility to lung cancer. Hum Mutat 31:851–857, 2010. © 2010 Wiley-Liss, Inc.

KEY WORDS: Asian populations; copy number variation; CNV; data resources; SNP array; PennCNV

Introduction

Copy number variations (CNVs) are defined as gains or losses in the number of copies of a segment of DNA (larger than 1 kb in length) when compared to a reference genome, and provide further insight into the complexity and diversity of genetic variations. Widespread deletions and duplications in the human genome were first reported in 2004 [Iafrate et al., 2004; Sebat et al., 2004] and many more have since been discovered [Conrad et al., 2009; McCarroll et al., 2008; Redon et al., 2006]. The comprehensive detection and characterization of CNVs is laying the foundation to improve our understanding of human genetic variation and is an important tool for deciphering the role of CNV in the risk of complex diseases. In fact, recent evidence has linked CNVs with complex diseases such as autoimmune disorders, HIV infection, schizophrenia, and autism [Wain et al., 2009].

To study the role of CNVs in the genetic risk of diseases in Asian populations, a more complete map of CNVs in these populations is needed. To date, there has been relatively little research on CNVs in Asian populations apart from the East Asians component of the Hapmap collections. Articles from other research group that report CNV in Asian populations tend to focus on East Asian ethnic groups [Li et al., 2009; Yim et al., 2010]. In this article, we explore the extent of CNVs in several Southeast Asian populations, namely, Chinese, Malay, and Asian Indian populations in Singapore. The subjects in this study are part of the Singapore Genome Variation Project (SGVP) [Teo et al., 2009]. The Chinese in Singapore are mostly Southern Chinese, which reflect the origins of most first-generation Chinese migrants in Singapore [Saw, 2007]. Previous research has shown that there is a north–south gradient population structure in the genetic structure of Han Chinese [Chen et al., 2009; Teo et al., 2009]. As such, we expect that the CNV characteristics of Southern Chinese to be different from Northern Chinese whom Beijing Chinese in the HapMap collection are part of. The Malays are the native population of Singapore, with close cultural and migration history with the Malays in the nearby Brunei, Indonesia, Malaysia, and Southern Thailand. In a broader scope, the Malays are part of the Austronesian people [Bellwood et al., 1995], which constitute the majority of ethnic groups in Brunei, Indonesia, Malaysia, and the Philippines, as well as forming a significant proportion of populations in Madagascar and Thailand. The Indians in Singapore are mostly descended from the Southern Indian ethnic groups of Tamils and Telugas [Saw, 2007]. The Southern Indians are genetically different from the Northern Indians, which in

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Agus Salim, Centre for Molecular Epidemiology, Department of Epidemiology and Public Health (MD3), Yong Loo Lin School of Medicine, National University of Singapore, 16 Medical Drive, Singapore 117597.
E-mail: ephaguss@nus.edu.sg

turn, is closer genetically to the HapMap CEU population [Reich et al., 2009].

In this article, we describe the CNV characteristics of these three populations and we believe that our findings can be used to complement CNV maps from HapMap populations to form a more global CNV map, as well to provide resources as a basis to further investigate the roles of CNV in the risk of diseases, especially in Asian populations. The data and information from this project is available online, and can be accessed at the SGVP Website (http://www.nus-cme.org.sg/SGVP/).

Materials and Methods

DNA Samples and Demographic Data

The genomic DNA samples used in this study was extracted from the peripheral blood samples of individuals recruited under a previous project that has been approved by the National University of Singapore-Institutional Review Board (NUS-IRB). The project recruited 600 unrelated and apparently healthy individuals (without clinical diseases) from the three major populations in Singapore (Chinese, Malay, and Asian Indian) for identification and characterization of novel genetic variants in drug transporters and ion channel genes. The DNA was directly extracted genomic DNA without whole genome amplification steps prior to genotyping to avoid any potential errors caused by uneven amplification throughout the genome. This study was approved by the NUS-IRB (Reference Code: 07-199E).

The DNA samples for this study ($n = 292$) were chosen using stratified random sampling from the pool of samples to ensure approximately equal representation of each population and gender. The selected samples were genotyped using Illumina® Human 1M Beadchip and Affymetrix Genome-Wide Human SNP Array 6.0. The Illumina chip was used to detect CNV loci in the study population, whereas the Affymetrix chip was used to characterize loci that are independently replicated (see Replicated CNV Loci section). The samples were anonymized, but basic demographic data such as gender, age, and self-reported ethnicity were retained. There were 99 Chinese, 98 Malay, and 95 Indian individuals in the genotyped samples.

Quality Control Measures

Pre-CNV calling

The genotyping using Illumina 1M chip for all DNA samples was done according to the manufacturer's protocol (Infinium II Assay Protocol). The SGVP applied several filtering criteria to identify and remove unsuitable samples [Teo et al., 2009] that we similarly adopted here. Five samples were removed because their genotype call rates were below the 98% threshold. One sample was excluded due to mislabeling of the sample ID during the genotyping process. A further five samples were suspected of having first-degree relative relationship and were also subsequently removed. Seven samples were removed because their principal component scores based on single nueleotide polymorphism (SNP) genotyping data (see Supp. Fig. S1) suggest population admixture or misclassification of self-reported ethnicity. One sample had to be removed because it failed to be genotyped during the validation study using the Affymetrix 6.0 array. Thus, of the initial 292 samples, 19 were filtered through these various criteria, leaving 273 samples for CNV calling.

Post-CNV calling

We used the PennCNV algorithm to identify both deletions and duplications for the 22 autosomes and the X chromosome [Wang et al., 2007]. The log R ratio (LRR) was calculated using HapMap samples as reference. By default, PennCNV does not limit its detection to aberrant regions greater than 1 kb in size, and in the spirit of maximizing the information we extract from our data, we decided to keep these small aberrant regions known as indels in our analysis. As a consequence, in the subsequent paragraphs whenever the term "CNV" is used it is implicitly assumed to include indels.

We applied a set of filtering criteria (as recommended by the algorithm) to exclude poor quality samples (criteria: LRR-standard deviation >0.28, BAF-median >0.55, BAF-median <0.45, or BAF-drift >0.002), which resulted in a further seven samples being excluded because their intensity data failed to conform to these criteria. Our final set for analysis included 266 samples that consisted of 93 Chinese, 88 Malay, and 85 Indian individuals. The mean age of subject was 22.8 ± 4.0 (age range: 19–48) for Chinese, 22.5 ± 4.2 (age range: 18–36) for Malay, and 22.7 ± 4.8 (age range: 18–42) for Indian. The gender proportion of male to female is almost equal in each population: 47:46 (Chinese), 41:47 (Malay), 43:42 (Indian).

For each sample, PennCNV returned a list of regions with abnormal copy number with their associated confidence scores. The score is a log Bayes Factor that measures the likelihood that the region harbors an abnormal copy number. A confidence score of 10 or larger has been suggested as a threshold to classify reliable CNV calls (Kai Wang, personal communication). In our case, we retained all CNV called with confidence scores higher than the median confidence score. This median score was calculated based on the confidence scores of CNVs detected in all individuals, and its value is approximately 12. Although the confidence score is only a statistical measure of true positive, empirical evidence (see Results section) shows that CNV regions with a higher confidence score are more likely to be detected consistently across both platforms. This empirical evidence further justifies our decision to retain only reliable CNV regions called with sufficient degree of confidence. For the subsequent analyses, only reliable CNV regions are included.

Construction of CNV Loci

Because CNV regions as called by PennCNV tend to overlap, we merged these regions into discrete, nonoverlapping loci with the boundaries of each locus determined by the union of all CNV regions that belong to that particular locus [Redon et al., 2006]. If both deletions and duplications were observed in a particular locus, two separate loci were identified for each form of CNV.

Replicated CNV Loci

To validate the CNV loci identified using the Illumina 1M platform, we genotyped the same 266 samples using Affymetrix SNP Array 6.0. The genotyping was done according to the manufacturer's protocol. The calculations of LRR and BAF are done according to PennCNV protocol (http://www.openbioinformatics.org/PennCNV/PennCNV_tutorial_affy_gw6.html). The signal intensity data were then analyzed using PennCNV with the same parameters as used for the Illumina samples.

The same confidence score threshold as used for the Illumina platform was used to filter the unreliable CNV calls. The reliable

CNV calls were then used to construct CNV loci. A CNV locus found using the Illumina platform was considered to be replicated if there was at least one overlapping CNV locus found using the Affymetrix platform. A CNV locus detected using Illumina is considered replicated if it shares at least 50% of its length with a CNV locus detected using Affymetrix platform.

Novel CNV Loci

To identify CNV loci that are novel, we compare our results with the CNV loci published in the Database of Genomic Variants (DGV) [Iafrate et al., 2004]. We used the most recent version of the DGV (variation.hg18.v8.aug.2009.txt and indel.hg18.v8.aug.2009.txt). These files were downloaded from the DGV Website (http://projects.tcag.ca/variation/). We classified a particular CNV locus identified using Illumina and subsequently replicated using Affymetrix platform as novel if it does not share at least 50% of its length with any established CNV loci in the DGV database.

Population Differentiation of CNV Loci

We used a V_{st} statistic [Redon et al., 2006] to describe the overall population differentiation due to CNV. For each locus, the V_{st} statistic was computed using $\log_2$ intensity data from the probe within that locus with the strongest signal of population differentiation. The strategy we adopted here follows the procedure described in Redon et al. [2006]. As the V_{st} statistic is not very sensitive in identifying loci with recent positive selection signals to identify specific loci with strong population differentiation, we also compared the distribution of integer copy numbers across the three ethnic groups using Fisher's exact test. A P-value < 0.001 from this test was used to identify loci that segregated at different frequency across the three populations. Because there is uncertainty with the estimated integer copy numbers, we only conducted Fisher's exact test at loci for which the distribution of copy numbers were estimated consistently across the two platforms. We defined these loci as loci for which the estimated proportion of subjects with CNVs according to the two platforms were not statistically different at a liberal significance level (α) of 0.10, in all three ethnic groups.

Mapping Against Annotated Genes and Disease-Associated CNV Loci

We used UCSC gene annotation (http://genome.ucsc.edu/) to identify genes that are located within or partially overlap with CNV loci. To identify loci that warrant further investigation for their roles in complex disease, we identified CNV loci that overlapped with genes listed in the Online Mendelian Inheritance in Man (OMIM) Morbid Map (http://www.ncbi.nlm.nih.gov/omim/). CNV loci showing strong population differentiation are especially of interest because of their potential role in causing differences in disease risk between populations.

Results

Characteristics CNV Regions and Loci

After filtering unreliable CNV calls, we discovered about 45 CNVs per individual with a ratio of deletions to duplications of approximately 4:1 (Table 1). The majority of individuals have 20–60 CNVs in their genome (Supp. Fig. S2). There is very little between-group variations in term of the average number of CNVs and the ratio of deletions to duplications. These findings on the average number of CNVs and the ratio of deletions to duplications agree quite well with recent publication studying Korean population [Yim et al., 2010], but we found more CNVs per genome compared to previous research that used lower resolution arrays [de Stahl et al., 2008; Redon et al., 2006]. The median size of CNV was 20.8 kb in Chinese, 25.1 kb in Malay, and 16.4 kb in Indian. About 70% of the CNV regions were <50 kb and approximately 35% were <10 kb. In each ethnic group, the median size of deletions is about four- to fivefold smaller than the median size of duplications (Table 1).

We merge overlapping individual regions and identify 1,841 deletion and 732 duplication loci in these populations, which cover approximately 82 Mb and 89 Mb of the nucleotide sequence, respectively. Some of the deletion and duplication loci overlap, so in total we identify 2,379 unique CNV loci. A large proportion of deletion loci (80%) tend to be small (<50 kb) (Table 2). Conversely, duplication loci were much larger, with the majority (52%) between 50 Kb to 1 Mb (Table 2). Less than 10% of deletion loci and less than 5% of duplication loci can be considered common (population frequency $>5\%$) across the three populations (Fig. 1).

Using Affymetrix 6.0 platform we identify 1,514 deletions and 560 duplications loci. Comparing the CNV loci identified by the two platforms we find 752 (40.8%) deletions and 422 (57.8%)

Table 1. Summary Statistics of CNVs

	High confidence CNVs		
Statistics	Chinese ($n = 93$)	Malay ($n = 88$)	Indian ($n = 85$)
Average number of CNVs/individual	41.5	46.0	45.3
Range of number of CNVs/individual	28–78	27–110	31–75
Average number of markers/CNV	15	17	14
Average size of CNVs (kb)	55.6	64.2	53.0
Median size of CNVs (kb)	20.8	25.1	16.4
Size range of CNVs	41 bp–1,823 kb	41 bp–3,414 kb	16 bp–3,066 kb
Proportion of CNVs <50 kb	0.71	0.66	0.71
Proportion of deletions	0.81	0.80	0.82
Average size (kb)	39.9	47.1	41.5
Median size (kb)	14.8	17.9	11.8
Size range	41 bp–946 kb	41 bp–3,066 kb	16 bp–3,066 kb
Proportion of duplications	0.19	0.20	0.18
Average size (kb)	118.9	132.0	104.4
Median size (kb)	58.8	71.3	59.8
Size range	1.8–1823 kb	1.9–3,414 kb	1.2–1,628 kb

Total sample size: 266. CNV, copy number variations.

Table 2. Distribution of CNV Loci by Types and Size

	Deletion loci			Duplication loci		
Size (kb)	Total (%)	Replicated (%)	Novel (%)	Total (%)	Replicated (%)	Novel (%)
<1	34 (1.8)	1 (0.1)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
1–10	658 (35.7)	160 (21.3)	63 (22.1)	57 (7.8)	6 (1.4)	0 (0.0)
10–50	784 (42.6)	343 (45.6)	137 (48.1)	287 (39.2)	125 (29.6)	53 (34.0)
50–100	185 (10.0)	115 (15.3)	48 (16.8)	160 (21.9)	109 (25.8)	38 (24.4)
100–1,000	177 (9.6)	131 (17.4)	36 (12.6)	220 (30.1)	177 (41.9)	64 (41.0)
$>1,000$	3 (0.2)	2 (0.3)	1 (0.4)	8 (1.1)	5 (1.2)	1 (0.6)
Total	1841	752	285	732	422	156

CNV, copy number variations.

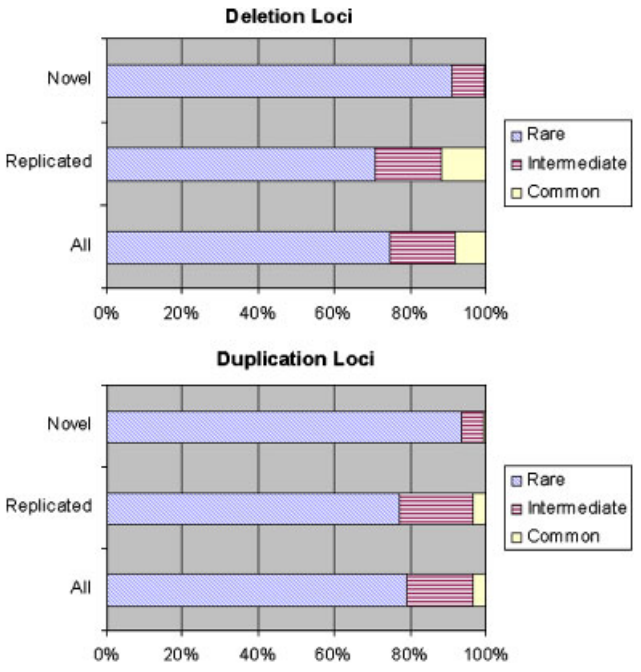


Figure 1. Distribution of CNV loci based on their frequency across the three populations (Common ≥ 5%, Intermediate = 1–5%, Rare ≤ 1%).

duplications loci identified using Illumina platform are replicated by Affymetrix platform (Supp. Table S1). The validation exercise was first done at population level (see Replicated CNV Loci section) without explicit requirement on the reproducibility at individual level. However, we did investigate the proportion of subjects who are detected to have CNV in the same locus by both platforms. For each replicated locus, we computed the proportion of subjects detected to have deletion (duplication) by Illumina and were also subsequently detected by Affymetrix to have the same form of CNV at the same locus. We found that across 752 replicated deletion loci, the average proportion is 89.1%. Meanwhile, across 422 replicated duplication loci, the average proportion is 75.4%. Hence, even though we did not explicitly require crossplatform reproducibility at sample level, it turns out that the rates of crossplatform reproducibility at sample level are relatively high.

Singletons constitute the majority of CNV loci that are not replicated, with 64.8% of nonreplicated deletion loci and 71.6% of nonreplicated duplication loci are singletons. There is a slight bias in the distribution of replicated deletion loci in that common and intermediate loci are more likely to be replicated (Fig. 1); however, almost no noticeable bias is observed for the duplication loci.

About five out of every six replicated deletion loci overlap by more than 80% of its length, with deletion loci independently detected using Affymetrix platform. The figure is even more impressive for replicated duplication loci, with almost 95% of them overlapping by more than 80% with their Affymetrix counterpart (Supp. Fig. S3). Hence, by using Affymetrix platform, we not only managed to replicate a significant proportion of the CNV loci but we are also able to confirm that the breakpoints for a vast majority of these loci are detected quite consistently across the two platforms.

The median confidence scores for nonreplicated deletion loci is significantly lower than the median confidence scores for the replicated loci (15.8 vs. 27.7, Kruskal-Wallis test, P-value < 0.001). The median confidence scores comparison for the duplication loci yields the same conclusion (15.9 vs. 28.9, Kruskal-Wallis test, P-value < 0.001). These results show that loci with higher confidence scores are more reliable and are more likely to be replicated. This provides an empirical justification to our decision in filtering unreliable calls based on the confidence score statistic.

A significant number of loci failed to be replicated. It is quite probable that this failure is partly due to differences in probe density across the genome; CNV loci located in the genomic area with dense markers tend to have their breakpoints well-estimated, and hence, more likely to be replicated across platforms. This is very possible given that our data show that in the nonreplicated deletion loci, on average, there is one marker every 22.3 kb, whereas in the replicated loci, there is more dense representation of markers, with a marker for every 2.8 kb (P-value < 0.001). The duplication loci also reveal a very similar pattern, with one marker in every 30.2 kb in the nonreplicated loci, whereas in the replicated loci, a marker is to be found for every 3.0 kb (P-value = 0.003).

Novel CNV Loci

We discover that almost 40% of the replicated deletion loci are novel, as only 467 out of the 752 loci are found in the DGV database (Supp. Table S1). However, out of the novel loci, only one locus in chromosome 1 can be considered common across the three populations with population frequency > 5% and a further 25 loci have population frequency between 1 and 5%. The large majority of these novel loci are relatively small, with just over 70% less than 50 kb.

Similar to the deletion loci, almost the same percentage of duplication loci is found to be novel (156/422). One of the novel duplication locus in chromosome 7 has a population frequency > 5%, with a further nine novel loci having a population frequency between 1 and 5%. Unlike the novel deletion loci, the novel duplication loci tend to be larger, with only 91 loci (58.3%) less than 100 kb.

Crucially these novel CNV loci are detected with the same degree of confidence as those CNV loci that overlap with previously published loci. In fact, the median confidence score for these 441 novel CNV loci (285 deletions, 156 duplications) is statistically higher than the other replicated loci (32.9 vs. 26.3, Kruskal-Wallis test, P-value = 0.001). This result indicates that these novel loci are unlikely to be false positives.

The vast majority of these novel loci are population specific (Fig. 2). The highest number of population-specific deletion loci is found in the Indian population (111 loci), whereas Malay have the highest percentage of novel duplication loci (75 loci). Only 30 deletion loci and 11 duplication loci exist in multiple populations, with more than half of the shared novel loci existing only in the Chinese and Malay.

Although the frequencies of the vast majority of these novel loci are relatively low across the three populations, there are still some loci that are common (> 5% frequency) in specific populations. Specifically, there are five novel loci common among the Indians, four novel loci common among the Malays, and three novel loci common among the Chinese (Table 3).

Population Differentiation

The median V_{st} statistic between Chinese and Malay populations, computed across 1,174 replicated loci is 0.016, lower than the corresponding comparisons for Chinese and Indian populations

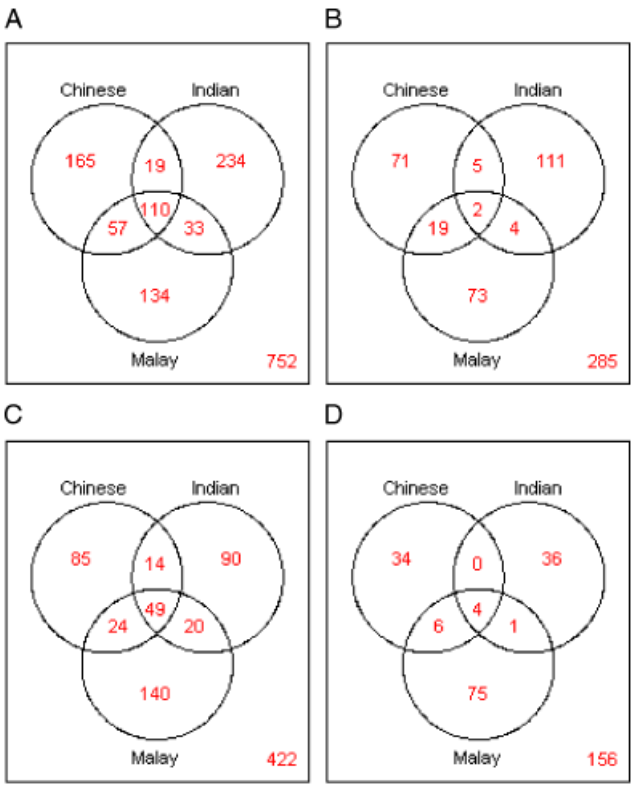


Figure 2. Venn diagrams showing distribution of replicated and novel CNV loci across the three ethnic groups. **A:** Replicated deletion loci, **B:** Novel deletion loci, **C:** Replicated duplication loci, **D:** Novel duplication loci.

Table 3. List of Novel CNV Loci Common to at Least One Population

Chr.	Start	End	Form of CNV	Chinese ($n = 93$)	Malay ($n = 88$)	Indian ($n = 85$)
1	173,064,490	173,135,447	Del	34	13	13
2	49,387,002	49,401,059	Del	8	4	0
2	72,102,919	721,33,022	Del	2	7	0
2	137,759,660	137,783,206	Del	0	1	6
6	40,174,655	40,204,896	Del	0	0	5
7	37,124	164,003	Dup	13	9	14
8	9,091,324	9,099,900	Del	0	0	7

(median = 0.035) and Indian and Malay (median = 0.028). Over the whole genome, Indian is more differentiated from Chinese and Malay populations (Supp. Fig. S4).

Using the PennCNV-inferred integer copy number we then compare the estimated proportion of subjects with copy-number variations within each ethnic group. We find that out of all replicated deletion and duplication loci, 676 deletion loci (89.9%) and 385 duplication loci (91.2%) have their copy-number frequency estimated consistently across the three ethnic groups. Among these loci with consistent estimated copy-number frequencies, we identified 27 deletion loci that segregate at significantly different frequency across the three populations (Supp. Table S1). Among the 27 deletion loci, there is only one novel locus that has significantly different copy-number distribution across the ethnic groups. This locus is an 8.6-kb deletion in Chromosome 8 that is found exclusively in the Indian population.

Mapping Against Annotated Genes and Disease-Associated CNV Loci

Compared to duplication loci, we find an appreciably lower percentage of deletion loci that overlap known genes or uncharacterized transcripts in the UCSC database (48.8 vs. 74.6%, P-value < 0.001). The percentage of novel deletion and duplication loci that overlap with UCSC genes is only slightly lower at 44.2 and 66.0%, respectively. This bias of deletion away from the genes is also observed previously by Redon et al. [2006] and Conrad et al. [2009]. Most of the 367 deletion loci that overlap UCSC genes are rare (66.8%); however, there are 66 (18.0%) of deletion loci with an intermediate frequency between 1 and 5% and 56 (15.3%) common deletion loci that also overlap UCSC genes. The number of deletion loci that overlap UCSC genes are highest among Indians (212 loci), followed by Chinese (182 loci) and Malays (176 loci). There are 62 novel deletion loci among Indians that overlap UCSC genes, whereas the number of novel deletion loci overlapping UCSC genes in Chinese and Malays are 41 and 40, respectively.

Likewise, we find 229 (72.7%) out of the 315 duplication loci overlapping UCSC genes are rare, 23.5% are intermediate, and 3.8% are common. Malays have the highest number of duplication loci overlapping UCSC genes (190 loci), followed by Chinese (134 loci) and Indians (132 loci). Among the 315 loci, there are 103 novel loci that overlap UCSC genes, with 61 of these novel loci are observed in Malay, 29 loci observed in Chinese and 24 loci observed in Indian.

We find several complex disease-associated genes overlapping the common CNV loci (population frequency > 5%), including *FCGR3B* (MIM# 610665) [Fanciulli et al., 2007], beta-defensin genes [Hollox et al., 2008], *UGT2B17* (MIM# 601903) [Park et al., 2006], *CCL3L1* (MIM# 601395) [Gonzalez et al., 2005], and a number of drug-related genes such as *CYP2A6* (MIM# 122720) and *CYP2A7* (MIM# 608054). From the list of deletion loci with strong population differentiation, several loci mapped to a few important genes involved in the metabolism of exogenous (drug) and endogenous compounds: *CYP2A6*, *CYP2A7*, *UGT2B17*, and *UGT2B15* (MIM# 600069). Previous research has shown these loci to be common [Ouahchi et al., 2006], and our findings confirm this (Supp. Table S2).

One of the novel deletion loci overlaps the *RABGAP1L* gene in chromosome 1 whose polymorphism has been suggested to be associated with increased risk of hypertension among Japanese [Oguri et al., 2010]. This deletion is particularly common among Chinese with a population frequency of 36.6%. Meanwhile, among Malays and Indians this deletion is only observed in 15–20% of the population.

Out of the 367 deletion loci that overlap UCSC genes, we also find 54 deletion loci (14.7%), with 24 of them being rare novel loci that overlap genes in OMIM Morbid Map. The common deletion on the *CYP2A6* gene is particularly interesting because it has been associated with reduced susceptibility to lung cancer in Japanese population [Miyamoto et al., 1999]. Interestingly, among our three populations the highest frequency of this deletion is found in the Malay population (35.2%), followed by the Chinese (18.3%) and Indian (7.1%). The role of this deletion in altering risk of lung cancer in the Singaporean population definitely needs to be further investigated, especially given that Singaporean Malay males have the highest smoking prevalence, yet their age-standardized lung cancer rates are lower than Chinese males [Singapore Cancer Registry, 2008].

The percentage is higher for duplication loci, with 71 (22.5%) out of 315 loci that overlap with UCSC genes also overlapping

with genes in the Morbid Map. Twenty-four of these duplication loci are novel; one locus with intermediate frequency in Chromosome 17 overlapped with *ASPSCR1* (MIM# 606236), a candidate gene for Alveolar soft part sarcoma.

Across the three ethnic groups, the number of deletion loci overlapping genes in the Morbid Map relative to the number of loci overlapping UCSC genes are 16.0% among Indians, and slightly lower among Malays (14.2%) and Chinese (13.7%). For duplication loci, the corresponding ethnic-specific percentage is 25.8% in the Malay population, followed by Indian (21.2%) and Chinese (20.1%).

Discussion

Using the Illumina 1M Beadchip array, we investigate copy-number characteristics of three South East Asian populations, namely, Chinese, Malay, and Asian Indian populations in Singapore. We genotype 266 individuals from the three populations and discover an average of 45 CNV regions per genome, with very little variations in terms of average CNV per genome between the three populations. This figure is higher than the number of CNVs found in previous studies that used lower density SNP arrays [Pinto et al., 2007; Redon et al., 2006; Wang et al., 2007]. We attribute our ability to detect more CNV regions mostly due to the higher density of markers used.

To filter unreliable CNV calls, we used confidence score statistics produced by the PennCNV software. Other studies have used different criteria to establish reliable CNV data. For example, Jakobsson et al. [2008] restricted analyses to CNVs with a minimum of 10 markers per CNV to minimize the number of false positives. Our empirical data show that higher confidence scores are associated with regions with higher number of markers as well higher marker density (number of markers per bp). Hence, placing a threshold on the confidence scores is in a way equivalent to indirectly placing a threshold on the number of markers. One advantage of our approach is that it enables us to identify relatively small CNV regions if the region is detected with high confidence, as long as there are at least three markers in the region, as this is a PennCNV default requirement for minimum number of markers. This identification would not have been possible if we had used number of markers as the threshold.

There are a higher proportion of duplications with confidence scores below the reliable threshold. This is due to the technological limitation of SNP array. Deletions are easier to detect than duplications, because the exponential of intensity data is linearly correlated with the copy number [Wang et al., 2007]. Consequently, the signal intensity difference between for example, one-copy deletion and diploid-copy is more pronounced than the difference between diploid-copy and one-copy duplication. Another important implication of this limitation is that heterozygous (one copy) deletion would be harder to detect, hence, more likely to be missed by the algorithm, than homozygous (two copies) deletion. In our study, we find an underrepresentation of heterozygous for small deletion regions. Approximately 82% of called deletions of size >10 kb are heterozygous but only approximately 73% of deletions <10 kb are heterozygous.

To overcome the challenge of estimating breakpoints, we merge overlapping CNV regions into nonoverlapping CNV loci, in a similar manner to Redon et al. [2006]. In total, we found 1,841 deletion and 422 duplication loci across the genome. About 40% of these deletion loci and close to 60% of these duplication loci found using the Illumina platform are replicated independently using the Affymetrix 6.0 platform.

We do not attempt to make joint-calling of CNV regions using data from both platforms simultaneously. Although joint-calling would probably increase the sensitivity of the CNV calls, our first priority is to reduce false discoveries, which we believe has been achieved by our more conservative method in defining replicated regions. Among the replicated loci, there are 285 novel deletion and 156 novel duplication loci not previously reported in DGV. Most of these novel loci are small (<50 kb) and are population specific; however, 30 novel deletion loci and 11 duplication loci are to be found in more than one population, with Chinese and Malay being the two populations that most frequently "share" a locus. Our findings that most of the novel loci are <50 kb support earlier prediction that there are likely to be a plethora of undiscovered CNV at this size [Estivill and Armengol, 2007].

We find tens of CNV loci that overlap genes in the OMIM Morbid Map with some of the loci being novel variants. Several of these CNV are quite common and interestingly segregate at different frequency across the three populations. These loci overlap with genes that have been previously linked to phenotypes such as drug metabolisms, lung cancer, and hypertension. This finding further highlights the need to better characterize the role of CNV in the aetiology of complex diseases and drug response. The fact that some of these loci are novel emphasizes the need to have a more complete CNV map for Asian populations, something to which we believe this article contributes. In addition, it is very likely that some of the CNV loci previously found in other populations have very different copy-number forms and distribution in our study population.

The individual-level CNV data from this study is downloadable from the SGVP Website and the population-level data are available in the Supp. Table S1. Researchers interested in obtaining the raw intensity data should e-mail us at ephcks@nus.edu.sg or statyy@nus.edu.sg.

We believe our findings and datasets can serve as resources to contribute further research into building a global CNV map as well as to stimulate research on the role of CNV in the genetic risk of complex diseases in Asian populations in general and South East Asia in particular.

Acknowledgments

The Yong Loo Lin School of Medicine, the Life Science Institute and the Office of Deputy President (Research and Technology), National University of Singapore. We also acknowledge the technical and financial support of the Genome Institute of Singapore, and Agency for Science, Technology and Research, Singapore. The authors appreciate the support of the Singapore Clinical Research Institute for the editorial assistance provided by Jon Kilner, MS, MA (Pittsburgh, PA).

References

Bellwood P, Fox JJ, Tryon D, editors. The Austronesians: historical and comparative perspectives. Canberra, Australia: Australian National University. 1995.

Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, Zhang F, Seielstad M, Zeng YX, Zhang X, Liu J. 2009. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet 85:775–785.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, The Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. 2009. Origins and functional impact of copy number variation in the human genome. Nature 2009 Oct 7 [Epub ahead of print].

de Stahl TD, Sandgren J, Piotrowski A, Nord H, Andersson R, Menzel U, Bogdan A, Thuresson AC, Poplawski A, von Tell D, Hansson CM, Elshafie AI, Elghazali G, Imreh S, Nordenskjöld M, Upadhyaya M, Komorowski J, Bruder CE,

Dumanski JP. 2008. Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array. Hum Mutat 29:398–408.

Estivill X, 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.

Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, Heward JM, Gough SC, de Smith A, Blakemore AI, Froguel P, Owen CJ, Pearce SH, Teixeira L, Guillevin L, Graham DS, Pusey CD, Cook HT, Vyse TJ, Aitman TJ. 2007. FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. Nat Genet 39:721–723.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'Connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440.

Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, van de Kerkhof PC, Traupe H, de Jongh G, den Heijer M, Reis A, Armour JA, Schalkwijk J. 2008. Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 40:23–25.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36:949–951.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003.

Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, He L, Zhu X, Deng H, Levy S, Papasian CJ, Drees BM, Hamilton JJ, Recker RR, Cheng J, Deng HW. 2009. Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. PLoS ONE 4:e7958.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40:1166–1174.

Miyamoto M, Umetsu Y, Dosaka-Akita H, Sawamura Y, Yokota J, Kunitoh H, Nemoto N, Sato K, Ariyoshi N, Kamataki T. 1999. CYP2A6 gene deletion reduces susceptibility to lung cancer. Biochem Biophys Res Commun 261:658–660.

Oguri M, Kato K, Yokoi K, Yoshida T, Watanabe S, Metoki N, Yoshida H, Satoh K, Aoyagi Y, Nozawa Y, Yamada Y. 2010. Assessment of a polymorphism of SDK1 with hypertension in Japanese individuals. Am J Hypertens 23:70–77.

Ouahchi K, Lindeman N, Lee C. 2006. Copy number variants and pharmacogenomics. Pharmacogenomics 7:25–29.

Park J, Chen L, Ratnashinge L, Sellers TA, Tanner JP, Lee JH, Dossett N, Lang N, Kadlubar FF, Ambrosone CB, Zachariah B, Heysek RV, Patterson S, Pow-Sang J. 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.

Pinto D, Marshall C, Feuk L, Scherer SW. 2007. Copy-number variation in control population cohorts. Hum Mol Genet 16:R168–R173.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME. 2006. Global variation in copy number in the human genome. Nature 444:444–454.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.

Saw SH. 2007. The population of Singapore, 2nd edition. Singapore: Institute of South East Asian Studies.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Mānér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. 2004. Large-scale copy number polymorphism in the human genome. Science 305:525–528.

Singapore Cancer Registry. 2008. Trends in cancer incidence in Singapore 2002–2006. Interim Report. Singapore: Singapore Cancer Registry.

Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, Small KS, Ku CS, Lee EJ, Seielstad M, Chia KS. 2009. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res 19:2154–2162.

Wain LV, Armour JA, Tobin MD. 2009. Genomic copy number variation, human health, and disease. Lancet 374:340–350.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. 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.

Yim SH, Kim TM, Hu HJ, Kim JH, Kim BJ, Lee JY, Han BG, Shin SH, Jung SH, Chung YJ. Copy number variations in East-Asian population and their evolutionary and functional implications. Hum Mol Genet. 2010 Jan 15 [Epub ahead of print].

Journal of Human Genetics (2011) 56, 524–533

www.nature.com/jhg

ORIGINAL ARTICLE

A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals

Shu-Mei Teo[1,2,3,5], Chee-Seng Ku[2,5], Nasheen Naidoo[2], Per Hall[1], Kee-Seng Chia[1,2,4], Agus Salim[4] and Yudi Pawitan[1]

The abundance of copy number variants (CNVs) and regions of homozygosity (ROHs) have been well documented in previous studies. In addition, their roles in complex diseases and traits have since been increasingly appreciated. However, only a limited amount of CNV and ROH data is currently available for the Swedish population. We conducted a population-based study to detect and characterize CNVs and ROHs in 87 randomly selected healthy Swedish individuals using the Affymetrix SNP Array 6.0. More than 600 CNV loci were detected in the population using two different CNV-detection algorithms (PennCNV and Birdsuite). A total of 196 loci were consistently identified by both algorithms, suggesting their reliability. Numerous disease-associated and pharmacogenetics-related genes were found to be overlapping with common CNV loci such as CFHR1/R3, LCE3B/3C, UGT2B17 and GSTT1. Correlation analysis between copy number polymorphisms (CNPs) and genome-wide association studies-identified single-nucleotide polymorphisms also indicates the potential roles of several CNPs as causal variants for diseases and traits such as body mass index, Crohn's disease and multiple sclerosis. In addition, we also identified a total of 14 815 ROHs ⩾500 kb or 2814 ROHs ⩾1M in the Swedish individuals with an average of 170 and 32 regions detected per individual respectively. Approximately 141 Mb or 4.92% of the genome is homozygous in each individual of the Swedish population. This is the first population-based study to investigate the population characteristics of CNVs and ROHs in the Swedish population. This study found many CNV loci that warrant further investigation, and also highlighted the abundance and importance of investigating ROHs for their associations with complex diseases and traits.
Journal of Human Genetics (2011) **56,** 524–533; doi:10.1038/jhg.2011.52; published online 2 June 2011

Keywords: Affymetrix SNP Array 6.0; Birdsuite; copy number variants; PennCNV; regions of homozygosity; Swedish population

INTRODUCTION

There is a growing body of copy number variant (CNV) maps covering different world populations.[1–5] Most of these newer studies used high-resolution methods for detecting CNVs, such as the Affymetrix SNP Array 6.0, which has a higher density of single-nucleotide polymorphism (SNP) and copy number probes than previous microarray-based methods. This has led to an improved performance of microarray-based methods to detect smaller CNVs (<50 kb).[1,6] In contrast, previous studies have used much lower resolution arrays, such as the bacterial artificial chromosome (BAC) clone or oligonucleotide comparative genomic hybridization arrays and SNP genotyping arrays.[7–10] Currently, there is only one CNV-detection study in a Swedish population,[10] but this was performed in a small sample size of 33 individuals and used a low-resolution 32-K bacterial artificial chromosome clone microarray. This has hampered the study from detecting less common and smaller CNVs and from estimating the population frequency of CNVs. The ability to detect smaller CNVs is critical as they are more numerous than the larger CNVs.[11]

In addition, the study by Díaz de Ståhl *et al.*[10] was unable to detect regions of homozygosity (ROHs) as the bacterial artificial chromosome clone microarray was unable to generate allelic intensity data. Research on ROHs has started to gain impetus, as evidenced by the increasing number of publications after the first study by Gibson *et al.*[12] reported the abundance of ROHs in the human genome of outbred populations. Further studies have investigated the population characteristics of ROHs in healthy individuals,[13–15] and also performed association analyses to identify ROHs that are associated with complex diseases and traits in a case–control study design.[16–18]

To circumvent the limitations of the previous study by Díaz de Ståhl *et al.*,[10] we conducted a study in a Swedish population by genotyping 100 individuals using the Affymetrix SNP Array 6.0 (Affymetrix, Santa Clara, CA, USA). The main aim of this study was to perform a more comprehensive detection of CNVs and ROHs in the Swedish

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore and [4]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[5]Joint first author.
Correspondence: C-S Ku, Center for Molecular Epidemiology, National University of Singapore, Singapore 117597, Singapore.
E-mail: csikcs@nus.edu.sg or Professor Y Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden.
E-mail: Yudi.Pawitan@ki.se
Received 18 January 2011; revised 12 April 2011; accepted 25 April 2011; published online 2 June 2011

population and to describe their population characteristics. Although several studies have been performed to detect and characterize CNVs and ROHs in multiple European populations, these studies have also documented the genetic differences among these populations.[14,15,19] The extension of the International HapMap Project to include an additional seven populations in Phase III further suggests that multiple populations from diverse ancestries or different geographical locations are needed to study their population genetics.[20] These previous studies have justified the need for a population-based study to characterize CNVs and ROHs in healthy Swedish individuals. We also compared the Swedish population with the HapMap phase III populations using principal component analysis.

MATERIALS AND METHODS

Samples and genotyping platform

A total of 100 randomly selected healthy Swedish individuals volunteering as controls in case–control studies were studied. Peripheral blood samples of the participants for genomic DNA extraction were drawn and stored at the Karolinska Biobank. Identities of the participants were kept anonymous and no personal identifiers were used. All 100 samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 as per the manufacturer's protocol. Two samples were removed from further analysis because their genotype call rates were below 98% and the remaining 98 samples were used for CNV detection.

CNV-detection algorithms and analyses

CNV calling using PennCNV. We used two CNV-detection algorithms, namely PennCNV[21] and Birdsuite,[22] for both comparison and validation. This study focused only on the CNVs in the 22 autosomes because of the inaccuracy of Birdsuite to detect CNVs in sex chromosomes. Log *R* ratio and B allele frequency were calculated according to the PennCNV algorithm (http://www.openbioinformatics.org/penncnv/penncnv_tutorial_affygw6.html). Smaller CNVs (<1 kb) were also included in our analysis, as PennCNV by default does not limit its detection to CNVs >1 kb in size. We applied a set of filtering criteria as recommended by the algorithm, namely Log *R* ratio-s.d >0.35, B allele frequency-median >0.55, B allele frequency-median <0.45 and B allele frequency-drift >0.006 to exclude samples with poor quality of signal intensity data (http://www.openbioinformatics.org/penncnv/). This resulted in a further exclusion of 11 samples, with the final set for analysis consisting of 87 samples. For each sample, PennCNV generated a list of CNVs with their confidence scores. The confidence score is a log Bayes factor that measures the likelihood that the locus harbors an abnormal copy number. A confidence score of $\geqslant 10$ has been recommended as the threshold to classify reliable CNVs. Therefore, we retained all CNVs called with confidence scores $\geqslant 10$ for subsequent analyses. Although the confidence score is only a statistical measure of a true positive, our previous study[5] found that CNVs with a higher confidence score are more likely to be detected consistently across two genotyping platforms. Therefore, this justifies our decision to retain only reliable CNVs called with a sufficient degree of confidence.

Construction of CNV loci using PennCNV output. The CNVs called by PennCNV were shown to overlap across samples. Thus, we merged or grouped these individual CNV calls into discrete, non-overlapping loci, with the boundaries of each locus determined by the union of all CNVs that belonged to that particular locus. This construction of CNV loci was needed to estimate the population frequencies and these steps were performed using the methods that we have developed previously.[5,23] We classified the status of these CNV loci into three categories, 'del' (loci containing deletions), 'dup' (loci containing duplications) and 'del/dup' (loci containing both deletions and duplications).

Copy number polymorphism (CNP) calling using Canary (Birdsuite). Birdsuite software[22] was also used to analyze the Affymetrix SNP Array 6.0 data. There are two components in the software for detecting copy number changes, namely Canary and Birdseye. Canary was used to determine the integer copy number at each of the predefined 1316 CNPs. The term 'CNPs' used by McCarroll *et al.*[1] is to describe common CNV loci. These 1316 CNPs were found in more than one HapMap II individual and their sizes were also accurately determined. Therefore, we used the Canary component in Birdsuite to determine the integer copy number of the 1316 CNPs in the 87 Swedish samples. These 1316 CNPs are distributed in all the autosomes and sex (X and Y) chromosomes. However, 25 CNPs located in the sex chromosomes were removed because the CNP calling in these chromosomes was less accurate. Thus, the results reported in this study comprised only 1291 CNPs in the 22 autosomes. Confidence statistics generated for the CNPs were also used to identify poor-quality calls, and only integer copy numbers detected with high confidence as recommended by the software (confidence score >0.1) were used for subsequent analyses.

Correlation analysis of CNPs. We performed a correlation analysis of CNPs and the nearby SNPs. Because the sizes of the CNPs were previously accurately determined by McCarroll *et al.*,[1] we restricted the analysis to only the CNPs detected by Canary. For each of the 1291 CNPs, SNPs within a 200-kb window from the start and end positions of the CNP were considered. We used the squared Pearson's correlation (r^2) for correlation analysis. The genotype calling of the Affymetrix SNP Array 6.0 was carried out using Birdsuite. In addition, to investigate the potential associations of CNPs with human diseases and traits, the same methods of r^2 calculations for the 1291 autosomal CNPs and the SNPs that were identified by genome-wide association studies (GWAS) were adopted. The list of GWAS-SNPs was downloaded from the National Human Genome Research Institute website (http://www.genome.gov/gwastudies/) on 26 October 2010.

CNV calling using Birdseye (Birdsuite). In addition to PennCNV, we also used another algorithm, Birdseye, to analyze the same set of data as different algorithms tend to have different sensitivities and specificities for detection of CNVs in different regions throughout the genome. As such, CNV loci detected by PennCNV and Birdseye can be cross-validated. Therefore, we used the Birdseye component in Birdsuite to detect additional CNVs throughout the genome, which was not restricted to the 1316 predefined CNPs. Similarly, only CNVs in autosomal chromosomes were used because of the inaccuracy of Birdseye in the sex chromosomes. CNVs with low confidence, as recommended by the software (confidence score $\leqslant 5$), were removed from subsequent analysis.

Construction of CNV loci using Birdseye output. We also constructed CNV loci based on the Birdseye output using methods similar to those applied to the PennCNV output. The cutoff for the confidence score used by PennCNV ($\geqslant 10$) and Birdseye ($\geqslant 5$) was recommended by both algorithms. This allowed for greater comparability between the CNV loci detected by these two algorithms.

Comparison of CNV loci detected by PennCNV and Birdsuite. The CNV loci identified by PennCNV and Birdseye were compared as a 'validation' step. We used a 'reciprocal 50% overlapping' method to compare the CNV loci detected by these two algorithms and considered a CNV locus 'found' by both algorithms when this locus was detected in both PennCNV and Birdseye with an overlap of $\geqslant 50\%$ of their lengths.

Novel CNV loci. To identify novel CNV loci, we compared the CNV loci detected by PennCNV and Birdseye with the data from the Database of Genomic Variants (DGV).[24] We used the latest data from the DGV (variation.hg18.v8.aug.2009.txt and indel.hg18.v8.aug.2009.txt) downloaded from the DGV Website (http://projects.tcag.ca/variation/). A CNV locus identified by PennCNV and Birdseye was considered novel if it did not share at least 50% of its length with any CNV loci cataloged in the DGV. All the downstream analyses after PennCNV and Birdsuite were performed using the statistical software package R (http://www.r-project.org/).

Comparison with HapMap phase III populations

The CEL files of the Affymetrix SNP Array 6.0 for the seven populations in HapMap phase III project were downloaded from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). The HapMap phase III populations studied are people of African ancestry in the southwestern USA (ASW), the Chinese community in Metropolitan Denver, Colorado, USA (CHD), Gujarati Indians in Houston, Texas, USA (GIH), the Luhya in Webuye,

Kenya (LWK), people of Mexican ancestry in Los Angeles, California, USA (MEX), the Maasai in Kinyawa, Kenya (MKK) and the Tuscans in Italy (TSI). All the samples were analyzed using Canary similarly to the analysis of the Swedish population. Only unrelated samples were included in our study, that is, family-related samples were removed using the 'relationships' file provided by the International HapMap Project. After the sample exclusion step, a total of 594 unrelated samples from the seven HapMap III populations were analyzed: ASW (n=52), CHD (n=89), GIH (n=89), LWK (n=90), MEX (n=53), MKK (n=132) and TSI (n=89). We performed principal component analysis to compare the Swedish population with the HapMap phase III populations using the CNP output generated by Canary.

ROH-detection algorithms and analyses

In addition to CNVs, we also detected ROHs using PennCNV in the 22 autosomes of the 87 Swedish individuals. However, we only focused on ROHs $\geqslant$500 kb, as this cutoff was adopted in a previous study.[18] For each of these we confirmed that they are ROHs by determining the genotypes of the SNPs that fall within each region. We then calculated the percentage of heterozygosity (number of heterozygotes/total number of heterozygotes and homozygotes). We also calculated the percentage of missingness genotypes (number of missingness/total number of SNPs in each ROH). First, we used an arbitrary cutoff of the median of the percentage of heterozygosity (2.5%) to allow for some heterozygote calls resulting from calling or genotyping errors. As a result, we removed half of the ROHs with a percentage $>$2.5%. Second, we removed ROHs with $>$1% for the missingness, to remove regions where genotype calling was problematic. Finally, for the remaining ROHs, we also ensured a density of one SNP per 10 kb to exclude those ROHs that could be spuriously detected by a sparse number of SNPs. As such, for a 500-kb ROH, a minimum of 50 SNPs is required. These three criteria were used as the filters to exclude less reliable ROHs. Several summary statistics were then computed to describe the characteristics of ROHs in the Swedish population.

RESULTS

Characteristics of CNVs identified by PennCNV

After filtering unreliable CNV calls, an average of approximately 36 CNVs per individual with a ratio of deletions to duplications of approximately 2.6:1 was discovered (Supplementary Table 1). The number of CNVs per individual ranged from 22 to 65. The median size of a CNV was 28.6 kb and approximately 66% of the CNVs were $<$50 kb and 26% were $<$10 kb (Supplementary Figure 1). The median size of deletions was approximately fourfold smaller than the median size of duplications.

Characteristics of CNV loci identified by PennCNV

We merged overlapping CNVs to construct CNV loci and identified 623 loci, of which 476 loci contained deletions ('del-loci'), 102 loci contained duplications ('dup-loci') and 45 loci contained both deletions and duplications ('del/dup-loci'; Table 1). These 623 loci covered approximately 61.52 Mb of the nucleotide sequence and the sum of the lengths for del-loci (19.83 Mb) was smaller than that for dup-loci (25.80 Mb). Similarly for the individual CNVs (Supplementary Table 1), the average size of del-loci (41.66 kb) was much smaller than that of dup-loci (252.93 kb; Table 1). More than 77% of the del-loci were $<$50 kb, and in comparison only 22.55% of dup-loci were within this size range. The majority (62.75%) of dup-loci ranged from 50 to 500 kb. In summary, there were far more del-loci, but their sizes tended to be smaller than those of dup-loci. A list of the 623 loci is shown in Supplementary Table 2.

Of the 623 CNV loci, 268 loci were detected in $\geqslant$2 individuals (Table 1). The remaining loci were detected in only one individual; these loci were not necessarily 'singleton loci' as we only studied

Table 1 Summary statistics of CNV loci constructed from PennCNV output

Summary statistics of CNV loci (PennCNV output)	*Total*	*Del*	*Dup*
Number of CNV loci	623	476 (76.40%)[a]	102 (16.37%)[a]
Number of CNV loci detected in $\geqslant$2 individuals	268 (43.02%)[b]	194 (40.76%)[b]	29 (28.43%)[b]
Sum of the length of loci (Mb)	61.52	19.83	25.80
Average length per locus (kb)	98.75	41.66	252.93
Average number of markers per locus	58	34	141
Size distribution			
$<$10 kb	141 (22.63%)	132 (27.73%)	6 (5.88%)
$\geqslant$10–$<$50 kb	265 (42.54%)	236 (49.58%)	17 (16.67%)
$\geqslant$50–$<$100 kb	79 (12.68%)	54 (11.34%)	21 (20.59%)
$\geqslant$100–$<$500 kb	110 (17.66%)	52 (10.92%)	43 (42.16%)
$\geqslant$500 kb	28 (4.49%)	2 (0.42%)	15 (14.71%)
Overlapping with DGV			
CNV loci that overlap	388 (62.28%)	298 (62.61%)	54 (52.94%)
CNV loci that did not overlap	235 (37.72%)	178 (37.39%)	48 (47.06%)
Overlapping with UCSC genes			
CNV loci that overlap	202 (32.42%)	135 (28.36%)	51 (50.00%)
CNV loci that did not overlap	421 (67.58%)	341 (71.64%)	51 (50.00%)
Overlapping with CNV loci from Birdseye data and consistent in CNV status that is, del/dup/del+dup			
CNV loci that overlap	196 (31.46%)	160 (33.61%)	30 (29.41%)
CNV loci that did not overlap	427 (68.54%)	316 (66.39%)	72 (70.59%)

Abbreviations: CNV, copy number variant; DGV, database of genomic variants; UCSC, University of California Santa Cruz genes.
[a]The percentage was calculated by dividing 623 loci.
[b]The percentage was calculated by dividing 623, 476 and 102 loci, respectively.
Note: As there are only 45 CNV loci (7.22%) with status del+dup, the summary statistics of these loci are not shown in the table. A full colour version of this Table is available at the Journal of Human Genetics Journal online.

87 individuals. The proportion of del-loci detected in $\geqslant 2$ individuals (40.76%) was much higher than the proportion for dup-loci (28.43%). Among the high-frequency CNV loci (loci that were detected in multiple individuals), several overlapped with disease-related genes such as *WWOX* and *ERBB4* (gastric and pancreatic cancers and melanoma)[25–27] and *CACNA1C* (bipolar disorder)[28] or drug-metabolizing genes such as *GSTT1*[29] (Supplementary Table 2). For example, a deletion locus overlapping with *WWOX* (a tumor suppressor gene) was detected in 24 of the 87 individuals (27.6%), and a deletion locus encompassing *GSTT1* was deleted at a population frequency of 13.8%. In addition, the proportion of del-loci encompassing the UCSC genes (28.36%) was much lower than dup-loci (50.00%) overall.

Detection of CNVs using microarrays is usually plagued with poor specificity or a high false-positive rate. In an effort to validate the 623 CNV loci constructed from the PennCNV output, we compared them with the CNV loci detected by Birdseye. We found 196 loci (31.46%) with $\geqslant 50\%$ reciprocal overlap with the Birdseye data and the status of 'del', 'dup' and 'del/dup' of the 196 loci were consistent with the Birdseye data. For the remaining 427 CNV loci that were not confirmed by Birdseye data, we found that 247 loci had been cataloged in the DGV (please see Materials and methods). Therefore, by applying two different ways of validation, 443 (71.1%) of the 623 CNV loci detected by PennCNV were considered reliable in this study (Table 1).

Characteristics of CNPs identified by Canary (Birdsuite)

Approximately 49.81% of the 1291 autosomal CNPs were non-polymorphic in the Swedish population (Supplementary Table 3). The population frequency distribution pattern of the 1291 CNPs is shown in Supplementary Figure 2. Among the polymorphic loci (648 CNPs) and non-polymorphic CNPs (643 loci) in the Swedish population, 289 loci (44.60%) and 255 loci (39.66%) overlapped with genes or entries from the UCSC annotation of the human genome, respectively. No substantial difference was observed between the polymorphic and non-polymorphic loci.

The majority of the 648 polymorphic CNPs were biallelic (545 CNPs or 84.1%), of which the integer copy numbers were either exclusively deletions, that is, copy number of 0 or 1 (387 CNPs or 59.7%), or exclusively duplications, that is, copy number of 3 or 4 (158 CNPs or 24.4%). Among the biallelic 545 CNPs, only one showed significant deviation from HWE at an FDR <0.01.

Numerous CNPs were found to overlap with important known disease- or pharmacogenetics-related genes (Table 2). The frequencies of these CNPs ranged from relatively uncommon (2.78% for CNP118) to completely polymorphic (100% for CNP88). For example, CNP88 overlapped with GSTM1 and GSTM2 was found to be completely deleted in the Swedish population, where all except one carried two-copy deletions. However, it is noteworthy that in approximately half of the sample (47 individuals), the integer copy numbers were successfully determined with high confidence scores. In addition, high deletion frequencies were also found for CNPs overlapping with other GST enzymes such as GSTT1 (60.00%), GSTT2, GSTT2B and GSTTP1 (98.65%). Two-copy deletion was common for these enzymes—17.6% of the individuals for GSTT1 (CNP2560) and 43.2% for the other GST enzymes (CNP2559).

Besides these phase II metabolizing enzymes, several disease-associated genes were also found to overlap with these CNPs, such as the FCG receptor genes (autoimmune or inflammatory diseases),[30] *TP63*[31] and *WWOX*[26] (lung adenocarcinoma, gastric, pancreatic and other cancers), *CFHR3* and *CFHR1* (age-related macular degeneration),[32] *UGT2B17* (prostate cancer and graft-versus-host disease),[33,34]

Table 2 CNPs that overlap with important and known disease- and pharmacogenetics-related genes

CNP ID	*CN=0*	*CN=1*	*CN=2*	*CN=3*	*CN=4*	*Frequency*	*Chromosome*	*Start*	*End*	*Length*	*UCSC gene (disease/trait)*
118	0	1	70	0	1	2.78	1	159 778 034	159 906 183	128 149	FCGR3A, FCGR2B, FCGR2C, FCGR3B (autoimmune or inflammatory diseases)
11164	0	1	83	2	0	3.49	6	162 658 558	162 660 430	1872	PARK2, parkin (Parkinson's disease)
530	1	10	71	0	0	13.41	3	190 846 372	190 847 332	960	TP63 (cancers)
147	3	31	53	0	0	39.08	1	194 997 658	195 068 695	71 037	CFHR3, CFHR1 (age-related macular degeneration)
603	8	33	46	0	0	47.13	4	69 043 083	69 168 574	125 491	UGT2B17 (prostate cancer, graft-versus-host disease)
2560	15	36	34	0	0	60.00	22	22 680 529	22 726 814	46 285	GSTT1 (phase II metabolizing enzyme)
2203	20	46	17	1	0	79.76	16	76 929 941	76 942 266	12 325	WWOX (cancers)
109	33	39	15	0	0	82.76	1	150 822 330	150 853 218	30 888	LCE3C, LCE3B (psoriasis)
2559	32	41	1	0	0	98.65	22	22 613 016	22 670 785	57 769	GSTT2, GSTT2B, GSTTP1 (phase II metabolizing enzyme)
88	46	1	0	0	0	100.00	1	110 025 907	110 044 476	18 569	GSTM2,GSTM1 (phase II metabolizing enzyme)

Abbreviations: CNPs, copy number polymorphisms; UCSC, University of California Santa Cruz genes.
A full colour version of this Table is available at the Journal of Human Genetics Journal online.

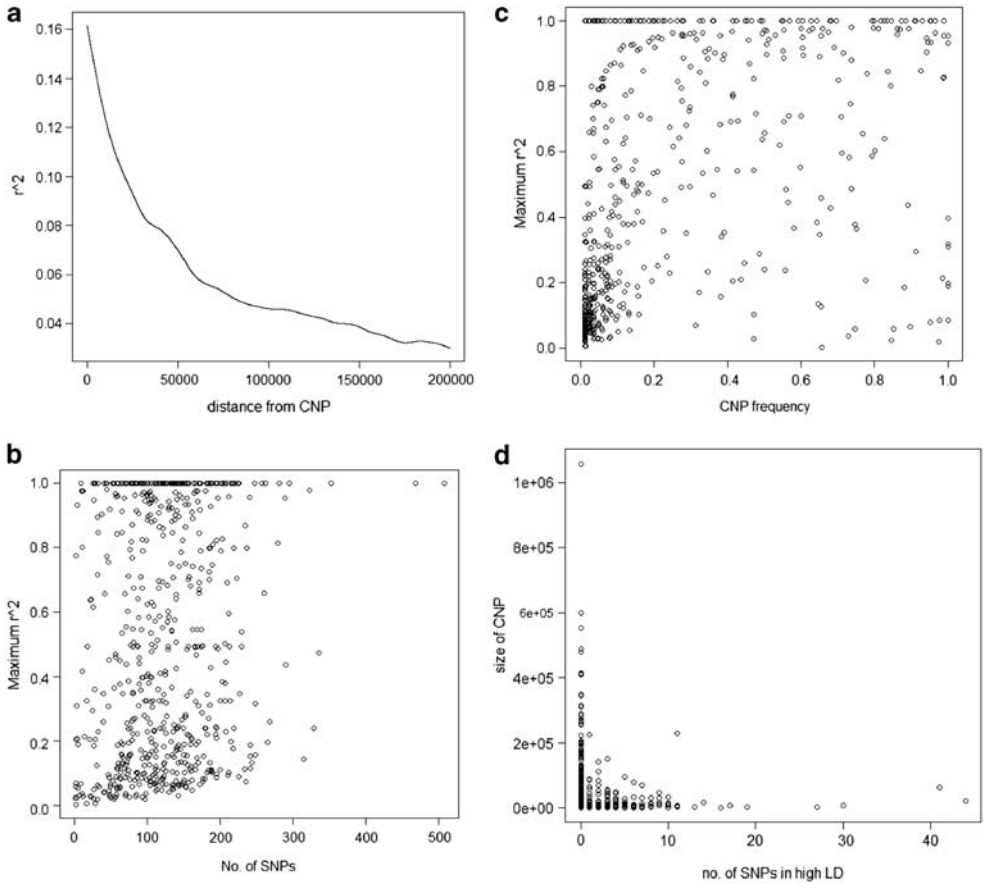


Figure 1 (**a**) The correlation between the r^2 and the distance between copy number polymorphism (CNP) and single-nucleotide polymorphism (SNP). (**b**) Maximum r^2 of CNP versus number of nearby SNPs in 200-kb windows. (**c**) Maximum r^2 of CNP versus CNP frequency. (**d**) Number of SNPs in strong correlation with the size of CNPs.

and *LCE3C* and *LCE3B* (psoriasis and rheumatoid arthritis) among others.[35,36] The high deletion frequency of loci overlapping with *LCE3C* and *LCE3B* (82.76%), *UGT2B17* (47.13%) and *WWOX* (79.76%) requires further studies to investigate their associations with complex diseases such as psoriasis, rheumatoid arthritis and graft-versus-host disease for hematopoietic stem cell transplantation patients. For example, the mismatch of the copy numbers of UGT2B17 was found to be associated with graft-versus-host disease in patients with hematopoietic stem cell transplantation.[34] Deletion of UGT2B17 was also associated with an increased risk for prostate cancer.[33]

Correlation analyses between CNPs and nearby SNPs

To study the correlation patterns with SNPs, we calculated the r^2 between the 648 polymorphic CNPs and nearby SNPs within a 200-kb window from the start and end positions of the CNP. The proportion of the CNPs with at least one SNP in strong correlation ($r^2 > 0.8$) was 31.9%, that is, 207 CNPs were found to be in strong correlation with at least one SNP. The median and maximum numbers of SNPs that were in strong correlation with the 207 CNPs were 3 and 44, respectively. This suggests that half of the 207 CNPs can be tagged by more than three SNPs and some of the CNPs were tagged by tens of SNPs. These results suggest that the majority of CNPs were not being well tagged by the nearby SNPs in the Affymetrix SNP Array 6.0. The strength of the r^2 value decreases with distance between the CNP and SNP (Figure 1a). We further investigated whether CNPs that were not well tagged tend to be located in the genomic regions where SNP markers are sparse. The correlation patterns do not appear to be affected by the number of nearby SNPs and the frequencies of CNPs (Figures 1b and c). In other words, there was no apparent difference in the number of nearby SNPs and the frequencies of CNPs between (a) the CNPs that were in strong correlation ($r^2 > 0.8$) and (b) CNPs that were not in strong correlation with SNPs (Figures 1b and c). However, smaller-sized CNPs were generally in strong correlation with more SNPs than the larger CNPs (Figure 1d).

Correlation analyses between CNPs and GWAS-SNPs

To investigate the potential role of CNPs in the etiology of complex diseases or traits, we computed the r^2 between CNPs and the SNPs on the NHGRI GWAS Catalog (http://www.genome.gov/gwastudies/). Of the >3000 GWAS-SNPs that have been found to be associated with various complex diseases and traits, only eight GWAS-SNPs were found to be in strong correlation with six CNPs (Table 3). Following the methods of Conrad *et al.*,[2] we define in our analysis a strong correlation as $r^2 > 0.5$. These eight SNPs were reported to be associated with five diseases or traits, namely body mass index, childhood acute lymphoblastic leukemia, early-onset myocardial infarction, Crohn's disease and multiple sclerosis. Several SNPs were in strong correlation with a single CNP, for example, three SNPs (rs13361189, rs1000113 and rs11747270) were found to be in strong correlation with CNP874.

The most notable SNP was rs2815752 near the *NEGR1* gene (associated with body mass index), which was in perfect correlation ($r^2 = 1$) with CNP60. This locus is a 42-kb deletion located in

Table 3 Correlation between CNPs and GWAS-SNPs at $r^2 > 0.5$

CNP ID	*Chromosome*	*Start position*	*End position*	*Length*	*GWAS-SNP*	*r^2 value*	*Gene*	*Complex disease/trait*
60	1	72 541 504	72 583 736	42 232	rs2815752	1	NEGR1	BMI
147	1	194 997 658	195 068 695	71 037	rs6428370	0.647399825	Intergenic	Acute lymphoblastic leukemia (childhood)
333	2	203 608 045	203 610 291	2246	rs6725887	0.84632626	WDR12	Myocardial infarction (early onset)
874	5	150 185 693	150 198 797	13 104	rs13361189	0.927251567	IRGM	Crohn's disease
874	5	150 185 693	150 198 797	13 104	rs1000113	0.927251567	IRGM	Crohn's disease
874	5	150 185 693	150 198 797	13 104	rs11747270	0.927251567	IRGM	Crohn's disease
877	5	155 409 350	155 415 307	5957	rs4704970	1	SGCD	Multiple sclerosis
933	6	32 539 530	326 81 749	142 219	rs3129934	0.664781909	HLA-DRB1	Multiple sclerosis

Abbreviations: BMI, body mass index; CNPs, copy number polymorphisms; GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.
A full colour version of this Table is available at the *Journal of Human Genetics* Journal online.

chromosome 1 that did not overlap with any of the UCSC genes and is located only 1.3 kb away from the SNP. The total deletion frequency in the Swedish population was high (Table 3 and Supplementary Table 4), of which 51.72% were one-copy deletions and 29.89% were two-copy deletions. CNP874 was found to be in nearly perfect correlation (r^2=0.93) with three GWAS-SNPs located near the *IRGM* gene, which is associated with Crohn's disease. However, in comparison with CNP60, the total deletion frequency for CNP874 was much lower, with only 11.90% one-copy deletions and 1.19% two-copy deletions. This locus spans 13 kb in chromosome 5 and does not overlap with any of the UCSC genes. The three GWAS-SNPs were located 4.8 kb (rs13361189), 21.4 kb (rs1000113) and 40.2 kb (rs11747270) away from the deletion. The CNP877 locus is implicated in multiple sclerosis, where it is in perfect correlation with the GWAS-SNP (rs4704970). None of the individuals were deleted in both copies, and 32.56% were one-copy deletions. The other CNPs were implicated in childhood acute lymphoblastic leukemia (CNP147) and early-onset myocardial infarction (CNP333). Interestingly, all the CNPs found to be in strong correlation with GWAS-SNPs had only deletions in the loci.

Characteristics of CNV loci identified by Birdseye (Birdsuite)

Similar to the PennCNV output analysis, we also merged overlapping CNVs to construct CNV loci for the Birdseye data and identified 641 loci, of which 451 were del-loci, 102 were dup-loci and the remaining 31 were del/dup-loci (Table 4). The proportion of del-loci (76.40%) identified by PennCNV data was higher than that for the Birdseye data (70.36%). In comparison, the Birdseye data identified a higher proportion of dup-loci (24.80%) than the PennCNV data (16.37%). However, these differences are not substantial.

The 641 loci identified by the Birdseye data cover approximately 35.23 Mb of the nucleotide sequence, and the sum of the length for del-loci (13.10 Mb) is smaller than that for dup-loci (15.06 Mb). Similar to PennCNV data, the average size of del-loci (29.04 kb) is much smaller than that of the dup-loci (94.70 kb). However, substantial differences were observed for these parameters between the PennCNV and Birdseye data (Tables 1 and 4). For example, the sum of lengths covering CNV loci detected by the PennCNV data (61.52 Mb) was approximately twice that for the Birdseye data (35.23 Mb), while they have an almost similar number of CNV loci.

More than 60% of del-loci were $<$10 kb, and in comparison, only 18.24% of dup-loci fall within this size range. The majority (52.20%) of dup-loci ranged from 10 to 100 kb. In summary, there were more del-loci, but their sizes tended to be smaller than those of the dup-loci. This is in agreement with the PennCNV data. However, the size distribution pattern of the CNV loci for the Birdseye data is skewed towards the 'smaller' end compared with the PennCNV data. This is apparent when comparing the proportions in the first two strata: (a) $<$10 kb and (b) $\geqslant$10–$<$50 kb between the two sets of data (Tables 1 and 4). The list of the 641 loci is shown in Supplementary Table 2.

Of the 641 CNV loci, 280 loci were detected in $\geqslant$2 individuals (Table 4), and the remaining loci in only one individual. The proportion of del-loci detected in $\geqslant$2 individuals (43.90%) was much higher than the proportion for dup-loci (32.08%). Among the high-frequency CNV loci (loci detected in multiple individuals), several overlapped with disease-associated or pharmacogenetics-related genes such as *WWOX* and *GSTT1*, which have also been observed in the PennCNV data (Supplementary Table 2). Furthermore, the deletion frequencies were comparable between the Birdseye and PennCNV data. For example, a deletion locus overlapped with *WWOX* was also found in the Birdseye data. It was detected in 29 of the 87 individuals (33.33%), and a deletion locus encompassing *GSTT1* was deleted at a population frequency of 11.49%. Among the 196 CNV loci (160 del-loci, 30 dup-loci and 6 del/dup-loci) that were detected by both the Birdseye and PennCNV data and consistent in their CNV status, only 21 loci differed significantly (FDR $<$0.01) in their frequencies estimated by both sets of data. In addition, the proportion of del-loci encompassing UCSC genes (24.83%) was much lower than dup-loci (45.28%); this finding is again consistent with the PennCNV data.

For the CNV loci detected with the Birdseye data, we also performed the 'validation' steps for overlap with the PennCNV data and the DGV. As mentioned earlier, we found 196 loci with $\geqslant$50% reciprocal overlap between the Birdseye and PennCNV data. For the remaining 445 CNV loci that were not confirmed by PennCNV data, we found that 322 loci have been cataloged in the DGV (please see Materials and methods). Therefore, by applying two different ways of validation, 518 (80.81%) of the 641 CNV loci detected by Birdseye were considered reliable in this study (Table 4).

Comparison with HapMap phase III populations

The principal component analysis showed distinct clusters for populations with different ancestries. The first two principal components (PC1 and PC2) separated the African (ASW, MKK and LWK) and non-African (CHD, GIH, MEX, SWED and TSI) populations (Figure 2a). This suggests that the CNP profiles of the African populations were substantially different from those of the non-African populations. From the second and fourth principal components (PC2 and PC4), three distinct clusters were observed (Figure 2b). The three African populations remained as a distinct cluster; however, CHD was separated from the European populations (MEX, SWED and TSI) and the Gujarati Indians (GIH). This indicates that the CNP profile of Gujarati Indians in Houston (Texas, USA) resembles that of the European populations. Principal component analysis was also performed by restricting only the 'European cluster' populations

Table 4 Summary statistics of CNV loci constructed from Birdseye (Birdsuite) output

Summary statistics of CNV loci (Birdseye output)	*Total*	*Del*	*Dup*
Number of CNV loci	641	451 (70.36%)[a]	159 (24.80%)[a]
Number of CNV loci detected in ⩾2 individuals	280 (43.68%)[b]	198 (43.90%)[b]	51 (32.08%)[b]
Sum of the length of loci	35.23 Mb	13.10 Mb	15.06 Mb
Average length per locus	54.96 kb	29.04 kb	94.70 kb
Average number of markers per locus	30	22	42
Size distribution			
<10 kb	303 (47.27%)	272 (60.31%)	29 (18.24%)
⩾10–<50 kb	193 (30.11%)	119 (26.39%)	63 (39.62%)
⩾50–<100 kb	52 (8.11%)	27 (5.99%)	20 (12.58%)
⩾100–<500 kb	79 (12.32%)	31 (6.87%)	40 (25.16%)
⩾500 kb	14 (2.18%)	2 (0.44%)	7 (4.40%)
Overlapping with DGV			
CNV loci that overlap	465 (72.54%)	335 (74.28%)	106 (66.67%)
CNV loci that did not overlap	176 (27.46%)	116 (25.72%)	53 (33.33%)
Overlapping with UCSC genes			
CNV loci that overlap	202 (31.51%)	112 (24.83%)	72 (45.28%)
CNV loci that did not overlap	439 (68.49%)	339 (75.17%)	87 (54.72%)
Overlapping with CNV loci constructed from Birdseye and consistent in CNV status, that is, del/dup/del+dup			
CNV loci that overlap	196 (30.58%)	160 (35.48%)	30 (18.87%)
CNV loci that did not overlap	445 (69.42%)	291 (64.52%)	129 (81.13%)

Abbreviations: CNV, copy number variant; DGV, database of genomic variants; UCSC, University of California Santa Cruz genes.
[a]The percentage was calculated by dividing 641 loci.
[b]The percentage was calculated by dividing 641, 451 and 159 loci, respectively.
Note: as there are only 31 CNV loci (4.84%) with status del+dup, the summary statistics of these loci were not shown in the table.
A full colour version of this Table is available at the Journal of Human Genetics Journal online.

(GIH, MEX, SWED and TSI) in PC2 versus PC4 (Figure 2b). More interestingly, we also found that the CNP profile of the Swedish population was substantially different from that of the other populations such as GIH and MEX, but it was also appreciably different from that of TSI (Figure 2c). These differences further justify the need to detect and characterize the CNV/CNP profile of the Swedish population.

Characteristics of ROHs

By restricting ROHs to ⩾500 kb, a total of 14 815 regions were found in the 87 Swedish individuals with an average of 170 ROHs (Supplementary Table 5). The number of ROHs ranged from 105 to 220. The majority of these ROHs were <1 Mb in length (Supplementary Figure 3). However, by restricting ROHs to ⩾1 Mb, 2814 ROHs with an average of 32 ROHs per individual were found. The median size of the ROHs was approximately 686 kb, with the largest ROH spanning a length of approximately 25 Mb in chromosome 11. This ROH contained 9034 homozygotes, 29 heterozygotes and 2 missing genotypes, and had a density of 3.6 SNPs per 10 kb. The second largest ROH was 12 Mb in length and was detected in a different individual. This ROH contained 1571 homozygotes and 19 heterozygotes and had a density of 1.3 SNPs per 10 kb. The sum of the length of ROHs in each individual (that is, the total length of all the ROHs in one individual) was then computed. It ranged from approximately 87 to 179 Mb with a median and mean of approximately 141 Mb, respectively. This finding suggests that, on average, 141 Mb or 4.92% of the human genome (2867 Mb) was homozygous in these Swedish individuals (Table 5).

The distribution pattern of these ROHs in the 22 autosomes was also studied. The larger chromosomes (chromosomes 1–8) tended to have a higher average number of ROHs per individual (Table 5). For example, these chromosomes had an average number of >9 ROHs per individual, and in contrast, an average number of <5 ROHs per individual was detected in chromosomes 16–22. As a result, chromosomes 1–8 also had a higher average sum of length of ROHs per individual (>7 Mb) than the smaller chromosomes, that is, <4 Mb for chromosomes 16–22. However, this pattern was less obvious when the parameters were adjusted for the sizes of the chromosomes. For example, the proportion of the chromosome encompassed by ROHs for the largest chromosome 1 (4.78%) was smaller than that for the other chromosomes such as chromosome 17 (5.14%). An apparent trend is not observed for the proportion of the chromosome encompassed by ROHs across the 22 autosomes. However, chromosomes 3, 4, 8 and 12 tended to have the highest proportions (5.90–6.16%), and, in contrast, chromosomes 16, 19, 21 and 22 had the lowest proportions (1.76–2.59%). These results were not due to differences in the density of SNPs across the 22 autosomes, as we found no substantial differences in the density of SNPs across the chromosomes (except for chromosome 19, which had a density of <2 SNPs per 10 kb when compared with the other chromosomes). Although chromosomes 3 and 4 had >6% of the proportion of the chromosome encompassed by ROHs, the density of SNPs of these chromosomes was similar to that of chromosome 16, where only approximately 2% of this chromosome was covered by ROHs (Table 5).

DISCUSSION

In this study, >600 CNV loci were detected in the Swedish population using two different CNV-detection algorithms, that is, PennCNV (623 loci) and Birdsuite (641 loci). From these, 196 loci were consistently identified by both algorithms, suggesting their reliability. In addition,

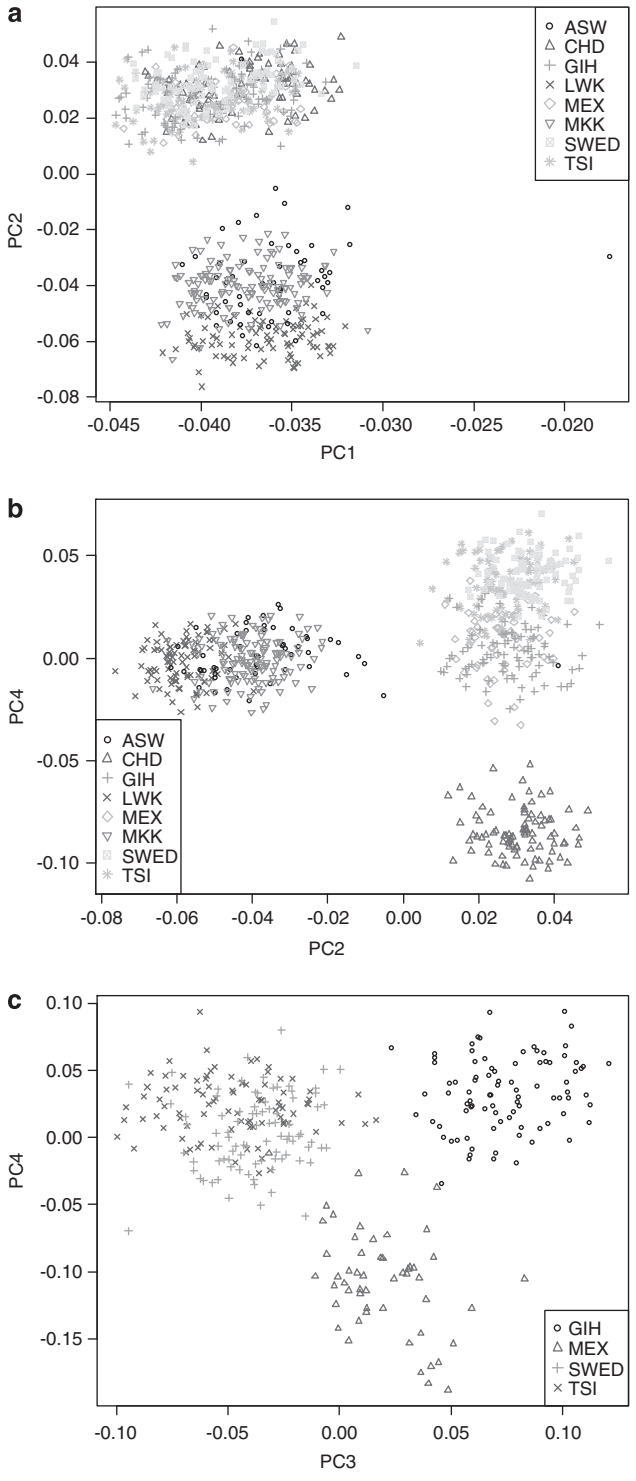


Figure 2 Principal component analysis comparing the populations. (**a**) Swedish and HapMap phase III populations—PC 1 versus PC 2. (**b**) Swedish and HapMap phase III populations—PC 2 versus PC 4. (**c**) Swedish and three HapMap III populations (GIH, MEX and TSI)—PC 3 versus PC 4.

we also identified a total of 14 815 ROHs $\geqslant$500 kb or 2814 ROHs $\geqslant$1 Mb in the Swedish individuals with an average of 170 and 32 regions detected per individual, respectively.

CNVs have been increasingly recognized as a significant source of genetic variation or diversity in human populations. Detection of CNVs using SNP genotyping arrays is more cost-effective and affordable for population-based studies as compared with sequencing-based methods, which are limited to only a few individuals.[37–39] This has enabled our study to investigate the population characteristics of CNVs. Although >600 CNV loci were identified, only 268 were detected in at least two individuals by PennCNV. Similarly, Birdseye also found 280 common CNV loci in the 87 Swedish individuals. More importantly, these common CNV loci were found to encompass several disease-related and important drug-metabolizing genes, suggesting that these loci warrant further characterization and study for their associations with the relevant diseases or traits.

We applied two different algorithms to detect CNV loci as a validation step; 196 loci were found by both the algorithms and these loci were also consistent in their CNV status ('del', 'dup' or 'del+dup'). In the majority of the 196 loci, the population frequencies were also in good agreement between PennCNV and Birdseye data, indicating that these CNV loci are highly reliable. In addition, most of the CNV loci detected by PennCNV (>70%) and Birdseye (>80%) can be 'validated' by comparing them with each other and with the DGV. The proportion of CNV loci overlapping with the DGV was approximately 62% and 72% for PennCNV and Birdsuite, respectively. These percentages could be overestimated because of the false-positive entries in the DGV. Of the 196 CNV loci that were identified by both algorithms, 53 loci had not been previously cataloged in the DGV, which represents a subset of reliable novel CNV loci identified in our study. The list of CNV loci in the DGV is not as yet complete as results from only 42 published studies were documented as of November 2010 (http://projects.tcag.ca/variation/).

On performing the correlation analysis between CNPs and GWAS-SNPs, our results also indicated that several CNPs could be potential causal variants because of their strong correlation with the GWAS-SNPs. Notably, the strong correlation between the CNPs and the GWAS-SNPs near NEGR1 and IRGM for body mass index and Crohn's disease, respectively, are consistent with previous studies.[40,41]

Our study has a higher sensitivity than the study by Díaz de Ståhl *et al.*,[10] which only detected an average of 15 CNVs per individual compared with our study, which detected an average of 36 CNVs per individual. An average of 4 clones per CNV was detected in the Díaz de Ståhl *et al.* study, whereas in our study, each CNV was detected by an average of 51 markers (Supplementary Table 1). The ability to detect smaller CNVs was also demonstrated in our study, because the average size of CNVs detected by Díaz de Ståhl *et al.* was approximately 3.5-fold (358 kb) larger than that in our study. Although Díaz de Ståhl *et al.* also clustered individual overlapping CNVs into loci, their analysis was performed using data from different ancestries (33 Europeans, 24 Africans and 14 Asians), whereas the CNV loci constructed in our study were based entirely on the data from 87 Swedish individuals. Therefore, our list of CNV loci and their frequencies was more representative of the Swedish population.

We did not compare our results with existing data from published studies because of the methodological issues in CNV and ROH detection in the different studies. As different studies have used different platforms, quality control criteria and methods to construct CNV loci and detect ROHs, comparisons with published studies would not be valid. Therefore, we would need to analyze the data from different populations with same analytical procedure. Furthermore, such a comparison is beyond the scope of the current paper and will be addressed in a future publication. However, to provide some preliminary insight into the population differences, we compared the CNP profiles of the Swedish population with the HapMap phase III populations. This comparison was appropriate as

Table 5 Distribution pattern of ROHs across the 22 autosomes

Chromosome	Total number of ROHs	Average number of ROHs per individual	Sum of length of ROHs (bp)	Average sum of length of ROHs per individual (bp)	Chromosome size (bp)[a]	Proportion (%) of chromosome encompassed by ROHs	Number of SNPs in Affymetrix 6.0	Density of SNPs per 10 kb
1	1243	14.3	1 029 256 231	11 830 531	247 249 719	4.78	73469	3.0
2	1491	17.1	1 223 537 523	14 063 650	242 951 149	5.79	75933	3.1
3	1256	14.4	1 069 972 110	12 298 530	199 501 827	6.16	62316	3.1
4	1246	14.3	1 015 875 656	11 676 732	191 273 063	6.10	57561	3.0
5	1021	11.7	859 950 902	9 884 493	180 857 866	5.47	57967	3.2
6	1008	11.6	834 180 388	9 588 280	170 899 992	5.61	57855	3.4
7	811	9.3	632 768 685	7 273 203	158 821 424	4.58	48419	3.0
8	896	10.3	762 529 281	8 764 704	146 274 826	5.99	50019	3.4
9	566	6.5	439 197 494	5 048 247	140 273 252	3.60	42710	3.0
10	722	8.3	612 229 774	7 037 124	135 374 737	5.20	49608	3.7
11	722	8.3	650 352 277	7 475 314	134 452 384	5.56	45944	3.4
12	725	8.3	679 233 723	7 807 284	132 349 534	5.90	43833	3.3
13	482	5.5	360 268 323	4 141 015	114 142 980	3.63	35158	3.1
14	571	6.6	448 210 796	5 151 848	106 368 585	4.84	28942	2.7
15	438	5.0	371 570 656	4 270 927	100 338 915	4.26	26905	2.7
16	192	2.2	159 973 057	1 838 771	88 827 254	2.07	28658	3.2
17	428	4.9	352 288 646	4 049 295	78 774 742	5.14	21347	2.7
18	330	3.8	234 464 335	2 694 992	76 117 153	3.54	27219	3.6
19	184	2.1	143 788 195	1 652 738	63 811 651	2.59	12419	1.9
20	271	3.1	220 116 198	2 530 071	62 435 964	4.05	23487	3.8
21	100	1.1	71 684 424	823 959	46 944 323	1.76	12948	2.8
22	112	1.3	100 622 242	1 156 577	49 691 432	2.33	12059	2.4

Abbreviations: ROHs, regions of homozygosity; SNPs, single-nucleotide polymorphisms; UCSC, University of California Santa Cruz genes.
[a]The size of chromosome was obtained from UCSC Genome Browser.
A full colour version of this Table is available at the Journal of Human Genetics Journal online.

we analyzed the CNP output for the HapMap III populations generated by Canary similar to the Swedish population output. As expected, the results of our analysis showed that the CNP profile of the Swedish population was substantially different from that of the African populations (ASW, MKK and LWK) and CHD. More interestingly, the CNP profile of the Swedish population was also considerably different from that of other European populations (MEX and TSI) and GIH. This further supports the importance of delineating the population characteristics of CNVs/CNPs in the Swedish population.

There are a number of limitations when using SNP genotyping arrays to detect CNVs and ROHs, and the CNV and ROH list reported in our study is not complete. Future studies will require higher sensitivity methods and larger sample sizes for a more thorough detection of CNVs and ROHs. Nevertheless, this is the first population-based study to investigate the population characteristics of CNVs and ROHs in the Swedish population. This study found many reliable CNV loci and also highlighted numerous loci that warrant further investigation for their medical or pharmacogenetic importance. The abundance of ROHs detected in the human genome also suggests the importance of studying their associations with complex phenotypes.

ACKNOWLEDGEMENTS

The Yong Loo Lin School of Medicine, the Life Science Institute and the Office of Deputy President (Research and Technology), National University of Singapore. We also acknowledge the support of the Genome Institute of Singapore, and Agency for Science, Technology and Research, Singapore.

1 McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40,** 1166–1174 (2008).
2 Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
3 Park, H., Kim, J. I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42,** 400–405 (2010).
4 Yim, S. H., Kim, T. M., Hu, H. J., Kim, J. H., Kim, B. J., Lee, J. Y. *et al.* Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum. Mol. Genet.* **19,** 1001–1008 (2010).
5 Ku, C. S., Pawitan, Y., Sim, X., Ong, R. T., Seielstad, M., Lee, E. J. *et al.* Genomic copy number variations in three Southeast Asian populations. *Hum. Mutat.* **31,** 851–857 (2010).
6 Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
7 Pinto, D., Marshall, C., Feuk, L. & Scherer, S. W. Copy-number variation in control population cohorts. *Hum. Mol. Genet.* **16,** R168–R173 (2007).
8 Zogopoulos, G., Ha, K. C., Naqib, F., Moore, S., Kim, H., Montpetit, A. *et al.* Germ-line DNA copy number variation frequencies in a large North American population. *Hum. Genet.* **122,** 345–353 (2007).
9 de Smith, A. J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N. A., Tsang, P. *et al.* Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* **16,** 2783–2794 (2007).
10 Díaz de Ståhl, T., Sandgren, J., Piotrowski, A., Nord, H., Andersson, R., Menzel, U. *et al.* Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array. *Hum. Mutat.* **29,** 398–408 (2008).
11 Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3,** 1787–1799 (2007).
12 Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15,** 789–795 (2006).
13 Li, L. H., Ho, S. F., Chen, C. H., Wei, C. Y., Wong, W. C., Li, L. Y. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum. Mutat.* **27,** 1115–1121 (2006).
14 McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Abdel-Rahman, R., Franklin, C. S., Pericic, M. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83,** 359–372 (2008).
15 Nothnagel, M., Lu, T. T., Kayser, M. & Krawczak, M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum. Mol. Genet.* **19,** 2927–2935 (2010).
16 Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl Acad. Sci. USA* **104,** 19942–19947 (2007).
17 Nalls, M. A., Guerreiro, R. J., Simon-Sanchez, J., Bras, J. T., Traynor, B. J., Gibbs, J. R. *et al.* Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10,** 183–190 (2009).
18 Yang, T. L., Guo, Y., Zhang, L. S., Tian, Q., Yan, H., Papasian, C. J. *et al.* Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *J. Clin. Endocrinol. Metab.* **95,** 3777–3782 (2010).

19 O'Dushlaine, C. T., Morris, D., Moskvina, V., Kirov, G., Consortium, I. S., Gill, M. *et al.* Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18,** 1248–1254 (2010).
20 International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
21 Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. *et al.* 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 (2007).
22 Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40,** 1253–1260 (2008).
23 Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. & Pawitan, Y. Identification of recurrent regions of copy-number variants across multiple individuals. *BMC Bioinformatics* **11,** 147 (2010).
24 Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36,** 949–951 (2004).
25 Aqeilan, R. I., Kuroki, T., Pekarsky, Y., Albagha, O., Trapasso, F., Baffa, Re *et al.* Loss of WWOX expression in gastric carcinoma. *Clin. Cancer Res.* **10,** 3053–3058 (2004).
26 Kuroki, T., Yendamuri, S., Trapasso, F., Matsuyama, A., Aqeilan, R. I., Alder, H. *et al.* The tumor suppressor gene WWOX at FRA16D is involved in pancreatic carcinogenesis. *Clin. Cancer Res.* **10,** 2459–2465 (2004).
27 Prickett, T. D., Agrawal, N. S., Wei, X., Yates, K. E., Lin, J. C., Wunderlich, J. R. *et al.* Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4. *Nat. Genet.* **41,** 1127–1132 (2009).
28 Ferreira, M. A., O'Donovan, M. C., Meng, Y. A., Jones, I. R., Ruderfer, D. M., Jones, L. *et al*. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat. Genet.* **40,** 1056–1058 (2008).
29 Ouahchi, K., Lindeman, N. & Lee, C. Copy number variants and pharmacogenomics. *Pharmacogenomics* **7,** 25–29 (2006).
30 Fanciulli, M., Norsworthy, P. J., Petretto, E., Dong, R., Harper, L., Kamesh, L. *et al.* FCGR3B copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39,** 721–723 (2007).
31 Miki, D., Kubo, M., Takahashi, A., Yoon, K. A., Kim, J., Lee, G. K. *et al.* Variation in TP63 is associated with lung adenocarcinoma susceptibility in Japanese and Korean populations. *Nat. Genet.* **42,** 893–896 (2010).
32 Spencer, K. L., Hauser, M. A., Olson, L. M., Schmidt, S., Scott, W. K., Gallins, P. *et al.* Deletion of CFHR3 and CFHR1 genes in age-related macular degeneration. *Hum. Mol. Genet.* **17,** 971–977 (2008).
33 Karypidis, A. H., Olsson, M., Andersson, S. O., Rane, A. & Ekström, L. Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. *Pharmacogenomics J.* **8,** 147–151 (2008).
34 McCarroll, S. A., Bradner, J. E., Turpeinen, H., Volin, L., Martin, P. J., Chilewski, S. D. *et al.* Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nat. Genet.* **41,** 1341–1344 (2009).
35 Docampo, E., Rabionet, R., Riveira-Muñoz, E., Escaramís, G., Julià, A., Marsal, S. *et al.* Deletion of the late cornified envelope genes, LCE3C and LCE3B, is associated with rheumatoid arthritis. *Arthritis Rheum.* **62,** 1246–1251 (2010).
36 de Cid, R., Riveira-Munoz, E., Zeeuwen, P. L., Robarge, J., Liao, W., Dannhauser, E. N. *et al.* Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41,** 211–215 (2009).
37 Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
38 Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
39 Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
40 Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41,** 25–34 (2009).
41 McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40,** 1107–1112 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Journal of Human Genetics (2011) 56, 552–560

www.nature.com/jhg

ORIGINAL ARTICLE

Copy number polymorphisms in new HapMap III and Singapore populations

Chee-Seng Ku[1,2,8], Shu-Mei Teo[1,2,3,8], Nasheen Naidoo[1,2], Xueling Sim[1,2], Yik-Ying Teo[1,2,4,5], Yudi Pawitan[6], Mark Seielstad[7], Kee-Seng Chia[1,2,6] and Agus Salim[1,2,8]

Copy number variations can be identified using newer genotyping arrays with higher single nucleotide polymorphisms (SNPs) density and copy number probes accompanied by newer algorithms. McCarroll *et al*. (2008) applied these to the HapMap II samples and identified 1316 copy number polymorphisms (CNPs). In our study, we applied the same approach to 859 samples from three Singapore populations and seven HapMap III populations. Approximately 50% of the 1291 autosomal CNPs were found to be polymorphic only in populations of non-African ancestry. Pairwise comparisons among the 10 populations showed substantial differences in the CNPs frequencies. Additionally, 698 CNPs showed significant differences with false discovery rate (FDR) < 0.01 among the 10 populations and these loci overlap with known disease-associated or pharmacogenetic-related genes such as *CFHR3* and *CFHR1* (age related macular degeneration), *GSTTI* (metabolism of various carcinogenic compounds and cancers) and *UGT2B17* (prostate cancer and graft-versus-host disease). The correlations between CNPs and genome-wide association studies–SNPs were investigated and several loci, which were previously unreported, that may potentially be implicated in complex diseases and traits were found; for example, childhood acute lymphoblastic leukaemia, age-related macular degeneration, breast cancer, response to antipsychotic treatment, rheumatoid arthritis and type-1 diabetes. Additionally, we also found 5014 novel copy number loci that have not been reported previously by McCarroll *et al*. (2008) in the 10 populations.
Journal of Human Genetics (2011) **56,** 552–560; doi:10.1038/jhg.2011.54; published online 16 June 2011

Keywords: Affymetrix SNP Array 6.0; Birdsuite software; copy number polymorphisms; International HapMap III populations; Southeast Asian populations

INTRODUCTION

The term copy number variation (CNV) was first introduced in 2006 and it is generally defined as additions or deletions in the number of copies of a particular segment of DNA (larger than 1 kb in length) when compared with a reference genome sequence.[1] The ubiquitous nature of CNVs in the human genome was underappreciated until 2004,[2,3] when these reports stimulated a series of efforts to detect and characterise CNVs in different populations.[4–8] This development has also resulted in several new terminologies such as copy number polymorphisms (CNPs), which have been defined as common CNVs with a population frequency of at least 1%.[4]

CNVs can be detected using microarray-based methods, but these have relatively poor resolution when compared with sequencing-based approaches.[9,10] The low resolution of microarray-based methods also led to imprecise mapping of the breakpoints. This is important when constructing copy number loci to estimate population frequencies. Newer genotyping arrays, such as the Illumina Human 1M Beadchip (Illumina, San Diego, CA, USA) and the Affymetrix SNP Arrays 6.0 (Affymetrix, Santa Clara, CA, USA), have higher single nucleotide polymorphisms (SNPs) density and copy number probes, resulting in improved performance of microarray-based methods to detect CNVs. However, even with higher resolution arrays, the challenge of identifying common breakpoints still remains. This is largely due to the early CNV-calling algorithms that identified breakpoints sample-by-sample, resulting in significant variation of breakpoints. The Canary algorithm in the Birdsuite software overcomes this problem by calling CNPs simultaneously across multiple individuals at pre-defined genomic locations.[11] McCarroll *et al*.[4] used the Canary algorithm to identify 1316 CNPs in the HapMap Phase II populations. These CNPs were well validated and their sizes were in agreement with the results from the fosmid paired-end sequencing experiment.[9]

[1]Centre for Molecular Epidemiology, National University of Singapore, Singapore, Singapore; [2]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore; [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore, Singapore; [4]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore; [5]Department of Statistics & Applied Probability, National University of Singapore, Singapore, Singapore; [6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden and [7]Laboratory Medicine, Institute of Human Genetics, University of California, San Francisco, CA, USA
Correspondence: C-S Ku or Assistant Professor A Salim, Centre for Molecular Epidemiology, Department of Epidemiology and Public Health (MD3), Yong Loo Lin School of Medicine, National University of Singapore, 16 Medical Drive, Singapore 117597, Singapore.
E-mail: g0700040@nus.edu.sg or ephaguss@nus.edu.sg
[8]These authors contributed equally to this work.
Received 26 November 2010; revised 3 May 2011; accepted 6 May 2011; published online 16 June 2011

To provide a more global map of CNPs, our study aims to determine integer copy numbers of the 1316 CNPs set of three Southeast Asian populations in Singapore, namely Chinese (Sing–Chinese), Malay (Sing–Malay) and Asian Indian (Sing–Indian), and the seven populations from the HapMap Phase III.[12] The HapMap III populations studied are people of African ancestry in the southwestern USA (ASW), the Chinese community in Metropolitan Denver, Colorado, USA (CHD), Gujarati Indians in Houston, Texas, USA (GIH), the Luhya in Webuye, Kenya (LWK), people of Mexican ancestry in Los Angeles, California, USA (MEX), the Maasai in Kinyawa, Kenya (MKK) and the Tuscans in Italy (TSI). The characteristics of CNPs in the 10 populations will be described and compared. In addition, the correlation between CNPs and SNPs in the 10 populations will also be characterised and compared. A special emphasis will be given to studying the correlation between SNPs in the genome-wide association studies (GWAS) catalog (GWAS–SNPs) and CNPs in the 10 populations. Additionally, novel copy number loci that have not been reported previously by McCarroll *et al*.[4] will also be reported on from the 10 populations.

MATERIALS AND METHODS

DNA samples and genotyping

The detailed information on the sources of DNA samples, demographic data of the samples, sample selection and the origin and migration history of the three Singapore populations (Chinese, Malay and Asian Indian) have been described in our previous publication.[8,13] This study was approved by the National University of Singapore Institutional Review Board (Reference Code: 07-199E). In total, 292 DNA samples (99 Chinese, 98 Malay and 95 Indian) were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0. Of the 292 samples, 27 were excluded from subsequent analysis. The final set of 265 samples (93 Chinese, 88 Malays and 84 Indians) was available for analysis using Birdsuite. There were 135 females and 130 males in the final dataset. The detailed information on the quality control and sample filtering have also been described in our previous papers.[8,13]

HapMap III samples

The CEL-files of the Affymetrix SNP Array 6.0 for the seven populations in HapMap III were downloaded from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/hapmap/raw_data/hapmap3_affy6.0/). All the samples were analysed by Birdsuite, with only unrelated samples included in our study; that is, family-related samples were removed using the 'relationships' file provided by the International HapMap Project. After the sample exclusion step, a total of 594 unrelated samples from the seven HapMap III populations were analysed: ASW (n=52), CHD (n=89), GIH (n=89), LWK (n=90), MEX (n=53), MKK (n=132) and TSI (n=89).

CNP calling using Canary

The Birdsuite software was used to analyse the Affymetrix SNP Array 6.0 dataset, which consisted of two components for detecting copy number changes. The first component, Canary, was used to determine the integer copy number at each of the predefined 1316 CNPs identified by McCarroll *et al*.[4] in the HapMap II samples. These CNPs were found in more than one HapMap II individual and the sizes of these CNPs were also determined. The 1316 CNPs were distributed in all the autosomes and sex chromosomes. However, 25 CNPs located in the sex chromosomes were removed, as CNP calling in sex chromosomes is more problematic and less accurate. Therefore, the results reported in this study comprised of only 1291 CNPs in the 22 autosomes. Confidence statistics was used to identify poor quality calls and only integer copy numbers detected with high confidence (confidence score <0.1) were reported and used for subsequent analyses. We performed the Hardy–Weinberg equilibrium analysis as a quality control measure for biallelic CNPs in all 10 populations. It is recommended that the samples should be analysed on the basis of the genotyping batches using Birdsuite; therefore, the samples for Singapore and HapMap III populations were analysed by batch without separating the samples into each specific population.

FDR correction for population comparisons of the integer copy numbers of the CNPs

Population differences in the integer copy numbers were examined using the Fisher's exact test as implemented by the 'fisher test' command in R. The false discovery rate (FDR) was used in place of the P-value to account for the multiple-testing problem. We calculated the FDR using the Benjamini and Hochberg method. We performed two different test procedures: (1) comparing the integer copy numbers among the 10 populations simultaneously and (2) pairwise comparisons of the integer copy numbers among the 10 populations. For each procedure, FDR was computed once to control for all the tests (that is, in the second procedure, we calculated the FDR once by combining the P-values from 45×1291 tests).

Correlation analysis

All the correlation analyses of CNPs and nearby SNPs were done separately for each of the 10 populations. For each autosomal CNP (restricted to biallelic CNPs with MAF$\geqslant$5%), SNPs in close proximity with the CNP; that is, within a 200-kb window from the start- and end-position of the CNP were considered. The square of the Pearson correlation coefficient (r^2) for each of the SNPs (excluding the SNPs used for CNP-calling) found within the 200-kb windows of the respective CNP was then calculated.

The r^2 is the square of the Pearson correlation coefficient between the copy number genotypes and the SNP genotypes. The copy number genotypes were obtained using Canary in the Birdsuite algorithm. The SNP genotypes were obtained using Larry Bird in the Birdsuite algorithms. Larry Bird outputs the number of allele A (0, 1, 2) and number of allele B (0, 1, 2) for each SNP. We used the number of allele A for the calculation. Larry Bird generates the number of allele A and number of allele B for each SNP. As each SNP has two alleles in total, knowing the number of allele A will inform the number of allele B; for example, if the number of allele A is 2, then number of allele B should be 0.

The same r^2 calculations used for the autosomal CNPs and the SNPs identified by GWAS were used to explore the potential associations of CNPs with human diseases and traits. The list of GWAS–SNPs was downloaded from the National Human Genome Research Institute's website (http://www.genome.gov/gwastudies/) on 24 May 2010.

Copy number loci calling using Birdseye and validation

The Birdseye component in Birdsuite was used to detect additional copy number loci located outside the 1316 CNPs in the 10 populations. Similarly, only the copy number loci in autosomal chromosomes were detected because of the inaccuracy of Birdseye in detecting copy number loci in the sex chromosomes. Copy number calls with low confidence (confidence score <5) were removed. On the basis of the copy number calls generated by Birdseye, we constructed novel copy number loci using the methods that we developed previously.[14] All the downstream analyses after Canary and Birdseye were performed using the software package R (http://www.r-project.org/). The novel copy number loci identified by Birdseye were compared with data from the Database of Genomic Variants (http://projects.tcag.ca/variation/) as a validation step. We defined a copy number locus overlapped with the Database of Genomic Variants, if the locus overlapped by $>50\%$ of its length with one or more entries in the Database of Genomic Variants.

RESULTS

Characteristics of CNPs in the 10 populations

In each of the 10 populations, among the polymorphic CNPs (Table 1), most were biallelic, where the integer copy numbers were either exclusively deletions (copy number=0, 1) or exclusively duplications (copy number=3, 4). Among the biallelic CNPs, the majority did not show significant deviation from Hardy–Weinberg equilibrium with less than 2% failing a Hardy–Weinberg equilibrium test at P-value <0.01 in all except three populations—Sing–Chinese (2.2%), ASW (4.2%) and LWK (2.8%).

Table 1 The number of loci (and the percentage) with varying population frequencies for the 1291 autosomal CNPs

CNP	*Sing–Chinese*	*Sing–Malay*	*Sing–Indian*	*ASW*	*CHD*	*GIH*	*LWK*	*MEX*	*MKK*	*TSI*
Not polymorphic (0%)	675[a] (52.29)[b]	663 (51.36)	670 (51.90)	341 (26.41)	688 (53.33)	677 (52.44)	487 (37.72)	681 (52.75)	460 (35.63)	650 (50.35)
Population frequencies $\leqslant$10%	335 (25.95)	342 (26.49)	324 (25.10)	592 (45.86)	330 (25.58)	318 (24.63)	458 (35.48)	336 (26.03)	507 (39.27)	355 (27.50)
Population frequencies $>$10–50%	155 (12.01)	158 (12.24)	170 (13.17)	242 (18.75)	141 (10.93)	174 (13.48)	229 (17.74)	152 (11.77)	208 (16.11)	157 (12.16)
Population frequencies $>$50%, $<$100%	109 (8.44)	113 (8.75)	109 (8.44)	103 (7.98)	109 (8.45)	106 (8.21)	101 (7.82)	105 (8.13)	99 (7.67)	113 (8.75)
Completely polymorphic (100%)	17 (1.32)	15 (1.16)	18 (1.39)	13 (1.01)	22 (1.71)	16 (1.24)	16 (1.24)	17 (1.32)	17 (1.32)	16 (1.24)

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.
[a]Number of loci.
[b]Percentage (number of loci/1291 autosomal CNPs).

In terms of the proportion of non-polymorphic loci and loci with varying population frequencies, the Singapore populations were similar to the HapMap III populations of non-African descent (CHD, GIH, MEX and TSI) (Table 1 and Supplementary Figure 1). More than half of the CNPs were non-polymorphic in the Singapore and HapMap III populations of non-African descent. This was in contrast to the populations of African descent (ASW, LWK and MKK), where only 26.41–37.72% of the CNPs were not polymorphic. They also had higher proportions of CNPs with frequencies ranging from 1 to 10%, ASW (45.86%), LWK (35.48%) and MKK (39.27%), compared with the other populations (ranging from 24.63 to 27.50%). In addition, among all the populations, there were no substantial differences in the proportion of CNPs with a population frequency $>$10%. The discrepancy between populations of African descent and others is largely due to these populations having a larger number of rarer CNPs with a population frequency $<$10%. Hence, the differences between populations of African descent and the others were primarily in the proportion of non-polymorphic loci and those with population frequencies $<$10%. It is also worth noting that the Sing–Indian and Sing–Chinese populations have almost similar distributions of polymorphic loci, when compared with the HapMap III populations with whom they share a similar ancestry (that is, GIH and CHD, respectively) (Table 1 and Supplementary Figure 1).

The proportion of common (MAF$\geqslant$0.05) biallelic CNPs that were highly correlated with at least one SNP ($r^2>0.8$) was approximately 50% for non-African populations, but a lower proportion for African populations; that is, ASW (35.34%), LWK (34.84%) and MKK (37.39%). The majority of the common biallelic CNPs were 'deletions'. There was a substantial difference in the proportion that was highly correlated with at least one SNP for CNPs categorised as 'deletions' and 'duplications'. However, this substantial difference could be biased because of the small number of 'duplications' (Table 2). The strength of correlation or the r^2 value decreased with distance between the CNP and SNP (Supplementary Figure 2).

We further investigated whether CNPs that were not well tagged were located in the genomic regions where SNP markers are sparse. The correlation patterns did not seem to be affected by the number of nearby SNPs and the MAF of CNPs. There was no apparent difference in the number of nearby SNPs and the MAF of CNPs between (a) the CNPs that were in strong correlation ($r^2>0.8$) and (b) CNPs that were not in strong correlation with SNPs (Supplementary Figures 3a and b). However, smaller sizes of CNPs were generally in strong correlation with more SNPs than the larger CNPs (Supplementary Figure 3c). These results were consistent across the 10 populations.

Population differences in the integer copy numbers of the CNPs

Out of the 698 CNPs (FDR$<$0.01) that differed between the 10 populations, several loci encompassed known disease- or traits-associated or pharmacogenetic-related genes (Supplementary Table 1). These included *WWOX*, *ERBB4* and *TP63* (cancers), *ADAMTSL3* (height), *CFHR3* and *CFHR1* (age-related macular degeneration), *GSTT1* (metabolism of various carcinogenic compounds and cancers), *UGT2B17* (prostate cancer and graft-versus-host disease) and *CYP2A6* (metabolism of various drugs). There was a large interpopulation difference in the frequencies of some of the CNPs overlapping these genes. For example, CNP2203, which overlaps with the tumour suppressor gene *WWOX*, was not polymorphic in CHD, whereas it had a deletion frequency of 2.38% in Sing–Chinese and 7.32% in Sing–Malay (Table 3 and Supplementary Table 2). In contrast, the deletion frequency was 51.81% in Sing–Indian and 48.86% in GIH. Similarly, CNP147, which overlaps with the *CFHR3* and *CFHR1* genes, had

Table 2 The number and proportion (%) of common (MAF ⩾ 0.05) biallelic (a) CNPs, (b) deletions, (c) duplications that were highly correlated with at least one SNPs ($r^2 > 0.8$)

Population	No. of CNPs (MAF⩾5%)	No. of CNPs correlated ($r^2 > 0.8$)	Proportion (%)	No. of deletions (MAF⩾5%)	No. of deletions correlated ($r^2 > 0.8$)	Proportion (%)	No. of duplications (MAF⩾5%)	No. of duplications correlated ($r^2 > 0.8$)	Proportion (%)
Sing–Chinese	194	104	53.61	174	103	59.20	20	1	5.00
Sing–Malay	190	106	55.79	170	105	61.76	20	1	5.00
Sing–Indian	210	115	54.76	190	112	58.95	20	3	15.00
ASW	266	94	35.34	241	94	39.00	25	0	0.00
CHD	201	112	55.72	181	110	60.77	20	2	10.00
GIH	216	117	54.17	197	117	59.39	19	0	0.00
LWK	263	89	33.84	242	87	35.95	21	2	9.52
MEX	229	105	45.85	204	104	50.98	24	1	4.17
MKK	230	86	37.39	210	86	40.95	20	0	0.00
TSI	205	105	51.22	183	103	56.28	22	2	9.09

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MAF, minor allele frequency; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; SNPs, single-nucleotide polymorphisms; TSI, Tuscans in Italy.
r^2, Square of the Pearson correlation coefficient.

Table 3 CNPs (FDR < 0.01) that overlap with known disease-associated or pharmacogenetic-related genes

CNP	Gene	Sing–Chinese	Sing–Malay	Sing–Indian	ASW	CHD	GIH	LWK	MEX	MKK	TSI
CNP2203	*WWOX*	2.38[a]	7.32	51.81	66.67	0.00	48.86	40.00	67.31	28.35	68.18
CNP340	*ERBB4*	0.00	2.33	12.05	7.69	0.00	17.24	0.00	0.00	0.00	4.49
CNP530	*TP63*	64.84	48.24	27.38	30.77	68.54	31.82	31.82	9.62	32.06	6.90
CNP2118	*ADAMTSL3*	67.05	46.84	11.54	38.46	51.19	4.49	49.40	24.32	48.80	19.51
CNP147	*CFHR3, CFHR1*	11.83	12.64	53.57	59.62	15.73	58.43	59.09	18.87	42.42	43.82
CNP2560	*GSTT1*	96.77	85.06	56.63	72.00	92.13	70.79	75.56	71.70	80.15	67.06
CNP603	*UGT2B17*	100.00	95.40	82.14	48.08	98.88	86.42	63.33	58.49	67.18	58.43
CNP2415	*CYP2A6*	18.89	36.25	5.13	6.00	23.86	11.49	8.05	2.04	8.80	4.60

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; FDR, false discovery rate; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.
[a]Population frequency (%)=deletion frequency+duplication frequency.

deletion frequencies in Sing–Chinese (10.75%), Sing–Malay (12.64%) and CHD (15.73%) that was substantially lower than the other populations.

Another CNP of interest was CNP2560, a 46-kb deletion that overlaps with *GSTT1*. *GSTT1* is an important detoxification enzyme and has a key role in metabolism of carcinogenic compounds. The total deletion frequency of this CNP was high in all the 10 populations ranging from 56.63 to 96.77% (Table 3 and Supplementary Table 2). Sing–Indians had a considerably lower total deletion frequency (56.63%) than Sing–Malays (85.06%) and Sing–Chinese (96.77%). This difference is attributable to two-copy deletion, as the difference in two-copy deletion frequency ranged from 15.66% in Sing–Indian, 32.18% in Sing–Malay and 46.24% in Sing–Chinese. The two Chinese populations had the highest two-copy deletion frequency (CHD, 41.57%). Conversely, both the Indian populations had the lowest two-copy deletion frequency (GIH, 17.98%).

CNP603 is a 125-kb deletion that overlaps with *TMPRSS11E* and *UGT2B17*. The entire *UGT2B17* gene is within the deletion locus, but only one exon from *TMPRSS11E* was deleted. The deletion frequency of CNP603 was very different in Asian and non-Asian populations (Table 3 and Supplementary Table 2). Asian populations (Sing–Chinese, Sing–Malay, Sing–Indian, CHD and GIH) had higher frequencies, which ranged from 82.14 to 100%, when compared with populations of European and African ancestry (48.08–67.18%). The differences were even more apparent for two-copy deletions with the highest frequencies in CHD (70.79%), Sing–Chinese (65.59%) and Sing–Malay (52.87%), followed by the two Indian populations, GIH (37.04%) and Sing–Indian (30.95%), whereas the European and African populations were in the lower end of the spectrum with frequencies <20%. Generally, this trend was reversed for the frequency of one-copy deletions especially in the Singapore populations (Sing–Chinese 33.33%, Sing–Malay 42.53% and Sing–Indian 51.19%).

The number of CNPs that showed significant differences (FDR < 0.01) in pairwise comparisons of the 10 populations are shown in Table 4. Only 19 CNPs showed significant differences between Sing–Chinese and CHD, and 12 CNPs between Sing–Indian and GIH, suggesting that the CNPs profile in the two Chinese and two Indian populations were very similar (Supplementary Figure 4). Through these pairwise comparisons (Table 4 and Supplementary Figure 4), the 10 populations can be divided into three groups representing Asian, European and African ancestry: (a) Sing–Chinese, Sing–Malay and CHD, (b) Sing–Indian, GIH, MEX and TSI, (c) ASW, LWK and MKK. The CNPs profiles of Sing–Indian and GIH were closer to European populations (MEX and TSI).

Correlation analysis between CNPs and GWAS-SNPs

To investigate the potential role of CNPs in the aetiology of complex diseases or traits, we computed the r^2 between CNPs and the SNPs in

Table 4 The number of CNPs that showed significant differences (FDR < 0.01) in the pairwise comparisons among the 10 populations

Population	*Sing–Chinese*	*Sing–Malay*	*Sing–Indian*	*ASW*	*CHD*	*GIH*	*LWK*	*MEX*	*MKK*	*TSI*
Sing–Chinese	—	6	84	137	19	106	209	81	199	141
Sing–Malay	—	—	46	125	26	72	197	59	180	126
Sing–Indian	—	—	—	93	88	12	186	32	147	54
ASW	—	—	—	—	132	95	13	69	18	90
CHD	—	—	—	—	—	113	196	77	192	130
GIH	—	—	—	—	—	—	170	35	155	52
LWK	—	—	—	—	—	—	—	123	33	176
MEX	—	—	—	—	—	—	—	—	97	27
MKK	—	—	—	—	—	—	—	—	—	146
TSI	—	—	—	—	—	—	—	—	—	—

Abbreviations: ASW, African ancestry in the southwestern USA; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; FDR, false discovery rate; GIH, Gujarati Indians in Houston, Texas, USA; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.

the National Human Genome Research Institute GWAS catalog. Out of the >2500 GWAS–SNPs that have been found to be associated with various complex diseases and traits, only 17 GWAS–SNPs were found to be in strong correlation with 12 CNPs (Table 5 and Supplementary Tables 3 and 4). In this analysis, we defined a strong correlation as $r^2 > 0.5$, following Conrad *et al*.[5] These 17 SNPs were reported to be associated with 14 diseases or traits and the notable phenotypes that were observed consistently across the populations were body mass index, Crohn's disease, multiple sclerosis, myocardial infarction and prostate cancer. Several SNPs were in strong correlation with a single CNP; for example, three SNPs (rs13361189, rs1000113, rs11747270) were found to be in strong correlation with CNP874. Of the 33 copy number loci identified by Conrad *et al*.,[5] which were in strong correlation with GWAS–SNPs, seven were also identified in our study which had >50% overlap in length. The remaining five CNPs in our study were associated with childhood acute lymphoblastic leukaemia, age-related macular degeneration, breast cancer, response to antipsychotic treatment, rheumatoid arthritis and type-1 diabetes (Table 5 and Supplementary Tables 3 and 4).

Several SNPs were consistently found to be in strong correlation with four CNPs (CNP60, CNP874, CNP877 and CNP333) in all populations. The most notable was rs2815752 near the *NEGR1* gene (associated with body mass index), which is in perfect correlation ($r^2=1$) with CNP60 in all the 10 populations (Table 5 and Supplementary Table 3). This locus is a 42-kb deletion located in chromosome 1, which did not overlap with any of the UCSC (University of California, Santa Cruz) genes and it is located only 1.3 kb away from the SNP. The total deletion frequency in the three Singapore populations was high (Figure 1a and Supplementary Table 5). There were, however, differences in the frequency of two-copy deletion. More than 80% of the Sing–Chinese and Sing–Malay samples were deleted in both copies, but only about 41% for the Sing–Indian samples. The pattern is similar between Sing–Chinese and CHD, as well as Sing–Indian and GIH. The frequency of two-copy deletion frequency varied substantially across the 10 populations, from the lowest in the LWK population (26.97%) to the highest in Sing–Chinese (87.10%). A significant difference in the two-copy deletion frequency of CNP60 was seen between Asian populations (>80% for Sing–Chinese, Sing–Malay and CHD) compared with African populations (<35% for ASW, LWK and MKK), whereas the frequency of the Sing–Indian and GIH resembles European populations (MEX and TSI) (Supplementary Table 5).

CNP874 was found to be in strong correlation with three GWAS–SNPs located near the *IRGM* gene, which is associated with Crohn's disease. This strong correlation pattern was consistent across the 10 populations (Table 5). Most of the individuals carried either deletions or had a diploid copy. This locus spans 13 kb in chromosome 5 and did not overlap with any of the UCSC genes. The SNPs were located 4.8 kb (rs13361189), 21.4 kb (rs1000113) and 40.2 kb (rs11747270) away from the deletion. The differences in the frequency of two-copy deletion of CNP874 appeared to divide the 10 populations into two clusters. The populations of European ancestry (MEX and TSI) and Indian populations (Sing–Indian and GIH) had a frequency ⩽6.41%, but the other populations had higher frequencies, which ranged from 10% to 20.69% (Figure 1b and Supplementary Table 5). We also found a substantially lower frequency of two-copy deletion in the Sing–Indian (6.41%) compared with the Sing–Chinese (15.22%) and the Sing–Malay (11.49%) populations.

The CNP877 locus has been implicated in multiple sclerosis. It was however not polymorphic in the Sing–Chinese (Figure 1c and Supplementary Table 5). The total deletion frequencies for Sing–Malay and CHD were 2.30 and 1.14%, respectively. However, we found a much higher total deletion frequency for the other seven populations, which ranged from 17.05 to 42.53%.

Novel copy number loci in the 10 populations

The second component of the Birdsuite software, Birdseye, was used to identify novel copy number loci in the 10 populations. We subsequently found 5947 copy number loci, of which 933 loci were excluded because of overlap with the 1291 autosomal CNPs identified by McCarroll *et al*.[4] As a result, only 5014 were novel copy number loci; that is, had not been previously found by McCarroll *et al*.[4] Of these, 1448 loci were detected in two or more individuals in the 10 populations (Table 6). The list of these loci is available in Supplementary Table 6. Using a more stringent definition of 'common' novel copy number loci (population frequency ⩾1%), there were only 170 loci and of these, 42 loci had a population frequency ⩾5%.

Of the 1448 novel copy number loci, 763 (52.69%) were found to overlap with the data from the Database of Genomic Variants. Although for the 170 loci, the overlap was 78.82% (Table 6). Additionally, we also found that 86.54% of the 1448 loci were biallelic; that is, these loci contained either deletions (48.76%) or duplications (37.78%). The remaining loci were found to have both deletions and duplications. The majority of these loci did not overlap with the UCSC genes (62.43%). Of the 170 loci, 37.06% contained both deletions and duplications and the majority of these loci also did not overlap with the UCSC genes (52.35%).

Table 5 Correlation between CNPs and GWAS–SNPs at $r^2 > 0.5$ in 10 populations

CNP	*Chr.*	*Start/end position*	*GWAS–SNPs*	*GWAS–SNPs position*	*Population*	*Gene*	*Disease/trait*
60	1	72 541 504 72 583 736	rs2815752	72 585 028	Sing–Chinese, Sing–Malay, Sing–Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	*NEGR1*	BMI
874	5	150 185 693 150 198 797	rs13361189	150 203 580	Sing–Chinese, Sing–Malay, Sing–Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	*IRGM*	Crohn's disease
874	5	150 185 693 150 198 797	rs1000113	150 220 269	Sing–Chinese, Sing–Malay, Sing–Indian, CHD, MEX, MKK, TSI	*IRGM*	Crohn's disease
874	5	150 185 693 150 198 797	rs11747270	150 239 060	Sing–Chinese, Sing–Malay, Sing–Indian, ASW, CHD, GIH, MEX, MKK, TSI	*IRGM*	Crohn's disease
877	5	155 409 350 155 415 307	rs4704970	155 433 570	Sing–Malay, Sing–Indian, ASW, CHD, GIH, LWK, MEX, MKK, TSI	*SGCD*	Multiple sclerosis
333	2	203 608 045 203 610 291	rs6725887	203 454 130	Sing–Chinese, CHD, LWK, MEX, MKK, TSI	*WDR12*	Myocardial infarction (early onset)
399	3	37 957 108 37 961 932	rs9311171	37 971 481	Sing–Chinese, Sing–Malay, CHD, MEX, TSI	*CTDSPL*	Prostate cancer
28	1	25 465 715 25 534 592	rs10903129	25 641 524	Sing–Indian, GIH	*TMEM57*	Total cholesterol
147	1	194 997 658 195 068 695	rs6428370	195 111 216	Sing–Indian, ASW, GIH, MEX, TSI	*Intergenic*	Acute lymphoblastic leukaemia (childhood)
147	1	194 997 658 195 068 695	rs10737680	194 946 078	GIH	*CFH*	Age-related macular degeneration
1491	9	98 700 200 98 729 161	rs10816533	98 578 959	CHD	*ZNP510*	Height
109	1	150 822 330 150 853 218	rs10888501	150 804 578	Sing–Malay, Sing–Indian	*Intergenic*	Response to antipsychotic treatment
12035	12	118 473 270 118 475 144	rs11064768	118 302 892	Sing–Chinese	*CCDC60*	Schizophrenia
2197	16	72 953 795 73 009 537	rs10871290	73 030 197	Sing–Indian	*GLG1*	Breast cancer
933	6	32 539 530 32 681 749	rs3135338	32 509 195	Sing–Malay, Sing–Indian	*HLA*	Multiple sclerosis
933	6	32 539 530 32 681 749	rs615672	32 682 149	Sing–Malay	*HLA-DRB1*	Rheumatoid arthritis
933	6	32 539 530 32 681 749	rs9272346	32 712 350	Sing–Malay	*MHC*	Type 1 diabetes

Abbreviations: ASW, African ancestry in the southwestern USA; BMI, body mass index; CHD, Chinese community in Metropolitan Denver, Colorado, USA; CNPs, copy number polymorphisms; GIH, Gujarati Indians in Houston, Texas, USA; GWAS–SNP, genome-wide association studies–single nucleotide polymorphisms; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California, USA; MKK, Maasai in Kinyawa, Kenya; Sing, Singapore; TSI, Tuscans in Italy.
r^2, Square of the Pearson correlation coefficient.

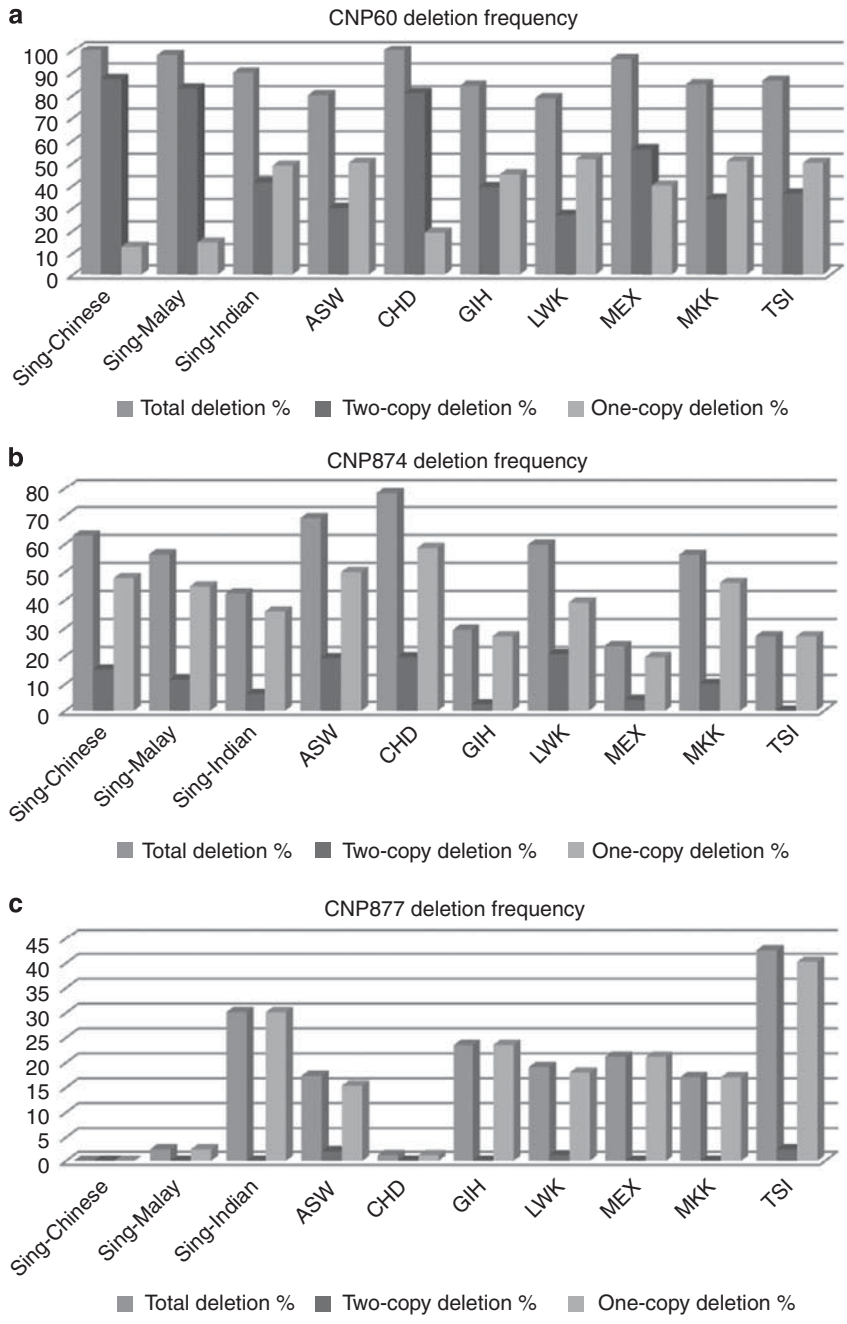


Figure 1 Total, two-copy and one-copy deletion frequencies of (**a**) CNP60, (**b**) CNP874 and (**c**) CNP877 in 10 populations.

DISCUSSION

The finding that approximately 50% of the CNPs identified by the McCarroll *et al*.[4] study were not polymorphic in all of the three Singapore populations and the HapMap III populations (CHD, GIH, MEX and TSI) suggests that the CNPs found in the 'reference' HapMap II populations are not necessarily polymorphic or common in other populations. This finding, together with the identification of novel copy number loci other than those found using the HapMap II populations, highlights the importance of characterising CNPs in different populations.

In addition, we also found several hundred CNPs that showed significant differences in integer copy numbers among the 10 populations. More interestingly, many of these loci encompass genes of medical relevance. For example, we found a markedly lower deletion frequency at CNP2203 (which is associated with the *WWOX* gene) in Sing–Chinese and Sing–Malay compared with other populations. *WWOX* is a tumour suppressor gene affected in multiple cancers.[15] On the other hand, deletion of the *UGT2B17* gene was also been found to be associated with an increased risk of prostate cancer.[16,17] The functional role of the *UGT2B17* enzyme is clear in prostate cancer, as it is involved in steroid hormone (androgen) metabolism. The mismatch of *UGT2B17* copy numbers in donors and recipients of stem cell transplantation were also associated with an increased risk of graft-versus-host disease.[18] This gene is contained within CNP603, which

Table 6 Characteristics of novel copy number loci identified in 10 populations using Birdseye

Detail	*Number (%)*	
General characteristics		
Novel copy number loci constructed from Birdseye	5014	
Number of loci that detected in ⩾two individuals	1448 (28.88)	
Number of loci that detected in ⩾1% of the studied sample size (859 samples); that is, detected in ⩾eight individuals	170 (3.39)	
Number of loci ⩾5%; that is, detected in ⩾43 individuals	42 (0.84)	
Focus on the loci detected (A) in ⩾two individuals and (B) in ⩾1% of the studied sample size	*(A) (n=1448 loci)*	*(B) (n=170 loci)*
Sum of the total length (Mb)	232.78	65.98
Average length per locus (kb)	160.76	388.11
Average number of markers per locus	82	143
Size distribution of loci		
<1 kb	56 (3.87)	5 (2.94)
1–<10 kb	325 (22.44)	31 (18.24)
10–<50 kb	420 (29.01)	46 (27.06)
50–<100 kb	165 (11.40)	10 (5.88)
100–<500 kb	354 (24.45)	35 (20.59)
500 kb–<1 Mb	91 (6.28)	22 (12.94)
>1 Mb	37 (2.56)	21 (12.35)
Deletion or duplication status		
Loci with only deletion	706 (48.76)	78 (45.88)
Loci with only duplication	547 (37.78)	29 (17.06)
Loci with deletion and duplication	195 (13.47)	63 (37.06)
Overlapping with DGV		
Loci that overlap with ⩾50% with the DGV	763 (52.69)	134 (78.82)
Loci that did not overlap with DGV	685 (47.31)	36 (21.18)
Overlapping with UCSC genes		
Loci that overlap with UCSC genes	544 (37.57)	81 (47.65)
Loci that did not overlap with UCSC genes	904 (62.43)	89 (52.35)

Abbreviations: DGV, Database of Genomic Variants; UCSC, University of California, Santa Cruz.

show substantial differences between the Singapore and HapMap III populations. Although a direct association between the CNPs and phenotypic differences is not established in our study, collectively our results suggest that CNPs distributions are substantially different between populations and thus, may account for phenotypic differences between them.

We found 12 CNPs that may have potential implications in various diseases and traits; however, only five of them have not been reported by Conrad *et al.*,[5] who found evidence of correlations for 33 copy number loci with GWAS–SNPs at $r^2>0.5$. The difference in the number of loci found to be in correlation with GWAS–SNPs between our study and the Conrad *et al.*[5] study is likely due to the limitation that we only focused on the 1291 CNPs, whereas Conrad *et al.*[5] studied the whole genome. Furthermore, it could also be due to the difference in the marker density of the microarrays used in our study and the Conrad *et al.*[5] study. We used the Affymetrix SNP Array 6.0, whereas they used a set of 20 oligonucleotide-CGH arrays, comprising 42 million probes. The differences in marker density will contribute to the differences in sensitivity of detection.[5]

Several previous studies have reported correlations between CNVs and GWAS–SNPs. For example, deletions near *IRGM* and *NEGR1* genes, which were in perfect linkage disequilibrium (LD) with the GWAS–SNPs, were identified for Crohn's disease and body mass index, respectively.[19,20] Our study also showed strong correlations between CNPs and GWAS–SNPs near *IRGM* and *NEGR1* in all 10 populations, but the deletion frequencies varied substantially among the populations. GWAS–SNPs are potentially indirect markers of disease variants, which include CNPs. This may have important clinical implications if these deletions are true disease variants.

A recent paper published by the International HapMap Consortium also studied CNPs in the HapMap III populations.[12] However, they merged and analysed the probe-level intensity data from both the Affymetrix SNP Array 6.0 and the Illumina 1M Beadchip arrays. In contrast, we only analysed the Affymetrix SNP Array 6.0 data and focused primarily on the 1291 CNPs identified previously, as only the raw signal intensity files of this array were available from the HapMap website. A total of 1610 CNPs with an estimated frequency of at least 1% of the cohort were identified in the HapMap III populations by the International HapMap Consortium. They also found that most CNPs also occurred at a low frequency.[12] This was consistent with our study where among the polymorphic CNPs, the majority also occurred at a low frequency (<10%). Similarly, the finding that the frequency spectrum of common CNPs (>10%) was similar across populations by the International HapMap Consortium was in good agreement with our results (Table 1).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was supported by the Yong Loo Lin School of Medicine, the Life Science Institute and the Office of Deputy President (Research and Technology), National University of Singapore. We also acknowledge the technical and financial support of the Genome Institute of Singapore and Agency for Science, Technology and Research, Singapore.

1 Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16,** 949–961 (2006).
2 Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36,** 949–951 (2004).
3 Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
4 McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40,** 1166–1174 (2008).
5 Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
6 Yim, S. H., Kim, T. M., Hu, H. J., Kim, J. H., Kim, B. J., Lee, J. Y. *et al.* Copy number variations in East-Asian population and their evolutionary and functional implications. *Hum. Mol. Genet.* **19,** 1001–1008 (2010).
7 Park, H., Kim, J. I., Ju, Y. S., Gokcumen, O., Mills, R. E., Kim, S. *et al.* Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.* **42,** 400–405 (2010).
8 Ku, C. S., Pawitan, Y., Sim, X., Ong, R. T., Seielstad, M., Lee, E. J. *et al.* Genomic copy number variations in three Southeast Asian populations. *Hum. Mutat.* **31,** 851–857 (2010).
9 Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453,** 56–64 (2008).

10 Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
11 Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40,** 1253–1260 (2008).
12 International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
13 Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E. *et al.* Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19,** 2154–2162 (2009).
14 Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. & Pawitan, Y. Identification of recurrent regions of Copy-Number Variants across multiple individuals. *BMC Bioinformatics* **11,** 147 (2010).
15 Lewandowska, U., Zelazowski, M., Seta, K., Byczewska, M., Pluciennik, E. & Bednarek, A. K. WWOX, the tumour suppressor gene affected in multiple cancers. *J. Physiol. Pharmacol.* **60,** 47–56 (2009).
16 Park, J., Chen, L., Ratnashinge, L., Sellers, T. A., Tanner, J. P., Lee, J. H. *et al.* Deletion polymorphism of UDP-glucuronosyltransferase 2B17 and risk of prostate cancer in African American and Caucasian men. *Cancer Epidemiol. Biomarkers Prev.* **15,** 1473–1478 (2006).
17 Karypidis, A. H., Olsson, M., Andersson, S. O., Rane, A. & Ekström, L. Deletion polymorphism of the UGT2B17 gene is associated with increased risk for prostate cancer and correlated to gene expression in the prostate. *Pharmacogenomics J.* **8,** 147–151 (2008).
18 McCarroll, S. A., Bradner, J. E., Turpeinen, H., Volin, L., Martin, P. J., Chilewski, S. D. *et al.* Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. *Nat. Genet.* **41,** 1341–1344 (2009).
19 McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40,** 1107–1112 (2008).
20 Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41,** 25–34 (2009).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Journal of Human Genetics (2011), 1–8

www.nature.com/jhg

ORIGINAL ARTICLE

Regions of homozygosity in three Southeast Asian populations

Shu-Mei Teo[1,2,3,4,5], Chee-Seng Ku[1,2,5], Agus Salim[2], Nasheen Naidoo[1], Kee-Seng Chia[1,2,4] and Yudi Pawitan[4]

The genomes of outbred populations were first shown in 2006 to contain regions of homozygosity (ROHs) of several megabases. Further studies have also investigated the characteristics of ROHs in healthy individuals in various populations but there are no studies on Singapore populations to date. This study aims to identify and investigate the characteristics of ROHs in three Singapore populations. A total of 268 samples (96 Chinese, 89 Malays and 83 Indians) are genotyped on Illumina Human 1 M Beadchip and Affymetrix Genome-Wide Human SNP Array 6.0. We use the PennCNV algorithm to detect ROHs. We report an abundance of ROHs ($\geqslant$500 kb), with an average of more than one hundred regions per individual. On average, the Indian population has the lowest number of ROHs and smallest total length of ROHs per individual compared with the Chinese and Malay populations. We further investigate the relationship between the occurrence of ROHs and haplotype frequency, regional linkage disequilibrium (LD) and positive selection. Based on the results of this data set, we find that the frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with the ROH loci. When we consider both the location of the ROHs and the allelic form of the ROHs, we are able to separate the populations by principal component analysis, demonstrating that ROHs contain information on population structure and the demographic history of a population.
Journal of Human Genetics advance online publication, 1 December 2011; doi:10.1038/jhg.2011.132

Keywords: PennCNV; regions of homozygosity; Singapore; Southeast Asian populations

INTRODUCTION

A region of homozygosity (ROH) is defined as a continuous stretch of DNA sequence without heterozygosity in the diploid state. All genetic variations such as single-nucleotide polymorphisms (SNPs) or microsatellites within the homologous DNA segments have two identical alleles that create homozygosity.[1] Currently, there is no consensus or standardized criteria to define an ROH. Previous studies focused on ROHs larger than 1 Mb which could have led to an underestimation of the true extent of homozygosity in the human genome,[2,3] whereas more recent studies define an ROH at a minimum length of 500 kb,[4] with the intention of avoiding this underestimation. This is of relevance as shorter ROHs are now also thought to be associated with complex phenotypes.[4]

The genomes of outbred populations were first shown in 2006 to contain ROHs of several megabases.[2,3,5] Their location is markedly nonrandom, where different individuals share similar region boundaries. Some loci are caused by a single common haplotype, whereas others are a consequence of several common haplotypes that could be markedly disparate.[6] Several mechanisms for the occurrence of ROHs have been suggested, including uniparental isodisomy (a chromosomal abnormality where a child inherits two identical copies of a chromosome from one parent and none from the other) and autozygosity (where a child inherits the same common ancestral haplotype chromosomal segment from both parents). Studies have found no significant violation of Mendelian transmission in these areas and concluded autozygosity as the most likely cause for the majority of ROHs observed.[7,8]

Previous studies have also investigated the population characteristics of ROHs in healthy individuals[9–11] and performed association analyses to identify ROHs that are associated with complex diseases and traits using a case–control study design.[4,12,13] However, the majority of these studies are conducted on European populations, and only a few on Asian populations. This study aims to identify and characterize ROHs in three Singapore populations, and to investigate their relationship to linkage disequilibrium (LD), haplotype frequency and positive selection.

[1]Centre for Molecular Epidemiology, National University of Singapore, Singapore; [2]Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore and [4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[5]These authors contributed equally to this work.
Correspondence: S-M Teo, NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore.
E-mail: g0801862@nus.edu.sg
or Dr Y Pawitan, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, PO Box 281, 17177 Stockholm, Sweden.
E-mail: Yudi.Pawitan@ki.se
Received 21 April 2011; revised 30 September 2011; accepted 24 October 2011

MATERIALS AND METHODS

Data

We use data from the Singapore Genome Variation Project (SGVP),[14] where a total of 292 DNA samples (consisting of 99 Chinese, 98 Malays and 95 Indians) are genotyped using the Illumina Human 1 M Beadchip and the Affymetrix Genome-Wide Human SNP Array 6.0. The characteristics of copy number variations of these populations have been investigated and reported.[15] The Chinese, Indians and Malays in Singapore descended from immigrants from neighboring countries such as China (mainly from southern provinces such as Fujian and Guangdong), India (majority from south-eastern India), Indonesia and Malaysia. The detailed information on the sources of DNA samples, demographic data of the samples, sample selection, and the origin and migration history of the three Singapore populations have been described in previous publications.[14,15] A total of 268 samples (consisting of 96 Chinese, 89 Malays and 83 Indians) are used in the subsequent analysis after removing samples on the basis of high rates of SNP missingness (greater than 2%), excessive heterozygosity or cryptic relatedness by excessive identity-by-states. Population membership is ascertained on the basis that all four grandparents belong to the same population, and samples that display either evidence of admixture or clear evidence of discordance between self-reported and genetically inferred population membership are excluded.

SNP genotypes are obtained from the SGVP website (http://www.nus-cme.org.sg/sgvp/). These SNPs have undergone a series of quality control measures,[14] including removing SNPs with SNP missingness $\geqslant 5\%$ and *P*-value <0.001 for a test of departure of Hardy–Weinberg Equilibrium (HWE), resulting in ~ 1.58 million SNPs per population remaining. Quality control measures were conducted seperately for each of the populations.

Identification of individual-specific ROHs

Individual-specific ROHs are identified using the PennCNV algorithm[16] for the Illumina and Affymetrix arrays based on the log R ratio and B allele frequency for each sample. The ROHs identified by PennCNV are copy neutral events, meaning that one-copy deletions are excluded. We exclude regions <500 kb. To further filter regions that may be called erroneously by PennCNV, we check the SNPs genotypes for the number of heterozygous genotypes within the region. Ideally, we would expect no heterozygous genotypes in the region, but we allow for some heterozygosity that may be due to genotyping errors or other causes.

We investigate the effect of allowing some heterozygosity on the relationship between ROH and LD. From a simulation (see Supplementary Methods, 'Simulation' section), we observe that ROH detection is very sensitive to heterozygosity present either due to mutation or genotyping errors, whereas the LD in the region is largely preserved despite the mutations introduced. By not allowing any heterozygosity, we miss detecting older ROHs in many individuals and this affects the formation of the common regions. So, to capture the LD/haplotype structure using ROHs, it is important to allow a small percentage of heterozygosity.

We use a binomial probability upper bound to calculate a confidence score for each region (see Supplementary Methods, 'Confidence scores calculation' section). The confidence score takes into account the amount of heterozygosity, as well as the SNP density, and is an indication of how confident we are that the ROH is true. In general, the confidence scores for regions detected by the Affymetrix platform are lower than that detected by the Illumina platform (see Supplementary Methods Figure S1). We decide to use the Illumina platform with more than 1 million SNPs for ROHs detection but still use the combined genotypes from 1.58 million SNPs from both platforms in the calculation of confidence scores. Several summary statistics are computed to describe and compare the characteristics of ROHs in the three Singapore populations.

Identification of common ROHs

We identify common ROH loci using a previously published method.[17] We define common loci as regions with consecutive probes where at least 5% of the subjects (that is, 13 subjects) have individual regions that overlap with the probes. Occasionally, individual regions within a common locus can show considerable variations in their boundaries, resulting in a heterogeneous region. To refine the identified common loci, we form clusters of regions by requiring all individual regions within a cluster to overlap by at least 80%. For each common locus, individual regions are said to be concordant if it overlaps with at least 80% of the length of the locus. Common loci with <2 concordant individuals or <500 kb or having a SNP density <0.2 (SNP per kb) are discarded. The common loci are further refined as the intersection of the concordant regions. We perform population comparisons and test of departure of HWE for each locus. For each set of tests, we account for multiple comparisons using the false discovery rate,[18] with results or discoveries considered interesting at false discovery rate of <0.01.

Quantification of regional LD

The two most widely used measures to quantify the amount of LD between two markers are the D' and r^2 statistics.[19] Here, instead of LD between two markers, we are interested in the amount of LD in a region. For all SNPs in a region, we calculate the pairwise D' (and r^2). We perform eigenvalue decomposition on the D' (r^2) matrix and calculate the percentage explained by the first eigenvalue (y). This percentage will take values between $100/n$ and 100, where n is the number of (polymorphic) SNPs in the region. To make the percentages comparable across regions with different number of SNPs, we scale it such that the value varies between 0 and 1. So, $y^*=(y-100/n)/(100-100/n)$. The higher the value of y^*, the stronger the LD in the region.

Haplotypes in ROH loci

For each common locus, we use phased genotypes (using the program fastPHASE version 1.3, see Supplementary Methods in Teo *et al.*[14] for details on the choice of parameters for phasing) to determine the different haplotypes present in the three populations. To reduce the dimension of the data, we consider only the top three most frequent haplotypes and combine the others as 'other haplotypes', that is, we categorize each region into four alleles (top three most common haplotypes and 'other' haplotypes). Each individual has two alleles for each region. For convenience, we will refer to the alleles as A, B, C and D.

Identification of regions with differential LD between populations

We use a previously published method, varLD,[20,21] to identify regions with differential LD between populations. Briefly, the method tests for equality between two LD matrices for a user-defined window size, shifting each window one SNP at a time. We calculate the varLD score for a window size of 50 SNPs for the signed r^2 matrices.[21] For each pair of populations, a region is considered to have differential LD if consecutive positions are above the 95th percentile of the genome-wide varLD score. We restrict to regions >500 kb for comparison with ROHs. We exclude the region if it overlaps by $>50\%$ with copy number variations previously reported for the same set of individuals,[14] as LD measures for regions that encapsulate copy number variations may not be reliable.[21]

RESULTS

Summary statistics of individual ROHs

We discard regions whose confidence scores are below the 25th percentile of the confidence scores. Table 1 summarizes the characteristics of ROHs. On average, the Indian population has lower number of ROHs compared with the Chinese and Malay populations. There is wide inter-individual difference in the number of ROHs, which ranges from 98 (sample 334_01 and 461_01) to 241 (sample 81_01). More than 80% of the ROHs are <1 Mb in length. The largest ROH spans a length of ~ 68.5 Mb, and is detected in one Indian individual (sample 408_01) in Chromosome 3. A total of 32 ROHs larger than 10 Mb are detected (Table 2). Interestingly, three Indian samples (397_01, 290_01 and 408_01) have five or more of these 'extremely long' ROHs. Figure 1 plots the number of ROHs versus the total length of ROHs in each individual. We see clusters of the three populations, indicating that number and length of ROHs differ among populations. This result was also observed by Kirin *et al.*[22]

Table 1 Characteristics of ROHs in three Singapore populations (using 1 029 591 SNPs from the Illumina 1 M platform)

Characteristics	*Chinese (n=96)*	*Malay (n=89)*	*Indian (n=83)*
Number of ROHs per individual			
Mean	207	179	126
Median	206	178	127
Minimum	157	123	98
Maximum	241	228	173
Length of ROHs (in kb)			
Mean	800.9	806.1	879.6
Median	670.5	672.8	666.3
Maximum	23 230	21 850	68 500
Total length of ROHs per individual (in Mb)			
Mean	166.1	144.6	111.2
Median	165.7	143.4	100.5
Minimum	115.4	90.49	73.91
Maximum	195.4	191.9	315.6
Size distribution of ROHs (proportion, %)			
500kb–1 Mb	83.0	82.5	83.5
$\geqslant$1 Mb	17.0	17.5	16.5

Abbreviations: ROHs, regions of homozygosity; SNPs, single-nucleotide polymorphisms.

Summary statistics of common ROHs

We identify 1256 common ROH loci in all three populations (Supplementary Table 1), where 90% of the loci overlap with UCSC genes (http://genome.ucsc.edu/), and 292 (23%) overlap with genes listed in the Online Mendalian Inheritance in Man Morbid Map (ftp://ftp.ncbi.nih.gov/repository/OMIM/ARCHIVE/morbidmap). For each locus, we test for differences among the three populations in terms of ROH frequencies and haplotype frequencies, and 47 loci ($<$4%) differ significantly in frequencies while 899 loci (69%) differ significantly in haplotype frequencies among the populations. Approximately 52% of the loci are detected in $>$5% (more common ROH loci) of individuals (Figure 2). Figure 3 shows the length distribution of the ROH loci; $\sim$78% of the ROH loci are $\leqslant$1 Mb, and majority of the long ROH loci ($>$1Mb) are in the range of 1–2 Mb. The proportion of the genome that is in the different ROH length categories differs among the three populations (Figure 4). The Chinese and Malays have more ROHs of shorter lengths compared with the Indians, while the Indians have more ROHs in the longer length categories ($>$4 Mb).

We compare the common loci we found to that published in previous studies.[10,23] Two regions are defined to overlap if the regions have a reciprocal overlap of at least 50%. Nothnagel *et al*.'s study[10] surveys ROHs in Europeans; we found that all 10 regions listed as 'ROH islands' (meaning they have a high population frequency) in their study overlap with an ROH loci found in this study, suggesting

Table 2 ROHs larger than 10 Mb

Chromosome	*Start*	*End*	*Length*	*Sample*	*Ethnicity*
1	120 837 663	143 420 875	22 583 213	108_01	Chinese
6	3 217 193	26 449 280	23 232 088	17_01	Chinese
16	34 034 376	45 968 704	11 934 329	131_01	Chinese
1	120 837 663	143 420 875	22 583 213	218_01	Chinese
8	41 842 707	52 102 021	1 025 9315	465_01	Malay
1	94 915 135	108 531 282	13 616 148	174_01	Malay
11	19 924 676	41 772 573	21 847 898	174_01	Malay
1	67 073 684	90 862 713	23 789 030	290_01	Indian
3	175 758 479	190 839 635	15 081 157	290_01	Indian
4	41 334 756	55 223 410	13 888 655	290_01	Indian
6	112 768 454	147 227 544	34 459 091	290_01	Indian
11	86 911 515	131 748 067	44 836 553	290_01	Indian
14	71 970 357	88 634 741	16 664 385	290_01	Indian
17	152 362	10 559 477	10 407 116	290_01	Indian
3	102 253 981	170 758 820	68 504 840	408_01	Indian
6	79 661	14 549 208	14 469 548	408_01	Indian
6	121 892 269	132 743 942	10 851 674	408_01	Indian
13	52 267 631	77 893 750	25 626 120	408_01	Indian
13	78 497 109	94 452 168	15 955 060	408_01	Indian
22	37 722 142	49 582 267	11 860 126	408_01	Indian
13	74 778 668	100 318 094	25 539 427	367_01	Indian
3	158 294 635	169 108 914	10 814 280	361_01	Indian
6	90 065 419	106 409 967	16 344 549	397_01	Indian
7	28 668 234	43 132 968	14 464 735	397_01	Indian
7	80 118 053	105 839 742	25 721 690	397_01	Indian
8	590 729	10 908 015	10 317 287	397_01	Indian
9	33 415 385	45 059 163	11 643 779	397_01	Indian
9	66 448 030	100 731 809	34 283 780	397_01	Indian
10	121 636	24 722 946	24 601 311	397_01	Indian
15	63 308 076	73 720 143	10 412 068	397_01	Indian
7	9 983 924	21 830 396	11 846 473	76_01	Indian
13	47 337 000	72 862 520	25 525 521	76_01	Indian

Abbreviations: ROHs, regions of homozygosity.

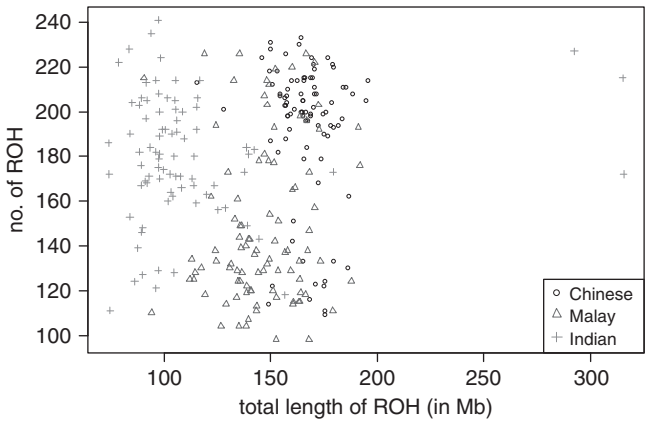


Figure 1 Number of ROH versus total length of ROHs in each individual. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

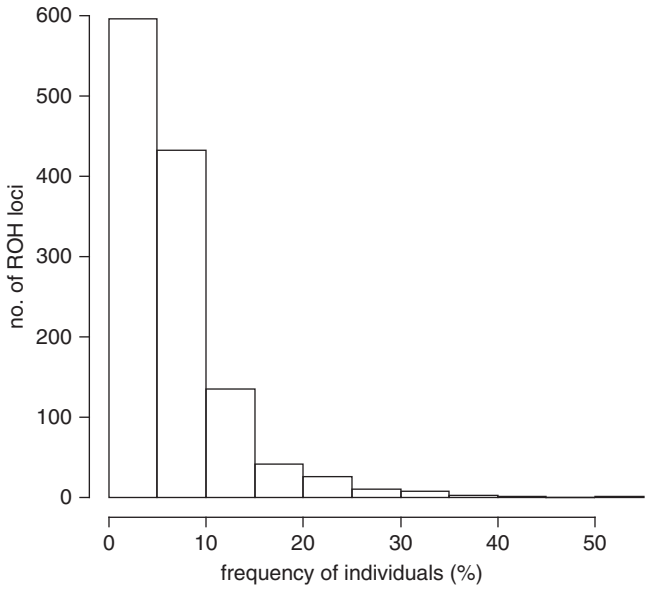


Figure 2 Number of ROH loci in the respective population frequency classes.

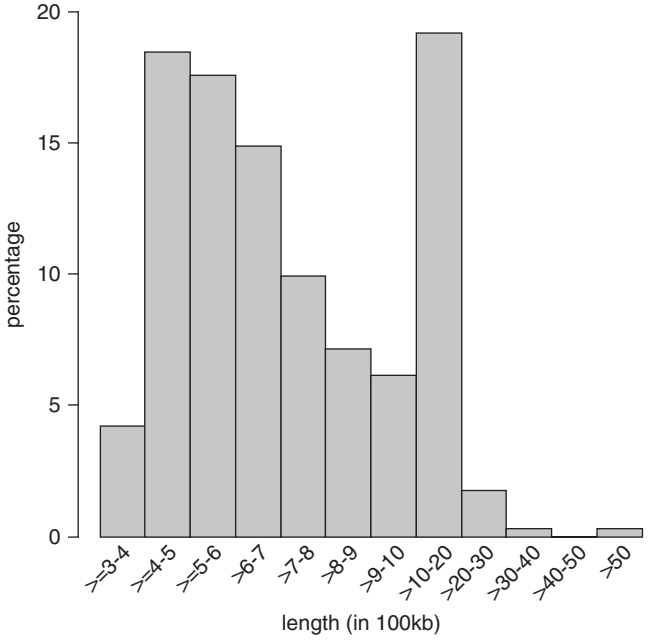


Figure 3 Percentage of ROH loci in the respective length classes.

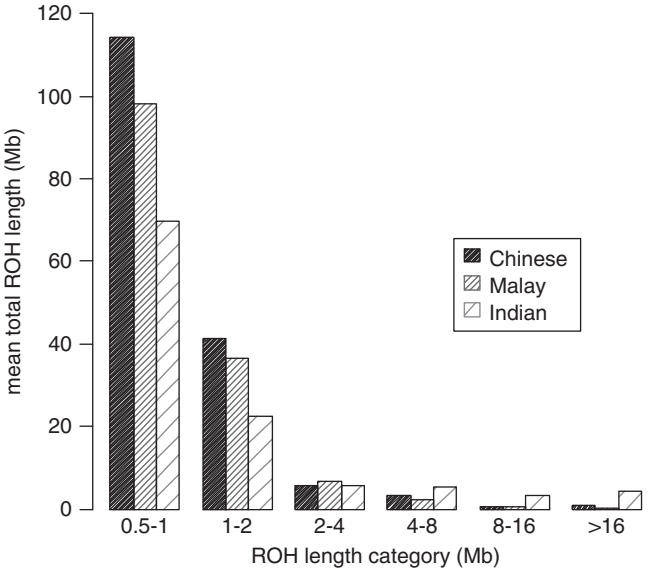


Figure 4 Percentage of ROH loci in the respective length classes. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

that these regions are not specific to Europeans (see Supplementary Methods Table S1). The population frequencies of these ROHs in our populations differ from that reported in Nothnagel *et al*.'s study,[10] but formal testing is inappropriate as the methods used to calculate the frequencies are different.

Auton *et al*.'s study[23] surveys ROHs in Mexicans, Europeans, East Asians and South Asians; we found that out of 34 high-frequency ROHs (defined as being found in at least 10% of individuals within a population) 11 overlap with an ROH locus found in our study (see Supplementary Methods Table S2). All the regions that overlap are found in the East Asian population, except for one region in Chromosome 4, which is present in all populations. The frequencies of these ROHs are, however, quite low in our population (1–4%).

Association with haplotype frequency and regional LD

Figure 5 shows that the frequency of an ROH is positively associated with the total frequency of the top three haplotypes (correlation of 0.69), and Figure 6 shows that as the frequency of an ROH increases, so does $y*_{D'}$ and $y*_{r^2}$ (figure is shown for the Malay population, similar figures for the Chinese and Indians are shown in Supplementary Methods Figures S2 and S3). If we assume random mating, the homozygosity of any region will be high when there are few haplotypes present at high frequency, thus it reinforces autozygosity as the mechanism for the occurrence of an ROH. These empirical results suggest that there is positive correlation between the frequency of an ROH and the frequency of the common haplotypes, and also between the frequency of an ROH and LD in the region.

Frequency of ROHs and frequency of haplotypes within ROHs

To assess if there is a difference in the location and frequency of ROHs among the populations, we perform principal component analysis

(PCA) using absence/presence of the common ROH loci. For each individual, we check if that individual has an ROH that is concordant with the common ROH. We can view the matrix input for the PCA analysis as a matrix of 1's and 0's where each row corresponds to an individual and each column corresponds to a common loci, so that the (i, j) entry indicates whether individual i has a concordant ROH at locus j. From Figure 7, we see that the Indians are quite well separated from the Chinese and Malays, and that there is some separation between the Chinese and Malays. This implies that the location and frequency of occurrence of ROHs differ among populations.

However, interestingly, populations can share the same (or similar) ROH location, but the common haplotypes driving the ROH can be markedly disparate. One example is a 700-kb ROH in Chromosome 16 (location 30,438,046–31,137,964) that overlaps with the Vitamin K epoxide reductase complex subunit 1 (*VKORC1*) gene (location 31,009,956–31,013,551). Genetic polymorphisms within the gene have been found to correlate with differences in warfarin dosage and response in many studies.[24–26] In the Singapore populations, the Indians were observed to display warfarin resistance, thus requiring a higher dose as compared with the Chinese and Malays.[26–29] There is no significant difference in ROH frequencies among the populations (ROH frequencies of 21, 13 and 20% for the Chinese, Malays and Indians, respectively). However, if we examine the haplotypes in this region, there is significant difference. Fisher's exact test performed on the frequencies of the top three most frequent haplotypes results in a P-value $< 10^{-6}$. In particular, the difference in haplotype frequencies of the Indians differs markedly from the Chinese and Malays. This is highlighted in Table 3, where haplotype A dominates in the Chinese and Malays but is almost absent in the Indians, while haplotype B dominates in the Indians but is almost absent in the Chinese and Malays. Haplotypes A and B differ at 104 locations out of the 158 SNPs in this region.

We also perform PCA on the allele counts of the haplotypes as described in the section 'Haplotypes in ROH loci'. The first two components separates the Indians from the Chinese and Malays while the third component further separates the Chinese from the Malays (see Figure 8). This suggests that ROH loci contain much genetic ancestral haplotype information of a population.

Testing departure from HWE

Using the estimated frequencies of the top three haplotypes, we are able to calculate the expected frequencies of the corresponding genotypes. For the observed frequencies, we use the unphased genotypes. For each individual, we can identify the haplotypes without phase information when all the SNPs in the region are homozygous

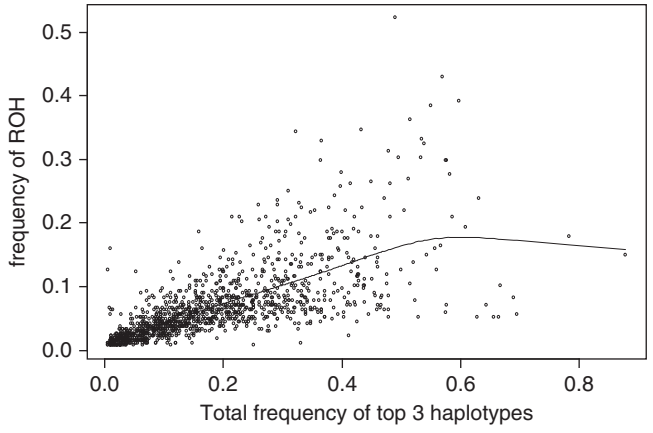


Figure 5 Frequency of ROH loci versus total frequency of top three haplotypes.

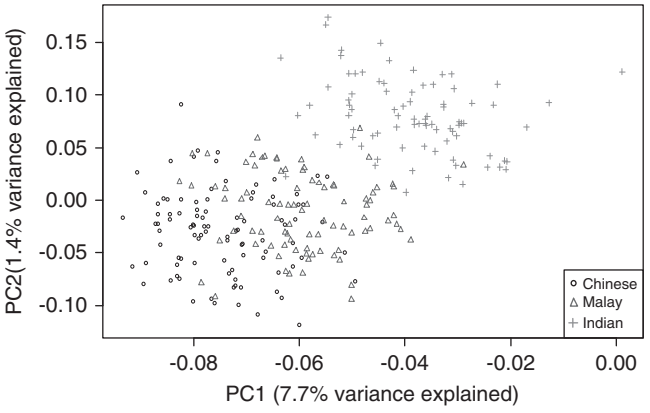


Figure 7 Principal component 2 versus principal component 1 using absence/presence of 1256 common ROHs. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

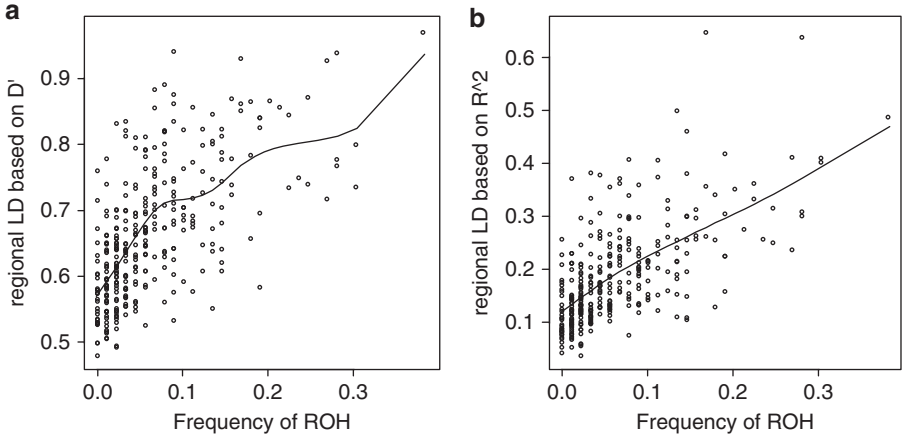


Figure 6 Regional LD versus frequency of ROH based on (**a**) D' matrix and (**b**) r^2 matrix. These results are based on the Malay population.

Table 3 Haplotype frequencies of three populations in a ROH in Chromosome 16 that overlaps with VKORC1

	Haplotype A	*Haplotype B*	*Haplotype C*
Chinese	0.31	0.0052	0.099
Malay	0.28	0.045	0.10
Indian	0.0060	0.34	0

Abbreviation: ROH, region of homozygosity.

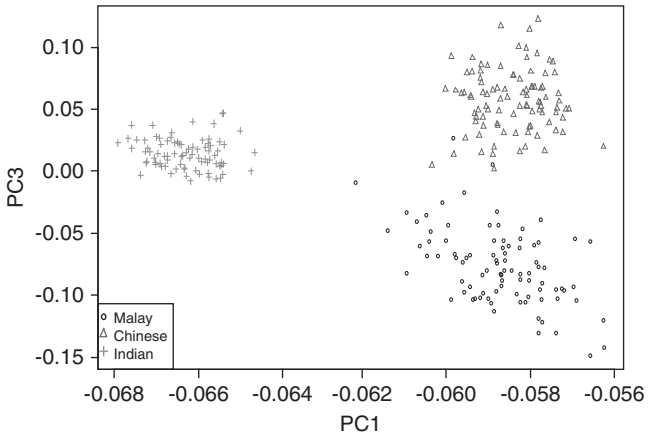


Figure 8 Results of PCA on haplotype frequencies of ROH regions. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

(removing SNPs where we had allowed heterozygosity in the detection of ROH). With that, we are able to obtain observed frequencies for the (A, A), (B, B) and (C, C) genotypes. We use the χ^2 test with three degrees of freedom to test if there is departure of the observed from the expected. A large majority of ROH loci ($>92\%$) adhere to HWE, suggesting that assumptions of autozygosity and random mating are true for most ROH loci. Of the regions that show departure from HWE (false discovery rate <0.01), majority show excess homozygosity than would be expected. The reasons for departure from HWE are not immediately clear, and could be due to various reasons such as positive selection (see section 'Comparison with regions associated with positive selection') or nonrandom mating.

Comparison with varLD

As described in the section 'Identification of regions with differential LD between populations', we identify 16, 10 and 13 regions with differential LD variation between the Chinese and Indian populations, Malay and Indian populations, and Chinese and Malay populations, respectively. Of the 16 regions, 14 overlap with a common ROH and 10 out of 14 show significant differences in haplotype frequency between the Chinese and Indian populations. Of the 10 regions, 7 overlap with a common ROH and 7 out of 7 show significant differences in haplotype frequency between the Malay and Indian populations. Of the 13 regions, 8 overlap with a common ROH and 8 out of 8 show significant differences in haplotype frequency between the Chinese and Malay populations.

We observe that the majority of regions (74%) that show LD differences between populations correspond to regions where ROHs are observed, and furthermore, the haplotype frequencies in these regions differ between the populations. These results indicate that ROH patterns explain a large proportion of LD variations.

Comparison with regions associated with positive selection

We investigate if the regions detected for recent positive natural selection overlap with ROHs. We consider the top 10 candidate regions for recent positive selection in each of the populations, as published in a previous study.[14] These regions were detected based on the clustering of SNPs with high integrated haplotype score.[30] Out of the 30 regions considered, 28 regions overlap with a common ROH defined in this study, with 20 regions completely within an ROH and the other 8 regions with a high percentage of overlap (at least 60%). This suggests the occurrence of ROHs as a possible consequence of positive selection, where the positively selected haplotypes rise to a high frequency, resulting in a high possibility of ROHs due to autozygosity.

Out of the 28 regions, 10 of them overlap with an ROH that failed HWE. Performing Fisher's exact test on a 2 by 2 table with indicators for departure from HWE and indicators for positive selected regions as rows and columns, we obtain an odds ratio of 1.89 (P-value=0.05). The departure from HWE may be a consequence of positive selection. An ROH that has a higher frequency than would be expected for its length may also be an evidence of positive selection (see Supplementary Methods Figure S8).

Effect of heterozygosity on the relationship between ROH and LD

When we filter the individual regions using a stricter confidence threshold of the 75th percentile (that is, allowing less heterozygosity), we identify 414 common regions, but the relationship of these regions with haplotype frequency, regional LD and positive selection is weak (see Supplementary Methods Figures S4 and S5 and section Comparison with VarLD (results based on these 414 common regions)). We also see poorer separation of the populations by PCA, but this is likely due to the fewer number of common regions identified. At the 25th percentile threshold, the percentage of heterozygosity is still kept low at $<5\%$ for a large majority of the regions (See Supplementary Methods Figure S9). With an overly strict confidence score threshold, many regions are omitted and this decreases the number of common regions formed from 1256 to 414. Allowing for some heterozygosity within the regions allows detection of older ROH loci (heterozygosity caused by recent recombination or mutation), which have a stronger relationship with LD and positive selection (see Simulation section in Supplementary Methods).

DISCUSSION

In summary, this study identifies and investigates the population characteristics of ROHs in three Singapore populations, Chinese, Malay and Indian. We report an abundance of ROHs, with an average of >100 ROHs per individual. On average, the Indians have lower numbers and total length of ROHs per individual than the Chinese and Malays, possibly indicative of a larger founder population. However, there are several Indians with multiple large ROHs, suggesting that they may be offsprings of parents who are close relatives. In India, consanguineous marriages are more prevalent in the South, especially in Tamil Nadu, from where many Singapore Indians descended. From the Consanguinity/Endogamy Resource (http://www.consang.net/index.php/Main_Page), data from a 1982 study have shown the prevalence of consanguineous marriages among Singapore Indians to be 4% compared with only 0.3% in Singapore Chinese. Published data have shown that the number of ROHs of several megabases increase markedly in the offsprings of consanguineous marriages,[3,24] with an average of 6.25% homozygosity expected in the genome of the offsprings of first cousin marriages.[7] Li *et al.*[3] have shown that in a family with four children from first cousin

marriages, multiple ROHs ranging from 3.06 to 53.17 Mb were observed in all the children. Woods *et al.*[24] have also shown a marked increase in homozygosity levels in individuals with a recessive disease whose parents were first cousins, where, on average, 11% of their genomes were homozygous.

In addition, we identify 1256 common ROH loci, and investigate the occurrence of ROHs and haplotype frequency, regional LD and positive selection. Based on the results for this data set, we find that the frequency of occurrence of ROHs is positively associated with haplotype frequency and regional LD. The preferential occurrence of ROHs in regions of high LD and low recombination has also been observed in other studies.[10] The majority of regions detected for recent positive selection and regions with differential LD between populations overlap with ROH loci. By considering both the location of the ROH and the allelic form of the ROH, we are able to separate the populations by PCA, demonstrating that ROHs contain information on population structure and the evolutionary and demographic history of a population.

The ability of genome-wide SNP markers for population structure analysis has been widely acknowledged. Here, we are not proposing the superiority of ROHs in population structure analysis. It is expected that using genome-wide SNP data allows very good separation of populations through PCA because of the amount of information it contains (see 14 for PCA analysis using SNPs on the same population samples). In this paper, we have shown that it is possible to distinguish populations using just $\sim$1000 segments of the genome. Comparatively, if we were to choose 1000 random segments of the genome and perform a similar analysis, we would not obtain as good a separation as with ROHs (see Supplementary methods Figure S7). The unique characteristics of ROHs allow us to study common haplotypes conveniently; it is complementary to SNP-based analysis. In SNP-based analysis, we simply compare SNP-level frequencies between populations but in ROH-based analysis, we are able to capture differences in LD or haplotype structures.

Majority of the ROH loci overlap with known genes but their association with complex phenotypes is still rudimentary. This warrants further characterization of ROHs in different populations, investigation of their roles in the genetics of complex phenotypes and further studies of population evolutionary genetics. These future studies will be of importance given the abundance of ROHs in the human genome and the differences of ROHs between populations.

A sufficiently large number of SNPs is required to accurately detect ROHs.[1,2] To this end, we have used two highly dense SNP arrays (Illumina 1 M and Affymetrix 6.0) with $>$1.58 million unique SNPs. Using a confidence score metric that takes into account percentage of heterozygosity as well as the number of SNPs in the region, we discard individual regions whose confidence scores are below the 25th percentile of the confidence scores. We use the PennCNV algorithm that relies on signal intensity data to detect putative ROHs. We then filter out false positives by checking SNP genotypes within the ROH. To our knowledge, most studies on ROHs use only SNP genotypes, but this approach may produce false positives caused by hemizygous deletions. On the other hand, due to the noise in signal intensity data, the regions called by PennCNV could also result in false-positive regions. We feel it is important to use a combination of the methods (that is, signal intensity data and genotype data) to minimize false-positive rates.

We also use PLINK, a widely used software for ROH detection, on genotypes from both platforms using the following parameters: 500 kb window with two heterozygous SNPs allowed, minimum length of 500 kb, 50 SNPs as minimum number of SNPs and minimum density of 1 SNP per 10 kb. We find that 75% of the regions found by PennCNV are detected by PLINK, suggesting that the results of the analysis will likely give similar conclusions using PLINK. A formal and systematic comparison of multiple algorithms for ROH detection will be interesting.

Potential biases in the detection of ROHs include false-negative regions due to ascertainment bias in SNP selection for the SNP arrays and false-positive regions due to the lack of minor allele frequency (MAF) criterion applied before the identification of ROHs. With regards to the former, SNPs from genotyping platforms are mostly tagged SNPs from the HapMap project, so populations that were not analyzed in the HapMap project will have less chance of their population-specific SNPs being included in the array. However, both the Illumina 1 M and Affymetrix 6.0 arrays have a high marker density and uniformity. With regards to the later, we do not expect our results to be affected considerably by not filtering SNPs with low MAF, for several reasons. First, we have very dense SNP genotyping data of $>$1.58 million SNPs, and as an ROH is defined as a region of consecutive homozygosity of $>$500 kb, it is unlikely that there exists a large number of consecutive low-MAF SNPs that cause a false-positive identification. In any case, these monomorphic/near monomorphic SNPs are uninformative and would not affect the haplotype analyses. It is of concern if the region is detected because the monomorphic/low-MAF SNPs are genotyped, whereas other SNPs present in the region are missed (due to ascertainment bias). However, as ROH detection is not reliant on a single SNP, but on many consecutive homozygous SNPs in a 500 kb region, we do not expect either issue to be of serious concern.

Some studies[23] have adopted the strategy of removing SNPs in high LD before defining an ROH (that is, thinning the data set but requiring a lower number of SNPs for the definition of ROH). However, we found poor correlation between the frequency of the ROHs we identified and the mean or median pairwise D' or r^2 statistics (for SNPs within the ROH, up to 250 kb apart, see Supplementary Methods Figure S6), meaning that a SNP being in high LD in the vicinity is not sufficient for its inclusion in an ROH, and a SNP in low LD is not sufficient for its exclusion in an ROH.

In conclusion, our study is one of the first to describe the population characteristics of ROHs in the three Singapore populations (Chinese, Malay and Indian). Our results are in support that ROHs contain population demographic and ancestral haplotype information.

ACKNOWLEDGEMENTS

We thank Dr Teo Yik Ying for helpful discussions related to this work and Rick Ong for identifying regions with differential LD between populations using the VarLD program. TSM acknowledges support from the National University of Singapore Graduate School for Integrative Sciences and Engineering (NGS) Scholarship.

1 Ku, C. S., Naidoo, N., Teo, S. M. & Pawitan, Y. Regions of homozygosity and their impact on complex diseases and traits. *Hum. Genet.* **129,** 1–15 (2011).
2 Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15,** 789–795 (2006).
3 Li, L. H., Ho, S. F., Chen, C. H., Wei, C. Y., Wong, W. C., Li, L. Y. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum. Mutat.* **27,** 1115–1121 (2006).
4 Yang, T. L., Guo, Y., Zhang, L. S., Tian, Q., Yan, H., Papasian, C. J. *et al.* Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. *J. Clin. Endocrinol. Metab.* **95,** 3777–3782 (2010).
5 Simon-Sanchez, J., Scholz, S., Fung, H. C., Matarin, M., Hernandez, D., Gibbs, J. R. *et al.* Genome-wide SNP assay reveals structural genomic variation, extended homo-

zygosity and cell-line induced alterations in normal individuals. *Hum. Mol. Genet.* **16,** 1–14 (2007).
6 Curtis, D., Vine, A. E. & Knight, J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann. Hum. Genet.* **72,** 261–278 (2008).
7 Broman, K. W. & Weber, J. L. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am. J. Hum. Genet.* **65,** 1493–1500 (1999).
8 Curtis, D. Extended homozygosity is not usually due to cytogenetic abnormality. *BMC Genet.* **8,** 67 (2007).
9 McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R, Franklin, C. S., Pericic, M., Barac-Lauc, L. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83,** 359–372 (2008).
10 Nothnagel, M., Lu, T. T., Kayser, M. & Krawczak, M. Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. *Hum. Mol. Genet.* **19,** 2927–2935 (2010).
11 O'Dushlaine, C. T., Morris, D., Moskvina, V., Kirov, G., Consortium, I. S., Gill, M. *et al.* Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18,** 1248–1254 (2010).
12 Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl Acad. Sci. USA* **104,** 19942–19947 (2007).
13 Nalls, M. A., Guerreiro, R. J., Simon-Sanchez, J., Bras, J. T., Traynor, B. J., Gibbs, J. R *et al*. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics* **10,** 183–190 (2009).
14 Teo, Y. Y., Sim, X., Ong, R. T., Tan, A. K., Chen, J., Tantoso, E. *et al.* Singapore genome variation project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19,** 2154–2162 (2009).
15 Ku, C. S., Pawitan, Y., Sim, X., Ong, R. T., Seielstad, M., Lee, E. J. *et al.* Genomic copy number variations in three Southeast Asian populations. *Hum. Mutat.* **31,** 851–857 (2010).
16 Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. *et al*. 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 (2007).
17 Mei, T. S., Salim, A., Calza, S., Seng, K. C., Seng, C. K. & Pawitan, Y. Identification of recurrent regions of copy-number variants across multiple individuals. *BMC Bioinformatics* **11,** 147 (2010).
18 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).
19 Lewontin, R. C. & Kojima, K. The evolutionary dynamics of complex polymorphisms. *Evolution* **14,** 458–472 (1960).
20 Ong, R. T. H. & Teo, Y. Y. varLD: A program for quantifying variation in linkage disequilibrium patterns between populations. *Bioinformatics* **26,** 1269–1270 (2010).
21 Teo, Y. Y., Fry, A. E., Bhattacharya, K., Small, K. S., Kwiatkowski, D. P. & Clark, T. G. Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res.* **19,** 1849–1860 (2009).
22 Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., Mckeigue, P. M. & Wilson, J. F. Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5,** e13996 (2010).
23 Auton, A., Bryc, K., Boyko, A. R., Lohmueller, k.E., Novembre, J., Reynolds, A. *et al.* Global distribution of genomic diversity underscores rich complex history of continental human populations. *Genome Res.* **19,** 795–803 (2009).
24 Woods, C. G., Cox, J., Springell, K., Hampshire, D. J., Mohamed, M. D., McKibbin, M. *et al.* Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am. J. Hum. Genet.* **78,** 889–896 (2006).
25 Aquilante, C. L., Langaee, T. Y., Lopez, L. M., Yarandi, H. N., Tromberg, J. S., Mohuczy, D. *et al.* Influence of coagulation factor, vitamin K epoxide reductase complex subunit 1, and cytochrome P450 2C9 gene polymorphisms on warfarin dose requirements. *Clin. Pharmacol. Ther.* **79,** 291–302 (2006).
26 Harrington, D. J., Underwood, S., Morse, C., Shearer, M. J., Tuddenham, E. G. D. & Mumford, A. D. Pharmacodynamic resistance to warfarin associated with a Val66Met substitution in vitamin K epoxide reductase complex subunit 1. *Thromb. Haemost.* **93,** 23–26 (2005).
27 Zhu, Y., Shennan, M., Reynolds, K. K., Johnson, N. A., Herrnberger, M. R., Valdes, R. Jr. *et al.* Estimation of wafarin maintenance dose based on VKORC1 (−1639 G > A) and CYP2C9 genotypes. *Clin. Chem.* **53,** 1199–1205 (2007).
28 Yuen, E., Gueorguieva, I., Wise, S., Soon, D. & Aarons, L. Ethnic differences in population pharmacokinetics and pharmacodynamics of warfarin. *J. Pharmacokinet. Pharmacodyn.* **37,** 3–24 (2009).
29 Lee, S. C. Inter-ethnic variability in warfarin requirement is explained by VKORC1 genotype in an Asian population. *Clin. Pharmacol. Ther.* **79,** 197–205 (2006).
30 Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of positive selection in the human genome. *PLoS Biol.* **4,** e72 (2006).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

Journal of Human Genetics (2010) 55, 403–415

www.nature.com/jhg

REVIEW

The discovery of human genetic variations and their use as disease markers: past, present and future

Chee Seng Ku[1], En Yun Loy[1], Agus Salim[1], Yudi Pawitan[2] and Kee Seng Chia[1,2]

The field of human genetic variations has progressed rapidly over the past few years. It has added much information and deepened our knowledge and understanding of the diversity of genetic variations in the human genome. This significant progress has been driven mainly by the developments of microarray and next generation sequencing technologies. The array-based methods have been widely used for large-scale copy number variation (CNV) detection in the human genome. The arrival of next generation sequencing technologies, which enabled the completion of several whole genome resequencing studies, has also resulted in a massive discovery of genetic variations. These studies have identified several hundred thousand short indels and a total of thousands of CNVs and other structural variations in the human genome. The discovery of these 'newer' types of genetic variations, indels, CNVs and copy neutral variations (inversions and translocations) has also widened the scope of genetic markers in human genetic and disease gene mapping studies. The aim of this review article is to summarize the latest developments in the discovery of human genetic variations and address the issue of inadequate coverage of genetic variations in the current genome-wide association studies, which mainly focuses on common SNPs. Finally, we also discuss the future directions in the field and their impacts on next generation genome-wide association studies.
Journal of Human Genetics (2010) **55,** 403–415; doi:10.1038/jhg.2010.55; published online 20 May 2010

Keywords: copy number variations; genome-wide association studies; human genetic variations; indels; loss of heterozygosity; restriction fragment length polymorphisms; single nucleotide polymorphisms; tandem repeats

INTRODUCTION

Human genetic variations are the differences in DNA sequence within the genome of individuals in populations. Genetic variations in the human genome can take many forms, including single nucleotide changes or substitutions; tandem repeats; insertions and deletions (indels); additions or deletions that change the copies number of a larger segment of DNA sequence; that is, copy number variations (CNVs); other chromosomal rearrangements such as inversions and translocations (also known as copy neutral variations); and copy neutral loss of heterozygosity (LOH) or homozygosity. These genetic variations span a spectrum of sizes from single nucleotides to megabases. Single nucleotide substitutions or alterations, as implied in the terminology, involve a change in a single nucleotide at a particular locus in the DNA sequence, such as restriction fragment length polymorphisms (RFLPs), single nucleotide polymorphisms (SNPs) and single nucleotide indels. On the other extreme, CNVs, inversions, translocations and LOHs encompass larger segments of DNA sequences that range from kilobases to megabases (>1 kb), whereas tandem repeats and indels fall in between the extremes (from >1 bp to 1 kb).

In general, these genetic variations take place naturally in the human genome, and they are the footprints of errors or mistakes that occur in DNA replication during cell division, although external agents, such as viruses and chemical mutagens, can also induce changes in the DNA sequence. The occurrence of each type of genetic variation is mediated by different mechanisms; nonetheless, most of these molecular events or processes are currently unclear and are still being investigated. For example, several mechanisms have been proposed to explain the widespread occurrence of CNVs in the human genome, such as nonallelic homologous recombination and nonhomologous end joining.[1] However, for copy neutral LOHs, the homozygosity could have resulted from uniparental isodisomy and autozygosity.[2] Regardless of the molecular mechanisms or processes that generated the genetic variations, they can be broadly classified as either somatic or germline variations depending on whether they arose from mitosis or meiosis, respectively.

The field of human genetic variations has advanced considerably over the past five years. It has added much information and deepened our knowledge and understanding of the complexity and diversity of genetic variations in the human genome. In addition to the physical mapping of different types of genetic variations, such as RFLPs in the 1980s,[3] tandem repeats in the 1990s,[4] and SNPs,[5,6] indels,[7] CNVs[8–10] and LOHs[2] after the new millennium, the data of their biological functional roles; for example, their effects on or associations with mRNA expression levels, alternative splicing processes and other

[1]Department of Epidemiology and Public Health, Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore and [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
Correspondence: CS Ku or Professor KS Chia, Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, National University of Singapore, 16 Medical Drive, Singapore 117597, Singapore.
E-mails: cmekcs@nus.edu.sg or ephcks@nus.edu.sg
Received 11 January 2010; revised 27 March 2010; accepted 11 April 2010; published online 20 May 2010

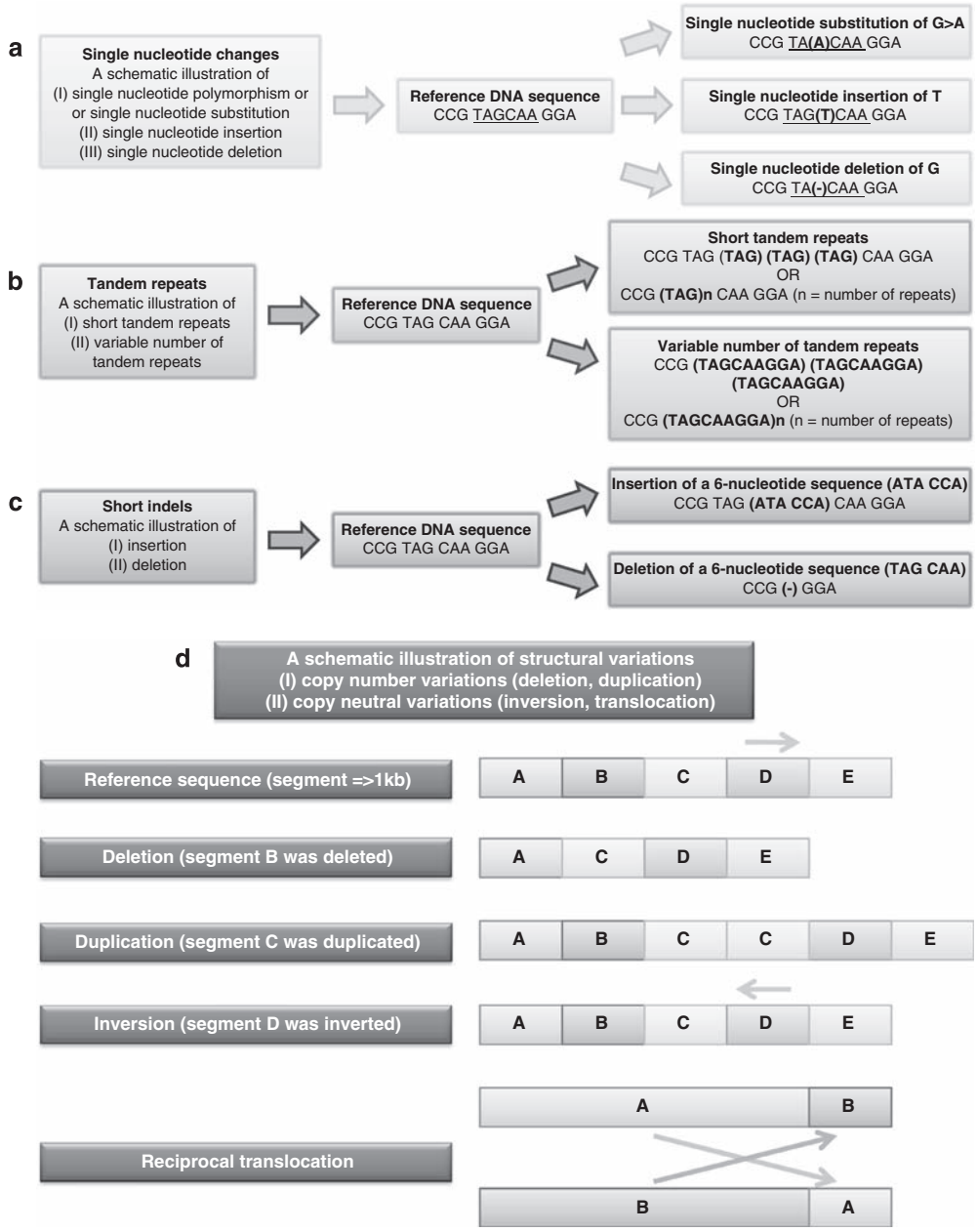


Figure 1 A schematic illustration of (**a**) single nucleotide changes; (**b**) tandem repeats; (**c**) short indels; (**d**) structural variations.

molecular and regulatory processes have also been accumulating.[11–14] Furthermore, these genetic variations were also found to be associated with various human diseases, including monogenic and complex diseases.[14–22]

Presently, research in genetic variations is drawing much attention and effort from the genetics community, as evident from the initiation of the 1000 Genomes Project, which has a major aim to construct the most detailed map of genetic variations in the human genome.[23,24] The non-SNP genetic variations certainly have the potential of becoming the next generation genetic markers in human genetic and disease gene mapping studies. The 'disease gene mapping' refers to mapping of genetic loci which may or may not contain genes that are associated with diseases. This review will focus on the discovery of different types of genetic variations and their use as genetic markers in disease gene mapping studies in the past, present and future.

CATEGORIES OF GENETIC VARIATIONS

There are issues and problems in categorizing genetic variations into distinct groups, and a clear consensus in defining genetic variations has not been achieved. As a result, the distinction for some of the genetic variations is rather vague at this time. Although SNPs are defined as single nucleotide substitutions, sometimes single nucleotide insertions or deletions also fall under this category (Figure 1a). In general, point mutations include both single nucleotide substitutions and single nucleotide indels, although they are only classified as such when their population frequencies are less than 1%. This is different from polymorphisms, terminology of which is reserved for those genetic variations with population frequencies higher than the arbitrary cutoff of 1% similar to SNPs.

Tandem repeats can be broadly divided into two classes: short tandem repeats (STRs) usually refer to tandem repeats in which the

sequence length is eight nucleotides or less, and longer tandem repeats are labeled as variable number tandem repeats (VNTRs; Figure 1b). They are also known as microsatellites and minisatellites, respectively. As such, it is apparent that the distinction between the two classes is solely based on the length of the repeated sequence, but it is only an arbitrary cutoff. The most common types of microsatellites are di-, tri- and tetra-nucleotide repeats. However, repeats of identical nucleotide of several bases or longer in the length; that is, consecutive identical nucleotides in the DNA sequence are known as homopolymer sequences; for example, GGGGG or AAAAA. Although the sequence in the tandem repeats is simple compared with other more complex DNA sequence changes or rearrangements, these simple sequences can be repeated from tens to hundreds of times, thus creating a high heterozygosity or allelic diversity.[25,26]

The boundary or distinction between CNVs and indels is even more obscure. In the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation/), deletions and duplications/insertions larger than 1 kb are classified as 'CNVs', whereas those between 100 bp to 1 kb are grouped as 'InDels'. As such, the remaining several hundred thousands of indels in the range of several nucleotides to tens of nucleotides, which were identified in the recent whole genome resequencing experiments, do not currently have their own category.[27–33] For example, Wang *et al.* (2008)[29] found ~140 000 indels within 1–3 bp in the Han Chinese YH genome, and ~400 000 indels defined from 1 to 16 bp were also detected in the African NA18507 genome by Bentley *et al.* (2008).[30] Perhaps a new category such as 'short indels' needs to be created to fit them in, and those indels between 100 bp to 1 kb should probably be renamed as 'intermediate indels' (Figures 1c and d). Similar to SNPs, common CNVs with population frequencies of 1% or higher are known as copy number polymorphisms. However, in some studies, CNVs that are detected in two or more individuals are also considered as copy number polymorphisms.[9]

However, apart from single nucleotide changes, such as SNPs, all the genetic variations can be broadly grouped under the umbrella of structural variations.[34] It is even more confusing when a variety of names are used to describe essentially the same genetic variation. For example, large-scale copy number variants and intermediate-sized variants have been used to describe CNVs before this terminology was introduced.[35] Some comparative genomic hybridization array-based studies used chromosomal gains and losses to indicate duplications and deletions, respectively.[36] Despite the various categories of genetic variations and terminologies that have been used, it is noteworthy that the definitions or sizes are rather arbitrary. Furthermore, classifications are without biological basis; that is, they are not classified by the mechanisms that mediated their occurrences. Instead, the classification is simply based on the patterns of DNA sequence changes and their sizes. As such, it is more important to describe the characteristics of the genetic variations that are being discovered and identified, rather than be concerned about their respective categories.

THE EVOLUTION OF GENETIC MARKERS IN DISEASE GENE MAPPING

Genetic variations in the human genome are useful as genetic markers for many applications in different areas, such as forensic investigations (for example, genetic or DNA fingerprinting), routine clinical tests (for example, human leucocyte antigen typing for hematopoietic stem cell or organ transplantation), prediction of drug responses or the tailoring of prescription doses (for example, genotyping tests for the SNPs in the thiopurine methyltransferase (*TPMT*) gene to predict patient responses to 6-mercaptopurine) and population genetics studies (for example, studies of human migration patterns).[37–40] Furthermore, they have also been widely used as genetic markers in disease gene mapping, such as family linkage and genetic association studies to identify the susceptibility loci or genes for monogenic and complex diseases.

Different genetic variations have different characteristics, and their applications are influenced by a number of factors. Tandem repeats such as minisatellites and microsatellites are highly variable or polymorphic in human populations, as such, they have higher allelic states and are more informative than the biallelic genetic markers, such as SNPs. Unlike SNPs in which a single nucleotide substitution will only give rise to two alleles, each repeat in minisatellites and microsatellites is considered as one allelic state. The genetic variations that occur in more than two allelic states are known as multiallelic markers. Owing to their inherent features, tandem repeats have been widely used in genetic fingerprinting and as the genetic markers in linkage studies to locate the chromosomal regions harboring the mutations or genes for monogenic or familial disorders, complex diseases and quantitative traits.[41–44] Although tandem repeats are more informative than SNPs at the individual marker level, their number is far less than the several million SNPs in the human genome. Thus, tandem repeats are not ideal genetic markers for applications that require high marker density or resolution, such as genome-wide association studies (GWASs), in which several hundred thousand of SNPs are needed. In GWAS, a large number of genetic markers are required spanning the whole genome, to achieve comprehensive coverage and adequate statistical power to detect unknown disease variants through linkage disequilibrium (LD).[45,46] In other words, the disease variants would not be detected if no markers in strong LD with them were genotyped.

Apart from the inherent characteristics of genetic variations such as their allelic diversity and abundance in the human genome, their applications are also influenced by technological developments. The rapid advances of high-throughput SNPs genotyping technologies have enabled the genotyping task of several hundreds of thousands to one million SNPs to be done efficiently on thousands of samples in GWAS. On the contrary, no high-throughput method was developed to assay microsatellites on a whole genome scale.[47–49] This technological development, together with their abundance in the human genome, have resulted in SNPs becoming the primary genetic markers used in more than 450 GWAS that have been published to date (A Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/26525384). In fact, almost all the GWAS have used the commercially available whole genome SNPs genotyping arrays from Illumina (San Diego, CA, USA), Affymetrix (Santa Clara, CA, USA).

In the past, researchers had relied solely on RFLPs and tandem repeats as the genetic markers in disease gene mapping studies. The RFLPs were used in linkage studies before the discovery of tandem repeats. Since the availability of the linkage map for microsatellites, RFLPs were mainly used as the genetic markers in candidate gene association studies, in which PCR–RFLP genotyping assay was commonly applied. However, microsatellites were widely used as the genetic markers in linkage studies.[41–44] These genetic variations have been used as the markers in human genetic studies for more than 20 years until the completion of the Human Genome Project [50] and the finding of millions of SNPs by the International SNP Map Working Group and other studies.[5,6] Thereafter, SNPs became the primary markers in genetic association studies, and also replaced microsatellites in some linkage studies.

Although SNPs have been studied in detail over the past decade, a comparable progress in the studies of other genetic variations, such as

indels, CNVs and LOHs has not been achieved. In fact, CNVs had only started gaining some attention from the genetics community when the finding of several hundreds of deletions and duplications was first reported in 2004.[51,52] Similarly, no large-scale attempt was made to identify indels until 2006, in which a study found several hundreds of thousands of indels in the human genome.[7] The commonness of LOHs or homozygosity regions in the genomes of outbred populations was also under appreciated until the first report appeared in 2006.[2] However, the richness of genetic variations in the human genome has recently been further corroborated by the several whole genome resequencing studies, revealing plenty of new SNPs, indels, CNVs and other structural variations.[27–33] The technological developments have facilitated and accelerated the process of identifying genetic variations, especially with the arrival of next generation sequencing technologies, which have made whole genome resequencing and the 1000 Genomes Project feasible.[53–55]

In recent years, many studies have been done to directly examine the associations of CNVs with complex diseases using SNP genotyping arrays. These studies have yielded some exciting results for several diseases, such as schizophrenia and autism.[56–58] Therefore, it further supports the use of CNVs as genetic markers to uncover new susceptibility loci for future disease association studies. Interestingly, genome-wide homozygosity mapping approaches have also been applied to dissect the genetic basis of complex diseases and have successfully identified a number of susceptibility loci for schizophrenia.[22] Conversely, short indels have not been directly interrogated in GWAS, but how much they can be tagged indirectly through LD by the SNPs in genotyping arrays is unclear. Unlike CNVs and homozygosity mapping, which can be studied by SNPs genotyping arrays, no high-throughput method has been designed and developed to investigate short indels on a genome-wide scale. Direct detection and interrogation of short indels requires sequencing-based methods as demonstrated in the whole genome resequencing studies. As a result they cannot be used effectively as genetic markers in GWAS at the time.

In the following sections, we will discuss the genetic variations and markers in the past (RFLPs and tandem repeats), present (SNPs) and future (CNVs, indels, inversions, translocations and LOHs). The use of 'past, present and future' genetic variations is only a 'time concept', to illustrate the time of their discoveries and the time when they are most commonly used as genetic markers. For example, RFLPs and tandem repeats were mainly discovered in 1980s and 1990s, so they are considered as the past genetic variations or markers, but this does not mean that they are totally obsolete nowadays or that they are no longer used in human genetic studies. However, although the commonness of CNVs, indels and LOHs in the human genome have already been reported several years ago, they are considered as future genetic variations or markers because they have yet to be 'intensively and completely' studied or discovered in the human genome. In addition, so far these newer genetic variations have not been widely used as markers in disease gene mapping.

PAST

Restriction fragment length polymorphisms

The RFLPs are single nucleotide substitutions that alter the cutting sites of restriction enzymes. They were one of the earliest genetic markers used in disease gene mapping. The genetic linkage map of RFLPs was constructed in the 1980s.[59] The use of RFLPs as genetic markers is based on their ability to create or eliminate the cutting sites of restriction enzymes to distinguish between two alleles. With the invention of the molecular technique PCR, alleles of RFLPs are usually determined by PCR-based methods, such as PCR–RFLP.

In PCR–RFLP assay, one set of probes or PCR primers (forward and reverse primers) are designed to amplify the DNA sequence that contains the RFLP. The PCR amplicons are then followed by restriction enzyme digestion and gel electrophoresis to separate the digestion products. As an example to illustrate the principle of the PCR–RFLP method, the PCR amplicons of G allele will be cut by the restriction enzyme but not for the C allele (a G>C substitution), assuming that there is only one cutting site in the PCR amplicon. Therefore, if all the PCR amplicons remain intact after restriction enzyme digestion (appearing as a single band in gel electrophoresis), this result shows the presence of two C alleles and the genotype is the homozygote CC. Conversely, all the PCR amplicons will be digested by the restriction enzyme for the homozygous GG genotype (two bands in gel electrophoresis for which the sizes are smaller than the PCR amplicon size), and a mixture of three bands suggests the presence of both alleles (Figure 2).

One of the major limitations of using RFLPs as genetic markers is that single nucleotide alterations do not necessarily alter the cutting sites of restriction enzymes. In other words, those single nucleotide substitutions that are not digested by restriction enzymes cannot be studied by PCR–RFLP method. As a result, their numbers are limited. Furthermore, PCR–RFLP is a tedious, laborious and low-throughput genotyping method. Nevertheless, PCR–RFLP has still been widely used in disease gene mapping studies at least before the arrival and feasibility of SNPs genotyping arrays or other higher throughput genotyping methods, such as MassARRAY iPLEX, Invader and SNPlex genotyping assays.[60–62] As RFLPs are single nucleotide substitutions, thus they are actually a subset of SNPs.

Tandem repeats

In addition to RFLPs, the earliest genetic markers also included tandem repeats. The more widespread distribution of microsatellites (>100 000) in the human genome and their higher allelic diversity than RFLPs have made them to be commonly used as the genetic markers in linkage studies for monogenic disorders and complex diseases. Similarly, microsatellite also out-performed VNTRs in terms of their numbers, where there are only a few thousand VNTRs in the human genome.[26] The availability of the genetic linkage map of microsatellites has resulted in the immense success of linkage studies in identifying genes for monogenic disorders.[4] In contrast, only limited success was achieved in dissecting the genetic basis of complex disease using linkage analysis. For complex diseases, the linkage regions identified were mostly irreproducible and inconsistent, and so far, only a handful disease associated genes, such as *CARD15*/*NOD2* (Crohn's disease), *PTPN22* (type-1 diabetes), *TCF7L2* (type-2 diabetes) and *STAT4* (rheumatoid arthritis and systematic lupus erythematosus), were identified through linkage and positional cloning strategies.[63–66]

The failure of linkage studies in interrogating the genetic basis of complex diseases is not due to the inappropriateness of the genetic markers (microsatellites) used to locate the genomic regions that harbor the disease genes, but is instead attributable to the study design. Linkage mapping is a powerful and effective approach to detect rare and highly penetrant mutations, and is best suited for diseases that segregate according to Mendelian inheritance. In contrast, complex diseases are characterized by genetic heterogeneity (multiple genetic variants with incomplete penetrance), and the phenotypes are consequences of complex interactions of genetic factors and environmental exposures.[67]

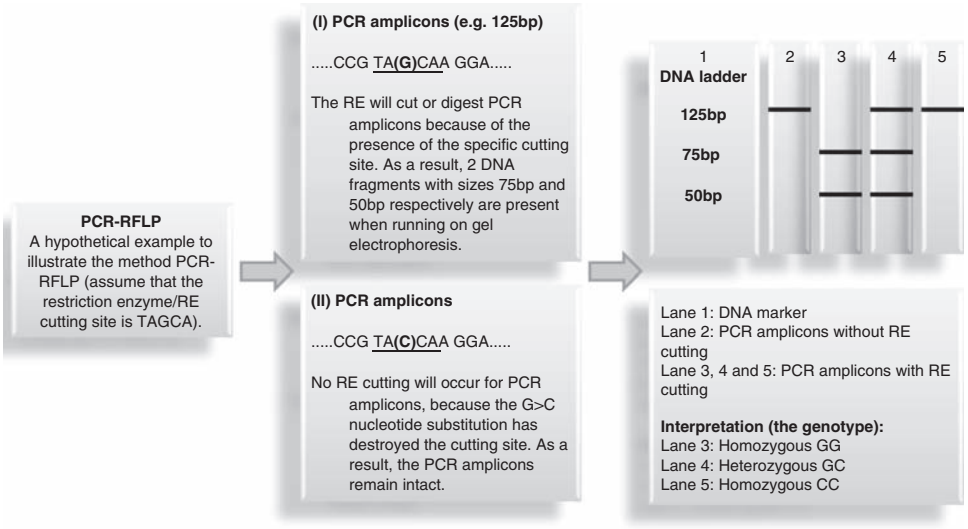


Figure 2 A schematic illustration for the method PCR–RFLP (restriction fragment length polymorphism).

The arrival of high-throughput SNP genotyping technologies and the ease of genotyping thousands of SNPs in a microarray have also replaced the use of microsatellites in some linkage studies.[68–71] In classical family linkage studies, a few hundred microsatellites are already sufficient to cover the whole genome. However, this number can be substituted by about 10 000 SNPs to provide a comparable or even greater amount of genetic information.[72,73] The need for a significantly larger number of SNPs is because of their lower heterozygosity as opposed to multiallelic genetic markers. Although microsatellite is more informative at the individual marker level, this can be superseded by a large number of SNPs.

Undoubtedly, microsatellites have been widely used in genome-wide linkage studies, but not in GWAS for complex diseases. Hitherto, there are only a few studies that have genotyped microsatellites in GWAS, and they have adopted a pooling strategy of DNA samples to reduce the amount of genotyping work.[74,75] This is mainly due to the need of genotyping a substantially larger number of microsatellites in GWAS ($\sim$20 000–30 000 markers) compared with linkage studies ($\sim$500 markers). The need for a larger number of microsatellites in GWAS is due to the weaker LD in unrelated individuals, as compared with family members in which there are only a limited number of recombination events. In addition, a larger sample size is also needed in GWAS to achieve adequate statistical power to detect genetic variants with modest effect sizes for complex diseases. Finally, there is a lack of high-throughput methods to assay microsatellites, and this is one of the major reasons that microsatellites have decreased in popularity in the SNP era. However, evidence is now increasing to support the potential functional roles of tandem repeats (tri-nucleotide repeats) and their variation could be associated with human complex diseases. Therefore, they should be reconsidered in the future genetic association.[16,76]

PRESENT

Single nucleotide polymorphisms

The completion of the Human Genome Project is a major scientific development in human genomics and biomedical sciences. The reference DNA sequence has provided the basis for studying genetic variations in the human genome among individuals in populations. While the Human Genome Project was about to be completed, genetic variations in particular SNPs were also being uncovered. In 2001, the International SNP Map Working Group identified 1.42 million SNPs in the human genome.[5] Currently, more than 17 million SNPs in human genome have been documented in the dbSNP. As a large number of SNPs has been reported, it is unavoidable that some of the entries in the database are actually errors or artifacts rather than 'genuine SNPs'. In fact, a false positive rate of 15–17% was estimated for dbSNP.[77] Therefore, large scale validation in population-based studies would be necessary and important to authenticate them. To bridge this gap of information, the International HapMap Project was conceived in 2003 with the aim to validate several million SNPs in the dbSNP, to obtain the SNP and genotype frequencies information, as well as to study their correlation or LD patterns in populations of European, Asian and African ancestry. These populations are the US Utah population with Northern and Western European ancestry (CEU), Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT) and Yoruba from Ibadan, Nigeria (YRI).[78]

In general, a SNP is defined as a single nucleotide substitution at one particular locus in the DNA sequence and this mutational event generates two alleles. To distinguish this from a point mutation, the frequency of the minor allele of a SNP has to be at least 1% in any population. Common SNPs are usually defined as those with minor allele frequency $>5\%$ and approximately 7 million of the SNPs in the human genome are common.[79] Therefore, for single nucleotide substitutions, where their population frequencies are yet to be determined, strictly, they should be labeled as single nucleotide variations (SNVs) to minimize confusion.[77] As a substantial fraction of entries in the dbSNP has not been validated in population-based studies, one has to bear in mind that not all the entries in the dbSNP are necessarily SNPs, as the name of database implies. As such, the several hundred thousand 'new SNPs' identified by whole genome resequencing studies[27–33] should probably be considered as 'new SNVs' instead, until their population frequency information is available (Figure 3a). The distinction between SNPs and SNVs should be emphasized to avoid misleading.

Single nucleotide polymorphisms are the most abundant type of genetic variation in the human genome in terms of their number. They occur at an interval of about one SNP in every kilobase of DNA sequence throughout the genome when the DNA sequences of any two

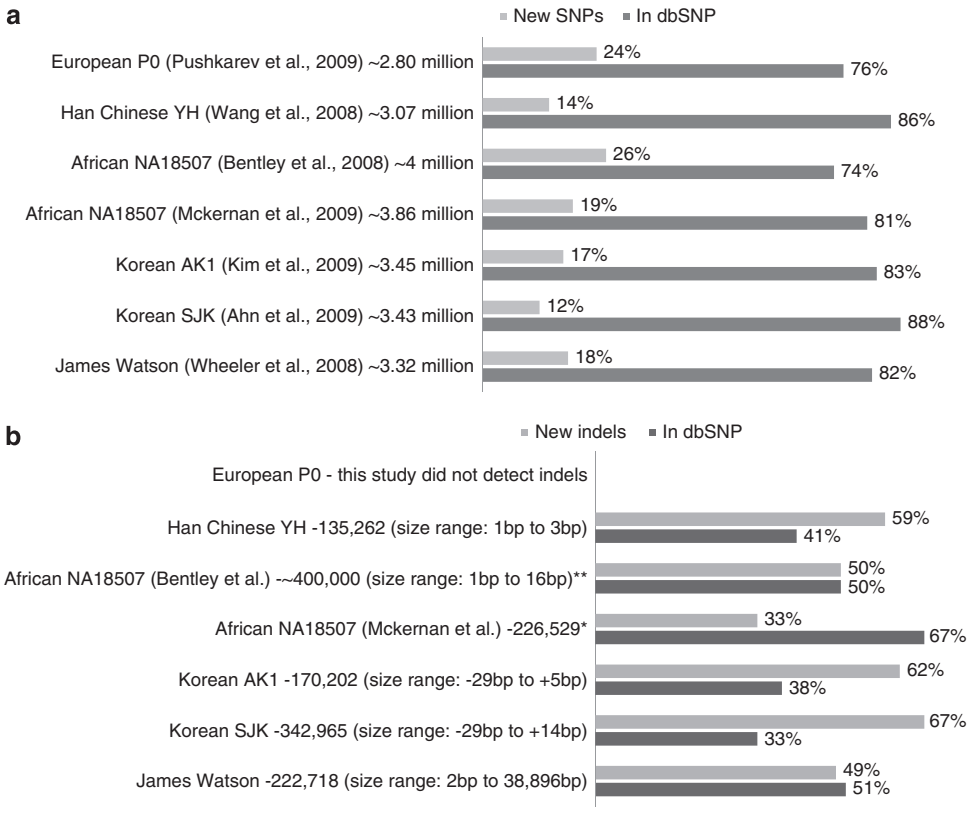


Figure 3 (**a**) The proportion of new SNPs identified in whole genome resequencing studies. (**b**) The proportion of new indels identified in whole genome resequencing studies. *89,679 insertions up to 3bp, 124,024 deletions up to 11bp, 12,826 larger indels. 67% of small indels in dbSNP (i.e. insertions up to 3bp and deletions up to 11bp). **Approximately 0.4 million indels were identified and it was reported that about half of the indels are corroborated by entries in dbSNP

individuals are compared. This is approximately equivalent to 3 million SNPs being carried by each individual genome. Therefore, the DNA sequence of any two genomes is estimated to be about 99.9% identical, and the 0.1% genetic variations that are mainly comprised of SNPs, are believed to be responsible for the phenotypic differences, such as physical traits (for example, height, hair and eye colors), disease susceptibility and drug responses, among individuals in populations. However, the finding of thousands of CNVs that collectively encompass hundreds of megabases of the genome[8–10] and the numerous short indels that are identified by whole genome resequencing studies[27–33] have thrown doubts to the estimation of '99.9% similarity'. The DNA sequences of individuals within and between populations are genetically more diverse and varied than previously thought.

Most of the SNPs are predicted to be neutral without functional effects and due to their abundance in the human genome; SNPs have become useful genetic markers in GWAS compared with other genetic variants such as microsatellites. In addition to the finding of a myriad of SNPs, some early reports have also documented the correlation patterns among the SNPs in parts of the human genome.[80–82] However, no large-scale effort was undertaken to study the LD patterns in the whole genome until the initiation of the International HapMap Project. So far, a total of >3 million SNPs have been genotyped and validated in the Phase I and Phase II of the project.[83,84]

The huge number of SNPs has also created a formidable task in genotyping because it is not technically feasible and cost effective to genotype several million of SNPs in a GWAS even with the latest genotyping technologies. Fortunately, SNPs are not completely independent of each other; instead they are correlated, as has been demonstrated by the International HapMap project. The existence of LD significantly reduces the number of SNPs that needs to be genotyped in a GWAS. The indirect association approach of GWAS is dependent on surrogate markers to locate disease variants through LD. As shown in the International HapMap Project and other published data, about half a million SNPs are already adequate to capture most of the SNPs that have been genotyped in the HapMap Project. However, the genome coverage of commercially genotyping arrays is population dependent. For example, Illumina HumanHap550 Beadchip, which contained $\sim$550 000 tagging SNPs, achieved genome coverage of 87 and 83% in CEU and CHB+JTP populations, respectively, but it was only 50% in YRI.[85–87]

The International HapMap Project has created a useful and valuable resource for GWAS. Furthermore, the availability of HapMap data has also driven the rapid developments in genotyping arrays, in which the data are used to guide the tagging SNPs selection. As the Phase I HapMap was completed in 2005, a number of genotyping arrays has been designed and introduced into the market, and the newer arrays have significantly improved in genome coverage and are also designed for CNVs detection, such as the Illumina Human 1 M Beadchip and Affymetrix 6.0 SNP Arrays.[49] Hence, the International HapMap Project was a key and essential component in making the GWAS a feasible approach.

Around the turn of millennium, there were also some intense debates about the genetic architecture of complex diseases.[88] It was polarized into two opposing models: the common-disease common-variant (CD/CV) versus multiple rare variant or common-disease

rare-variant hypothesis.[89] However, the CD/CV model formed the basis of the International HapMap Project; it was clearly shown in the Phase I HapMap, in which common SNPs have become the main focus. Over one million SNPs with minor allele frequency $>5\%$ were genotyped in 270 DNA samples collected from the four populations. Even in the Phase II HapMap, common SNPs remained as the focus; however, SNPs within minor allele frequency of 1–5% were also chosen to be genotyped.[83,84] As the HapMap data was used to develop commercial genotyping arrays, the SNP selection has been largely influenced by the CD/CV hypothesis. Therefore, the current GWAS are mainly interrogating the association of common SNPs with various complex diseases and traits.

The reason that the CD/CV model trumped the opposing model was also due to the technologies that were available at that time. Sanger dideoxynucleotide sequencing did not allow the survey of rarer SNPs or point mutations in the whole genome to be carried out efficiently. With the arrival of next generation sequencing technologies, whole genome sequencing is practical now, but still prohibitively expensive to be done in a large sample set for association studies. Instead, targeted sequencing of certain regions identified by GWAS, as well as exomes, is more feasible at the moment.[90,91] This approach has been advocated by genetics community as a temporary alternative to searching for rarer SNPs before we reach the goal of 1000 dollars per genome, enabling thousands of cases and controls to be sequenced. In contrast, the convenient high-throughput genotyping platforms have enabled an efficient interrogation of several hundred thousand to one million SNPs directly throughout the genome, which eventually captured almost all the SNPs in the International HapMap Project indirectly. Furthermore, it is more affordable to genotype (rather than to sequence) the whole genome of several thousand cases and controls for a statistically powerful association study.

FUTURE

Copy number variations

The term CNV was first introduced in 2006, and it is generally defined as additions or deletions in the number of copies of a particular segment of DNA (larger than 1 kb in length) when compared with a reference genome sequence.[35] The commonness of CNVs in the human genome was under-appreciated until the first reports in 2004. The findings have also stimulated a lot of enthusiasm and interest in the research of genetic diversity in the human populations and resulted in a series of effort to detect CNVs in different populations. The number of publications of CNVs studies has indeed increased greatly over the past few years.

In contrast to SNPs that have already been relatively well-cataloged in the dbSNP, and well-studied by the International HapMap Project, a lot more remains unclear for other types of genetic variations and to what extent they are present in the human genome. Although the ubiquity of CNVs in the human genome was reported several years ago, and many more have since been found, most of the studies used array-based detection methods that have relatively poor sensitivity compared with sequencing-based approaches.[8,9,36,92–95] These array-based methods include bacterial artificial chromosome clones and oligonucleotides comparative genomic hybridization arrays and SNPs genotyping arrays. These methods are not sensitive enough to detect smaller sizes of CNVs that are less than 50 kb in size due to the limitations in array density or resolution.[96] However, the number of smaller CNVs is estimated to be more abundant than the larger CNVs in the human genome.[97]

The poor sensitivity of array-based methods becomes apparent when their results are to be compared with the sequencing studies. The number of CNVs found in most of the array-based studies was in the range of tens to several hundred per genome on average, which is several fold lesser than the numbers that were reported in the whole genome resequencing studies. In each of the studies, several thousands of CNVs have been found;[29–32] for example, Ahn *et al.* identified 2920 deletions and 963 insertions in the Korean SJK genome. Although the improvements in SNPs density and inclusion of copy number probes in newer genotyping arrays, such as Illumina Human 1 M Beadchip and Affymetrix 6.0 SNP Arrays, have undoubtedly increased the performance of array-based methods to detect CNVs, the methods overall still suffer from poor sensitivity to detect CNVs smaller than 5–10 kb.[9,98] This was again clearly shown in the findings from whole genome resequencing studies. For example, a total of 2682 structural variations (dominated by deletions) were detected in the Han Chinese YH genome with a median length of about 0.5 kb.[29] In contrast, the median length found by array-based methods was in the range of tens to hundreds of kilobases depending on the resolution of the arrays. This indicates that sequencing-based methods have much higher sensitivity to detect smaller CNVs. This also suggests that the overall larger number of CNVs found in whole genome resequencing studies was attributed to the better sensitivity in detecting more CNVs of smaller sizes. In addition, it is worthwhile noting that if the arbitrary cutoff of 1 kb is applied here, at least half of the reported CNVs by Wang *et al.*[29] should be labeled as indels. This further illustrates the problems in classifying CNVs and indels into distinct categories.

Indels

In addition to CNVs, the several whole genome resequencing studies also identified hundreds of thousands of short indels.[27–32] The numbers reported in each study are not directly comparable, because the analyses, detection methods and criteria used are different between the studies. For example, for the two Korean genomes, the number of indels found in one study is twice another one. Ahn *et al.*[32] identified 342 965 indels within a size range of -29 to $+14$ bp, whereas Kim *et al.*[31] only found 170 202 indels within -29 to $+5$bp. Collectively these studies have uncovered plenty of short indels in the human genome. Moreover, the number of indels found is likely to represent only a fraction of the total number of indels in the human genome, because a rather narrow size range was defined in each of the studies. In summary, the several whole genome resequencing studies have further revealed the richness of genetic variations in the human genome and their numbers are more abundant than previously expected.

It is estimated that there are 1.6–2.5 million indels in human populations. However, no large-scale attempt was made to identify indels until 2006, in which a study identified 415 436 indels with about equal numbers of insertions and deletions.[7] The sizes of these indels ranged from 1 bp to $\sim$10 kb (which span the '1 kb boundary'), thus suggesting that the dataset is actually a mixture of indels and CNVs. In addition, the study also found over 148 000 indels located within known genes and several thousands of them are found in the promoter regions and exons of genes. This means that these indels could potentially alter gene expression levels or affect protein structure or function. Similarly in the whole genome resequencing studies, several hundreds of indels were also found to overlap with coding sequences.[28,31] Despite some differences in the number of indels found in each study that overlapped with coding sequences, these studies have provided evidence to support their putative functional roles and also underscores the importance of investigating them in disease association studies. The discovery effort for indels is not keeping

pace with that of SNPs, as indels have not been well cataloged in the dbSNP. This can be clearly shown from the proportion of new indels found in the whole genome resequencing studies; about 50% or more of the identified indels are not in dbSNP. In contrast, less than 30% of the SNPs identified in the studies are new (Figures 3a and b).

Though findings from whole genome resequencing studies have broadened our knowledge in human genetic variation, all of them only sequenced one individual genome, rendering them unable to investigate the population genetics of the identified genetic variants, such as frequencies and LD patterns. This piece of information is crucial and would be needed for future disease association studies. Moving towards this goal, and to accelerate the process of discovery of various genetic variations in the human genome, the 1000 Genomes Project was conceived and initiated in 2008. This project is currently on-going and the aim is to eventually sequence at least 1000 individual genomes from different populations worldwide. The ultimate goal is to build a useful resource of human genetic variations for future disease association studies. The availability of these resources and the genetic variations maps will certainly drive the technological development of new microarrays or other high-throughput methods to capture the non-SNP genetic variations in the near future, and it will bring another revolution to the genetic studies of complex diseases.

Copy neutral variations—inversions and translocations

The discovery of CNVs in the human genome of healthy populations has advanced rapidly over the last few years. However, an equivalent progress has not been seen for the detection of copy neutral variations; this is largely due to the lack of a powerful and efficient method for a genome-wide discovery of inversions and translocations. Unlike CNVs that can be studied by microarrays, the detection of copy neutral variations usually requires sequencing-based methods, and the high-throughput sequencing technologies that have only recently been made more accessible. In addition, inversions and translocations are technically more difficult to detect. A relatively slower progress in the studies of copy neutral variations is evident from the data entries recorded in the DGV, in which more than 29 000 CNVs and nearly 20 000 indels have been reported in the database, whereas less than a thousand inversions have been found, and no data is available for translocations in the DGV at the moment. However, one should be cautious with this interpretation because the numbers are not proportions. As the total number of CNVs, indels and inversions in the human genome is still unknown, therefore, the proportions of these genetic variations that have been discovered are also unknown. The data in the DGV are so far derived from the results of 35 studies using array-based and sequencing-based detection methods, and other approaches. In fact, more than this number of studies have been performed and published for CNVs detection in various populations; but not all their results have been cataloged in the DGV. As such, it is apparent that the entries in the database are still far from complete.

Most of the CNV data were generated by array-based methods (comparative genomic hybridization and SNP arrays), in which the signal intensity information is used to detect deletions and duplications, which relied on differences in signal intensities. As a result, these methods are unsuitable for detecting inversions and translocations (also known as balanced chromosomal rearrangements) because they do not lead to gain or loss of chromosomal or DNA segments. Rather, several different strategies and approaches have been taken to try to identify inversions in the human genome. For example, Feuk *et al.*[99] discovered regions that are inverted between the chimpanzee and human genomes by performing comparative analysis of their DNA sequence assemblies. In the study, they identified about 1600 putative regions of inverted orientation in the genomes that covered >150 megabases of DNA sequence. The inverted regions are distributed throughout the genomes and span the sizes from 23 bp to 62 Mb in length. A number of inverted regions were also selected to be validated by using PCR and fluorescence *in situ* hybridization, and out of the 23 experimentally validated inversion regions, 3 of them were found to be polymorphic (>1%) in a panel of human samples, and were known as inversion polymorphisms.

However, a statistical method has also been developed to identify large inversion polymorphisms using high-density SNP genotyping data in which it is based on unusual LD patterns. The method was developed to detect chromosomal regions that are inverted in a majority of the chromosomes in a population with respect to the reference human genome sequence. Although this method has worked using the International HapMap Project data to detect inversion polymorphisms, it has not been widely used by other studies. In any case, this study was able to identify 176 inversions ranging from 200 kb to several megabases in length using the Phase I data. However, their results were not placed in the DGV.[100] This, together with the study by Feuk *et al.* (2005)[99], also provided some supporting evidence that a considerable portion of their detected inversions were flanked by highly homologous repeats or segmental duplications. This suggests that segmental duplications could be the favorite spots mediating the chromosomal rearrangements that generate inversions.

The breakthrough in the discovery of inversions was credited to the development of a sequencing-based method known as paired-end mapping, and the concurrent advances in next generation sequencing technologies. The paired-end mapping method also contributed greatly to the mapping of CNVs in the human genome. In the paired-end mapping method, both ends of the DNA fragments with known sizes would be sequenced and then aligned to the human reference genome. The principle of the paired-end mapping to detect various structural variations is simple in theory; it is based on the discordances in size or orientation of the DNA fragments that are to be aligned to the reference genome. When both ends of the DNA fragments that map to the reference genome show discordances in terms of size, this is an indication for deletion and insertion, whereas discordances in orientation suggests the presence of inversion.[101]

The power of this method to detect inversions was first demonstrated in the study by Tuzun *et al.*[102] by sequencing the fosmid paired-end sequences. The study successfully identified 56 inversion breakpoints. The same strategy of fosmid clones sequencing was also used by Kidd *et al.*[103] to detect structural variations in eight individual genomes, and a total of 224 inversions were also identified. However, this study is only the preliminary phase of a larger project that will eventually construct and sequence the fosmid clone libraries (~40 kb inserts) prepared from the genomic DNA of 48 unrelated females, and bacterial artificial chromosome clone libraries (~150 kb inserts) from 14 unrelated males in the International HapMap Project.[104] Therefore, more inversions are expected to be discovered when the project is finished. The fosmid paired-end sequencing work of these studies was completed by traditional Sanger sequencing methods.

The first proof-of-concept study using next generation sequencing technologies in paired-end mapping to detect structural variations was published in 2007.[105] In the study, libraries of 3-kb fragments for two female samples from the International HapMap Project were prepared and sequenced by Roche 454 sequencing, and they found 1297 structural variations, including 122 inversions. Using the same approach, hundreds of inversions were also uncovered by whole genome resequencing studies; for example, 91 and 415 inversions

were detected in the African NA18507 genome and Korean SJK genome, respectively.[32,106] Although the progress in the discovery of inversions is moving at a slower pace than CNVs, there is already evidence to support their roles in human diseases.[107,108]

Loss of heterozygosity and homozygosity

Copy neutral LOH defines a continuous stretch of DNA sequence without heterozygosity. It is different from a single copy deletion which could also lead to the absence of heterozygosity. More specifically, extended homozygosity is essentially copy neutral LOH, but it encompasses a large region of at least 1 Mb. Again, the distinction between the two categories is solely based on the length of DNA sequence without heterozygosity. Currently, there is no consensus on the definition of extended homozygosity. Previous studies have focused on homozygosity regions larger than 1 Mb, so the true level of homozygosity in the human genome could be underestimated.[2,109]

The information regarding the extent of LOHs in the human genome is even less compared with indels and CNVs, but their potential impact on complex diseases could also be as much as other genetic variations. Although the biomedical significance of regions of homozygosity to complex diseases remains largely unexplored, some schizophrenia studies have already shown significant differences in homozygosity regions between cases and controls in a genome-wide study.[22] More importantly is that the study has demonstrated the feasibility of using the homozygosity mapping approach to identify susceptibility loci and genes for complex diseases. This also highlights the need to further investigate and catalog the extent of LOH and homozygosity in the human genome. Similar to other genetic variations, LOHs definitely have the potential of being the genetic markers in future GWAS. Although homozygosity mapping has not been widely applied for most of the complex diseases, this approach is commonly used to interrogate the genetic basis of cancers to identify cancer-associated genes.[110,111]

The ubiquity of homozygosity in the genomes of outbred populations has not been well documented. Previously, only a few studies reported an abundance of homozygosity in the human genome with frequent occurrence in genomic regions with extensive LD and low recombination rates.[2,109] Three widely discussed possibilities that led to the commonness of homozygosity are parental consanguinity, uniparental disomy and autozygosity. One previous study had demonstrated that the number of homozygosity regions increased markedly in the offspring of consanguineous marriages.[112] However, this is unlikely in outbred populations in which parental consanguinity is rare.

Uniparental disomy can be divided into two types: uniparental isodisomy and uniparental heterodisomy. Only the former situation can cause homozygosity as the child inherits two identical copies of a chromosome segment from only one parent.[113] This is also an unlikely explanation for the abundant homozygosity given that uniparental disomies are rare genetic abnormalities that can cause severe and rare genomic disorders, such as Prader–Willi Syndrome and Angelman Syndrome. This assumption is further supported by previous research that found extended homozygosity to be generally not due to genetic abnormalities.[114] Using this reductionist approach, autozygosity seems to be the most likely process responsible for the commonness of homozygosity in the human genome. Autozygosity is a situation in which common ancestral haplotypes are inherited from both parents. Hence, extended homozygosity seems likely to have occurred as a result of common haplotypes, present in high frequencies in the population, which are passed on by chance from both parents to the child. This is further supported by previous findings of no excess apparent deviation from Mendelian transmission in extended homozygosity.[109,114]

THE FUTURE GENETIC VARIATIONS MAP

The significance of the 1000 Genomes Project for future disease association studies is tremendous. Although SNPs have been widely used as the genetic markers in GWAS to search for disease variants, evidence has started accumulating to suggest that (common) SNPs alone are unlikely to account for all the heritable risk of complex diseases. Concurrently, the amount of data showing the associations of CNVs with complex diseases has been growing.[19–21] Similarly, the importance of rare variants in complex diseases is also being recognized.[56,90,115,116] This implies that future disease association studies need to interrogate non-SNP and rare genetic variations as well, and for this to be feasible, a detailed catalog of human genetic variations is a prerequisite. Common SNPs are well documented in the dbSNP, but rarer SNPs (or lower frequency SNPs) are still under-represented in the database and the information of indels and structural variations is far from complete.

Unlike the whole genome resequencing studies of individual genomes, the 1000 Genomes Project is a large scale population-based sequencing study that enables studies of the population properties of genetic variations and their LD patterns. This information will be required to design next generation genotyping arrays to select surrogate markers that are not only able to tag for SNPs, but also to efficiently to capture indels and CNVs as well. This development will certainly widen the scope of genetic variations interrogated in GWAS. In fact, data have shown that CNVs could be tagged by SNPs through LD,[9,10,117] but a detailed and in-depth investigation of their LD patterns can only be done when most of the SNPs, indels, CNVs and other genetic variations have been identified. In-depth studies of LD among different genetic variations is important, as the finding of the 20-kb deletion located upstream of the *IRGM* gene for Crohn's disease has demonstrated the efficiency of using SNPs as surrogate markers to identify non-SNP genetic variants.[118] Other examples include the finding of a 45-kb deletion that is in perfect LD with BMI-associated SNPs in NEGR1.[119]

It is less likely that the number of indels and CNVs will reach several millions similar to the SNPs, but the total number of nucleotides encompassed by these genetic variations has already far exceeded that of the SNPs. Given their abundance in the human genome as found by the whole genome resequencing studies, their total nucleotide composition and functional impact on gene expression levels,[11,120,121] they could potentially account for some or even a substantial portion of the inherited risk of complex diseases.

A comprehensive interrogation of genetic variations is essential because GWAS is an indirect approach to identify disease variants; therefore, its success is dependent on whether surrogate markers that are in strong LD with the disease variants are included in the studies. The LD information between SNPs, indels, CNVs and other genetic variations is valuable because it is more efficient to interrogate or capture indels and CNVs through LD by genotyping a number of SNPs, rather than by locating the probes within the copy number variable regions and detecting them through signal intensity differences. If the number or fraction of 'untaggable' indels and CNVs is considerable, then other high-throughput methods or microarrays can be developed to complement the content of next generation SNPs genotyping arrays. Besides driving the development of more efficient genotyping arrays to interrogate SNPs and non-SNP genetic variations, the data from the 1000 Genomes Project will also accelerate the fine mapping work in the regions identified by GWAS and improve

the imputation powers because a much more complete reference set of genetic variations will be available for imputing.

THE CURRENT STATUS OF GWAS

Genome-wide association study is a comprehensive and biologically agnostic approach to searching for unknown disease variants, and as demonstrated in more than 450 studies, this strategy has been very successful in identifying new genetic loci for various human complex traits. Most of the genes and loci that have been identified are not previously thought to be associated with their respective diseases.[122–125] More importantly, the GWAS findings have also provided new insights into the molecular pathways of complex diseases even when most of the disease causative variants remain to be discerned from the neighboring correlated markers. For example, the three new genes that have been linked to Crohn's disease: *IL23R*, *ATG16L1* and *IRGM* have highlighted the importance of interleukin-23 receptor and autophagy pathways underlying the pathophysiology of this chronic inflammatory bowel disease.[126,127] Notably, GWAS have been making some significant advances in our understanding and knowledge of the genetic basis of human complex diseases compared with the pre-GWAS approaches (that is, the candidate gene association and linkage studies).

Most of the risk alleles that have been identified by GWAS are common (allele frequency $>5\%$) and confer small effect sizes (odds ratio <1.5).[17,18] However, this observation is not really reflecting the true allelic frequency spectrum of complex diseases. This is because for any given sample size, association studies have higher statistical power to find associations with common SNPs. The other reason is that the rarer SNPs (allele frequency $<5\%$) are not well-covered either directly or indirectly through LD by the markers in Illumina and Affymetrix genotyping arrays, so they remain unexplored for disease association. The design of GWAS and SNPs selection in commercial genotyping arrays have been largely driven by the CD/CV hypothesis.

Due to their small effect sizes, collectively the identified risk alleles only explain a small portion of the total inherited risk for the diseases. For example, all the type-2 diabetes risk alleles that are identified by GWAS cumulatively only account for $\sim 5\%$ of the heritability, and similarly for other diseases, only a small proportion of the heritability was accounted for.[128] The unexplained or missing heritability has been a major concern in the field, leading to the skepticism of the promise of GWAS to fully decipher the genetic basis of complex diseases. Nevertheless, it is noteworthy that GWAS have only interrogated a fraction of the total genetic variations in the human genome.

The genetic architecture of complex diseases remains elusive; it is unclear how much each type of genetic variation contributes to inherited risk and the relative proportion of rare versus common variants. If non-SNP genetic variants or rarer SNPs constitute most of the genetic component of complex diseases, then GWAS using the current genotyping arrays would be likely to miss them, simply because they are not covered directly by the genotyping arrays. How much they can be tagged through LD by the markers on the arrays still needs further investigation. Regardless, it is important to continue investigating other genetic variations to discover additional disease associated variants to explain the heritability.

INADEQUATE COVERAGE OF GENETIC VARIATIONS IN GWAS

All the GWAS rely heavily on the commercial genotyping arrays from Illumina and Affymetrix to comprehensively genotype several hundred-thousand of common SNPs. These genotyping arrays have near complete coverage of the >3 million SNPs genotyped by the International HapMap Project in CEU and CHB+JPT populations.[85–87] The HapMap Project SNPs are either genotyped directly or tagged indirectly through LD with one or more SNPs on the arrays. Nevertheless, the HapMap SNPs are only a subset of the entire collection in the dbSNP, and currently there are more than 10 million SNPs cataloged in the database. More than half of the SNPs in dbSNP have not been studied for association with complex diseases directly and the number of these SNPs that are covered indirectly through LD by the genotyping arrays is unclear. It is noteworthy that the current GWAS only investigate a portion of the SNPs and the non-SNP genetic variations are likely not well studied for disease associations.

Furthermore, SNPs are not the only type of genetic variation in the human genome. Although the roles of non-SNP genetic variations in disease susceptibility remain largely unexplored, associations of CNVs with complex diseases such as schizophrenia, autism, autoimmune disorders, HIV infection and cancers have already been established from both candidate gene and genome-wide approaches.[56,115,129–132] The amount of evidence is expected to increase in the near future, when we have a better understanding of the characteristics of non-SNP genetic variations and a more comprehensive map of them constructed upon the completion of 1000 Genomes Project, and when more efficient and accurate methods are available to detect and study them. One major limitation of the current GWAS using the commercial genotyping arrays is that it covers only a portion of the total genetic variations, thus a substantial false negative rate is likely due to incomplete interrogation of all the genetic variations for disease association. For future studies, the focus should be directed on studying other genetic variations that have not yet been interrogated by the GWAS, such as tandem repeats, indels, inversions and CNVs, although it is highly dependent on the development of the technologies and methods of detection and analysis.

It is also obvious from the results of GWAS that the common SNPs are unable to account for the total inherited risk of a complex disease. However, it is not clear how much heritability can be attributed to rarer SNPs (<1–5%) at the time. Rarer SNPs are not well-covered by the GWAS or the genotyping arrays, as a result, they have not been intensively studied for disease association. Fortunately, the current genotyping arrays seem to work fine for detecting rare CNVs for diseases.[56,115] The evidence linking complex diseases and traits to multiple rare variants has also been growing; for example, for schizophrenia,[56,115] high-density lipoprotein cholesterol level[133,134] and type-1 diabetes.[90] This implies that the rare variants (both SNP and non-SNP) should not be neglected in future studies. Sequencing approaches will improve their detection, and consequently offer a better understanding of the genetic architecture of complex diseases. The advances in sequencing technologies enable researchers to study a wider spectrum of genetic variants compared with genotyping methods.

CONCLUSIONS

The ultimate goal of GWAS is to correlate the genotype with disease phenotype, and to identify all the genetic variations that are associated with the diseases. To achieve this, most of the genetic variations in the human genome have to be first identified. It is essential to identify and validate all the genetic variations in the human genome in population-based studies, and catalog them properly in databases, so they can be used as the genetic markers for future disease association studies. Currently, we are moving towards these goals with the on-going 1000 Genomes Project, and only with the availability of a very detailed and near complete map of all genetic variations will it be feasible to perform a truly comprehensive search for the disease causing variants throughout the human genome.

1 Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. *Nat. Rev. Genet.* **10,** 551–564 (2009).
2 Gibson, J., Morton, N. E. & Collins, A. Extended tracts of homozygosity in outbred human populations. *Hum. Mol. Genet.* **15,** 789–795 (2006).
3 Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32,** 314–331 (1980).
4 Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P. *et al.* A second-generation linkage map of the human genome. *Nature* **359,** 794–801 (1992).
5 Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
6 Haga, H., Yamada, R., Ohnishi, Y., Nakamura, Y. & Tanaka, T. Gene-based SNP discovery as part of the Japanese Millennium Genome Project: identification of 190 562 genetic variations in the human genome. Single-nucleotide polymorphism. *J. Hum. Genet.* **47,** 605–610 (2002).
7 Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16,** 1182–1190 (2006).
8 Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
9 McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.* **40,** 1166–1174 (2008).
10 Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).
11 Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315,** 848–853 (2007).
12 Gilad, Y., Rifkin, S. A. & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24,** 408–415 (2008).
13 Fraser, H. B. & Xie, X. Common polymorphic transcript variation in human disease. *Genome Res.* **19,** 567–575 (2009).
14 Usdin, K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.* **18,** 1011–1019 (2008).
15 Haberman, Y., Amariglio, N., Rechavi, G. & Eisenberg, E. Trinucleotide repeats are prevalent among cancer-related genes. *Trends Genet.* **24,** 14–18 (2008).
16 Hannan, A. J. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet.* **26,** 59–65 (2010).
17 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).
18 Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106,** 9362–9367 (2009).
19 Wain, L. V., Armour, J. A. & Tobin, M. D. Genomic copy number variation, human health, and disease. *Lancet* **374,** 340–350 (2009).
20 Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61,** 437–455 (2010).
21 Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10,** 451–481 (2009).
22 Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M. *et al.* Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc. Natl. Acad. Sci. USA* **104,** 19942–19947 (2007).
23 Kaiser, J. A plan to capture human diversity in 1000 genomes. *Science* **319,** 395 (2008).
24 Kuehn, B. M. 1000 Genomes Project promises closer look at variation in human genome. *JAMA* **300,** 2715 (2008).
25 Schlötterer, C. The evolution of molecular markers—just a matter of fashion? *Nat. Rev. Genet.* **5,** 63–69 (2004).
26 Nakamura, Y. DNA variations in human and medical genetics: 25 years of my experience. *J. Hum. Genet.* **54,** 1–8 (2009).
27 Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
28 Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
29 Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
30 Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
31 Kim, J. I., Ju, Y. S., Park, H., Kim, S., Lee, S., Yi, J. H. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460,** 1011–1015 (2009).
32 Ahn, S. M., Kim, T. H., Lee, S., Kim, D., Ghang, H., Kim, D. S. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19,** 1622–1629 (2009).
33 Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27,** 847–852 (2009).
34 Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7,** 85–97 (2006).
35 Freeman, J. L., Perry, G. H., Feuk, L., Redon, R., McCarroll, S. A., Altshuler, D. M. *et al.* Copy number variation: new insights in genome diversity. *Genome Res.* **16,** 949–961 (2006).
36 de Stahl, T. D., Sandgren, J., Piotrowski, A., Nord, H., Andersson, R., Menzel, U. *et al.* Profiling of copy number variations (CNVs) in healthy individuals from three ethnic groups using a human genome 32 K BAC-clone-based array. *Hum. Mutat.* **29,** 398–408 (2008).
37 Tamaki, K. & Jeffreys, A. J. Human tandem repeat sequences in forensic DNA typing. *Leg. Med. (Tokyo)* **7,** 244–250 (2005).
38 Petersdorf, E. W. HLA matching in allogeneic stem cell transplantation. *Curr. Opin. Hematol.* **11,** 386–391 (2004).
39 Karas-Kuzelicki, N. & Mlinaric-Rascan, I. Individualization of thiopurine therapy: thiopurine *S*-methyltransferase and beyond. *Pharmacogenomics* **10,** 1309–1322 (2009).
40 HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. *Science* **326,** 1541–1545 (2009).
41 Feng, B. J., Huang, W., Shugart, Y. Y., Lee, M. K., Zhang, F., Xia, J. C. *et al.* Genome-wide scan for familial nasopharyngeal carcinoma reveals evidence of linkage to chromosome 4. *Nat. Genet.* **31,** 395–399 (2002).
42 Bakker, S. C., van der Meulen, E. M., Buitelaar, J. K., Sandkuijl, L. A., Pauls, D. L., Monsuur, A. J. *et al.* A whole-genome scan in 164 Dutch sib pairs with attention-deficit/hyperactivity disorder: suggestive evidence for linkage on chromosomes 7p and 15q. *Am. J. Hum. Genet.* **72,** 1251–1260 (2003).
43 Garner, C. P., Ding, Y. C., Steele, L., Book, L., Leiferman, K., Zone, J. J. *et al.* Genome-wide linkage analysis of 160 North American families with celiac disease. *Genes Immun.* **8,** 108–114 (2007).
44 López, S., Buil, A., Ordoñez, J., Souto, J. C., Almasy, L., Lathrop, M. *et al.* Genome-wide linkage analysis for identifying quantitative trait loci involved in the regulation of lipoprotein a (Lpa) levels. *Eur. J. Hum. Genet.* **16,** 1372–1379 (2008).
45 Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6,** 109–118 (2005).
46 Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6,** 95–108 (2005).
47 Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E. *et al.* Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* **1,** 109–111 (2004).
48 Steemers, F. J., Chang, W., Lee, G., Barker, D. L., Shen, R. & Gunderson, K. L. Whole-genome genotyping with the single-base extension assay. *Nat. Methods* **3,** 31–33 (2006).
49 Ragoussis, J. Genotyping technologies for genetic research. *Annu. Rev. Genomics Hum. Genet.* **10,** 117–133 (2009).
50 International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
51 Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
52 Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36,** 949–951 (2004).
53 Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).
54 Mardis, E. R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9,** 387–402 (2008).
55 Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
56 International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237–241 (2008).
57 Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S. *et al.* Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* **459,** 569–573 (2009).
58 Cook, E. H. Jr. & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455,** 919–923 (2008).
59 Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K. *et al.* A genetic linkage map of the human genome. *Cell* **51,** 319–337 (1987).
60 De la Vega, F. M., Lazaruk, K. D., Rhodes, M. D. & Wenz, M. H. Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP genotyping assays and the SNPlex genotyping system. *Mutat. Res.* **573,** 111–135 (2005).
61 Olivier, M. The invader assay for SNP genotyping. *Mutat. Res.* **573,** 103–110 (2005).
62 Ragoussis, J., Elvidge, G. P., Kaur, K. & Colella, S. Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genet.* **2,** e100 (2006).
63 Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J. P., Belaiche, J. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411,** 599–603 (2001).
64 Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M. *et al.* A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat. Genet.* **36,** 337–338 (2004).
65 Grant, S. F., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Manolescu, A., Sainz, J. *et al.* Variant of transcription factor 7-like 2 (*TCF7L2*) gene confers risk of type 2 diabetes. *Nat. Genet.* **38,** 320–323 (2006).
66 Remmers, E. F., Plenge, R. M., Lee, A. T., Graham, R. R., Hom, G., Behrens, T. W. *et al.* STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.* **357,** 977–986 (2007).

67 Hirschhorn, J. N. Genetic approaches to studying common diseases and complex traits. *Pediatric Res.* **57,** 74–77 (2005).
68 Kemp, Z., Carvajal-Carmona, L., Spain, S., Barclay, E., Gorman, M., Martin, L. *et al.* Evidence for a colorectal cancer susceptibility locus on chromosome 3q21–q24 from a high-density SNP genome-wide linkage scan. *Hum. Mol. Genet.* **15,** 2903–2910 (2006).
69 Sellick, G. S., Goldin, L. R., Wild, R. W., Slager, S. L., Ressenti, L., Strom, S. S. *et al.* A high-density SNP genome-wide linkage search of 206 families identifies susceptibility loci for chronic lymphocytic leukemia. *Blood* **110,** 3326–3333 (2007).
70 Stanford, J. L., FitzGerald, L. M., McDonnell, S. K., Carlson, E. E., McIntosh, L. M., Deutsch, K. *et al.* Dense genome-wide SNP linkage scan in 301 hereditary prostate cancer families identifies multiple regions with suggestive evidence for linkage. *Hum. Mol. Genet.* **18,** 1839–1848 (2009).
71 Gao, X., Martin, E. R., Liu, Y., Mayhew, G., Vance, J. M. & Scott, W. K. Genome-wide linkage screen in familial Parkinson disease identifies loci on chromosomes 3 and 18. *Am. J. Hum. Genet.* **84,** 499–504 (2009).
72 Sellick, G. S., Longman, C., Tolmie, J., Newbury-Ecob, R., Geenhalgh, L., Hughes, S. *et al.* Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays. *Nucleic Acids Res.* **32,** e164 (2004).
73 John, S., Shephard, N., Liu, G., Zeggini, E., Cao, M., Chen, W. *et al.* Whole-genome scan, in a complex disease, using 11 245 single-nucleotide polymorphisms: comparison with microsatellites. *Am. J. Hum. Genet.* **75,** 54–64 (2004).
74 Yatsu, K., Mizuki, N., Hirawa, N., Oka, A., Itoh, N., Yamane, T. *et al.* High-resolution mapping for essential hypertension using microsatellite markers. *Hypertension* **49,** 446–452 (2007).
75 Tamiya, G., Shinya, M., Imanishi, T., Ikuta, T., Makino, S., Okamoto, K. *et al.* Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. *Hum. Mol. Genet.* **14,** 2305–2321 (2005).
76 Kozlowski, P., de Mezer, M. & Krzyzosiak, W. J. Trinucleotide repeats in human genome and exome. *Nucleic Acids Res.* (2010) [e-pub ahead of print].
77 Day, I. N. dbSNP in the detail and copy number complexities. *Hum. Mutat.* **31,** 2–4 (2010).
78 International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
79 Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10,** 241–251 (2009).
80 Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29,** 229–232 (2001).
81 Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J. *et al.* Linkage disequilibrium in the human genome. *Nature* **411,** 199–204 (2001).
82 Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).
83 International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
84 International HapMap ConsortiumFrazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
85 Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38,** 659–662 (2006).
86 Eberle, M. A., Ng, P. C., Kuhn, K., Zhou, L., Peiffer, D. A., Galver, L. *et al.* Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet.* **3,** 1827–1837 (2007).
87 Li, M., Li, C. & Guan, W. Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur. J. Hum. Genet.* **16,** 635–643 (2008).
88 Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69,** 124–137 (2001).
89 Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19,** 212–219 (2009).
90 Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324,** 387–389 (2009).
91 Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–276 (2009).
92 Zogopoulos, G., Ha, K. C., Naqib, F., Moore, S., Kim, H., Montpetit, A. *et al.* Germ-line DNA copy number variation frequencies in a large North American population. *Hum. Genet.* **122,** 345–353 (2007).
93 de Smith, A. J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N. A., Tsang, P. *et al.* Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum. Mol. Genet.* **16,** 2783–2794 (2007).
94 Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K. *et al.* High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res.* **19,** 1682–1690 (2009).
95 Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84,** 148–161 (2009).
96 Scherer, S. W., Lee, C., Birney, E., Altshuler, D. M., Eichler, E. E., Carter, N. P. *et al.* Challenges and standards in integrating surveys of structural variation. *Nat. Genet.* **39,** S7–S15 (2007).
97 Estivill, X. & Armengol, L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* **3,** 1787–1799 (2007).
98 Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E. & Nickerson, D. A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.* **40,** 1199–1203 (2008).
99 Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1,** e56 (2005).
100 Bansal, V., Bashir, A. & Bafna, V. Evidence for large inversion polymorphisms in the human genome from HapMap data. *Genome Res.* **17,** 219–230 (2007).
101 Medvedev, P., Stanciu, M. & Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6,** S13–S20 (2009).
102 Tuzun, E., Sharp, A. J., Bailey, J. A., Kaul, R., Morrison, V. A., Pertz, L. M. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37,** 727–732 (2005).
103 Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature.* **453,** 56–64 (2008).
104 Human Genome Structural Variation Working GroupEichler, E. E., Nickerson, D. A., Altshuler, D., Bowcock, A. M., Brooks, L. D., Carter, N. P. *et al.* Completing the map of human genetic variation. *Nature* **447,** 161–165 (2007).
105 Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
106 McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19,** 1527–1541 (2009).
107 Feuk, L. Inversion variants in the human genome: role in disease and genome architecture. *Genome Med.* **2,** 11 (2010).
108 Antonacci, F., Kidd, J. M., Marques-Bonet, T., Ventura, M., Siswara, P., Jiang, Z. *et al.* Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.* **18,** 2555–2566 (2009).
109 Curtis, D., Vine, A. E. & Knight, J. Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. *Ann. Hum. Genet.* **72,** 261–278 (2008).
110 Assie, G., LaFramboise, T., Platzer, P. & Eng, C. Frequency of germline genomic homozygosity associated with cancer cases. *JAMA* **299,** 1437–1445 (2008).
111 Bacolod, M. D., Schemmann, G. S., Wang, S., Shattock, R., Giardina, S. F., Zeng, Z. *et al.* The signatures of autozygosity among patients with colorectal cancer. *Cancer Res.* **68,** 2610–2621 (2008).
112 Li, L. H., Ho, S. F., Chen, C. H., Wei, C. Y., Wong, W. C., Li, L. Y. *et al.* Long contiguous stretches of homozygosity in the human genome. *Hum. Mutat.* **27,** 1115–1121 (2006).
113 Ting, J. C., Roberson, E. D., Miller, N. D., Lysholm-Bernacchi, A., Stephan, D. A., Capone, G. T. *et al.* Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. *Hum. Mutat.* **28,** 1225–1235 (2007).
114 Curtis, D. Extended homozygosity is not usually due to cytogenetic abnormality. *BMC Genet.* **8,** 67 (2007).
115 Walsh, T., McClellan, J. M., McCarthy, S. E., Addington, A. M., Pierce, S. B., Cooper, G. M. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008).
116 Bochukova, E. G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463,** 666–670.
117 Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* **38,** 82–85 (2006).
118 McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40,** 1107–1112 (2008).
119 Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41,** 25–34 (2009).
120 Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F. *et al.* Segmental copy number variation shapes tissue transcriptomes. *Nat. Genet.* **41,** 424–429 (2009).
121 Cahan, P., Li, Y., Izumi, M. & Graubert, T. A. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat. Genet.* **41,** 430–437 (2009).
122 Mohlke, K. L., Boehnke, M. & Abecasis, G. R. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Hum. Mol. Genet.* **17,** R102–R108 (2008).
123 Easton, D. F. & Eeles, R. A. Genome-wide association studies in cancer. *Hum. Mol. Genet.* **17,** R109–R115 (2008).
124 Lettre, G. & Rioux, J. D. Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.* **17,** R116–R121 (2008).
125 Ku, C. S., Loy, E. Y., Pawitan, Y. & Chia, K. S. The pursuit of genome-wide association studies: where are we now? *J. Hum. Genet.* **55,** 195–206 (2010).

126 Cho, J. H. The genetics and immunopathogenesis of inflammatory bowel disease. *Nat. Rev. Immunol.* **8,** 458–466 (2008).
127 Mathew, C. G. New links to the pathogenesis of Crohn disease provided by genome-wide association scans. *Nat. Rev. Genet.* **9,** 9–14 (2008).
128 Maher, B. The case of the missing heritability. *Nature* **456,** 18–21 (2008).
129 Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316,** 445–449 (2007).
130 Hollox, E. J., Huffmeier, U., Zeeuwen, P. L., Palla, R., Lascorz, J., Rodijk-Olthuis, D. *et al.* Psoriasis is associated with increased beta-defensin genomic copy number. *Nat. Genet.* **40,** 23–25 (2008).
131 Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307,** 1434–1440 (2005).
132 Shlien, A. & Malkin, D. Copy number variations and cancer susceptibility. *Curr. Opin. Oncol.* **22,** 55–63 (2010).
133 Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R. & Hobbs, H. H. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305,** 869–872 (2004).
134 Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H. *et al.* Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.* **39,** 513–516 (2007).

Hum Genet (2011) 129:1–15
DOI 10.1007/s00439-010-0920-6

REVIEW PAPER

Regions of homozygosity and their impact on complex diseases and traits

Chee Seng Ku · Nasheen Naidoo · Shu Mei Teo · Yudi Pawitan

Received: 8 August 2010 / Accepted: 4 November 2010 / Published online: 23 November 2010

Abstract Regions of homozygosity (ROHs) are more abundant in the human genome than previously thought. These regions are without heterozygosity, i.e. all the genetic variations within the regions have two identical alleles. At present there are no standardized criteria for defining the ROHs resulting in the different studies using their own criteria in the analysis of homozygosity. Compared to the era of genotyping microsatellite markers, the advent of high-density single nucleotide polymorphism genotyping arrays has provided an unparalleled opportunity to comprehensively detect these regions in the whole genome in different populations. Several studies have identified ROHs which were associated with complex phenotypes such as schizophrenia, late-onset of Alzheimer's disease and height. Collectively, these studies have conclusively shown the abundance of ROHs larger than 1 Mb in outbred populations. The homozygosity association approach holds great promise in identifying genetic susceptibility loci harboring recessive variants for complex diseases and traits.

Introduction

Human genetic variations are the differences in DNA sequences within the genome of individuals within populations. These variations can take many forms, including single nucleotide variants or substitutions, tandem repeats (short tandem repeats and variable number of tandem repeats), small indels (insertions and deletions of a short DNA sequence), duplications or deletions that change the copy number of a larger segment of a DNA sequence ($\geq$1 kb) i.e. copy number variations (CNVs), and other chromosomal rearrangements such as inversions and translocations (also known as copy-neutral variations) (Nakamura 2009; Frazer et al. 2009; Ku et al. 2010a). The amount of genetic variation in the human genome is more abundant than previously thought, and this has been further corroborated with the findings from whole genome resequencing studies where several million single nucleotide polymorphisms (SNPs) and several hundred thousand indels and structural variants were identified (Wheeler et al. 2008; Bentley et al. 2008; Wang et al. 2008; Kim et al. 2009). In addition to SNPs (Altshuler et al. 2008; Hindorff et al. 2009), other genetic variations have also been found to be associated with various complex diseases and traits (Haberman et al. 2008; Hannan 2010; Wain et al. 2009; Stankiewicz and Lupski 2010).

By comparison, the region of homozygosity (ROH) is not currently classified as a type of genetic variation as there is no consensus on whether it should be classified as one type of 'structural' genetic variation. The reasons for this are two fold: (a) the ROH is not a 'genetic alteration' of the DNA sequence like other genetic variations and, (b) the research on their genome-wide mapping is still relatively new. However, the extent of ROHs varies among individuals and between different populations. In comparison to other types of genetic variations where the inter-population differences have been well documented (International HapMap Consortium 2005, 2007; Jakobsson et al. 2008; Teo et al. 2009), published data has increasingly shown the

C. S. Ku (✉) · N. Naidoo · S. M. Teo
Department of Epidemiology and Public Health,
Centre for Molecular Epidemiology,
Yong Loo Lin School of Medicine,
National University of Singapore, Singapore, Singapore
e-mail: g0700040@nus.edu.sg

Y. Pawitan (✉)
Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, Stockholm, Sweden
e-mail: yudi.pawitan@ki.se

inter-individual and inter-population variations in the profiles of homozygosity (Gibson et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; O'Dushlaine et al. 2010).

Research on ROHs has started to gain impetus, as is evidenced by the increasing numbers of publications after the first study by Gibson et al. (2006) reporting its abundance in the human genomes of outbred populations. Further studies have investigated the population genetics aspects of ROHs in healthy individuals (Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b), and also performed association analyses to identify ROHs that are associated with complex diseases and traits in a case–control study design (Lencz et al. 2007; Nalls et al. 2009a; Vine et al. 2009; Yang et al. 2010b).

The aim of this paper is to review the recent progress and to elaborate on the issues and challenges in genome-wide mapping of ROHs in the human genome using high-density SNPs genotyping arrays in normal populations and in disease association studies. We also highlight the findings showing associations between ROHs and complex phenotypes. Finally, we discuss the future directions and the potential applications of ROHs as surrogate markers in identifying recessive loci for complex phenotypes. This approach is also known as 'genome-wide homozygosity association' and could be a promising alternative to finding the 'missing heritability' for complex phenotypes (Manolio et al. 2009). Population genetics and selection pressure on ROHs are briefly discussed, as these topics are beyond the scope of this review paper. Other interesting areas of ROHs research such as studies of homozygosity in inbreeding and isolated populations and findings from animal and plant genetics deserve to be reviewed in a separate paper.

What is a region of homozygosity?

A ROH defines a continuous or uninterrupted stretch of a DNA sequence without heterozygosity in the diploid state, that is in the presence of both copies of the homologous DNA segment. Thus, all the genetic variations, such as SNPs (biallelic marker) or microsatellites (multiallelic marker) within the homologous DNA segments have two identical alleles that create homozygosity (Gibson et al. 2006). The ROH is different from one-copy deletion (or hemizygous deletion), which could also lead to the homozygosity, e.g. in genome-wide SNPs genotyping data. However this is considered as a 'spurious homozygosity' because only one allele of the SNPs is present in the deleted region for one-copy deletions. Thus, the DNA fragments with only the single allele are hybridized on the genotyping array. As a result, the signal intensity of only one allele is measured and subsequently used in genotype calling, and hence it would be incorrectly labeled a homozygote genotype. Therefore, the result of 'homozygosity' is due to the absence of the other allele, instead of 'true homozygosity' where two identical alleles are present (Peiffer et al. 2006). The distinction between 'true homozygosity' as opposed to 'spurious homozygosity' due to one-copy deletion is difficult to determine just by inspection of the genotype data alone. The allelic signal intensity ratio (the relative ratio of the fluorescent signals between two probes/alleles at each SNP) is needed to differentiate between the two types of homozygosity (Peiffer et al. 2006; Wang et al. 2007). Therefore, for studies that used only SNPs genotype data to identify the ROHs, i.e. to screen regions with a minimum consecutive homozygote SNPs, the possibility that some regions are caused by one-copy deletion cannot be firmly excluded, because deletions are also widespread in the human genome (McCarroll et al. 2008; Conrad et al. 2010).

Cytogenetic abnormalities such as uniparental isodisomy can also result in homozygosity where two copies of a single parental homologous DNA segment are inherited from one parent. As such it cannot be distinguished from homozygosity resulting from other factors such as parental consanguinity using the allelic signal intensity ratio as in the case of one-copy deletion. Thus for studies that involved unrelated samples where checking the Mendelian transmission errors in the ROHs is not possible, the possibility of uniparental isodisomy leading to homozygosity cannot be definitively ruled out. Assessing the transmission errors requires data from trios or families. However, the likelihood that a considerable fraction of ROHs will be accounted for by uniparental isodisomy is low given that this cytogenetic abnormality is rare (Curtis 2007).

Currently, there is no consensus or standardized criteria used to define the ROH. However, previous studies have focused on regions $\geq$1 Mb, and thus the true extent of homozygosity in the human genome could be underestimated (Gibson et al. 2006; Li et al. 2006). More recent studies have defined a ROH at a minimum length of 500 kb (Yang et al. 2010b) with the intention of avoiding underestimation of the numbers of regions in the human genome. This is because shorter ROHs are now also thought to be associated with complex phenotypes. However, setting a shorter length for definition will increase the number of false positive signals i.e. increase the sensitivity at the expense of specificity. Therefore, in discovery studies, balancing both the sensitivity and specificity when setting the criteria to identify ROHs is critical.

By focusing only on regions $\geq$500 kb or 1 Mb, the 'noise' introduced by one-copy deletions is likely to be minimal, thus reducing the potential to cause spurious homozygosity. This is because large deletions of $\geq$500 kb are relatively rare in the human genome—as supported by data from high-resolution genome-wide mapping of CNVs

studies (McCarroll et al. 2008; Conrad et al. 2010; Ku et al. 2010b; Park et al. 2010a; Yim et al. 2010). Therefore, a critical issue to be addressed in future homozygosity mapping studies is determining the optimal cutoff of the length of the ROH to be adopted, as this will avoid over-estimating the homozygosity when the length is set too low and which can then be easily confounded by one-copy deletion of hundreds of kilobases or smaller. Although some studies have reduced the cutoff length to 500 kb (Yang et al. 2010b), it is still uncertain whether this new cutoff can readily reflect the true extent of homozygosity in the human genome.

Defining criteria and terminologies

Before the term 'copy number variation (CNV)' was first introduced in 2006 (Freeman et al. 2006), various different terms were used to describe these copy number variable regions such as 'large-scale copy number variants' and 'intermediate-sized variants' (Sebat et al. 2004; Iafrate et al. 2004). To date, various terminologies have also been used to describe the ROHs such as 'extended tracts of homozygosity' (Gibson et al. 2006), 'long contiguous stretches of homozygosity' (Li et al. 2006), 'runs of homozygosity' (Nothnagel et al. 2010; McQuillan et al. 2008), 'autozygosity regions' (Nalls et al. 2009b) and 'homozygosity-by-descent' (Polasek et al. 2010). Different studies have used their own criteria in identifying ROHs with some studies employing more stringent criteria compared to others applying a more liberal definition (Gibson et al. 2006; Li et al. 2006; Nothnagel et al. 2010; McQuillan et al. 2008; Nalls et al. 2009b; Curtis et al. 2008). For example, Curtis et al. (2008) used their own developed software and the criteria of a minimum of 10 consecutive, homozygous SNPs extending over 1 Mb. In comparison, other studies employed the default definition implemented in the 'Runs of homozygosity' function in the PLINK software (http://pngu.mgh.harvard.edu/~purcell/plink/). These criteria are (a) the length of the ROH $\geq$1 Mb, (b) a minimum of 100 SNPs per ROH, and (c) a density of at least 1 SNP per 50 kb (Nothnagel et al. 2010). As all the studies are referring to the same type of 'DNA sequence feature' it is essential to standardize the terminology to be used in describing these regions to avoid confusion.

Polymorphic markers used to detect ROHs

Although long continuous ROHs have been documented a decade ago in reference families from the Centre D'etude Du Polymorphisme Humain (CEPH) (Broman and Weber 1999), no large-scale population-based study had been performed to interrogate the extent of ROHs in the human genome until the first study by Gibson et al. (2006). The recent advances in genome-wide mapping or detection of ROHs have been driven mainly by the availability of highly accurate SNPs databases such as the International HapMap Project, and the technology to genotype several hundred thousand to several million SNPs throughout the human genome (International HapMap Consortium 2005, 2007; Gibbs and Singleton 2006; Ragoussis 2009). The early study in the CEPH families used approximately 8,000 short tandem repeat markers and detected long continuous ROHs. In contrast, subsequent studies have applied SNPs as the polymorphic markers to detect the ROHs (Gibson et al. 2006; Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b). At the single marker level, short tandem repeats are more informative than SNPs because they are multiallelic markers. However, SNPs are more numerous and collectively can yield more information than short tandem repeats and offer a higher resolution compared to other genetic markers—both of which are important to accurately identify the numbers and sizes of ROHs.

Genotyping a large number of SNPs in a microarray platform presents a powerful tool to detect ROHs comprehensively across the whole genome (Gibbs and Singleton 2006; Ragoussis 2009). This also enables investigation into the number, length or size, and location or distribution of the ROHs in the human genome in a more unbiased manner compared to microsatellite markers (Gibson et al. 2006; Li et al. 2006; McQuillan et al. 2008; Nothnagel et al. 2010; Nalls et al. 2009b). The SNPs genotyping platforms also allow studies of the relationship between ROHs and recombination or linkage disequilibrium (LD) patterns, as the SNPs data can be used for haplotype analyses and to calculate the recombination rates (Curtis et al. 2008). The ability to investigate the co-occurrence of ROHs in the areas with extensive LD or low recombination is important in investigating the mechanisms contributing towards the high frequency of ROHs in the human genome.

Genotyping of a sufficiently large number of SNPs is required to accurately detect the ROHs. The study by Gibson et al. (2006) used data from the International HapMap Phase I Project comprising of approximately 1 million SNPs (International HapMap Consortium 2005), whilst other studies have used lower density genotyping arrays ranging from 300,000 to 550,000 SNPs. The importance of having high-density polymorphic markers was shown by Gibson et al. (2006) who found the largest ROH of 17.9 Mb containing 3,922 SNPs from the SNPs data from HapMap Phase I. However, using the data from HapMap Phase II comprising of >3 million SNPs (International HapMap Consortium 2007), a total of 12,778 SNPs were found in the region with 11 heterozygotes. These heterozygotes interrupted the ROH and have divided it into 12 smaller

segments (Gibson et al. 2006). However, it is unclear whether these 11 heterozygotes are genotyping errors or true heterozygotes occurring as a result of recent mutations. Thus, to account for genotyping errors, studies have allowed some missing genotypes and heterozygotes for each ROH to avoid artificially splitting the region (Table 1).

This hints that the sizes of ROHs may be over-estimated in previous studies when using lower density SNPs genotyping arrays. Therefore, the numbers and sizes of ROHs identified by previous studies are likely to be different or altered when higher density SNPs data is available for analysis on the same samples. This also implies that a cautious interpretation should be imposed for ROHs of several megabases for studies using lower resolution SNPs data. A higher density of SNPs is needed for a definitive assessment of ROHs. Although the SNPs genotyping array is an invaluable tool to detect ROHs, it is not without limitations. Similar to CNV detection using SNPs genotyping platforms, the boundaries of the ROHs cannot be determined accurately at a single nucleotide resolution, as accuracy depends on the SNPs resolution. Therefore, like CNVs, the sizes of ROHs could be inflated, i.e. the ROHs detected in previous studies could be smaller than currently estimated. However, there is currently no data supporting this speculation for ROHs as compared to CNVs (McCarroll et al. 2008; Perry et al. 2008).

Methods of detecting ROHs

Several targeted and genome-wide molecular methods are available to detect structural variations such as CNVs (deletions and duplications) and copy-neutral variations (translocations and inversions). However, unlike with structural variations, ROHs cannot be detected with technologies used in molecular genetics such as fluorescence in situ hybridization (FISH) and bacterial artificial chromosome (BAC) clone or oligonucleotide-based comparative genomic hybridization (CGH) arrays (Carson et al. 2006; Feuk et al. 2006; Carter 2007). Furthermore, several new sequencing-based approaches for detecting structural variations such as paired-end sequencing mapping and depth-of-coverage of the sequence read are also unfit to detect ROHs (Korbel et al. 2007; Kidd et al. 2008; Yoon et al. 2009).

The genome-wide mapping of ROHs can only be done using SNPs genotyping arrays or direct sequencing. The whole-genome resequencing or de novo genome assembly using the next or third generation sequencing technologies will offer an almost complete solution to detecting most of the genetic variations including ROHs within the human genome. However, these high-throughput sequencing technologies were not readily available until recently, and the cost is still prohibitively expensive to sequence the whole human genome in a population-based study (Mardis 2008; Metzker 2010). As a result, SNPs genotyping arrays are the main tools for ROH mapping. The SNPs data can be used in two different ways to detect the ROHs. The first approach is to screen the whole genome in a sliding window manner for consecutive SNPs showing homozygotes over a certain length such as 1 Mb, as implemented in PLINK (Purcell et al. 2007). Since this approach only uses genotype data, it is unable to distinguish between true homozygosity and the spurious homozygosity caused by one-copy deletion without further investigations of CNVs in the samples.

This limitation has been overcome by the second approach which relies on the signal intensity data. Two types of signal intensity data are generated by the SNPs genotyping array: (a) the total signal intensity or log R ratio (LRR) and (b) the allelic intensity ratio or B allele frequency (BAF). The combination of LRR and BAF can be used to determine several different states of copy numbers such as homozygous and hemizygous deletions, and one-copy and two-copy duplications, and ROHs as implemented in the PennCNV algorithm. The BAF is needed to differentiate between ROH from normal diploid copies and one-copy deletion (Wang et al. 2007). Figure 1 illustrates the difference in LRR and BAF patterns between ROH and one-copy deletion. For the one-copy deletion, there is a decrease in LRR in addition to the absence of heterozygosity as shown in the BAF panel. Conversely, no reduction in LRR will be seen for ROH, but the absence of heterozygosity is observed. Most of the genome-wide studies of ROHs have used SNPs genotyping arrays. In comparison, the commonly used oligonucleotide-based CGH arrays in detecting CNVs produced only total signal intensity data. This renders them unable to be used for identifying ROHs.

In addition to the most commonly used PLINK software for detecting and analyzing ROHs (Table 1), other methods have also been recently developed for these purposes (Seelow et al. 2009; Browning and Browning 2010; Polasek et al. 2010). The development of powerful and accurate tools or methods for the detection and analysis is a prerequisite for the success of research into ROHs. Furthermore, new algorithms to identify disease-related segments based on homozygosity using case–control data have also been developed. This will enhance studies to identify ROHs that differ between cases and controls, as these regions may contain recessive variants underlying the diseases (Wang et al. 2009). All the ROHs detection methods have their own strengths and limitations with varying rates of false-positive and false-negative results and as such, a combination of methods would be more ideal to minimize these limitations.

Table 1 Summaries of genome-wide association studies of ROHs and complex phenotypes using high-density SNP genotyping arrays

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Schizophrenia (Lencz et al. 2007)	178 cases and 144 controls Affymetrix 500K	Software • Whole-genome homozygosity analysis (WGHA) performed with customized python scripting in the HelixTree environment Criteria • ROH—any window of 100 or more consecutive SNPs that are homozygous, not receiving a heterozygous call • Common ROHs—only those ROHs in which 10 or more subjects share ≥100 identical homozygous calls were retained for further analysis Association analysis • Case–control comparisons of frequency of presence for each common ROH were examined by using χ^2 test	• A total of 339 common ROHs were identified • Schizophrenia cases demonstrated a significantly greater number of common ROHs than controls • 9 ROHs significantly differed in frequency between cases and controls
Bipolar disorder (Vine et al. 2009)	553 cases and 547 controls Affymetrix 500K	This study applied the WGHA approach as demonstrated in the Lencz et al. (2007) study	• A total of 239 common ROHs were identified • The total number of common ROHs did not differ between cases and controls • 7 common ROHs were significant at $p < 0.05$
Late-onset Alzheimer's disease (Nalls et al. 2009a)	837 cases and 550 neurological normal controls Affymetrix 500K	Software • PLINKv1.02 • A sliding window of 50 SNPs, allowing at most 2 missing genotypes and 1 heterozygote call per ROH Criteria • ROH—at least 1 Mb of consecutive homozygous genotypic calls •Minimum SNP density coverage—at least 50 SNPs per megabase Association analysis • 1,090 consensus regions from overlapping ROHs were defined •Each consensus region was found in no less than 10 participants • The consensus ROHs were analyzed using the maxT permutation test algorithm for case/control studies in PLINKv1.02	• One homozygous consensus region in chromosome 8 was found to be significantly overrepresented in cases when compared to controls • The cases presented a slightly higher degree of extended homozygosity when compared with the control group

Table 1 continued

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Height (Yang et al. 2010b)	Discovery study 998 US Caucasian subjects Affymetrix 500K Replication study 8,385 Caucasian subjects from the Framingham Heart Study Affymetrix 500K plus 50K supplemental array	Software • PLINKv1.01 • A sliding window of 5 Mb (minimum 50 SNPs), allowing 5 missing SNPs and 1 heterozygous site per window Criteria • A minimum of 100 consecutive SNPs in a ROH • Minimum length for a ROH, 500 kb • Minimum density in a ROH, 50 kb per SNP • Maximum gap between 2consecutive homozygous SNPs—100 kb Association analysis • Individual ROHs were divided into different ROH groups using the homozyg-group command in the Runs of Homozygosity program • For each ROH group containing >50 subjects—Student's *t* test to compare the adult height of subjects with this ROH group to the height of subjects without this ROH group	Discovery study • 113,910 individual ROHs in 998 subjects • For the association analyses between human adult height and ROHs, 3,322 ROH groups containing more than 50 individual ROHs • 80 ROH groups overlapped with copy number polymorphisms and were excluded from the subsequent association analyses. • One ROH group (ROH 12q21.31) was significantly associated with adult height even after Bonferroni correction Replication study • A significant association with adult height was successfully replicated for the ROH group by FBAT analysis
Colorectal cancer (Spain et al. 2009)	921 cases and 929controls Illumina Infinium Human Hap550 BeadChips	Software • PLINKv1.05 • A sliding window of 50 SNPs, allowing 2% heterozygous SNPs and 5 missing calls in each window Criteria • This study initially analyzed ROHs that were ≥50 SNPs in length • Repeated the analysis using a number of different criteria to define a ROH (≥30 SNPs, ≥40 SNPs, ≥60 SNPs, ≥2 Mb, ≥4 Mb, and ≥10 Mb) Association analysis • Statistical analyses were performed using packages available in R	• No evidence was found for an association between total size of the ROHs in each individual and colorectal cancer • This study calculated the frequencies of cases and controls in which one or more ROHs of ≥4 Mb were detected • 159 of 921 (17%) cases and 142 of 929 (15%) controls had ROHs ($p = 0.14$, Fisher's exact test)
Childhood acute lymphoblastic leukemia (Hosking et al. 2010)	824 cases and 2,398 controls Illumina Infinium Human370 Duo BeadChips	Software • PLINKv1.06 • A sliding window of SNPs across the entire genome, 2% heterozygous SNPs were allowed in each window, 5 missing calls per window Criteria • ROH, ≥75 consecutive SNPs • Only ROHs which occurred in ≥10 persons were retained for analysis Association analysis • Subsequent statistical analyses were performed using packages available in R • Comparison of the distribution of categorical variables was performed using the χ^2 test	• A total of 396 ROHs were identified • Patients and controls showed no significant difference in the average number of ROH • 4 ROHs differed significantly ($p < 0.01$) between cases and controls

Table 1 continued

Phenotype and study	Sample size and genotyping platform	Software, criteria of ROHs, association analysis	Major results
Breast and prostate cancer (Enciso-Mora et al. 2010)	Breast cancer 1,183 cases and 1,185 controls Illumina Infinium Human550 Duo BeadChips Prostate cancer 1,177 cases and 1,149 controls Illumina Infinium Human217 and Human 317 BeadChips	Software • PLINK v1.06 • A sliding window of SNPs across the genome, 2% heterozygous SNPs were permitted in each window, 5 missing calls per window Criteria • ROH, ≥80 consecutive SNPs • Only considered ROH that occurred in ≥ 10 individuals Association analysis • Subsequent statistical analyses were performed using packages available in R • Comparison of the distribution of categorical variables was performed using the χ^2 test	• A total of 415 and 426 ROHs were identified in breast cancer and prostate cancer series, respectively • 6 ROHs differed significantly ($p < 0.01$) between breast cancer cases and controls. • 4 ROHs differed significantly ($p < 0.01$) between prostate cancer cases and controls

Different studies have applied different filtering or quality control criteria of the genome-wide SNP data and samples before the data was used for ROH analysis and association studies

Mechanisms generating ROHs

Several mechanisms and factors have been postulated to explain the high frequency of ROHs in the human genome namely, parental consanguinity, uniparental isodisomy and the presence of 'common extended haplotypes'. One of the most common and well established mechanisms leading to ROHs of several megabases is parental consanguinity, in which the offspring inherits chromosomal segments that are identical-by-descent from each parent. Published data has shown that the number of ROHs of several megabases increased markedly in the offspring of consanguineous marriages (Li et al. 2006; Woods et al. 2006) with up to 6% of homozygosity anticipated in the genome of the offspring of first cousin marriages (Broman and Weber 1999). Li et al. (2006) showed that in a family with 4 children from first cousin marriages, multiple ROHs ranging from 3.06 to 53.17 Mb were observed in all the children. Woods et al. (2006) also showed a marked increase in homozygosity levels in individuals with a recessive disease whose parents were first cousins, where 11% of their genomes were homozygous on average. Additionally, the cumulative length of ROHs per genome was found to be larger in two isolated rather than in two more cosmopolitan (non-isolated) European populations (McQuillan et al. 2008). Therefore, when compared to outbred populations, there is an expected increase in the level of homozygosity or number of ROHs in populations where consanguineous marriages are prevalent, as well as in isolated populations where limited random mating or a restricted mate choice has taken place. However, this is unlikely to be the main factor responsible for the high frequency of ROHs in outbred populations in which parental consanguinity is uncommon.

Another widely discussed mechanism is cytogenetic abnormalities such as uniparental disomy, which can be divided into uniparental isodisomy and uniparental heterodisomy. Only uniparental isodisomy can cause homozygosity as the offspring inherits two identical copies of a homologous chromosomal segment from only one parent. As a result, no heterozygosity would be observed in that particular homologous chromosomal segment (Ting et al. 2007). Similarly, this is also an unlikely explanation for the abundance of ROHs reported in the literature; given that uniparental disomies are rare genetic abnormalities that can cause severe and rare genomic disorders when their locations affect imprinted genes. Examples of these disorders are Prader–Willi Syndrome, Angelman Syndrome and Silver–Russell syndrome (Gurrieri and Accadia 2009; Van Buggenhout and Fryns 2009; Abu-Amero et al. 2008). This is further supported by previous studies concluding that the ROHs are not due to genetic abnormalities as no excess apparent deviation from Mendelian transmission was observed. More specifically, transmis-

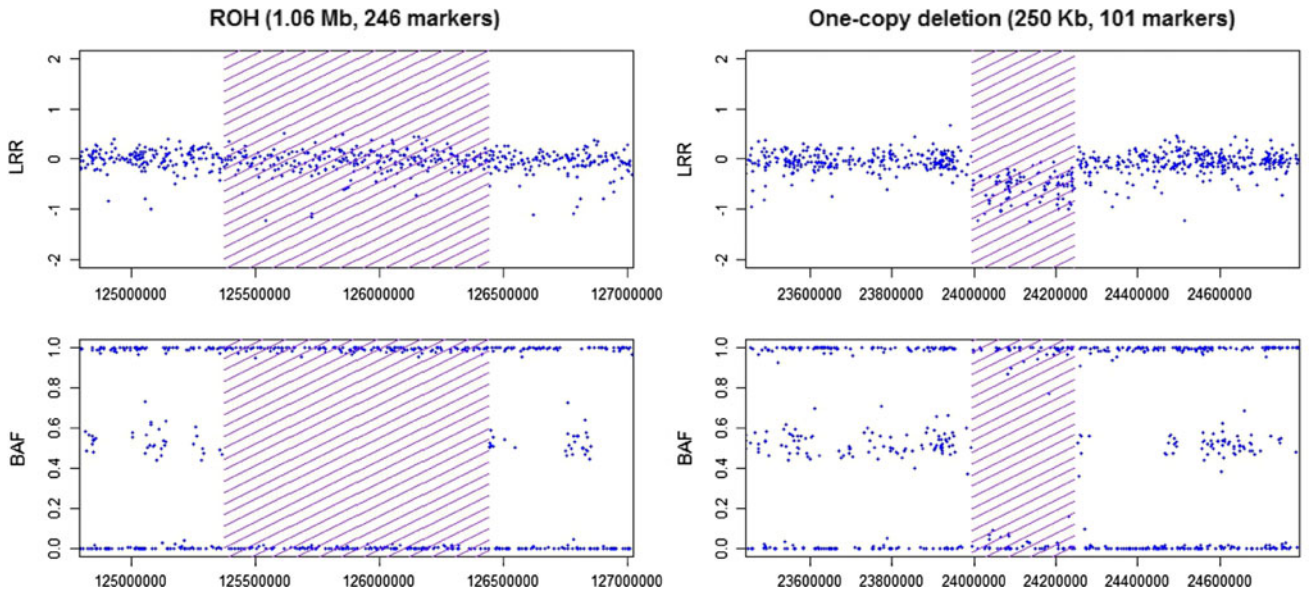


Fig. 1 Plots of the differences in the LRR (Log R Ratio) and BAF (B Allele Frequency) patterns for the ROH (*left panels*) and one-copy deletion (*right panels*) generated from a sample derived from our previous study (Ku et al. 2010b) and genotyped by the Illumina 1 M Beadchip. The ROH and one-copy deletion were detected using the LRR and BAF information by PennCNV algorithm (LRR: total fluorescent intensity signals from both sets of probes/alleles at each SNP, BAF: the relative ratio of the fluorescent signals between two probes/alleles at each SNP) (Wang et al. 2007). The size of the ROH is approximately 1.06 Mb (1,064,933 bases) spanning from 125374832 to 126439764 in chromsome 2. This region contains 246 markers. The size of the one-copy deletion is approximately 250 kb (250,186 bases) spanning from 23994408 to 24244593 in Chromosome 22. This region contains 101 markers. The regions affected by the ROH and one-copy deletion were shaded and the blue dots represent markers in the genotyping array

sion errors occur more rarely in ROHs than would be expected by chance as shown by the observed number of Mendelian transmission errors within a ROH which is less than the expected number (Curtis 2007). Since this study has clearly demonstrated that the ROHs are not usually due to cytogenetic abnormalities, it then indirectly supports the presence of common extended haplotypes as the mechanism contributing toward the high frequency of ROHs in human genomes.

The presence of common extended haplotypes therefore becomes the most likely factor responsible for the high frequency of ROHs which are passed on from both parents to the offspring in the genomes of outbred populations. Data demonstrating the co-occurrence of ROHs in regions with extensive LD and low recombination rates also support the hypothesis of common extended haplotypes in generating homozygosity in the genomes of outbred populations (Gibson et al. 2006; Curtis et al. 2008). A further process believed to be driving the increasing frequency of common extended haplotypes is positive selection. ROHs resulting from common extended haplotypes may be indicative of positive selection pressure of functional importance of these regions. Several methods have been used to quantify the positive selection pressure on ROHs namely, the integrated haplotype score (iHS), Tajima's *D* test and the Fixation index (F_{ST}). Numerous large (several megabases) and common (>25%) ROHs were found to have high values for these metrics indicating the signal for positive selection (Enciso-Mora et al. 2010; Hosking et al. 2010).

Genome-wide mapping of ROHs in the human genome

It was not previously expected that the genomes of outbred populations contain ROHs of several megabases until the first few early reports in 2006 and 2007 (Gibson et al. 2006; Li et al. 2006; Simon-Sanchez et al. 2007). One study found ROHs of >5 Mb in 26 of the 272 unrelated samples assessed (Simon-Sanchez et al. 2007). Similarly, another study performed in Han Chinese also observed the high frequency of ROHs, where 34 of the 515 unrelated individuals contained ROHs ranging from 2.94 to 26.27 Mb (Li et al. 2006). While Gibson et al. (2006) studied the samples from the International HapMap Projects and identified 1,393 ROHs exceeding 1 Mb in 209 unrelated HapMap individuals. Several hundreds of ROHs were found in each of the HapMap populations, and the average number of ROHs (>1 Mb) per individual was found to be lowest in the Yoruba Ibadan Nigerian (YRI) population compared to other populations within the HapMap Phase I Project (Gibson et al. 2006). In addition to demonstrating that ROHs are remarkably common, even in the unrelated individuals from the apparently outbred populations, Gibson et al. (2006) also demonstrated the value of including diverse

populations to examine the differences in ROHs. In the YRI population, the samples have the least number of ROHs per individual. This finding is expected, because the populations of African ancestry are older in human history and hence have more generations and a higher number of recombination events than other populations (recombination occurs during meiosis in each generation). Recombination is one of the important processes to interrupt the long continuous ROHs into smaller segments over the generations. Population differences in ROHs have also been well documented in other studies (Nothnagel et al. 2010).

Each of the previous studies identified a different number of ROHs per individual (Li et al. 2006; Nothnagel et al. 2010; McQuillan et al. 2008; Nalls et al. 2009b; Curtis et al. 2008). These differences are likely reflective of technical and methodological variations such as differing genotyping platforms or SNPs data, differing defining criteria and differing analytical techniques used. Both the genotyping platform and defining criteria can significantly influence the profile of ROHs by way of number, size, cumulative length and genomic distributions. Slight alterations in defining criteria can substantially affect the number of ROHs detected and as a result comparisons between studies are difficult. Therefore, it is critical to develop a set of standardized criteria in identifying ROHs and to establish a database to catalog these regions in the human genome from published studies, similar to other databases developed for SNPs and structural variants (CNVs) such as the dbSNP and Database of Genomic Variants, respectively (Day 2010; Iafrate et al. 2004). This database will enable researchers to quickly compare their results with published data. Consensus on defining the ROHs and the construction of a database to serve as a reference will help in expediting research in ROHs.

LD-pruning of SNPs in mapping of ROHs

The SNPs genotyping data is undoubtedly invaluable for identifying ROHs. However, there is an issue of whether pruning the list of SNPs to remove local LD (i.e. to remove SNPs that are in strong LD) should be done before the data can be used for ROHs. The idea of LD-pruning of SNP data is that the LD between the SNPs can inflate the chance of occurrence of biologically meaningless ROHs. However, there are still uncertainties with regards to the LD-pruning step such as the optimal cutoff of LD (measured by r^2) to be used, although some studies have used the conventional and arbitrary cutoff of $r^2 > 0.8$. More importantly, it is unclear about the quality and performance in terms of sensitivity and specificity for mapping ROHs using LD-pruning SNPs data compared to data without the LD-pruning step. This is an interesting research subject worth pursuing and studies should be done to assess the importance of this LD-pruning step. However, unless significant differences in the sensitivity and specificity are shown using LD-pruning SNP data, the LD-pruning step may not necessarily be needed.

Some of the studies using whole-genome SNPs genotyping arrays have omitted the LD-pruning step before the data was used for mapping ROHs, even though Gibson et al. (2006) used the SNP data from the International HapMap Project where the LD information is readily available. However, others have taken the LD between SNPs into account and used the pairwise LD SNP pruning function in PLINK, with a default value of $r^2 > 0.8$ (Enciso-Mora et al. 2010; Hosking et al. 2010). For example, one study found 370,611 separate tag groups which is a 27.6% reduction of information compared with the original number of SNPs. To account for this, the study adopted a more stringent cutoff of a minimum of 80 consecutive SNPs (instead of 58) to identify ROHs (Enciso-Mora et al. 2010). Similarly Lencz et al. (2007) also took into consideration the LD between the SNPs through setting a more stringent threshold of 100 consecutive SNPs that are homozygous. In comparison, another study removed SNPs in LD with $r^2 < 0.1$ leaving only 30,307 SNPs to form the 'low-LD panel' for some analyses (Spain et al. 2009). Although these studies have taken LD between SNPs into account, it is unclear whether an improvement in sensitivity and specificity was achieved by implementing this LD-pruning step since no evaluation was done to directly compare the differences between the ROHs profile with and without the LD-pruning step. Therefore, the LD-pruning step is conceptually correct; however to warrant this step to be performed in future genome-wide mapping of ROHs, more published data demonstrating its advantages is needed.

Implications on complex diseases and traits

Many novel pathogenic genes or mutations underlying autosomal recessive disorders have been identified through homozygosity mapping. This approach has been shown to be powerful and is particularly useful in investigating autosomal recessive disorders especially in populations with a high prevalence of consanguinity. This is evident from the enormous number of studies identifying causal mutations for autosomal recessive disorders in consanguineous families (Abu Safieh et al. 2010; Harville et al. 2010; Walsh et al. 2010; Pang et al. 2010; Lapunzina et al. 2010; Nicolas et al. 2010; Uz et al. 2010; Iseri et al. 2010; Collin et al. 2010). However, the first study applying the homozygosity association approach at the genome-wide scale for complex diseases only appeared in 2007 (Lencz et al. 2007). Table 1 summarizes the

genome-wide ROH association studies of complex phenotypes using high-density genotyping arrays.

The 'homozygosity analysis' has been shown to be useful for the identification of disease susceptibility genes in both monogenic and complex diseases (Miyazawa et al. 2007; Jiang et al. 2009). The effects of inbreeding or consanguinity and recessive variants or heterozygosity levels on the risk of complex phenotypes (diseases and quantitative traits) have been previously well established (Rudan et al. 2003a, 2003b, 2006; Campbell et al. 2007). A strong linear relationship between the inbreeding coefficient and blood pressure was found and several hundred recessive loci were predicted as contributing to blood pressure variability. Recessive or partially recessive genetic variants account for 10–15% of the total variation in blood pressure (Rudan et al. 2003a). Higher levels of relative heterozygosity were shown to be associated with lower blood pressure and total and low-density lipoprotein cholesterol by measuring genome-wide heterozygosity (Campbell et al. 2007). In addition to quantitative traits, inbreeding was also found to be a significant positive predictor for a number of late-onset complex diseases such as coronary heart diseases, stroke, cancer and asthma (Rudan et al. 2003b). These studies have strongly supported the hypothesis that the genetics of complex phenotypes include a component of recessively acting variants; however, these studies did not directly investigate the associations of complex phenotypes with ROHs detected using polymorphic markers.

Although the information regarding the extent of ROHs in the human genome is still limited compared with SNPs, indels and CNVs, their potential impact on complex diseases and traits could also be significant as other genetic variations. The importance of ROHs to complex phenotypes remains largely unexplored; however, several studies have shown significant differences in ROHs between cases and controls in a genome-wide investigation for schizophrenia (Lencz et al. 2007), late-onset Alzheimer's disease (Nalls et al. 2009a) and height (Yang et al. 2010b). The idea underlying the homozygosity association approach is to uncover recessive variants contributing to complex phenotypes. The success of this approach has been demonstrated in several studies. Nine common ROHs significantly differentiated schizophrenia cases from controls. More interestingly, four of the regions contained or were located near to the genes that are known to be associated with schizophrenia such as *NOS1AP*, *ATF2*, *NSF*, and *PIK3C3* (Lencz et al. 2007). This proof-of-principle study has demonstrated the applications of the whole-genome homozygosity association approach in identifying genetic risk loci for complex phenotypes and it represents an alternative and new avenue in addition to SNPs analysis.

Similarly in a large-scale association study involving 837 late-onset Alzheimer's disease cases and 550 controls, one ROH on chromosome 8 was identified, and three of the genes (*STAR, EIF4EBP1* and *ADRB3*) in the region are biologically plausible candidates (Nalls et al. 2009a). Success was also achieved for complex quantitative traits such as height (Yang et al. 2010b), where strong statistical evidence showing association of one ROH with height was obtained in a total sample size of >10,000 in both the genome-wide discovery and replication studies. The height of individuals with the particular ROH was significantly higher (increased by 3.5 cm) than the individuals without the region. The identification of this ROH added further support to the contribution of recessive loci to adult height variation (Kimura et al. 2008; Xu et al. 2002). Nonetheless, other studies produced negative results, as no evidence of homozygosity was found for bipolar disorder (Vine et al. 2009).

To date, the results showing the association between homozygosity with various cancers are also controversial (Hosking et al. 2010; Assié et al. 2008; Enciso-Mora et al. 2010). For example, two studies investigating the homozygosity in colorectal cancers derived an opposing conclusion which is likely due to the differences between the two studies such as the sample sizes, the density of genotyping platforms and the analysis (Bacolod et al. 2008; Spain et al. 2009). Although studies have found statistically negative results after imposing the stringent Bonferroni correction for multiple-testing, a number of ROHs warrant further investigation as these regions overlapped with biologically plausible genes for the phenotypes. One ROH was found to encompass the gene encoding erythropoietin receptor (*EPOR*) protein. Over-expression of this protein has been documented in acute lymphoblastic leukemia (Hosking et al. 2010).

Many reasons can be speculated for the inconsistencies as to why associations of ROHs were only found in some diseases or studies but not others. This could also indicate that the effects of homozygosity on the risk of complex phenotypes may be disease or trait-dependent, for example some quantitative traits have shown significant variance due to recessive alleles such as systolic blood pressure, total cholesterol and low-density lipoprotein cholesterol. This implies that the effects of homozygosity may be greater in influencing the variation of these traits than others (Campbell et al. 2009). On the other hand, it could also be population-dependent since differences in homozygosity between populations have been documented. Although a number of genome-wide homozygosity association studies have been performed, the optimum study design or analysis methods for assessing the associations or effects of ROHs on the disease risk has not yet been well established. This is, however, vital before breakthrough discoveries can be made in this research area.

The idea for using the homozygosity association approach to dissect the genetics of complex phenotypes is

to reveal the recessive loci that only express their effects (or increase the risk of complex diseases) in the presence of two deleterious recessive alleles, in a recessive disease model. In addition to autosomal recessive disorders, complex diseases can also be affected by recessive variants. The conventional single-SNP analysis approach applied in GWAS may not be statistically powerful enough to identify recessive alleles with small effect sizes and moreover, the recessive model is not usually tested. Until the effect of homozygosity on complex phenotypes is better understood, it is premature to make any conclusions, as the field is still in its infancy compared to association studies between SNPs and CNVs for complex diseases and traits. However, collectively these studies have demonstrated the feasibility of using the homozygosity association approach to identify susceptibility loci for complex phenotypes and have produced encouraging results. This also further underscores the need to further investigate and catalog the extent of ROHs in different populations. Similar to the other genetic variations, ROHs have the potential of becoming the genetic markers in GWAS. In fact, homozygosity mapping has been commonly used to identify the loci for recessive diseases in consanguineous families.

Strengths and shortcomings of genome-wide homozygosity association studies

From the statistical analysis point of view, the advantage of the genome-wide homozygosity association approach is that it suffers lesser penalty from Bonferroni correction for multiple-testing as significantly fewer ROHs are involved compared to the number of SNPs tested in GWAS. Thus, it needs a less stringent *p* value cutoff to declare genome-wide significance. Thus, the genome-wide ROHs association approach has a higher statistical power or requires a fewer number of samples in the studies than the 'conventional GWAS'.

GWAS is an indirect approach that relies on LD to identify the causal variants, thus the results from GWAS are pinpointing genetic loci rather than revealing the causal variants directly (Wang et al. 2005; Hirschhorn and Daly 2005). Similarly in genome-wide homozygosity association studies, one or more ROHs are identified as susceptibility risk loci rather than revealing the actual recessive variants causing the disease. For example, the homozygous consensus region in chromosome 8 was found to be associated with late-onset Alzheimer disease contains seven genes. However, the number of recessive variants within these genes or this region responsible for this 'statistical association signal' and which are functionally important in causing the diseases is unknown (Nalls et al. 2009a). The approaches to be taken from identifying the disease or trait-associated ROHs to locating the functional recessive variants is also unclear. Moreover, the sizes of ROHs are many folds larger than the LD blocks detected by conventional SNP analysis in GWAS, thus making the fine mapping of recessive variants harder. Therefore, the genome-wide association of ROHs, at best, can only pinpoint to a relatively large region harboring as yet to be identified recessive variants.

One common issue and problem in case–control association studies of CNVs and ROHs is how to construct the common CNV and ROH regions in the first place. This step is required to group the individual CNVs or ROHs into a common and discrete region. Similar to CNVs, it is unclear how to partition the individual ROHs into ROH groups so that the frequencies can be used for association analysis. This represents an important analytical challenge in these studies. Genome-wide studies investigating the association of common CNVs with complex phenotypes have so far yielded limited successes (Wellcome Trust Case Control Consortium 2010). As for ROHs, different studies have used their own methods to define ROH groups as no standardized criteria are available. Alternatively this step can be easily performed as the individual ROHs can be divided into different ROH groups by using the 'homozyg-group' command in the 'Runs of Homozygosity' program in PLINK. As a result, each ROH group is actually the overlapping region among all the individual ROHs in the group i.e. the consensus region (the region shared by all overlapping ROHs) (Fig. 2). Using this approach, Yang et al. (2010b) identified 3,322 ROH groups containing more than 50 individual ROHs. While Nalls et al. (2009a) identified 1,090 consensus regions from overlapping ROHs, but each consensus region was found in 10 or more individuals.

Besides identifying the ROH groups for association analysis, attempts were also made to compute other parameters such as the total length of the genome comprised by ROHs (the sum of the length of all ROHs), average length of each ROH (the total length divided by the number of ROHs) and the number of ROHs per individual and

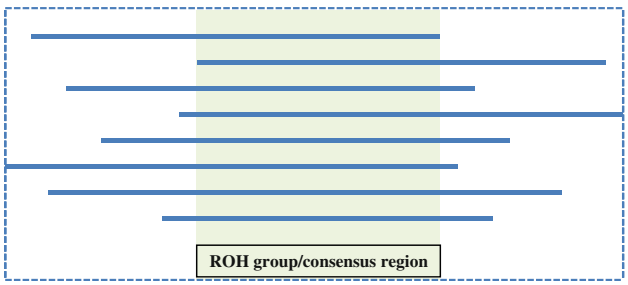


Fig. 2 Schematic diagram illustrating the ROH group or consensus region (*shadowed rectangle*) of several individual ROHs (*blue line*). Only 8 individual ROHs are shown for illustrative purposes with each individual ROH extending in both directions from the consensus region

compare these parameters between cases and controls. Nonetheless, no significant result was observed for late-onset Alzheimer disease (Nalls et al. 2009a). Likewise, no significant difference was found in the average number of ROHs between acute lymphoblastic leukemia, breast and prostate cancers with their controls (Hosking et al. 2010; Enciso-Mora et al. 2010).These analyses may not be very fruitful and have a limited interpretation. Even though significant results were obtained for all the three parameters, the findings are not informative in pointing to specific ROHs that are important to the disease. It can only be concluded that the overall extent of homozygosity is significantly greater in cases compared to controls and thus some recessive variants may be predisposed to the disease risk.

Conclusions

Published data have conclusively demonstrated the high frequency of ROHs in the genomes of outbred populations, and previous studies have also successfully unraveled the associations between ROHs and several complex phenotypes such as schizophrenia, late onset Alzheimer's diseases and height. These studies have shown the promise of the homozygosity association approach in identifying recessive loci for complex phenotypes. However, to what extent this approach contributes toward dissecting the genetics of complex phenotypes is yet to be determined. The analysis of ROHs is now feasible and convenient given the readily available high-density SNPs genotype data and the powerful detection tools such as the PLINK and PennCNV algorithms. Cataloging ROHs in different populations is important, as it lays the foundation for exploring the recessive variants for complex phenotypes.

Currently, the results from GWAS focusing on SNPs analysis alone, explains only a small fraction of the heritability of complex phenotypes (Manolio et al. 2009). Several reasons accounting for the missing heritability have been postulated (Eichler et al. 2010). The missing heritability has challenged the validity of the common-disease common variant (CD/CV) hypothesis (Schork et al. 2009), and has also diverted the research focus to rare variants (Bodmer and Bonilla 2008; Gorlov et al. 2008; Dickson et al. 2010). However, more recent studies have shown that common variants, or more specifically common SNPs, can explain a greater proportion of the heritability than what has been accounted for by GWAS done to date. These SNPs, however, are hidden within the GWAS data, and require larger sample sizes to be discovered (Yang et al. 2010a; Park et al. 2010b). The homozygosity association approach will offer an additional avenue to discovering genetic risk loci that may be missed by the conventional SNPs analysis in GWAS. The homozygosity analysis can be 'easily' performed using the SNPs genotype data and the available detection algorithms, and this is also in line with the ethos of maximizing the information from the GWAS dataset. However several issues and problems still remain as has been discussed.

The power of the homozygosity mapping approach in identifying genes and mutations for autosomal recessive disorders has been previously shown, but currently available data is limited in order to evaluate the success of this approach when applied to complex phenotypes. Hence more studies are needed in the future. Finally we advocate the use of the homozygosity association approach as an additional method of identifying loci harboring recessive variants for complex diseases and traits, which may have been undetected when conventional SNPs analysis was performed alone. The success of this approach has been demonstrated in several complex phenotypes applying the approach. The results so far are encouraging enough to warrant further studies on ROHs to investigate their impacts on complex phenotypes.

Cataloging the ROHs in human genomes and investigating their associations with complex phenotypes should build on the existing GWAS data and these are important areas to pursue in future. The contribution and the role of ROHs in complex phenotypes have been considerably neglected in GWAS; therefore we encourage researchers to explore the associations of ROHs with various phenotypes using their existing SNP data. As the high-density SNPs genotype data have already been generated by several hundred GWAS, the studies of ROHs should be relatively uncomplicated. The availability of these SNP datasets will facilitate the assessment of the roles that ROHs have in complex phenotypes.

References

Abu Safieh L, Aldahmesh MA, Shamseldin H, Hashem M, Shaheen R, Alkuraya H, Al Hazzaa SA, Al-Rajhi A, Alkuraya FS (2010) Clinical and molecular characterisation of Bardet-Biedl syndrome in consanguineous populations: the power of homozygosity mapping. J Med Genet 47:236–241

Abu-Amero S, Monk D, Frost J, Preece M, Stanier P, Moore GE (2008) The genetic aetiology of Silver–Russell syndrome. J Med Genet 45:193–199

Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322:881–888

Assié G, LaFramboise T, Platzer P, Eng C (2008) Frequency of germline genomic homozygosity associated with cancer cases. JAMA 299:1437–1445

Bacolod MD, Schemmann GS, Wang S, Shattock R, Giardina SF, Zeng Z, Shia J, Stengel RF, Gerry N, Hoh J, Kirchhoff T, Gold B, Christman MF, Offit K, Gerald WL, Notterman DA, Ott J, Paty PB, Barany F (2008) The signatures of autozygosity among patients with colorectal cancer. Cancer Res 68:2610–2621

Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59

Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40:695–701

Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 65:1493–1500

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86:526–539

Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, Barac L, Pericic M, Janicijevic B, Smolej-Narancic N, Polasek O, Kolcic I, Weber JL, Hastie ND, Rudan P, Wright AF (2007) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. Hum Mol Genet 16:233–241

Campbell H, Rudan I, Bittles AH, Wright AF (2009) Human population structure, genome autozygosity and human health. Genome Med 1:91

Carson AR, Feuk L, Mohammed M, Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. Hum Genomics 2:403–414

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16–S21

Collin RW, Safieh C, Littink KW, Shalev SA, Garzozi HJ, Rizel L, Abbasi AH, Cremers FP, den Hollander AI, Klevering BJ, Ben-Yosef T (2010) Mutations in C2ORF71 cause autosomal-recessive retinitis pigmentosa. Am J Hum Genet 86:783–788

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712

Curtis D (2007) Extended homozygosity is not usually due to cytogenetic abnormality. BMC Genet 8:67

Curtis D, Vine AE, Knight J (2008) Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations. Ann Hum Genet 72:261–278

Day IN (2010) dbSNP in the detail and copy number complexities. Hum Mutat 31:2–4

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450

Enciso-Mora V, Hosking FJ, Houlston RS (2010) Risk of breast and prostate cancer is not associated with increased homozygosity in outbred populations. Eur J Hum Genet 18:909–914

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. Genome Res 16:949–961

Gibbs JR, Singleton A (2006) Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. PLoS Genet 2:e150

Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15:789–795

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–112

Gurrieri F, Accadia M (2009) Genetic imprinting: the paradigm of Prader–Willi and Angelman syndromes. Endocr Dev 14:20–28

Haberman Y, Amariglio N, Rechavi G, Eisenberg E (2008) Trinucleotide repeats are prevalent among cancer-related genes. Trends Genet 24:14–18

Hannan AJ (2010) Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet 26:59–65

Harville HM, Held S, Diaz-Font A, Davis EE, Diplas BH, Lewis RA, Borochowitz ZU, Zhou W, Chaki M, MacDonald J, Kayserili H, Beales PL, Katsanis N, Otto E, Hildebrandt F (2010) Identification of 11 novel mutations in eight BBS genes by high-resolution homozygosity mapping. J Med Genet 47:262–267

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

Hosking FJ, Papaemmanuil E, Sheridan E, Kinsey SE, Lightfoot T, Roman E, Irving JA, Allan JM, Taylor M, Tomlinson IP, Greaves M, Houlston RS (2010) Genome-wide homozygosity signatures and childhood acute lymphoblastic leukemia risk. Blood 115:4472–4477

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

International HapMap Consortium, Frazer KA, Ballinger DG et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

Iseri SU, Wyatt AW, Nürnberg G, Kluck C, Nürnberg P, Holder GE, Blair E, Salt A, Ragge NK (2010) Use of genome-wide SNP homozygosity mapping in small pedigrees to identify new mutations in VSX2 causing recessive microphthalmia and a semidominant inner retinal dystrophy. Hum Genet 128:51–60

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451:998–1003

Jiang H, Orr A, Guernsey DL, Robitaille J, Asselin G, Samuels ME, Dubé MP (2009) Application of homozygosity haplotype analysis to genetic mapping with high-density SNP genotype data. PLoS One 4:e5280

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

Kim JI, Ju YS, Park H et al (2009) A highly annotated whole genome sequence of a Korean individual. Nature 460:1011–1015

Kimura T, Kobayashi T, Munkhbat B, Oyungerel G, Bilegtsaikhan T, Anar D, Jambaldorj J, Munkhsaikhan S, Munkhtuvshin N, Hayashi H, Oka A, Inoue I, Inoko H (2008) Genome-wide association analysis with selective genotyping identifies candidate loci for adult height at 8q21.13 and 15q22.33–q23 in Mongolians. Hum Genet 123:655–660

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426

Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS (2010a) The discovery of human genetic variations and their use as disease markers: past, present and future. J Hum Genet 55:403–415

Ku CS, Pawitan Y, Sim X, Ong RT, Seielstad M, Lee EJ, Teo YY, Chia KS, Salim A (2010b) Genomic copy number variations in three Southeast Asian populations. Hum Mutat 31:851–857

Lapunzina P, Aglan M, Temtamy S, Caparrós-Martín JA, Valencia M, Letón R, Martínez-Glez V, Elhossini R, Amr K, Vilaboa N, Ruiz-Perez VL (2010) Identification of a frameshift mutation in Osterix in a patient with recessive osteogenesis imperfecta. Am J Hum Genet 87:110–114

Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci USA 104:19942–19947

Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, Hung SI, Chung WH, Pan WH, Lee MT, Tsai FJ, Chang CF, Wu JY, Chen YT (2006) Long contiguous stretches of homozygosity in the human genome. Hum Mutat 27:1115–1121

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40:1166–1174

McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Wright AF, Campbell H, Wilson JF (2008) Runs of homozygosity in European populations. Am J Hum Genet 83:359–372

Metzker ML (2010) Sequencing technologies—the next generation. Nat Rev Genet 11:31–46

Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Huqun, Kyo S, Okazaki Y, Hagiwara K (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. Am J Hum Genet 80:1090–1102

Nakamura Y (2009) DNA variations in human and medical genetics: 25 years of my experience. J Hum Genet 54:1–8

Nalls MA, Guerreiro RJ, Simon-Sanchez J, Bras JT, Traynor BJ, Gibbs JR, Launer L, Hardy J, Singleton AB (2009a) Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics 10:183–190

Nalls MA, Simon-Sanchez J, Gibbs JR, Paisan-Ruiz C, Bras JT, Tanaka T, Matarin M, Scholz S, Weitz C, Harris TB, Ferrucci L, Hardy J, Singleton AB (2009b) Measures of autozygosity in decline: globalization, urbanization, and its implications for medical genetics. PLoS Genet 5:e1000415

Nicolas E, Poitelon Y, Chouery E, Salem N, Levy N, Mégarbané A, Delague V (2010) CAMOS, a nonprogressive, autosomal recessive, congenital cerebellar ataxia, is caused by a mutant zinc-finger protein, ZNF592. Eur J Hum Genet 18:1107–1113

Nothnagel M, Lu TT, Kayser M, Krawczak M (2010) Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans. Hum Mol Genet 19:2927–2935

O'Dushlaine CT, Morris D, Moskvina V, Kirov G, Consortium IS, Gill M, Corvin A, Wilson JF, Cavalleri GL (2010) Population structure and genome-wide patterns of variation in Ireland and Britain. Eur J Hum Genet 18:1248–1254

Pang J, Zhang S, Yang P, Hawkins-Lee B, Zhong J, Zhang Y, Ochoa B, Agundez JA, Voelckel MA, Fisher RB, Gu W, Xiong WC, Mei L, She JX, Wang CY (2010) Loss-of-function mutations in HPSE2 cause the autosomal recessive urofacial syndrome. Am J Hum Genet 86:957–962

Park H, Kim JI, Ju YS, Gokcumen O, Mills RE, Kim S, Lee S, Suh D, Hong D, Kang HP, Yoo YJ, Shin JY, Kim HJ, Yavartanoo M, Chang YW, Ha JS, Chong W, Hwang GR, Darvishi K, Kim H, Yang SJ, Yang KS, Kim H, Hurles ME, Scherer SW, Carter NP, Tyler-Smith C, Lee C, Seo JS (2010a) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42:400–405

Park JH, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, Chatterjee N (2010b) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42:570–575

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16:1136–1148

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82:685–695

Polasek O, Hayward C, Bellenguez C, Vitart V, Kolcić I, McQuillan R, Saftić V, Gyllensten U, Wilson JF, Rudan I, Wright AF, Campbell H, Leutenegger AL (2010) Comparative assessment of methods for estimating individual genome-wide homozygosity-by-descent from human genomic data. BMC Genomics 11:139

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a toolset for whole genome association and population based linkage analyses. Am J Hum Genet 81:559–575

Ragoussis J (2009) Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet 10:117–133

Rudan I, Rudan D, Campbell H, Carothers A, Wright A, Smolej-Narancic N, Janicijevic B, Jin L, Chakraborty R, Deka R, Rudan P (2003a) Inbreeding and risk of late onset complex disease. J Med Genet 40:925–932

Rudan I, Smolej-Narancic N, Campbell H, Carothers A, Wright A, Janicijevic B, Rudan P (2003b) Inbreeding and the genetic complexity of human hypertension. Genetics 163:1011–1021

Rudan I, Campbell H, Carothers AD, Hastie ND, Wright AF (2006) Contribution of consanguinuity to polygenic and multifactorial diseases. Nat Genet 38:1224–1225

Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19:212–219

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P (2009) HomozygosityMapper—an interactive approach to homozygosity mapping. Nucleic Acids Res 37:593–599

Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Craw-

ley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet 16:1–14

Spain SL, Cazier JB, CORGI Consortium, Houlston R, Carvajal-Carmona L, Tomlinson I (2009) Colorectal cancer risk is not associated with increased levels of homozygosity in a population from the United Kingdom. Cancer Res 69:7422–7429

Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61:437–455

Teo YY, Sim X, Ong RT, Tan AK, Chen J, Tantoso E, Small KS, Ku CS, Lee EJ, Seielstad M, Chia KS (2009) Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res 19:2154–2162

Ting JC, Roberson ED, Miller ND, Lysholm-Bernacchi A, Stephan DA, Capone GT, Ruczinski I, Thomas GH, Pevsner J (2007) Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. Hum Mutat 28:1225–1235

Uz E, Alanay Y, Aktas D, Vargel I, Gucer S, Tuncbilek G, von Eggeling F, Yilmaz E, Deren O, Posorski N, Ozdag H, Liehr T, Balci S, Alikasifoglu M, Wollnik B, Akarsu NA (2010) Disruption of ALX1 causes extreme microphthalmia and severe facial clefting: expanding the spectrum of autosomal-recessive ALX-related frontonasal dysplasia. Am J Hum Genet 86:789–796

Van Buggenhout G, Fryns JP (2009) Angelman syndrome (AS, MIM 105830). Eur J Hum Genet 17:1367–1373

Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, Robinson M, Lawrence J, Anjorin A, Sklar P, Gurling HM, Curtis D (2009) No evidence for excess runs of homozygosity in bipolar disorder. Psychiatr Genet 19:165–170

Wain LV, Armour JA, Tobin MD (2009) Genomic copy number variation, human health, and disease. Lancet 374:340–350

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M (2010) Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet 87:90–94

Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6:109–118

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M (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

Wang J, Wang W, Li R et al (2008) The diploid genome sequence of an Asian individual. Nature 456:60–65

Wang S, Haynes C, Barany F, Ott J (2009) Genome-wide autozygosity mapping in human populations. Genet Epidemiol 33:172–180

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME et al (2010) Genome-wide association study of CNVs in 16, 000 cases of eight common diseases and 3, 000 shared controls. Nature 464:713–720

Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, Stern R, Raymond FL, Sandford R, Malik Sharif S, Karbani G, Ahmed M, Bond J, Clayton D, Inglehearn CF (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. Am J Hum Genet 78:889–896

Xu J, Bleecker ER, Jongepier H, Howard TD, Koppelman GH, Postma DS, Meyers DA (2002) Major recessive gene(s) with considerable residual polygenic effect regulating adult height: confirmation of genomewide scan results for chromosomes 6, 9, and 12. Am J Hum Genet 71:646–650

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010a) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yang TL, Guo Y, Zhang LS, Tian Q, Yan H, Papasian CJ, Recker RR, Deng HW (2010b) Runs of homozygosity identify a recessive locus 12q21.31 for human adult height. J Clin Endocrinol Metab 95:3777–3782

Yim SH, Kim TM, Hu HJ, Kim JH, Kim BJ, Lee JY, Han BG, Shin SH, Jung SH, Chung YJ (2010) Copy number variations in East-Asian population and their evolutionary and functional implications. Hum Mol Genet 19:1001–1008

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19:1586–1592

Characterising Structural Variation by Means of Next-Generation Sequencing

Chee Seng Ku, *National University of Singapore, Singapore*
Nasheen Naidoo, *National University of Singapore, Singapore*
Shu Mei Teo, *National University of Singapore, Singapore*
Yudi Pawitan, *Karolinska Institutet, Stockholm, Sweden*

Advanced article

Article Contents

Online posting date: 15th February 2011

A new era of copy number variants (CNVs) discovery began when two separate studies, published concurrently in 2004, identified several hundred deletions and duplications in the human genome. Over the past several years, most of the CNV data were generated by microarrays. These methods have several shortcomings, such as the inability to detect copy-neutral variants (e.g. inversions and translocations), limited sensitivity to detect smaller CNVs and poor resolution in determining CNV breakpoints especially with lower resolution microarrays. A paradigm shift in the discovery of copy-neutral variants was attributed to the development of a sequencing-based method known as paired-end mapping. This method was first demonstrated to be powerful in detecting structural variants using next-generation sequencing technologies in 2007. Further studies have also leveraged an important feature of sequencing data, where several hundred million short sequence reads are produced by next-generation sequencers, to detect CNVs based on the abundance or density of the sequence reads aligned to a reference genome. This approach is known as depth-of-coverage. These emerging sequencing-based methods will continue playing an important role in the discovery of structural variants until *de novo* genome assembly becomes more feasible.

ELS subject area: Genetics and Disease

How to cite:
Ku, Chee Seng; Naidoo, Nasheen; Teo, Shu Mei; and Pawitan, Yudi (February 2011) Characterising Structural Variation by Means of Next-Generation Sequencing. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0023399

Introduction

A new era of *copy number variants (CNVs)* discovery began when two separate studies, published concurrently in 2004, identified several hundred deletions and duplications in the human genome (Sebat *et al*., 2004; Iafrate *et al*., 2004). However, these genetic abnormalities were documented decades ago in clinical cytogenetics studies and found to cause various genomic or cytogenetic disorders (Lee *et al*., 2007). The distinguishing feature of the recent studies were that these CNVs were more prevalent in the human genome than expected. These changes in copies number also did not result in any apparent phenotype or disorder and these regions of variable copies were found in the genomes of phenotypically normal individuals (Sebat *et al*., 2004; Iafrate *et al*., 2004). As these submicroscopic (<3–5 Mb) deletions and duplications are beyond the detection limit of traditional cytogenetics tools such as molecular fluorescence *in situ* hybridisation (FISH), these recent discoveries can be credited to the use of whole genome microarray technologies (Carter, 2007). **See also**: Copy Number Variation in the Human Genome; Genetic Variation: Human; Relevance of Copy Number Variation to Human Genetic Disease

Whole Genome Microarray and Sequencing Technologies and Their Progress

The early whole genome microarray studies discovered several hundred CNVs (Sebat *et al*., 2004; Iafrate *et al*., 2004), for example, Sebat *et al*. (2004) detected a total of 221 CNVs in 20 individuals with an average CNV length of 465 Kb. However, it was widely believed that the number of CNVs detected is likely to be underestimated. These

studies used 'low-resolution' microarrays such as ROMA (representational oligonucleotide microarray analysis) containing 85 000 probes with a resolution of approximately one probe for every 35 Kb (Sebat *et al*., 2004) and the BAC-CGH (bacterial artificial chromosome-comparative genomic hybridisation) array with a resolution of approximately one probe for every 1 Mb (Iafrate *et al*., 2004). Furthermore, these studies investigated a small sample size of only tens of individuals which limits the detection of less common CNVs. CNVs smaller than 50–100 Kb will also not be detected as their size is below the resolution limits of these microarrays. Thus, both the sample size and the resolution of microarray are critical factors in determining the discovery of less common and smaller CNVs.

A later study by Tuzun *et al*. (2005) showed that approximately 85% of the 297 identified *structural variants* (139 insertions, 102 deletions and 56 inversions) were not detected by earlier studies. However, this study used a *sequencing*-based method, where the fosmid paired-end sequences were sequenced, instead of microarrays. Many of the structural variants that are being identified using this sequencing-based method are beyond the resolution limit of ROMA and the BAC-CGH microarrays. Inversions are also undetected by microarrays (Tuzun *et al*., 2005; Sebat *et al*., 2004; Iafrate *et al*., 2004). The discovery of many novel structural variants is likely due to the difference between the resolution of sequencing- and microarray-based methods in detecting structural variants.

The contribution of CNVs as a significant source of genetic variation in human populations has since been appreciated despite the limitations using microarrays. This is evident from the enormous amount of interest and efforts generated towards mapping CNVs in different populations (Redon *et al*., 2006; Zogopoulos *et al*., 2007; Wong *et al*., 2007). The first comprehensive mapping of CNVs in the 270 samples from the International HapMap I Project was completed in 2006 (Redon *et al*., 2006). 'Human Genetic Variation' was then recognised as the 'Breakthrough of The Year' in 2007 by the journal *Science*. This was partly accomplished due to the significant progress made in the research of CNVs in addition to the numerous single nucleotide polymorphisms (SNPs) identified by genome-wide association studies for complex phenotypes (Pennisi, 2007). The limitations of ROMA and the BAC-CGH arrays have been overcome in later studies by using higher resolution microarrays and larger sample sizes of several hundred samples (McCarroll *et al*., 2008; Matsuzaki *et al*., 2009; Conrad *et al*., 2010; Park *et al*., 2010; Yim *et al*., 2010; Ku *et al*., 2010). For example, a set of 20 high-resolution oligonucleotide-CGH microarrays comprised of 42 million probes with a median spacing of 56 bases was designed and used by Conrad *et al*. (2010) in mapping CNVs in the HapMap samples (Conrad *et al*., 2010). Other studies have also used the highest resolution SNP microarrays that are commercially available such as the Affymetrix SNP Array 6.0 and the Illumina Human 1M BeadChip (McCarroll *et al*., 2008; Ku *et al*., 2010).

Other types of chromosomal rearrangements, particularly inversions and balanced translocations, have received relatively less attention (Feuk *et al*., 2006; Feuk, 2010; Stankiewicz and Lupski, 2010). Inversions and translocations are also known as 'copy-neutral variants' or 'balanced chromosomal rearrangements' and do not involve changes in copies number (or losses or gains of deoxyribonucleic acid (DNA) sequences). Collectively these copy number and copy-neutral variants are broadly classified as 'structural variants'. The genome-wide mapping or detection of CNVs in different populations has advanced considerably since 2004 and was driven mainly by high-resolution microarray technologies such as oligonucleotide-CGH and SNP microarrays. In contrast, the pace in identifying inversions and translocations in the human genome has been slower as more powerful and effective methods were not available until the advent of *next-generation sequencing (NGS) technologies* (Mardis, 2008; Shendure and Ji, 2008; Metzker, 2010).

Although sequencing-based methods such as *paired-end mapping (PEM)*, which uses cloning and Sanger sequencing methods to sequence the fosmid paired-end sequences, have been shown to be powerful in identifying copy-neutral variants, this method is laborious and expensive (Tuzun *et al*., 2005). Even with the arrival of NGS technologies, PEM has still not as yet been applied in population-based studies (Korbel *et al*., 2007), as opposed to microarrays which are commonly applied to several hundred or thousand samples for CNV detection. However, it is foreseeable in the near future that sequencing-based methods will eventually be routinely and widely applied in large-scale population-based studies when the cost of sequencing becomes more affordable and the challenges in the analysis have been addressed.

The mechanisms that generate structural variants such as nonallelic homologous recombination and non-homologous end joining are beyond the scope of this article (Hastings *et al*., 2009). Similarly, genome-wide detection of CNVs in population-based studies and the population characteristics of CNVs or structural variants, and their associations with various complex diseases or genomic disorders have been reviewed extensively in several excellent review papers (Conrad and Hurles, 2007; McCarroll and Altshuler, 2007). This article will focus on the new and emerging research on structural variants using high-throughput sequencing technologies (Mardis, 2008; Shendure and Ji, 2008; Metzker, 2010; Schadt *et al*., 2010; Gupta, 2008). We also discuss the relative strengths and weaknesses of sequencing-based approaches in comparison to microarrays, and elucidate the potential approaches for a more comprehensive and thorough detection of structural variants in the human genome before *de novo* genome assembly becomes more practical (Li *et al*., 2010a, b; Paszkiewicz and Studholme, 2010).

Microarray-based Methods

Over the past few years, most of the CNV data were generated by CGH and SNP microarrays where fluorescence

signal intensity information was used to detect deletions and duplications. These microarrays are highly accessible and affordable for population-based studies. Additionally, the analysis methods and tools for detecting CNVs using microarray data have been well-developed (Wang *et al*., 2007; Korn *et al*., 2008). This has enabled studies of population characteristics of CNVs in many different populations (McCarroll *et al*., 2008; Matsuzaki *et al*., 2009; Yim *et al*., 2010; Ku *et al*., 2010). However, because of the reliance on the relative or difference in signal intensity compared to a reference in inferring regions with copy number changes, this has hindered microarrays from detecting copy-neutral variants (Carter, 2007). Furthermore, due to the limitations in marker density or resolution of microarrays used in the previous studies, these methods had poor sensitivity to detecting smaller CNVs (<50 Kb) (Redon *et al*., 2006). However, the ability to detect smaller CNVs is critical as they are known to be more numerous than the larger CNVs (Estivill and Armengol, 2007). The accuracy in determining the sizes or breakpoints of CNVs is highly dependent on the resolution of the microarrays as the sizes of CNVs found by previous studies were frequently over-estimated. It is notable that 88% of 1153 CNV loci were smaller than sizes reported in the Database of Genomic Variants and that a reduction of $>50\%$ in size was observed for 76% of the CNV loci (Perry *et al*., 2008).

The latest developments in SNP microarrays such as an increase in marker density and uniformity of distribution in the genome and copy number probes to cover regions with sparse SNPs have improved the sensitivity of microarrays. Nonetheless, these SNP microarrays still lack the sensitivity to detect CNVs smaller than 5–10 Kb even with use of the highest resolution microarrays such as the Illumina Human 1M Beadchip and the Affymetrix SNP Array 6.0 (McCarroll *et al*., 2008; Cooper *et al*., 2008). Although designing a set of high-resolution CGH microarrays comprising tens of millions of probes offers an unprecedented resolution, this method is more costly for several hundred samples (Conrad *et al*., 2010). However, these improvements in microarrays are still unable to detect copy-neutral variants. Thus, developments of other methods that can overcome the limitations of microarrays and simultaneously detect both CNVs and copy-neutral variants are needed.

Sequencing-based Methods

Several previous studies have used sequencing data to detect structural variants. For example, a study by Feuk *et al*. (2005) discovered regions that are inverted between the chimpanzee and human genomes by performing a comparative analysis of their DNA sequence assemblies. This study identified approximately 1600 putative regions of inverted orientation in the genomes (Feuk *et al*., 2005), whereas Khaja *et al*. (2006) identified various types of genetic variants, including structural variants, through comparison of two human assemblies (Khaja *et al*., 2006). However, the paradigm shift in the discovery of copy-neutral variants was attributed to the development of the PEM and concurrent advances in NGS technologies (Korbel *et al*., 2007). The PEM method has also contributed greatly to the discovery of CNVs in the human genome (Wang *et al*., 2008; Ahn *et al*., 2009). **See also**: Comparing the Human and Chimpanzee Genomes; Human Genome Project: Importance in Clinical Genetics; Sequencing the Human Genome: Novel Insights into its Structure and Function

Further studies have also leveraged on an important feature of sequencing data generated by NGS technologies where several hundred million short sequence reads are produced per instrument run to detect CNVs. It is based on the abundance or density of the sequence reads aligned to the reference genome. This approach is known as *depth-of-coverage (DOC)* and is similar to microarray-based methods in that it is also unable to detect copy-neutral variants (Yoon *et al*., 2009). Although *de novo* genome assembly is still developing, the established PEM and DOC methods will continue to play important roles in identifying new structural variants. **Table 1** shows the comparison between microarrays and sequencing-based methods for detecting structural variants.

Paired-end Mapping

Principle

In the PEM method, a library of DNA fragments with a fixed insert size is prepared and both ends of the DNA fragments are sequenced to generate 'paired-end sequences' (the sequences at both ends of the DNA fragments). This sequence information is then aligned against the reference genome. The underlying principle of PEM to detect structural variants is reliant on the discrepancy or discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome to infer 'simple' deletion, insertion and inversion. The use of the term 'simple' is to distinguish from other more complex structural variants such as 'everted duplication', 'linked insertion' and 'hanging insertion'. Thus, the terms deletion, insertion and inversion used throughout this paper refer to the 'simple' types unless otherwise specified (Tuzun *et al*., 2005; Korbel *et al*., 2007).

When paired-end sequences that are being aligned to the reference sequence display discordance from the expected insert size or distance, this is an indication of deletion and insertion, whereas discordance in orientation suggests the presence of inversion (i.e. paired-end sequences are incorrectly oriented comparing to the reference genome). Since the insert size of the DNA fragment library is known, when paired-end sequences that align to the reference are substantially shorter than expected, this indicates the presence of insertion. Conversely, a longer than the expected insert size suggests the presence of deletion while other more complicated patterns of discordance when aligning the

Table 1 Comparison between microarrays and sequencing-based methods for detecting structural variants

	Microarrays[a]	PEM[b]	DOC
Principle	Based on the relative or difference in florescence signal intensity compared to a reference (one sample or a set of samples) to infer CNVs	Based on the discrepancy or discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome to infer 'simple' deletion, insertion and inversion	Based on the density of sequence reads being aligned to the reference genome to infer CNVs
Ability to detect CNVs	Yes	Yes	Yes
Ability to detect copy-neutral variants	No	Yes	No
Reliably detecting CNVs	Multiple or tens of probes	Multiple discordant pairs	A high density of sequence reads
Application to population-based studies	Commonly applied to several hundred or thousand samples	Has not yet been applied	Has not yet been applied
Sensitivity to detect smaller CNVs e.g. <10 Kb	Generally poor, but depends on the resolution of the microarrays, e.g. a set of oligonucleotide CGH arrays containing 42 million probes has provided an unprecedented resolution	Yes, preparation of several libraries of different insert sizes are able to detect insertions and deletions of varying sizes, but the detection of insertions is limited by the insert sizes	It may not be powerful enough to detect smaller CNVs (related to the strength of DOC signatures and the coverage of the sequencing data or the number of sequence reads)
Sensitivity to detect larger CNVs	Yes, even low resolution BAC clone CGH arrays (with a resolution of approximately one probe for every 1 Mb) have been used to detect CNVs of several hundred kilobases to megabases	Yes, however, the detection of insertions is limited by the insert sizes, thus preparation of fosmid or BAC clone libraries with larger insert sizes are needed for detecting larger insertions	Yes, the DOC signatures will be stronger for larger CNVs
Precision in mapping breakpoints	Generally poor, however, it can be improved by increasing the resolution of microarrays	Good, theoretically the breakpoints can be mapped to a single nucleotide resolution	The precision to map the breakpoints can be improved by increasing the density or coverage of sequence reads
Role in 'discovery' and 'genotyping'	Can be used as an effective method to genotype newly discovered and known CNVs in population-based studies	Powerful for discovery of new structural variants	Discovery of CNVs especially in regions such as segmental duplications where PEM is less effective
Weakness as a result of technology limitation	Generally have poor signal-to-noise ratios for oligonucleotide-CGH and SNP microarrays compared to BAC clone CGH arrays	Short sequence reads are less specific in aligning uniquely to the reference genome especially in segmental duplications	Sequencing biases may lead to certain regions of the genome being over or under-sampled resulting in spurious DOC signatures

(*Continued*)

Table 1 Continued

	Microarrays	PEM	DOC
Scalability of sample throughput by technology	High sample throughput, for example, several hundred samples can be genotyped by SNP arrays per week as evident in genome-wide association studies	Tens of gigabases of sequencing data can be produced per instrument run in several days by NGS technologies, and the sample throughput can be scaled up by 'barcoding' i.e. labelling the samples by barcodes	Tens of gigabases of sequencing data can be produced per instrument run in several days by NGS technologies, and the sample throughput can be scaled up by 'barcoding' i.e. labelling the samples by barcodes
Level of analytical and computational challenges	Lesser, analytical methods for detecting CNVs using microarray data are well-developed	Greater, an emerging and maturing method leveraging on the large amount of NGS data	Greater, an emerging and maturing method leveraging on the large amount of NGS data
Difficulty in sample preparation	Easier in processing the samples for hybridisation on the microarrays	More challenging in preparing sequencing libraries especially clone-based libraries	More challenging in preparing sequencing libraries

[a]Whole genome oligonucleotide-CGH and SNP microarrays.
[b]Paired-end and mate-pair libraries and clone-based libraries (such as fosmid and BAC clones) for PEM.

paired-end sequences provide hints at more complex rearrangements or structural variants (Tuzun *et al*., 2005; Korbel *et al*., 2007; Medvedev *et al*., 2009).

As such, the paired-end sequences are usually classified as 'concordant pairs' or 'discordant pairs' and only the discordant pairs are informative for inferring structural variants. The presence of both concordant and discordant pairs spanning a locus suggests a heterozygote state with respect to the structural variant, for example a deletion occurs only in one homologous chromosome. In addition, usually multiple paired-end sequences are needed to reliably infer if a locus is harbouring a structural variant. The requirement of multiple paired-end sequences spanning a locus to detect structural variants will reduce the number of false-positive signals. It will also minimise the false-negative rate, for example, a heterozygous deletion will be missed by the presence of one concordant pair. However, with multiple paired-end sequences, it is more likely that both the concordant pair and the discordant pair will be observed to detect the heterozygous deletion. As a result, a sufficient amount of sequencing is needed to ensure that there are multiple paired-end sequences spanning across the genome. This also means that a substantial amount of sequencing is needed for the PEM method and thus this method will be more costly using Sanger sequencing compared to NGS technologies (Tuzun *et al*., 2005; Korbel *et al*., 2007; Medvedev *et al*., 2009).

The detection of structural variants using PEM 'signatures' depends on the clustering strategies and criteria used in the analysis, and the results can be varied for the same dataset by applying different strategies and criteria. 'Clustering' refers to steps to group PEM signatures (e.g. several discordant pairs) that support the presence of a structural variant into clusters. As such, clustering will improve reliability in inferring or predicting structural variants and also increase the precision in estimating breakpoints or the sizes of structural variants. The important criteria to be determined in clustering are (a) the minimum number of discordant pairs for a cluster and (b) the number of standard deviations of the insert size to distinguish between concordant and discordant pairs. The strategies and criteria used will then affect the sensitivity and specificity in detecting structural variants (Tuzun *et al*., 2005; Korbel *et al*., 2007; Medvedev *et al*., 2009).

Physical coverage and mate-pair library

'Physical coverage' is important in detecting structural variants using PEM. Physical coverage measures the number of fragments spanning a site and this affects the ability to detect structural variants. It is different from 'sequence coverage' which measures the number of sequence reads that cover a site and this sequence coverage affects the ability to detect single nucleotide variants or point mutations. Thus, physical coverage can be increased by creating a library of larger insert sizes. When preparing a 'shotgun library' using standard methods, the sizes of DNA fragments are usually several hundred bases, with approximately tens of bases on both ends of the DNA fragments sequenced using NGS technologies (Meyerson *et al*., 2010).

However, the insert size can be increased to several kilobases by creating a 'jumping library' or a '*mate-pair library*'. Additional steps are involved in preparing a mate-pair library in comparison to a paired-end library, where both ends of the DNA fragments of several kilobases (e.g.

3 Kb in the Korbel *et al*. (2007) study) were first ligated with biotinylated hairpin adapters. The DNA fragments were then circularised and randomly sheared. The fragments attached to biotinylated hairpin adapters were isolated to form a mate-pair library and then followed by sequencing (Korbel *et al*., 2007). Mate-pair library construction enables sequencing at both ends of longer DNA fragments of several kilobases. The mate-pair library with a larger insert size will increase the physical coverage of the genome. For example, by sequencing 50 bases from both ends of the DNA fragments from a library with a 3-Kb insert size, the physical coverage of the genome is 10-fold higher than that from a library with a 300-bp insert size. However, the sequence coverage is similar between both libraries as only 50 bases of paired-end sequences were generated with regards to the library insert size (Meyerson *et al*., 2010).

Thus the paired-end and mate-pair libraries differ only in the steps of constructing these libraries, as the sequencing and aligning of the paired-end sequences to the reference to detect structural variants follow the same principle. Although creating a mate-pair library increases physical coverage, a larger insert size is less sensitive in detecting smaller structural variants because of the difficulty in tightly controlling the sizes of the DNA fragments in the library. Therefore, depending on the 'tightness' or 'narrowness' of the distribution pattern (standard deviation) of the insert sizes in the library, it can be difficult to distinguish a true PEM signature caused by a small indel (i.e. indel of several or tens of bases) because of the variance in insert sizes in the library. This is because it is not practically possible to generate an exact similar size for each of the DNA fragments when preparing a library (Medvedev *et al*., 2009).

Strengths and weaknesses

In comparison to microarray-based methods, PEM has a higher sensitivity to detect smaller CNVs in addition to identifying copy-neutral variants, and it also has a greater precision in determining the breakpoints or boundaries of structural variants. For example, the PEM method has been applied in a number of whole genome resequencing studies where several thousand structural variants were detected (Wang *et al*., 2008; Ahn *et al*., 2009). Wang *et al*. (2008) identified a total of 2682 structural variants (the majority were CNVs) in the Han Chinese Yan Huang (YH) genome with a median length of approximately half a kilobase. These sizes are much smaller than those identified by microarrays ranging from tens to hundreds of kilobases depending on their resolution (Redon *et al*., 2006; Zogopoulos *et al*., 2007; Wong *et al*., 2007). This has clearly shown the greater sensitivity of PEM to detect smaller structural variants.

Nonetheless, this method could be biased against detection of duplications or insertions. This has been clearly shown in the YH genome, where most of the identified CNVs are deletions, namely 2441 deletions compared to 33 duplications. This is because PEM is unable to detect insertions larger than the insert size of the library. This also reveals the major limitation of PEM with a fixed insert size in detecting insertions (Wang *et al*., 2008). Deletions are easier to be detected because they are identified by a longer than expected insert size when aligned to the reference, whereas detection of insertions is restricted by the insert size. This means that insertions larger than the insert size are beyond the detection range. Therefore, several paired-end and mate-pair libraries with short and long insert sizes will be needed to capture structural variants of varying sizes. This will also nevertheless increase the sequencing costs several fold depending on the number of libraries. For the YH genome, the two paired-end libraries had a small insert size of 135 and 440 bp (Wang *et al*., 2008). Since the bias against detection of insertions is partly due to the small insert size, larger insert sizes of several kilobases should improve the ability to detect more insertions. Indeed, this has been demonstrated by Korbel *et al*. (2007) who prepared libraries of 3 Kb insert size for two individuals and found 1297 structural variants, including 853 deletions, 322 insertions and 122 inversions (Korbel *et al*., 2007). Although the number of deletions is still higher than insertions, it is significantly less biased compared to the numbers detected by Wang *et al*. (2008).

Human Genome Structural Variation Working Group

The PEM method to detect structural variants was first demonstrated by Tuzun *et al*. in 2005 by mapping paired-end sequences data from a human fosmid DNA genomic library. The average insert size of a fosmid library is approximately 40 Kb. However, sequencing of fosmid clones is laborious and costly using Sanger sequencing (Tuzun *et al*., 2005). These limitations have been overcome by NGS technologies which directly sequence the paired-end or mate-pair libraries without the need for cloning steps (Korbel *et al*., 2007). Both of these studies applied the PEM approach to investigate structural variants in the same sample (NA15510) from the International HapMap Project. However, their library insert sizes differed and this has enabled a comparison of the sensitivity between these studies. Korbel *et al*. (2007) were able to confirm 41% of all deletion and inversion events detected by fosmid paired-end sequencing. Additionally, they identified an additional 407 structural variants in NA15510 that had not been previously detected by fosmid paired-end sequencing (Korbel *et al*., 2007; Tuzun *et al*., 2005). This further suggests that several libraries with different insert sizes are needed to increase the sensitivity of PEM. The majority of structural variants detected by PEM were relatively small where approximately 65% were <10 Kb and 30% were <5 Kb (Korbel *et al*., 2007). This represents a significant improvement in resolution over microarrays.

In addition to these studies, a large-scale effort is currently being undertaken by the Human Genome Structural

Variation Working Group to comprehensively map structural variants in phenotypically normal individuals using the PEM approach as demonstrated by Tuzun *et al*. (2005) (Eichler *et al*., 2007). More specifically, the objective is to characterise the pattern of human structural variants at the nucleotide level from a collection of 48 individuals of European, Asian and African ancestry. This project plans to make fosmid clone libraries of approximately 40 Kb insert size from the genomic DNA of 48 unrelated females. These samples have already been genotyped in the HapMap Project. A larger insert size of approximately 150 Kb prepared from BAC clone libraries will also be constructed from 14 unrelated HapMap males. This will aim to provide sequence information on structural variants that are too large to be included in the fosmid libraries, such as those associated with segmental duplications (Eichler *et al*., 2007). As such, both the fosmid and BAC libraries will ensure a comprehensive capture of structural variants of varying sizes across the human genome. A preliminary report was published for eight individuals (Kidd *et al*., 2008).

Depth-of-coverage

Principle, strengths and weaknesses

Depth-of-coverage (DOC) is another method using the NGS data for CNVs detection. As the name implies, this method is based on the depth of coverage of the sequence reads to infer deletions and duplications. The DOC method is enabled by the production of several hundred million short sequence reads per instrument run by NGS technologies. The principle underlying the DOC approach is based on the assumptions that the sequencing process is uniform so that the number of sequence reads mapping to a region follows a Poisson distribution. As such, the number of sequence reads should be proportional to the number of times that a particular region appears in the genome. Therefore, it is expected that a duplicated region will have more reads aligned to it, with the converse true for deletions (Yoon *et al*., 2009; Medvedev *et al*., 2009). However, the assumption that the sequencing process is uniform may not be valid. This is because of the sequencing bias of the NGS technologies which leads to certain regions of the genome being over or under-sampled resulting in spurious signals (Harismendy *et al*., 2009).

Based on the principle of the DOC method, the strength of a DOC signature (i.e. 'gain' or 'loses') is thus directly related to the coverage of the sequencing data (the number of sequence reads) and also to the size of the CNVs. This means that the DOC signatures will be stronger for larger CNVs, and is thus more powerful for detecting larger CNVs compared to PEM. In contrast, unlike PEM, the DOC method cannot detect copy-neutral variants. Moreover, the DOC method may not be powerful enough to identify smaller CNVs (related to the strength of DOC signatures) and it is also limited in defining breakpoints (Medvedev *et al*., 2009). In comparison to microarrays, copies number can only be inferred to four (CN = 4) as the upper boundary for SNP microarray or copy number changes will be denoted as 'gain' or 'loses' for CGH microarrays (McCarroll *et al*., 2008; Wang *et al*., 2008). The DOC method is also more robust and accurate at determining higher copies number.

Merging DOC with PEM

Studies comparing the results between the DOC and PEM methods found that only a small fraction of the CNVs overlap between these methods. Furthermore, the identified CNVs that are specific to the DOC method are more enriched in segmental duplications than the PEM-specific CNVs. This is complementary to the PEM method as it has difficulty detecting structural variants in segmental duplications because the paired-end sequences from these repetitive regions cannot uniquely map to a single site or location in the genome, especially for short sequence reads. In comparison, this problem is less significant for DOC as this method does not rely on uniquely mapping sequence reads to a region to infer CNVs. This suggests that a combination of the methods is ideal to further improve the sensitivity of detection throughout the genome. In fact, both methods have their own advantages and limitations (Yoon *et al*., 2009; Medvedev *et al*., 2009). As discussed earlier, the main assumption of the DOC method may not be valid because of the sequencing biases that cause certain regions to be over or under-sampled. To overcome this limitation, a recent study by Medvedev *et al*. (2010) has developed a method to detect CNVs by supplementing the DOC with the PEM data by integrating both types of sequencing data. Using this integrative method, the discordant pairs will be used to indicate the presence of CNVs for DOC. It has been shown that PEM can improve both the sensitivity and the specificity of the DOC method. Several advantages of integrating the DOC and PEM data have also been demonstrated which addresses some of the limitations of each method when used independently. For example, by using this integrative approach, the size of the variants that can be detected is no longer limited by the insert size of library and this approach is also more robust in detecting variants in segmental duplications (Medvedev *et al*., 2010).

Choosing a Sequencing Platform for PEM and DOC

The applications of high-throughput sequencing technologies that are commercially available and accessible by end-users or researchers for PEM and DOC will be further discussed. It is noteworthy that the development of numerous other sequencing technologies such as single molecule real time (SMRT) sequencing (to be marketed commercially soon) are on the horizon (Schadt *et al*., 2010). Although others such as nanopore sequencing may take several years to become a mature technology (Branton

et al., 2008). In comparison, companies such as Complete Genomics provides a sequencing service rather than selling their sequencing machines to end-users (Drmanac *et al*., 2010). The sequencing technologies that are currently available can be broadly grouped into NGS technologies such as the Roche 454 Genome Sequencer FLX (GS FLX) System, Illumina Genome Analyzer (GA) and Applied Biosystems (ABI) Supported Oligonucleotide Ligation Detection System (SOLiD) and *third generation sequencing (TGS) technologies* such as the HeliScope Single Molecule Sequencer which is now commercially marketed by Helicos Biosciences. **See also**: Next Generation Sequencing Technologies and Their Applications; Whole Genome Resequencing and 1000 Genomes Project

Although Roche 454 GS FLX, Illumina GA and ABI SOLiD are classified as NGS technologies, several features differ substantially between them. They are characterised by the ability of parallel sequencing of a very large number of sequence reads. However, the Roche 454 GS FLX can only generate approximately one million sequence reads per instrument run, in comparison to the Illumina GA and ABI SOLiD where several hundred million sequence reads are produced. Similarly, the HeliScope Single Molecule Sequencer can also produce several hundred million sequence reads (Mardis, 2008; Shendure and Ji, 2008; Metzker, 2010; Li and Wang, 2009). One of the major distinctions between NGS and TGS is that TGS requires no whole genome amplification steps such as emulsion polymerase chain reaction and bridge amplification compared to NGS. Therefore, TGS has the potential to further increase the number of sequence reads or throughput per instrument run than their current capacity. Therefore, the Illumina GA, ABI SOLiD and HeliScope Single Molecule Sequencer provide an advantage for the DOC method that requires a high density of sequence reads to infer CNVs. The specificity of DOC to detect CNVs and the precision to map the breakpoints can be improved by increasing the density or coverage of sequence reads (Yoon *et al*., 2009; Medvedev *et al*., 2009). However, the length of sequence reads produced by Roche 454 GS FLX is on average 400–500 bp, which is substantially longer than that for the other three sequencing technologies which range from 32 to 125 bp (Li and Wang, 2009). Although PEM and DOC methods are targeting large structural variants, the sequence read length produced by Roche 454 GS FLX is better for detecting small indels of several to tens of bases. Moreover, the longer sequence read length of Roche 454 GS FLX may also be more suitable for *de novo* genome assembly before read lengths of several kilobases is generated by future sequencing technologies.

The PEM method, when applying it alone rather than integrated with DOC data, must ensure that the paired-end sequences are uniquely aligned to the reference genome to infer structural variants compared to ambiguous paired-end sequences which align to multiple locations. As such, shorter sequence read lengths may be less specific in aligning against the reference genome especially in repetitive regions such as segmental duplications. Moreover, the number of paired-end sequences is also important as usually multiple discordant pairs are needed to reliably detect structural variants. In terms of preparing the PEM libraries for sequencing, all three NGS technologies are able to generate both paired-end and mate-pair libraries, thus allowing for sequencing of short and longer insert sizes (Robison, 2010; Koboldt *et al*., 2010). Each of the sequencing technologies has its own strengths and weaknesses, and a combination of these technologies in an experiment may be the ideal approach to detecting new structural variants and also to address the systematic biases in sequencing (Harismendy *et al*., 2009).

A Comprehensive Detection of Structural Variants in the Human Genome

Currently no single approach can detect all CNVs or structural variants within a human genome. A combination of different approaches is thus ideal where both microarrays and sequencing-based methods can be utilised for this purpose before *de novo* genome assembly is feasible. In comparison to whole genome resequencing that relies on a reference genome for aligning the sequence reads (Wang *et al*., 2008; Bentley *et al*., 2008; Ahn *et al*., 2009), *de novo* genome assembly will enable a more thorough and comprehensive detection of various genetic variants in the human genome ranging from single nucleotide variants, small indels (insertions and deletions) to large structural variants. Currently *de novo* genome assembly is challenging and less practical because of the short sequence reads generated by NGS technologies especially the Illumina Genome Analyzer and Applied Biosystems SOLiD. However, recent studies have attempted to perform *de novo* human genome assembly using short sequence reads with limited success (Li *et al*., 2010a, b; Paszkiewicz and Studholme, 2010). *De novo* genome assembly will become more feasible with longer sequence read lengths of several to tens of kilobases generated by future sequencing technologies. The number of *de novo* genome assembly studies is anticipated to increase exponentially with the arrival of third generation or single-molecule sequencing technologies in the next few years (Schadt *et al*., 2010; Gupta, 2008; Branton *et al*., 2008).

In anticipation, a recent study has used sequencing and microarray-based strategies to detect various genetic variants which complement the results of the assembly comparison approach used in the HuRef genome (Craig Venter) (Levy *et al*., 2007). This study detected genetic variants by aligning the original Sanger sequence reads generated for the HuRef genome to the reference genome (NCBI build-36 assembly). In addition, high density microarrays were custom-designed to probe the HuRef genome to identify variants in regions where sequencing-based approaches may have difficulties. Thousands of new structural variants (i.e. copy number and copy-neutral variants) were discovered and approximately 1.58% (48.8 Mb) of the HuRef haploid genome consisted of structural variants. In

addition, the study also found biases in each method in detecting these variants. This further justifies the need to combine different methods for a more thorough detection of structural variants (Pang *et al*., 2010).

Conclusions

Microarrays have been widely used in the discovery of CNVs over the last several years. However, with the development of PEM and DOC, this raises the question of whether these sequencing-based methods will eventually replace microarrays in structural variant research. The answer is likely to be a resounding 'yes', but at present the microarrays and sequencing-based methods are proving to be valuable by being complementary to each other in population studies of structural variants. The role of microarrays will likely need to be switched from that of 'discovery' to 'genotyping'. Although sequencing-based methods are more powerful in the discovery of new structural variants, these methods are costly for several hundred or thousand samples especially when several libraries of different insert sizes are needed for PEM. This would limit the number of future studies of population characteristics and disease association. However, the newly discovered and the currently known structural variants can be characterised in population-based studies for investigating their associations with diseases using custom-designed oligonucleotide microarrays. However, this is limited to CNVs which are believed to be in the majority in structural variants. Thus other high-throughput methods to assay newly discovered and known copy-neutral variants need to be developed.

Although the PEM and DOC methods have overcome the major shortcomings of microarrays in detecting structural variants, these methods have their own weaknesses. Nevertheless, these emerging sequencing-based methods will continue to play a role in the discovery of structural variants until *de novo* genome assembly is more feasible (Li *et al*., 2010a, b; Paszkiewicz and Studholme, 2010). *De novo* genome assembly will be more practical with the promise of third generation sequencing technologies to increase the sequence read length to tens of kilobases so that a full human genome can be assembled (Schadt *et al*., 2010; Gupta, 2008; Branton *et al*., 2008). In addition to advancing the knowledge of human genetic variation, these methods are also useful in dissecting somatically acquired rearrangements in cancer genomes (Campbell *et al*., 2008; Stephens *et al*., 2009). Finally, the discovery of various genetic variants including structural variants in the human genome has been greatly accelerated by *1000 Genomes Project* (Genomes Project Consortium, 2010; Sudmant *et al*., 2010).

References

1000 Genomes Project Consortium, Durbin RM, Abecasis GR *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Ahn SM, Kim TH, Lee S *et al*. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research* **19**: 1622–1629.

Bentley DR, Balasubramanian S, Swerdlow HP *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Branton D, Deamer DW, Marziali A *et al*. (2008) The potential and challenges of nanopore sequencing. *Nature Biotechnology* **26**: 1146–1153.

Campbell PJ, Stephens PJ, Pleasance ED *et al*. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics* **40**: 722–729.

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics* **39**: S16–S21.

Conrad DF and Hurles ME (2007) The population genetics of structural variation. *Nature Genetics* **39**: S30–S36.

Conrad DF, Pinto D, Redon R *et al*. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712.

Cooper GM, Zerr T, Kidd JM *et al*. (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics* **40**: 1199–1203.

Drmanac R, Sparks AB, Callow MJ *et al*. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78–81.

Eichler EE, Nickerson DA, Altshuler D *et al*. (2007) Completing the map of human genetic variation. *Nature* **447**: 161–165.

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

Feuk L (2010) Inversion variants in the human genome: role in disease and genome architecture. *Genome Medicine* **2**: 11.

Feuk L, Carson AR and Scherer SW (2006) Structural variation in the human genome. *Nature Reviews. Genetics* **7**: 85–97.

Feuk L, MacDonald JR, Tang T *et al*. (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genetics* **1**: e56.

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* **26**: 602–611.

Harismendy O, Ng PC, Strausberg RL *et al*. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biology* **10**: R32.

Hastings PJ, Lupski JR, Rosenberg SM and Ira G (2009) Mechanisms of change in gene copy number. *Nature Reviews. Genetics* **10**: 551–564.

Iafrate AJ, Feuk L, Rivera MN *et al*. (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949–951.

Khaja R, Zhang J, MacDonald JR *et al*. (2006) Genome assembly comparison identifies structural variants in the human genome. *Nature Genetics* **38**: 1413–1418.

Kidd JM, Cooper GM, Donahue WF *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**: 56–64.

Koboldt DC, Ding L, Mardis ER *et al*. (2010) Challenges of sequencing human genomes. *Briefings in Bioinformatics* **11**: 484–498.

Korbel JO, Urban AE, Affourtit JP *et al*. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Korn JM, Kuruvilla FG, McCarroll SA *et al*. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**: 1253–1260.

Ku CS, Pawitan Y, Sim X *et al*. (2010) Genomic copy number variations in three Southeast Asian populations. *Human Mutation* **31**: 851–857.

Lee C, Iafrate AJ and Brothman AR (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nature Genetics* **39**: S48–S54.

Levy S, Sutton G, Ng PC *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biology* **5**: e254.

Li R, Zhu H, Ruan J *et al*. (2010a) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**: 265–272.

Li Y, Hu Y, Bolund L and Wang J (2010b) State of the art de novo assembly of human genomes from massively parallel sequencing data. *Human Genomics* **4**: 271–277.

Li Y and Wang J (2009) Faster human genome sequencing. *Nature Biotechnology* **27**: 820–821.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387–402.

Matsuzaki H, Wang PH, Hu J *et al*. (2009) High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians. *Genome Biology* **10**: R125.

McCarroll SA and Altshuler DM (2007) Copy-number variation and association studies of human disease. *Nature Genetics* **39**: S37–S42.

McCarroll SA, Kuruvilla FG, Korn JM *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166–1174.

Medvedev P, Fiume M, Dzamba M *et al*. (2010) Detecting copy number variation with mated short reads. *Genome Research* September 21 [Epub ahead of print].

Medvedev P, Stanciu M and Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13–S20.

Metzker ML (2010) Sequencing technologies – the next generation. *Nature Reviews. Genetics* **11**: 31–46.

Meyerson M, Gabriel S and Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews. Genetics* **11**: 685–696.

Pang AW, MacDonald JR, Pinto D *et al*. (2010) Towards a comprehensive structural variation map of an individual human genome. *Genome Biology* **11**: R52.

Park H, Kim JI, Ju YS *et al*. (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nature Genetics* **42**: 400–405.

Paszkiewicz K and Studholme DJ (2010) De novo assembly of short sequence reads. *Briefings in Bioinformatics* **11**: 457–472.

Pennisi E (2007) Breakthrough of the year. *Human Genetic Variation. Science* **318**: 1842–1843.

Perry GH, Ben-Dor A, Tsalenko A *et al*. (2008) The fine-scale and complex architecture of human copy-number variation. *American Journal of Human Genetics* **82**: 685–695.

Redon R, Ishikawa S, Fitch KR *et al*. (2006) Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Robison K (2010) Application of second-generation sequencing to cancer genomics. *Briefings in Bioinformatics* **11**: 524–534.

Schadt EE, Turner S and Kasarskis A (2010) A window into third-generation sequencing. *Human Molecular Genetics* **19**: R227–R240.

Sebat J, Lakshmi B, Troge J *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Shendure J and Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.

Stankiewicz P and Lupski JR (2010) Structural variation in the human genome and its role in disease. *Annual Review of Medicine* **61**: 437–455.

Stephens PJ, McBride DJ, Lin ML *et al*. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**: 1005–1010.

Sudmant PH, Kitzman JO, Antonacci F *et al*. (2010) Diversity of human copy number variation and multicopy genes. *Science* **330**: 641–646.

Tuzun E, Sharp AJ and Bailey JA (2005) Fine-scale structural variation of the human genome. *Nature Genetics* **37**: 727–732.

Wang J, Wang W, Li R *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Wang K, Li M, Hadley D *et al*. (2007) 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.

Wong KK, deLeeuw RJ, Dosanjh NS *et al*. (2007) A comprehensive analysis of common copy-number variations in the human genome. *American Journal of Human Genetics* **80**: 91–104.

Yim SH, Kim TM, Hu HJ *et al*. (2010) Copy number variations in East-Asian population and their evolutionary and functional implications. *Human Molecular Genetics* **19**: 1001–1008.

Yoon S, Xuan Z, Makarov V *et al*. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**: 1586–1592.

Zogopoulos G, Ha KC, Naqib F *et al*. (2007) Germ-line DNA copy number variation frequencies in a large North American population. *Human Genetics* **122**: 345–353.

Further Reading

Alkan C, Kidd JM, Marques-Bonet T *et al*. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genetics* **41**: 1061–1067.

Carson AR, Feuk L, Mohammed M and Scherer SW (2006) Strategies for the detection of copy number and other structural variants in the human genome. *Human Genomics* **2**: 403–414.

Hormozdiari F, Alkan C, Eichler EE and Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Research* **19**: 1270–1278.

Kidd JM, Sampas N, Antonacci F *et al*. (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nature Methods* **7**: 365–371.

Wain LV, Armour JA and Tobin MD (2009) Genomic copy number variation, human health, and disease. *Lancet* **374**: 340–350.

Next Generation Sequencing Technologies and Their Applications

Ku Chee-Seng, *Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Loy En Yun, *Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Pawitan Yudi, *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

Chia Kee-Seng, *Centre for Molecular Epidemiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Advanced article

Article Contents

Online posting date: 19th April 2010

The advances in next generation sequencing (NGS) technologies have tremendous impacts on the studies of structural and functional genomics. Sequencing-based approaches like ChIP-Seq and RNA-Seq have started taking the place of microarray experiments to study protein–DNA (deoxyribonucleic acid) interactions and transcriptomic profiling, respectively. The arrival of NGS technologies has also enabled several whole human genome resequencing studies to be completed efficiently at an affordable price. The major strengths of NGS technologies are their ultra high-throughput production, characterized by their ability to generate several hundred megabases to tens of gigabases of sequencing data per instrument run, and more importantly, the steep reduction in cost compared to the traditional Sanger sequencing method. Hence, NGS technologies have rapidly become the primary choice for large scale as well as genome-wide sequencing studies. The new sequencing-based approaches to explore structural and functional genomics have produced important information and significantly expanded our knowledge in these areas.

ELS subject area: Genetics and Disease

How to cite:
Chee-Seng, Ku; En Yun, Loy; Yudi, Pawitan; and Kee-Seng, Chia (April 2010) Next Generation Sequencing Technologies and Their Applications. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0022508

Introduction

The rapid developments in *sequencing* technologies have transformed the approaches in the studies of structural and functional genomics. The studies of structural genomics focus on identifying various genetic variations or mutations, whereas functional genomics studies aim to interrogate and annotate the functional and regulatory elements or sequences in the human genome. The *next generation sequencing (NGS) technologies* have started substituting traditional Sanger sequencing methods in many large scale or genome-wide sequencing studies. These new sequencing technologies have been attracting a considerable amount of interest from researchers since they have been commercially marketed. The major attractions are their ultra high-throughput production, characterized by their ability to simultaneously sequence millions of DNA (deoxyribonucleic acid) fragments and produce gigabases of sequencing data per instrument run, and more importantly, the steep reduction in cost compared to the traditional sequencing method.

Revolution in the Approaches for Genomics Studies

Previously, the molecular genomics studies mainly relied on microarray technologies such as gene expression microarrays and the ChIP-chip method (i.e. *chromatin immunoprecipitation* coupled with microarray) for genome-wide interrogation. However, this was swiftly replaced by sequencing-based methods, namely RNA-Seq (to measure transcripts or ribonucleic acids (RNAs) expression levels) and ChIP-Seq (to study protein–DNA

interactions like identifying transcription factor-binding sites and interrogating histone modifications), respectively (Wang *et al*., 2009; Park, 2009).

There are a number of limitations in using microarrays compared to sequencing-based methods. For example, conventional microarrays do not allow a truly comprehensive interrogation of the whole genome, because the selection of probes to be synthesized and immobilized on the solid surface of microarrays requires some prior knowledge and reference genome sequences are also needed. The probes are needed to detect and measure the abundance of DNA or RNA targets through hybridization. In other words, microarray-based methods are limited to interrogating those genomic regions that are probed by the microarrays. It is obvious from the conventional gene expression microarray studies where the gene expression levels could not be measured unless there are probes to capture them, and the probes are usually synthesized to capture known annotated protein-coding genes. Therefore unknown transcripts or those transcripts from noncoding sequences in the *transcriptome* could not be assessed. Similarly for ChIP-chip experiments, the DNA fragments that are pulled down by immunoprecipitation would be undetected if no complementary probes are designed to capture them. On the contrary, theoretically sequencing-based approaches are able to capture all the DNA fragments that are isolated by immunoprecipitation (ChIP-Seq), and all the transcripts (coding and noncoding transcripts) that are available in the transcriptome (RNA-Seq) including the low abundance transcripts, if the sequencing depth is sufficient (Wang *et al*., 2009; Park, 2009).

Likewise in structural genomics studies, microarray-based methods such as comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) arrays have poor sensitivity to detect smaller sizes of copy number variations (CNVs) like those of < 10 kb, and these methods are unable to detect copy neutral variations like balanced translocations and inversions. Furthermore, microarray-based methods have limited resolution to define the breakpoints of CNVs and *structural variations*. However, these limitations have been overcome by sequencing-based methods like paired-end mapping (Korbel *et al*., 2007). These new and innovative sequencing-based approaches to studying structural and functional genomics have produced important information and have significantly expanded our knowledge in each area.

Next Generation Sequencing Technologies

Sanger dideoxynucleotide or chain termination sequencing has been the most widely used sequencing method for the past three decades since it was invented in late 1970s until the first NGS platform was marketed in 2005. Sanger sequencing has been used for various applications such as mutations discovery, genotyping and *serial analysis of gene expression* (*SAGE*) for measuring gene expression levels, and more importantly, it was used to complete the Human Genome Project (International Human Genome Sequencing Consortium, 2004). **See also**: Human Genome Project: Importance in Clinical Genetics; Sequencing the Human Genome: Novel Insights into its Structure and Function; Whole Genome Resequencing and 1000 Genomes Project

Shortly after the first next generation sequencer was introduced by Roche® 454 Life Science, the Genome Sequencer 20 (GS 20) System (it was subsequently replaced by GS FLX System with further improvements, i.e. higher throughput and longer sequence read length, to the preceding system), another two biotechnology companies also marketed their sequencing platforms: Illumina® Genome Analyzer (GA) and Applied Biosystems® (ABI) Supported Oligonucleotide Ligation Detection System (SOLiD). The simultaneous advent of several next generation sequencers created intense competition in the sequencing market; with each technology having its own strengths and limitations. This article focuses on the NGS technologies because they have been widely used for various applications unlike the newer third generation sequencing instrument, the Heliscope Single Molecule Sequencer, which has only recently been introduced. The following sections described the main features of NGS technologies.

Sequencing throughput and cost

Currently, Sanger sequencing machines (e.g. ABI® 3730xl) have been largely supplanted by next generation sequencers in many large genomics institutes worldwide. This was mainly due to the ultra high-throughput production of NGS technologies which is several orders of magnitude higher than Sanger sequencing method. One of the major differences between modern and traditional sequencing is the ability of next generation sequencers to simultaneously sequence one million to several hundred millions of DNA fragments in contrast to the 96-capillary Sanger sequencer. Therefore, NGS is also known as massively parallel sequencing technologies. This feature has enormously increased the amount of the production or the number of nucleotides or bases that it can sequence compared to the Sanger sequencer in one experiment or per instrument run. For example, the latest developments in Illumina® GA and ABI® SOLiD have further increased the throughput production generating more than 10 gigabases of sequencing data per instrument run in a few days, whereas Roche® GS FLX can generate several hundred megabases per run in 10 h. In contrast, Sanger sequencer like ABI® 3730xl which is commonly used in most of the research laboratories can only produce $\sim$100 kb per run in 3 h (see **Table 1** for the summary of the features of NGS technologies) (Shendure and Ji, 2008; Tucker *et al*., 2009).

The sequencing chemistry of NGS technologies together with their ultra high-throughput production has also reduced the sequencing cost significantly, making large-

Table 1 Summary of the features of NGS technologies

Feature	Roche® 454 GS FLX	Illumina® GA	ABI® SOLiD
The year of the first sequencer that commercially marketed	2005	2006	2007
Current generation of the sequencer	Roche® 454 GS FLX Titanium	Illumina® GA II	ABI® SOLiD 3.0
Massively parallel sequencing (number of DNA fragments)	Several hundred thousand to one million	Several hundred millions	Several hundred millions
Sequencing throughput per instrument run	Several hundred megabases per run in 10 h	>10 Gb per run in a few days	>10 Gb per run in a few days
Sequencing cost per megabase (US$)	$\sim$ \$80	$\sim$ \$6	$\sim$ \$6
Differences in cost in relative to Sanger sequencing ($500 per megabase)	$\sim$ 6-fold	$\sim$ 80-fold	$\sim$ 80-fold
In vitro amplification method	Emulsion PCR	Bridge amplification on solid surface	Emulsion PCR
Sequencing approach	Sequencing by synthesis mediated by polymerase – pyrosequencing	Sequencing by synthesis mediated by polymerase – sequencing by reversible terminator chemistry	Sequencing by ligation of dinucleotide probes mediated by ligase
Sequencing reagent	Four types of dNTPs	Four types of ddNTPs labelled by four different fluorescent colours	16 types of dinucleotide probes labeled by 4 different fluorescent colours
Detection method of the incorporated nucleotides	Emission of chemiluminescent light	Fluorescent colours	Fluorescent colors
Sequence read length	400–500 bases	75–125 bases	50 bases
Read base or base calling error rate	0.5–1.5%	0.2–2%	<0.1%
Error type	Insertion or deletion of nucleotides in homopolymer sequences	Substitution of nucleotides	Substitution of nucleotides

scale sequencing studies affordable nowadays. Currently, Illumina® GA and ABI® SOLiD have already achieved a sequencing cost of $6 per megabase as compared to Roche® GS FLX, which is offered at $80 per megabase. In general, the sequencing cost of NGS technologies was substantially decreased by several folds to nearly 100-fold compared to Sanger sequencing, which costs about $500 for the same amount of sequencing data (Shendure and Ji, 2008; Tucker *et al*., 2009). It is noteworthy that the cost of sequencing is changing continuously; therefore the prices cited here may not be the latest in the market. Regardless, this provides some useful information on differences in sequencing cost between Sanger sequencing and NGS. Undoubtedly, both sequencing production and cost would be continuously improved. The developments of third generation sequencing technologies are on the horizon and the instruments are expected to be marketed soon which would certainly decrease the sequencing cost further and eventually achieve the ultimate goal of $1000 per genome sequencing (Von Bubnoff, 2008).

On top of the considerations of sequencing throughput and cost, the other concern is logistics. As the amount of sequencing data produced by a next generation sequencer is equivalent to tens of Sanger sequencers, a large area or space would be needed to accommodate the instruments. This can only be feasibly attained by large genomics laboratories or institutes. Furthermore, the maintenance of tens of sequencing instruments will also be substantial and this has not taken into account costs of labour or man-power to operate the instruments.

Sequencing chemistry: *in vitro* amplification

The advances in sequencing technologies have enabled several whole human diploid genome resequencing studies to be completed efficiently. Besides the genome of James Watson (Wheeler *et al*., 2008), several genomes of anonymous individuals have also been sequenced; they are two Koreans (AK1 and SJK) and one individual each of

Han Chinese (YH), African (NA18507) and European (P0) ancestries (Kim *et al*., 2009; Ahn *et al*., 2009; Wang *et al*., 2008; Bentley *et al*., 2008; Mckernan *et al*., 2009; Pushkarev *et al*., 2009). All these genomes were sequenced by NGS technologies except the genome of the European individual P0 which was sequenced by Heliscope Single Molecule Sequencer. In contrast, the diploid genome of Craig Venter was sequenced by the Sanger sequencing method (Levy *et al*., 2007). The whole genome resequencing studies using next and third generation sequencing technologies were completed at a cost of tens of thousands to several hundred thousands of dollars compared to Venter's genome which cost millions of dollars.

One of the major limitations in whole genome resequencing using the Sanger sequencing method is the *in vivo* amplification of DNA fragments using bacterial cloning. This is unlike targeted sequencing studies, where conventional polymerase chain reaction (PCR) is commonly used to amplify the regions of interest to be sequenced. The bacterial cloning procedures can introduce host cloning-related biases; for example, it could affect the genome representation in the sequencing of organism genomes because some of the DNA fragments failed to be cloned. Moreover, these steps are tedious and labour intensive. However, this method has since been eliminated and is replaced by the *in vitro* amplification of millions of DNA fragments simultaneously by NGS technologies, that is emulsion PCR for Roche® GS FLX and ABI® SOLiD, and bridge amplification on solid surface for Illumina® GA (Mardis, 2008; Strausberg *et al*., 2008; Ansorge, 2009).

In emulsion PCR, the single-stranded DNA fragments or templates are attached to the surface of beads using adaptors or linkers, and one bead is attached to a single DNA fragment from the DNA library. The DNA library is generated through random fragmentation of the genomic DNA. The surface of the beads contains oligonucleotide probes with sequences that are complementary to the adaptors binding the DNA fragments. After that, the beads will be compartmentalized into separate water-oil emulsion droplets. In the aqueous water-oil emulsion, each of the droplets capturing one bead will serve as a PCR microreactor for amplification steps to take place and produce clonally amplified copies of the DNA fragment.

However, for bridge amplification on solid surface for Illumina® GA, the single-stranded DNA fragments are first attached to a solid surface known as a flowcell using adaptors with complementary probes on the flowcell. Then, the other unattached end of the DNA fragments will create a 'bridge-like structure' by bending over and also hybridize to the probes on the flowcell, which form the template for amplification to generate clonally amplified copies of the DNA fragments on the surface of the flowcell. However, this third generation sequencing is characterized by single DNA molecule sequencing without the need for amplification steps. The first third generation sequencing instrument – Heliscope Single Molecule Sequencer – is now commercially marketed by Helicos Biosciences.

Sequencing chemistry: massively parallel sequencing

The sequencing approaches for NGS technologies can be broadly divided into sequencing-by-synthesis mediated by polymerase enzymes (pyrosequencing for Roche® GS FLX and sequencing by reversible terminator chemistry for Illumina® GA) and sequencing-by-ligation mediated by ligase enzymes (ABI® SOLiD) (Mardis, 2008; Strausberg *et al*., 2008; Ansorge, 2009).

In pyrosequencing, the adding of dNTPs (deoxynucleotide triphosphate) and reagents for cyclic sequencing is controlled, where each of the four types of dNTPs will flow through the picotiter plate consecutively or sequentially. This means that only one type of dNTP is present per cycle of sequencing or synthesis, followed by another type of dNTP in the next cycle and the cycles repeat. This is totally different from the reversible terminator chemistry sequencing for Illumina® GA where all the four types of ddNTPs labelled by different fluorescent colours are present in each cycle of sequencing. A picotiter plate contains more than one million wells where the beads (attached to clonally amplified copies of DNA fragments) are situated, and one well holds one bead. As such, it allows parallel sequencing of an enormous number of DNA fragments.

The polymerase-based synthesis or incorporation of the complementary dNTPs to the DNA templates will cause the release of inorganic pyrophosphate triggering a series of downstream reactions which eventually produce chemiluminescent light which is captured by a detection system (CCD camera). The detection system records the intensity of light emitted from each well that corresponds to a single DNA fragment. In summary, generally each cycle of sequencing consists of dNTPs incorporation, pyrosequencing reactions and emission of chemiluminescent light and measurement of the light intensity. The sequencing reagents of the previous cycle are washed away before next cycle of sequencing takes place.

The intensity of chemiluminescence is proportional to the amount of inorganic pyrophosphate released and thus the number of dNTPs incorporated to the DNA template. Owing to this factor, pyrosequencing is more susceptible to insertion deletion (indel) errors in homopolymer sequences (i.e. DNA sequences of consecutive identical nucleotides like GGGGG or AAAAA) because of less accurate estimation of the length or the number of nucleotides in homopolymer sequences. This is especially problematic for homopolymers with more than six bases. In pyrosequencing, several dNTPs can be incorporated when there are consecutive identical nucleotides in the sequences; this is in contrast to the sequencing by reversible terminator chemistry where only one ddNTP (dideoxynucleotide triphosphate) is incorporated to the DNA templates per cycle of sequencing. To further illustrate this, for example, for homopolymer GGGGG, five dCTPs (deoxycytidine triphosphate) will be incorporated for pyrosequencing at one time, whereas only one ddCTP (dideoxycytidine

triphosphate) for reversible terminator chemistry sequencing and another ddCTP will be incorporated in the next four cycles of sequencing.

Like Roche® GS FLX, Illumina® GA also employs the sequencing-by-synthesis approach, although it is totally different from pyrosequencing. In reversible terminator chemistry sequencing, all the four types of ddNTPs and sequencing reagents are added onto the flowcell, and these ddNTPs are labelled by four different fluorescent colours corresponding to the four different nucleotides. One flowcell has several hundred million clusters and each cluster contains clonally amplified copies from a single DNA fragment. Similar to the Roche® GS FLX picotiter plate, the format of the flowcell also allows simultaneous sequencing of an enormous number of DNA fragments. However, it is noteworthy that the difference in the number of DNA fragments that gets sequenced in parallel between the two platforms is about several hundred-folds.

The ddNTPs are reversible terminators, allowing for the synthesis of DNA templates in the next cycle of sequencing for the incorporation of other ddNTPs. In this cyclic sequencing approach, one complementary ddNTP will be incorporated to the DNA template at one time, followed by washing steps to remove the excess sequencing reagents. This is then followed by the imaging of the fluorescence signals across the whole flowcell. After imaging, the fluorescent labels will be removed and the 3′ blocking group of the ddNTPs is also chemically removed. These steps are then repeated. Since only one ddNTP is incorporated at one time, and the base calling is not proportional to light intensity but is dependent on the fluorescent colours, the reversible terminator chemistry does not have problems in the sequencing of homopolymer sequences. However, it is more prone to substitution errors because all the four types of ddNTPs are present in each cycle of sequencing, unlike in pyrosequencing, where only one specific type of dNTP is present.

It is worthwhile to note that in pyrosequencing, dNTPs are used, whereas in reversible terminator chemistry sequencing, ddNTPs are used and they are reversible terminators. However, Sanger sequencing requires a mixture of both dNTPs and ddNTPs, and the ddNTPs are non-reversible terminators. Although these sequencing approaches are generally based on sequencing-by-synthesis, it is obvious that the sequencing chemistries and approaches are very different. In pyrosequencing, the identity of nucleotides that are incorporated into DNA templates is determined by emission of chemiluminescent light; however, the nucleotides are determined by different fluorescent colours for reversible terminator chemistry and Sanger sequencing.

The sequencing approach of ABI® SOLiD is based on sequencing-by-ligation. Like Roche® GS FLX, ABI® SOLiD also employs emulsion PCR for amplification. The beads containing DNA fragments are then deposited on a glass slide. The sequencing of DNA templates is mediated by ligase. In brief, the sequencing is based on sequential ligation of dinucleotide probes which are labelled by four different fluorescent colours. There are 16 possible combinations of two nucleotides, and these dinucleotide probes will compete for incorporation into the DNA templates. As such, ligation of one probe will query two nucleotides in the DNA templates. The sequencing of DNA templates is completed by seven ligation cycles for each of the five rounds of primer reset, and at the end produces a sequence read length of 35 bases. Using this unique sequencing approach, every single position or base in the DNA template is interrogated twice, and allowing for distinction between true genetic variations and errors.

Sequence read length and error

The NGS technologies have a number of advantages over Sanger sequencing, but they are not without limitations. The new sequencing technologies are characterized by shorter sequence read lengths compared to Sanger sequencing, that is 125 bases or less for Illumina® GA and ABI® SOLiD, as well as for Heliscope Single Molecule Sequencer. As a result, NGS technologies are not suitable for *de novo* sequencing of large and complex genomes like the human genome as the assembly of billions of short sequence reads into large contigs would be difficult and challenging. Relatively longer sequence read lengths are needed to obtain larger contigs with fewer gaps in between in the assembled consensus sequence. However, the latest improvements in sequencing chemistry and system have enabled Roche® GS FLX to achieve sequence read lengths of 400–500 bases on average, but it is still half of that that can be achieved by Sanger sequencing, which is approximately 800 bases to 1 kb in length (Mardis, 2008; Strausberg *et al.*, 2008; Ansorge, 2009).

The feature of short sequence read lengths also makes NGS technologies like Illumina® GA and ABI® SOLiD inadequate for metagenomics studies in investigating bacterial diversity. It is crucial to have longer sequence read lengths to achieve sufficient discriminatory power of the sequences derived from different bacterial species in a sample, determining the presence of diverse species by mapping the sequences against different reference bacterial genomes. As a result, Roche® GS FLX has become the primary choice for this kind of studies. Nevertheless, the feature of short sequence read length is just nice for 'sequence census methods or applications' like ChIP-Seq and RNA-Seq. These sequence census methods do not require full sequence or long sequence read lengths, but rather, lengths sufficient to align or map the sequences uniquely to the reference genome sequence (Wold and Myers, 2008).

Although Illumina® GA and ABI® SOLiD are less suitable for metagenomics studies at the time, they appear to be more ideal for studies like ChIP-Seq and RNA-Seq compared to Roche® GS FLX. This is because of their ability to generate several hundred millions of short sequence reads compared to several hundred thousand to one million longer sequence reads for Roche® GS FLX. In the applications like ChIP-Seq and RNA-Seq, the number of sequence reads is more crucial than the length of sequence reads for 'counting' purposes, as far as the length is sufficient to align uniquely to the reference genome sequence.

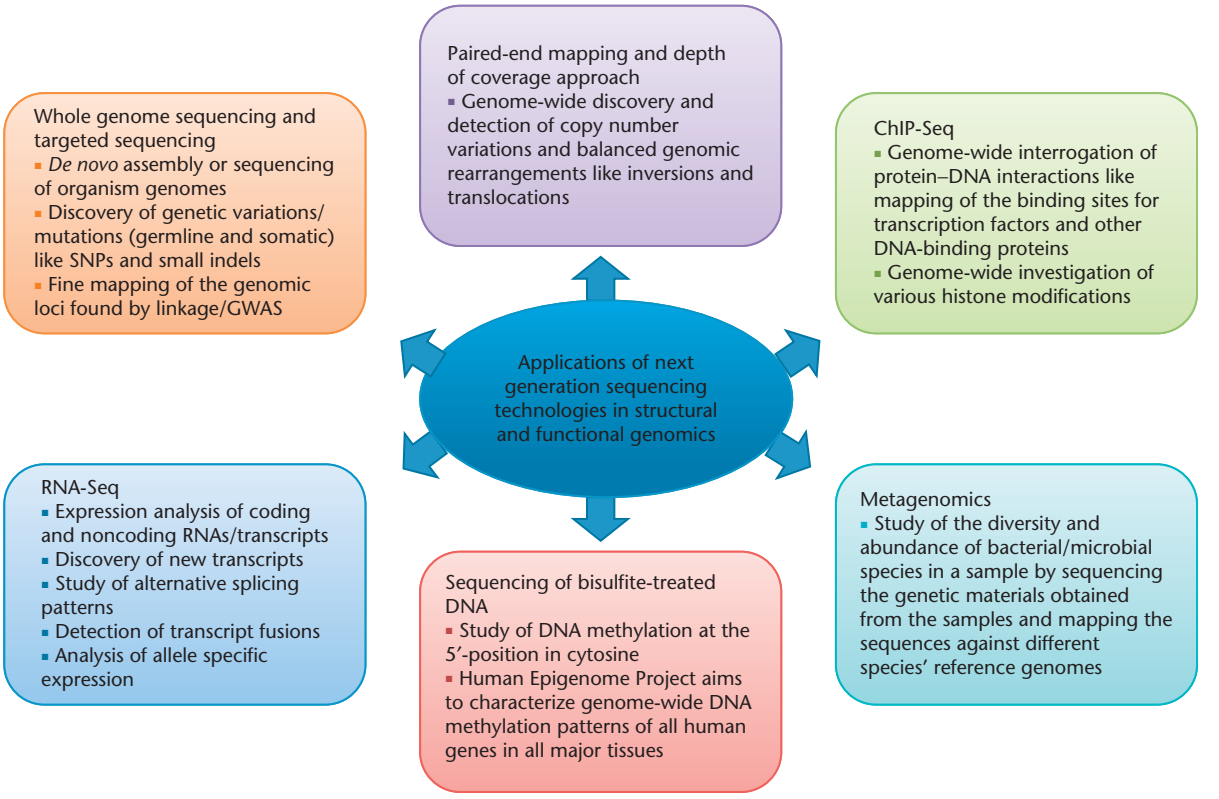


Figure 1 Application of next generation sequencing technologies in structural and functional genomics.

In addition to the limitation in sequence read length, the NGS technologies were also reported to have higher read base or base calling error rates, although it has been improving. ABI® SOLiD has achieved the highest accuracy with <0.1% error rate among the NGS technologies, whereas the read base error rates for Illumina® GA and Roche® GS FLX are within 0.2–2% and 0.5–1.5%, respectively (Li and Wang, 2009). The differences seem to be small and insignificant in terms of the percentage, but when the error rates are transmitted to whole genome sequencing of six billion bases for a human diploid genome, it will generate hundreds of thousands to millions of errors in base calling and this will cause a detrimental effect in identifying genetic variations like SNPs. Fortunately, results from whole genome resequencing studies suggest that the SNP calling error rate decreases significantly with greater sequencing depth (Wang *et al*., 2008). Therefore, it seems that the remedy is to increase the sequencing depth, but one has to bear in mind that this will also add to the sequencing cost.

Applications in Structural and Functional Genomics Studies

Since the arrival of first NGS technology in 2005, these new sequencing platforms have contributed much to the progress in the research of structural and functional genomics. The NGS technologies have been used in various research areas besides the standard sequencing applications such as whole genome sequencing; they have also been increasingly applied in detecting structural variations (paired-end mapping), studies of protein–DNA interactions and histone modifications (ChIP-Seq), and transcriptomic profiling of messenger RNAs (mRNAs) and noncoding RNAs (RNA-Seq). These are the most common applications built on the NGS data and will be the focus of our discussion (**Figure 1**). This article also focuses on these applications in human genomics studies, although NGS technologies have also been widely used for genomics studies of plants and other model organisms. The new and innovative applications of NGS technologies have contributed remarkably to the advancement in human genomics studies.

Whole genome sequencing

The completion of several whole human genome resequencing studies has yielded important scientific findings and new insights into *human genetic variations* (Wheeler *et al*., 2008; Kim *et al*., 2009; Ahn *et al*., 2009; Wang *et al*., 2008; Bentley *et al*., 2008; Mckernan *et al*., 2009; Pushkarev *et al*., 2009). It is equally important that they also served as proof-of-concept studies demonstrating the feasibility of using NGS and third generation sequencing technologies

to decode the DNA sequence of human genome efficiently and at an affordable price per genome. Moreover, these studies have also addressed important questions and issues surrounding the experimental design and data analysis, such as the preparation of DNA libraries for sequencing, assessment of the sequencing depth that is needed to provide adequate coverage of the reference genome sequence and to minimize SNP calling error rate, and the quality control criteria for the detection of genetic variations like SNPs, indels and structural variations. For example, Wang *et al.* (2008) found that at a sequencing depth of greater than 10-fold, the assembled consensus sequence covered ~83% of the NCBI human reference genome using single-end reads and ~95% coverage using paired-end reads, and greater sequencing depth has minimally increased in the coverage. However, the SNP calling error rate decreases significantly with greater sequencing depth.

The findings from several whole genome resequencing studies have also deepened our understanding of human genetic variations. These studies revealed an abundance of various genetic variations in the human genome, namely SNPs, indels and structural variations. Although the finding of several million SNPs in each individual genome is not new, more interesting is the fact that the studies have identified several hundreds of thousands of new SNPs that have not been catalogued in dbSNP. For example, about one million new SNPs were identified in the African genome (NA18507) and approximately half a million SNPs for the other genomes of Caucasian and Asian ancestry (Bentley *et al.*, 2008; Wheeler *et al.*, 2008; Wang *et al.*, 2008; Kim *et al.*, 2009; Ahn *et al.*, 2009).

Apart from SNPs, whole genome resequencing studies also identified several hundred thousand of short indels with sizes ranging from several bases to tens of bases. The Han Chinese (YH) genome contained approximately 135 000 indels within 1–3 bp, and approximately 400 000 indels defined from 1 to 16 bp were found in the African NA18507 genome. However, Ahn and colleagues identified the indels within a size range from −29 to +14 bp and found nearly 343 000 entries for the Korean genome SJK (Bentley *et al.*, 2008; Wang *et al.*, 2008; Ahn *et al.*, 2009). The effort to catalogue short indels in the human genome was far less devoted than that for SNPs, where more than 50% of the identified indels have not been catalogued, whereas only less than 30% of the identified SNPs are new. Similarly some new discoveries have also been made for structural variations, where several thousands of them were identified. The large-scale sequencing studies like whole genome resequencing and *1000 Genomes Project* would not have been feasible without the advances in NGS technologies. **See also**: Copy Number Variation in the Human Genome; Genetic Variation: Human; Single Nucleotide Polymorphism (SNP)

In addition to the aforementioned whole genome resequencing of nondisease genomes, the cancer genome of acute myeloid leukaemia has also been sequenced to study the *de novo* somatic mutations (Ley *et al.*, 2008). Apart from germline genetic variations, the importance of somatic mutations in carcinogenesis is also well established. Therefore, focusing merely on germline genetic variations will not be sufficient to fully decipher the genetic basis of cancers. It is noteworthy that the genome-wide association studies (GWAS) only interrogated the germline genetic variations of cancer and that the whole genome SNP genotyping arrays used in GWAS are not designed to study somatic mutations. Direct sequencing is required for detecting somatic mutations; hence, sequencing approach provides an additional advantage in dissecting the cancer genome compared to genotyping.

Paired-end mapping of structural variations

The ubiquity of CNVs in the human genome was first reported several years ago (Sebat *et al.*, 2004; Iafrate *et al.*, 2004), and many more have since been found. Previous studies have used poor sensitivity methods to detect CNVs leading to high false negative rates (Scherer *et al.*, 2007). Most of the CNV data were generated by microarray-based methods such as CGH and SNP arrays where the signal intensity information is used to detect deletions and duplications. Because of the reliance on relative or differences in signal intensities to detect copy number variable regions, these methods are unsuitable for detecting other structural variations like inversions and translocations (also known as balanced chromosomal rearrangements). Furthermore, due to the limitations in density or resolution of CGH and SNP arrays, the methods are lacking in sensitivity to detect smaller sizes of CNVs (< 50 kb). The discovery of smaller sizes of CNVs is crucial as they are predicted to be more abundant than the larger CNVs (Estivill and Armengol, 2007). The latest developments in SNP genotyping arrays, namely increased probe density and uniformity of distribution in the genome, and also included copy number probes to cover regions lacking of SNPs, have improved the sensitivity compared to earlier arrays. Nonetheless, the SNP arrays still suffer from poor sensitivity to detect CNVs smaller than 5–10 kb even using the highest density SNP arrays such as Illumina® Human 1 M Beadchip and Affymetrix® 6.0 SNP Arrays (McCarroll *et al.*, 2008; Cooper *et al.*, 2008). Therefore, higher resolution and sensitivity methods are needed to detect CNVs and also balanced structural variations.

The proof-of-concept study using NGS technologies to detect structural variations was published in 2007, and the sequencing-based method was known as paired-end mapping (Korbel *et al.*, 2007). In this method, a library of DNA fragments of fixed insert sizes is prepared, both ends of the DNA fragments are sequenced, and the sequence information is used to map against the human reference genome. The underlying principle of the paired-end mapping approach to detect structural variations is simple; it is based on the discrepancies in length or orientation of the DNA fragments to be sequenced. In other words, when both ends of the DNA fragment that map against the reference sequence show discordances in terms of size or

length, this is an indication for deletion and insertion, whereas discordance in orientation suggests the presence of inversion. Since the insert size of the library is known, both ends of DNA fragments that map to the reference is shorter than expected; this indicates the presence of insertion; conversely, longer than the insert size suggests the presence of deletion. Korbel and colleagues prepared libraries of 3 kb insert size for two female individuals, and using the aforementioned mapping approach and Roche® 454 sequencing, they found 1297 structural variations, including 853 deletions, 322 insertions and 122 inversions. After this study, several whole genome resequencing studies have also used the paired-end mapping strategy and identified thousands of structural variations (Wang *et al*., 2008; Ahn *et al*., 2009).

Furthermore, the paired-end sequencing method has also been used to interrogate somatic genomic rearrangements in cancer (Campbell *et al*., 2008). In the study, Illumina® GA was used to perform the sequencing of both ends of DNA fragments derived from the genomes of two individuals with lung cancer, and they identified 306 germline structural variants and 103 somatic rearrangements to the single nucleotide level of resolution. The cancer genome is well characterized by genomic instability, with the presence of numerous structural variations and complex genomic rearrangements, and these genetic aberrations are not well captured by microarray hybridization methods. However, this study has now shown the feasibility and advantages of paired-end sequencing method to decipher the cancer genome. This mapping approach is undoubtedly a promising strategy to harvest new cancer genes. The paired-end sequencing approach takes the advantages of the short sequence reads produced by NGS technologies to map against human reference genome, it is an application of 'census sequence methods'. Nonetheless, one major limitation of paired-end mapping is the inability to detect insertions larger than the insert size of the library.

Recently a new and innovative method of using NGS data to detect CNVs has been developed. The approach is based on the depth of coverage of the sequence reads, and some promising results have been obtained showing that it is effective to search for copy number variable regions. The principle underlying the depth of coverage approach is not complicated. This approach assumes that the sequencing is uniform, and that the number of sequence reads mapping to a region follows a Poisson distribution. As such, the number of reads should be proportional to the number of times that a particular region appears in the genome. Therefore, it is expected that a duplicated region will have more number of reads mapping to it, and the converse is true for deletions (Yoon *et al*., 2009; Medvedev *et al*., 2009).

Studies comparing the results between the depth of coverage approach and the paired-end mapping approach found that only a minority of the CNVs had overlapped between the two methods. Furthermore, the identified CNVs that are specific to the former method are more greatly enriched in segmental duplications than the paired-end mapping-specific CNVs. This suggests that both methods in identifying CNVs are complementary to each other and that the combination of the methods will certainly further improve the sensitivity of detection throughout the genome. In fact, both methods have their own advantages and limitations (Yoon *et al*., 2009; Medvedev *et al*., 2009).

ChIP-seq for studying protein–DNA interactions and histone modifications

Previously, the studies of protein–DNA interactions like identifying transcription factor binding sites, have relied on some low-throughput methods, and focused on some specific genomic regions. However, with the advent of microarray technologies, for the first time, a comprehensive interrogation of the whole genome has become feasible. In the era of microarrays, the genome-wide studies of protein–DNA interactions and histone modifications were performed using a method known as ChIP-chip.

The ChIP or chromatin immunoprecipitation experiment consists of several steps. First, the protein (e.g. a transcription factor of interest) and its binding DNA sequences or genomic regions are chemically cross-linked by treating the cells with formaldehyde. Then the genomic DNA is extracted and fragmented before adding the specific antibody interacting with the protein of interest. The function of the antibody is to selectively isolate the antibody–protein–DNA complexes by immunoprecipitation. After the immunoprecipitation, the cross-linking between protein and DNA is reversed to obtain the DNA sequences. The identity of isolated DNA sequences can be determined by methods such as Southern blot, quantitative PCR (qPCR), microarray (ChIP-chip) or sequencing (ChIP-Seq). Chromatin immunoprecipitation requires a highly specific antibody for the DNA-binding protein of interest.

Before microarrays were available, most of the studies of protein–DNA interactions were designed to answer simple questions like whether a genomic region (e.g. the promoter region of a gene of interest) is bound to a transcription factor thus regulating the transcription levels, that is locus-specific experiment. These studies require some prior knowledge to design the experiments and the immunoprecipitated DNA sequences are analysed by Southern blot or qPCR to determine whether the genomic region was indeed immunoprecipitated. However, the arrival of microarray technologies has enabled a different question to be asked. Since the scope of ChIP-chip experiments is not restricted to specific regions, the question that is posed is where the transcription factor binds to in the human genome, that is to identify all the regions where the transcription factor might have regulatory roles. In ChIP-chip experiments, the isolated DNA fragments are labelled fluorescently and hybridized to the probes on microarrays. Undeniably, the developments of microarrays have enabled interrogation on a genome-wide scale, but the

detection of the isolated DNA sequences is still dependent on the availability of the probes to capture them. Although the developments of high-density tiling arrays, where oligonucleotide probes are placed in high density throughout the whole genome, have improved the sensitivity of the ChIP-chip, the cost for such tiling arrays is expensive especially for large genomes like the human genome.

In contrast, for ChIP-Seq, the isolated DNA sequences are not hybridized on microarrays (hence avoiding the inherent problems in probe hybridization experiments); instead they are directly sequenced to detect their presence and abundance. This allows detection of all the DNA fragments or sequences that are isolated in the sample without biases of probe selection. Actually, both the methods, microarray- and sequencing-based experiments, rely on the reference genome sequence, the former method requiring it for synthesizing the probes, and the later method requiring the reference genome for alignments of DNA sequences that it sequenced (Park, 2009; Farnham, 2009).

The earliest two ChIP-Seq studies were first published in 2007 to identify the genome-wide binding sites for DNA-binding proteins, NRSF (neuron restrictive silencer factor) and STAT1 (signal transducer and activator of transcription 1) (Johnson *et al*., 2007; Robertson *et al*., 2007). These papers served as proof-of-concept studies for the new approach in studying protein–DNA interactions. Both studies used Illumina® GA to sequence the immunoprecipitated DNA sequences. The identification of the previously known binding sites in both the studies serves as the validation of the approach, and the detection of novel-binding sites shows the higher sensitivity of ChIP-Seq compared to ChIP-chip. The studies have shown some promising results; for example, a total of 1946 locations were identified in the human genome for NRSF, and more importantly, the sequencing data provide a sharp resolution of the binding sites. This approach will certainly facilitate the annotation of the binding sites in the genome for other DNA-binding proteins as well.

The first paper investigating histone methylations using ChIP-Seq also appeared in 2007 (Barski *et al*., 2007). The study performed genome-wide mapping of 20 different types of histone modifications in the human genome and also used Illumina® GA to perform the sequencing. The high-resolution maps of histone modifications generated by sequencing methods are important in expanding our knowledge on how this mechanism regulates the expression of genes in the human genome. The development of ChIP-Seq is a major stride in functional genomics as the studies of genome-wide protein–DNA interactions like transcription factor binding sites and studies of epigenetics like histone modifications are essential in our understanding of the transcriptional regulatory network. Nonetheless, ChIP-Seq is not without its own challenges and limitations (Park, 2009; Farnham, 2009).

Transcriptomic profiling

Studies of gene expression are important because they are the immediate molecular traits that are directly affected by genetic variations in DNA sequence and epigenetics regulations. The term gene expression usually refers to expression levels of protein-coding genes, or mRNAs. Previous studies were mainly focused on mRNAs expression, because this class of RNAs is important as they serve as the templates to synthesize proteins through the process of translation, and proteins are the functional molecules involved in diverse cellular functions and biological processes. However, this perception has been changed after the completion of the pilot phase of the ENCODE (Encyclopedia of DNA Elements) Project. The project revealed a pervasive transcription pattern in the 1% of the human genome that was interrogated (ENCODE Project Consortium, 2007; Carninci and Hayashizaki, 2007). It had been previously thought that only the protein-coding regions or sequences (i.e. genes) will undergo transcription followed by translation. However, the ENCODE Project showed that transcription also occurs in nonprotein coding regions as well.

Following the findings, the importance and existence of noncoding RNAs is getting appreciated and research has been devoted to identify and characterize them in the transcriptome. In contrast to mRNAs, the noncoding RNAs only undergo transcription, but are not translated into protein. As such, the transcriptome profiling encompassed both the coding RNAs (mRNAs) and noncoding RNAs. One of the well-known noncoding RNAs is *microRNAs*.

Traditionally, gene expression levels were measured by the Northern blot method and reverse transcription quantitative PCR (RT-qPCR) before the introduction of microarray technologies. Nevertheless, both of them are low-throughput methods where expression profiling of all the known annotated genes in the human genome is not feasible.

The arrival of microarray technologies has enabled for the first time the interrogation of several thousand genes simultaneously in a single experiment, and whole genome expression studies of all the known genes have also become feasible. Although microarrays have been the method of choice for whole genome gene expression profiling for more than a decade, there are a number of inherent limitations or problems in microarray studies. The conventional gene expression microarrays are mainly focused on the expression levels of known annotated protein-coding genes. Like the ChIP-chip experiment, the developments of tiling arrays where the probes were designed to cover the genome systematically in high resolution regardless of the gene annotation have been used in discovering unknown or novel transcripts, although the cost for tiling arrays is expensive. Besides gene expression microarrays, further developments have also enabled microarrays to be used for studies of alternative splicing and microRNAs expression. Currently, a variety of microarrays is commercially available for transcriptomic applications by companies like Affymetrix® and Illumina®.

The microarray method is based on the hybridization of fluorescent labelled targets and probes, and the expression

levels are inferred indirectly from fluorescent intensity. Therefore, the method suffers from certain levels of cross hybridization and noise, generating artefacts which will complicate the interpretation of results. Sequencing-based approach like SAGE was developed before the arrival of NGS technologies and offered a number of advantages over microarrays, such as the ability to detect novel transcripts and the direct measurement of the abundance of transcripts instead of relying on hybridization intensities. In SAGE, the abundance of mRNAs is estimated or measured by counting of sequence tags derived from the 3′ end of mRNAs. Nonetheless, this method also has several major limitations such as the costly Sanger sequencing and laborious cloning procedure.

The arrival of NGS technologies has brought about another breakthrough in the approaches to explore the transcriptome, and this sophisticated method is known as RNA-Seq. RNA-Seq is based on NGS technologies that offered several advantages over the previous methods like microarray or SAGE for transcriptomic studies. First, unlike the microarray hybridization method, the detection capability of RNA-Seq is not limited by the probes synthesized on the microarray to capture the corresponding transcripts in the transcriptome, but it is instead influenced by the depth of sequencing. Secondly, since RNA-Seq is not based on hybridization to detect and measure transcript expression, it avoids the background noise resulting from cross-hybridization.

Furthermore, RNA-Seq provides the highest resolution to a single-base resolution which precisely maps the transcription boundaries, and it can also identify sequence variations like SNPs in the transcribed regions. In addition, RNA-Seq can be used to study fusion transcripts and alternative splicing. Although special microarrays like exon microarrays where the probes are designed to span exon junctions can be used to study alternative splicing as well, they are subject to inherent limitations of microarray methods.

RNA-Seq directly sequences and maps the transcripts to the reference genome to measure transcript expression by counting the number of sequence reads. Therefore, RNA-Seq has the largest dynamic range of expression levels, from low abundance to highly expressed transcripts (if the sequencing depth is sufficient for low-abundance transcripts). The number of sequence reads that map to a genomic region corresponds to the level of expression from that region. The performance of RNA-Seq has also been evaluated by benchmarking against the gold standard method, that is RT-qPCR for measuring the expression levels, and has been shown to be highly accurate. Besides this parameter, high reproducibility of the results obtained from RNA-Seq has also been shown. Finally, RNA-Seq allows the studying of the expression of mRNAs and noncoding RNAs, and is also able to detect and identify new transcripts (coding and noncoding) that have not been annotated. However, no method is perfect; RNA-Seq is also not without its problems and challenges (Wang *et al*., 2009; Morozova *et al*., 2009).

In summary, besides gene expression profiling, the applications of sequencing-based approaches in transcriptomic studies have been expanded to genome annotation, discovery of new transcripts, investigation of the alternative splicing patterns, detection of gene fusions in cancer, allele-specific expression analysis, as well as the discovery and measurement of noncoding RNA expression (Denoeud *et al*., 2008; Pan *et al*., 2008; Maher *et al*., 2009; Heap *et al*., 2009; Bar *et al*., 2008; Morin *et al*., 2008). The number of publications using sequencing for transcriptomic applications has been growing rapidly. The high-throughput production and significantly cheaper cost are the main factors for NGS technologies to be quickly adopted in transcriptomic studies. The ability to study coding and noncoding RNA expression, alternative splicing, protein–DNA interactions and histone modifications effectively on a genome-wide scale holds a great promise to significantly advance our knowledge in this complex field of transcriptional regulations.

Future Perspectives and Summary

Large-scale sequencing studies have become more feasible and affordable nowadays. In the recent few years since the NGS technologies were introduced, we have seen their tremendous impacts on transforming the approaches in the studies of structural and functional genomics. Moreover, sequencing-based approaches have already yielded numerous novel and important findings in research areas like genome-wide mapping of histone modifications and protein–DNA interactions, discovery of genetic variations, and transcriptomics studies even though the approaches are still new and maturing. These new sequencing technologies have enabled researchers to answer old questions in unprecedented detail and have raised new questions. It has also allowed researchers to design various experiments which were unthinkable just a few years ago with Sanger sequencing.

Further improvements in various aspects of current NGS technologies such as throughput, read length and accuracy and reduction in cost are anticipated. The NGS technologies have shown their potential of being dominant in future genomics studies. It is evident from several international projects using NGS technologies like the ENCODE Project, 1000 Genomes Project and cancers sequencing project by the International Cancer Genome Consortium. Each of the projects has its own specific aim: the ENCODE project aims to annotate all the functional elements, whereas 1000 Genomes Project aims to construct the most comprehensive map of genetic variations in the human genome. The cancers sequencing project intends to study somatic genetic aberrations like point mutations and chromosomal rearrangements in the cancer genome. It is clear that the approaches based on NGS fit in all research areas, and in fact NGS have become an indispensable tool for genomics studies.

It is only a matter of time before achieving the goal of $1000 per whole genome sequencing. This should not be

too far from now given the progresses in the development of third generation sequencing technologies. The arrival of single molecule DNA sequencing technologies like nanopore sequencing will certainly bring about another breakthrough (Gupta, 2008). In fact, a recent study has shown that whole human genome sequencing can be done at a cost of US$4400 using a new sequencing platform (Drmanac *et al*., 2009). Although the $1000 genome will technically make sequencing of thousands of human genomes a reality, the substantial cost that will be incurred for data storage, powerful computational packages and analytical softwares has to be borne in mind. However, beyond affordability, what are left behind are the bioinformatics challenges in processing and analysing the huge amount of sequencing data (Flicek and Birney, 2009; Pepke *et al*., 2009; Pop and Salzberg, 2008).

References

Ahn SM, Kim TH, Lee S *et al*. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research* **19**: 1622–1629.

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology* **25**: 195–203.

Bar M, Wyman SK, Fritz BR *et al*. (2008) MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem Cells* **26**: 2496–2505.

Barski A, Cuddapah S, Cui K *et al*. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* **129**: 823–837.

Bentley DR, Balasubramanian S, Swerdlow HP *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Campbell PJ, Stephens PJ, Pleasance ED *et al*. (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics* **40**: 722–729.

Carninci P and Hayashizaki Y (2007) Noncoding RNA transcription beyond annotated genes. *Current Opinion in Genetics and Development* **17**: 139–144.

Cooper GM, Zerr T, Kidd JM *et al*. (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics* **40**: 1199–1203.

Denoeud F, Aury JM, Da Silva C *et al*. (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biology* **9**: R175.

Drmanac R, Sparks AB, Callow MJ *et al*. (2009) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**: 78–81.

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

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

Farnham PJ (2009) Insights from genomic profiling of transcription factors. *Nature Reviews. Genetics* **10**: 605–616.

Flicek P and Birney E (2009) Sense from sequence reads: methods for alignment and assembly. *Nature Methods* **6**: S6–S12.

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* **26**: 602–611.

Heap GA, Yang JH, Downes K *et al*. (2009) Genome-wide analysis of allelic expression imbalance in human primary cells by high throughput transcriptome resequencing. *Human Molecular Genetics* **19**: 122–134.

Iafrate AJ, Feuk L, Rivera MN *et al*. (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949–951.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Johnson DS, Mortazavi A, Myers RM and Wold B (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* **316**: 1497–1502.

Kim JI, Ju YS, Park H *et al*. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**: 1011–1015.

Korbel JO, Urban AE, Affourtit JP *et al*. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Levy S, Sutton G, Ng PC *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biology* **5**: e254.

Ley TJ, Mardis ER, Ding L *et al*. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**: 66–72.

Li Y and Wang J (2009) Faster human genome sequencing. *Nature Biotechnology* **27**: 820–821.

Maher CA, Kumar-Sinha C, Cao X *et al*. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**: 97–101.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387–402.

McCarroll SA, Kuruvilla FG, Korn JM *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166–1174.

McKernan KJ, Peckham HE, Costa GL *et al*. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* **19**: 1527–1241.

Medvedev P, Stanciu M and Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods* **6**: S13–S20.

Morin RD, O'Connor MD, Griffith M *et al*. (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research* **18**: 610–621.

Morozova O, Hirst M and Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. *Annual Review of Genomics and Human Genetics* **10**: 135–151.

Pan Q, Shai O, Lee LJ *et al*. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**: 1413–1415.

Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews. Genetics* **10**: 669–680.

Pepke S, Wold B and Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. *Nature Methods* **6**: S22–S32.

Pop M and Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24**: 142–149.

Pushkarev D, Neff NF and Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology* **27**: 847–852.

Robertson G, Hirst M, Bainbridge M *et al*. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**: 651–657.

Scherer SW, Lee C, Birney E *et al*. (2007) Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**: S7–S15.

Sebat J, Lakshmi B, Troge J *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Shendure J and Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.

Strausberg RL, Levy S and Rogers YH (2008) Emerging DNA sequencing technologies for human genomic medicine. *Drug Discovery Today* **13**: 569–577.

Tucker T, Marra M and Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. *American Journal of Human Genetics* **85**: 142–154.

Von Bubnoff A (2008) Next-generation sequencing: the race is on. *Cell* **132**: 721–723.

Wang J, Wang W, Li R *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Wang Z, Gerstein M and Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics* **10**: 57–63.

Wheeler DA, Srinivasan M, Egholm M *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–876.

Wold B and Myers RM (2008) Sequence census methods for functional genomics. *Nature Methods* **5**: 19–21.

Yoon S, Xuan Z, Makarov V *et al*. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**: 1586–1592.

Further Reading

ABI® The SOLiD System: http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/index.htm

Illumina® Sequencing Technology: http://www.illumina.com/technology/sequencing_technology.ilmn

Kahvejian A, Quackenbush J and Thompson JF (2008) What would you do if you could sequence everything? *Nature Biotechnology* **26**: 1125–1133.

MacLean D, Jones JD and Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Reviews. Microbiology* **7**: 287–296.

Morozova O and Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**: 255–264.

Roche® 454 Sequencing: http://www.454.com/

Whole Genome Resequencing and 1000 Genomes Project

Ku Chee-Seng, *Centre for Molecular Epidermiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Loy En Yun, *Centre for Molecular Epidermiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Pawitan Yudi, *Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden*

Chia Kee-Seng, *Centre for Molecular Epidermiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Advanced article

Article Contents

Online posting date: 19th April 2010

The recent advances in sequencing technologies have enabled the whole human genome to be sequenced within weeks. To date, several human diploid genomes have been sequenced and the number of genomes being sequenced is expected to increase in the years to come. In fact, a 3-year international collaborative project, the 1000 Genomes Project, was initiated in 2008 to sequence at least 1000 individual genomes from different populations around the world. The aim is to create the most detailed and comprehensive map of genetic variations in the human genome for future disease-association studies and biomedical research. While waiting for this ambitious project to be completed, several whole genome sequencing studies have already provided some exciting results, where hundreds of thousands of new SNPs and short indels have been identified. In addition, these studies also address many important questions and issues in the experimental design and data analysis of whole genome sequencing.

Introduction

The arrival of *next generation sequencing* (*NGS*) and third generation sequencing technologies has enabled the *sequencing* of the whole human genome to be completed within weeks compared to more than a decade taken by the Human Genome Project (HGP). The first human whole genome sequencing (WGS) study using a next generation sequencer was completed in 2008, which marked the beginning of a new era in personalized genome sequencing. To date, seven WGS studies have been done using NGS technologies; other than James Watson's genome, all are anonymous individuals. The number of genomes being sequenced is expected to increase in the coming years when sequencing technologies and analysis tools are further advanced and become even more feasible and affordable. These next generation technologies enable WGS to be finished at an unprecedented speed. More importantly, the NGS technologies also allow the whole human genome to be sequenced at a cost of a few hundreds of thousand dollars or less which is only a small fraction of the three billion US dollars spent by the HGP. Certainly, the cost of WGS will decline steeply in the years to come, especially with forceful competition from several third generation sequencers which are expected to be in the market soon, as ultimately the goal is to reduce the cost to 1000 dollars per genome sequencing.

Nevertheless, it is noteworthy that the WGS studies would not have been feasible today without the human reference genome or DNA (deoxyribonucleic acid) sequence provided by the HGP. The reference genome sequence is needed for alignments of the massive amount of sequence reads produced by the next generation sequencers. These studies are not *de novo* assembly of the human genome sequence, but rather, resequencing studies. In spite of the hefty cost of the HGP, we have amassed extensive knowledge from the project, and rapid developments have occurred in the studies of structural and functional genomics as the finished human reference genome sequence was in hand. Besides the human genome, genomes of plants, animals and microorganisms were also getting sequenced. However, this article focuses on human genome sequencing and discusses the deliverables and impacts of the studies.

ELS subject area: Genetics and Disease

How to cite:
Chee-Seng, Ku; En Yun, Loy; Yudi, Pawitan; and Kee-Seng, Chia (April 2010) Whole Genome Resequencing and 1000 Genomes Project. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0022507

Human Genome Sequencing: Past, Present and Future

The field of molecular genetics and genomics has been progressing rapidly after the completion of the HGP, which provided a reference DNA sequence of the human genome (International Human Genome Sequencing Consortium, 2004). It was a major scientific development in human genomics and biomedical sciences. This ambitious work was started in 1990 and took approximately 13 years to complete (**Figure 1**). On its completion, HGP offered the first glimpse of the number of protein-coding genes; it was estimated to be 20 000–25 000 genes embedded in the ~3 billion nucleotides which comprise the human *haploid genome*. Recent data seem to suggest that the 'real number' is approximately 20 500 genes. Withal, none of the genome experts have cast a good guess close to this number back in the year 2000; even the lowest estimated number was still several thousands of genes higher than this real number (Pennisi, 2007). The massive sequencing task in the HGP was completed by the Sanger dideoxynucleotide or chain termination sequencing method which was developed in the late 1970s.

In competition with the International Human Genome Sequencing Consortium (IHGSC), a private company – Celera Genomics – also finished their human genome sequencing project in the same year, where both published their draft sequences and data analysis in 2001 (International Human Genome Sequencing Consortium, 2001; Venter *et al*., 2001). Nevertheless, the draft or initial sequences were not flawless; several imperfections in the draft sequences were generated by both groups. The draft sequences were far from perfect because of the incomplete coverage of the *euchromatic regions or euchromatin*, where ~10% of these regions were missing. The coverage was even less when the whole genome is considered which includes the heterochromatic regions; some 30% of the genome was not covered. Furthermore, there were an excessive number of gaps between the *contigs* which made the genome sequences patchy or discontinuous. Therefore, there was a need to improve the draft sequence, and subsequently the IHGSC published an improved version of the human genome sequence in 2004 and the Human Genome Project was deemed to complete at that time. This improved version has achieved a nearly complete coverage of all the euchromatic regions in the human genome (~99%) and also significantly reduced the number of gaps to 341 from the initial hundreds of thousands of gaps (International Human Genome Sequencing Consortium, 2004). **See also**: Comparing the Human and Chimpanzee Genomes; Human Genome Project: Importance in Clinical Genetics; Sequencing the Human Genome: Novel Insights into its Structure and Function

Although both the HGP and Celera Genomics only sequenced the human haploid genome, the availability of the reference DNA sequence has marked an important milestone in genomic research. It has initiated a new era in the studies of *genetic variations* and the functional

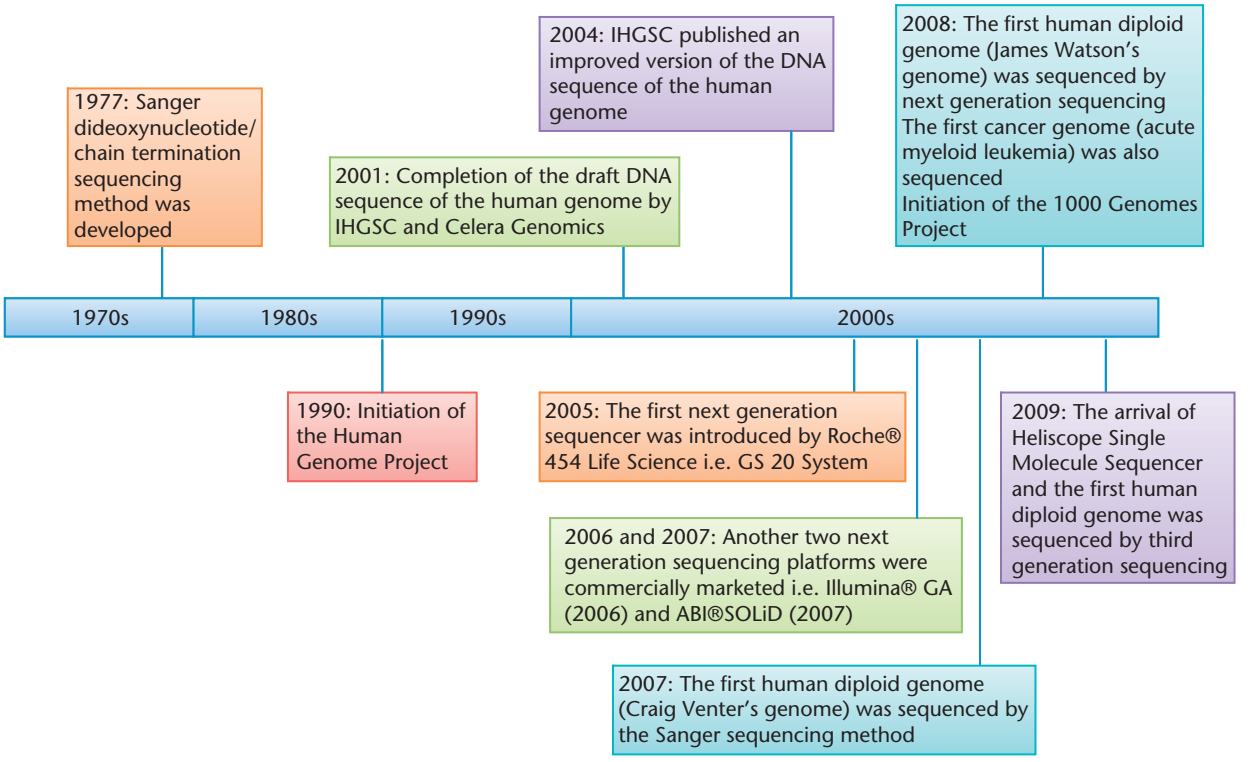


Figure 1 The developments of sequencing technologies and whole human genome resequencing studies.

characterization of the human genome. The two global projects that ensued are the International HapMap (Haplotype Map) Project and the ENCODE (Encyclopedia of DNA Elements) Project (International HapMap Consortium, 2003; ENCODE Project Consortium, 2004). The aim of the HapMap initiative is to validate millions of single nucleotide polymorphisms (SNPs) that are identified during and after the completion of the HGP, and then to characterize their correlation or linkage disequilibrium (LD) patterns in populations of European, Asian and African ancestry. Apart from decoding the blueprint of life and identifying genetic variation (particularly the SNPs), it is also crucial to understand the biological code and message embraced in the ~ 3 billion nucleotides. To advance knowledge in these areas, the ENCODE Project was conceived to identify all the functional and regulatory elements in the human genome. Both these projects have achieved significant successes and generated important information in each area (International HapMap Consortium, 2007; ENCODE Project Consortium, 2007). **See also**: HapMap Project

The first human diploid genome sequence appeared in 2007 and it was the Craig Venter's genome that was sequenced by the Sanger sequencing method (Levy *et al*., 2007). In the subsequent year, the genome of James Watson, who discovered the double-helix structure of DNA molecule half a century ago (Wheeler *et al*., 2008), was also sequenced. In contrast to Venter's genome, Watson's genome was sequenced using NGS technologies. Today, on top of the two scientists' genomes, an additional six human genomes have also been fully sequenced, five of them being anonymous individuals with no phenotypic or medical information, and one being a patient with acute myeloid leukaemia (AML) (Ley *et al*., 2008). These anonymous individuals are one Caucasian/European (Pushkarev *et al*., 2009), one African (NA18507) from the International HapMap Project who was sequenced in two studies (Bentley *et al*., 2008; Mckernan *et al*., 2009), two Koreans (Ahn *et al*., 2009; Kim *et al*., 2009) and one Han Chinese (Wang *et al*., 2008). The sequencing work of all these individual genomes was accomplished by next generation sequencers, except one that was done by a 'next-next' or third generation sequencer which is based on single DNA molecule sequencing (Pushkarev *et al*., 2009). The NGS technologies are Roche® 454 Life Science Genome Sequencer FLX (GS FLX), Illumina® Genome Analyzer (GA) and Applied Biosystems® (ABI) Supported Oligonucleotide Ligation Detection System (SOLiD).

Next Generation Sequencing Technologies

Sanger dideoxynucleotide chain termination sequencing has been the most common sequencing method used for the past 30 years since it was invented in the late 1970s until the first NGS technologies were introduced to the market in 2005. Sanger sequencing is used for various applications such as mutation detection and SAGE (serial analysis of gene expression) for measuring transcript levels, and more importantly, it was used to complete the HGP. Shortly after the first next generation sequencer was introduced by Roche® 454 Life Science, that is Genome Sequencer 20 (GS 20) System, which was subsequently replaced by GS FLX System with further improvements to the preceding system, another two biotechnology companies also launched their sequencing platforms, that is Illumina® GA and ABI® SOLiD. The simultaneous advent of several next generation sequencers has created intense competition in the sequencing market, with each technology having its own strengths and limitations.

Currently, Sanger sequencing machines have been largely supplanted by next generation sequencers in many large genomics institutes worldwide. This is mainly due to the ultra high-throughput production of NGS technologies which is several orders of magnitude higher than traditional sequencing. One of the major differences between the two of them is the ability of next generation sequencers to simultaneously sequence millions of DNA fragments; therefore, NGS is also known as massively parallel sequencing technologies. This feature has enormously increased the throughput production or the number of nucleotides that one can sequence when compared to the Sanger sequencer in one experiment or per instrument run. The sequencing chemistry of NGS technologies, together with their ultra high-throughput production, has also reduced sequencing cost significantly, making large-scale sequencing studies affordable nowadays. The development of third generation sequencing technologies is expected to further decrease in the sequencing cost and eventually achieve the goal of 1000 dollars per sequenced genome (Mardis, 2006).

One of the tedious steps in WGS using the Sanger method is the *in vivo* amplification step using bacterial cloning. This has been substituted by *in vitro* amplification of millions of DNA fragments by NGS technologies using emulsion polymerase chain reaction (PCR) (Roche® GS FLX and ABI® SOLiD) or bridge amplification on solid surface (Illumina® GA). The third generation sequencing is characterized by single DNA molecule sequencing without the need of amplification. The first third generation sequencing instrument – Heliscope Single Molecule Sequencer – is now commercially marketed by Helicos Biosciences. The sequencing chemistry or approach for NGS technologies can be broadly divided into sequencing-by-synthesis (pyrosequencing for Roche® GS FLX, and sequencing by reversible terminator chemistry for Illumina® GA) and sequencing-by-ligation (ABI® SOLiD).

Nowadays WGS is able to be completed quickly, but it would not be possible without the HGP that provides the template for alignment of billions of short sequence reads produced by next generation sequencers. This is because the NGS technologies are characterized by short sequence read length, less than 100 bp for Illumina® GA and ABI® SOLiD, as well as for the third generation sequencer. This

feature makes the *de novo* sequencing or assembly of billions of short sequence reads into large contigs a difficult task, especially for large and complex genomes like the human genome. Longer read length is crucial to obtain larger contigs with fewer gaps between them during the assembly steps. Although the latest improvements in sequencing chemistry and system allow Roche® GS FLX to achieve a sequence read length of 500 bp on average, there is still a sizable gap from the length that can be achieved by Sanger sequencing, which is approximately 800 bp–1 kb (Rothberg and Leamon, 2008; Shendure and Ji, 2008; Gupta, 2008). NGS technologies have many advantages over the traditional sequencing method, but they are not without limitations. In addition to short read length, they have higher sequence error rates, although this has been gradually improving.

NGS technologies have not only accelerated the sequencing speed exponentially, they have also remarkably changed the approaches in genomics studies. In addition to their applications in DNA sequencing, they also have innovative applications in other areas such as genome-wide mapping of transcription factor binding sites and histone modifications (*ChIP-Seq*), transcriptomic profiling of mRNAs (messenger ribonucleic acid) and noncoding RNAs (RNA-Seq), studies of alternative splicing events, detection of chromosomal alterations or rearrangements (paired-end sequencing), direct sequencing of CpG methylation sites and metagenomics (Morozova and Marra, 2008; Mardis, 2008).

Deliverables from Whole Genome Sequencing

Recent advances in sequencing technologies have permitted WGS to be done efficiently. With current and future sequencing technologies, there should be no major obstacle in getting anyone's genome sequenced quickly. But what is more important is to be aware of the value and significance of getting these genomes sequenced, what can be learned from the sequencing data and how the results will impact future genetic studies.

Experimental design and data analysis

Several important findings and favourable outcomes have been delivered by the WGS studies. These studies have clearly demonstrated the feasibility of using all the NGS and third generation sequencing technologies to decode the DNA sequence of human genome efficiently and at an affordable price per genome. This should be scalable to sequence hundreds or thousands of genomes when the cost drops further. In addition, these studies have also addressed many important questions and issues surrounding the experimental design and data analysis, such as preparation of single- and paired-end DNA libraries for sequencing, assessment of the sequencing depth that is needed to achieve near complete coverage of the reference genome sequence and to minimize SNP calling error rate, and the quality control criteria for detection of genetic variations (SNPs, *indels* and *structural variations*). For example, Wang *et al*. (2008) found that at a sequencing depth of greater than 10-fold, the assembled consensus sequence covered $\sim$83% of the NCBI human reference genome sequence using single-end reads and $\sim$95% coverage using paired-end reads, and greater sequencing depth has minimally increased the coverage. However, the SNP calling error rate decreases significantly with greater sequencing depth. Similar conclusions were also derived from other studies such as that of Bentley *et al*. (2008), who reported that the discordances between sequence-based SNP calls and data from genotyping arrays reduced with increasing sequencing depth. The discordance genotypes are mostly heterozygotes that were under-called due to low sequencing depth, because a sufficient sequencing depth is needed to detect both the alleles. This piece of information would be useful and helpful for future studies using the same sequencing platform to balance between coverage, error rate in SNP calling and cost of sequencing.

The subsequent WGS studies that used shorter read lengths have higher average sequencing or coverage depth than the first study using Roche® GS FLX. Average coverage depths of 36-fold and 27.8-fold were achieved in the YH and AK1 genomes, respectively, using Illumina® GA, compared to the genomes that were sequenced by longer read length, such as 7.4-fold for Watson's genome and 7.5-fold for Venter's genome. Higher sequencing depth is needed for next generation sequencers that produce shorter read lengths to attain more complete genome coverage. With that sequencing depth, the aforementioned two studies achieved almost complete coverage of the NCBI human reference genome. Although, it is bit lower for the P0 genome that was sequenced by the third generations sequencer, it was reported that only $\sim$90% of the reference genome sequence was covered at 28-fold average coverage.

In addition, several single- and paired-end libraries (with different insert sizes) are usually prepared and sequenced to minimize the systematic biases in genome representation. Using the same studies as examples, Wang *et al*. (2008) prepared and sequenced eight single- and two paired-end libraries for the YH genome. Similarly Kim *et al*. (2009) also prepared several libraries with different insert sizes for the AK1 genome to provide even coverage or to minimize coverage biases, whereas three paired-end libraries with span sizes of 100, 200 and 300 bp were prepared for the other Korean genome (SJK).

The studies also examined the performance of next generation sequencers from several aspects, such as the percentage of uniquely aligned sequence reads to distinguish from sequence reads that aligned to multiple sites in the reference genome sequence. This is because only the uniquely aligned sequence reads are used to build the consensus sequence and to detect genetic variations, and multiple-site aligned reads are filtered because they are ambiguous. This aspect is crucial for next generation

sequencers, because aligning short sequence reads uniquely to the reference genome could be problematic. It is wasteful if a large fraction of the sequence reads generated by NGS is aligning to multiple sites. Nevertheless, we have seen some favourable outcomes as Wang *et al*. (2008) reported that approximately 86% of the sequence reads that mapped to the reference genome could be uniquely aligned.

In addition, these studies have also led to the developments of many algorithms for aligning the sequence reads and detecting the genetic variations. Some studies have used their in-house developed algorithm, such as SOAP (Short Oligonucleotide Alignment Program) to align billions of sequence reads to the reference, and the improved version has significantly increased the alignment speed (Li *et al*., 2009a). Others have used alignment tools that are developed externally, like the MAQ (Mapping and Assembly with Qualities) (Ahn *et al*., 2009). Besides the advent of efficient alignment softwares, there is also a surge in the publications of analysis tools for processing and quality assessment of the sequencing data (Martinez-Alcantara *et al*., 2009; Lassmann *et al*., 2009), methods for detecting CNVs and structural variations using NGS data (Yoon *et al*., 2009; Chen *et al*., 2009) and SNP detection methods (Li *et al*., 2009b). If the amount of genotyping data in genome-wide association studies (GWAS) can be described as 'drinking from the fire hose' (Hunter and Kraft, 2008), then in the WGS era, perhaps it is 'drinking from the waterfall'. It is indeed true that one of the greatest challenges in the WGS era is the handling of massive amounts of sequencing data, and fortunately, the developments of analysis methods and programs are keeping pace with the increasing throughput production of next generation sequencers. These achievements, together with the established experimental procedures and analyses pipelines, have greatly contributed to the maturity of the field.

Richness of genetic variations

The more significant finding from the WGS studies is that they have conclusively revealed the richness of genetic variations in the human genome (**Table 1**). Although the ubiquity of CNVs in the human genome has been reported back in 2004 (Sebat *et al*., 2004; Iafrate *et al*., 2004), the CNVs found in those studies were far less than the numbers reported in the WGS studies. Most of the CNV data were generated by array-based methods (CGH and SNP arrays) where the signal intensity information is used to detect deletions and duplications. Because of the reliance on relative or differences in signal intensities, these methods are unsuitable for detecting other structural variations like inversions and translocations (also known as balanced chromosomal rearrangements). Furthermore, due to limitations in marker density or resolution, these methods are not sensitive enough to detect smaller sizes of CNVs (<50 kb) (Scherer *et al*., 2007), which are predicted to be more abundant than the larger CNVs (Estivill and Armengol, 2007). Even using the highest density SNP arrays such as Illumina® Human 1M Beadchip and Affymetrix® 6.0 SNP Arrays, the method still suffers from poor sensitivity to detect CNVs smaller than 5–10 kb (McCarroll *et al*., 2008a; Cooper *et al*., 2008). Thankfully, the completion of several WGS studies has provided much information on the genetic diversity of human genome.

Several thousands of structural variations are found in all the WGS studies although they have used different detection methods and criteria. In total, 2682 structural variations (dominated by CNVs) are detected in the YH genome with a median length of ~ 0.5 kb; their sizes are much smaller than those identified by array-based methods. These results show that paired-end sequencing methods have much higher sensitivity to detect smaller CNVs. Nonetheless, this method could be extremely biased towards detection of deletions, where most of the identified CNVs are deletions, which are 2441 deletions versus 33 duplications. Moreover, this clearly reveals the limitations of paired-end sequencing methods with certain insert sizes in detecting CNVs. Deletions are more likely to be detected because they are identified by longer than expected paired-end insert sizes when mapped to the reference sequence, whereas detection of duplication is restricted by the paired-end library span size. This means that duplications or insertions larger than the paired-end insert size are undetected. Therefore, several paired-end libraries with different (short and long) insert sizes would be needed to capture duplications or insertions and deletions of varying sizes. For the YH genome, the two paired-end libraries had a span size of 135 and 440 bp, respectively. The bias against detection of duplications is partly due to the paired-end insert size; therefore, larger insert sizes of several kilobases should improve the ability to detect more duplications. **See also**: Copy Number Variation in the Human Genome; Epigenetic Variation in Humans; Genetic Variation: Human; Relevance of Copy Number Variation to Human Genetic Disease; Single Nucleotide Polymorphism (SNP)

In addition to structural variations, the studies also identified hundreds of thousands of indels (**Table 1**). The numbers reported are not directly comparable to each other, as the analyses, detection methods and criteria used are different between the studies. For the two Korean genomes, the number of indels found in one study is twice the other one. Ahn and colleagues identified the indels within a size range from -29 to $+14$ bp, whereas the other study detected the indels within -29 to $+5$ bp. The YH genome contained approximately 135 000 indels within 1 to 3 bp, and approximately 400 000 indels defined from 1 to 16 bp were found in the NA18507 genome. All these four genomes were sequenced using Illumina® GA. Regardless of the numbers, collectively these sequencing studies have definitely revealed plenty of short indels in the human genome. The reported numbers are likely representing only a portion of the total number of indels in the human genome, because a rather narrow size range was defined in each of the studies described here to identify indels.

Table 1 Summary of genetic variations identified in the whole genome sequencing studies

Study, individual genome and sequencing technology	Total number of SNPs detected	In dbSNP new SNPs	Total number of indels detected	In dbSNP new indels	Total number of copy number or structural variations detected
Levy *et al*. (2007) Craig Venter Sanger sequencing	3 213 401	91% in dbSNP 9% New	292 102 Heterozygous indels (size range: 1–571 bp) 559 473 Homozygous indels (size range:1–82 711 bp)	–	62 Copy number variable regions (microarray-based methods) 90 Inversions 53 823 Block substitutions or multinucleotide polymorphisms (size range: 2–206 bp)
Wheeler *et al*. (2008) James Watson Roche® GS FLX	3.32 Million	2.72 Million in dbSNP (82%) 0.61 Million new (18%)	222 718 (65 677 Insertions, 157 041 deletions, size range: 2–38 896 bp)	113 539 in dbSNP (51%) 109 179 new (49%)	No mate-paired reads for detecting structural variants 23 CNV regions (size range: 26 kb–1.6 Mb) were detected by a CGH array
Ahn *et al*. (2009) Korean – SJK Illumina® GA	3 439 107	3 019 024 in dbSNP (88%) 420 083 New (12%)	342 965 (Size range: −29–+14 bp)	113 534 in dbSNP (33%) 229 431 New (67%)	2920 Deletions 415 Inversions 963 Insertions
Kim *et al*., 2009 Korean – AK1 Illumina® GA	3 453 653	2 863 078 in dbSNP (83%) 590 575 New (17%)	170 202 (Size range: −29–+5 bp)	38% in dbSNP 62% New	1237 CNV regions (deletions) 77 Copy number gains
Mckernan *et al*., 2009 African – NA18507 ABI® SOLiD	3 866 085	3 131 423 in dbSNP (81%) 734 662 New (19%)	226 529 89 679 Insertions of up to 3bp 124 024 Deletions of up to 11bp 12 826 Larger indels	67% of the small Indels found (insertions up to 3 bp and deletions up to 11 bp) are present in dbSNP	5590 Indels between mate-paired reads (1515 Insertions and 4075 deletions) 91 Inversions
Bentley *et al*. (2008) African – NA18507 Illumina® GA	4 Million	74% in dbSNP 26% New	0.4 Million (Size range: 1–16 bp)	Half of the indels are corroborated by entries in dbSNP	5704 Structural variants
Wang *et al*. (2008) Han Chinese – YH Illumina® GA	3.07 Million	2 657 081 in dbSNP (86%) 417 016 New (14%)	135 262 (Size range: 1–3 bp)	55 390 in dbSNP (41%) 79 872 New (59%)	2682 Structural variants
Pushkarev *et al*. (2009) European – P0 Heliscope Single Molecule Sequencer	2 805 471	76% in dbSNP 24% New	–	–	752 Regions of CNV

Although, the finding of several million SNPs in each individual genome is not new; more interesting is the identification of several hundred thousands of new SNPs in all the studies that have not been catalogued in dbSNP. Bentley *et al*. (2008) found about one million new SNPs in the NA18507 genome, and more or less half a million for other genomes. Therefore, these results still suggest a lack of completeness of the current dbSNP, even though SNPs are the most well characterized genetic variations in the human genome. Most of the common SNPs in human populations have already been captured; thus the new SNPs identified in each study are likely representing those from the spectrum of lower frequencies. The information about the population frequencies of the new SNPs is unavailable, since they are individual genome sequencing studies, but it should be available through the 1000 Genomes Project when it is ready.

The WGS studies also assessed the accuracy of their SNP calls by comparing with the data from genotyping arrays, and found an excellent concordance between them. The accuracy of SNP calling in the SJK genome was assessed using Illumina® 1M and Affymetrix® 6.0 arrays, and it was found that >99% of the SNP calls were in agreement. However, Mckernan *et al*. (2009) used a different approach instead of relying on commercial genotyping arrays to validate the SNP calls. It is arguable that the SNPs in genotyping arrays are well selected; therefore, it could bias the comparisons. Therefore, they chose to focus on validation of new SNPs, instead of the known SNPs in the genotyping arrays. They randomly selected a small fraction of the novel SNPs and validated them using SNPlex genotyping assays. They found >95% agreement between the sequencing detected genotypes and the SNPlex genotypes, but >99% agreement with the HapMap genotype data. Finally, Bentley *et al*. (2008) used both approaches to examine the accuracy of SNP calling, by comparing the sequence-based SNP calls with genotyping arrays, HapMap genotype data, as well as validating a subset of new SNPs. Good concordance was found in all the three analyses.

The differences in the number of genetic variations identified in the WGS studies are likely to be the consequence of technical and analytical differences rather than due to genuine interindividual variability. It is apparent from the NA18507 genome which was sequenced by Illumina® GA and ABI® SOLiD; both studies found an appreciable difference in the number of SNPs, indels and structural variations. The number of SNPs identified by both studies can differ as many as 0.2 million SNPs even if it is the same individual genome that gets sequenced twice (**Table 1**).

Although technical and analytical differences between the studies have to be borne in mind, there is still a significantly higher number of SNPs found in the African's genome (NA18507) by the two studies when compared to other genomes. Yet, this is not an unexpected finding because the Africans are ancient populations, and have greater genetic diversity than European and Asian populations. Other interesting results include the finding of highly significant pairwise correlations of SNP and indel densities throughout the genome (Kim *et al*., 2009).

Comparing the genetic variations between genomes

Besides identifying genetic variations in each genome, the studies also compared the genetic variation data with other genomes, such as the proportion of SNPs that are shared with other genomes and the number of SNPs that are unique to each genome, to provide some insights to the genetic differences between individuals from distinct ancestries. For instance, Kim *et al*. (2009) found that 21% of the AK1's SNPs were unique, that were not found in the other four genomes that were sequenced before it (Venter, Watson, NA18507 and YH), although 8% of SNPs were shared by all. Ahn and colleagues also made similar comparisons for the other Korean genome; SJK shared 50–60% of the SNPs with the genomes of YH, Venter, Watson and NA18507. Owing to the limited number of genomes sequenced, for the time being, it is only feasible to compare between individual genomes. More meaningful comparisons should instead involve a group of individuals from each population to interrogate the extent of similarities and differences in genetic variations between different populations like the International HapMap Project. The International HapMap only focused on (common) SNPs, but the 1000 Genomes Project should enable a much more detailed population comparison of various genetic variations.

Genetic variations and complex diseases

Since SNPs have been well studied for disease association in both genome-wide and candidate gene association studies, it would be interesting to know whether the SNPs identified in these genomes are associated with any of the diseases. In fact, a list of such disease-associated SNPs has been compiled by the studies. More than 100 SNPs were identified in the genome of AK1 that have shown association with various complex diseases such as 90 SNPs for cancers, 34 SNPs for type 2 diabetes, 13 with Alzheimer disease and 7 for rheumatoid arthritis. Although the findings are exciting, their interpretations could be challenging, unless most of the disease genetic variants (both the protective and predisposing risk factors) are known, and their interactions with environment factors are also being characterized. This is because the development of complex diseases depends on both the genetic and environment factors and their interactions. Nevertheless, these findings further support the need to catalogue all the genetic variations in the human genome as the first step for future disease association studies.

New sequences

The WGS studies also found a portion of the sequence reads that could not map to the NCBI human reference

genome, indicating that some sequences are missing from the reference. Wheeler and colleagues found 1.5 million reads (approximately 1.4% of the total sequence data) that did not map to the reference. These 'unmappable' sequence reads were then assembled into ~170 000 contigs spanning 48 Mb. Even after the removal of contigs that were less than 100 bp in size, there were still ~110 000 contigs spanning 29 Mb. This agrees with the estimation of 25 Mb of euchromatic sequence that is absent from the reference. For the SJK genome, ~6% of the sequence reads were not mapped to the reference. Meanwhile, Wang *et al*. (2008) also assembled approximately 1.7 million reads into nearly 21 000 contigs with lengths larger than 100 bp. Although several technical factors could be responsible for the unmappable sequence reads, it is equally likely to be due to the missing sequences in the reference.

1000 Genomes Project

The 1000 Genomes Project was initiated in 2008 to sequence the genomes of at least 1000 individuals from different populations around the world (http://www.1000genomes.org/page.php). The major aim of this international collaborative project is to provide the most detailed and comprehensive map of human genetic variations. The participants are anonymous individuals and no medical information is collected since the aim of this project is to build a useful resource of human genetic variations for future disease association studies, rather than correlating the genetic information with disease phenotypes. Like the International HapMap Project, the data from this project will also be made publicly available to researchers and the scientific community (Kaiser, 2008).

The significance of this project for future disease association is tremendous. Owing to the ease of large-scale genotyping, SNPs have been widely used as genetic markers in GWAS to search for disease variants. Moreover, evidence has been accumulating to suggest that (common) SNPs alone are unlikely to account for all the heritable risk of complex diseases. Concurrently, the amount of data showing the associations of CNVs with complex diseases has been growing (Wain *et al*., 2009). Similarly, the importance of rare variants in complex diseases is also being recognized (Nejentsev *et al*., 2009). This indicates that future disease association studies need to interrogate non-SNP and rare genetic variants, and for this to be feasible, a detailed catalogue of human genetic variations is needed. Common SNPs have been well documented in the dbSNP, but rarer SNPs (or lower frequencies SNPs) are still under-represented in the database and the information of indels and structural variations are far from complete.

Unlike the WGS of one individual genome, the 1000 Genomes Project is a large-scale population-based sequencing study which enables studies of the population properties of genetic variations and their LD patterns. This information will be required to design next generation genotyping arrays to select surrogate markers that are not only able to tag for SNPs through LD, but also to efficiently capture indels and CNVs. This development will certainly broaden the scope of genetic variations interrogated in GWAS. In fact, there is already some evidence to show that CNVs could be tagged by SNPs through LD (Hinds *et al*., 2006; McCarroll *et al*., 2008a), but a detailed and in-depth investigation of their LD patterns can only be done when most of the SNPs, indels and structural variations are identified. The in-depth studies of LD between different genetic variations is important, as the finding of the 20 kb deletion located upstream of the IRGM gene for Crohn disease has demonstrated the efficiency of using SNPs as surrogate markers to identify non-SNP genetic variants (McCarroll *et al*., 2008b).

Sequencing of Cancer Genome

Many complex diseases have been interrogated in GWAS over the past few years; these diseases include many types of cancers such as prostate, breast, lung, ovarian and colorectal cancers, acute lymphoblastic leukaemia, follicular lymphoma and glioma among others. These studies have successfully compiled a list of germline SNPs that confer susceptibility to various cancers. However, most of the risk alleles have small effect sizes (odds ratio <1.5) which only explain a small fraction of the total heritable risk (Easton and Eeles, 2008). **See also**: Genome-Wide Association Studies

Cancers are different from other complex diseases in several aspects. In addition to germline genetic variations, the importance of *somatic mutations* in carcinogenesis is well established and recognized. Cancer is caused by the accumulation of genetic lesions or somatic mutations (over the lifetime of patient) in the genome of the cell where the cancer originates from. Therefore, focusing merely on germline genetic variations will not be sufficient to fully decipher the genetic basis of cancers. Nonetheless, identifying somatic mutations or variations is challenging because it requires the specific cell population for investigation. Unlike germline variations which are inherited in all cells of the human body, the acquired mutations are not shared by all types of cells. As a result, to study somatic mutations, DNA has to be obtained from the specific cell population where the disease arises. The original cells where most cancers arise are known; but for other diseases like schizophrenia or hypertension, it is unclear which cell population to be studied for somatic mutations. Furthermore, tissue specimens are readily available for cancers (compared to other diseases) where DNA is extracted. All these factors provide opportunities to study somatic mutations in cancers, but the genotyping arrays used in GWAS are designed to interrogate germline SNPs. Therefore, direct sequencing of the cancer genome would be needed to study somatic mutations.

In the recent years, a number of targeted sequencing studies have been undertaken and have identified an enormous number of somatic mutations in various cancers

(Wood *et al*., 2007; Ding *et al*., 2008; Prickett *et al*., 2009). One of the notable studies was the Cancer Genome Project which conducted a systematic sequencing of the exons of protein kinase genes in various human cancers and found more than one thousand somatic mutations (Greenman *et al*., 2007). Previous studies have used PCR to isolate and amplify the targeted regions. This method is tedious, laborious and time consuming for isolating large genomic regions of hundreds of megabases. Nevertheless, the developments of several sequences capture or enrichment methods to isolate the targeted genomic regions for sequencing has allowed large-scale targeted sequencing studies to be done more efficiently (Summerer, 2009). Moreover, several international collaborative projects have also been initiated to decipher the cancer genome through large-scale sequencing approaches such as the International Cancer Genome Consortium (ICGC). The ICGC aims to eventually sequence the full genome of many thousands of cancer samples of various types, but for the near term, a targeted approach is being used to sequence only the exons (Maher, 2009).

Whole genome sequencing is the ultimate and complete solution to unravel disease genetic variants regardless of whether they are germline polymorphisms or somatic mutations, SNPs or non-SNP genetic variants. Although sequencing cost is decreasing, the WGS approach is still prohibitively expensive to be applied in a large sample set of cancer and noncancer samples; therefore, a targeted sequencing strategy is more feasible and being advocated at the moment. So far, there is only one WGS study of cancer. The WGS of the AML genome has provided some preliminary yet exciting findings, and also demonstrated the power of this approach to discover new cancer-associated mutations. For the time being, they focused primarily on the coding sequences of annotated genes, and detected eight new heterozygous and nonsynonymous somatic mutations (single nucleotide changes) in the AML genome. Interestingly, some of the mutations are found in the genes involved in several pathways that are known to contribute to cancer pathogenesis. For example, four of the mutations are found in the genes that are strongly associated with cancer progression (i.e. protein tyrosine phosphatase, receptor type, T (PTPRT), cadherin 24, type 2 (CDH24), protocadherin LKC (PCLKC) and solute carrier family 15 (oligopeptide transporter), member 1 (SLC15A1)). In addition to providing some preliminary results, this study can be used as a reference for experimental design and data analysis in the future cancer WGS studies (Ley *et al*., 2008).

Summary

There is still a considerable huge gap to move from 'personalized genome sequencing' to 'personalized medicine'. Personalized genome sequencing is able to provide full DNA sequences and identify the enormous number of genetic variations in the genome. However, personalized medicine aims to predict individual susceptibility risks to various diseases and responses to drug therapies using the genetic variation information. Therefore, it is essential to know beforehand which genetic variations are neutral and which of them are disease-causing variants. To bridge the gap, the first steps are to detect and validate all the genetic variations in the human genome in population-based studies, and catalogue them properly in databases, so they can be used as the genetic markers for future disease association studies. Currently, we are moving towards these goals with the on-going 1000 Genomes Project. These recourses would be needed for future GWAS, and only with the availability of a very detailed and near complete map of all genetic variations will it be feasible to perform a truly comprehensive search for the disease causing variants throughout the human genome. The current GWAS that target only common SNPs have a limited representation of the total genetic variations.

Identifying all the 'single effect' disease variants is the preceding step to study their 'combined effect' or gene–gene interactions. There will only be some hope for personalized medicine such as disease risk prediction when most, if not all, of the disease variants are known. Sequencing-based methods will become an indispensable tool for future genetic association studies as well as functional genomics.

References

Ahn SM, Kim TH, Lee S *et al*. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Research* **19**: 1622–1629.

Bentley DR, Balasubramanian S, Swerdlow HP *et al*. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Chen K, Wallis JW, McLellan MD *et al*. (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**: 677–681.

Cooper GM, Zerr T, Kidd JM *et al*. (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics* **40**: 1199–1203.

Ding L, Getz G, Wheeler DA *et al*. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**: 1069–1075.

Easton DF and Eeles RA (2008) Genome-wide association studies in cancer. *Human Molecular Genetics* **17**: R109–R115.

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636–640.

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799–816.

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

Greenman C, Stephens P, Smith R *et al*. (2007) Patterns of somatic mutation in human cancer genomes. *Nature* **446**: 153–158.

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends in Biotechnology* **26**: 602–611.

Hinds DA, Kloek AP, Jen M *et al*. (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genetics* **38**: 82–85.

Hunter DJ and Kraft P (2008) Drinking from the fire hose: statistical issues in genomewide association studies. *New England Journal of Medicine* **357**: 436–439.

Iafrate AJ, Feuk L, Rivera MN *et al*. (2004) Detection of large-scale variation in the human genome. *Nature Genetics* **36**: 949–951.

International HapMap Consortium (2003) The International HapMap Project. *Nature* **426**: 789–796.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431**: 931–945.

Kaiser J (2008) A plan to capture human diversity in 1000 genomes. *Science* **319**: 395.

Kim JI, Ju YS, Park H *et al*. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**: 1011–1015.

Lassmann T, Hayashizaki Y and Daub CO (2009) TagDust: a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* **25**: 2839–2840.

Levy S, Sutton G, Ng PC *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biology* **5**: e254.

Ley TJ, Mardis ER, Ding L *et al*. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**: 66–72.

Li R, Yu C, Li Y *et al*. (2009a) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966–1967.

Li R, Li Y, Fang X *et al*. (2009b) SNP detection for massively parallel whole-genome resequencing. *Genome Research* **19**: 1124–1132.

Maher B (2009) Exome sequencing takes centre stage in cancer profiling. *Nature* **459**: 146–147.

Mardis ER (2006) Anticipating the 1,000 dollar genome. *Genome Biology* **7**: 112.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* **9**: 387–402.

Martinez-Alcantara A, Ballesteros E, Feng C *et al*. (2009) PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics* **25**: 2438–2349.

McCarroll SA, Huett A, Kuballa P *et al*. (2008b) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature Genetics* **40**: 1107–1112.

McCarroll SA, Kuruvilla FG, Korn JM *et al*. (2008a) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166–1174.

McKernan KJ, Peckham HE, Costa GL *et al*. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* **19**: 1527–1241.

Morozova O and Marra MA (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**: 255–264.

Nejentsev S, Walker N, Riches D *et al*. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**: 387–389.

Pennisi E (2007) Working the (gene count) numbers: finally, a firm answer? *Science* **316**: 1113.

Prickett TD, Agrawal NS, Wei X *et al*. (2009) Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4. *Nature Genetics* **41**: 1127–1132.

Pushkarev D, Neff NF and Quake SR (2009) Single-molecule sequencing of an individual human genome. *Nature Biotechnology* **27**: 847–852.

Rothberg JM and Leamon JH (2008) The development and impact of 454 sequencing. *Nature Biotechnology* **26**: 1117–1124.

Scherer SW, Lee C, Birney E *et al*. (2007) Challenges and standards in integrating surveys of structural variation. *Nature Genetics* **39**: S7–S15.

Sebat J, Lakshmi B, Troge J *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–528.

Shendure J and Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* **26**: 1135–1145.

Summerer D (2009) Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. *Genomics* **94**: 363–368.

Venter JC, Adams MD, Myers EW *et al*. (2001) The sequence of the human genome. *Science* **291**: 1304–1351.

Wain LV, Armour JA and Tobin MD (2009) Genomic copy number variation, human health, and disease. *Lancet* **374**: 340–350.

Wang J, Wang W, Li R *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Wheeler DA, Srinivasan M, Egholm M *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872–876.

Wood LD, Parsons DW, Jones S *et al*. (2007) The genomic landscapes of human breast and colorectal cancers. *Science* **318**: 1108–1113.

Yoon S, Xuan Z, Makarov V *et al*. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research* **19**: 1586–1592.

Further Reading

Frazer KA, Murray SS, Schork NJ *et al*. (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews. Genetics* **10**: 241–251.

Mardis ER and Wilson RK (2009) Cancer genome sequencing: a review. *Human Molecular Genetics* **18**: R163–R168.

Tucker T, Marra M and Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. *American Journal of Human Genetics* **85**: 142–154.

High-Throughput Single Nucleotide Polymorphisms Genotyping Technologies

KU Chee Seng, *Centre for Molecular Epidemiology, Department of Epidemiology and Public Health (MD3), Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

KASIMAN Katherine, *Centre for Molecular Epidemiology, Department of Epidemiology and Public Health (MD3), Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

CHIA Kee Seng, *Centre for Molecular Epidemiology, Department of Epidemiology and Public Health (MD3), Yong Loo Lin School of Medicine, National University of Singapore, Singapore*

Advanced article

Article Contents

Online posting date: 15th September 2009

Genome-wide association studies have successfully identified many novel genetic loci for various human complex diseases and quantitative traits. There are several important factors contributing to the feasibility of this approach; one of them is the rapid advancement in high-throughput single nucleotide polymorphisms (SNPs) genotyping technologies which has enabled researchers to comprehensively interrogate the entire human genome. Almost all the studies that have been published up to date used commercially available whole-genome genotyping arrays from Illumina® and Affymetrix®. The most prominent feature of these high-throughput genotyping platforms is the ability to interrogate several hundred thousands to one million SNPs simultaneously in a microarray. The application of genotyping arrays is not only limited to association studies, but it has also been applied to many other human genetic studies. However, the rapid developments of sequencing technologies have started replacing the microarray experiments for both structural and functional genomics studies.

Introduction

Over the past three years, we have been seeing the success of *genome wide association studies* (*GWAS*) in identifying an enormous number of novel genetic loci and implicating new biological pathways for various human complex diseases and quantitative traits (Ku and Chia, 2008). Many genes which were identified by GWAS are not candidate genes previously thought to be associated with diseases such as the two novel genes and biological pathways linked to Crohn disease: *IL23R* (interleukin-23 receptor) and *ATG16L1* (autophagy) pathways (Mohlke *et al*., 2008; Easton and Eeles, 2008; Lettre and Rioux, 2008). The paradigm shift in genetic approach – from candidate–gene association and family linkage studies to GWAS – has been attributed to several important developments, notably the rapid advancement in high-throughput *single nucleotide polymorphisms* (*SNPs*) *genotyping* technologies which has enabled researchers to interrogate several hundred thousands to one million SNPs simultaneously in a microarray. GWAS is a comprehensive and agnostic approach in the search for unknown disease variants, as such; the ability to interrogate large number of SNPs covering the entire human genome is a prerequisite to this study design. In parallel with decreasing cost of genotyping, it is currently practical to genotype thousands of samples in GWAS. Both the technological feasibility and affordable genotyping cost have been one of the primary driving forces for the rapid publications of GWAS. **See also**: Genome-Wide Association Studies

ELS subject area: Genetics and Disease

How to cite:
Chee Seng, KU; Katherine, KASIMAN; and, Kee Seng, CHIA (September 2009) High-Throughput Single Nucleotide Polymorphisms Genotyping Technologies. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0021631

Overview of Genotyping Platforms

To date, more than 200 GWAS have been published since 2007 (http://www.genome.gov/26525384), and almost all

the studies had used commercially available *whole-genome SNPs genotyping arrays* from Illumina® (San Diego, California, USA) and Affymetrix® (Santa Clara, California, USA). Currently, Illumina® and Affymetrix® are the only two companies in the market who design and provide whole-genome genotyping arrays for human genetic studies. Two common features of these high-throughput genotyping platforms are the ability to interrogate several hundred thousands to one million SNPs simultaneously in a microarray, and the allelic discrimination or scoring is based on fluorescent signal intensity. On the contrary, several aspects are distinct between the two genotyping platforms, especially the chemistry of genotyping assays, experiment protocol and principle of allelic scoring.

Other low-throughput genotyping platforms (i.e. the number of SNPs that can be genotyped in one experiment per sample is less than several hundreds) are Sequenom® MassARRAY iPLEX, Invader® assay, Applied Biosystems® SNPlex genotyping system and TaqMan genotyping assay. The principle of allelic scoring for MassARRAY iPLEX is based on the mass of allele-specific-extended product which is generated by each of the two alleles for one particular SNP. This extended product is then separated by MALDI-TOF-MS (matrix-assisted laser desorption/ionization, time-of-flight mass spectrometry). This is totally different from most of the genotyping platforms or assays which are based on fluorescent signal intensity measurement. Low-throughput genotyping platforms are commonly applied for replication studies of several ten-to-hundred SNPs identified from initial genome-wide scanning in GWAS, fine mapping of linkage regions or candidate gene association studies.

SNPs Selection Approaches

There are two commonly adopted approaches in SNPs selection for whole-genome genotyping arrays: direct and indirect. Direct approach focuses on selecting SNPs that are of putative functional importance, for example, SNPs in coding, promoter regions and splicing sites. These SNPs are predicted to alter the structure or function of proteins, gene expression (transcription) or pre-messenger ribonucleic acid (mRNA) splicing process. This selection approach is applied in gene-centric association studies which directly interrogate the putative functional SNPs as opposed to GWAS, which is based on *linkage disequilibrium* (*LD*) to indirectly locate the disease variants. The logic for this approach, which assumed that SNPs within or near the genes have higher prior probability being the disease variants, is clear (Jorgenson and White, 2006). **See also**: Single Nucleotide Polymorphism (SNP)

The indirect approach selects SNPs despite their functionality. The approach can be further divided into LD-based and random-based approaches. The LD-based approach, which is employed by Illumina®, selects SNPs based on the LD pattern or information and the selected SNPs to be genotyped on the array are called the *tagging SNPs*. The LD-based approach relies on a metric called *correlation coefficient* (r^2) (Carlson *et al*., 2004). Genotyping for the SNPs which are in strong LD ($r^2 > 0.8$) is redundant in terms of obtaining more information. LD-based is a more efficient approach (in contrast to random-based approach) in SNPs selection because fewer (tagging) SNPs need to be selected to sufficiently provide information for other untyped SNPs in regions with strong LD. The LD-based approach is feasible and practical on the completion of International HapMap Project which characterized LD patterns in the human genome (International HapMap Consortium, 2005, 2007). This global effort revealed that most regions in the human genome are in strong LD, and that recombination events which decay the correlation between SNPs are not a random process but rather clustered in certain hotspots, that is, the recombination hotspots. **See also**: HapMap Project

The random-based approach employed by Affymetrix® selects SNPs randomly distributed throughout the human genome without taking LD patterns into account. As a result, for equal number of SNPs being genotyped, the LD-based approach tends to provide higher genome coverage than the random-based approach because the tagging SNPs provide additional information for other SNPs. From the aspect of genome coverage, random-based approach is less efficient because there are redundancies of SNPs being genotyped in regions with strong LD since the SNPs are randomly and evenly spaced throughout the genome. However, there will be inadequacies of SNPs covering weak LD regions or recombination hotspots. The LD-based approach also encounters the latter problem because SNPs in the weak LD regions are not 'informative' (i.e. in providing information for other SNPs) from the perspective of this approach. As such, to optimize the efficiency of LD-based approach, these SNPs are not prioritized during SNPs selection. In addition to being more efficient and cost-effective, the LD-based approach also alleviates the statistical problem of multiple-hypothesis testing in GWAS as lesser number of SNPs is genotyped.

Chemistry of Genotyping Assays and Principle for Allelic Scoring

Whole-genome sampling assay

The Affymetrix® genotyping assay is known as the whole-genome sampling assay (WGSA) which involves genome complexity reduction step, that is, restriction enzyme digestion and polymerase chain reaction (PCR) amplification. A total genomic deoxyribonucleic acid (DNA) of 250 ng is digested by a restriction enzyme for each genotyping assay. Different restriction enzymes are used for different whole-genome genotyping genechips, for example, GeneChip Human Mapping 100 K Set uses *XbaI* and *HindIII*, whereas GeneChip Human Mapping 500K

Set, Affymetrix® SNP Array 5.0 and 6.0 use *NspI* and *StyI*. All the digested DNA fragments are ligated to an adaptor. Subsequently, these ligated DNA fragments are amplified by a universal primer in PCR, and only DNA fragments within a certain size range are selected for amplification. The PCR amplification step is important in obtaining sufficient amount of DNA fragments (targets) for hybridization to the probes immobilized on genechip to produce signal intensity later. The PCR step is then followed by purification, fragmentation, labelling, denaturation and hybridization. Finally, washing and staining steps eliminate any weak or unspecific bindings, and amplify signal intensities respectively.

The genome complexity reduction step is critical for Affymetrix® genotyping assay because the allelic discrimination is based on the allele-specific hybridization principle. For Human Mapping 100K and 500K, there are perfect match (PM) and mismatch (MM) oligonucleotide probes for each of the SNPs capturing two possible alleles denoted as A allele and B allele. For one particular SNP with homozygote AA genotype, DNA fragment where the SNP is located will not hybridize to PM probes for B allele. As a result, the signal intensity of PM probes for A allele is maximal, whereas the signal intensity of PM probes for B allele is minimal. Based on this complementary and hybridization between targets and probes, the allelic discrimination or genotype is determined by comparing signal intensities between the pair of PM probes after subtracting background noise. For heterozygote genotype, signal intensities of PM probes for both the A and B alleles are expected to be equal. The function of MM probes is to measure background noise (Kennedy *et al*., 2003; Matsuzaki *et al*., 2004a, 2004b).

The genome complexity reduction approach selectively amplifies a portion of the human genome for hybridization. Hence, only SNPs located in the PCR-amplified DNA fragments, that is, only SNPs in a subset of genome can be selected for genotyping. This constraint precludes the flexibility of selecting any SNPs in the human genome for genotyping. As a result, the LD-based approach is not suitable for this genotyping assay. Reduction of the genome complexity is essential to minimize noise level. This is because the allelic discrimination is based on the allele-specific hybridization principle. Genome complexity reduction is crucial in reducing cross hybridizations between targets and probes because the differentiation of two alleles takes place during hybridization step.

Infinium assay

The two genotyping assays for Illumina® whole-genome genotyping beadchips are Infinium I and Infinium II assays. Infinium assays involved several simple and straightforward steps in the protocol. Total genomic DNA is first amplified through a whole-genome amplification step to harness sufficient amount of DNA for hybridization to the probes on beadchip. This step substantially increases the amount of DNA by more than 1000-folds. This amplification step is completely different from Affymetrix® genotyping assay which utilizes PCR to amplify only a subset of genome. The amplified DNA is fragmented and then precipitated to remove any impurities. The DNA pellet is then dissolved in a resuspension buffer, and denatured before overnight hybridization on the beadchip which is then followed by washing and staining in the next day.

Infinium I assay employs two bead types and one-colour chemistry to assay one SNP; each bead-type corresponds to one allele. The allelic discrimination or scoring is based on the allele-specific primer extension (ASPE) principle. Two allele-specific oligonucleotide probes are designed to assay two of the alleles for one particular SNP. The two probes are identical except at their terminal 3′ nucleotide (i.e. the most extreme nucleotide at the 3′ end) which is designed to perfectly match to one allele or the other. For a hypothetical example where the alleles are denoted as A allele and B allele, for homozygote AA genotype, there will be perfect complementarity between target and A allele-specific probe, and one base mismatch with B allele-specific probe. Perfect complementarity will allow primer extension of several nucleotides by polymerase enzyme, and followed by signal amplification for A allele-specific probe during the washing and staining steps, but minimal signal intensity for B allele-specific probe. The genotype calling is based on this signal intensities information and is done by Illumina® software – BeadStudio.

The Infinium II assay, however, employs one bead type and two-colour chemistry. The dideoxy (dd) adenosine triphosphate (ATP) and dideoxy thymidine triphosphate (ddTTP) are labelled by one colour whereas dideoxy guanosine triphosphate (ddGTP) and dideoxy cytosine triphosphate (ddCTP) are labelled by another colour. As such, this assay is unable to differentiate between A to T nucleotide substitution and G/C SNP. Infinium II assay is based on the principle of single base extension (SBE) for allelic scoring. No allele-specific probes are designed for this assay; instead, one locus-specific probe is designed for each SNP. This locus-specific probe binds to the target before the SNP position, that is, immediately adjacent to the SNP locus. Addition of one dideoxy nucleoside triphosphate (ddNTP) will terminate the extension reaction, that is, SBE principle. For Infinium assays, the allelic discrimination or scoring does not occur during hybridization, but takes place posthybridization in an enzymatic-based extension step (Gunderson *et al*., 2005; Steemers *et al*., 2006; Gunderson *et al*., 2006; Steemers and Gunderson, 2007). This is different from the allele-specific hybridization principle employed by Affymetrix® genotyping assay. The Infinium HD assay for the latest beadchips – Human 1M-Duo and Human660W-Quad – is able to assay all types of SNPs.

The two-step involvement in allelic scoring minimizes the noise and increases accuracy of genotype calling. The first step is accomplished by hybridization between target and probe (selectivity) whereas the second step is accomplished by enzymatic-based primer extension (specificity) in generating signal intensities for both the ASPE and SBE.

Generally, Infinium assays employ whole-genome amplification step, which ultimately hybridize the 'entire human genome' on beadchips. As such, these assays enable unconstrained SNPs selection and are suitable for designing a variety of arrays for different applications, from whole-genome genotyping arrays and focused-content arrays such as HumanCVD Genotyping BeadChip to custom-designed array, that is, iSelect Custom Genotyping BeadChip. The SNPs selection is only restricted by probes representation on beadchip. Theoretically, any SNPs can be selected as long as there are probes designed to capture them.

Whole-Genome SNPs Genotyping Arrays

Many products have been introduced by Affymetrix® and Illumina® over the past few years, but only several commonly used genotyping arrays in GWAS have been selected for discussion. A considerable emphasis is placed on these aspects – genome coverage and marker density of genotyping arrays – as genome coverage is a critical factor in contributing to a successful GWAS, and marker density is important for comprehensive detection and discovery of *copy number variations* (*CNVs*). The number of SNPs/markers and genome coverage for each genotyping array are summarized in **Table 1** and **Figure 1**.

Affymetrix

GeneChip Human Mapping 100 K Set was the first whole-genome genotyping product introduced to the market by Affymetrix®. This set comprised of two arrays and each could interrogate greater than 50 000 SNPs. This genechip was used in the first GWAS published in 2005 and uncovered the association between an SNP in complement factor H gene and age-related macular degeneration (Klein *et al*., 2005).

Since the number of SNPs which can be selected by Affymetrix® WGSA relies on restriction enzyme cutting and the size range of DNA fragments amplified by PCR, to include additional SNPs to genotype, different restriction enzymes and PCR conditions are needed. The second product launched was GeneChip Human Mapping 500K Set. This set also comprised of two arrays and used two different restriction enzymes compared to the earlier product. In comparison to other arrays with a comparable number of SNPs, for example, HumanHap550 or even lesser number of SNPs such as HumanHap300, Human Mapping 500K achieved considerably lower genome coverage for HapMap CEU and CHB + JPT (**Table 1**) (Barrett and Cardon, 2006; Li *et al*., 2008a, 2008b). This demonstrated the power and efficiency of LD-based approach over

Table 1 Number of SNPs and genome coverage of whole-genome genotyping arrays

Genotyping array	Number of SNPs/ markers	Genome coverage (%) ($r^2 > 0.8$)		
Illumina		CEU	CHB + JPT	YRI
Human-1 BeadChip	> 109 000 SNPs	26[a]	28[a]	12[a]
HumanHap300 BeadChip	> 317 000 SNPs	75[a]	63[a]	28[a]
		77[b]	66[b]	29[b]
HumanHap550 BeadChip	> 550 000 SNPs	87[b]	83[b]	50[b]
HumanHap650Y BeadChip	> 650 000 SNPs	87[b]	84[b]	60[b]
Human 1M Beadchip	1 072 820 markers (< 5% are copy-number probes) Median spacing between markers: 1.7 kb	93[b]	92[b]	68[b]
Affymetrix				
Human Mapping 100 K Set	> 110 000 SNPs	31[a]	31[a]	15[a]
Human Mapping 500 K Set	> 500 000 SNPs	65[a]	66[a]	41[a]
SNP Array 5.0	500 568 SNPs 420 000 copy-number probes	64[b]	66[b]	41[b]
SNP Array 6.0	906 000 SNPs 946 000 copy-number probes Median spacing between markers: < 700 bp	83[b]	84[b]	62[b]

[a]Barrett and Cardon (2006).
[b]Li *et al*. (2008b).

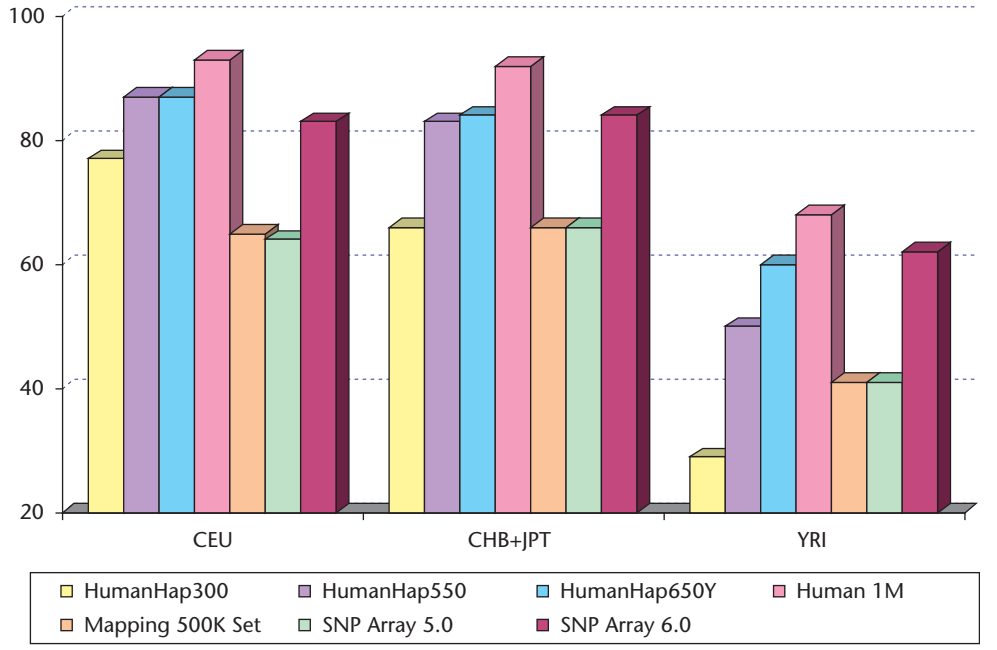


Figure 1 Comparisons of genome coverage across several whole-genome genotyping arrays from Illumina® and Affymetrix® in International HapMap populations. The *y*-axis is the genome coverage in percentage. (see Li *et al.*, (2008b); Barrett and Cardon, (2006)).

random-based approach in SNPs selection in optimizing for coverage. Simultaneous interrogation of more than 500 000 SNPs enabled the Wellcome Trust Case Control Consortium (WTCCC) to conduct a high-powered and well-designed GWAS. Seventeen thousand samples (14 000 cases and 3000 controls) were genotyped using Human Mapping 500K Set which led to the identification of many novel genetic loci for the seven complex diseases (Wellcome Trust Case Control Consortium, 2007).

Illumina

The first product introduced by Illumina® was Human-1 Genotyping BeadChip. SNPs located directly within coding, promoter and highly conserved regions were selected. The evolutionary-conserved sequences were predicted to be functionally important. The SNPs selection for this product was mainly driven by direct approach as greater than 70% of SNPs are located in *transcripts* or within 10 kb of the exons.

On completion of the International HapMap Project and the release of data, Illumina® launched their second line of whole-genome genotyping products – HumanHap300, HumanHap550 and HumanHap650Y BeadChips. Selection of tagging SNPs for HumanHap300 was based on Phase I HapMap data, whereas the latter two products were developed using combined Phase I and II HapMap data. HumanHap300 was designed mainly to capture common SNPs in Caucasians, and in such, performed excellently well when it was evaluated in HapMap CEU population; HapMap CHB + JPT population only achieved moderate coverage.

To further increase genome coverage for Asian and African populations, more tagging SNPs were selected and the resultant product – HumanHap550 – achieved considerably good coverage for HapMap CEU and CHB + JPT. The genome coverage for HapMap YRI was much lower; this is due to a greater genetic diversity and weaker LD in African populations. HumanHap650Y is an extension of the content in HumanHap550 with an additional approximately 100 000 tagging SNPs specifically chosen to improve genome coverage for Africans. Sixty percent was achieved for HapMap YRI, whereas no increment in genome coverage was observed for other HapMap populations as expected (Eberle *et al.*, 2007). The genome coverage was computed by setting a threshold of $r^2 > 0.8$ and measured using International HapMap data as reference. Theoretically, setting a less stringent r^2 threshold will increase genome coverage, but this will also decrease the statistical power to detect disease variants through LD. Sample size is inversely proportional to r^2 value.

As far as the same method in computing genome coverage and SNPs selection approach are concerned for Illumina® beadchips, further increase in the number of tagging SNPs to genotype in GWAS will cause 'diminishing return' effect especially for Caucasian and Asian populations. This was well-demonstrated in Human 1M Beadchip, which doubled the number of tagging SNPs to one million, adding a slight increment to genome coverage (**Table 1**). However, it is unclear how much improvement it has in terms of capturing other common SNPs which are not validated by the International HapMap Project but are genuine and deposited in the database of SNPs (dbSNP).

Latest genotyping arrays

The latest whole-genome genotyping arrays are Illumina® Human 1M BeadChip and Affymetrix® SNP Array 6.0 which enable genotyping of up to one million SNPs. Unlike the earlier version of beadchips, Illumina® Human 1M is not focused on tagging SNPs selection alone from International HapMap Project, but other SNPs from dbSNPs have also been selected to further increase genome coverage and marker uniformity across the genome. Approximately 950 000 are HapMap tagging SNPs and 100 000 are non-HapMap SNPs. This new product used combined direct and indirect LD-based SNPs selection approaches which could be more powerful for disease variants discovery in GWAS.

The latest genotyping genechip from Affymetrix® contains more than 1.8 million markers; half of the content is SNPs and the remaining is nonpolymorphic or copy-number probes to enhance power for detection of CNVs (Shen *et al*., 2008). On the contrary, only a small portion of the content in Human 1M Beadchip is copy-number probes. The SNPs for Array 6.0 is built on the content from Human Mapping 500K, with additional SNPs chosen from the International HapMap Project (tagging SNPs), with more SNPs on sex chromosomes and mitochondrial SNPs.

Copy-number probes were deliberately selected to cover regions lacking of SNPs or regions where SNPs are difficult to assay such as the repetitive sequences in segmental duplications. In addition, markers were also chosen to target known CNV regions reported in the Database of Genomic Variants. With such design, these genotyping arrays enabled researchers to discover novel CNVs as well as to validate known CNVs which were previously identified. These latest arrays are designed for both applications: SNPs GWAS and CNVs detection.

Genome Coverage

GWAS are indirect association studies relying on tagging SNPs that are genotyped to detect disease variants which are not tested directly. Fine mapping is needed to capture the disease or functional variants in regions revealed by GWAS. Since indirect association studies are reliant on LD to find disease variant, genome coverage of commercially available genotyping arrays is critical and is a key factor for the success of GWAS.

The statistical power of genetic association studies is basically a function of sample size, magnitude of genetic effect size and allele frequency. As the latter two factors are unknown until the genetic variants are uncovered, sample size is the only controllable factor determining the statistical power. In addition, power also depends on genome coverage. Ideally, increasing both sample size and genome coverage will increase the statistical power of a study. High genome coverage is important because the underlying principle of this approach is based on LD in the detection of disease variants. In SNPs-scarce regions, genuine disease variants could be missed because they are not in strong LD with any of the SNPs that are genotyped on the array. Whole-genome genotyping arrays like Human 1M and SNP Array 6.0 offer almost complete genome coverage for HapMap CEU and CHB + JPT.

Genome coverage is an estimate of the proportion of SNPs (using the International HapMap data as reference) that can be captured by the SNPs which directly genotyped in an array with a preset r^2 threshold. Usually a threshold of 0.8 is used to estimate genome coverage. This is also known as global genome coverage. This estimate differs from local coverage and gene coverage of genotyping arrays. The latter two estimates provide more information of coverage at a finer scale (Li *et al*., 2008a, 2008b).

Most studies estimated genome coverage using International HapMap populations; therefore, whether the same coverage is achieved in other populations is unclear especially for Illumina® beadchips which selected SNPs based on tagging SNPs approach. However, there have been a number of studies which demonstrated that HapMap tagging SNPs are broadly transferable in many other populations which are not part of the HapMap Project (Lundmark *et al*., 2008; Xing *et al*., 2008). Based on these findings, a comparable coverage is likely to be achieved in other populations. In fact, these genotyping arrays have been shown to perform equally well in other populations (Magi *et al*., 2007). Genome coverage may vary from one population to another, but would not be in a large magnitude.

Since the SNPs selection for these genotyping arrays especially the Illumina beadchips was based on the International HapMap data, genome coverage could be overestimated when the HapMap data was used as the reference as well. This is known as data 'over-fitting'. In fact, it was well-demonstrated by using a set of resequencing data as the reference; the genome coverage of common SNPs by both Illumina® and Affymetrix® arrays were appreciably lower. Coverage was about 17% lower in both Human 1M and SNP Array 6.0 when compared to the HapMap-based estimates for CEU population (Bhangale *et al*., 2008). These differences clearly highlight the bias introduced when HapMap data was used as the reference to compute genome coverage for commercial genotyping arrays. The same conclusions were also derived from other studies using independent set of SNPs data and samples in the evaluation of genome coverage, accounting for both data and sample 'over-fitting' (Hao *et al*., 2008).

Other Applications

Application of whole-genome genotyping arrays is not only limited to GWAS for complex diseases and traits, but it has also been applied for many other human genetic studies such as population genetics and population structure analysis (Li *et al*., 2008a, 2008b; Jakobsson *et al*., 2008), identification of eQTL (expression quantitative trait loci) and pQTL (protein

quantitative trait loci) (Schadt *et al*., 2008; Melzer *et al*., 2008) and other studies (Cooper *et al*., 2008a; Kong *et al*., 2008). Most importantly, Illumina® and Affymetrix® were the main genotyping platforms used to complete the International HapMap Project (International HapMap Consortium, 2005, 2007).

Whole-genome genotyping arrays have also been increasingly used for the detection and analysis of CNVs in population-based studies. Both Human Mapping 500K and SNP Array 6.0 were used for CNVs characterization on International HapMap samples (Redon *et al*., 2006; McCarroll *et al*., 2008). The significant increase in resolution and marker uniformity on the Illumina® and Affymetrix® latest arrays – minimizing large gaps devoid of markers – is critical in ensuring a comprehensive and less-biased CNVs discovery throughout the genome. Even with these notable improvements, a considerably large fraction of CNVs was still missed by these arrays when compared to a sequencing method (Cooper *et al*., 2008b). The application of genotyping arrays is beyond CNVs detection and discovery studies; it has also extended to disease association studies using CNVs as the markers for identifying novel genetic loci (Weiss *et al*., 2008; Stefansson *et al*., 2008). **See also**: Copy Number Variation in the Human Genome; Relevance of Copy Number Variation to Human Genetic Disease

Challenges from Whole-Genome Sequencing Technologies

Expeditious developments of sequencing technologies have started threatening the market of microarrays. Currently, the three next-generation sequencing platforms are Illumina® Genome Analyzer, Roche® 454 GS-FLX Sequencer and ABI® SOLiD Sequencer (Mardis, 2008). Sequencing approach has been quickly adopted for various applications in structural and functional genomic studies, for example, ChIP-Seq (the combined chromatin immunoprecipitation technique and sequencing method) in the identification of transcription factor binding sites in a genome-wide scale (Johnson *et al*., 2007). This approach had started replacing the preceding method, ChIP-chip, which was based on microarray hybridization. Moreover, the applications have also been extended to transcriptome profiling (RNA-Seq) substituting the conventional gene-expression microarray which dominated the field over the past decade (Sultan *et al*., 2008). Paired-end sequencing method has been increasingly applied to characterize structural variations (Korbel *et al*., 2007; Campbell *et al*., 2008) which were previously studied using comparative genomic hybridization (CGH) arrays. The strengths and advantages of these innovative sequencing methods over previous microarray approaches were clearly demonstrated in these studies. Towards the end of 2008, we have also seen the completion of sequencing two diploid genomes (Wang *et al*., 2008; Bentley *et al*., 2008).

So, is there a future for SNPs genotyping arrays in the era of next-generation sequencing technologies? Currently, whole-genome sequencing has not been applied in association studies of complex diseases which require a minimum of thousands of samples. There are several hurdles to this: cost is still prohibitively expensive for whole-genome sequencing to be applied in large sample size, even though it is rapidly decreasing. In addition, there are statistical and computational challenges in analyzing tremendous amount of data and this is undoubtedly several orders of magnitude beyond what was described in the *Nature* article 'drinking from the fire hose' by Hunter and Kraft (2008). Finally, technical problems and issues of sample throughput and data quality for studying thousands of genomes still exist. Nonetheless, with the current momentum of progress, it is foreseeable that within the next few years whole-genome sequencing will start to play a key role in association studies. Sequencing method would not only be restricted to common SNPs, but uncommon SNPs, rare mutations, structural variations and other genetic variations can also be studied simultaneously. Thus this would provide a comprehensive picture for the genetic landscape of complex diseases in comparison to genotyping which can only produce a 'snapshot' picture. The competitive force from sequencing technologies was well-addressed in the 16 October issue of *Nature News* – The death of microarrays? – in 2008. The 1000 Genomes Project will further drive the rapid technological developments of sequencing and their applications in genomics research.

Issues and Factors for Choosing a Genotyping Platform

Selecting genotyping array for GWAS is one of the critical factors determining the success of GWAS. Many factors needed to be taken into account. At the stage of study design, factors such as genome coverage, marker density and robustness of genotyping assay, cost and the application are important. For instance, should the major aim of GWAS is for SNPs association analysis in a Caucasian or Asian population, HumanHap500 may be the wise choice as it had already provided fairly good genome coverage in these populations. However, if the application is for CNVs detection on top of the SNPs analysis, perhaps Human 1M or SNP Array 6.0 is a better option as higher marker density enhanced the power for CNVs studies. Robustness of genotyping assay, sample throughput and ease of experiment protocol are other factors which need to be considered. Sample throughput is an important element in GWAS which genotyped thousands of samples, for example, the multisample format beadchips increase the throughput by several folds. Experiment protocols that involve only several simple and straightforward steps are important in minimizing sample contamination and ensuring a high sample genotyping success rate. Whole-genome genotyping arrays have been playing an important

part in GWAS for discoveries of novel genetic loci. It will continue to make contributions in this field until we reach the $1000 whole-genome sequencing era and the associated hurdles are conquered.

References

Barrett JC and Cardon LR (2006) Evaluating coverage of genome-wide association studies. *Nature Genetics* **38**: 659–662.

Bentley DR, Balasubramanian S, Swerdlow HP *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.

Bhangale TR, Rieder MJ and Nickerson DA (2008) Estimating coverage and power for genetic association studies using near-complete variation data. *Nature Genetics* **40**: 841–843.

Campbell PJ, Stephens PJ, Pleasance ED *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics* **40**: 722–729.

Carlson CS, Eberle MA, Rieder MJ *et al.* (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics* **74**: 106–120.

Cooper GM, Johnson JA, Langaee TY *et al.* (2008a) A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood* **112**: 1022–1027.

Cooper GM, Zerr T, Kidd JM *et al.* (2008b) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genetics* **40**: 1199–1203.

Easton DF and Eeles RA (2008) Genome-wide association studies in cancer. *Human Molecular Genetics* **17**: R109–R1115.

Eberle MA, Ng PC, Kuhn K *et al.* (2007) Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genetics* **3**: 1827–1837.

Gunderson KL, Kuhn KM, Steemers FJ *et al.* (2006) Whole-genome genotyping of haplotype tag single nucleotide polymorphisms. *Pharmacogenomics* **7**: 641–648.

Gunderson KL, Steemers FJ, Lee G *et al.* (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics* **37**: 549–554.

Hao K, Schadt EE and Storey JD (2008) Calibrating the performance of SNP arrays for whole-genome association studies. *PLoS Genetics* **4**: e1000109.

Hunter DJ and Kraft P (2008) Drinking from the fire hose – statistical issues in genomewide association studies. *New England Journal of Medicine* **357**: 436–439.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**: 1299–1320.

International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

Jakobsson M, Scholz SW, Scheet P *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998–1003.

Johnson DS, Mortazavi A, Myers RM *et al.* (2007) Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316**: 1497–1502.

Jorgenson E and White JS (2006) A gene-centric approach to genome-wide association studies. *Nature Reviews Genetics* **7**: 885–891.

Kennedy GC, Matsuzaki H, Dong S *et al.* (2003) Large-scale genotyping of complex DNA. *Nature Biotechnology* **21**: 1233–1237.

Klein RJ, Zeiss C, Chew EY *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385–389.

Kong A, Thorleifsson G, Stefansson H *et al.* (2008) Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science* **319**: 1398–1401.

Korbel JO, Urban AE, Affourtit JP *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420–426.

Ku CS and Chia KS (2008) The success of the genome-wide association approach: a brief story of a long struggle. *European Journal of Human Genetics* **16**: 554–564.

Lettre G and Rioux JD (2008) Autoimmune diseases: insights from genome-wide association studies. *Human Molecular Genetics* **17**: R116–R121.

Li JZ, Absher DM, Tang H *et al.* (2008a) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100–1104.

Li M, Li C and Guan W (2008b) Evaluation of coverage variation of SNP chips for genome-wide association studies. *European Journal of Human Genetics* **16**: 635–643.

Lundmark PE, Liljedahl U, Boomsma DI *et al.* (2008) Evaluation of HapMap data in six populations of European descent. *European Journal of Human Genetics* **16**: 1142–1150.

Magi R, Pfeufer A, Nelis M *et al.* (2007) Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation. *BMC Genomics* **8**: 159.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics* **24**: 133–141.

Matsuzaki H, Dong S, Loi H *et al.* (2004a) Genotyping over 100 000 SNPs on a pair of oligonucleotide arrays. *Nature Methods* **1**: 109–111.

Matsuzaki H, Loi H, Dong S *et al.* (2004b) Parallel genotyping of over 10 000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Research* **14**: 414–425.

McCarroll SA, Kuruvilla FG, Korn JM *et al.* (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics* **40**: 1166–1174.

Melzer D, Perry JR, Hernandez D *et al.* (2008) A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genetics* **4**: e1000072.

Mohlke KL, Boehnke M and Abecasis GR (2008) Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Human Molecular Genetics* **17**: R102–R108.

Redon R, Ishikawa S, Fitch KR *et al.* (2006) Global variation in copy number in the human genome. *Nature* **444**: 444–454.

Schadt EE, Molony C, Chudin E *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biology* **6**: e107.

Shen F, Huang J, Fitch KR *et al.* (2008) Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. *BMC Genetics* **9**: 27.

Steemers FJ, Chang W, Lee G *et al*. (2006) Whole-genome genotyping with the single-base extension assay. *Nature Methods* **3**: 31–33.

Steemers FJ and Gunderson KL (2007) Whole genome genotyping technologies on the BeadArray platform. *Biotechnology Journal* **2**: 41–49.

Stefansson H, Rujescu D, Cichon S *et al*. (2008) Large recurrent microdeletions associated with schizophrenia. *Nature* **455**: 232–236.

Sultan M, Schulz MH, Richard H *et al*. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**: 956–960.

Wang J, Wang W, Li R *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

Weiss LA, Shen Y, Korn JM *et al*. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *New England Journal of Medicine* **358**: 667–675.

Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* **447**: 661–678.

Xing J, Witherspoon DJ, Watkins WS *et al*. (2008) HapMap tagSNP transferability in multiple populations: general guidelines. *Genomics* **92**: 41–51.

Further Reading

De la Vega FM, Lazaruk KD, Rhodes MD *et al*. (2005) Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP genotyping assays and the SNPlex genotyping system. *Mutation Research* **573**: 111–135.

Fan JB, Chee MS and Gunderson KL (2006) Highly parallel genomic assays. *Nature Reviews Genetics* **7**: 632–644.

Olivier M (2005) The invader assay for SNP genotyping. *Mutation Research* **573**: 103–110.

Ragoussis J, Elvidge GP, Kaur K *et al*. (2006) Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genetics* **2**: e100.

Syvänen AC (2005) Toward genome-wide SNP genotyping. *Nature Genetics* **37**: S5–S10.

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

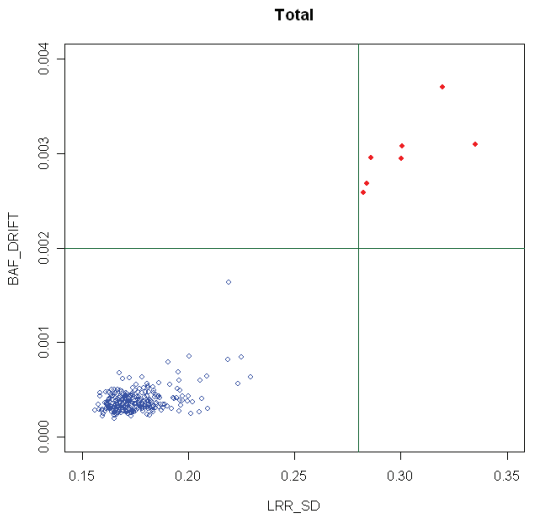
Samples

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.

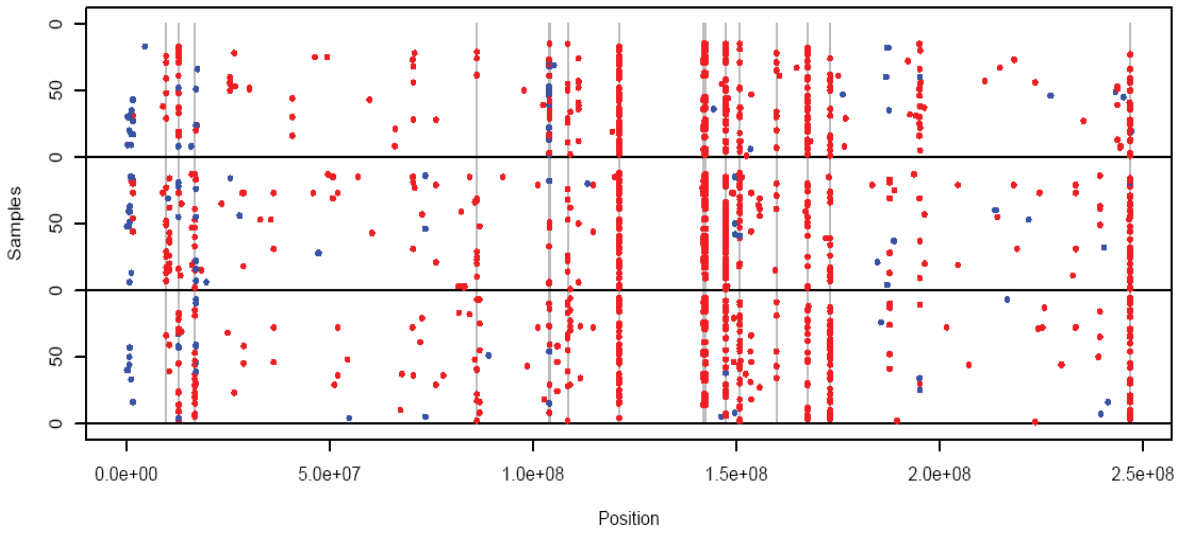
The goal of this study was to interrogate the *de novo* occurrence of CNV events compared to inherited germ line CNVs. CNVs detected in offspring but absent in their parents, or differences in the CNVs between a pair of monozygotic twins, are indications of putative *de novo* CNVs. Due to the noise inherent 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

illumina®

APPENDICES
(Additional relevant publications)

Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations

Yik-Ying Teo, Xueling Sim, Rick T.H. Ong, et al.

Genome Res. 2009 19: 2154-2162 originally published online August 21, 2009
Access the most recent version at doi:10.1101/gr.095000.109

Resource

Singapore Genome Variation Project: A haplotype map of three Southeast Asian populations

Yik-Ying Teo,[1,2,3,7] Xueling Sim,[1,7] Rick T.H. Ong,[1,4,7] Adrian K.S. Tan,[4] Jieming Chen,[4] Erwin Tantoso,[4] Kerrin S. Small,[3] Chee-Seng Ku,[1] Edmund J.D. Lee,[5] Mark Seielstad,[4,8] and Kee-Seng Chia[1,6,8,9]

[1]*Centre for Molecular Epidemiology, National University of Singapore, Singapore 117597;* [2]*Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546;* [3]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, United Kingdom;* [4]*Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore 138672;* [5]*Department of Pharmacology, National University of Singapore, Singapore 117597;* [6]*Department of Epidemiology and Public Health, National University of Singapore, Singapore 117597*

The Singapore Genome Variation Project (SGVP) provides a publicly available resource of 1.6 million single nucleotide polymorphisms (SNPs) genotyped in 268 individuals from the Chinese, Malay, and Indian population groups in Southeast Asia. This online database catalogs information and summaries on genotype and phased haplotype data, including allele frequencies, assessment of linkage disequilibrium (LD), and recombination rates in a format similar to the International HapMap Project. Here, we introduce this resource and describe the analysis of human genomic variation upon agglomerating data from the HapMap and the Human Genome Diversity Project, providing useful insights into the population structure of the three major population groups in Asia. In addition, this resource also surveyed across the genome for variation in regional patterns of LD between the HapMap and SGVP populations, and for signatures of positive natural selection using two well-established metrics: iHS and XP-EHH. The raw and processed genetic data, together with all population genetic summaries, are publicly available for download and browsing through a web browser modeled with the Generic Genome Browser.

[Supplemental material is available online at http://www.genome.org.]

The detailed survey of human genomic variation across four populations globally from the International HapMap Project (The International HapMap Consortium 2005, 2007) has yielded valuable insights into the design (de Bakker et al. 2005; Pe'er et al. 2006) and analysis (Marchini et al. 2007) of studies that examine the entire genomic landscape for correlation with the onset of diseases or traits. These genome-wide association studies (GWAS) typically detect indirect associations, where the identified genetic variants by themselves are not biologically functional but are in the neighborhood and thus are correlated or are in linkage disequilibrium (LD) with the causal polymorphisms. Commercial genotyping arrays for genome-wide studies utilize these informative markers for providing suitably dense genomic coverage, which with the appropriate use of sophisticated imputation methods can increase the effective genomic coverage of these arrays to that of the HapMap by statistically inferring the genotypes of the remaining unobserved markers in the HapMap (Marchini et al. 2007; Servins and Stephens 2007). The accuracy of genotype imputation, however, relies on having reference databases that are representative of the target populations to be imputed. While it has been shown that tagging SNPs identified from the HapMap are expected to be portable across other non-African populations (de Bakker et al. 2006; Conrad et al. 2006; Huang et al. 2009), imputation performance is expected to be optimized if local reference haplotypes are used (Huang et al. 2009; Jallow et al. 2009). The ability to reproduce an association finding in other populations through replication studies or meta-analyses is a prerequisite to validating the authenticity of the discovery (NCI-NHGRI Working Group on Replication in Association Studies 2007), and this fundamentally relies on having a similar LD structure between the identified variant and the functional polymorphism in these populations (Teo et al. 2009a). The success of imputation procedures, meta-analyses, and replication studies thus hinges critically on possessing sufficient knowledge on the extent of genomic variation between multiple populations. The Singapore Genome Variation Project (SGVP) is established with this aim of characterizing genomic variation and positive natural selection in three major population groups in Asia.

Singapore is a relatively young country with a migratory history predominantly consisting of immigrants with Chinese, Malay, and Indian genetic ancestries from neighboring countries such as China, India, Indonesia, and Malaysia (Saw 2007). The Chinese community consists mainly of descendents of Han Chinese settlers from the southern provinces of China, such as Fujian and Guangdong, and currently represents the dominant racial population in Singapore, accounting for 76.7% of the resident population from the Singapore Census conducted in 2000 (Saw 2007). While Han Chinese represents the largest ethnic group amongst the Chinese globally, there are a considerable number of sub-ethnicities within the Han classification with a diverse range of dialects and cultural diversity, with established genetic heterogeneity following a geographical north–south cline (Chu et al. 1998; Wen et al. 2004). The majority of the early Chinese immigrants to

[7]These authors contributed equally to this work.
[8]These authors jointly directed the project.
[9]Corresponding author.
E-mail ephcks@nus.edu.sg; fax 65-6-7791489.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.095000.109. Freely available online through the *Genome Research* Open Access option.

19:2154–2162 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

Singapore were mainly attributed to the dialect groups of Hokkien, Teochew, Cantonese, Hakka, and Hainanese (Saw 2007) that are predominantly found in Southern China. While Malays formed the dominant race in Singapore prior to the colonization by British settlers, the proportion of indigenous Malays has been surpassed by migrant Malays from Peninsula Malaysia, as well as Javanese and Boyanese people from Indonesia. Cultural and religious similarities have resulted in intermarriages between the immigrant and local Malays, whose descendents are now collectively known as Malays and account for 13.9% of the Singapore population (Saw 2007). The British colonization of Singapore also brought Indian migrants from the Indian subcontinent, with the majority consisting of Telugas and Tamils from southeastern India and a minority of Sikhs and Pathans from north India. The definition of Indians in Singapore comprises people with paternal ancestries tracing back to the Indian subcontinent, and, as a race, Indians represent 7.9% of the Singapore population. Cumulatively, the SGVP resource has the potential for representing the genetic diversity across multiple large populations in Asia while serving as a useful complement to the HapMap database.

This paper aims to describe the SGVP resource, which genotyped in excess of 2 million polymorphisms across 99 Chinese, 98 Malay, and 95 Indian individuals. The genotype data, phased haplotypes, and other data summaries for this resource have been modeled after the format of the International HapMap Project and are publicly available online. In addition, this paper details the extent of population differences between the SGVP, the HapMap, and the populations from the Human Genome Diversity Project (HGDP) (Rosenberg et al. 2002; Jakobsson et al. 2008; Li et al. 2008). We also compared the diversity of SNPs and haplotypes between the populations in the HapMap and SGVP, with a particular focus on the extent of LD variations between these populations. A genome-wide survey for candidate signatures of recent positive natural selection was also performed in the SGVP populations, replicating a number of previous findings from HapMap while identifying novel candidates, particularly in the Malay and Indian population groups.

Results

Sample and SNP quality control

A total of 292 individuals comprising of 99 Chinese, 98 Malays, and 95 Indians were genotyped across 2,007,788 SNPs on the Affymetrix SNP6.0 and Illumina 1M arrays, of which 268,667 SNPs overlap between the two platforms. The fidelity and accuracy of the genotype data are of paramount importance in establishing reference haplotype maps. We implemented a hierarchical quality control (QC) procedure that begins with an initial round of SNP QC to identify a set of "pseudo-cleaned" SNPs for detecting problematic samples. Samples with high levels of missingness, potential relatedness, and discordance between self-reported and genetically inferred population membership were identified and excluded from further analyses (Supplemental Table S1). A final round of SNP QC was performed within each population separately on the basis of missingness, departures from Hardy–Weinberg equilibrium (HWE), excessive discordance in the genotypes for the duplicated samples, and annotation failures. A total of 96 Chinese, 89 Malays, and 83 Indians remained after merging the SNP data from both arrays. Here, we further excluded SNPs that were common on both arrays but with <95% concordant genotypes, and SNPs that mapped to different alleles on the forward strand according to the SNP manifests from Affymetrix and Illumina. This yielded a final post-QC set with 1,584,040 autosomal SNPs for Singapore Chinese (CHS); 1,580,905 SNPs for Singapore Malays (MAS); and 1,583,454 SNPs for Singapore Indians (INS) (Supplemental Table S1), with an average inter-SNP distance of 2 kb across most of the genome (Supplemental Figs. S1, S2). The overall concordance in the genotype calls for the sample duplicates was 99.899%, at an overall call rate of 99.285%. Details of the QC process can be found in the Methods and Supplemental material.

Population structure

Principal components analysis (Price et al. 2006) and Wright's F_{ST} statistic (Wright 1951) were used to explore the extent of population differentiation between the SGVP, HapMap, and HGDP populations (Supplemental Table S2).

In the context of global genetic diversity, Singapore Chinese, the HapMap Han Chinese in Beijing, China (CHB), and HapMap Japanese in Tokyo, Japan (JPT) were virtually indistinguishable, while Singapore Malays were observed to be highly similar to the East Asian populations in general (Fig. 1A). Singapore Indians were comparable to samples from Central and South Asia, and genetically closer to the samples with European ancestries than to the East Asian samples from HGDP. As with the non-African populations in HapMap and HGDP, all three Singapore groups were considerably distinct from the HapMap Yoruba samples from the Ibadan region of Nigeria (YRI) and the African samples in the HGDP. The first axis of variation at this global level effectively distinguished samples from the Far East from Africans, while the second axis of variation addressed the difference between European and African ancestries. Comparing between the East Asian populations, the first axis separated the Yakut people of Siberia from Chinese sub-ethnic groups mainly located in Southern China (Dai, Lahu) and Southeast Asia (Cambodian, CHS) (Fig. 1B). When we consider only the HapMap and SGVP populations, the third axis of variation separated INS from the HapMap Utah samples with ancestry from Northern and Western Europe (CEU), while MAS was differentiated from the Far East Asian cluster (comprising CHB, CHS, and JPT) by the fourth axis of variation (Fig. 1C).

Comparing within the three Far East Asian populations from HapMap and SGVP, the JPT samples were clearly more different from the two Chinese cohorts (F_{ST} = 0.3% with CHB; 0.4% with CHS) than between the two Chinese cohorts themselves, although substantial dissimilarities exist to distinguish between the two Chinese cohorts (F_{ST} = 0.2%; Fig. 1D). In the latter analysis, a few CHB samples were clustered together with most of the CHS samples and vice versa (see also Supplemental Fig. S3). The separation seen between samples from CHB and CHS may be indicative of a north–south genetic cline, as Singapore Chinese are predominantly descendents of immigrants from southern provinces in China, while we expect the HapMap Han Chinese in Beijing samples to mainly reflect the genetic ancestry from northern China. It is possible that the HapMap Han Chinese samples from Beijing have included individuals with genetic ancestries more commonly seen in Southern China, and likewise with the Singapore Chinese samples, as it is evident from the Chinese samples in HGDP (Fig. 1B) that the designation of Han Chinese encompasses people from genetically distinguishable sub-groups or sub-ethnicities. Within the SGVP populations, the INS was more differentiated compared with CHS (F_{ST} = 3.9%) and MAS (F_{ST} = 2.7%), than between the Chinese and the Malay samples (F_{ST} = 0.6%, Fig. 1E).

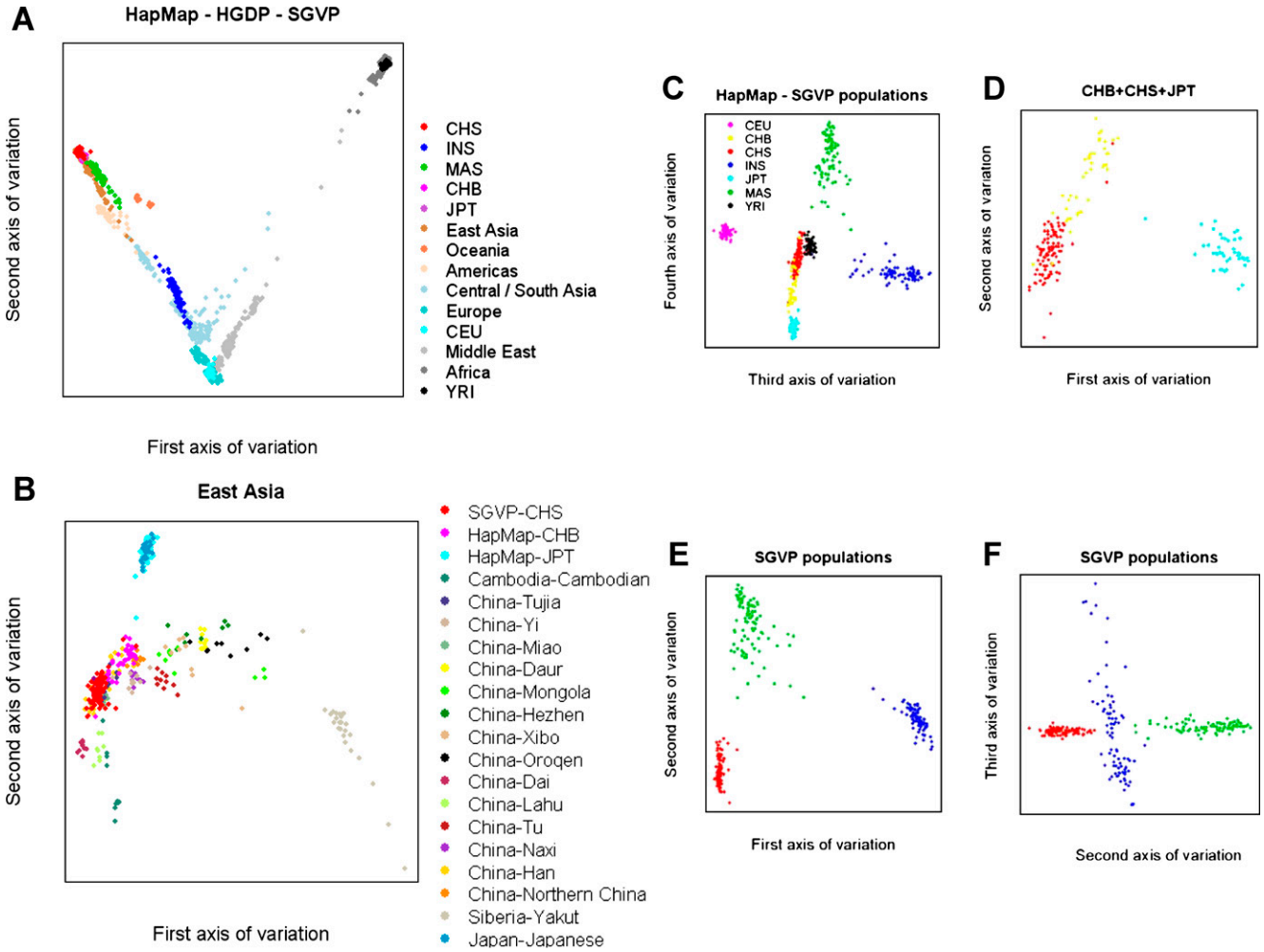


Figure 1. Principal component analysis plots of genetic diversity across HapMap, HGDP, and SGVP populations. Each figure represents the genetic diversity seen across the populations considered, with each sample mapped onto a spectrum of genetic variation represented by two axes of variations corresponding to two eigenvectors of the PCA. (*A*) Individuals from each population in the HapMap and SGVP are represented by a unique color, while samples from HGDP are broadly grouped by geography in which a unique color is assigned to each geographical location. (*B*) Comparison between CHS and samples from Far East Asia found in the HapMap and HGDP. (*C*) A plot of the third and fourth axes of variation for the seven populations from HapMap and SGVP. (*D*) A plot of the first two axes of variation when the PCA is run on only the three Far East Asian populations comprising the Singapore Chinese, HapMap Han Chinese in Beijing, China, and Japanese in Tokyo, Japan. (*E*) A plot of the first two principal components in a separate analysis within the three SGVP populations. (*F*) A plot of the second and third principal components within the SGVP populations. The same color scheme has been used in *C*–*F*; the legend for the color assignment can be found in *C*.

Interestingly, the third axis of variation indicated there was substantial genetic variability within the Indian samples (Fig. 1F), which may be attributed to the numerous ethnicities that comprise the Indian population.

SNP and haplotype diversity

The availability of accurate genome-wide data allows the assessment of genetic diversity across the SGVP populations. At the SNP level, there was considerably less variance in the allelic spectrum between CHS and MAS, relative to comparisons between either population and INS (Fig. 2A–C), while, expectedly, CHS was most similar to CHB (Fig. 2D; Supplemental Figs. S4, S5). In a genome-wide survey for regions that are highly differentiated in the SGVP populations, the top 10 regions were attributed mainly to allele frequency variations between INS and the two other populations and encapsulated well-documented regions of genomic differentiation between East Asian and other global populations, including *EDAR* (Sabeti et al. 2007) and *VKORC1* (Lal et al. 2006; Lee et al. 2006) (Table 1).

To investigate the extent of haplotype diversity across the seven SGVP and HapMap populations, we calculated the percentage of the chromosomes within each population that can be accounted for by a specified number of distinct haplotypes across 22 regions of 500 kb. We observed that there was considerably higher haplotype diversity in YRI compared with the rest, while the populations with Far East Asian ancestries (CHB, CHS, and JPT) have the lowest haplotype diversity (Supplemental Fig. S6). For example, 12 haplotypes accounted for only 43% of the YRI chromosomes, and between 73% (for JPT) and 79% (for CHS) for the three populations with Far East Asian ancestries. Among the SGVP populations, INS has the greatest haplotype diversity, with 12 haplotypes accounting for 57% of the INS chromosomes. This is followed by MAS, with 68% of the chromosomes accounted for by

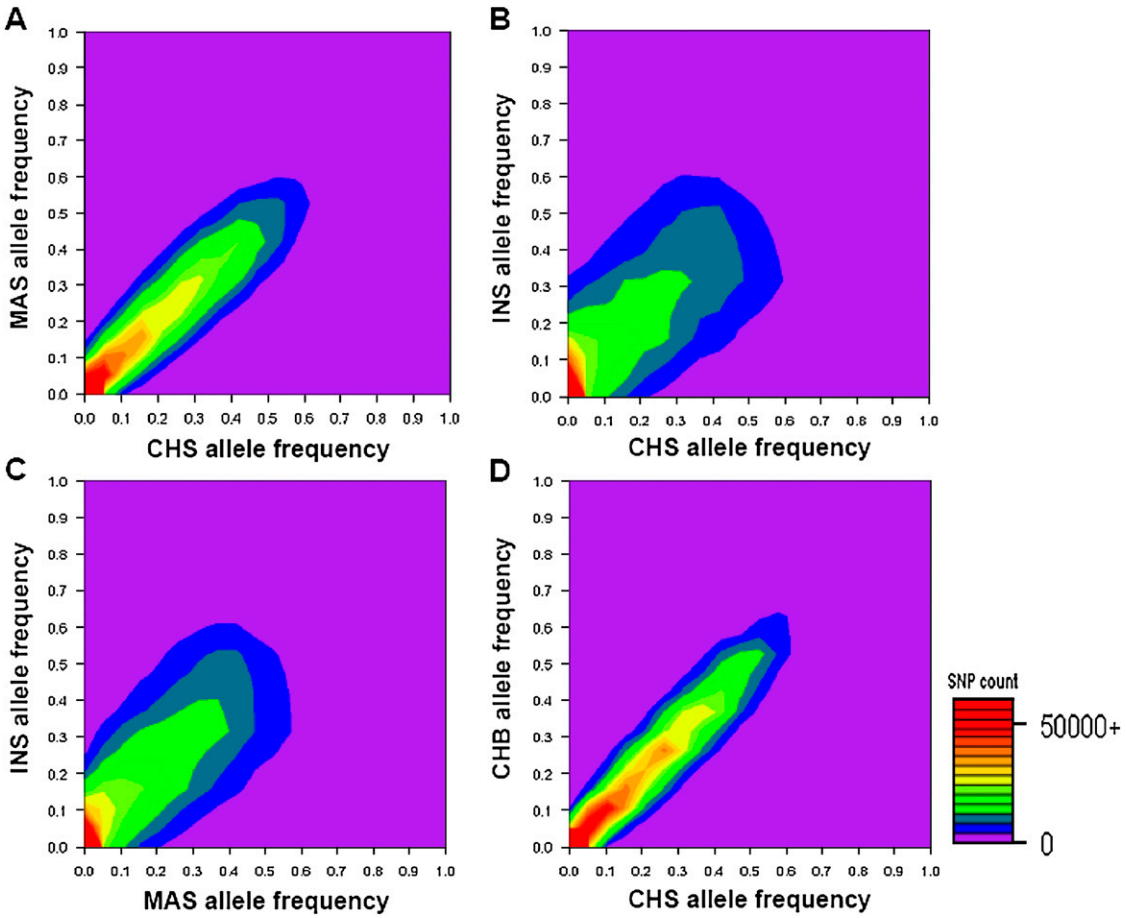


Figure 2. Allele frequency comparison between pairs of populations. The axes in each figure represent the allele frequencies for each of the two represented populations. For each SNP, we define the minor allele after agglomerating the genotype data from all three SGVP populations and subsequently calculate the frequency of this allele in each population. Twenty allele frequency bins each spanning 0.05 units are constructed for each population, and we tabulate the number of SNPs found in each bin. The intensity of the contour represents the number of SNPs that displayed the corresponding allele frequencies in the two populations, from a low number of SNPs (purple) to a higher number of SNPs (red). The figure panels compare the allelic spectrum among CHS-MAS (*A*), CHS-INS (*B*), MAS-INS (*C*), and CHS-CHB (*D*).

12 haplotypes. The pattern of haplotype sharing between these populations was very similar across the 22 regions, and we illustrate this with chromosome 1 (Supplemental Fig. S6). A high degree of haplotypes was shared between CHB, CHS, and JPT, and it was evident that there were different haplotypes present in CEU, INS, and YRI that were either absent or at low frequencies in the rest of the populations. These analyses concurred with the observations from the analysis of population structure that CHB, CHS, and JPT are more genetically similar compared with the rest of the populations, with INS being the most genetically diverse among the SGVP populations.

Linkage disequilibrium, tagging efficiency, and LD variation

One important utility of the SGVP resource is the comparison of the extent of LD between the SGVP and HapMap populations, as this reflects the tagging efficiency for genotyping arrays that were designed using the patterns of LD that were observed in the HapMap populations. Overall, the SGVP populations exhibited similar rates of LD decay with increasing distance as compared with the HapMap non-African populations, with CHS and INS having the greatest and least conservation of LD, respectively, with distance amongst the three SGVP populations (Fig. 3). This is similarly reflected in the number of tagging SNPs that are required to capture all the common SNPs in the SGVP panels at a pairwise r^2 threshold of 0.8, where between 349,800 and 406,900 SNPs are required for CHS and INS, respectively (Table 2). For comparison, the corresponding range for the HapMap populations is between 358,800 and 546,300 for JPT and YRI, respectively. Intriguingly, we observed the number of tagging SNPs required at a pairwise r^2 threshold of 1 for each SGVP population is almost comparable to the number required for YRI, although this is likely to be a consequence of designing commercial genotyping microarrays utilizing LD patterns observed in the HapMap populations.

One of the factors that affects the reproducibility of the association results from GWAS is the degree of similarity in the correlation structure between the causal variants and the reported SNPs in these populations (Teo et al. 2009a). By comparing the extent of LD differences in a sliding-window approach between any two populations, we identified the regions that are found in the top 5% of the distribution of LD differences as candidate regions of LD variation, where consecutive signals in the top 5% within 25 kb are binned as a single region (see Methods). As a significant proportion of GWAS has been performed in populations

Table 1. Top 10 regions across the genome with strongest signals of genetic differentiation (F_{ST}) across all three SGVP populations

Chromosome	Region (start–end)	No. of SNPs	Genes	Top SNP	Minor allele frequency[a]						
					CHS	MAS	INS	CEU	CHB	JPT	YRI
1	203,126,372	1	*NFASC*	rs7541623	0.185	0.185	0.886	0.017	0.244	0.300	0.508
2	16,659,951–16,660,077	2	*FAM49A*	rs751192	0.063	0.180	0.801	0.867	0.100	0.136	0.508
2	108,305,167–108,956,812	16	*SULT1C4, GCC2, LIMS1, RANBP2, CCDC138, EDAR*	rs3827760	0.083	0.573	0.994	1.000	0.044	0.205	1.000
2	215,991,803–216,030,633	6	*FN1*	rs1437787	0.036	0.225	0.801	0.771	0.034	0.067	0.850
3	81,515,924–81,742,773	5	*GBE1*	rs276105	0.042	0.114	0.693	0.542	0.044	0.125	0.342
6	131,499,350	1	*AKAP7*	rs6569733	0.109	0.163	0.807	0.862	0.100	0.182	0.508
11	134,012,618	1	—	rs3017964	0.100	0.303	0.873	0.883	0.156	0.200	0.692
12	111,440,158–111,465,954	9	*PTPN11*	rs6489847	0.078	0.219	0.837	0.879	0.089	0.133	0.678
14	96,394,042–96,429,553	2	*VRK1*	rs12434466	0.104	0.315	0.861	0.992	0.131	0.179	0.945
16	30,364,851–31,055,049	27	*ZNF(768, 747, 764, 689, 629, 668, 646), ITGAL, PRR14, FBRS, SRCAP, PHKG2, RNF40, BCL7C, CTF1, FBXL19, ORAI3, SETD1A, STX4, BCKDK, PRSS8, MYST1, VKORC1, PRSS36*	rs11864054	0.078	0.203	0.855	0.578	0.056	0.080	1.000

[a]The minor allele is defined with respect to CHS (Singapore Chinese). (MAS) Singapore Malays, (INS) Singapore Indians, (CEU) Utah samples with ancestry from Northern and Western Europe, (CHB) Han Chinese in Beijing, (JPT) Japanese in Tokyo, (YRI) Yoruba samples from the Ibadan region of Nigeria; (SGVP) Singapore Genome Variation Project.

of European descent, Supplemental Table S3 shows the top 10 candidate regions of LD variation between each SGVP population and CEU, while a complete listing of the identified regions in the top 0.1% of the distribution between pairs of populations from SGVP and HapMap can be found in Supplemental Table S4. Perhaps unsurprisingly, one of these regions observed between INS and CEU spans the *SLC24A5* gene, which has been established to be functionally involved with skin pigmentation (Lamason et al. 2005). A region that shows considerable signals of LD variations from multiple pairs of populations and that coincided with reported association signals from GWAS spans the *CDKAL1* gene, which has been implicated with Type 2 diabetes in populations with European ancestry (Saxena et al. 2007; Scott et al. 2007; Steinthorsdottir et al. 2007; Zeggini et al. 2007) and also in Asian populations such as the Chinese (Liu et al. 2008; Wu et al. 2008), Koreans (Ng et al. 2008), and Japanese (Tabara et al. 2009). Our analysis indicates that the implicated variant rs7754840 is found in a region with extensive LD differences between multiple groups (Fig. 4). The population-specific recombination profiles differed between the SGVP and HapMap populations as the higher SNP density from the HapMap data allowed inference of the recombination rates at a finer scale compared with the SGVP (Myers et al. 2005).

Comparing the genome-wide LD patterns between the two Chinese populations (CHB and CHS), the top 10 regions identified contain an olfactory cluster on chromosome 1 as well as two *HLA* gene clusters in the major histocompatibility complex (*MHC*) region on chromosome 6 (Supplemental Table S5), suggesting that these regions are highly polymorphic even between two relatively homogeneous populations. Intriguingly, we observed that three regions outside the top 10 were in the vicinity of candidate genes for common metabolic disorders (*FABP2*, *PCSK1*, *CLOCK*) that have been implicated for climate adaptations (Hancock et al. 2008). The frequencies of the derived allele associated with greater tolerance to cold climate at the A54T (rs1799883) polymorphism in *FABP2* were significantly lower in CHS (22.4%) and MAS (15.7%) when compared with CHB (31.4%) and JPT (30.0%), consistent with reported findings of a significant correlation with latitude (Hancock et al. 2008). For comparison, the frequencies for CEU, INS, and YRI were 37.3%, 30.7%, and 20.8%, respectively.

Signatures of positive natural selection

Genome-wide data on the three SGVP populations also permit the survey of signatures of recent positive natural selection through the detection of uncharacteristically long haplotypes in the genome. Using the single-SNP integrated haplotype score (iHS) and the XP-EHH score (see Methods and Supplemental material), we observed that most of the signals detected by iHS in the SGVP populations concur with those established in the HapMap populations, particularly for signals that span multiple SNPs (Supplemental Table S6). Novel candidates for positive selection were identified in each of the three SGVP populations, with the largest number observed in INS. Supplemental Table S7 lists the top 10 candidate regions for recent positive selection in each SGVP population. Across the genome, selection signals that corroborated with earlier findings from the HapMap in genes with well-documented

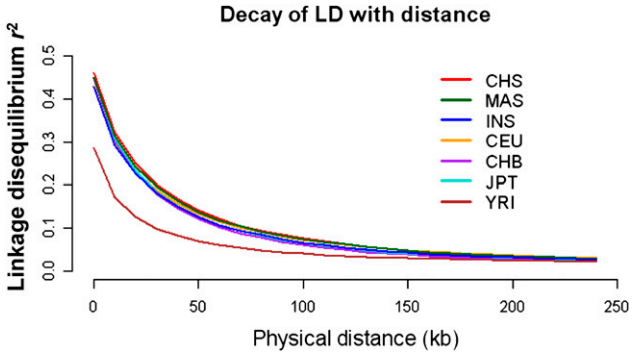


Figure 3. Decay of LD with distance. Decay of LD as measured by the r^2 statistic with increasing distance up to 250 kb for each of the HapMap and SGVP populations, where 90 chromosomes were chosen from each population to perform the LD calculation. Only SNPs with minor allele frequencies ≥5% in each population were considered in this analysis.

Table 2. Number of tagging SNPs required to capture all 979,573 common SNPs in each of the SGVP and HapMap populations

	SGVP			HapMap			
r^2 threshold	CHS	MAS	INS	CEU	CHB	JPT	YRI
$r^2 \geq 0.5$	195,462	205,927	228,701	211,011	209,167	205,956	367,593
$r^2 \geq 0.8$	349,814	371,631	406,814	370,941	364,540	358,898	546,250
$r^2 = 1.0$	633,161	670,423	680,740	562,479	547,233	530,642	679,687

A common SNP is defined as one with a minor allele frequency of ≥5% in all three SGVP populations. The HapMap panels are thinned to contain the same set of SNPs for comparison. See Table 1 for definitions of abbreviations.

functions include the alcohol dehydrogenase (*ADH*) gene cluster in CHS and INS, genes involved in skin pigmentation (*SLC24A5* in INS, *OCA2* in CHS and MAS, *TYRP1* in CHS and INS, *MYO5A* in all three populations), sucrose metabolism (*SI* in CHS and MAS), brain development and function (*CENPJ* in CHS and INS, *MCPH1* in MAS and INS, *CDK5RAP2* in all three), regulation of energy and appetite (*LEPR* in CHS and MAS), and low-density lipoprotein cholesterol (*LDLR* in CHS and INS, *APOB* in all three) (Supplemental Table S8). The concurrence of positive selection across multiple populations is reassuring, although we advocate caution in drawing immediate relevance to the biological interpretations.

Discussion

The aim of the Singapore Genome Variation Project is primarily to provide a genomic resource for studying the three major population groups in Southeast Asia that can be used in the design and analysis of studies for discovering the genetic basis of complex traits and common diseases. This is achieved by making a database of 1.6 million polymorphisms assayed in 268 individuals publicly available for researchers, with the data intentionally formatted in a similar structure to complement the International HapMap Project.

It has been historically documented that Chinese migrants into early Singapore predominantly consisted of people from the southern provinces of China. Our analysis of population structure in East Asia where CHS clustered together with Chinese subethnicities from Southern China and Southeast Asia supported this claim, together with the observation at *FABP2* that Singapore Chinese are less likely to carry the genetic variant that confers greater tolerance to cold climates compared with the Han Chinese in Beijing from Northern China. As this variant is similarly found at low frequency in the Malays with equatorial habitats, this suggests

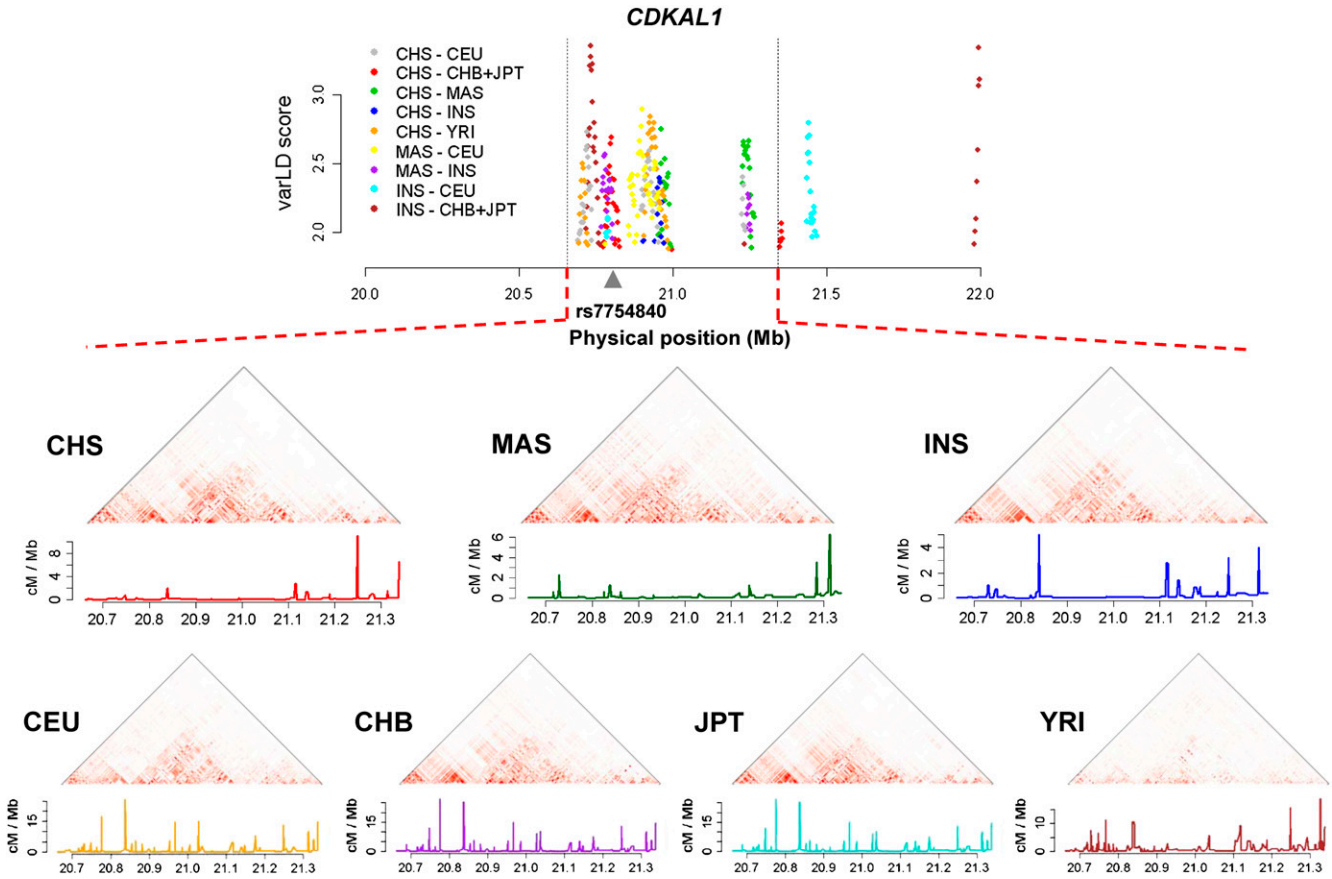


Figure 4. LD variation and population-specific recombination rates at *CDKAL1*. The extent of LD variation between pairs of SGVP and HapMap populations at the *CDKAL1* gene, with separate LD heatmaps and recombination rates estimated from genotype data at each population. Population-specific recombination rates are shown except for CHB and JPT, where the same HapMap estimated recombination rates for JPT+CHB are used.

that the difference between Singapore Chinese and the Han Chinese from Beijing reflects the genetic diversity found between the northern and southern parts of China.

One of the main motivations in establishing this genomic resource is to explore the possibility of localizing functional polymorphisms through combining association signals across populations with diverse genetic backgrounds. Preliminary findings from targeted sequencing and sequence-resolution imputation studies have suggested that the presence of long LD in populations of European and East Asian descent is a hindrance to this process of fine-mapping, as what emerges from these sequencing studies are sets of SNPs in perfect or almost perfect LD that are virtually impossible to distinguish between for isolating the causal variants. However, as the patterns of LD between the causal variants and the neighboring SNPs can vary across populations, pooling GWAS results with dense population-specific reference haplotypes across multiple populations can be expected to minimize the number of SNPs that are potential candidates to be functional. While the full merit of such transethnic fine-mapping approaches will only be realized with sequence-level haplotypes in the relevant populations, we expect the availability of dense genome-wide data for more populations will at least serve a few purposes: (1) to serve as reference panels to impute against for the purpose of extending the coverage of current genome-wide experiments in Southeast Asia to at least 1.6 million SNPs; (2) to prioritize SNPs that emerged from genome-wide scans for replication in Southeast Asia; and (3) to perform genome-wide comparisons of LD between populations, which will be valuable in identifying regions where transethnic fine-mapping holds the greatest promise.

To date, most genetic research and genomic databases (other than the HapMap) have either focused on populations of European descent or have surveyed comparatively few samples in each Asian population (e.g., the Human Genome Diversity Project). The SGVP provides a timely complement to these databases by providing a publicly available resource of 1.6 million polymorphisms genotyped in 268 samples from three major population groups in Asia. To facilitate the access, analysis, and display of the SGVP data, we have designed a genome browser that is publicly available at http://www.nus-cme.org.sg/SGVP/ (Supplemental Fig. S7). We expect this resource will be valuable for advancing genetic and genomics science in Asia.

Methods

Samples

Subjects enrolled in the SGVP were originally recruited for an interpopulation study on the genetic variability to drug response, where 100 individuals from each of the Chinese, Malay, and Indian population groups were anonymously and randomly chosen from the manifest to partake in SGVP, with only gender and population information. Of these 300 samples, genomic DNA samples for 99 Chinese, 98 Malay, and 95 Indians were chosen for genotyping. Population membership was ascertained on the basis that all four grandparents belong to the same population group. Ethical consent for the original study on drug response and further ethical approval for the extension to genome-wide genotyping were granted by two independent Institutional Review Boards at the National University Hospital (Singapore) and the National University of Singapore, respectively.

SNP genotyping

Genomic DNA for all 292 individuals was assayed on the Affymetrix SNP6.0 Genotyping Chip and the Illumina 1M-single DNA Analysis BeadChip. Preliminary genotypes for 3022 control probes on the Affymetrix array were called using the DM algorithm (Di et al. 2005) for sample QC. The set of genotype data from the Affymetrix array used in downstream analyses was called using the BirdSeed algorithm (Korn et al. 2008). Genotypes for the Illumina array were assigned using the proprietary calling algorithm GenCall in the BeadStudio Suite (Oliphant et al. 2002; Fan et al. 2004). We implemented a threshold of 0.15 on the GC score during the calling process: a valid genotype was assigned if the GC score was ≥0.15; otherwise, a missing genotype was assigned.

Quality assessment

The quality of the genotypes for data from both arrays was assessed independently, in the following four phases in sequential order: (1) preliminary SNP QC on the autosomal chromosomes to identify a set of "pseudo-cleaned" SNPs for sample QC; (2) sample QC to remove sample duplicates, related samples, or samples with high rates of missing data; (3) identification of samples with inconsistent population membership or inconsistent gender when comparing between the self-reported and genetically inferred data; (4) another round of SNP QC after excluding samples identified by (2) and (3) to yield the set of SNPs for inclusion in the SGVP database. Post-QC data for both arrays were available for 96 Chinese, 89 Malay, and 83 Indian samples. For SNPs that are common to both Affymetrix and Illumina, only those with ≥95% concordant genotypes between the two arrays were retained.

Assessing population structure

Population structure between the HapMap and SGVP populations was assessed by principal components analysis (PCA) with EIGENSTAT (Price et al. 2006). We thinned the available SNPs by using every tenth SNP out of the 1,423,464 SNPs that were common between HapMap and SGVP, consisting of 142,347 SNPs, to reduce the extent of LD between the SNPs used in the PCA. The F_{ST} calculation uses the same formula as that used by the International HapMap Project (The International HapMap Consortium 2005), which accounts for the different number of samples in each population (see Supplemental material).

Haplotype phasing and LD calculation

The software *fastPhase* (Scheet and Stephens 2006) was used to perform the phasing of the genotype data within each population separately. The parameters used in the analysis were optimized to yield minimal error rates within realistic running time of the analysis. The LD between a focal SNP and any SNP found within 250 kb upstream and downstream of the focal SNP was calculated using the software Haploview (Barrett et al. 2005). LD was measured by the square of the genetic correlation coefficient r^2, D', and the LOD score, and was calculated off the phased haplotype data. Comparisons of LD across populations utilized 45 samples from each population to avoid the effects of different sample sizes.

Comparing allele frequency spectrum

We considered the same set of SNPs that passed QC across all the SGVP panels. For each SNP, the minor allele was identified after agglomerating the genotypes from all three SGVP populations. The frequencies of the minor alleles were subsequently calculated within each SGVP populations and categorized in 20 bins of size 0.05 spanning 0 to 1.

Quantifying haplotype diversity

For each chromosome, we randomly selected a 500-kb region, avoiding centromeres and genomic regions with low SNP density.

For an unbiased comparison across all seven population panels from HapMap and SGVP, we considered only the SNPs that were common to all seven panels. In each of the 500-kb regions, we identified the number of distinct haplotype forms. We then quantified haplotype diversity by the proportion of chromosomes from each population that had been accounted for by a specific number of haplotypes. This procedure is similar to that established for quantifying haplotype diversity across multiple populations (Bonnen et al. 2006). In order to investigate the extent of haplotype sharing, chromosomes from the region in chromosome 11 were clustered and visualized with the use of *haplosim* and *hapvisual* from the R package *haplosuite* (Teo and Small 2009). Briefly, *haplosim* identifies the canonical haplotypes in each region across all seven populations, where each canonical haplotype is defined as a specific haplotype configuration to which a substantial proportion of the individuals are highly similar. Each chromosome is subsequently mapped either uniquely to one of these canonical haplotypes, or as a mosaic of these haplotypes. We explicitly chose to implement an upper limit of seven possible canonical haplotypes in our analysis of the HapMap and SGVP populations. The outcome of the haplotype clustering was subsequently fed into *hapvisual*, which produced a visualization of the haplotype clustering for each population, where each canonical haplotype is assigned a unique color that remains consistent across the populations.

Analysis of LD variation

Comparison of regional LD between two populations was performed with the *varLD* algorithm (Teo et al. 2009b). Briefly, we considered windows of 50 consecutive SNPs found in both populations and calculated the signed r^2, defined as the r^2 with the sign of the D' metric, between all possible pairs of these SNPs. Consequently, we constructed a 50×50 symmetric matrix for each population where the $(i, j)^{\text{th}}$ element represents the signed r^2 metric between the i^{th} and j^{th} SNPs calculated. We compared the equality between the two matrices by comparing the extent of departures between the eigenvalues, given by the sum of the absolute difference between the ranked eigenvalues for the two matrices that yields a score for each window of 50 SNPs. The extent of LD differences in each window was assessed by comparing the relative rank of the score obtained against the distribution of scores in the genome, and we identified regions that constituted the top 5% of the distribution of the scores. For visualizing the signals from comparisons across multiple population pairs, we standardized the scores to have a mean of zero and a standard deviation of one. Signals in the top 5% of the distribution were binned into regions if two consecutive signals were found within 25 kb.

Detecting signatures of positive selection

We used the single-SNP integrated haplotype score (iHS) statistic introduced by Voight et al. (2006) to identify signals of positive selection within each of the HapMap and SGVP populations. This analysis followed the set-up described in Sabeti et al. (2007). To compare signals of positive natural selection that differ between populations, we used the XP-EHH test with the same set-up as introduced and described by Sabeti and colleagues (Sabeti et al. 2007).

A full description of the methods with additional figures and tables for the methodologies can be found in the Supplemental material.

Acknowledgments

We thank three anonymous reviewers and E.S. Tai for their insightful comments that helped improve the manuscript. We thank all the subjects in this study for their participation. This project also acknowledges the support of the Yong Loo Lin School of Medicine, the National University Health System, the Life Science Institute and Office of Deputy President (Research and Technology) from the National University of Singapore. We also acknowledge the support of the Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore.

Author contributions: Y.Y.T., M.S., and K.S.C. jointly conceived and designed the experiment; Y.Y.T., X.S., and R.T.H.O. wrote the paper; Y.Y.T., X.S., R.T.H.O., A.K.S.T., C.S.K., E.T., K.S.S., and J.C. analyzed the data; R.T.H.O. and X.S. designed the website; E.J.D.L. contributed samples; C.S.K. and M.S. coordinated the genotyping; M.S. and K.S.C. jointly directed the project.

References

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Bonnen PE, Pe'er I, Plenge RM, Salit J, Lowe JK, Shapero MH, Lifton RP, Breslow JL, Daly M, Reich DE, et al. 2006. Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. *Nat Genet* **38:** 214–217.

Chu JY, Huang W, Kuang SQ, Wang JM, Xu JJ, Chu ZT, Yang ZQ, Lin KQ, Li P, Wu M, et al. 1998. Genetic relationship of populations in China. *Proc Natl Acad Sci* **95:** 11763–11768.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, Prichard JK. 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38:** 1251–1260.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* **37:** 1217–1213.

de Bakker PI, Burtt NP, Graham RR, Guiducci C, Yelensky R, Drake JA, Bersaglieri T, Penney KL, Butler J, Young S, et al. 2006. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat Genet* **38:** 1298–1303.

Di X, Matsuzaki H, Webster TA, Hubbell E, Liu G, Dong S, Bartell D, Huang J, Chiles R, Yang G, et al. 2005. Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. *Bioinformatics* **21:** 1958–1963.

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen MS, Steemers F, Butler SL, Deloukas P, et al. 2004. High parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* **68:** 69–78.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard J, Coop G, Di Reinzo A. 2008. Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet* **4:** e32. doi: 10.1371/journal.pgen.0040032.

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. 2009. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84:** 235–250.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **427:** 1299–1320.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung H, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451:** 998–1003.

Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, Kivinen K, Bojang KA, Conway DJ, Pinder M, et al. 2009. Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet* **41:** 657–665.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40:** 1253–1260.

Lal S, Jada SR, Xiang X, Lim WT, Lee EJ, Chowbay B. 2006. Pharmacogenetics of target genes across the warfarin pharmacological pathway. *Clin Pharmacokinet* **45:** 1189–1200.

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, et al. 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science* **310:** 1782–1786.

Lee SC, Ng SS, Oldenburg J, Chong PY, Rost S, Guo YJ, Yap HL, Rankin SC, Khor HB, Yeo TC, et al. 2006. Inter-ethnic variability in warfarin requirement is explained by VKORC1 genotype in an Asian population. *Clin Pharmacol Ther* **79:** 197–205.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319:** 1100–1104.

Liu Y, Yu L, Zhang D, Chen Z, Zhou DZ, Zhao T, Li S, Wang T, Hu X, Feng GY, et al. 2008. Positive association between variations in CDKAL1 and type 2 diabetes in Han Chinese individuals. *Diabetologia* **51:** 2134–2137.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39:** 906–913.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310:** 321–324.

NCI-NHGRI Working Group on Replication in Association Studies. 2007. Replicating genotype-phenotype associations. *Nature* **447:** 655–660.

Ng MC, Park KS, Oh B, Tam CH, Cho YM, Shin HD, Lam VK, Ma RC, So WY, Cho YS, et al. 2008. Implications of genetic variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2 and FTO in type 2 diabetes and obesity in 6,719 Asians. *Diabetes* **57:** 2226–2233.

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. 2002. BeadArray technology: Enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* **32:** S56–S61.

Pe'er I, de Bakker PI, Maller J, Yelensky R, Altshuler D, Daly MJ. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* **38:** 663–667.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* **298:** 2381–2385.

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449:** 913–918.

Saw SH. 2007. *The population of Singapore,* 2nd edition. Institute of South East Asian Studies, Singapore.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316:** 1331–1336.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78:** 629–644.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316:** 1341–1345.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet* **3:** e114. doi: 10.1371/journal.pgen.0030114.

Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, et al. 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* **39:** 770–775.

Tabara Y, Osawa H, Kawamoto R, Onuma H, Shimizu I, Miki T, Kohara K, Makino H. 2009. Replication study of candidate genes associated with type 2 diabetes based on genome-wide screening. *Diabetes* **58:** 493–498.

Teo YY, Small KS. 2009. A novel method for haplotype clustering and visualization. *Genet Epidemiol* doi: 10.1002/gepi.20432.

Teo YY, Small KS, Fry AE, Wu Y, Kwiatkowski DP, Clark TG. 2009a. Power consequences of linkage disequilibrium variation between populations. *Genet Epidemiol* **33:** 128–135.

Teo YY, Fry AE, Bhattacharya K, Small KS, Kwiatkowski DP, Clark TG. 2009b. Genome-wide comparisons of variation in linkage disequilibrium. *Genome Res* **19:** 1849–1860.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4:** e72. doi: 10.1371/journal.pbio.0040072.

Wen B, Li H, Lu D, Song X, Zhang F, He Y, Li F, Gao Y, Mao X, Zhang L, et al. 2004. Genetic evidence supports demic diffusion of Han culture. *Nature* **431:** 302–305.

Wright S. 1951. The genetical structure of populations. *Ann Eugen* **15:** 323–354.

Wu Y, Li H, Loos RJ, Yu Z, Ye X, Chen L, Pan A, Hu FB, Lin X. 2008. Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8, and HHEX/IDE genes are associated with type 2 diabetes and impaired casting glucose in a Chinese Han population. *Diabetes* **57:** 2834–2842.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JRB, Rayner NW, Freathy RM, et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316:** 1336–1341.

Received April 15, 2009; accepted in revised form August 10, 2009.

Mei et al. BMC Bioinformatics 2010, 11:147
http://www.biomedcentral.com/1471-2105/11/147

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Identification of recurrent regions of copy-number variants across multiple individuals

Teo Shu Mei[1,2,5], Agus Salim[1,2], Stefano Calza[3], Ku Chee Seng[2], Chia Kee Seng[1,2], Yudi Pawitan[4*]

Abstract

Background: Algorithms and software for CNV detection have been developed, but they detect the CNV regions sample-by-sample with individual-specific breakpoints, while common CNV regions are likely to occur at the same genomic locations across different individuals in a homogenous population. Current algorithms to detect common CNV regions do not account for the varying reliability of the individual CNVs, typically reported as confidence scores by SNP-based CNV detection algorithms. General methodologies for identifying these recurrent regions, especially those directed at SNP arrays, are still needed.

Results: In this paper, we describe two new approaches for identifying common CNV regions based on (i) the frequency of occurrence of reliable CNVs, where reliability is determined by high confidence scores, and (ii) a weighted frequency of occurrence of CNVs, where the weights are determined by the confidence scores. In addition, motivated by the fact that we often observe partially overlapping CNV regions as a mixture of two or more distinct subregions, regions identified using the two approaches can be fine-tuned to smaller sub-regions using a clustering algorithm. We compared the performance of the methods with sequencing-based results in terms of discordance rates, rates of departure from Hardy-Weinberg equilibrium (HWE) and average frequency and size of the identified regions. The discordance rates as well as the rates of departure from HWE decrease when we select CNVs with higher confidence scores. We also performed comparisons with two previously published methods, STAC and GISTIC, and showed that the methods we consider are better at identifying low-frequency but high-confidence CNV regions.

Conclusions: The proposed methods for identifying common CNV regions in multiple individuals perform well compared to existing methods. The identified common regions can be used for downstream analyses such as group comparisons in association studies.

Background

Copy-number variants (CNVs) are genomic regions that contain an abnormal number of copies. In humans, we normally expect two copies of each autosomal region, but in CNV regions we may observe copy gains or losses. Current common technology used for CNV detection are high-density single nucleotide polymorphism (SNP) arrays or array comparative genomic hybridization (aCGH) arrays. Detection of CNVs from aCGH arrays is mostly based on locating change-points in intensity-ratio patterns that would partition each chromosome into several discrete segments [1-5]. On the other hand, the hidden Markov model (HMM) is particularly popular for detection of CNVs from SNP arrays, where the hidden states provide a natural way of combining information from the total signal intensity and the allele frequency values (see for example, [6,7]). These approaches detect CNVs sample-by-sample, and because of the high noise level in the intensity values, especially for SNP array data, the boundaries of the detected CNVs tend to vary among individuals. However, in a homogenous population, common CNV regions are likely to occur at the same genomic locations across different individuals. Our focus in this paper is to identify common CNV regions in multiple individuals from a given population.

Common CNV detection algorithms for SNP arrays report the log Bayes factor as a confidence score for each identified region; this provides a measure of the

* Correspondence: yudi.pawitan@ki.se
[4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, Stockholm 17177, Sweden

reliability of a detected CNV within an individual. Previous methods developed to identify recurrent CNV regions (see [8] for a review) were primarily developed for aCGH data and hence did not incorporate confidence scores. For example, a previously published method, STAC [9], uses two statistics to identify recurrent CNV regions. These statistics are based on the frequency of occurrence of the regions and the alignment of the regions. However, since the method does not incorporate confidence scores, every individual region contributes equally to the statistic, whereas in fact, inter-sample variability is bound to exist, where some regions are more likely to be true/false positives. Furthermore, STAC requires each chromosome to be split into non-overlapping windows of a user-defined fixed size. The algorithm then searches for evidence of common CNV regions within each window. The weakness of this is that the output from such an approach will only provide evidence of whether each window harbours a common CNV, but will not indicate the breakpoints of the CNV. Although we may decrease the window size to improve the resolution, in practice, doing so will incur an enormous computational burden.

In this paper, we investigated two different methods to detect common CNV regions. The methods take segmented data as the input. The first method estimates a statistic based on the frequency of occurrence of reliable CNVs, where reliability is determined by a high confidence score. The second method is based on a weighted frequency of occurrence of CNVs, where the weights are determined by the confidence scores. Figure 1 illustrates a common CNV region in chromosome 22, identified using the first method, and shows evidence of several distinct subregions within the identified common region. Hence, in addition to these methods, we also investigated the use of a clustering algorithm to split the common regions into smaller subregions.

To assess the performance of the methods, we ran the algorithms on 112 HapMap samples from the Illumina iControl database, composed of individuals from three populations (Yoruba, Caucasian and Asian). We compared the regions we identified to the regions identified using sequencing [10]. In general, the discordance rates with sequencing-based CNV regions as well as the rates of departure from HWE decreased when we filtered the individuals with a stricter confidence score threshold. To benchmark the proposed methods to currently available methods, we performed comparisons with STAC [9] and GISTIC [11] and found that the proposed methods outperformed both STAC and GISTIC in identifying low-frequency but high-confidence CNV regions.

Methods

Data Structure

We assume that the raw intensity data have been processed by a CNV detection algorithm. Denote by $R_i = \{R_{i1}, R_{i2}..., R_i = \{R_{i1}, R_{i2}, \ldots, R_{i\ell_i}\}\}$ the collection of CNV regions detected in individual i, for $i = 1,...,n$. A region is defined by its start and end probe locations, and its CNV type (duplication or deletion). For each region, we assume we have a confidence score statistic that measures the likelihood that the detected region is real. An example of this statistic is the log Bayes Factor (see [6]). For region j detected in individual i, we denote this statistic as C_{ij}.

Cumulative Overlap Using Very Reliable Regions (COVER)

Our confidence in a CNV region depends on the within- and between-subject information; our methods shall utilize both information. The within-subject information comes from the strength of the signal within an individual CNV region, and this is measured by the confidence score. The between-subject information comes from the consistency of the CNVs across different individuals. Intuitively, we have less confidence in a CNV that occurs in one individual than one that occurs in many individuals. However, a single occurrence of CNV might still be a true discovery if it is associated with a high confidence score, i.e., it is based on a strong signal.

Since individual CNV regions span different probes, the number of individual regions that overlap each probe varies. However, common CNV regions tend to occur at almost the same genomic locations across multiple individuals. Hence, we expect the common regions to be identified by consecutive probes where a 'significant' number of individuals have an overlapping CNV region. Furthermore, we also expect the confidence score of the individual region to be relatively high.

Let Z_{ijk} be the indicator that region j detected in individual i overlaps with probe k. For each probe k, we calculate the Cumulative Overlap using Very Reliable Regions (COVER) statistic y_k, defined as

$$y_k = \sum_{i=1}^{n} \sum_{j=1}^{\ell_i} (Z_{ijk} \times I_{C_{ij} \geq c}),$$

where $I_{C_{ij}>c}$ is the indicator function for regions detected with a confidence score above a certain threshold c. The common CNV regions are then defined by

$$\Re = \left\{ \left[l_m, l_{m'} \right], y_k \geq u, \forall_k \in \left[m, m' \right] \right\},$$

representing sets of consecutive probes for which y_k is consistently greater than or equal to a specified

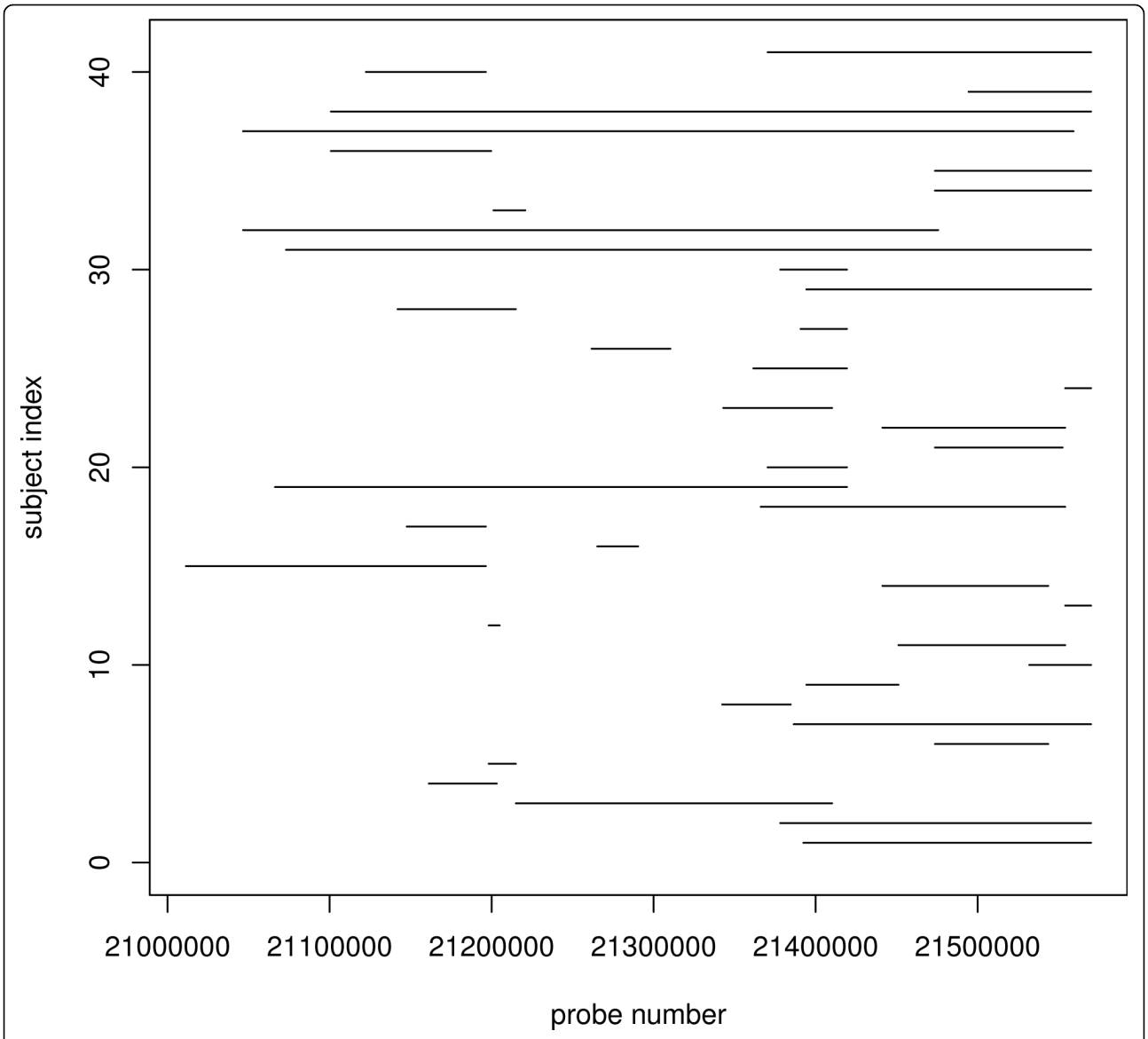


Figure 1 An example of a common CNV region found based on COVER method with threshold $u = 2$ and $c = 60$. This figure illustrates a common CNV region in part of chromosome 22, found using the COVER method with threshold $u = 2$ and confidence cutoff at 60th percentile. 41 out of 112 individuals have CNVs that overlap with this common region, indicated by the horizontal lines. We can see that despite being identified as a common region, the individual regions still portray a mixture phenomenon of several distinct subregions.

threshold u. l_m is the genomic position of probe m and it is implicitly understood that the cardinal position of the probe reflects its relative position in the chromosome so that when there are M probes in a chromosome, $l_1 < l_2 < ... < l_M$.

Using COVER, we can identify multiple common CNV regions within a chromosome. Furthermore, different subsets of individuals may contribute to different common regions, hence allowing COVER to identify regions that are common to only a subset of individuals. By only considering individual regions that are detected with high reliability, we also incorporate the uncertainty associated with each individual region in the identification of common regions. If this is not taken into account, then all regions would be treated equally despite the fact that some are more likely to be true than the others. Figure S4 in the [Additional File 1] gives an illustration of how COVER works.

Cumulative Composite Confidence Scores (COMPOSITE)

In COVER, regions with low confidence are given zero weights and they do not contribute to the COVER statistic. The within-subject confidence is not fully exploited when computing the COVER statistic: regions

that are detected with low confidence but nonetheless detected consistently across a large number of subjects might be missed.

This limitation is addressed in the second method. For probe k the composite confidence score (COMPOSITE) statistic is defined as,

$$s_k = \sum_{i=1}^{n} \sum_{j=1}^{\ell_i} (Z_{ijk} \times C_{ij}).$$

This formula is in fact similar to COVER statistic, where instead of using the indicator function $I_{C_{ij}>c}$ as weights, now all detected individual regions contribute to the COMPOSITE statistic, with the amount of their contribution proportional to their confidence scores.

Using COMPOSITE, the common CNV regions are then defined as

$$\mathfrak{R} = \left\{ \left[l_m, l'_m \right], s_k \geq v, \forall_k \in \left[m, m' \right] \right\},$$

representing sets of consecutive probes for which s_k is consistently greater than or equal to a specified threshold v. Figure S4 in [Additional file 1] gives an illustration of how COMPOSITE works.

Clustering of Individual CNV Regions within a Common Region (CLUSTER)

Cluster analysis has been used in the analysis of gene expression and aCGH data (see for example, [12-14]). Here, the motivation for CLUSTER stems from the observation that within a common CNV region identified by COVER or COMPOSITE, a complex mixture phenomenon can still be observed (see Figure 1).

Figure 2 depicts the hypothetical situation where a common region of length L bases has been identified by COVER or COMPOSITE. Four individual regions overlap with the common region and from the figure, it is clear that the first two regions are clustered to the left while the last two are clustered to the right. The two groups may form two distinct subregions and these

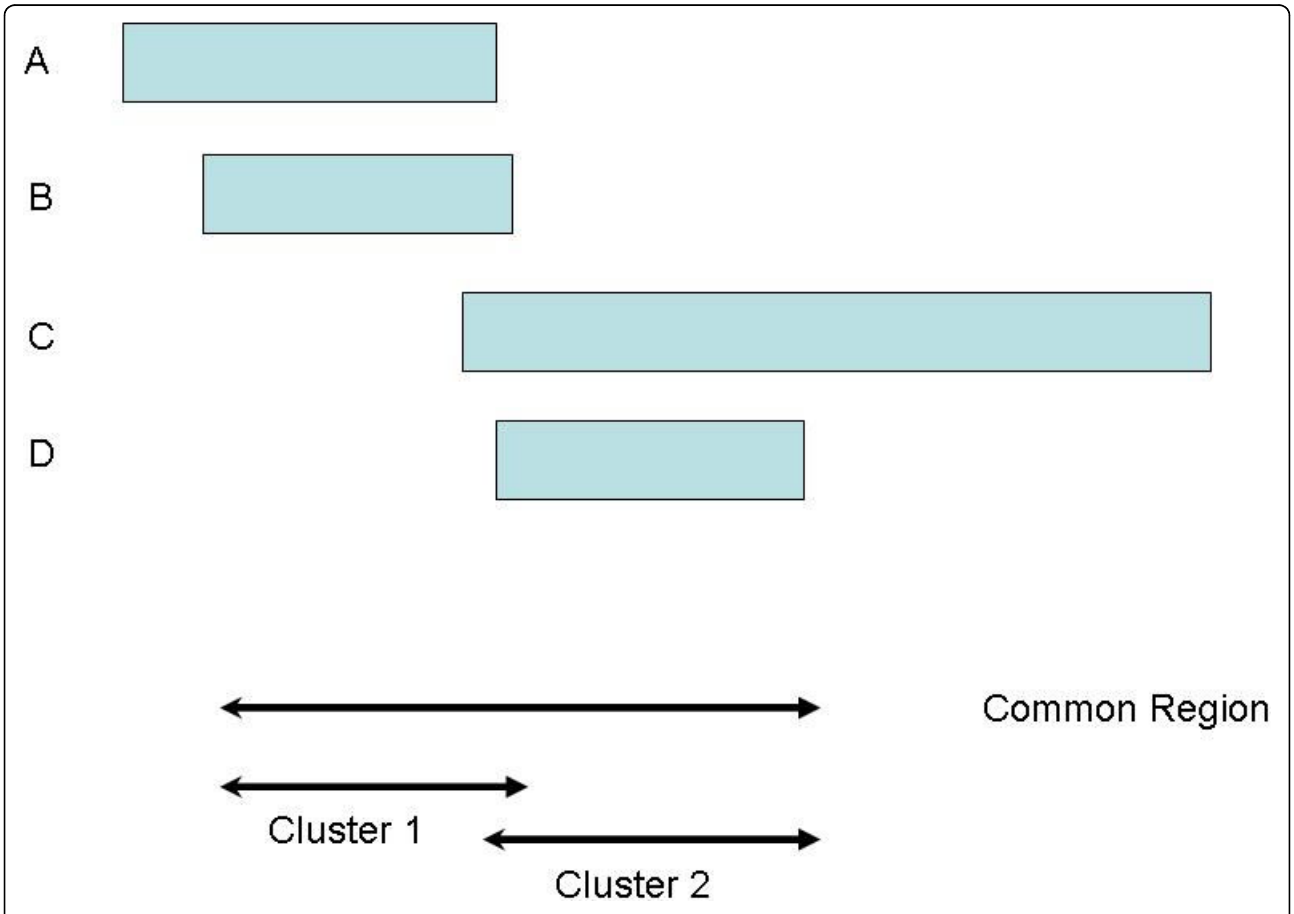


Figure 2 Hypothetical example of a identified common CNV region with 2 distinct clusters. Hypothetical situation where an identified common CNV region is common to four individuals. From the figure, it is clear that the common region consists of two partially overlapping regions. The first two individual regions are clustered together to the left of the common region, while the last two individual regions are clustered to the right.

subregions could differ biologically. In reality, the situation is more complex than the hypothetical example here (see for example Figure 1).

To find the subregions inside this common region, we first perform pairwise comparisons of the individual regions that overlap with the common region. For example, the comparison of two regions A and B can be summarized into 4 values (a, b, c, d), where a is the number of bases for which both A and B overlap with the common region, b is the number of bases where A overlaps with the common region but B does not, c is the number of bases where B overlaps with the common region but A does not, and $d = L - a - b - c$.

The (dis)similarity index can be computed using a variety of distance metrics appropriate for binary data such as the Manhattan, Canberra or Jaccard distance [15]. The Jaccard distance is particularly attractive for our case; it is defined by $a/(a + b + c)$ and can be interpreted as the percentage of common overlap of the two regions relative to the union of the overlaps of the two regions with the common region. We then construct a dissimilarity matrix as input to a hierarchical clustering algorithm. The number of clusters will be determined by the amount of within-cluster similarity we require. The boundaries of each subregion will be the minimum and maximum positions of all individual regions that belong in that cluster. If these bounds overshoot the boundaries of the initially identified region, then the boundaries will be reset to the boundaries of the initial region.

Results and Discussion

Assessment and Comparison

Datasets

We studied the performance of the proposed procedures by varying the corresponding threshold parameters in each approach. 112 HapMap samples, comprising 46 Caucasian (CEU), 29 Beijing Chinese and Tokyo Japanese (CHBJPT) and 37 Yoruban (YRI) individuals were used in the analysis. These samples are part of the Illumina iControl Database. Each sample was genotyped using the Illumina 1M chip, and PennCNV [6] was used to detect the individual CNV regions.

Comparison with Sequenced Regions

We compared the common regions we identified to a list of reference CNVs identified in eight HapMap samples using sequencing data [10]. For each of the eight samples, we calculated the discordance rates by recording the proportion of common CNV regions (found using our methods) for that sample that were not concordant with the sample-specific reference CNVs. To be 'concordant' with a reference CNV, a region has to be either contained within the reference CNV or it has to overlap with at least 50% of the reference region. It is important to note however that it is difficult to get a gold standard for common CNV boundaries; even the sequencing-based CNV regions cannot be expected to have 100% sensitivity and specificity in genotype calling and certainly not in boundary calls for common CNVs.

Comparison with other Array-based Regions

We compared the regions found using our methods to the regions found by two other groups using array-based methods. We compared with McCarroll *et al*. [16], where the regions were identified using the Affymetrix SNP 6.0 arrays on 270 HapMap samples. To minimize false discoveries, they ran two independent experiments and require a CNV to be observed in both experiments. We also compared our regions to the regions found by Conrad *et al*. [17]. These regions were identified using tiling oligonucleotide microarrays, comprising of 42 million probes, on 41 HapMap samples. A total of 11,700 CNVs were identified, and 8,599 were validated using a set of stringent criteria including (i) additional measurements by Agilent 105K CGH arrays, (ii) overlap with previous studies and (iii) other quality-control filters. For our comparisons, we used only the 8,343 validated CNVs in the autosomal regions.

Comparison to other approaches

We compared our approaches to previous common CNV detection methods, STAC: Significance Testing for Aberrant Copy number [9] and GISTIC: Genomic Identification of Significant Targets in Cancer [11].

Briefly, STAC takes segmented data as input and estimates two statistics: 1. A frequency statistic, which estimates the frequency of aberration at each location across all individuals. 2. A footprint statistic, which uses a subset search methodology and counts the number of locations c such that c is contained in a set of intervals (see [9] for more details). It then uses a permutation test to assess the significance of the observed region. STAC requires each chromosome to be split into non-overlapping regions of a user-defined fixed size. The algorithm looks for evidence of common CNV regions within each window, and reports the associated frequency and footprint p-values.

GISTIC first calculates a 'G score' that is associated with both the frequency of occurrence as well as the amplitude of the aberration. Then, it calculates the probability (q-value) of the observed region occurring by chance via a permutation test. One can either input the log intensity ratios, where the GLAD algorithm [18] will be used to segment the data, or input pre-segmented data using other algorithms.

We had also planned to make comparison to another method called MSA [19], but failed because the software, which is part of the GenePattern module, did not work properly. MSA can be viewed as an improvement over STAC, where it extends the notions of frequency

and footprint statistics using original intensity ratio data instead of segmented data [8]. We also tried a comparison to RJaCGH [2], which uses a non-homogenous Hidden Markov Model fitted via the Reversible-Jump Markov Chain Monte Carlo method to estimate the probability that a region has copy number alterations; the method also allows the identification of minimal common regions of copy number changes among multiple individuals.

Unfortunately, with our samples, the algorithm did not converge, so we could not proceed with the comparison.

Testing Hardy-Weinberg Equilibrium

It has been observed that the majority of common CNV regions are inherited [20]. Hence, for a population of normal (healthy) individuals, we expect, for most of the common regions, the integer copy numbers to be in Hardy-Weinberg equilibrium (HWE). The small number of regions that depart from HWE can be attributed to factors such as recent mutations. For example, McCarroll *et al*. [16] found that about 98% of common diallelic CNV regions do not show significant departure from HWE. In principle, HWE applies to both diallelic CNVs (where only loss or gain of copy numbers are present in addition to normal copies) and multi-allelic CNV regions (where both loss and gain of copies are present).

For diallelic CNVs with only loss and normal-copy numbers (copy-number = 0,1,2), the HWE test can be conducted by treating '0' copies as minor allele homozygous, '1' copy as heterozygous and '2' copies as reference homozygous. Similarly, for CNVs with only gain and normal-copy numbers (copy-number = 2,3,4), we treat '2' copies as reference homozygous, '3' copies as heterozygous and '4' copies as minor-allele homozygous. For multi-allelic CNVs, a model with three or more alleles is needed. However, the HWE test cannot be performed directly on the unphased copy-number because there is an issue with different combinations of alleles producing the same copy-number. For example, in a 3-allele model, a copy-number of 2 can be produced by a combination of '0' and '2' copies or two '1' copy alleles.

When dealing with samples from healthy individuals, we propose to use the outcome of the HWE tests to select 'optimal' parameter thresholds (e.g., c in COVER and v in COMPOSITE). If we observe a large number of common CNV regions with significant departure from HWE (after accounting for population stratification), it could mean that the parameters we choose are not optimal. When dealing with a mixture of healthy and diseased individuals such as in association studies, it is expected that the CNVs among the diseased individuals will show some degree of departure from HWE as some of the CNVs could be due to recent abberations. We propose performing HWE tests only among the healthy individuals to select the optimal threshold parameters.

Results

COVER results

Figure 3 shows the results for COVER. The discordance rates with Kidd *et al*.'s [10] reference CNVs (see Comparison with Sequencing Results) can be as high as 80% when we include all CNV calls in identifying the common regions. The discordance rates decrease when we exclude CNVs whose confidence scores are below a certain percentile; more severe filtering generally reduces the discordance rates. The lowest discordance rates of about 55% were achieved when we excluded individual regions whose confidence scores were below the 80th percentile. Surprisingly, increasing the required minimum number of individuals inside a region (u) does not seem to have an effect on the discordance rates.

However, the required minimum number of individuals (u) does affect the rates of HWE violation (calculated as the percentage of diallelic CNVs whose p-value from the HWE test is < 0.01 in at least one of the three ethnic groups). (Some HapMap individuals were related; the HWE test in each ethnic group was carried out on unrelated individuals only.) There is an overall increasing trend for the proportion of common CNV regions that violate HWE when we increase the minimum number of individuals (Figure 3(b)). This is partly due to the fact that with increasing number of individuals, we detect CNV regions with larger minor allele frequencies (see Figure 3(c)), hence the test for HWE will be more powerful. Generally, the rates of departure from HWE are less than 10% and can be lowered by filtering out individuals with lower quality regions. A steeper reduction in the rates of departure from HWE can be observed when only individual regions whose confidence scores are above the 60th percentile are considered (Figure 3(b)).

The sizes of the identified common regions generally increase when we filter lower quality individual regions (Figure 3(d)), reflecting the fact that smaller regions with fewer overlapping probes would tend to have lower confidence scores. By choosing confidence score thresholds (c) anywhere up to the 60th percentile, the average size of the common regions are approximately the same or slightly smaller than the average size that Kidd *et al*. [10] obtained using sequencing methods (solid horizontal line in Figure 3(d)). The dashed horizontal line in Figure 3(d) shows that the median size of CNV regions identified using the 500K EA chip [21] is much larger than what we observe using our methods.

For this dataset, setting the confidence score threshold to the 60th percentile seems to be the optimum choice. With this setting, the discordance rates are around 60%

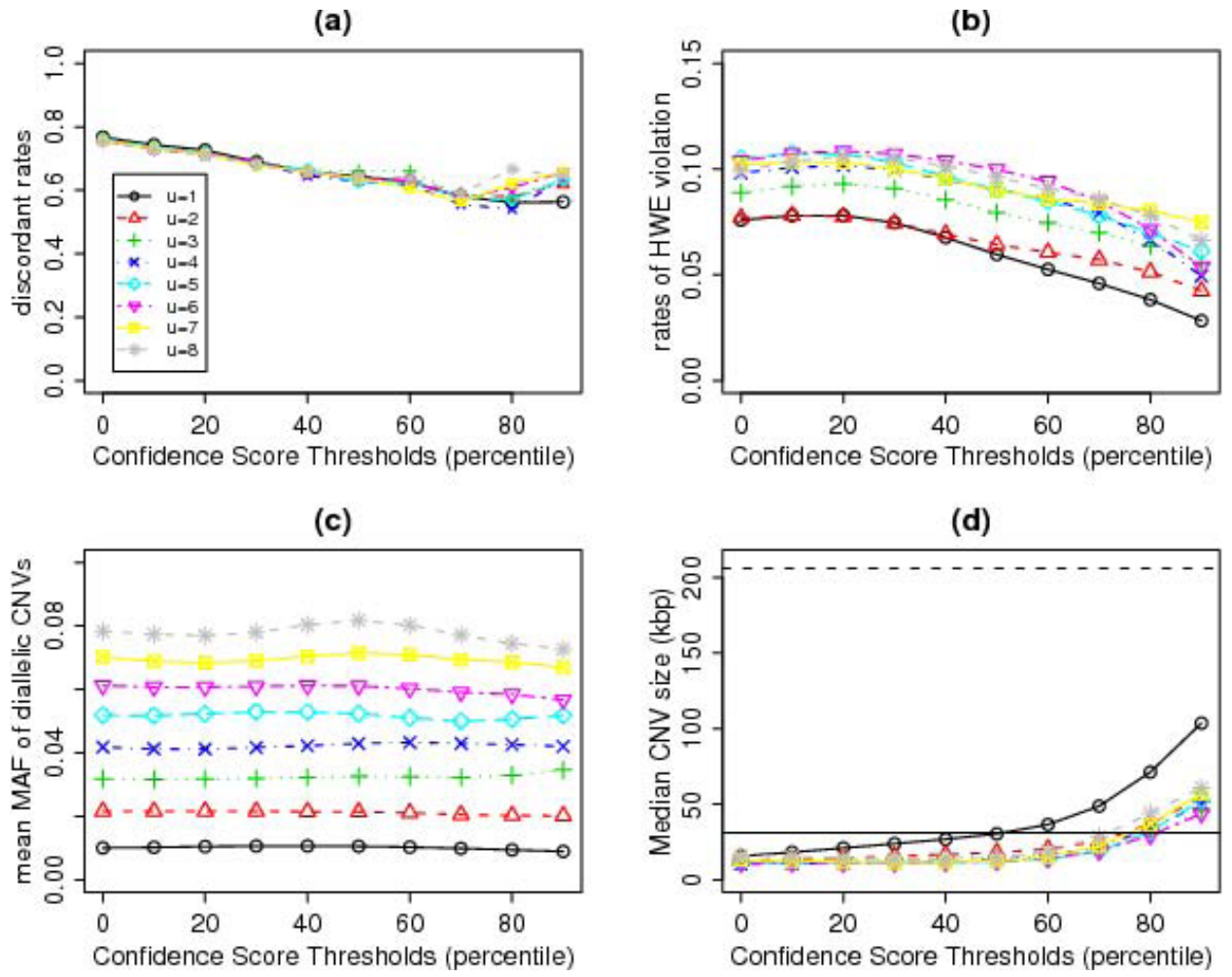


Figure 3 Results of COVER method. (a) Discordant Rates, (b) Proportion of diallelic CNVs that failed HWE, (c) mean minor allele frequency (MAF) of diallelic CNVs and (d) Mean CNVs size (kilo-bases) as a function of confidence scores cut-off points and minimum number of individuals.

and the proportion of diallelic CNVs that violate HWE is kept at around 8%. The choice of u is more subjective, as it depends on our definition of 'common' regions. For example, if we require each common region to overlap with at least three individual regions and set c to the 60th percentile, we will identify 443 common CNV regions (see [Additional file 2]).

COMPOSITE results

A total of 89% of the probes does not contain any individual CNV regions and thus their composite scores are zero. So, if we set the threshold v at the 89th percentile of the composite scores, we do not filter out any individual regions and this approach is essentially the same as using $u = 1$ and $c = 0$ in COVER.

Figures 4(a) and 4(b) show that, as we increase the threshold, the discordance rates as well as the rates of HWE violation decrease steadily. Unlike the COVER approach, where increasing the confidence score threshold does not result in lower ability to detect rarer CNVs, increasing the composite score threshold does result in fewer rare CNVs being detected (Figure 4(c)). This is because the composite score is a function of both the confidence score and the number of individuals within a common region. By increasing the threshold, we are implicitly requiring more individuals within a common region.

The increasing trend of mean minor allele frequency (MAF) is consistently seen when the threshold is increased to the 96th percentile. Beyond this, the mean MAF decreases because large regions with higher MAF may be split into several subregions with smaller MAF. This observation is consistent with the pattern of median size of CNV regions (Figure 4(d)). Generally, we are losing the smaller regions with low composite scores as we increase the threshold. However, beyond the 96th

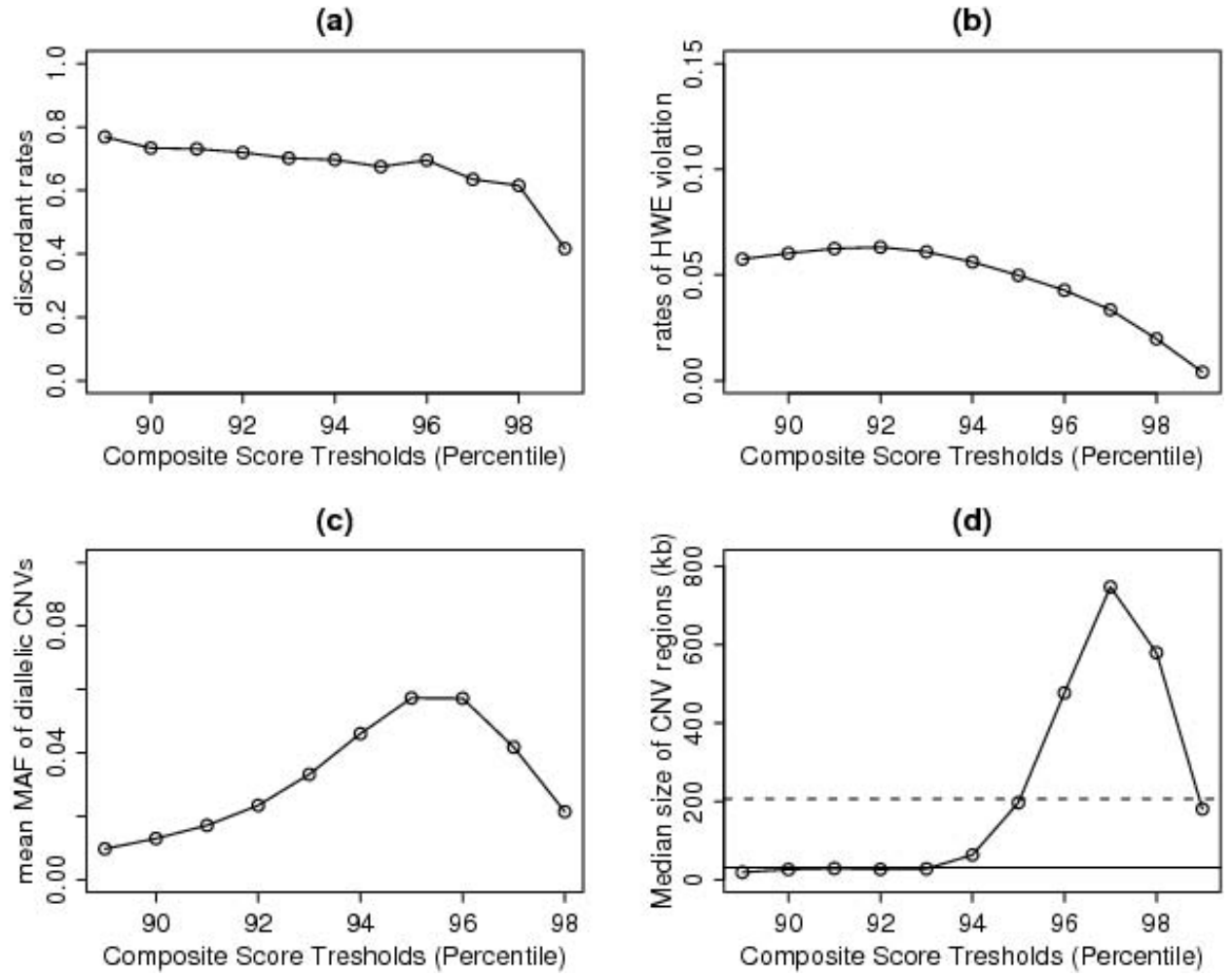


Figure 4 Results of COMPOSITE method. (a) Discordant Rates, (b) Proportion of diallelic CNVs that failed HWE, (c) mean minor allele frequency (MAF) of diallelic CNVs and (d) Median size of CNV regions (kb) as a function of composite confidence scores cut-off points. Solid line is median CNV size found by Kidd *et al*.

percentile, the median region size decreases again due to the splitting of the large regions.

The optimal setting is to set the threshold to the 94th percentile, where the proportion of regions that failed HWE is around 5% (Figure 4(c)). Using this setting, we are able to detect 491 CNV regions (see [Additional file 3]) with median CNV size slightly larger then the median size found by Kidd *et al*. [10]. The discordance rates among the eight HapMap samples are approximately 70%, higher than what can be achieved by COVER. Hence, although COMPOSITE can pick up more regions, a higher percentage of these regions is likely to be false discoveries.

CLUSTER results

The common regions identified using either COVER or COMPOSITE can be further refined into distinct subregions using CLUSTER. Here, we present the results of applying CLUSTER to the common regions identified by COVER. We choose the CLUSTER parameters so that regions will be clustered together if they are at least 60% similar. Complete linkage is used so that the distance between any pair of clusters is defined as the maximum distance between a pair of members drawn one from each cluster. Single or average linkage can also be used. Since single linkage defines the distance between any pair of clusters as the minimum between a pair of members from the clusters, it generally tends to produce clusters that are more similar to each other, and when the same similarity cut-off point is used, it tends to produce fewer clusters than complete linkage. Meanwhile, using average linkage gives more clusters than single linkage, but fewer than complete linkage. In the [Additional file 1], we compare the three linkage measures for a sample region.

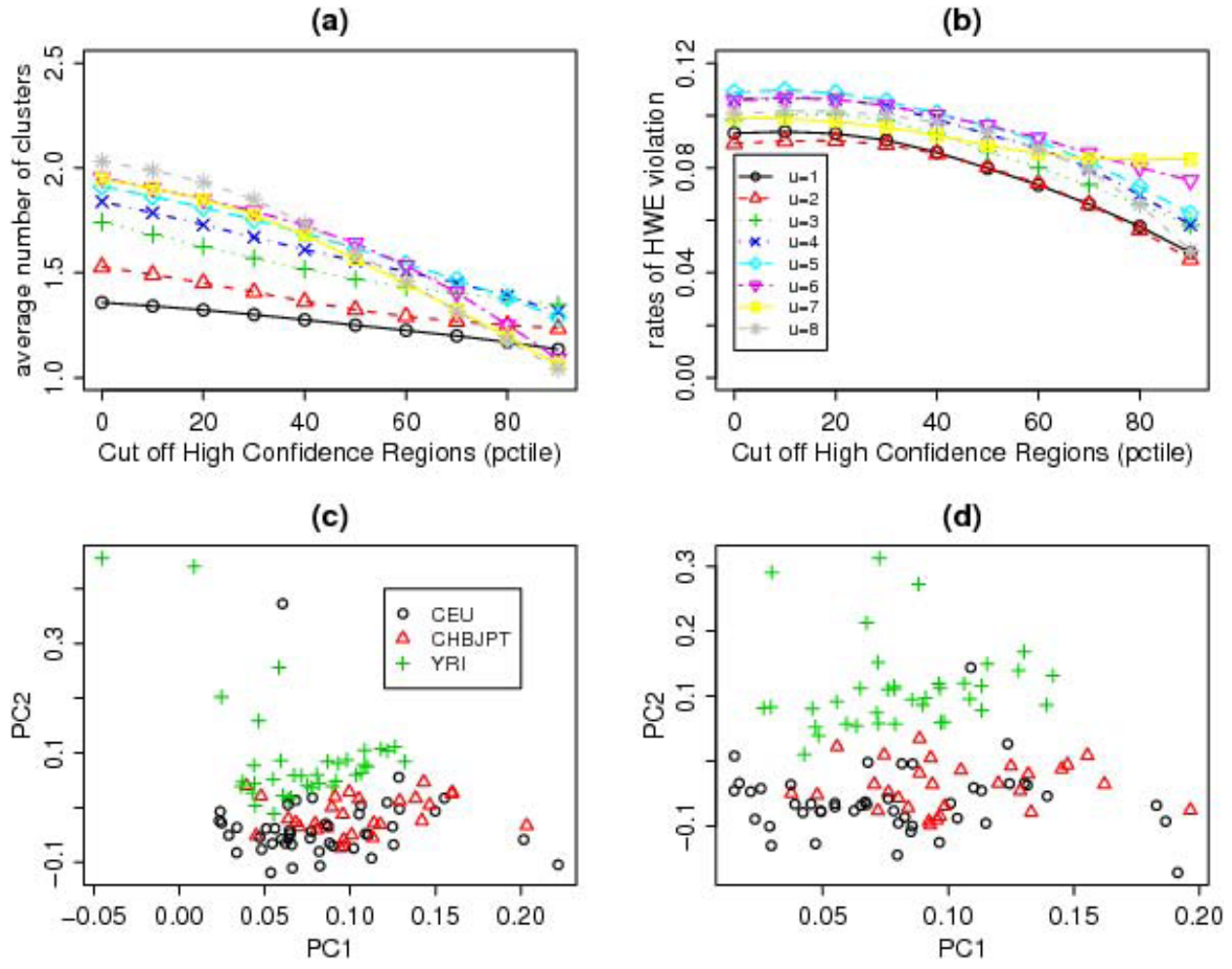


Figure 5 Results of applying CLUSTER to common regions identified by COVER method. (a) Average number of clusters, (b) rates of departure from HWE, (c) First and second components of PCA based on subjects' integer copy-number calls at common regions found using COVER (with $u = 3$ and $c = 60$), (d) First and second components of PCA based on subjects' integer copy-number calls at common regions found using complete-linkage CLUSTER (`with cluster.limit = 0.6`).

Figure 5(a) shows that the number of clusters decreases when we increase the confidence score threshold. But even when we consider CNVs with confidence scores above the median, the clustering effect is still evident with 1.3 to 1.7 clusters found for each common region, depending on which threshold value u is used. For the optimum parameters $u = 3$ and $c = 60$, on average, 1.5 clusters are found per common region. The rates of departure from HWE (Figure 5(b)) are approximately the same as in Figure 3(b) and increasing the confidence-score threshold lowers the rates.

Once the common regions are identified, it is straightforward to perform a number of downstream analyses. For example, a principal component analysis (PCA) can done based on subjects' integer copy-number calls at these regions (see Section 'Principal Component Analysis of CNV Profiles' for more details). In the HapMap dataset, CLUSTER clearly improves the separation between the Yoruba and the other two populations based on the subjects' common CNV region profiles (compare Figure 5(c) vs 5(d)). This result suggests that different ethnic groups have more subtle differences in the breakpoints of CNV regions.

Comparisons

McCarroll et al.'s versus Kidd et al.'s Results Using the Affymetrix 6.0 arrays, McCarroll *et al.* [16] employed a set of strict criteria based on duplicate experiments to identify the CNV regions. For each of the eight samples sequenced by Kidd *et al.* [10], we calculated the discordance rates with McCarroll *et al.*'s CNVs and they range from 71% for sample NA12878 to 84% for sample NA18517. On average, across the eight samples, 76% of the regions found by McCarroll *et al.* are discordant with the regions found by Kidd *et al.*

[10]. In comparison, using COVER, the discordance rates are around 60% (see Section "COVER Results"). Thus, the methods described in this paper, using only data from a non-duplicated experiment, actually perform better in terms of discordance rates against sequencing data.

McCarroll et al.'s versus COVER/COMPOSITE Results We also compared the regions identified by our approaches to the list of common CNV regions identified by McCarroll *et al*. [16]. Figure 6(a) shows that by using COVER, the discordance rates can be lowered by either increasing the confidence-score threshold, placing a higher limit on the minimum number of individuals (u), or both. For the best scenario, the discordance rate is about 15%. Using COMPOSITE, the discordance rates can be reduced by increasing the composite-score threshold, but even for the best scenario, the discordance rate is around 25% (see Figure 6(b)).

Comparison to Conrad et al.'s regions Treating the set of 8,343 validated autosomal CNVs found by Conrad et al. [17] as reference CNVs, we calculate the discordance rates against this reference list. Using the optimal parameters for COVER/COMPOSITE for this dataset, we obtain discordance rates of 42% and 31% for COVER and COMPOSITE respectively. By refining the regions using CLUSTER, the discordance rate for COVER decreases to 34% and that for COMPOSITE remains about the same, at 33%. These are better than McCarroll *et al*.'s [16] regions, which have a discordance rate of 44%.

Comparison to GISTIC As input to GISTIC, we used CNV calls from PennCNV for the same Hapmap samples as described in the Datasets Section. Using the default parameters of GISTIC, with the q-value threshold set at 0.25, we obtained 342 significant common regions with a mean frequency of 0.106 and a median confidence score of 15.7. For comparison with COVER and COMPOSITE, we chose threshold parameters to give the closest number of common regions to that detected by GISTIC. For COVER, this corresponded to the choice of u = 3 and c = 70th percentile, which yielded 329 regions with a mean frequency of 0.065 and median confidence of 32.3. For COMPOSITE, the threshold was chosen to be the 94.5th percentile, and this yielded 360 regions with a mean frequency of 0.121 and median confidence of 27.6.

For each region identified by COVER, we checked if it was concordant with any region identified by GISTIC. Concordance is defined in the same way as in the Section 'Comparison with Sequencing Results'. The COVER-identified regions can hence be divided into two groups: those that are concordant with at least one GISTIC region and those that are not. For each group, we computed the mean frequency and median confidence score, as well as the discordance rates with Kidd et al.'s regions. We did the same for each region identified by GISTIC, checking if the region was concordant with any region identified by COVER. Similar analysis was done comparing COMPOSITE and GISTIC.

Table 1, for COVER, shows that regions that are concordant with GISTIC regions have higher frequencies but moderate confidence scores, while those that are not concordant with GISTIC regions have lower frequencies but higher confidence scores. The concordant

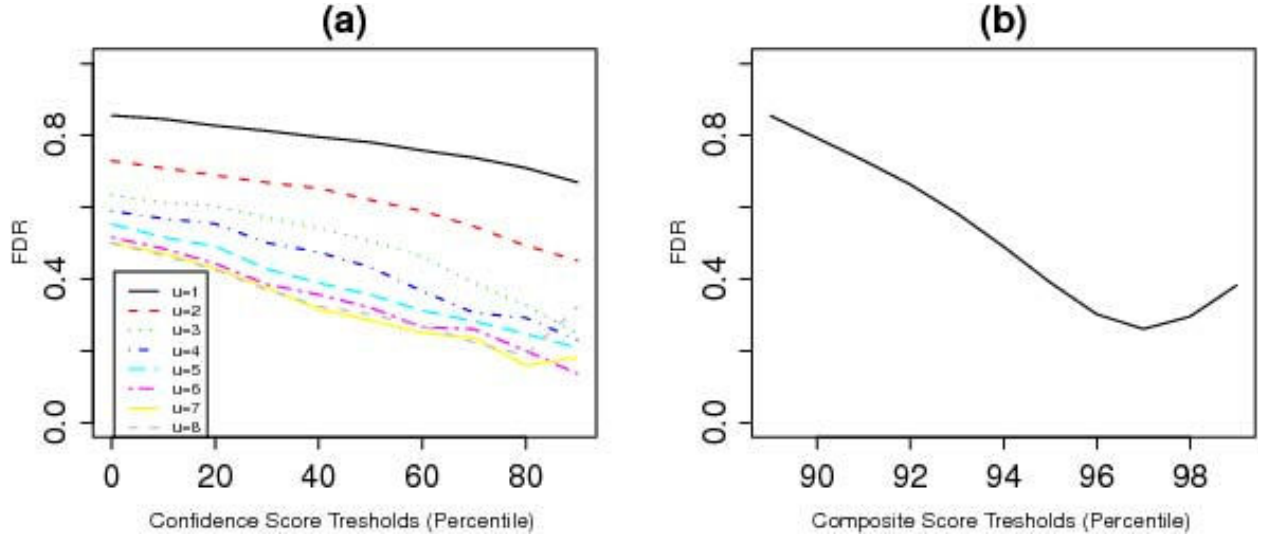


Figure 6 Comparison to McCarroll's CNVs. (a) Discordance rates when comparing regions found using COVER and those found by McCarroll *et al*., plotted against confidence score thresholds for different values of u. (b) Discordance rates when comparing regions found using COMPOSITE and those found by McCarroll *et al*., plotted against composite score thresholds.

Table 1 Comparison with GISTIC.

regions found by	overlap?	no. of regions	mean freq	median conf	discordance**
COVER	✓ GISTIC	139	0.10	30	62%
	✗ GISTIC	190	0.037	37.5	87%
COMPOSITE	✓ GISTIC	162	0.21	20.0	64%
	✗ GISTIC	198	0.048	72.8	75%
GISTIC	✓ COVER	153	0.15	22.3	56%
	✗ COVER	189	0.072	8.8	84%
	✓ COMPOSITE	173	0.15	20.6	61%
	✗ COMPOSITE	169	0.058	8.8	82%

✓ - overlap

✗ - no overlap

** discordance rates with Kidd's sequencing results.

This table shows a summary of the results obtained from comparing COVER/COMPOSITE to GISTIC.

regions have lower discordance rates with sequenced-based results. Similar patterns in frequencies, confidence scores and discordance rates are also seen for the regions found by COMPOSITE. We deduce that GISTIC misses regions that are of low frequencies but high confidence scores. Hence, it seems that COVER/COMPOSITE can identify the low-frequency CNVs better. In addition, of the regions found by GISTIC, those that are concordant with COVER or COMPOSITE have high frequencies and moderate confidence scores while those that are not concordant have low frequencies and low confidence scores. Again, the concordant regions have lower discordance rates with sequenced-based results. From this, we deduce that the regions identified by GISTIC but missed by our methods are those with low frequencies and low confidence scores, and hence more likely to be false positives.

Comparison to STAC For the purpose of analysis using STAC, we split each chromosome into 1450-1500 fixed-size windows with the size of the windows varying from 165 kb for chromosome 1 down to 24 kb for chromosome 22, resulting in a total of 32780 windows across chromosome 1-22. (We tried a smaller window size but the computational burden became too large, where even after 48 hours the algorithm was still running in a 3 GHz windows PC with 4 Gb RAM). We used 0.05 as a cut-off to declare windows with significant frequency or footprint p-values, and obtained 868 significant windows with a mean frequency of 0.155. Each significant fixed-size window will be taken as a significant region.

To compare the regions found by STAC to the regions found using COVER and COMPOSITE, we chose threshold parameters to give a number of common regions closest to that detected by STAC. For COVER, this corresponded to the choice of u = 2 and c = 60th percentile, and for COMPOSITE, the 93th percentile. We obtained 777 and 805 common regions respectively. We performed similar analysis as in the comparison to GISTIC.

A summary of this comparison is shown in Table 2a. We observe similar results as in the comparison to GISTIC: regions that were identified by STAC but that were missed by COVER/COMPOSITE have low frequencies and low confidence scores, but regions identified by COVER/COMPOSITE that were missed by STAC have low frequencies but high confidence scores, and were thus more likely to be true positives.

We also investigated if the relative performance of STAC would improve if we manually filtered out individual regions with lower confidence scores. We decided to use only individual regions whose confidence scores were above the median confidence score of all reported regions. Using this filtered input, STAC identified 654 significant windows. Using u = 2 and c = 70th percentile for COVER and the 93.5th percentile for COMPOSITE, we identified a similar number of common regions (615 for COVER and 610 for COMPOSITE). Table 2b summarizes the results of this comparison and our conclusions are similar to those with the unfiltered input data.

We conclude that COVER and COMPOSITE are able to detect the majority of the regions found by STAC, and in addition they also detect common high-confidence CNV regions that occur in a smaller number of subjects that were missed by STAC.

Implementation

The methods are implemented in an R package `cnvpack`. The main input is a list of detected individual CNV regions with the following information: Sample name, chromosome number, detected integer copy number, start and end genomic locations and a confidence score. The package can be downloaded from http://www.meb.ki.se/~yudpaw.

Table 2 Comparison with STAC.

STAC input: all data regions found by	overlap?	no. of regions	mean(freq)	median(conf)
COVER	✓ STAC	301	0.084	25.6
	✗ STAC	476	0.021	31.2
COMPOSITE	✓ STAC	372	0.14	18.6
	✗ STAC	433	0.023	52.5
STAC	✓ COVER	609	0.15	23
	✗ COVER	259	0.11	8.1
	✓ COMPOSITE	727	0.15	20.5
	✗ COMPOSITE	141	0.07	7.21
STAC input: filtered data regions found by	overlap?	no. of regions	mean(freq)	median(conf)
COVER	✓ STAC	294	0.068	30.2
	✗ STAC	321	0.020	37.6
COMPOSITE	✓ STAC	297	0.14	23.1
	✗ STAC	313	0.045	65.2
STAC	✓ COVER	585	0.14	28.1
	✗ COVER	69	0.07	16.1
	✓ COMPOSITE	595	0.14	26.8
	✗ COMPOSITE	59	0.06	20.2

✓ - overlap

✗ - no overlap

This table shows a summary of the results obtained from comparing COVER/COMPOSITE to GISTIC.

Downstream analyses

CNV-association analysis

One important use of the identified common CNV regions is for group comparisons in association studies. For each region we test whether certain CNVs are over-represented in one group compared to the others. Typically, the Fisher's exact test or chi-squared test for contingency tables can be used. The test can be carried out for all identified common CNV regions and the issue of multiple testing can be dealt with using the false discovery rate (FDR) assessment. (See [Additional file 1] on how to use the package for such analyses.)

As an illustration we performed an association analysis on the common regions identified in the 112 control subjects using the optimal parameters for COVER and COMPOSITE. The subjects were grouped by ethnicity (YRI, CHBJPT and CEU). Both methods showed that there were a number of highly-significant CNV regions with p-value < 1e-06. Two of these regions were detected by both methods. The first one is a 16.2 kb deletion in chromosome 2 (genomic positions 203,004,035 to 203,020,242). This region occurs exclusively in the Yoruba population (17/37) and overlaps with the BMPR2 gene that has been linked to primary pulmonary hypertension [22]. The second region is a 4.6 kb deletion in chromosome 4 (genomic positions 20,982,707 to 20,987,259) that occurs among Yoruban (19/37) and CHBJPT (4/29). This region overlaps with the KCNIP4 gene that is known to interact with presenilin, a protein that has been reported to be involved in early-onset Alzheimer's disease [23].

Principal component analysis of CNV profiles

We also perform principal component analyses (PCA) to obtain informative plots of population differentiation in the CNV profiles (see [Additional file 1] for more information). For the HapMap samples, the first two components obtained using the optimal COVER parameters separate the Yoruba population (YRI) from the Caucasian(CEU) and Asian(CHBJPT) populations, but the other two populations are not very well separated (Figure S1 in the [Additional file 1]). A better separation between the CEU and CHBJPT populations is achieved using the third and fourth components(see Figure 7(a)) and the separation is further improved when we use CLUSTER to refine the CNV regions identified by COVER (Figure 7(b)).

Conclusions

We have described and compared two different methods for identifying common CNV regions. Using 112 HapMap samples, we have shown that these methods produce common CNV regions that mostly follow Hardy-Weinberg equilibrium (HWE). For the eight HapMap samples where we compared the regions we identified to the reference CNV regions found by sequencing [10], the discordance rates can be as high as 80%, but this can be reduced to 60% by considering CNVs with higher confidence scores, thus showing the importance of

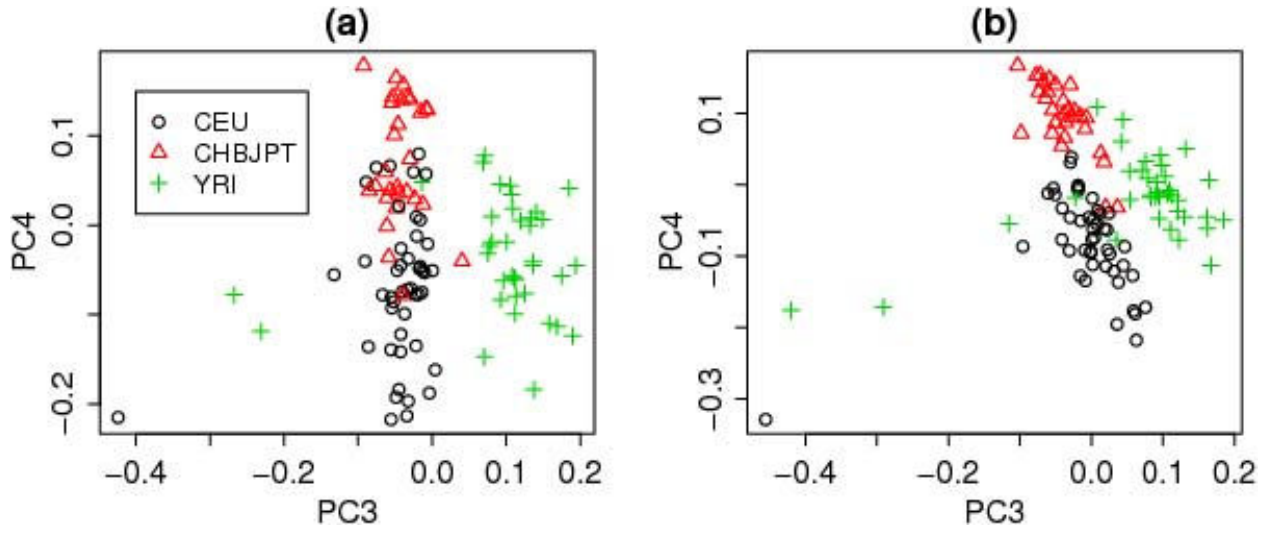


Figure 7 The third and fourth principal components. (a) Using COVER (with u = 3 and c = 60). (b) the same as (a) but using the output of complete-linkage CLUSTER (`with cluster.limit = 0.6`).

further processing of the CNVs. The high level of discordance itself might be due to an inherent limitation in the SNP array as the platform for CNV detection, but perhaps also due to imperfection in the sequencing-based results. Further works are needed to explain the discordance level.

When we compared our methods to previously published methods, STAC and GISTIC, we found that our methods are better at identifying low-frequency CNVs. Moreover, STAC is rather rigid and insensitive to the actual breakpoints of a CNV region, because if two consecutive windows are reported as significant, we do not know if there is one large CNV which spans both windows, or two separate and distinct CNVs. Although we can decrease the window size to increase the resolution, in practice, decreasing the window size beyond a certain point will incur too much computational burden. Another limitation of previous methods is the lack of consideration of individual-specific confidence scores. This means that all samples contribute equally to the calculation of the statistic used to identify the common regions, while in fact, there is bound to be inter-sample variability, where some CNVs are more likely to be true positives than others.

The results of COVER and COMPOSITE are similar in terms of discordance rates and HWE violation rates, but COMPOSITE appears to be better at identifying rare regions. The HWE violation rates are useful in determining the choice of parameter values for COVER and COMPOSITE. For this particular data set, we observed a steeper reduction in HWE violation rates when we used COVER with a confidence score threshold set above the median or higher. For COMPOSITE, a more noticeable reduction in HWE violation rates was observed when we set v to the 94th percentile. For a new dataset, we encourage users to choose the confidence score and COMPOSITE score parameter thresholds for which steeper reduction in HWE violation rates can be observed.

When using COVER, the minimum number of individuals inside a common region (u) needs to be specified as well. If we are interested in rare variants in addition to the common variants, then it makes sense to set u = 1. Otherwise, $u \geq 2$ should be used. A higher u will result in the identification of fewer, but more highly-recurrent CNV regions. In our experience with the HapMap samples, clustering results produce better separation of the ethnic groups than indicated by the initial common CNV regions. In comparison with the highly-validated CNVs from Conrad et al. [17], the concordance rate of COVER also improves after refinement with CLUSTER. So, in summary, we recommend users to further refine the identified common CNV regions using CLUSTER.

Additional file 1: The supplementary report documents details on how to use the R package cnvpack for the various analyses described in this paper.

Additional file 2: This table shows details of the regions found by COVER.

Additional file 3: This table shows details of the regions found by COMPOSITE.

Acknowledgements
This work was supported by a grant from the Swedish Research Council and National University of Singapore (NUS) Start-up Grant No. R-186-000-103-133.

Author details
[1]Department of Epidemiology and Public Health, National University of Singapore, 16 Medical Drive, 117597, Singapore. [2]Centre for Molecular Epidemiology, National University of Singapore, 30 Biopolis Street, 138671, Singapore. [3]Department of Biomedical Sciences and Biotechnology, University of Brescia, Viale Europa, 11 25123 Brescia, Italy. [4]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, Stockholm 17177, Sweden. [5]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, 28 Medical Drive, 117456, Singapore.

Authors' contributions
TSM and AS contributed equally to this work; TSM, AS, SC and YP conceived the study, performed data analysis and wrote the manuscript. KCS and CKS conceived the study. All authors read and approved the final manuscript.

Received: 15 October 2009 Accepted: 22 March 2010
Published: 22 March 2010

References

1. Olshen AB, Venkatraman ES, Lucito R, Wigler M: **Circular binary segmentation for the analysis of array-based DNA copy number data.** *Biostatistics* 2004, **5**:557-572.
2. Rueda OM, Diaz-Uriarte R: **Flexible and accurate detection of genomic copy-number changes from aCGH.** *PLoS Computational Biology* 2007, **3(6)**: e122.
3. Erdman C, Emerson JW: **A fast Bayesian change point analysis for the segmentation of microarray data.** *Bioinformatics* 2008, **24**:2143-2148.
4. Pique-Regi R, *et al*: **Sparse representation and Bayesian detection of genome copy number alterations from microarray data.** *Bioinformatics* 2008, **24**:309-3182.
5. Pique-Regi R, *et al*: **Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA.** *Bioinformatics* 2009, **25(10)**:1223-1230.
6. 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* 2007, **17**:1665-167.
7. Colella S, *et al*: **QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data.** *Nucleic Acids Research* 2007, **35**:2013-2025.
8. Rueda OM, Diaz-Uriarte R: **Finding Recurrent Regions of Copy Number Variation: A Review.** *Collection of Biostatistics Research Archive* 2008, Art42.
9. Diskin SJ, *et al*: **STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments.** *Genome Research* 2006, **16**:1149-1158.
10. Kidd JM, *et al*: **Mapping and sequencing of structural variation from eight human genomes.** *Nature* 2008, **453**:56-64.
11. Beroukhim R, *et al*: **Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma.** *PNAS* 2007, **104(50)**:20007-20012.
12. Van Wieringen WN, Wiel Van De MA, Ylstra B: **Weighted clustering of called array CGH data.** *Biostatistics* 2008, **9(3)**:484-500.
13. Eisen MB, *et al*: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:14863-14868.
14. Jong K, *et al*: **Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors.** *Oncogene* 2007, **26**:1499-1506.
15. Everitt BS, *et al*: *Cluster Analysis* Arnold, 4 2001.
16. McCarroll SA, *et al*: **Integrated detection and population-genetic analysis of SNPs and copy number variation.** *Nature Genetics* 2008, **40**:1166-1174.
17. Conrad DF, *et al*: **Origins and functional impact of copy number variation in the human genome.** *Nature* 2009.
18. Hupe P, *et al*: **Analysis of array CGH data: from signal ratio to gain and loss of DNA regions.** *Bioinformatics* 2004, **20(18)**:3413-3422.
19. Guttman M, *et al*: **Assessing the significance of conserved genomic abberations using high resolution genomic microarrays.** *PLoS Genetics* 2007, **3(8)**:e143.
20. Locke DP, *et al*: **Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome.** *American Journal of Human Genetics* 2006, **79**:275-290.
21. Redon R, *et al*: **Global variation in copy number in the human genome.** *Nature* 2006, **444**:444-454.
22. Lane KB, Consortium TIP, *et al*: **Heterozygous germline mutations in BMPR2, encoding a TGF-beta receptor, cause familial primary pulmonary hypertension.** *Nature Genetics* 2000, **26**:81-84.
23. Hutton M, Hardy J: **The presenilins and Alzheimer's disease.** *Human Molecular Genetics* 1997, **6**:1639-1646.

doi:10.1186/1471-2105-11-147
Cite this article as: Mei *et al.*: Identification of recurrent regions of copy-number variants across multiple individuals. *BMC Bioinformatics* 2010 **11**:147.