Copy Number Variants in the human genome and their association with quantitative traits

By

WANTING CHEN (BSc, MSc)

College of Medicine and Veterinary Medicine

A thesis presented for the degree of Doctor of Philosophy at the University of Edinburgh

2011

To my family

Declaration

I hereby declare that this thesis has been composed by myself and that it has not been accepted in any previous application for a Doctor of Philosophy degree. This work, of which is a record, except where specifically acknowledged, has been carried out entirely by myself. All sources of information have been specifically acknowledged by means of references.

Wanting Chen

………………………..

Abstract

Copy number Variants (CNVs), which comprise deletions, insertions and inversions of genomic sequence, are a main form of genetic variation between individual genomes. CNVs are commonly present in the genomes of human and other species. However, they have not been extensively characterized as their ascertainment is challenging.

I reviewed current CNV studies and CNV discovery methods, especially the algorithms which infer CNVs from whole genome Single Nucleotide Polymorphism (SNP) arrays and compared the performance of three analytical tools in order to identify the best method of CNV identification. Then I applied this method to identify CNV events in three European population isolates—the island of Vis in Croatia, the islands of Orkney in Scotland and villages in the South Tyrol in Italy - from Illumina genome-wide array data with more than 300,000 SNPs. I analyzed and compared CNV features across these three populations, including CNV frequencies, genome distribution, gene content, segmental duplication overlap and GC content. With the pedigree information for each population, I investigated the inheritance and segregation of CNVs in families. I also looked at association between CNVs and quantitative traits measured in the study samples.

CNVs were widely found in study samples and reference genomes. Discrepancies were found between sets of CNVs called by different analytical tools. I detected 4016 CNVs in 1964 individuals, out of a total of 2789 participants from the three population isolates, which clustered into 743 copy number variable regions (CNVRs). Features of these CVNRs, including frequency and distribution, were compared and were shown to differ significantly between the Orcadian, South Tyrolean and Dalmatian population samples. Consistent with the inference that this indicated population-specific CNVR identity and origin, it was also demonstrated that CNV variation within each population can be used to measure genetic relatedness. Finally, I discovered that individuals who had extreme values of some metabolic traits possessed rare CNVs which overlapped with known genes more often than in individuals with moderate trait values.

Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisors, Professor David Porteous and Harry Campbell. Their continuous support and advice on both my research and student life have greatly helped me through all the highs and lows during my PhD study. They have devoted time and energy supervising every stage of my research and actively helped me outsourcing external assistance when needed during my research. They frequently inspired me on scientific ideas and processes of thinking. Without the generous support and patience from these two mentors, the completion of this thesis would not have been possible.

Many thanks to Dr. Andy MacLeod who always generously helped me with bioinformatics. I would also like to thank Dr. Veronique Vitart, Dr. Caroline Hayward, Dr. Sara Knott and Dr. Pippa Thomson for their help with data preparation, advice and discussion; Dr. Rehab Rahman for her help in the early stages of my research; Archie Campbell, Dave Liewald and Stuart Morris for IT support and all the other members of Generation Scotland team, who inspired me with their professionalism and devotion to a scientific career.

I would also like to acknowledge my research collaborators Dr. Jim Wilson of the ORCADES study, Dr. Igor Rudan of the CROAS study, Prof. Peter Pramstaller and Dr. Andrew Hicks of the MICROS study for the provision of SNP data and useful discussion.

I am grateful to Generation Scotland and the College of Medicine and Veterinary Medicine (University of Edinburgh) for providing financial support. Last but not least, I sincerely thank my parents for their unconditional love and my friends for making my life in Edinburgh wonderful. Special thanks go to my fiancé Jacky who was always there for me.

Publications from this thesis

Papers

Wanting Chen, Igor Rudan, Caroline Hayward, Veronique Vitart, Jim F Wilson, Alan Wright, Sara Knott, Sarah H Wild, Andrew A Hicks , Peter P Pramstaller and David J Porteous. (2011). Copy Number Variation across European Populations. PLoS ONE 6(8):e23087

Abstracts and oral presentations

Wanting Chen, Rehab Abdel-Rahman, Igor Rudan, David Porteous, Harry Campbell. (2008) Copy Number Variation in Croatian Population. Wellcome Trust School of Human Genomics, 17-21 August 2008, Hinxton, UK

Abbreviations

aCGH	Array-based Comparative Genome Hybridization
BAC	Bacterial Artificial Chromosome
BAF	B Allele Frequency
BF	Bayes Factor
CBS	Circular Binary Segmentation
CGH	Comparative Genome Hybridization
CMT	Charcot-Marie-Tooth
CNAG	Copy Number Analyzer for GeneChip arrays
CNP	Copy Number Polymorphisms
CNV	Copy Number Variation
CNVR	Copy Number Variable Region
CROAS	the Croatian Anthropogenetic Study
Cy	Cyanine
DECIPHER	Database of Chromosomal Imbalances Using Ensembl Resources
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic Acid
DOP-PCR	Degenerated Oligonucleotide Primer PCR
DSB	Double Strand Breaking
DSL	Disease Susceptibility Loci
EM	Expectation Maximization
ESP	End-sequence Profiling
FISH	Fluorescence in situ Hybridization
FoSTeS	Fork Stalling and Template Switching
GADA	Genome Alteration Detection Algorithm
GWAS	Genome Wide Association Studies
HERVs	Human Endogenous Retroviruses
HGSC	Human Genome Sequencing Consortium
HMM	Hidden Markov Model
INDELS	small insertion and deletions

L1	Long Interspersed Element-1
LCRs	Low Copy Repeats
LD	Linkage Disequilibrium
LLR	LogR Ratio
LOH	Loss of Homozygosity
MICROS	the Genetic Study of Three Population Microisolates in South Tyrol
NAHR	Non-allelic Homologous Recombination
NGS	Next-generation Sequencing
NHEJ	Non-homologous End Joining
OaCGH	Oligonucleotide Array CGH
OB-HMM	Objective Bayes Hidden Markov Model
OMIM	Online Mendelian Inheritance in Man
ORCADES	the Orkney Complex Disease Study
PCR	Polymerase Chain Reaction
PEM	Paired-end Read Mapping
PES	Paired-end Read Sequencing
qPCR	Quantitative Fluorescent real-time PCR
QTN	Quantitative Trait Nucleotide
RD	Read Depth
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
SBE	Single-base Extension
SD	Segmental Duplications
SNP	Single Nucleotide Polymorphisms
SNV	Single Nucleotide Variants
SV	Structural Variants
T2D	Type 2 Diabetes
WGSA	Whole Genome Sampling Assay
WGTP	Whole-genome Tiling Path
WTCCC	Welcome Trust Case Control Consortium

List of Figures

List of Tables

Table of Contents

Chapter 1

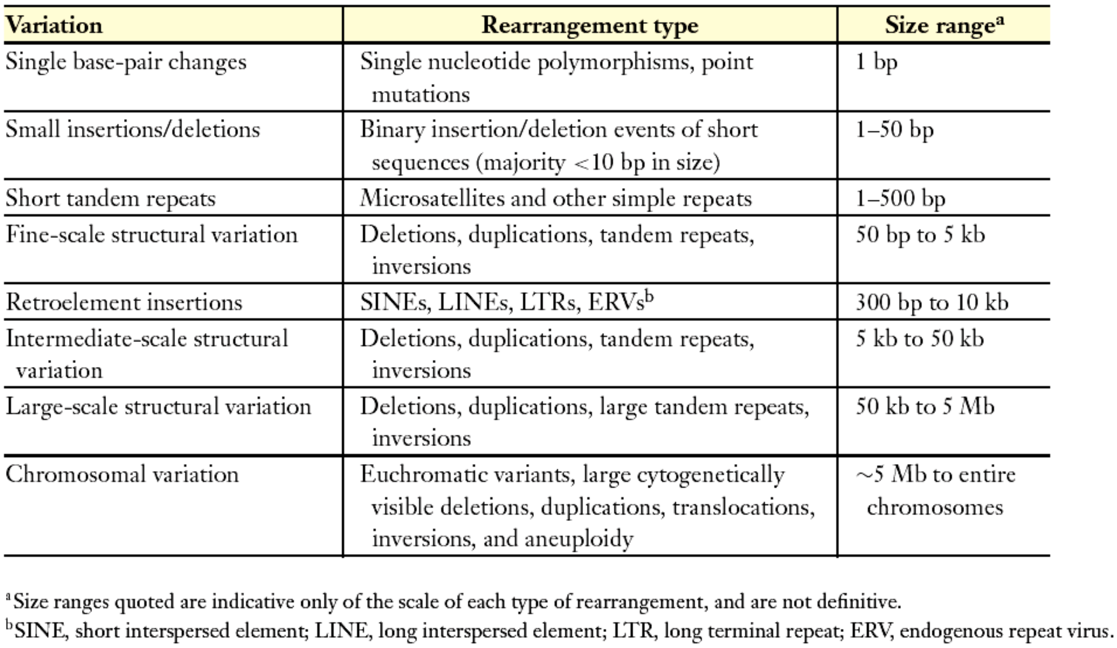
Introduction

1.1 Copy Number Variation as a source of genetic variation

Genetic variation in the human genome takes many forms, from single nucleotide changes (including single nucleotide polymorphisms, SNPs), to fine-scale copy number changes such as small insertions and deletions (INDELS), microsatellite and minisatellite repeats, larger scale structural variations such as inversions, translocations and copy number variations (CNVs), to large microscopically visible chromosome anomalies (Sharp et al., 2006, Table 1). The above variations and polymorphisms constitute the architecture of the human genome and underline the differences between individuals at the DNA (Deoxyribonucleic Acid) level.

SNPs are common polymorphic markers that are uniformly distributed throughout the genome. With the guidance of the DNA reference sequence (e.g. HapMap (The International HapMap Consortium, 2005)) and the use of high-throughput genome-wide SNP microarrays, numerous studies have successfully found genetic determinants of human complex traits in a fast and economic way. SNPs may be associated with phenotypic variation either through direct causal effects or by indicating the location of causal variants which are in linkage disequilibrium with them (Stranger et al., 2007). So far, SNPs are considered to be the most common form of genomic variation and to account for a large proportion of normal phenotypic variation. During the last few years, structural variants such as copy number variants have attracted much attention, as they are also found to be widely spread all over the genome. Despite the advance of technologies to detect variants in different forms in the genome, genomic rearrangements of median size, between 500bp and 5Mb, have remained largely unnoticed until recently. That has changed with the advent of studies that have discovered an abundance of submicroscopic copy number variation of DNA segments.

Table 1.1 Type of genetic variants and their relative sizes (Sharp et al., 2006)



Variation	Rearrangement type	Size range[a]
Single base-pair changes	Single nucleotide polymorphisms, point mutations	1 bp
Small insertions/deletions	Binary insertion/deletion events of short sequences (majority <10 bp in size)	1–50 bp
Short tandem repeats	Microsatellites and other simple repeats	1–500 bp
Fine-scale structural variation	Deletions, duplications, tandem repeats, inversions	50 bp to 5 kb
Retroelement insertions	SINEs, LINEs, LTRs, ERVs[b]	300 bp to 10 kb
Intermediate-scale structural variation	Deletions, duplications, tandem repeats, inversions	5 kb to 50 kb
Large-scale structural variation	Deletions, duplications, large tandem repeats, inversions	50 kb to 5 Mb
Chromosomal variation	Euchromatic variants, large cytogenetically visible deletions, duplications, translocations, inversions, and aneuploidy	~5 Mb to entire chromosomes

[a] Size ranges quoted are indicative only of the scale of each type of rearrangement, and are not definitive.

[b] SINE, short interspersed element; LINE, long interspersed element; LTR, long terminal repeat; ERV, endogenous repeat virus.

Copy number variation (CNVs, or copy number changes, copy number alterations), are defined as a DNA segment that is 1kb or larger and present at variable copy number in comparison with a reference genome (Redon et al., 2006), exclusive of the insertion/deletion events caused by transposable elements (Freedman et al., 2004). CNVs are discovered as deletions, insertions, duplications and complex multi-site variants. They are a subset of structural variation, which is defined as genomic alterations that involve segments of DNA that are larger than 1kb or translocation between one chromosome and another that may result in no loss of genetic material, but do involve loss of genetic coding information.

CNVs have been studied as gene copy number differences between individuals at specific loci, especially for rare diseases for some considerable time , for example large deletions on chromosome 15 which caused Prader-Willi syndrome, and hemizygosity at the elastin locus in Williams syndrome, The common α-globin gene deletions widely detected in different isolated populations and the well established study of identical duplications of gene PMP22 in autosomal dominant and sporadic forms of Charcot Marie-Tooth disease type 1 are other examples (Sebat, 2007). However, the knowledge of such genomic rearrangements was limited until recently, and the prevalence and impact of CNVs was assumed to be small. Since 2004, a wave of structural variant studies in normal human individuals leads to a whole new understanding of CNVs, revealing them as a major source of human genetic variation (Conrad et al., 2006; Iafrate et al., 2004; Redon et al., 2006; Sebat et al., 2004; Tuzun et al., 2005; McCarroll et al., 2005).

The advance of microarray technology and availability of complete human genome sequencing enabled researchers to capture genome wide CNVs in multiple individuals, and also shed light on their location and frequencies. The first attempt was made in 2004 by two independent groups, Iafrate et al. (2004) and Sebat et al (2004). Both groups surveyed the genome wide copy number changes in a number of human genomes, and revealed the extent of this category of genetic variation at a previously unanticipated level. Iafrate et al.(2004) detected 255 loci which contain genomic imbalances among 55 unrelated individuals, while Sebat et al. (2004) found 221 CNVs which represented 76 Copy Number Polymorphisms (CNP, defined as CNVs at >1% frequency) in 20 individuals. Some of these loci influence genes that have important biological roles such as neurological functions and metabolism (Iafrate et al., 2004; Sebat et al., 2004). More studies followed in the subsequent years, which made use of existing SNP genotyping data and clone paired-end sequencing data to detect CNVs (Tuzun et al., 2005; Conrad et al., 2006; McCarroll et al., 2005).

In 2006, a first generation CNV map across the whole human genome was published by Redon et al., (2006). This map was based on 270 individuals from four different ethnic populations of European, Asian or African descents, each of whom contributed to the original International HapMap Project. 1447 copy number variable regions (CNVRs) were reported, which in total covered 12% of the genome. They also observed enrichment of genes in functional categories such as cell adhesion, sensory perception of smell and of chemical stimulus, and also neurophysiologic processes (Redon et al., 2006). Other studies utilizing HapMap data also emerged. Comprehensive investigation of CNVs in this representative sample set of humans revealed a series of characteristics of CNVs, such as chromosomal distribution, correlation with segmental duplications (SDs) in the genome and gene content (Kohler and Cutler, 2007; Korn et al., 2008; Lin et al., 2008; Locke et al., 2006; McCarroll et al., 2008; Redon et al., 2006; Wang et al., 2007). Although platform choice and algorithm differences in these studies resulted in a degree of discrepancy of CNVs detected even for the same individuals (see Chapter 5 for details), these studies have greatly broadened the dimension of human genetics research and have brought a population perspective into such investigations.

Whilst conventional CNV detection methods such as mining SNP genotyping data or using array-based Comparative Genome Hybridization (array-CGH) were widely adopted in CNV studies, some early fruits were harvested from fine-scale CNV detection based on whole genome sequencing technologies, which enabled detection of CNVs at unprecedented resolution. To date, whole genome sequencing has been reported in at least 33 individual genomes; 30 of them have had CNVs determined (Ahn et al., 2009; Bentley and et al., 2008; Drmanac et al., 2010; Kim et al., 2009; Levy et al., 2007; Lupski et al., 2010; McKernan et al., 2009; Pelak et al., 2010; Pushkarev et al., 2009; Schuster et al., 2010; Wang et al., 2008; Wheeler et al., 2008. See Chapter 4 for details). These studies have demonstrated that compared to SNP, CNVs confer higher level differences between individuals. Pioneer study of CNVs detected from two partially sequenced genomes estimated the total amount of sequence divergence to be 0.5%, with the majority of variation being due to CNVs (Korbel et al., 2008), which is contrary to our traditional view that human genomes share 99.9% similarity. An analysis of deletions detected from whole genome sequence for five individuals (Ahn et al., 2009; Bentley and et al., 2008; Levy et al., 2007; Lupski et al., 2010; McKernan et al., 2009; Wang et al., 2008; Wheeler et al., 2008) showed that there was only a small fraction of total detected CNV loci shared between any two individuals among these five (see Chapter 4 for details). In the future, the human genome similarity rate may further be revised, with knowledge of accurate whole genome CNV calls from more sequenced individuals.

The presence of copy number variation is not limited to human genomes, but also widely found in genomes of other species. Genome-wide CNVs have been characterized in great apes, chimpanzees, mice, dogs and drosophila (Perry et al., 2007; Dopman and Hartl, 2007; Pielberg et al., 2002; Locke et al., 2003; Li et al., 2004; Chen et al., 2009; Snijders et al., 2005). The impact of specific CNVs on phenotypes of other species, for example domestic pigs and black sheep, have also been studied (Norris and Whan, 2008; Pielberg et al., 2002).

Since CNVs are of thousands to millions base pairs long, they frequently span entire genes leading to different gene copy numbers between individuals, or alter the intron/exon structure of genes by disrupting exons or fusing genes together (Korbel et al., 2008), therefore CNVs could contribute to disease susceptibility or phenotypes through alterations in gene dosage. Disease relevance of DNA copy number alteration is not a new topic. Recurrent deletions of tumor suppressors and amplifications of oncogenes have been investigated in cancer studies for a long time. The discovery of CNVs in germline DNA in diseased and healthy individuals extends the frontier of such researches from cancer to more and more common and rare diseases. Since then, a number of CNVs were shown to be associated with disease (see Chapter 1.5). Of equal importance to SNPs, CNVs as a major source of genetic variation has brought a new dimension to genomics and disease genetics.

1.2 Characteristics of copy number variants

1.2.1 How many CNVs are there in the human genome?

Before 2004, the scale of copy number variants in human genome was underestimated. Ever since the pioneer genome-wide CNV discoveries, more and more new CNVs have been revealed from various studies. Redon et al. in 2006 estimated 12% of the human genome is covered by CNVs (Redon et al., 2006); now the proportion of genome regions showing evidence of copy number variation has been revised to be 35.07%, according to Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/).

DGV is a database where publicized CNV results from various platforms are centralized and deposited. Up to early 2011, it has received 42 publications with available CNV data. It now contains 66741 CNVs underlying 15963 loci (CNVRs); the number of identified CNVRs has dramatically increased over the last few years (Figure 1.1).

Not only has the total number of CNV's discovered increased, but it is also clear that only a small portion of CNVRs identified in each new study overlap those found by others (Freedman et al., 2004; McCarroll and Altshuler, 2007; Redon et al., 2006; Smith et al., 2007), indicating that the catalogue of copy number variability remains incomplete. This may be due to 1) the difference of CNV size detected by use of different methods, 2) CNV frequency differences between different population, 3) the limitation to detect rare CNVs and variable false positive rates and 4) false negative rates to detect CNVs across different studies (Smith et al., 2007). Therefore there is still a great potential to identify novel CNVs and make a contribution to the building of a more complex genetic architecture of human genome.

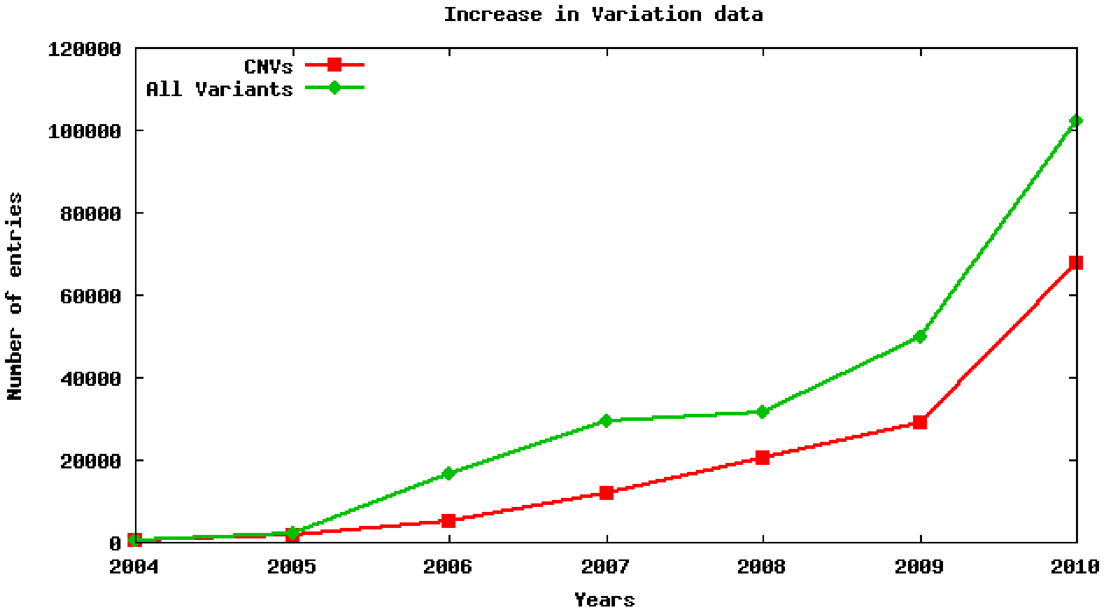


Figure 1.1 Increase in published CNV and InDel data that have been added to DGV since 2004. The numbers reflect the year of publication. The studies published are all included in the 2004 total. (From http://projects.tcag.ca/variation/, last accessed in April 2011)

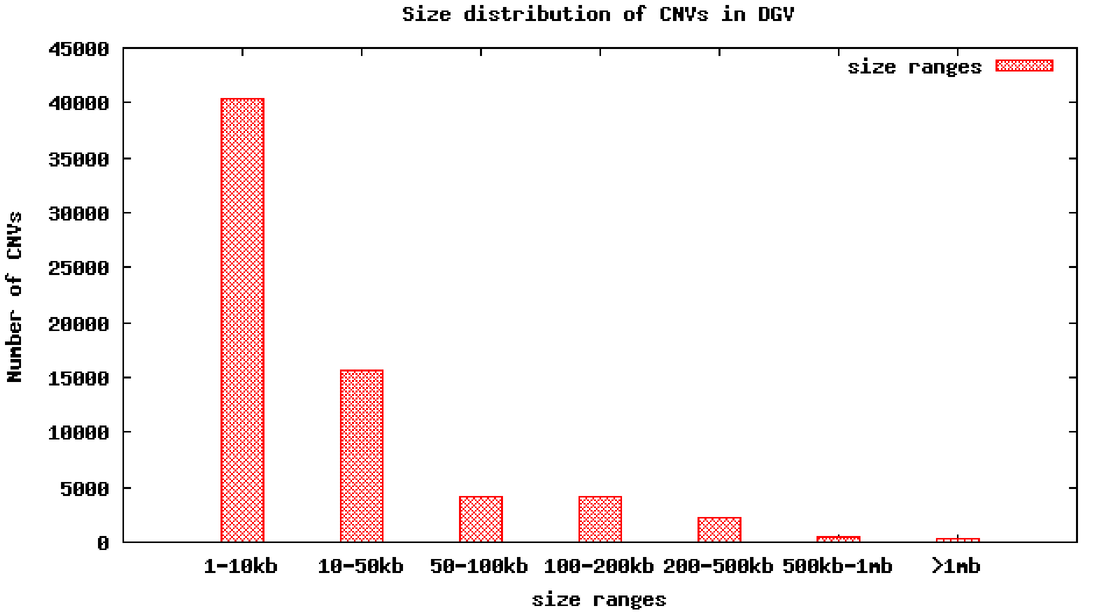


Figure 1.2 CNV size distributions in DGV (From http://projects.tcag.ca/variation/, last accessed in April 2011)

1.2.2 Length of CNVs

The size distribution of the CNVs in DGV is shown in Figure 1.2. Most CNVs are of small to median sizes, while only a small portion of CNVs are large.

The detection method of choice might result in different size ranges of predicted CNVs. First of all, the targeted length of CNVs is limited by resolution of assays. The insertion size of BAC (Bacterial Artificial Chromosome) CGH arrays is typically 80-200kb, therefore the identification of CNVs smaller than 50 kb is difficult. Fosmid and cosmid clones of approximately 40 kb in length improve the resolution to about 20kb. Resolution of SNP arrays is variable across the genome and for many array types there is a lower limit of 10-40kb (Carter, 2007). Whole genome sequencing potentially provides the highest resolution, but the short read length limits the size of detected CNVs. Secondly, density and coverage of probes in an array is positively correlated to the number of CNVs detected. Third, breakpoint determination affects the size of actual CNVs. For example, the size of CNVs detected by BAC arrays is significantly overestimated, due to long insert sizes of BACs (but breakpoints are determined as the boundaries of the first and last probes in a region showed signal change).

1.2.3 Chromosomal location of CNVs

The distribution of CNVs on chromosomes is not uniform (Fig 1.3). Enrichment of CNVs is observed in peri-telomeric and/or sub-centromeric regions for most chromosomes. The CNVRs cluster in functional categories such as cell adhesion, sensory perception of smell and of chemical stimulus, as well as neurophysiologic processes (Nguyen et al., 2006).

1.2.4 CNVs and segmental duplications

Copy number changes are found to preferentially enrich near segmental duplications (SD), which are defined as duplicated sequences of >1kb with 90% or more sequence identity in the reference human genome assembly (Bailey et al., 2002). A number of studies which specifically focus on structural variants in SDs also argue that these loci are hotspots for chromosomal rearrangement and copy number variations (Locke et al., 2006; Sharp et al., 2005).

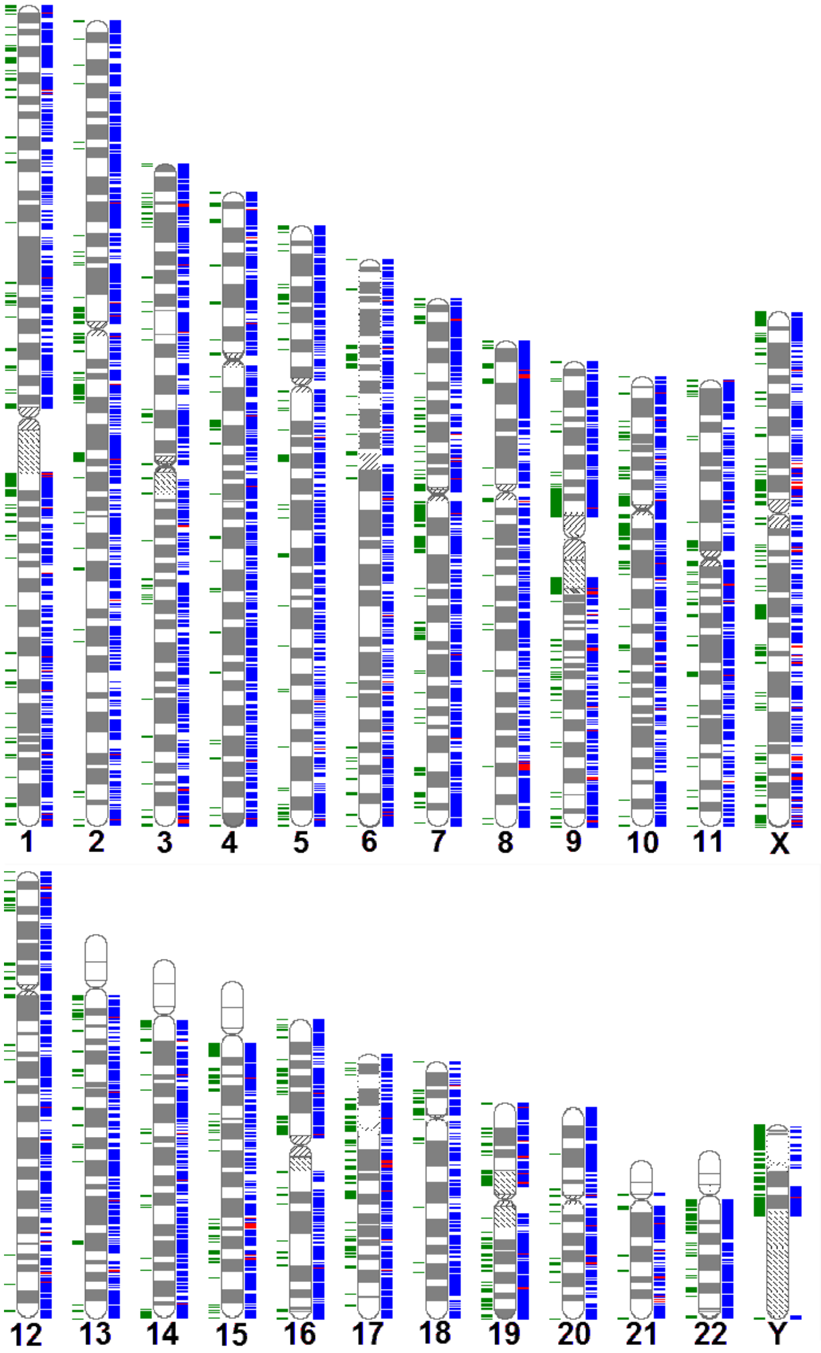


Figure 1.3 Genome-wide view of CNVs. Blue bars indicate reported CNVs, red bars indicate reported inversion breakpoints, green bars to the left indicate segmental duplications. (Adapted from http://projects.tcag.ca/variation/, last accessed in April 2011)

1.3 Detection of copy number variants

1.3.1 Classical cytogenetic techniques to detect structural variants

Cytogenetics is the study of chromosome structure. Classical cytogenetic techniques include routine Giemsa-banding (G-banding) to reveal large structural variations which are microscopically visible. Another category of cytogenetics is molecular cytogenetics. One well-known example of classical cytogenetics application is the diagnosis of trisomy 21, which leads to Down syndrome. This procedure was established in 1950s and has been widely used in routine prenatal examinations. A large number of abnormalities of chromosome structure (partial deletion, duplication or inversion) or number (aneuploidy) have been described to be associated with a very wide range of congenital abnormalities (James et al., 1971).

1.3.2 Fluorescence in situ hybridization (FISH) and chromosome based comparative genome hybridization (CGH)

Fluorescence *in situ* hybridization (FISH) is a type of molecular cytogenetic techniques introduced in the 1980s. It is an *in situ* hybridization technique in which a labeled probe of specific DNA sequences, for example a Bacterial Artificial Chromosome (BAC) clone or fosmid clone, is hybridized to a preparation of metaphase chromosomes or interphase DNA, which is usually attached to solid media (e.g. a glass slide). Then the chromosome DNA and probe mixture is denatured, so that the single-stranded probe and single stranded DNA can re-anneal, and at the same time the probe hybridize to the complementary DNA sequences, and now a double stranded molecule is reformed. Following hybridization, unbound probes are washed away, and the hybridized probes can be visualized directly if they were tagged with fluorochromes such as Cyanine (Cy) or Alexa Fluor dyes, or can be detected by antibodies against hapten-tagged probes, or affinity agents such as avidin or streptavidin if probes are labeled with the biotin and digoxigenin systems (Bauman et al., 1980).

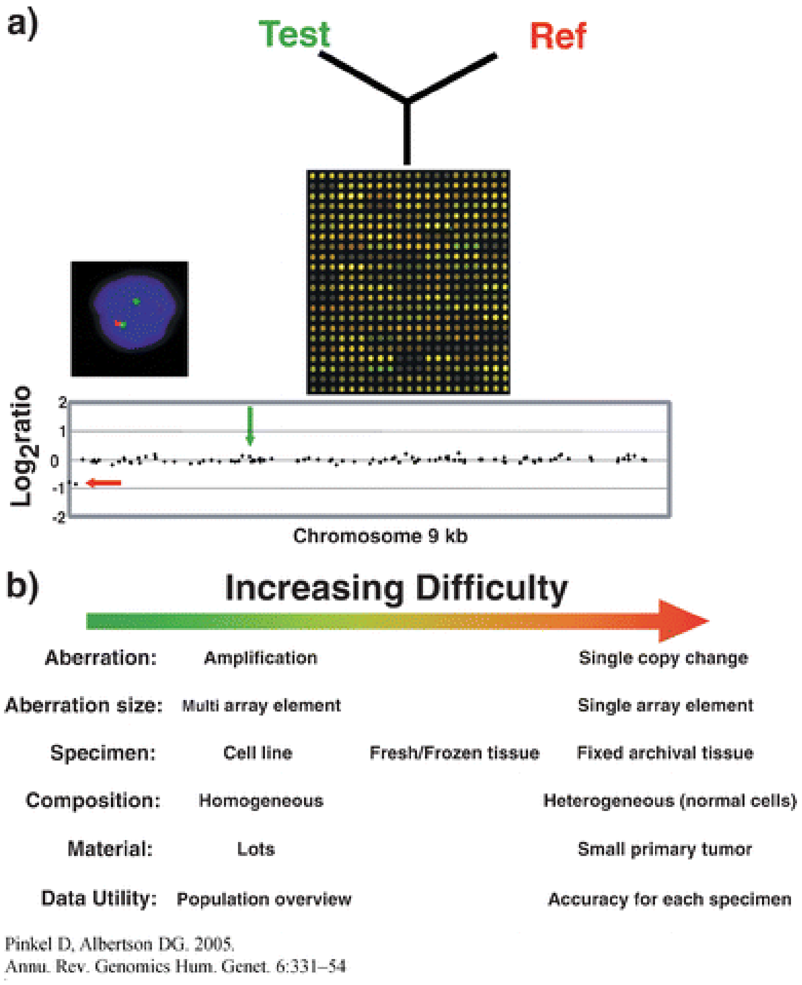
Compared to classical cytogenetic karyotyping, FISH has the advantage of increased the resolution to ~100kb, which enables the detection to submicroscopic copy number changes. Over the years FISH techniques have evolved, for instance multi-colour hybridization which allow visualizing of rearrangements involving multiple chromosomes and Fiber-FISH which benefit from extended chromatin fibers in the DNA preparation to detect small copy number changes. Fiber-FISH can detect deletion gaps or tandem duplications down to the size of a fosmid clone (40kb) or even smaller; it was utilized in the survey of and related segmental duplication in human and chimpanzees (Perry et al., 2008).

Another type of molecular cytogenetic techniques is chromosome-based comparative genome hybridization (CGH). The first step of CGH is to labeled test and reference DNA differentially, then those DNA simultaneously hybridize to chromosome metaphase spreads (at the same time unlabelled Cot-1 DNA is also hybridized to block DNA repeats). The hybridization will then be detected with two fluorochromes, and genomic regions of gains or losses would be illustrated as changes in the ratio of intensities of the two fluorophores along the chromosomes. The technique was first introduced in the analysis of amplification of the *myc* locus at 8q24 in tumor cell line (Kallioniemi et al., 1992). Thereafter, it has been widely applied in the analysis of tumor chromosomal aberrations. The advance of CGH is that it allows whole-chromosome or whole-genome surveys of chromosomal rearrangement and aberrations, while previous approaches only target specific genomic regions. However, the use of metaphase chromosomes largely limits the resolution of CGH.

1.3.3 Array-based comparative genome hybridization (array-CGH)

Array-based comparative genome hybridization (array-CGH) is an upgrade to the traditional CGH technique. By applying comparative genome hybridization to a solid surface with immobilized DNAs targeted in small spots, it greatly enhanced the resolution and application range for copy number detection.

A CGH array consists of mapped DNA sequences spotted or directly synthesized onto a solid surface, for exame a glass slide. Different sources of DNA sequences which could be generally classified as genomic inserts are used in array approaches, such as BAC, cosmid or fosmid clones, cDNA clones, genomic polymerase chain reaction (PCR) products or oligonucleotides. For the hybridization experiment, uniquely labeled subject and control DNA are co-hybridised onto arrays with Cot-1 blocking agent which would count out signal from common repetitive sequences (Figure 1.4 a)). Then the test and reference DNA signal intensity is recorded for all probes on the array. Significant deviation from the test/reference ratio of 1 (equivalent to $\log_2$ (test/reference)=0) for a probe (or a series of consecutive probes) would be interpreted as DNA copy number changes (Figure 1.4 b)) (Solinas-Toldo et al., 1997; Pinkel and Albertson, 2005).



Aberration:	Amplification		Single copy change
Aberration size:	Multi array element		Single array element
Specimen:	Cell line	Fresh/Frozen tissue	Fixed archival tissue
Composition:	Homogeneous		Heterogeneous (normal cells)
Material:	Lots		Small primary tumor
Data Utility:	Population overview		Accuracy for each specimen

Pinkel D, Albertson DG. 2005.
Annu. Rev. Genomics Hum. Genet. 6:331–54

Figure 1.4 (*a*) Array comparative genomic hybridization (CGH). Genomic DNA from two cell populations is differentially labeled and hybridized to a microarray. The fluorescent signal intensity ratios measured at each array spot are normalized so that the median $\log_2$ ratio is 0. Plotting of the data for chromosome 9 from pter to qter shows that most elements have a ratio near 0. The two elements nearest pter have a ratio near −1, indicating a reduction of a factor of two in copy number. Fluorescent in situ hybridization (FISH) with a red-labeled probe for the deleted region and a green-labeled control probe (genome locations indicated by the *red and green arrows* on the ratio profile) shows that the cells contain two copies of the green probe and only one for the red, consistent with the array CGH analysis (adapted from Pinkel and Albertson, 2005)

1.3.3.1 BAC array CGH

Bacterial Artificial Clones array CGH is the first generation CGH, and is also used for the earliest genome-wide human CNV surveys (Iafrate et al., 2004; Redon et al., 2006; Solinas-Toldo et al., 1997). BAC arrays targets specific regions of the genome, or tiling the genome at an average resolution of about 1 Mb. Now the BAC arrays can be synthesized with over 30,000 features at a tiling path resolution 80-150kb, due to the availability of overlapping sequencing clone contigs generated for public domain of the Human Genome Project (Fiegler et al., 2006). The DNA preparation for BAC arrays was problematic, because substantial effort was required in bacterial culture handling to extract enough DNA from the BAC clones; this problem was tacked later by applying DNA amplification methods such as rolling circle replication, linker adaptor PCR or degenerated oligonucleotide primer PCR (DOP-PCR). The whole-genome tiling path (WGTP) array platform was generated using the DOP-PCR strategy, with BAC DNA amplified using three different, specifically designed degenerated oligonucleotide primers. At that point complete amplification of the clone DNA was achieved and the effect of *E. coli* host vector DNA contamination was minimized. After years of improvement, the technique of BAC array CGH has matured and it has been a sensitive and reliable platform for the detection of genomic aberrations (Fiegler et al., 2006).

1.3.3.2. Oligonucleotide Array CGH (OaCGH)

Alongside BAC arrays, oligonucleotide-based arrays are also among the most popular methods to detect genomic imbalances. 25~85 mer oligonucleotides are synthesized in-situ onto the solid base of the array which serve as the probes or features for CNV detection. Different oligo arrays are combined with different labeling and hybridization techniques to yield high-resolution copy number measurements. Oligonucleotide arrays are usually commercially available, from suppliers such as Affymetrix, Roche Nimblegen and Agilent Technologies. Non-commercial oligonucleotide array CGH platforms have also been applied , where 60~70 mer oligonucleotide libraries are spotted as elements on the arrays (Ylstra et al., 2006).

Compared to BAC arrays, early oligonucleotide arrays had low signal-to-noise ratio, which leads to variable reported signals for CNV detection (Carter, 2007). Another issue which hindered the wider application of OaCGH is the higher costs to purchase commercial arrays and experimental reagent; therefore OaCGH were mostly used for validation (Conrad et al., 2006; Locke et al., 2006; Wang et al., 2007) or breakpoint mapping (Sharp et al., 2005) rather than whole genome discovery of CNVs, in the early years of its application. New technologies, such as the use of digital mask photolithography, has allowed oligo arrays to be constructed at much higher density, providing better resolution and precision for CNV detection. Flexibility of array design also expanded the usage of oligo arrays: the arrays can be optimized to avoid highly repetitive regions, but can also cover low copy repeats such as segmental duplications where CNVs are abundant. Designs can also be customized to target clinically relevant regions for disease studies. With improved signal-to-noise ratio, enhanced reproducibility, better quality control and the reducing cost, oligo arrays are now recognized as an accurate method for high resolution CNV detection (Cronin et al., 2008; Ylstra et al., 2006; Urban et al., 2006). The latest commercially available oligo arrays include the Agilent SurePrint G3 Human CGH Microarray 1M with 963,029 biological features and NimbleGen Human CGH 4.4M Whole-Genome Tiling Array.

Representational oligonucleotide microarray analysis (ROMA) is a member of the OaCGH technology family. “Representations” of the test and reference genomes are prepared by digesting the genomes with restriction enzymes. After differential PCR amplifications, representations of the entire genome are amplified to show relative increases, decreases or remain equal copy number in the two genomes. Oligonucleotide probes from human genome sequence are hybridized with differently labeled test and reference genome fragments, so that change of copy numbers could be detected from signal intensity changes (Lucito et al., 2003). ROMA, which was developed from a previous method called Representational Difference Analysis, has the advantage of reducing the complexity of a genome with taking restriction enzyme cleaved DNA fragments. This approach has been used in several CNV discovery and disease association studies (Sebat et al., 2004; Sebat et al., 2007; Walsh et al., 2008).

1.3.4 CNV detection using Single Nucleotide Polymorphism data

Single Nucleotide Polymorphism is the most well studied genetic variants in human genomes. Since the 1990s, enormous effort has been put into the SNP discovery, validation and characterization. The International HapMap project is an important landmark in the history of human genetics studies, which provides a catalogue and database of well-characterized SNPs in sampled human individuals from four major ethnic populations (The International HapMap Consortium, 2005). The release of reference SNP map in human genome has enabled the development of high-throughput array technologies for SNP genotyping, for example SNP genotyping platforms by Affymetrix and Illumina Inc. With the guidance of DNA reference sequence and the use of high-throughput genome-wide SNP microarrays, numerous studies showed success in finding genetic determinants of human complex traits in a fast and economic way.

Many genome-wide SNP association studies with large sample sizes (hundreds to thousands of participants) are already established. Intuitively one may argue whether these SNP genotyping data can be mined for copy number analysis at no extra cost and further, make the data a potential source for CNV-disease association studies complementing SNP-disease association, which may aid understanding of the contribution of a higher level of genetic variants to the disease susceptibility or disease related traits.

A number of algorithms have been introduced to indicate CNVs from SNP array genotyping results (Gunderson and Peitter, 2006; Li and Wong, 2001; Colella et al., 2007; McCarroll et al., 2008; Korn et al., 2008).Compared with array-CGH, SNP genotyping provides extra information to estimate copy number by combining the normalized intensities and allelic ratio. SNP genotyping platform are different in that a combination of two genotyping parameters is analyzed: normalized intensity measurement and allelic ratio. Together, these parameters provide a more sensitive and precise profile of chromosomal aberrations. SNP array data also provides genetic information (haplotypes) of the involved locus. Importantly, the SNP genotyping platform has the capability of identifying copy-neutral LOH (Loss of Heterozygosity) events, such as gene conversion, which cannot be detected with array-CGH (Gunderson and Peitter, 2006).

1.3.4.1 SNP-tagging based on linkage disequilibrium

It is hypothesized tht SNPs could tag adjacent common copy number changes, or CNPs, based on linkage disequilibrium (LD), therefore LD could be utilized for CNV investigation. However, whether SNP can serve as a good proxy for CNV detection still remains unclear (Redon et al., 2006; McCarroll and Altshuler, 2007). Only a small proportion of CNVs has to date been accurately genotyped, making the assessment of linkage disequilibrium around CNVs difficult. Some studies suggested that deletion polymorphisms are generally in strong linkage disequilibrium and segregate on ancestral SNP haplotypes (Hinds et al., 2006; McCarroll et al.,

2005) while some others argued although a number of CNVs are in strong linkage disequilibrium with nearby markers, accurate genotypes can only be captured for small proportion of the tested CNVs (Redon et al., 2006). CNVs are commonly found in regions rich in segmental duplications, but those regions are not favored in SNP selection in commercial SNP genotyping assays, whose design is based on SNP tagging which aims to capture most common variants in the genome, because regions harboring CNVs usually tend to be gene-poor and common CNPs usually cause SNP genotyping assays to fail Hardy-Weinberg and Mendelian inheritance checks. Consequently, these regions are often filtered out during the selection of high-performance SNP assays (McCarroll and Altshuler, 2007; Locke et al., 2006). Other reasons may be high recombination rate in regions of CNV or high rate of spontaneous recurrence of CNVs (Lee and Jeon, 2008; Lee et al., 2007).

1.3.4.2 CNV detection using whole-genome SNP genotyping arrays

SNP arrays originally designed to genotype SNPs in genome-wide association studies can be used to estimate copy number variations. Hybridization signals of probes at each SNP locus can be compared to those from a single or a group of references genomes hybridized on the same array type. The CNV calls can thus be generated (Figure 1.5). Apart from directly utilizing signal intensities, CNVs, in particular deletions, can be inferred from regions with extended loss of heterozygosity (LOH), non-Mendelian inconsistency among families and enrichment of Hardy-Weinberg disequilibrium (Conrad et al., 2006; McCarroll et al., 2005)

1.3.4.2.1 CNV studies utilizing whole-genome SNP data

The first SNP based CNV studies were carried out in 2005, with Affymetrix GeneChip Human 10k arrays to analyze copy number changes from a variety of different sources, including primary tumors, cell lines and blood from patients with unbalanced translocations (Conrad et al., 2006; Herr et al., 2005; McCarroll et al., 2005). Thereafter the resolution of SNP arrays gradually increased, and the companion analytical tools also developed. For example Affymetrix's Genome-Wide Human SNP Array 6.0 which consists of more than 906,000 SNP and 946,000 CNV probes are now available for underlying CNV profile in human samples (Kidd et al., 2008).

The SNP data from International HapMap project was often mined by CNV researchers to generate example maps of CNVs in human genome from multiple population cohorts or served as reference for testing SNP-base CNV detection algorithms (Komura et al., 2006; Conrad et al., 2006; Locke et al., 2006; Redon et al., 2006; Kohler and Cutler, 2007; Ting et al., 2007; Wang et al., 2007; Korn et al., 2008; Lin et al., 2008; McCarroll et al., 2008; Rigaill et al., 2008; Sanders et al., 2008; Shen et al., 2008; Cooper et al., 2008; Pique-Regi et al., 2009; LaFramboise et al., 2005). SNP genotyping data from already established association studies has been routinely recycled for CNV determination (see 1.5 for references). The advantage of such a study design is that knowledge of SNPs, CNVs and other clinically useful data such as uniparental disomy (which is a copy number neutral LOH detectable by analyzing hybridization signals) (Dunbar et al., 2008) can all be gained from the same array simmutanously.

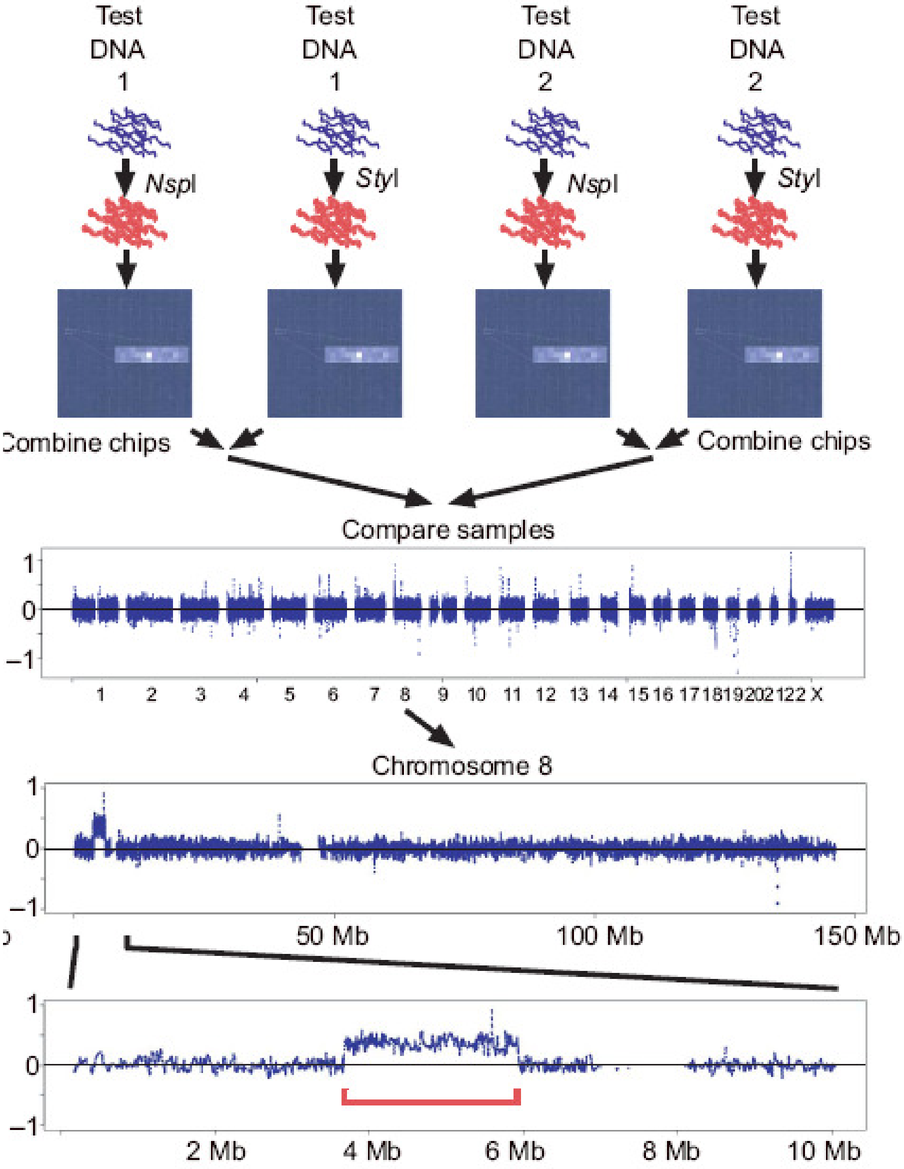


Figure 1.5 Protocol outline of CNV detection from hybridization signal intensity data of SNP arrays (Adapted from Redon et al., 2006)

The major problem of using earlier generation SNP arrays to detect CNVs is uneven probe spacing, with particularly low density of SNPs near or at repeat-rich regions such as segmental duplication, centromere and telomeres. This will directly cause inaccurate CNV calls which in practice are inferred from signal intensity data for SNPs. The later SNP genotyping platforms with denser SNPs perform better in detecting CNVs. Several recent SNP genotyping platforms have taken copy number detection into account, which include non-polymorphic probes specifically selected for their genomic positions and for linear response to copy number changes. For example Affymetrix Genome-wide Human SNP array 5.0 and 6.0, and Illumina Human 1M-Duo Bead Chip, with over 1.2 million markers including probes designed to target known CNV regions and gaps between HapMap SNPs. These platforms have greatly enhanced power to detect CNVs in association studies, integrating both SNP and CNV assessments in the pipeline of such researches (Korn et al., 2008).

1.3.4.2.2 SNP-based CNV detection algorithms

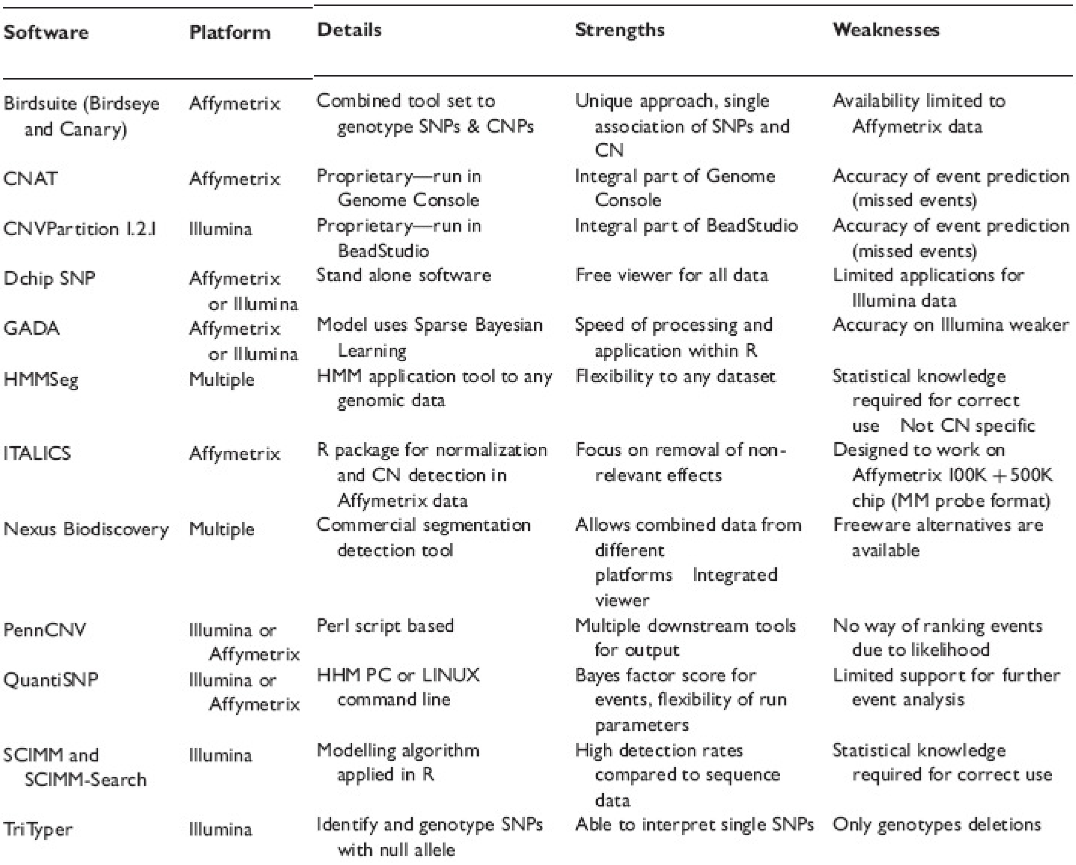
A number of bioinformatics tools have been designed to detect CNVs using the intensity data from hybridization of sample DNA to the probes on the array. LogR Ratio (LLR) and B Allele Frequency (BAF) are the two most important parameters of singnal intensity for CNV detection. The detection algorithms fall into generally two major categories: Hidden Markov model (HMM) and circular binary segmentation (CBS). Building upon the statistical principles of HMM and CBS a number of algorithms have been developed; some examples can be viewed in Table 1.2.

The assumption of HMM is that the observed intensities of each SNP probe are related to an unobserved copy number state at each locus, so that a DNA segment of copy number change can be determined if consecutive probes within this segment all show the same non-neutral copy number state (Details in Chapter 2).

With prior knowledge of modeling statistics, algorithms have been developed to infer copy number changes with genomic SNP data. HMMseg is one of the earliest algorithms designed for this purpose, which is command line operated (Day et al., 2007). However application of correct modeling procedures is not an obvious process to non-statisticians. For these reasons software with user-friendly interface has been developed which allows guided applications of HMM methods (Winchester et al., 2009). QuantiSNP and PennCNV are two academically developed software tools that are freely available for CNV prediction. Users can apply HMM to their own data using these tools.

QuantiSNP (Colella et al., 2007) was initially designed for Illumina Infinium array platforms, but the later versions of this software have been proved to have satisfactory accuracy on Affymetrix and Illumina GoldenGate data where SNP coverage is suitable. The output of QuantiSNP gives a log Bayes factor with its prediction which is a post-process parameter to indicate the likelihood of the results. The user can rank events in order of Bayes factor and chose a satisfactory cutoff to define their list of CNVs. QuantiSNP is widely used in CNV discovery and disease associations, as highlighted by CNV studies in Autism spectrum disorders and schizophrenia (Moreno-De-Luca et al., 2010; Pinto et al., 2010; Stefansson et al., 2008; Vrijenhoek et al., 2008; Wang et al., 2007; Wang K. et al., 2010).

Table 1.2 Examples of SNP-based CNV detection algorithms (Winchester et al., 2009)



Software	Platform	Details	Strengths	Weaknesses
Birdsuite (Birdseye and Canary)	Affymetrix	Combined tool set to genotype SNPs & CNPs	Unique approach, single association of SNPs and CN	Availability limited to Affymetrix data
CNAT	Affymetrix	Proprietary—run in Genome Console	Integral part of Genome Console	Accuracy of event prediction (missed events)
CNVPartition 1.2.1	Illumina	Proprietary—run in BeadStudio	Integral part of BeadStudio	Accuracy of event prediction (missed events)
Dchip SNP	Affymetrix or Illumina	Stand alone software	Free viewer for all data	Limited applications for Illumina data
GADA	Affymetrix or Illumina	Model uses Sparse Bayesian Learning	Speed of processing and application within R	Accuracy on Illumina weaker
HMMSeg	Multiple	HMM application tool to any genomic data	Flexibility to any dataset	Statistical knowledge required for correct use Not CN specific
ITALICS	Affymetrix	R package for normalization and CN detection in Affymetrix data	Focus on removal of non-relevant effects	Designed to work on Affymetrix 100K + 500K chip (MM probe format)
Nexus Biodiscovery	Multiple	Commercial segmentation detection tool	Allows combined data from different platforms Integrated viewer	Freeware alternatives are available
PennCNV	Illumina or Affymetrix	Perl script based	Multiple downstream tools for output	No way of ranking events due to likelihood
QuantiSNP	Illumina or Affymetrix	HHM PC or LINUX command line	Bayes factor score for events, flexibility of run parameters	Limited support for further event analysis
SCIMM and SCIMM-Search	Illumina	Modelling algorithm applied in R	High detection rates compared to sequence data	Statistical knowledge required for correct use
TriTyper	Illumina	Identify and genotype SNPs with null allele	Able to interpret single SNPs	Only genotypes deletions

PennCNV (Wang et al., 2007) was also tailored for Illumina genotyping platforms at first, and later modified to be compatible to Affymetrix platforms in a version called "PennCNV-Affy". It has a number of downstream analyses including the use of family trio data in analysis, in which the family information (when applicable) is taken into account to give more confidence of a CNV detected to be passed on from parent(s) to offspring. It also includes a number of options to handle results such as scripts which allow the viewing of PennCNV results in BeadStudio Chromosome Browser or the web-based UCSC browser (http://genome.ucsc.edu/). The application of PennCNV to detect CNVs can be found in over 60 published articles (http://www.ncbi.nlm.nih.gov/pubmed, last accessed April 2011), in population CNV profiling (Cooper et al., 2008; McQuillan et al., 2008; Perry et al., 2008), tumor studies (Cooper et al., 2008; Jacobs et al., 2007; McQuillan et al., 2008; Perry et al., 2008; Toujani et al., 2009), and CNV association with diseases (Glessner et al., 2009a; Wang K. et al., 2010; Glessner et al., 2009b; Glessner et al., 2010; Bademci et al., 2010).

Other HMM based programmes include: the Birdsuite package which integrate calling of common CNVs (with the knowledge of categorized CNVs from publications) and the discovery of rare CNVs (McCarroll et al., 2008); dChip software which was originally developed for Affymetrix platforms and outputs and LOH score alongside each prediction (Li et al., 2008); Copy Number Analyzer for GeneChip arrays (CNAG) (Nannya et al., 2005), and many more.

Another category of algorithms are based on Circular Binary Segmentation (CBS). These algorithms were originally developed for arrayCGH analysis to convert noisy intensity values into regions of equal copy number (Olshen et al., 2004), but have been modified for SNP genotyping arrays. CBS continuously divides a region into segments until it finds a segment with a different copy number compared to the neighboring region. This detection of change-point along chromosomes is designed to identify all the places which partition the chromosome into segments of the same copy number. Segment ends were joined to form a circle to allow a further likelihood ratio test. The median LLR values are then given to the final set of segments within the region, and this median values are used to define copy number status of each segment.

DNAcopy and cnvPartition are two packages among many CBS based algorithms. Traditionally, CBS scans a chromosome multiple times and generate a permutation reference distribution to obtain the corresponding P value for each segment of CN change, which is time consuming. As the number of markers increases, the number of computations increase exponentially, which is not favoured in CNV detection from newer arrays which contain hundreds of thousands markers. DNAcopy implements a ‘stopping rule’ into the basic CBS algorithm; this will stop a computation process early when there is strong evidence for the presence of copy change of the segment being assessed (Venkatraman and Olshen, 2007a). cnvPartition was developed by Illumina for their proprietary software BeadStudio. As a plug-in, it can easily be applied to LRR and BAF data organized and analyzed in BeadStudio. cnvPartition contains two modules: one for breakpoint identification using LRR, and another for assigning copy number to the regions between identified breakpoints with information of BAF (BeadStudio TechNotes of CNV algorithms). Other examples of computational tools utilizing CBS are The Genome Alteration Detection Algorithm (GADA) (Pique-Regi et al., 2008), and commercial software Partek (Partek Inc., St. Louise, MO) and HelixTree (Golden Helix, Inc.).

1.3.4.2.3 Choice of algorithm for SNP based CNV detection

Accurate CNV prediction from hybridization singnal intensities largely relies on the performance of sophisticated algorithms or statistical methods to discover copy number changes. To date a number of commercial and in-house CNV detection software have been developed to process SNP array data, which one or ones to chose from the large collection should be an important concern in the primary stages of any study that intend to mine SNP data for CNV discovery. It has been shown that output from different algorithms showed large scale variability (Korn et al., 2008) , but it is only recently that researchers have started to assess the impact of algorithm choices on resulting CNV calls in a systematic way (Dellinger et al., 2010; Tsuang et al., 2010; Winchester et al., 2008; Zhang et al., 2011).

No consistent conclusion could be made for the algorithms considered in the already published comparison studies. For example, all four studies assessed QuantiSNP but its ranking amongst all algorithms tested was different in each study. The large scale of discrepancy of results from different methods makes one question the power and accuracy of these algorithms. To increase confidence of CNV predictions in the data, it is recommended using a second algorithm on the single dataset to produce the most informative results (Winchester et al., 2009). But it should be noted that, the trade-off of taking overlap of two algorithms is the chance of missing true positive calls only made by one algorithm.

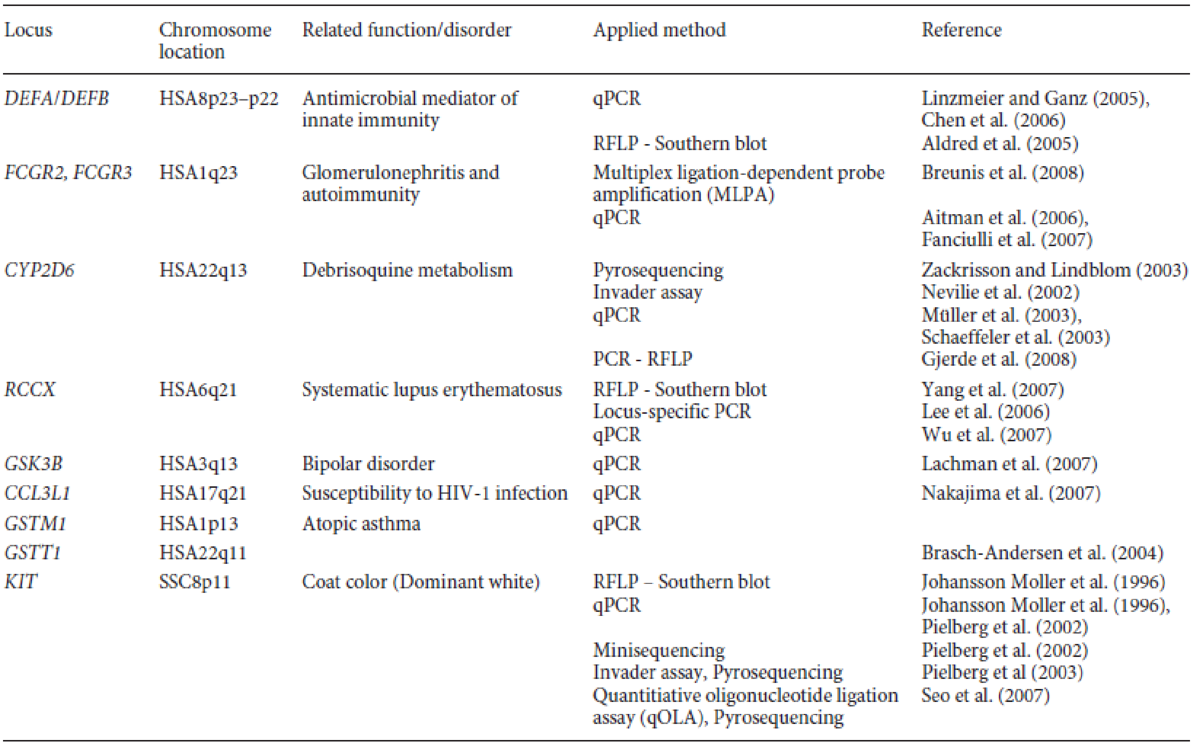
1.3.5 Validation and detection of locus-specific CNVs

The CNV prediction is often followed by quantitative or semi-quantitative measurements of copy variation at selected targeted loci. These measurements serve as independent platforms to validate array-based CNV discoveries and they can precisely determine breakpoints of CNV regions. The detection/confirmation of CNVs at targeted loci can be expended in a wider context, where candidate CNV loci can be genotyped in a large number of samples for disease association studies or in clinical diagnostic settings. Table 1.3 shows some example of locus-specific CNV detection.

Conventional methods such as FISH, RFLP (Pulsed field gel electrophoresis)-Southern Blot, PFGE and long range PCR can be applied for validating a small number of CNVs in limited samples. However, these methods are low-throughput and technically demanding.

Newer techniques such as quantitative fluorescent real-time PCR (qPCR), multiplex quantitative Fluorescent Real-time PCR, pyrosequencing, invader assays and Ligation detection reaction (LDR) have been developed along with the progress of CNV discoveries. These new methods are more cost effective and faster in detecting CNVs at targeted loci, and have the potential to be applied in large-scale investigations.

Table 1.3 Example of locus-specific CNVs and the methods applied for the detection. (Lee and Jeon, 2008)



Locus	Chromosome location	Related function/disorder	Applied method	Reference
DEFA/DEFB	HSA8p23–p22	Antimicrobial mediator of innate immunity	qPCR	Linzmeier and Ganz (2005), Chen et al. (2006)
			RFLP - Southern blot	Aldred et al. (2005)
FCGR2, FCGR3	HSA1q23	Glomerulonephritis and autoimmunity	Multiplex ligation-dependent probe amplification (MLPA)	Breunis et al. (2008)
			qPCR	Aitman et al. (2006), Fanciulli et al. (2007)
CYP2D6	HSA22q13	Debrisoquine metabolism	Pyrosequencing	Zackrisson and Lindblom (2003)
			Invader assay	Nevilie et al. (2002)
			qPCR	Müller et al. (2003), Schaeffeler et al. (2003)
			PCR - RFLP	Gjerde et al. (2008)
RCCX	HSA6q21	Systematic lupus erythematosus	RFLP - Southern blot	Yang et al. (2007)
			Locus-specific PCR	Lee et al. (2006)
			qPCR	Wu et al. (2007)
GSK3B	HSA3q13	Bipolar disorder	qPCR	Lachman et al. (2007)
CCL3L1	HSA17q21	Susceptibility to HIV-1 infection	qPCR	Nakajima et al. (2007)
GSTM1	HSA1p13	Atopic asthma	qPCR	
GSTT1	HSA22q11			Brasch-Andersen et al. (2004)
KIT	SSC8p11	Coat color (Dominant white)	RFLP – Southern blot	Johansson Moller et al. (1996)
			qPCR	Johansson Moller et al. (1996), Pielberg et al. (2002)
			Minisequencing	Pielberg et al. (2002)
			Invader assay, Pyrosequencing	Pielberg et al (2003)
			Quantitiative oligonucleotide ligation assay (qOLA), Pyrosequencing	Seo et al. (2007)

1.3.6 Whole genome sequencing and detection of CNVs

The sequence of original consensus human genome is the bedrock of most CNV discoveries to date. Genomic insert clones, for example BACs or oligonucleotides are representative segments of the reference genome sequence compiled by the HGSC (Human Genome Sequencing Consortium, 2004; Human Genome Sequencing Consortium, 2001). The other major CNV detection platform, namely SNP genotyping arrays, have depended upon SNP information derived from the International HapMap Project, which has identified and catalogued common genetic polymorphisms at single nucleotide level.

On the other hand, the option of directly detecting copy number change of DNA segments from sequence data is considered in a few whole genome sequencing studies. Thus, 666 CNVs were found in the first published DNA sequence from a sole individual, Dr J. C. Venter (Levy et al., 2007), using first-generation shotgun sequencing technology; 602 CNVs detected in Dr J. D. Watson's genome on Roche/454 platform (Wheeler et al., 2008); 5704 in the genome of an African individual (Bentley and et al., 2008) and 2682 in the first Asian genome (Wang et al., 2008), both on Illumina Genome Analyzer platform.

The computational tools to detect CNVs from sequence data fall into two main categories, paired-end read mapping/paired-end read sequencing/end-sequence profiling (PEM/PES/ESP), and read depth (RD) (Koboldt et al., 2010). Whole genome sequencing allows detection of structural variants at unprecedented resolution, however it is still too early to announce unbiased whole genome CNV profiling can be achieved by solely analyzing DNA sequencing data. The state-of-art bioinformatics methods developed for this purpose each has its drawbacks. First of all, due to the short length of sequenced bases, many reads cannot be uniquely mapped to the genome. Second, the alignment is particularly problematic at segmental duplication rich regions; read-depth methods could detect variants at those locations, but their resolution is relatively poor. Third, PEM-based methods have the advantage to detect dosage-invariant SVs, but these algorithms have limited power in detecting insertions larger than the insert size. Fourth, the G+C content throughout the genome, amplification error and uneven likelihood of fragmentation all may cause different representation of certain regions compared to others. Last but not least, many of the data sets do not have sufficient coverage to infer all SVs with statistical significance (Xi et al., 2010).

Therefore more advanced and sophisticated platforms and algorithms for detecting SVs from sequence data are required. It is also advisable that the results from direct DNA sequencing are combined with conventional platforms such as CGH and SNP array which has a better specificity in detecting longer CNVs, so that the relative advantages of these methods can complement each other and the power of revealing genome-wide CNVs can be maximized.

1.4 Mechanisms for the formation of genomic rearrangements

One fundamental question to be asked in CNV studies is what mechanisms contribute to the generation of CNVs. The knowledge of underlying mutation processes for CNV will yield insights into the genomic distribution, evolution and frequency of CNVs in the human population. Four major mechanisms for the formation of genomic rearrangement have been proposed; they are non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS) and L1 retrotransposition.

1.4.1 Non-allelic homologous recombination

It has long been observed that copy number variable regions often coincide with repeat sequences, for instance segmental duplications (Sharp et al., 2005) or Alu repeats (Lee et al., 2007). Some genetic disorders, known to be caused by large scale microdeletion or microduplication of the genome, such as Williams Syndrome and Charcot Marie Tooth Disease, often have breakpoints in or around highly homologous segmental duplications or low copy repeats (LCRs) (Lee et al., 2007). It has been suggested that SDs or LCRs can serve as substrates for NAHR, and that NAHR account for most cases of recurrent CNV formations (Stankiewicz and Lupski, 2006; Lee et al., 2007; Stankiewicz and Lupski, 2006). Other repeats such as SINEs (eg. Alu), LINEs and human endogenous retroviruses (HERVs) can all act as substrates for NAHR, which are observed in less recurrent CNV events (Cooper et al., 2007).

NAHR happens during meiosis, when two sister chromosomes align. In this process, misalignments or unequal cross-over of the homologous sequences will lead to germline rearrangements (Figure 1.6a). Different distribution and participation of the homologous sequences cause different type of CNVs in NAHR, from simple deletion or duplication to more complex rearrangement, such as tandemly duplicated arrays, or other complicated structural variations involving multiple homologous duplicons (Zhang et al., 2009).

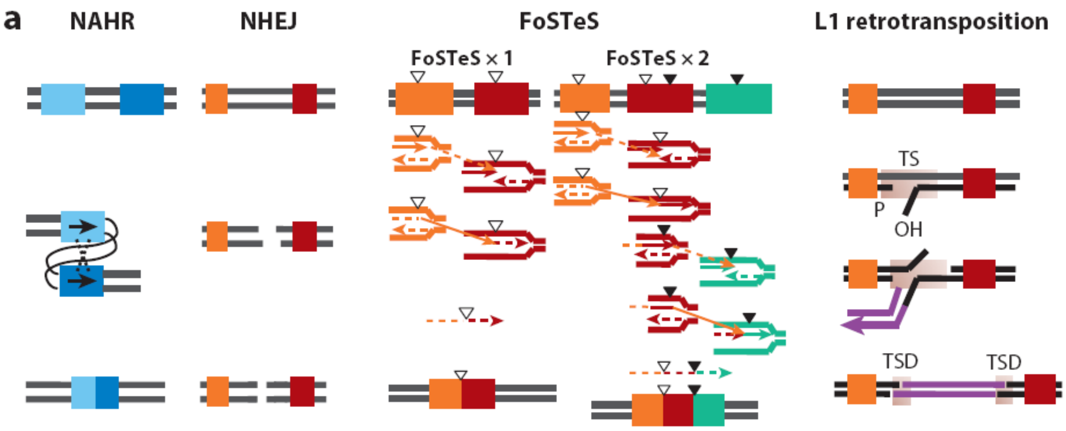
1.4.2 Non-homologous end joining

NHEJ is an alternative mechanism for repair of DNA double strand breaking (DSB) in cells. When there are no extensive homologous sequences to act as the repair template for NAHR near a random DSB, the nucleases will remove the broken ends which is followed by filling of missing nucleotides by the Pol X family of DNA polymerases(Zhang et al., 2009) (Figure 1.6a). It is an error-prone repair mechanism, which often generates gains and losses of nucleotides at the junctions. Compared to NAHR, the knowledge of NHEJ is limited (Lee et al., 2007).

NHEJ more usually appears in unstable regions of the genome, for example subtelomeric regions (Kim et al., 2008). Moreover, many 17p translocations and other nonrecurrent disease-causing deletions have one of their breaking points in LCRs (Stankiewicz and Lupski, 2006). These may implicate the contribution of NHEJ to CNV formation with the absence of NAHR substrates.

1.4.3 Fork stalling and template switching

In 2007, Lee et al. proposed a new model apart from NAHR and NHEJ as a mechanism for human genomic rearrangements (Lee et al., 2007).



b	NAHR	NHEJ	FoSTeS	Retrotransposition
Structural variation type	dup, del, inv	dup, del	dup, del, inv, complex	ins
Homology flanking breakpoint (before rearrangement)?	Yes (LCR/SD, *Alu*, L1, or pseudogene)	No	No	No
Breakpoint	Inside homology	Addition or deletion of basepairs, or microhomology	Microhomology	No specification
Sequence undergoing SV	Any	Any	Any	Transcribed sequences

Figure 1.6 Comparisons and characteristics of the four major mechanisms underlying human genomic rearrangements and CNV formation. (a) Models: NAHR, between repeat, sequences (LCRs/SDs, Alu, or L1 elements); NHEJ, recombination repair of double strand break; FoSTeS, multiple FoSTeS events (×2 or more) resulting in complex rearrangement and single FoSTeS event (×1) causing simple rearrangement; and retrotransposition. TS, target site; TSD, duplicated target site. Thick bars of different colors indicate different genomic fragments; completely different colors (as orange and red or orange/red/green in FoSTeS×2) indicate that no homology between the two fragments is required. The two bars in two similar shades of blue indicate that the two fragments involved in NAHR should have extensive homology with each other. The triangles symbolize short sequences sharing microhomologies. Each group of triangles (either filled or empty) indicates one group of sequences sharing the same microhomology with each other. (b) Characteristic features for each rearrangement mechanism: variation type that each mechanism generates; type of homology sequences flanking breakpoint; the way the breakpoints formed and favorable sequence feature recognized by each mechanism. Specific features of certain mechanisms are shown in red. Abbreviations: dup, duplication; del, deletion; inv, inversion; ins, insertion (Zhang et al., 2009).

In this process, when a DNA replication fork stalls, the lagging strand detangles from the original template and switches to another replication fork if it finds a micro homologous sequence in the new fork. Then DNA synthesis restarts on the new fork and finally a rearrangement is made between the sites of original and new replication fork (Figure 1.6a). The difference between NAHR and FoSTeS are the required size of homologous sequences (intensive repeats for NAHR and microhomology for FoSTeS) and that NAHR happens in chromosome recombination process while FoSTeS in DNA replication process(Figure 1.6b).

1.4.4 Retrotransposition

Transposons are one type of mobile genetic elements in the human genome which can move, or transpose, themselves to new positions within the same genome. They act either by transcribing a segment of targeted DNA into RNA (Ribonucleic Acid), then from RNA back to DNA at a different location by reverse transcription and therefore result in a duplication of the targeted sites (named retrotransposition, and known as "copy and paste"), or by cutting out a DNA segment and inserting it into a new site in the genome via transposase activity (known as "cut and paste"). Long interspersed element-1 (L1) elements are a major class of retrotransposons. They are abundant in human genome, and are the only currently active autonomous retrotransposons (Zhang et al., 2009).

In a survey of eight end-pair sequenced genomes, it was declared that retrotransposition accounted for 30% of all the detected SV indels (Kidd et al., 2008). Another study which analyzed SVs associated with mobile element in J. C. Venter's genome claimed that about 10% of the indels of >100 bp were associated with transposable DNA sequences, including L1, Alu and SVA (composite retrotransposon) (Kidd et al., 2008; Xing et al., 2009)

1.5 Disease and phenotypic relevance of copy number variation

1.5.1 Approaches to identify the genetic components of phenotypic variation

It has been shown that most medical disorders, including cancer, heart disease and mental illness, have a significant genetic component. However, the isolation of disease susceptibility loci (DSL) remains difficult (Smith et al., 2006).

Linkage studies have proven to be very powerful in localising genes for monogenic Mendelian disorders, but are poor for fine mapping genes of small effect size or low penetrance, thus may be much less effective for common complex diseases.

It is widely accepted that association studies, which use large numbers of SNPs or other markers that are genotyped in known linkage regions or candidate genes, are an important complement to linkage studies, in the attempt to localize genes for complex traits. This method involves mapping hundreds of thousands, even millions of single nucleotide polymorphisms (SNPs) throughout the whole genome of multiple individuals in either case-control or population based study design, comparing frequencies of different alleles or haplotypes at the same genetic variant locus in people with the disease (cases) and similar people without (controls). The allele frequency differences together with other information is analyzed with statistical techniques, if one allele or haplotype appears more often in one group (cases or controls) than the other, this variant is suggested to be associated with an elevated risk of or protection against this disease. The association study design has greater power to detect smaller effects compared to linkage analysis, but requires many more markers to be examined. In recent years, with the identification of tightly spaced SNP variants through the human genome, the improvements in genotyping technology and associated cost reduction have made high resolution association studies practical, and given them greater power and resolution for gene mapping than linkage studies.

In the last few years, accumulating evidence has been found to reveal the universal existence of CNVs distributed in human genome and it has been suggested that CNVs may serve as a useful class of markers in genetic association studies (Redon et al., 2006; Sharp et al., 2006; Shelling and Ferguson, 2007).To carry out an association analysis using CNVs as markers, all levels of copy number should be measured. However, only a small percentage of common CNVs can yield genotypes of the quality that could be used for linkage disequilibrium analysis (Redon et al., 2006). And in the existing collection of CNVRs identified, only a very small number of CNVs have been genotyped (McCarroll and Altshuler, 2007). This is inadequate to establish a robust and comprehensive CNV marker map.

The current methodologies of CNV genotyping fall into two categories: use raw copy-number measurements such as $\log_2$R ratio (Stranger et al., 2007) or dichotomize over CNVRs and define the CNV variations as 'gain' or 'lose'. It has been argued that summarizing raw copy-number measurements into such 'calls' may lose information present in the original measurements and is of uncertain relationship to the true genotype (McCarroll and Altshuler, 2007).

The genetic variance underlying heritable traits related to complex disease can be partly explained by common variants of small effect, which form the basic principle of association studies: common-variant-common-disease hypothesis. However, the phenotypic variation will also have a contribution from rarer variants. A spectrum of the observed allelic frequency of 22 functional quantitative trait nucleotide (QTN) variants influencing the trait (Figure 1.7), obtained in various resequencing studies, suggests that rare variants (with frequencies<0.05) may be very important causal factors in quantitative trait variation (Blangero, 2004). On the other hand, it is shown in an L-shaped or exponential distribution of mutation effect sizes that there are many variants with small effects, a small number with intermediate effects and relatively few with large effects (Wright et al., 2003). It could be argued that a collection of multiple rare variants could make a very significant contribution to human phenotypic variation.

However, the current popular methods to detect genetic variants have limitations when it comes to the case of multiple rare variants. Linkage studies conducted among families with multiple cases of disease were successful in identifying variants of large effects with high penetrance. Association studies conducted in general populations samples using common genetic markers typically find low penetrate variants with (very) small effects. It is not unexpected given that these common genetic variants are ancient and will have been subject to some selective pressure over time. Usually the rare SNPs are not in linkage disequilibrium (LD) with any other variants within the gene and are uncorrelated with any common haplotype, so the HapMap strategy base solely on tagging common SNP variants could easily fail even with very large sample sizes (Blangero, 2004).

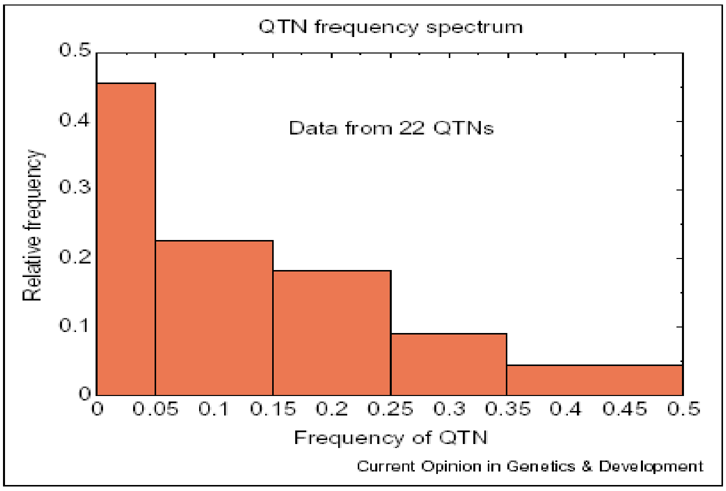


Figure 1.7 Observed allelic frequency spectrum from 22 QTNs obtained from 7 different resequenced genes (Blangero, 2004).

1.5.2 The study of CNV in human traits and diseases

1.5.2.1 Single causative CNVs in human traits and diseases

The first attempt to link structural variants to human disease was made by Lupski et al in the early 1990's. Their study revealed an association between a gene duplication on chromosome 17p and a common inherited neurological disorder, Charcot-Marie-Tooth (CMT) disease type 1A (Lupski et al., 1991). However, the publication of that pioneering study was not without difficulty. Both *Science* and *Nature* rejected Lupski's submission without even sending it out for review. At that time, "there was no appreciation that copy number was a mechanism of disease", said Lupski (Cohen, 2007). This study finally appeared in *Cell* later that year.

The study of association between gene copy number and CMT brought a refreshing new way of thinking to all geneticists. This discovery not only offered understanding to the etiology of the devastating disease but it also started the search for finding connections between genetic disease and a wider variety of genetic variation, which can be anything from a single base pair to very long stretches of DNA on chromosomes. This marked the opening of a new era in human genetics.

Since then, alongside the progress of technology to detect structural variants, more and more scientific groups studied the CNV-disease association and several fruits were presented in this field. An important discovery of CNV association in 2005 should not be missed out here. Gonzalez et al. discovered significant interindividual and interpopulation differences in the copy number of a duplication influencing CCL3L1 gene, and found out lower copy of CCL3L1 was associated with higher HIV/AIDS susceptibility (Gonzalez et al., 2005). This became a classical example of CNV-disease association via gene dosage effect in the early years.

1.5.2.2 Genome-wide CNV association

Genome-wide association studies (GWAS) have been proposed as a powerful strategy to detect potential genes which may account for complex disease outcomes or phenotypes. The first genome-wide association study was carried out in 94 myocardial infarction patients and 658 controls, based on 92,788 SNPs (Ozaki et al., 2002). Two SNPs in the lymphotoxin-alpha (LTA) gene were found to be significantly associated with myocardial infarction. One of them has a functional impact on transforming amino acid residual from threonine to asparagine (Thr26Asn) while another in intron 1 of the LTA gene influences the level gene transcription. Since then, the number of GWAS has grown nearly exponentially. To date, dbGaP by NCBI (Database of Genotypes and phenotypes, http://www.ncbi.nlm.nih.gov/gap, last accessed in March 2011) categorized 5684 analyses on 124656 clinical variables from 132 studies. Numerous novel genetic loci underlying disease susceptibility have been discovered, and many of these associations hold up to rigorous standards for replication. NHGRI's 'A Catalog of Published Genome-Wide Association Studies' listed 817 publications which attempted to assay at least 100,000 SNPs in the initial stage, as of the time of writing. In total, 3998 SNPs were reported to be significantly ($P<10^{-5}$) associated with disease or disease-related traits (http://www.genome.gov/gwastudies/).

A database called DECIPHER (Database of Chromosomal Imbalances using Ensembl Resources) was developed to catalog CNVs identified by array CGH which links to disease, using a variety of bioinformatics applications (https://decipher.sanger.ac.uk/application/syndrome/). To date, 59 syndromes for over 4200 consented patients investigated in more than 150 institutions worldwide were categorized in DECIPHER. These diseases included Charcot-Marie-Tooth syndrome (CMT), Adult-onset autosomal dominant leukodystrophy (ADLD), Miller-Dieker syndrome, Angelman syndrome and many more.

SNP genotyping platforms are widely used to construct CNV genotypes that are used in association and linkage studies, to map chromosomal regions containing genetic variants account for complex phenotypes and diseases in GWAS. A number of studies have used SNP array to detect CNVs and performed association studies between CNVs and disease outcome. Table 1.4 lists some CNV GWAS with positive results between 2005 and 2009. Only one association of a common CNV reached significance out of all the 26 studies (Bae et al., 2008); most studies with positive findings claimed a general enrichment of CNVs in cases, especially *de novo* or rare CNVs. It should be noted that although CNVs are ubiquitous in the human genome, the frequencies of this type of genetic variation are often observed to be low, therefore only a small portion of CNVs can reach acceptable marker allele frequency for GWAS. For the rare ones only the general enrichment could be tested; more analysis and experiments are needed to address the biological roles of each rare CNVs.

Table1.4 Examples of genome-wide CNV association studies. Associated risk: "Multiple CNVs" indicates overall CNV burden is associated with disease.

Disease	Associated condition	Chromsome	Gene or location	Associated risk	Study type	Genotyping platform	Reference
Autism spectrum disorders	ascertainment		neurexins	Multiple CNVs	Familial	Affymetrix 10K	Szatmari et al., 2007
	ascertainment	Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 550K	Glessner et al., 2009
	ascertainment	Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 500K	Marshall et al., 2008
		16	16p11.2	Duplication			
Schizophrenia	ascertainment	22	22q11.2	Hemizygous deletion	Case control	Affymetrix 250K	Bassett et al., 2008
	acertainment	Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 5.0/6.0	ISC, 2008
	acertainment	Multiple	Multiple	Multiple CNVs	Case control	Illumina 550K	Kirov et al., 2009
		17	17p12	Deletion			
		22	22q11.2	Deletion			
	acertainment	11	chr11:112772031-	Deletions in this region	Case control	Illumina 550/610K	Need et al, 2009
	acertainment	1	1q21.1	de novo deletions	Case control	Illumina 300/550K, Affy 6.0	Stefansson et al., 2008
		15	15q11.2				
		15	15q13.3				
		Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 250K	Vrijenhoek et al., 2008
	acertainment	Multiple	Multiple	Novel SVs	Case control	Affymetrix 500K	Walsh et al., 2008
Bipolar disorder	acertainment	6	6q27	Duplication	Familial	Illumina 550K	Yang S et al., 2009
		9	9q21.11	Duplication			
		12	12p13.31	Duplication			
		15	15q11.2	Deletion			
	acertainment	Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 6.0	Zhang et al., 2009
Autosomal recessive mental retardation	ascertainment	8	TUSC3	Homozygous deletion	Familial	Affymetrix 250K	Garshasbi et al., 2008
Amyotrophic lateral sclerosis	acertainment	Multiple	Multiple	Heterozygous deletion	Case control	Illumina 300K	Blauw et al., 2008
Blepharophimosis-Ptosis-Epicanthus inversus syndrome	ascertainment	Multiple	Multiple	Multiple CNVs	Patients	Affy262K, Illumina 300/370K	Gijsbers et al., 2008
Myelodysplastic/myeloproliferative disease	survival	Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 250K	Gondek et al., 2008
	ascertainment	4	TET2	Various deletions	Patients	Affymetrix 250K	Langemeijer et al., 2009
	acertainment	5	del(5q)	Deletion	Case control	Affymetrix 50K	Wang L et al., 2008
		Multiple	Multiple	Multiple CNVs			
Autosomal recessive juvenile nephronophthisis	acertainment		NPH1	Heterozygous deletion	MR patients	Affymetrix 100K	Hoyer et al., 2007
Systemic lupus erythematosus	acertainment	6	C4	Deletion	Case control	unkown	Kamatani et al., 2008
Osteoporosis	acertainment	4	4q13.2	Deletion	Case control	Affymetrix 500K	Yang T et al., 2009
Acute myeloid leukemia	survival	Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 250K	Gondek et al., 2008
Subarachnoid aneurysmal hemorrhage	acertainment	14	14q31.1	Heterozygous deletion	Case control	Illumina 300K	Bae et al., 2008
Li-Fraumeni syndrome	acertainment	Multiple	Multiple	Multiple CNVs	Case control	Affymetrix 250K	Shlien et al., 2008
Renal cell carcinomas	expression level	Multiple	Multiple	Duplication	Gene expression	Affymetrix 100K	Cifola et al., 2008
Neuroblastoma	expression level	Multiple	Multiple	Duplication	Gene expression	Affymetrix 100K	Fix et al., 2008

It can be predicted that resequencing studies will identify many variants, including rare variants of intermediate effect associated with common complex disease. This paradigm shift has already begun with the seminal work of Cohen et al., who compared non-synonymous sequence variations in individuals at the extremes of the population distribution of HDL-cholesterol levels, and determined that a significant fraction of genetic variance is due to multiple alleles with intermediate effects that are present at low frequencies in the population (Cohen et al., 2004). Furthermore, genotyping only the individuals with extreme phenotypic values is proved to greatly increase power while reducing cost, in linkage and association studies(Abecasis et al., 2001). Romeo et al. has demonstrated that targeting extremes is a powerful strategy to identify rarer variants (Romeo et al., 2007). Until many more such studies are reported it would be premature to decide on the relative importance of the common variant-common disease model and the alternative rare variant-common disease model which states that disease susceptibility to common diseases is the result of multiple low frequency/rare variants with larger phenotypic effects. Although individually rare, these variants may be collectively common in the population (Cohen et al., 2004).

1.5.3 The ways that CNVs influence phenotypic variation

Genetic changes at the DNA level could alter gene expression and eventually confer phenotypic effects. CNVs discovered as deletions, insertions, duplications and complex multi-site variants often span thousands of base pairs, therefore they can potentially influence gene expression. Stranger et al. surveyed the impact of CNV on expression patterns by examining mRNA levels in lymphoblastoid cell lines from 210 HapMap individuals from four ethnic groups (Stranger et al., 2007), with knowledge of the CNVs from the same set of samples (Redon et al., 2006). Copy number changes were found to account for 8.75% of variation at expression levels of 972 genes.

Effects of some CNVs on gene expression were found in all four ethnic groups, while some were population specific. More than half of the expression probes associated with CNVs were away from the CGH clone harboring the CNV, some were as far away as 2Mb apart, indicating distant regulation (Stranger et al., 2007). The major molecular mechanisms suspected include: 1) gene dosage effect, 2) interruption of a gene, e.g. interrupting protein coding sequences, 3) influencing regulatory sequence, 4) gene fusion, 5) unmasking mutations or functional SNPs in the remaining allele.

The effect of CNVs on dosage-sensitive genes is the most prominent. For instance gene copy number of the human salivary amylase gene (AMY1) can vary from 2 to 15. Populations which bear a diet habit to consume more starch were found to have higher AMY1 copy numbers. A correlation of levels of mRNA and AMY1 protein level for different copy numbers was observed; it was therefore argued that this CNV locus influence gene expression at both transcriptional and translational levels (Perry et al., 2007).

CNVs may disrupt protein coding sequences of a gene to cause functional loss or modification of that gene. One example comes from a schizophrenia study, where genome-wide CNVs were determined in cases and controls. Multiple CNVs were found to be enriched in schizophrenia patients, including a deletion of 400 kb in size which disrupts the ERBB4 (receptor tyrosine-protein kinase erbB-4) gene (Walsh et al., 2008). cDNA ends amplification experiment confirmed absence of erbB-4's receptor intracellular kinase domain in the mutant transcript by the authors (Walsh et al., 2008).

CNV can also have a long range position effect on gene expression. Velagaleti et al. discovered a pair of translocations whose breakpoints was 900 kb upstream and 1.3 Mb downstream of SOX9 gene, respectively, could cause compomelic dysplasia, a disease which has been proved previously to be associated with mutations within SOX9 (Velagaleti et al., 2005).

Genomic rearrangements can cause fusion of different genes or their regulatory sequences, thus generating a gain-of-function mutation. This mechanism is prominently observed in cancers associated with specific somatic chromosomal translocations (Zhang et al., 2009), and also found in disease studies. Glucocorticoid-remediable aldosteronism (GDA) is an autosomal dominant disorder which has syndrome of hypertension with variable hyperaldosteronism. NAHR causes gene fusion of two GDA candidate genes on chromosome 8q, one encoding aldosterone synthase and another coding steroid 11 beta-hydroxylase. This fusion can account for hypertension in animals and humans (Lifton et al., 1992).

Hemizygous deletion at a locus may diminish one allele and unmask another recessive allele or functional polymorphism. For example, the plasma coagulation factor 12 (FXII) is a gene that underlies Sotos syndrome. The activity of FXII is low in normal individuals but high in individuals who have a single deletion at this loci, remaining only one copy of FXII allele (Kurotaki et al., 2005).

1.5.4 CNV and evolution

CNVs can influence human traits and disease susceptibility, therefore they can be potentially exposed to selection pressure during evolution. Both positive and negative selection on CNV have been discovered in humans and other species (Zhang et al., 2009).

CNVs are usually found to be located out of functional sequences in human genomes, which suggests purifying selection on CNVs. Conrad et al. (2006) investigated the SNP density within deletions in coding sequence and introns, and observed a strong underrepresentation of SNPs in deletion regions compared with the HapMap average. Many of the genes containing deletions had disease-associated OMIM (Online Mendelian Inheritance in Man) entries (Conrad et al., 2006). The gene-poor phenomenon in CNV regions and the lower number of deletions than duplications that overlap disease-related genes are also confirmed in a number of other studies (Nguyen et al., 2008; Redon et al., 2006).

Gene duplication has long been considered to be a main mechanism driving positive selection. The variable copy of AMY1 in different human populations in response to changing diet is a good example (Perry et al., 2007), a duplicated gene and its regulatory elements could be modified for new functions, resulting diversification in a species.

CNVs could also be under reduced purifying selection, resulting in more variants in nonessential genes. The frequency of neutral CNVs will fluctuate under genetic drift. There is good evidence suggesting that CNVs significantly enrich in regions with genes that respond to the environment, such as sensory perception and immunity (Redon et al., 2006). Such variants can arise and remain in the genome by reduced purifying selection as in an unstable or changing environment so selective advantage of any one copy number state may vary (Nguyen et al., 2008; Redon et al., 2006).

1.6 CNVs at familial and population level

CNV should behave in just the same way as other genetic variants, in terms of segregation between individuals across generations. Redon et al. investigated heritability of 67 bi-allelic CNVs in 90 HapMap parents-offspring trios (30 trios each from three human populations). They showed only 0.2% of the CNVs exhibited Mendelian discordance; they argued the small proportion probably reflected genotyping error rate rather than the rate of de novo events at these loci (Redon et al., 2006). In a study of association between CNVs and schizophrenia, Xu et al. detected CNVs in parents-offspring trios in 200 affected families and 152 control families. Out of the total 11,268 unique CNVs identified, only 19 were *de novo* (Bademci et al., 2010; Xu et al., 2008). Due to the complex inheritance of CNVs, SNPs located within CNV regions often display inconsistency of Mendelian inheritance or are not in Hardy-Weinberg equilibrium. This special behavior of markers has been used to successfully identify polymorphic deletions and inversions (Conrad et al., 2006; Hinds et al., 2006; McCarroll et al., 2005).

Studying CNV at population level is useful in a number of respects: first, recurrent CNVs detected in a large number of individuals can be used to construct a human CNV map; second, CNVs provide detailed information about haplotype structure that could facilitate the selection of tagging SNPs for use in association studies; last but not least, cataloging CNV frequency, distribution and map location between different populations may help geneticists to understand human migrations, recent natural selection and evolution of recombination hotspots. The HapMap data derived from samples from four ethnic groups (European, African, Asian-Chinese and Asian-Japanese) was first to be used to demonstrate variation of copy number at population level (Redon et al., 2006). Later on, a couple of studies extended high-resolution surveys of variation in genotype, haplotype and copy number to a wider range of human population groups.

Jacobsson et al. (2008) surveyed 396 CNV loci in a worldwide sample of 29 populations. They found the frequencies of CNVs were generally low (only one CNV frequency exceeded 10%), and CNVs private to one population were more common than private SNP alleles. Partial similarity was found between population structure inferred for CNVs and that inferred from SNP and haplotype data sets (Jakobsson et al., 2008). Li et al. discovered and characterized 3900 CNVs from 985 Caucasians and 692 Asians. Many CNVs showed significant ethnic differences in frequencies (Li et al., 2009). However, comparisons of CNVs in multiple populations are still rare. First, there are few well designed population genetics studies, which have data available for CNV analysis and at the same time involve multiple populations with family information; second, the discrepancies in study design, platform choice and analytical methods between studies, leads to difficulties when merging CNV information from different studies each focus on different populations. Here, I take advantage of data collected from three different study cohorts, each from an isolate population, all with family data and all genotyped with the same SNP GWAS panel. This has allowed me to undertake a captive analysis of CNV frequency and distribution within and between cohorts.

1.7 Aims of this thesis

The aim of this thesis is to a) address the advantages and limitations of current copy number variants (CNV) detection methods, b) apply and compare alternative calling algorithms for CNV detection in three European populations from Illumina 300K whole-genome genotyping data, and c) investigate association between CNVs and quantitative traits in the study samples. This thesis comprises the following experimental sections:

In Chapter 3, I demonstrate the extent of copy number variation between individuals. Using downloaded information from five human individuals who had their genomes sequenced, I compared CNVs called directly from whole-genome sequencing data for these five genomes. I also examined the characteristics of these CNVs and discuss limitations of the methods which categorize CNVs from raw DNA sequence data.

In Chapter 4, I investigate population level CNV profiles, drawing on published studies, focusing on those that detected CNVs from whole genome SNP genotyping data. A structured review was constructed to retrieve CNVs identified specifically in HapMap samples, from various types of genotyping platforms and using different calling algorithms. Two HapMap samples were in common to six of the identified studies and used to evaluate the impact of platform choice on CNV calling. Also the CNVs detected in SNP genotyping studies were compared against those identified by another physical mapping technique on the same individuals, to address the robustness and / or limitations of CNV detection methods using SNP genotyping data.

Chapter 5 compares CNVs called by four different algorithms from whole-genome genotyping data, for a subset of our study samples. The false positive and false negative rates of each algorithm were estimated from the extent of overlap in CNV detection and the relative performance of these four algorithms evaluated.

In Chapter 6, CNVs of individuals from three European population isolates were determined, using algorithms chosen on the basis of method evaluation concluded in Chapter 5. The characteristics of these CNVs are described, and genetic difference at population level in context of copy number difference will be examined. Also I investigate the inheritance of CNVs with the knowledge of pedigree information for the three study populations.

Chapter 7 comprises an analysis of association between CNVs and seven metabolic-related quantitative traits. The overall burden of CNVs on metabolic phenotypes is assessed. Regression analysis was performed for common CNV association and for the rare CNVs, to test whether there is an enrichment of rare CNVs in metabolic pathways in individuals with extreme trait values will be explored.

Chapter 2

Materials and Methods

2.1 Study populations

The study samples come from three population genetics projects: CROAS in Island of Vis, Croatia, ORCADES in Orkney Isles, UK and MICROS in South Tyrol, Italy, (Figure 2.1) which are under the banner of a large collaboration project, EUROSPAN.

2.1.1 EUROSPAN

The EUROSPAN project (http://homepages.ed.ac.uk/s0565445/index.html) was initiated in 2006 and involves five population isolates from Italy (MICROS), Croatia (CROAS), Scotland (ORCADES), Sweden, and the Netherlands. The project aims at assessing the genetic structure of European isolates and at identifying genes underlying common traits, taking advantage of the genetic and environmental homogeneity that usually characterizes population isolates. In the current context, population isolates are "*secondary isolates*", i.e. groups that, for some reasons, detached or were detached from larger populations. In particular, EUROSPAN cohorts were derived from small population samples which have grown slowly, with little influx from outside the groups.

As inclusion criteria these populations had to a) represent local populations occupying
a well defined geographic area and have limited exchange with surrounding populations enabling the inclusion of large family groups, b) have historical records enabling their genealogy and migration patterns to be defined, c) span the geographic range and environmental diversity inhabited by European populations. The Scottish, South Tyrolean and Croatian populations are small populations founded by a limited number of individuals and/or have undergone a population bottleneck, followed by long isolation and very low immigration (Marroni et al., 2006).

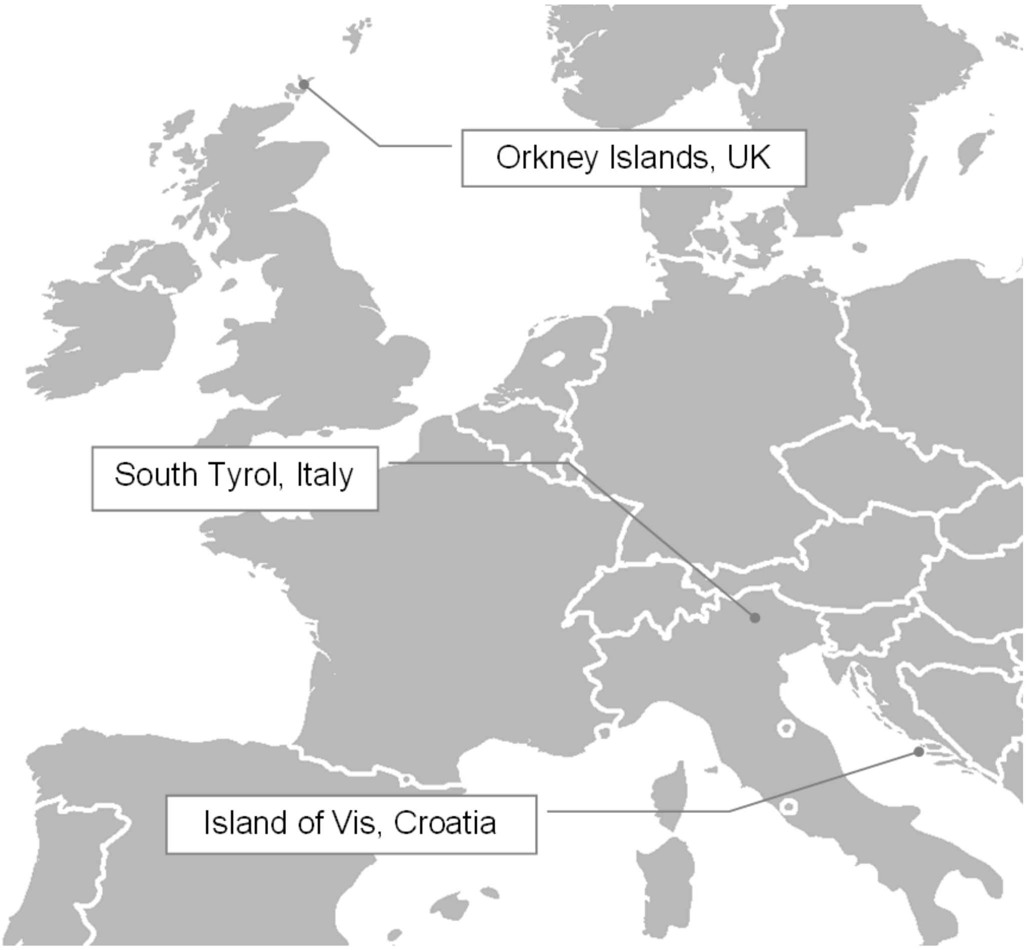


Figure 2.1 Geographic distribution of study samples

2.1.2 The Dalmatian samples: CROAS study

The Croatian Anthropogenetic Study (CROAS) is funded by the Medical Research Council (MRC) and involves the Universities of Zagreb and Split, the Institute for Anthropological Sciences, Zagreb, the University of Edinburgh and the MRC Human Genetics Unit, Edinburgh.

The research focuses particularly on the isolated population on the island of Vis. Residents from the villages Komiza and Vis on Island of Vis were asked to volunteer and give their consent to take part in the study. The Komiza survey was carried out in May 2003 and the Vis survey in May 2004. 1062 volunteers were recruited (584 from Komiza and 478 from Vis), which represented a very high proportion of the permanent resident adult population (65-70%). Data were successfully recorded for 1030 of these individuals. The volunteers ranged from 19 to 93 years of age and all gave informed consent (426 males and 604 females). All participants were asked to give some basic family information, such as the names of their parents and any siblings. More extensive genealogical data were extracted from the Komiza and Vis parish registers which dated back to 1838. Using both sources of information, pedigrees were constructed. 134 participants could be joined up into a single pedigree. 588 of phenotyped and genotyped individuals could be placed in 125 pedigrees, of which the largest pedigree links 134 individuals. The remaining individuals were singletons (those who were unable to be connected to any relatives).

2.1.3 The Orcadian samples: ORCADES study

The Orkney Complex Disease Study (ORCADES) is an ongoing, family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in the population isolate of the Orkney Isles in northern Scotland. The North Isles of Orkney, the focus of this study, consist of a subgroup of ten inhabited islands with census populations varying from ~30 to ~600 people on each island. The North Isles have experienced a period of severe population decline over the last 150 years, fueled by high emigration and low fertility. The population fell from an estimated peak of 7700 in the 1860s to 2217 by 2001. Endogamous marriage was widespread during the nineteenth century and into the twentieth century. The combined effects of steep population decline and endogamy have led to inflated levels of parental relatedness in the current population. Data collection was carried out in Orkney between 2005 and 2007.

2.1.4 The South Tyrolean samples: MICROS study

Samples from South Tyrol were collected as part of the Genetic Study of Three Population Microisolates in South Tyrol (MICROS) from settlements in Venosta Valley. The MICROS study was an extensive survey carried out in Val Venosta (South Tyrol, Italy) in 2001. Participants were from three isolated villages located in the Italian Alps, in a German-speaking region bordering with Austria and Switzerland. Due to geographical, historical and political reasons, the entire region experienced prolonged isolation from surrounding populations. Information on participants' health status was collected through a standardized questionnaire. Laboratory data were obtained from standard blood analyses. Initially 1175 people were enrolled in the study, which has subsequently risen to approximately 1400.

2.2 Ethical approval and consent

Ethical approval was given for the patient recruitment in Vis, Orkney and South Tyrol by the relevant Research Ethics Committee of the Faculty of Medicine, University of Zagreb, Croatia and Local Research Ethics Committee of NHS Lothian; the Local Research Ethics Committee of NHS Orkney and the North of Scotland Research Ethics Committee in Aberdeen; and the Local Research Ethics Committee South Tyrol, respectively. In all three sites, volunteers gave written informed consent to all parts of the study, with the research medical doctors or research nurse or research co-ordinator present to answer questions. They were made aware that they need not take part in all parts of the study and were free to withdraw at any time without consequences for them. In Orkney and Tyrol, volunteers chose whether to consent to their family doctor being contacted in the event of incidental findings coming to light (Mascalzoni et al., 2009).

2.3 Phenotyping

Clinical history questionnaires were filled out and core quantitative traits were measured for participants in Vis and Orkney. Clotted blood was obtained and the following traits were a subset of those measured: height, weight, body mass index (BMI), waist and hip circumference, subscapular skinfold thickness and suprailiac skinfold thickness. Levels of fasting fasting glucose and insulin were measured in the NHS Orkney laboratories and laboratory of Dr Salzer from the University of Zagreb, for Orcadians and Dalmatians, respectively.

2.4 Genotyping

Fasting blood samples were taken from all participants (EDTA blood to be used for DNA extraction and clotted blood for serum biochemistry). DNA samples were genotyped according to the manufacturer's instructions (http://www.illumina.com/) on Illumina Infinium HumanHap300v1 for Dalmatian samples and on Illumina Infinium HumanHap300v2 for Orcadians and South Tyroleans, by technicians at the Wellcome Trust Clinical Facilities in Edinburgh (for Vis and Orkney samples) and University of Lübeck in Germany (for MICROS samples). The first and second version of HumanHap 300 arrays had 317,525 and 318,235 SNPs, respectively. SNPs with >10% missing genotypes were excluded. After data cleaning, 308,300 and 318,237 autosomal SNPs remained for version 1 and version 2 arrays, respectively.

Samples in Vis and Orkney with a call rate below 95% and those in South Tyrol with a call rate below 98% were excluded from the analysis. Sex checking was performed with PLINK, and individuals with discordant pedigree and genomic data were removed. In total, on completion of data-cleaning and quality-control procedures, SNP genotyping data from a total of 2861 individuals from three populations were available.

2.4.1 Infinium II Whole Genome Genotyping Assay

The Infinium^{TM} II Assay enables flexible SNP selection using a tagSNP content strategy and provides even coverage across the genome (http://www.illumina.com/). The Infinium^{TM} II procedure takes three days to complete and has four automated steps that are all performed on a Tecan robot system: (1) whole genome amplification, (2) hybridization to an oligonucleotide probe array (performed offline in a hybridization oven), (3) array-based SNP scoring assay, and (4) signal amplification. For the Infinium II assay, one bead type is used per SNP and the alleles are scored by SBE (Single-Base Extension) using differentially-labeled terminators (Figure 2.2).

Illumina manufactures several high density tag SNP genotyping arrays including Sentrix® HumanHap300 BeadChip. BeadChips are constructed by random assembly of bead pools into micro-well patterned stripes on a silicon substrate.

Each stripe is loaded with a unique bead pool composed of tens of thousands of different bead types for a total complexity of hundreds of thousands of bead types across the BeadChip. Each bead type is immobilized with a decoding sequence and a locus-specific 50-mer oligonucleotide probe. The identity is determined by a hybridization-based decoding procedure. These BeadChips utilize the single-base extension (SBE) Infinium II assay to genotype tag SNPs selected from Phase I and II of the HapMap project. The median SNP spacing on the

HumanHap300 is 5 kb and the HumanHap550 is 2.8 kb, enabling an effective resolution of ~50 kb and ~28 kb (10 SNP smoothing), respectively. (Gunderson and Peiffer, 2006)

2.4.2 BeadChip Imaging

After completion of the assay, the BeadChips are scanned with a two-color confocal Illumina BeadArray™ Reader at a 0.84–1.0-µm pixel resolution. Image intensities are extracted and genotypes are determined using Illumina's BeadStudio software. Aberrations were detected by visualizing SNP-CGH data in the Illumina Genome Viewer and Chromosome Browser.

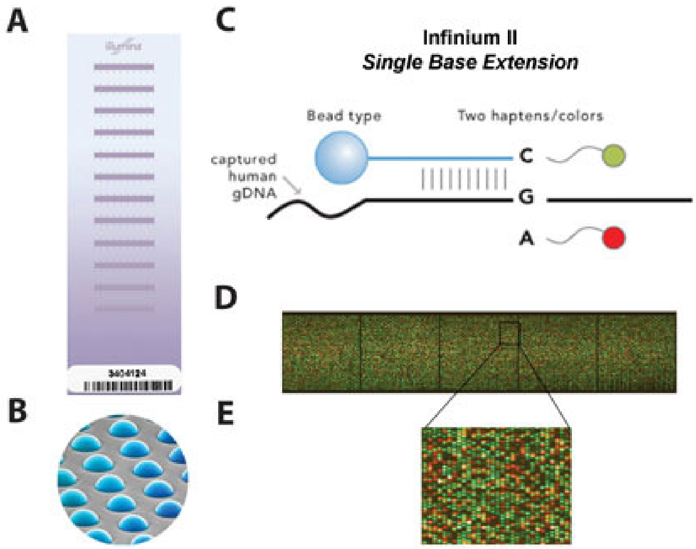


Figure 2.2 Infinium II Single-Base Extension. (Adapted from (Gunderson and Peiffer, 2006)

2.4.3 Analyzing SNP-CGH Data in BeadStudio

Genotyping data consist of two channel intensity data corresponding to the two alleles, allele A and allele B. The raw intensity of allele A and B (X and Y, respectively) of each SNP are imported to BeadStudio software (Illumina Manual 1) and normalized using a proprietary algorithm. The normalized X and Y are then combined as raw A versus raw B allele intensities to produce a value called "normalized R".

Next, the normalized R intensities were compared to 120 normal reference HapMap individuals with GenTrain software. Finally, the data were converted to polar coordinates R and Theta (Fig. 2.3). Theta (θ) represents the angle deviation from pure A signal, where 0 represents pure A signal and 1.0 represents pure B signal. It can be calculated by the equation: θ= (2 / ∏)* arctan(B/A). Each genotype has a specific expected cluster position from the reference data set as in figure 2.3. Each SNP genotype of each individual was determined according to its relation to these clusters. Values which lie between these clusters were discarded and labelled as NC (no call). Theta was then transformed to allele frequency (e.g. B Allele Frequency), which is more discriminative, using linear interpolation of the canonical clusters (Gunderson and Peiffer, 2006).

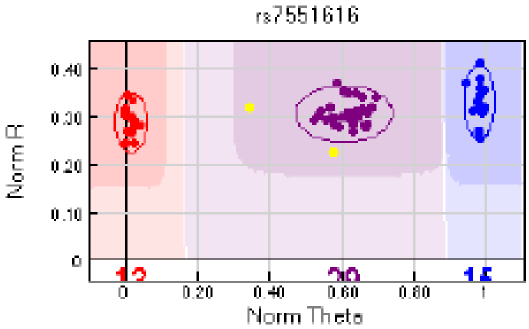


Figure2.3 Polar plotting of genotyping clusters. The AA genotype lies near the zero theta value and BB near theta of 1, with AB in between both genotypes. The yellow dots indicates an excluded individual with values lie away from expected cluster positions

2.5 CNV detection

Copy number variation was determined for study samples using genotyping data on the Illumina platforms. Log R ratio (LRR) and B allele frequency (BAF) at each SNP loci are two important types of values for CNV detection. LRR and BAF information was analyzed with various methods in the current study, which included Hidden Markov Model (HMM) based algorithms QuantiSNP and PennCNV, and Circular Binary Segmentation (CBS) based algorithms DNAcopy and cnvPartition. QuantiSNP, cnvPartition and DNA copy were run on a desktop computer with a 2.76GHz processor and 4GB of RAM. PennCNV were performed on a computer of 1.77GHz processor and 3GB of RAM.

2.5.1 Log R Ratio and B Allele Frequency

The comparison of normalized intensities between a reference and subject sample is the foundation of traditional array-CGH. SNP-CGH is different in that a combination of two genotyping parameters is analyzed: normalized intensity measurement and allelic ratio. Collectively, these parameters provide a more sensitive and precise profile of chromosomal aberrations.

Illumina has developed two modes of SNP-CGH analysis. The first is a single sample mode in which reference values are derived from canonical genotyping clusters created from clustering on normal reference samples. This mode is often used in general copy number detections, and is the mode chosen for the analysis in this thesis. The second is a paired sample mode in which direct intensity comparisons between a subject sample and its corresponding matched pair are performed. This mode is often applied in copy number change discoveries in paired tumor and normal samples.

Normalized intensity measurement, log R ratio (LRR), and allelic ratio, B allele frequency (BAF), are two key parameters to be considered in detecting chromosomal aberrations. LRR is the log, base 2, of the normalized signal intensity (Normalized R) of subject versus normalized signal intensity (Normalized R) of the reference. An LLR value of zero indicates no difference between individual and reference values. A significant departure from the zero value is an indicator of copy number changes (a gain if the R ratio is greater than one and a loss if the R ratio is less than one). BAF is the ratio of intensity at the B allele versus that at the A allele; for example BAF of 0 indicates an absent B allele at this locus (genotype AA or A_ etc.). The combination of the two parameters provides information to predict copy number change (Figure 2.4): LRR indicates overall copy number of alleles at a locus and BAF indicates which allele is deleted or amplified. The effect size that is required to depart from normality is not a constant; it changes with different numbers of the SNPs involved in a specific window size.

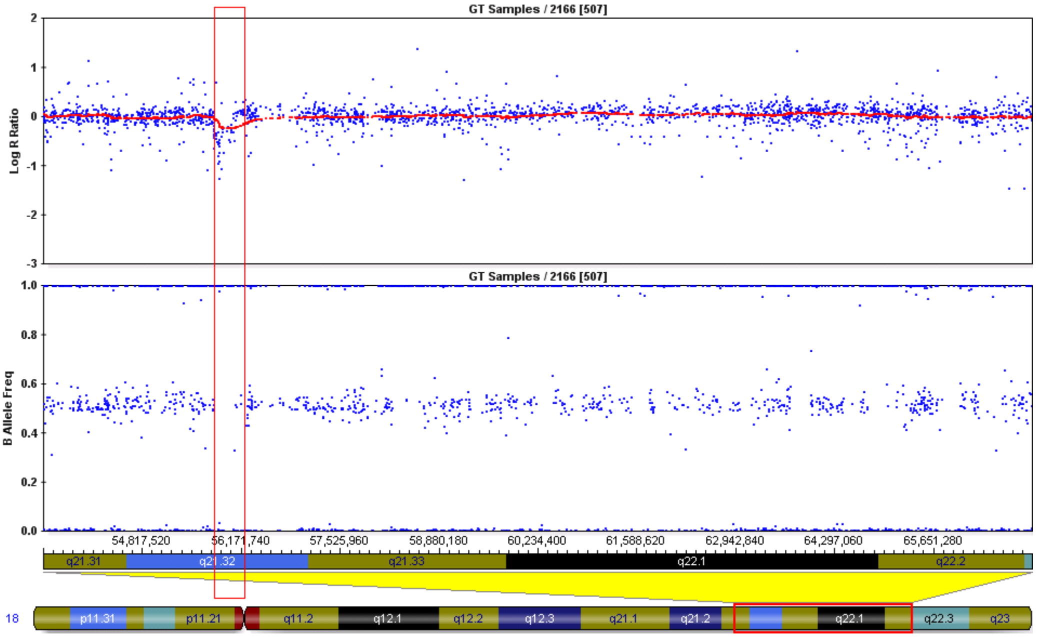


Figure 2.4 Shift in log R ratio (LRR) and B allele frequency (BAF) value plots infer copy number change. LRR (top) and BAF (bottom) plotted from one individual for each SNP on chromosome 18. A deletion on the q arm can be identified by the shift in the LRR downwards and the loss-of-heterozygosity indicated by the disappearance of heterozygous state (0.5) in the BAF (as indicated by the area within the red rectangle). Chromosome location coordinates in hg18 (Build 36).

2.5.2 Hidden Markov Model based CNV detection algorithms

2.5.2.1 Markov Chain and Hidden Markov Model

A Markov Chain, named after Russian mathematician Andrey Markov, is a discrete random process with the Markov property. A discrete random process is a system which can be in various states, and which changes randomly in discrete steps. The random process has the Markov property if the conditional probability distribution of future states in the process depends only upon the present state, in other words, given the present, the future only depends on the present state but does not depend on the previous states (Rabiner, 1989).

Since the Markov property is a simple and mathematically tractable relaxation of the assumption of independence, it is natural to consider discrete-time Markov Chains on a finite state space as possible models for sequential data, which are not independent from each other, in that space.

The Markov Chain can be represented in terms of a graphical model, as shown in Figure 2.5 Each node represents a random variable, and the edges indicate conditional dependence structure. $X_0, X_1, \ldots X_{n-1}, X_n$ are a sequence of random variables on a Markov Chain, which are all with the Markov Property; i.e. given the present state, the future and past states of each variable are independent:

$$\Pr(X_{n+1} = x | X_1 = x_1, X_2 = x_2 \ldots, X_n = x_n) = \Pr(X_{n+1} = x | X_n = x_n).$$

The behaviour of a Markov Chain resembles a random walk on the graph shown in Figure 2.5.The possible values of X_i comes from a finite set S, which is called the “state space” of the chain. The changes of state on the chain are called transitions (represented by the horizontal arrows in Figure 2.5.), and the probabilities associated with various state-changes are called transition probabilities. The transition probability matrix of the Markov Chain is denoted as T. Suppose X_0 is drawn from a distribution λ. Initially X_0 is chosen according to λ at that state S_{x0}; at time t the current position is S_{xt}. From any position (state) there are two possible transitions, either to the next integer or to the previous integer. The transition probabilities at S_{xt}, denoted as T_{xt} (the X_t-th row of the transition matrix T), depend only on the current state; therefore the values of X_{t-1} and X_{t+1} are chosen only with respect to T_{xt}, regardless of any prior positions.

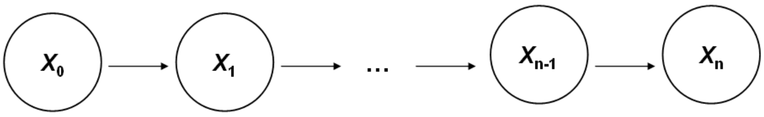


Figure 2.5 Graphical illustration of a Markov Chain

λ and T are important parameters of a Markov Chain, by studying λ and T one can identify many properties of Markov Chain. For example, the distribution of X_0 is determined by λ, while the distribution of X_1 is determined by λT_{x1}, etc.

In statistics, a Hidden Markov Model (HMM) describes a probability distribution over a potentially infinite number of sequences. The name ‘Hidden Markov Model’ comes from the fact that the state sequence is a first-order Markov chain, but only the symbol sequence is directly observed. It is an extension of a Markov Chain which is able to capture the sequential relations among hidden variables.

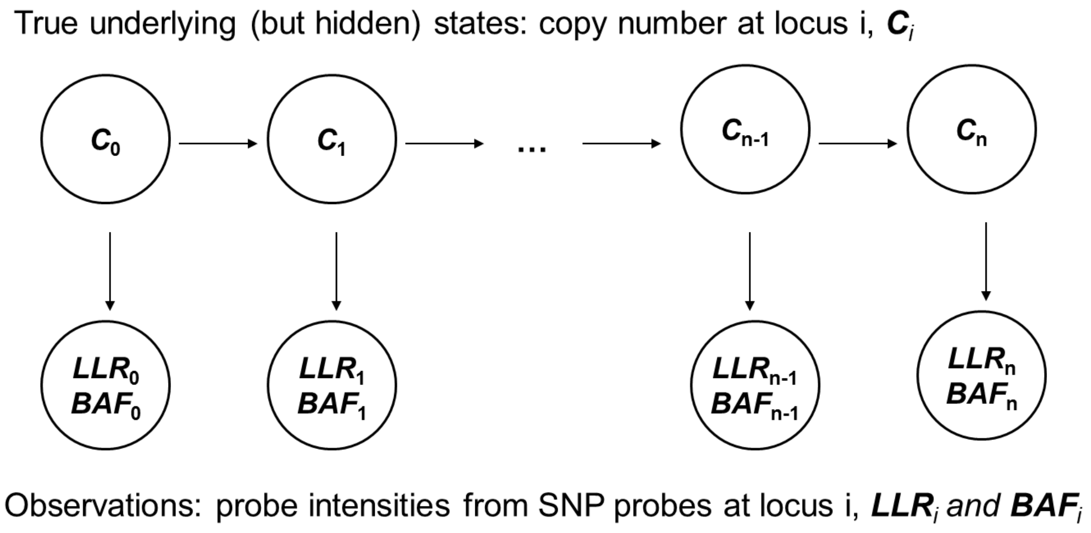

True underlying (but hidden) states: copy number at locus i, C_i
C_0
C_1
...
C_{n-1}
C_n
LLR_0 BAF_0
LLR_1 BAF_1
LLR_{n-1} BAF_{n-1}
LLR_n BAF_n
Observations: probe intensities from SNP probes at locus i, LLR_i and BAF_i

Figure 2.6 Graphical illustration of a Hidden Markov Model

The illustration of a HMM is shown in Figure 2.6. $X_0, X_1, \ldots X_{n-1}, X_n$ is a Markov Chain and at time t, for t=0,1,…,n. Y_t is independent of all other variables given X_t. From the graphical model for HMM, we can easily see the conditional independence structure of all variables (X_0, Y_0),…, (X_n, Y_n). Here the observed sequences $Y_0, Y_1, \ldots, Y_{n-1}, Y_n$ are influenced by a hidden Markov chain $X_0, X_1, \ldots X_{n-1}, X_n$; and the observed sequences are used to infer the hidden sequences. The horizontal arrows represent state transition on the Markov Chain. The vertical arrows represent the relationships between each pair of X_t and Y_t , called emission probability. The emission probability matrix is denoted as Γ (Zhang 2008).

HMM is widely applied in the detection of copy number variation from genotyping data. Theoretically, information at each SNP loci along the chromosomes can be considered as a node from a sequence of data. The hidden state is the true copy number of the individual's genome; the observed states are the normalized intensity measurements of each probe on the array (Figure 2.7). The emission probability of allelic intensities for an underlying hidden state of 2 copies (normal state), <2 copies (copy number loss) and >2 copies (copy number gain) is modelled and the modelling parameters are determined from calculations which are based on a training set of data. The transition probabilities between underlying copy number states is asserted such that transitioning out of a state reflecting 2 copies is low, while transitioning within the same state or returning to normal copy number is relatively high.

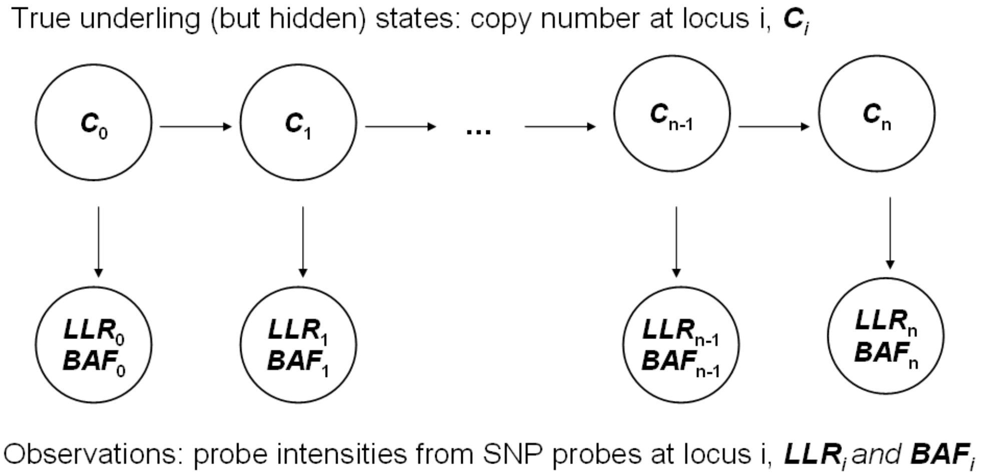

True underling (but hidden) states: copy number at locus i, C_i
C_0
C_1
...
C_{n-1}
C_n
LLR_0 BAF_0
LLR_1 BAF_1
LLR_{n-1} BAF_{n-1}
LLR_n BAF_n
Observations: probe intensities from SNP probes at locus i, LLR_i and BAF_i

Figure 2.7 HMM modelling in detection of copy number variation

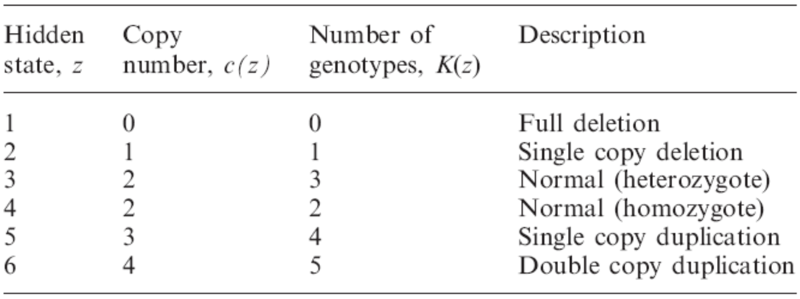
2.5.2.2 QuantiSNP:

QuantiSNP is an analytical tool to analyze copy number variation using whole genome SNP genotyping data. It was originally developed for Illumina arrays, but later versions of this software support Affymetrix data with additional data conversion steps (Collella et al., 2007).

QuantiSNP uses an Objective Bayes Hidden-Markov Model (OB-HMM) to automatically infer regions of segmental copy number abnormalities from genotyping data. The OB-HMM is claimed to be highly suited to the analysis of high-throughput genomic data when one of the hidden states has special status as a ‘null’ or normal state, in which case OB-HMM allows for setting of parameters which ensure certain frequentist coverage properties for excursions of the model out of the null state, while benefiting from Bayesian marginal inference.

Six hidden states used in HMM for QuantiSNP are listed in Table 2.1. *A priori* probability that hidden state change occurs between adjacent SNP loci (the transition probability) is defined by an exponential function. The emission probabilities that a set of LLR and BAF values of a SNP predict hidden states of the same SNP are defined as a mixture of Gaussian and uniform distributions. Most of the hyper-parameters in the above model are estimated via maximum marginal likelihood techniques on a training data set, then a expectation maximization (EM) algorithm is used to find maximum marginal *a posteriori* estimates for the parameters of the emission distributions, followed by a Viterbi algorithm which to compute the sequence of hidden states with highest probability. After corrected for type I error and multiple sample influence, the copy number of a DNA segment is determined with a Bayes Factor (BF), which is a measurement of confidence of the region being in hidden state in comparison to all other sequences in which no part of this region is in this hidden state. The higher BF indicates significance of events (Collella et al., 2007) (Table 2.2).

Table 2.1 Hidden states, associated copy numbers and biological interpretation



Hidden state, z	Copy number, $c(z)$	Number of genotypes, $K(z)$	Description
1	0	0	Full deletion
2	1	1	Single copy deletion
3	2	3	Normal (heterozygote)
4	2	2	Normal (homozygote)
5	3	4	Single copy duplication
6	4	5	Double copy duplication

Each hidden state z is associated with a given copy number c(z) and genotype number K(z). For each copy number there can be a number of genotypes. For example, for copy number3 there can be one of four genotypes {AAA, AAB, ABB, BBB}. The genotype number gives the number of components in the mixture distribution of B allele frequencies for that state.

Table 2.2 Combination of LRR and BAF indicates six hidden states of copy number

Copy number	Normal(2 copies): AA,AB,BB	Copy-neutral LOH(2 copies): AA,BB	Copy number gain		Copy number loss	
			3:AAA,AAB,ABB,BBB	4:AAAA,AAAB,AABB,ABBB,BBBB	0 copy	1 copy: A_, B_
Log2R Ratio	0	>0		<0		

B allele frequency	0(AA)
	0.5 (AB)
	1(BB)
		0(AA)
		1(BB)
			0(AAA)
			0.33(AAB)
			0.67(ABB)
				0(AAAA)
				0.25(AAAB)
				0.5(AABB)
				0.75(ABBB)
				1(BBBB)
					0.5	0(A_)
						1(B_)

QuantiSNP version 1.0 (Windows command line based) was downloaded from http://www. well.ox.ac.uk/~ioannisr/quantisnp/ after a licence agreement signed between University of Edinburgh Medical Genetics Section and University of Oxford, in 2008. Copy number analysis was carried out in QuantiSNP following instructions described in the user manual (Yau 2007). LLR and BAF data for each individual in the study sample were exported from BeadStudio and processed with R software. Individuals with standard deviation of LLR>0.3 were excluded. Parameter settings for QuantiSNP analysis were set as: defined length of a CNV—no more than 3,000,000 bp; maximum number of optimisation steps of expectation maximisation—25; correction for local GC content—yes; array data library—Illumina HapMap 550K (which is compatible with HapMap 300K array). After calculation, QuantiSNP outputs a list of CNVs with chromosomal location, assigned copy number and a Log Bayes Factor (LBF) for each detected segment. LBF threshold of 30 was selected; any CNVs with a LBF<30 were excluded.

2.5.2.3 PennCNV:

PennCNV is also an algorithm based on HMM. It was originally designed for Illumina assays, the later versions also support data from Affymetrix platforms. It incorporates multiple sources of information, including LRR and BAF at each SNP marker, the distance between neighbouring SNPs, the allele frequency of SNPs, and family trio information where available.

The modelling of six hidden states in CNV is the same as that in QuantiSNP (Table 2.1 and Table 2.2). The first-order HMM is used to predict copy number states at each SNP. The emission probabilities are underlined as that a set of LLR and BAF values of a SNP predicts hidden states of the same SNP. These are defined in a manner very similar to that in QuantiSNP, but uses uniform distribution to model both random fluctuation of signal measures in chemical assays and the possible genome misannotation and misassembly. The modelling and interpretation of LRR and BAF values for chromosome X is treated differently, because of the hemizygous state in males. The transition probability of hidden states is constructed differently from that in QuantiSNP, which incorporates unknown parameters to be resolved by a HMM learning process using the Baum-Welth algorithm. The Viterbi algorithm is used to infer the most likely path (state sequences for all SNPs along each chromosome) (Wang et al., 2007). PennCNV has a module to validate CNVs detected in the former steps using family information, specifically the parent-offspring trio information. However, only a small fraction of the individuals in the study sample (10.13%) could be placed in family trios. Also this process takes 5*5*5 CNV matrices to compute and is time consuming. For these reasons this module was not utilized in the CNV analysis in this thesis.

PennCNV was downloaded from http://www.openbioinformatics.org/penncnv/. To create a Perl environment on Windows system for PennCNV to be performed, ActivePerl was downloaded from http://www.activestate.com/activeperl and installed. CNV detection with PennCNV used individual LRR and BAF files exported from BeadStudio, and the analysis was performed following PennCNV tutorial (http://www.openbioinformatics.org/penncnv/penncnv_beadstudio_tutorial.html). CNVs containing ≤2 SNPs were excluded.

2.5.3 Circular binary segmentation based CNV detection algorithms

2.5.3.1 Circular binary segmentation

Circular binary segmentation (CBS) is a modified version of binary segmentation method to find change points in a sequence of data. For example, it can be applied in aberrant DNA copy number detection with chromosomal SNP genotyping data. It allows for tertiary splits by connecting the two chromosomal ends (thus called "circular") (Olshen et al., 2004).

CNVs occur in contiguous regions of the chromosome that often cover multiple markers. The markers within a CNV region display aberrant copy number (normal copy number=2); therefore the beginning and ending of this region are "change points" on the chromosome underlying change of copy number status in the region compassed by the two change points. The CBS algorithm provides a natural way to segment a chromosome into contiguous regions and bypasses parametric modelling of the data with its use of a permutation reference distribution. The SNP array data to be used for change-point detection are the log ratio of normalized intensities indexed by the marker locations. There may be multiple change-points in a given chromosome, each corresponding to a change in the copy number in the test sample

(Venkatraman and Olshen, 2007a). The goal of CBS is to identify all the change-points which will then partition the chromosome into segments where copy numbers are constant. Once the chromosome is partitioned the copy numbers of the segments can be estimated with the help of additional information. This will provide the locations of copy number aberrations.

2.5.3.2 DNAcopy

DNAcopy is a non-parametric method which is based on circular binary segmentation (CBS). It splits the chromosomes into contiguous regions of equal copy number by modeling discrete copy number gains and losses. Using a permutation reference distribution, it bypasses parametric modeling of the data for assessing significance of the proposed splits. The model selection is done in the forward way by repeatedly splitting each contiguous segment until no significant splits are found, using a maximal t-statistic with a permutation reference distribution to assess statistical significance of differences in the LRR values within a segment compared to those in the adjacent segments. The computational time required for permutation is exponentially correlated with number of markers considered. To tackle the problem of long computational time when applying CBS to high-density array data, two speed enhancements to the original CBS algorithm are incorporated in DNAcopy: 1) a hybrid approach for the computation of the p-value of the maximal t-statistic using a tail probability approximation for the maximal of a Gaussian random field; 2) a sequential testing approach for deriving a stopping rule that reduces the number of permutations when there is strong evidence for the existence of a change-point. DNAcopy outputs the predicted mean LRR of each predicted segment. Because only LRR are used in the detection, no specific copy numbers can be assigned for each predicted segment; the segments are thus generally classified as “copy number gains” and “copy number losses”.

DNAcopy version 1.12.0 was downloaded from http://www.bioconductor.org/packages/2.3/bioc/html/DNAcopy.html. It was installed and run under the R environment. The analysis was performed according to the user manual (Venkatraman et al. 2007b). A smoothing step was implemented before segmentation.

2.5.3.3 cnvPartition

The cnvPartition algorithm quickly scans genotyping data sets to identify the existence and location of copy number aberrations. This algorithm is best conceptualized as two modules: one for breakpoint identification, and another for assigning copy numbers to the regions lying between these breakpoints (Illumina Manual 2).

The breakpoint detection module is similar in style to the circular binary segmentation algorithm for copy number analysis, but processes samples at a faster rate. For breakpoint detection, two hypothetical breakpoints are placed at the 5' and 3' ends of a chromosomal region, respectively. The algorithm then tries to find one internal breakpoint such that the mean log R ratio between the breakpoints is maximally different from the mean LRR on either side of the breakpoints. Then, maximal binary splits are determined on both sides of the first single breakpoint in the hypothetical region, resulting in the labeling of two more putative breakpoints. The segment between two of the 5 breakpoints described above which has the highest difference in LRR compared to other segments is identified as a putative chromosomal aberration. Once a putative aberration is identified, the significance of its splits is assessed and a confidence score is given for each of such putative CNV segment.

Following partitioning, copy number estimates are assigned to each identified segment. The first step is to compute the median of the segment's LRR and their robust standard deviation (median absolute deviation—MAD). If a segment's LRR estimate is more than a threshold value specified by the programme, it is called a copy number gain, otherwise it is called a copy number loss. This is followed by the discrimination between single-copy and homozygous deletions using BAF. Specifically, the number of SNPs in this region which have extreme BAF values (<0.25 or >0.75) is determined. If this number is greater than what would be expected by chance ($p < 0.01$, sign test), the segment is assigned a copy number of one; otherwise it is assigned a copy number of zero.

cnvPartition 1.0.2 as a plug-in for BeadStudio (Illumina Manual 3) was downloaded from Illumina Connect website at http://www.illumina.com/software/illumina_connect.ilmn and installed on BeadStudio platform. The parameter settings applied for the CNV analysis were: minimum probe count (the minimum number of SNPs to define a CNV): 3, because CNVs containing ≤2 SNPs are more likely to contain a high fraction of false positives; detect extended homozygosity: true; minimum homozygosity region size: 10 Mb; confidence threshold: 35. The programme gave an output of a list of detected CNVs with their chromosomal locations, estimated copy number and confidence score for each putative CNV. The CNVs with confidence score <35 were excluded.

2.6 Data cleaning

SNP coverage in centromeric regions is very low, thus CNVs called in these regions are much more likely to represent false positive calls. For this reason all the CNVs spanning centromeres were excluded from the analysis (according to the coordinates of centromeres on each chromosome). CNVs on chromosome X and chromosome Y were excluded due to the complications of hemizygosity in males and X-chromosome inactivation in females. CNVs represented by a single SNP were excluded. It is decided that a CNV region detected by any method based on genotyping array data must span at least two markers; therefore any CNV reflecting only one marker (length of 1bp) was excluded.

The Croatian samples were genotyped on an earlier version (V1) of the Illumina HumanHap 300K platform which were designed based upon the Human May 2004 (hg17) sequence assembly, while the Orcadian and South Tyrolean samples were genotyped Illumina HumanHap 300K V2 platform based upon the Human March 2006 (hg18) assembly. To match the single Orcadian sample with the Croatian data, the coordinates of each CNV detected for the Orcadian and South Tyrolean sample were converted to hg17 using liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver).

Chapter 3

Copy Number Variation in Individual Genomes

3.1 Preface

Completion of the first finished sequence of human genome, a global collaboration of scientists under the banner of International Human Genome Project (Human Genome Sequencing Consortium, 2004; Human Genome Sequencing Consortium, 2001), was a major technical achievement that provided a common start point for a wide range of basic science, biology and medicine. It was achieved through automated Sanger sequencing, a robust but relatively costly and time consuming chemistry for genome analysis. Once the first consensus human sequence was assembled, acquiring complete new human genome sequence was much easier and faster, but still costly and unaffordable on a large scale (Human Genome Sequencing Consortium, 2004; Levy et al., 2007). Over the past five years, the technical advances in nucleotide chemistry combined with array based methods for templating target sequences, plus massively parallel, detection of sequence reactions have led to a revolutionary shift in genome sequencing , so called next-generation sequencing (NGS). NGS platforms include Roche 454, Illumina GenomeAnalyzer, Life Technologies SOLiD, Helicos Bioscience HeliScope and Complete Genomics. As of February, 2011, at least 33 individual genomes have been fully sequenced (Table 4.1), one of which was sequenced twice on two different platforms (Bentley and et al., 2008; McKernan et al., 2009).

One valuable application of massively parallel sequencing is variant discovery by sequencing targeted regions of interest or whole genome. Those genetic variants include single nucleotide variants (SNV) and structural variants (SV). To identify SVs using NGS platforms, two categories of computational approaches have been developed. The first category are based on paired-end read mapping/paired-end read sequencing/end-sequence profiling (PEM/PES/ESP), which detects insertions and deletions by comparing the distance or orientation between mapped read pairs to the average insert size of the genomic library (Koboldt et al., 2010). The second category detects discrepancies of read depth (RD) between sample genome and reference genomes for a DNA fragment, to determine events with either increased or decreased copy number (Yoon et al., 2009). In addition to SNV discovery, most of the published individual genome sequencing studies has utilized either PEM or RD methods to investigate SV of the study genome(s) (Table 3.1).

Whole genome sequencing enables detection of SVs with essentially single base resolution, substantially higher resolution than achievable by Comparative Genomic Hybridization (CGH) or Single Nucleotide Polymorphism (SNP) genotyping platforms. It is also able to detect copy-invariant structural variants, such as inversions, and can pin point structural break-points, which always remain problematic for array-based platforms (CGH and SNP).

Table 3.1 Complete individual genomes sequenced on massively parallel sequencing platforms, and the CNVs detected from sequencing data

Sample ID	Ancestry	Gender	Disease State	Study	Sequencing platform	Max read length (bp)	Fold coverage	SNPs (m)	dbSNP (%)	Definition of CNV	Calling method	num CNVs by sequencing
Hemo0001	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	30.4×	3.38	88	>2kb	read depth	746
Hemo0004	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	23×	3.29	88	>2kb	read depth	788
Hemo0005	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	36.2×	3.39	88	>2kb	read depth	847
Hemo0006	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	51×	3.46	88	>2kb	read depth	863
Hemo0007	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	34.2×	3.37	88	>2kb	read depth	847
Hemo0011	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	31.6×	3.27	88	>2kb	read depth	918
Hemo0017	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	33.4×	3.47	88	>2kb	read depth	770
Hemo0019	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	20.2×	3.29	88	>2kb	read depth	823
Hemo0020	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	36.4×	3.44	88	>2kb	read depth	776
Hemo0022	European	Male	Hemophilia A	Pelak *et al.* 2010	Illumina GA	2×25	38.7×	3.40	88	>2kb	read depth	884
Control 1	European	Female	Epilepsy	Pelak *et al.* 2010	Illumina GA	2×25	32.3×	3.58	88	>2kb	read depth	739
Control 2	Hispanic American	Male	Epilepsy	Pelak *et al.* 2010	Illumina GA	2×25	28×	3.74	85	>2kb	read depth	782
Control 3	European	Male	Control Individual	Pelak *et al.* 2010	Illumina GA	2×25	23.6×	3.36	88	>2kb	read depth	819
Control 4	Hispanic American	Male	Schizophrenia	Pelak *et al.* 2010	Illumina GA	2×25	30.5×	3.66	87	>2kb	read depth	765
Control 5	European	Male	Schizophrenia	Pelak *et al.* 2010	Illumina GA	2×25	27.4×	3.42	89	>2kb	read depth	814
Control 6	African American	Male	Schizophrenia	Pelak *et al.* 2010	Illumina GA	2×25	24.9×	4.02	78	>2kb	read depth	809
Control 7	European	Male	Extreme Memory	Pelak *et al.* 2010	Illumina GA	2×25	23.3×	3.40	88	>2kb	read depth	822
Control 8	European	Male	Extreme Memory	Pelak *et al.* 2010	Illumina GA	2×25	26.9×	3.42	88	>2kb	read depth	790
Control 9	European	Female	Cold Urticaria	Pelak *et al.* 2010	Illumina GA	2×25	29×	3.58	88	>2kb	read depth	760
Control 10	European	Female	Metachondromatosis	Pelak *et al.* 2010	Illumina GA	2×25	31.4×	3.54	88	>2kb	read depth	743
Lupski	European	Male	CMT	Lupski *et al.* 2010	Life SOLiD	2×25~50	29.6×	3.42	83	N.A.	PEM	84
NA07022	European (CEU)	Male	N.A.	Drmanac *et al.* 2010	Complete Genomics	2×35	87×	3.08	90	N.A.	PEM	2125
NA19240	African (YRI)	Male	N.A.	Drmanac *et al.* 2010	Complete Genomics	2×35	63×	4.04	81	N.A.	none	none
NA20431 (Church)	European	Male	none	Drmanac *et al.* 2010	Complete Genomics	2×35	45×	2.91	90	N.A.	none	none
ABT	African	Male	N.A.	Schuster *et al.* 2010	Life SOLiD	2×25	30×	3.62	89	N.A.	none	none
KB1	African	Male	N.A.	Schuster *et al.* 2010	Roche/454	1×350	10.2×	4.05	82	>20kb, duplication only	read depth	886
NA18507	African (YRI)	Male	N.A.	McKernan *et al.* 2009	Life SOLiD	2×25	18×	3.87	81	100bp-100kb	PEM	4660
P0 (Quake)	European	Male	none	Pushkarev *et al.* 2009	Helicos	1×70	28×	2.81	76	>1kb	read depth	752
AK1	Asian (Korean)	Male	N.A.	Kim *et al.* 2009	Illumina GA	2×106	28×	3.45	83	duplication only	PEM	24
SJK	Asian (Korean)	Male	N.A.	Ahn *et al.* 2009	Illumina GA	1×75	29×	3.44	88	0.1-100kb	PEM	3335
YH (Wang)	Asian (Chinese)	Male	N.A.	Wang *et al.* 2008	Illumina GA	2×35	36×	3.07	86	>100bp	PEM	2682
Watson	European	Male	none	Wheeler *et al.* 2008	Roche/454	1×250	7.4×	3.32	82	1kb-3Mb	N.A.	602
NA18507	African (YRI)	Male	N.A.	Bentley *et al.* 2008	Illumina GA	2×35	41×	3.45	74	50bp-35kb	read depth	5704
Venter	European	Male	N.A.	Levy *et al.* 2007	Shotgun	1×800	7.5×	3.07	95	N.A.	N.A.	666

‘Sequencing Platform’: Illumina GA--Illumina Genome Analyzer.’SNPs’ indicates the number of single nucleotide polymorphisms called, excluding small indels. ‘dbSNP’ indicates the proportion of SNPs that were present in dbSNP build 126 or later. ‘Definition of CNV’ indicates what the authors define as a CNV in each study; if such definition is absent in the article but CNVs recorded in DGV, the DGV deification of a CNV (1kb-3Mb) is used. ‘num CNVs by sequencing’ indicates total number of CNVs detected from sequencing data (validated and unvalidated), with definition of a CNV by author for the according study.

Copy number variation, structural variants of median length (1kb to 3Mb), has been intensively studied in recent years. However, CNV detection from SNP genotyping suffers from limited SNP coverage in most SNP genotyping platforms and lack of (a) robust algorithm(s) for detection, resulting in high error rates and poorly assayed regions of the genome. Some basic but important features of CNVs including breakpoint definition and chromosomal distribution are inadequate or partial. A complete map of all CNVs from individual genomes would be a valuable resource to catalogue Copy Number Polymorphism (CNP) and to understand the origin and formation mechanism with the knowledge of genomic features within and spanning CNV regions. Although CNVs were reported in most individual whole genome sequencing studies, parallel comparisons of CNVs with other studies and with the CNVs in public databases were only mentioned briefly in a few of them (Schuster et al., 2010; Wang et al., 2008).

In this Chapter, I describe a) how I brought together CNVs directly and unambiguously called from whole genome sequencing data from all available published resources, b) compared the distribution and other features of deletions in five individual genomes of three ethnic origins (European, African and Asia), in the aim of revealing the difference of panorama CNV make up at individual level and group level and c) address current limitations of CNV cataloguing using NGS.

3.2 Materials and Methods

3.2.1 CNV data

Whole genome CNVs called from sequencing data, from 12 published studies (Ahn et al., 2009; Bentley and et al., 2008; Drmanac et al., 2010; Kim et al., 2009; Levy et al., 2007; Lupski et al., 2010; McKernan et al., 2009; Pelak et al., 2010; Pushkarev et al., 2009; Schuster et al., 2010; Wang et al., 2008; Wheeler et al., 2008), were either obtained from the published paper, supplementary information or downloaded from Database of Genomic Variants (http://projects.tcag.ca/variation/, version 10 of all recorded structural variants, last updated in November 2010). The total number of each CNV called from sequencing data (CNVs called by further validation methods such as CGH and SNP array were not included), calling method, definition of CNV in each of the 12 studies were recorded.

3.2.2 Analysis of deletions

Deletions within the length range of 1kb to 3Mb in five individual genomes, from six studies were selected for comparison. Every genomic location (chromosome coordinates in Build 36) was recorded. Deletions detected primarily by CGH or SNP array were not included, as the following comparison of deletions focuses on CNV results directly drawn from DNA sequencing data. Information of genes overlapping each deletion was obtained from DGV (version 10, last updated in November 2010).

A deletion region was defined as the maximum genomic region which was shared among all samples carrying a deletion at the same locus. The deletion regions for the five individual genomes were determined with in-house scripts compiled by myself, using R.

The Mann-Whitney test was performed to test the difference in a) mean lengths of deletion and b) gene content in the deletion regions, between groups of individuals. The significance level was set to 0.05.

3.3 Results

3.3.1 Overview of CNV detection in completely sequenced individual genomes

To date, 12 whole individual genome sequencing studies reported their CNV findings from part or whole genome sequence data (Table 3.1). Kim et al. used a small number of CNV calls generated with sequence data from part of the genome as a training set, while the main method they adopted to detect genome-wide CNVs was CGH. Lupski et al. also utilized CGH to detect CNVs in sample genome and used sequence data as a complementary method. All other studies produced CNV calls directly from sequence data, with other CNV detection methods to validate or to complement those CNVs. In these 10 studies, Levy et al. (2007), Wheeler et al. (2008) and Bentley et al. (2008) did not describe method of CNV detection from sequence data in their research article; the CNV results from these three studies were later deposited to DGV, which were not available in the original research article or supplementary information. Drmanac et al. (2010) and Schuster et al. (2010) each selected one sample genome (NA07022 and KB1 respectively) from their studies to investigate structural variants. 4 studies chose read depth (RD) methods to detect CNVs from sequence data; the other 6 studies used end-pair mapping (EPM) methods (Table 3.1).

The number of detected CNVs ranged from 84 to 5704 for each genome, and the definition of a CNV by length also varied across the studies (Table 3.1). Schuster et al. and Kim et al. only investigated duplications, but not deletions.

3.3.2 Comparison of deletions in five individual genomes

The CNVs detected directly from sequencing data were at a higher resolution and with more complete genome coverage, compared to other CNV detection methods (WTCCC, 2010). The released CNV data from genome sequence studies made it possible to reveal some features of CNVs from a relatively accurate and reliable source. Among the 12 studies which reported sequencing of individual genomes, 7 released lists of CNVs which were obtained from the raw sequence data (Ahn et al., 2009; Bentley and et al., 2008; Drmanac et al., 2010; Kim et al., 2009; Levy et al., 2007; Lupski et al., 2010; McKernan et al., 2009; Pelak et al., 2010; Pushkarev et al., 2009; Schuster et al., 2010; Wang et al., 2008; Wheeler et al., 2008). Ahn et al. (2009) only tested for duplications and Bentley et al. (2008) only released data of deletions. To generate a data set of CNVs comparable across the most of available genomes, only deletions of a certain range of length (>1kb and <3Mb) were considered. In all, six genome sequencing studies on five human genomes, Venter, Watson, NA18507, YH and SJK, met the starting criteria for analysis and comparison. (Table 3.2)

Table 3.2 Deletions in five sequenced genomes from six studies

Sample ID	Ancestry	Gender	Study	Sequencing platform	Calling method	# Deletions by sequencing	Length range (bp)	Median length (bp)
Venter	European	Male	Levy *et al.* 2007	Shotgun	N.A.	320	1006-19710	2115
YH (Wang)	Asian (Chinese)	Male	Wang *et al.* 2008	Illumina GA	ESP	487	1004-124100	2443
Watson	European	Male	Wheeler *et al.* 2008	Roche/454	N.A.	602	1007-38900	4022
NA18507	African (YRI)	Male	Bentley *et al.* 2008	Illumina GA	read depth	693	1002-50123	2620
NA18507	African (YRI)	Male	McKernan *et al.* 2009	ABI SOLiD	ESP	4125	1103-937300	1616
SJK	Asian (Korean)	Male	Ahn *et al.* 2009	Illumina GA	ESP	988	1001-99440	2624

The number of deletions between 1kb and 3Mb varies greatly among individuals (from 320 to 4125). It is noted that even for the same individual, NA18507, the choice of sequencing platform affected detection. Bentley et al. (2008) sequenced the genome of NA18507 using Illumina Genome Analyzer in 2008 and reported 693 deletions; a year later McKernan et al. (2009) sequenced the same individual on ABI SOLiD platform, while the number of deletions in the same size range they found was about 7 times (4125) that from Bentley et al. (2008) study. A comparison of deletions from these two studies for NA18507 was carried out. I found that 594 deletions were found by both studies, corresponding to 85.7% of the deletions in Bentley et al. (2008) and 14.4% deletions in McKernan et al.(2009). The number of deletions per genome for NA18507 in Bentley et al.(2008) study was more similar to that of other four genomes; and also according to a study which constructed a reference set of structure variants for two HapMap individuals based on data from four sequencing studies, the average deletion variants in those two individuals were about 680 per genome (Mills et al., 2011). Therefore data from Bentley et al. (2008) was selected to be included in further analysis, representing deletion profile of NA18507. Subsequently a final set of totally 3090 deletions for JCV (J.C.Venter), JDW (J.D.Watson), NA18507, YH and SJK were determined.

The distribution of deletion lengths for the five genomes showed an 'L' shape: the majority of deletions were small in length (<10kb) and only a few detected deletions were large (11.4% deletions >10kb). A concordance of trend in deletion length distribution was observed among all five individuals (Figure 3.1). The median length of deletions for SJK, NA18507 (2008) and YH, whose genomes were sequenced on Illumina GA platform, were similar at around 2500bp (Table 3.1).

The 3090 deletions were grouped into 2053 non-redundant copy number variant regions (CNVRs). A CNVR is the maximum region shared among all individuals carrying a CNV at the same locus. Among the total 2053 CNVRs, 1502 (73.1%) were detected in only one individual,; 295 (14.3%) shared by two individuals; 126 (6.1%) by three individuals; 86 (4.2%) by four individuals; and 44 (2.1%) by all five individuals. For each individual, except YH, about half of the deletions detected did not overlap those in any other genomes (Table 3.3).

The overlap rate of deletion CNVRs in each two genomes selected from the five sequenced genomes ranged from 12.9% to 61.7%. The concordance of deletions between the African sample NA18507 and each of the two Asian samples (27.8% and 36.8% of deletions from NA18507 overlapped those from YH and JSK, respectively) was higher, than that between NA18507 and each of the two European samples (15.4% and 22.5% of deletions from NA18507 overlapped those from JCV and JDW, respectively). The concordance of deletions in the two Asian samples was high (61.7% of YH's and 30.6% of JSK's deletions were detected in each other's genome), while the concordance was low in the two European samples (28.4% of JCV's and 15.9% of JDW's deletions were detected in each other's genome).

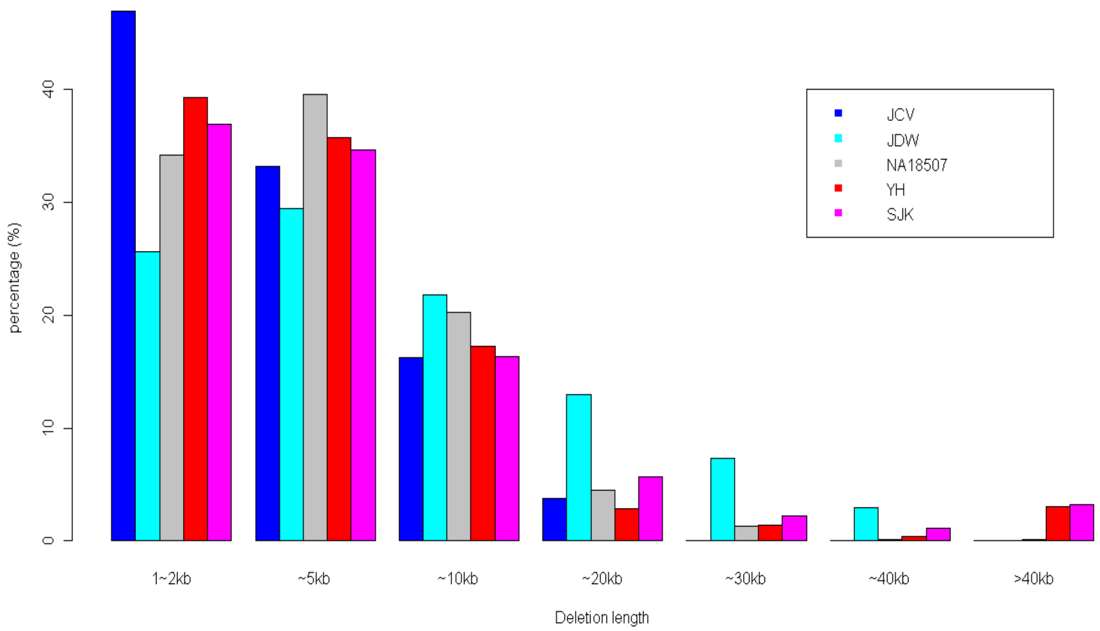


Figure 3.1 Deletion length distribution for five sequenced genomes: JCV, JDW, NA18507, YH and SJK. Each vertical bar represents deletions in the indicated length range, in proportion to total number of deletions for an individual genome.

Table 3.3 Overlap of deletion CNVRs in paired genomes.

	NA18507	JCV	JDW	YH	JSK
#CNVRs	679	320	573	485	977
Overlap of CNVRs with other genomes (percentage):					
NA18507	-	105(32.8%)	153(26.7%)	189(40.0%)	250(25.6%)
JCV	105(15.4%)	-	91(15.9%)	97(20%)	126(12.9%)
JDW	153(22.5%)	91(28.4%)	-	141(29.1%)	178(18.2%)
YH	189(27.8%)	97(30.3%)	141(24.6%)	-	299(30.6%)
JSK	250(36.8%)	126(39.4%)	178(31.1%)	299(61.7%)	-
private	349(51.4%)	156(48.8%)	329(57.4%)	137(28.2%)	531(54.3%)

Each column presents comparison of deletion CNVRs in one genome to each of other four genomes, and also the CNVRs found exclusively in this genome (private CNVRs). In brackets are the percentages of CNVRs overlapped by each of the other four genomes, compared to total CNVRs in an individual genome.

The overlap of deletion CNVRs in the five samples in the context of their ethnic origins is shown in Figure 3.2. The 'Asian' deletion set comprised 1163 deletion CNVRs detected in one or both of YH and JSK's genomes; the 'European' deletion set comprised 802 deletion regions from JCV and JDW, whilst the 'African' set, represented by NA18507 comprised 679 deletions.

Of the total 2053 deletions, 153 were observed in all three ethnic groups. The number of deletions shared between African and European groups (197) was lower than that for African and Asian (286), or Asian and European (261). (Figure 3.2)

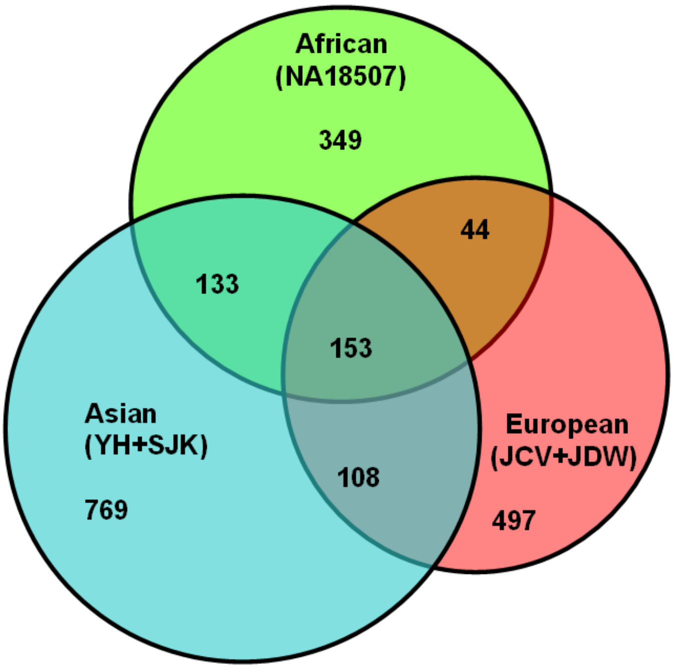


Figure 3.2 Deletions in sequenced individuals from African, Asian and European origins.

The gene content in shared and private deletion regions was investigated. A third (1059) of the 3090 deletions for the five genomes overlapped genes. Gene count in these deletions ranged from 1 to 8. The mean number of genes in 1502 deletions private to one individual was 0.33, which was significantly less than the 0.44 genes on average contained in the 551 shared deletions (P=0.0001314).

3.4 Discussion

The advance of ultra-high-throughput sequencing technologies enabled and greatly accelerated the study of full spectrum of genomic variants, from single base change (single nucleotide polymorphism) to large scale structural variants, including copy number variants. At the time of writing, at least 12 studies have reported whole genome sequencing of 33 human genomes, however the focus of variant discovery was still on single nucleotide variants; the description and discussion of SV detected directly from sequence data was superficial and mostly without comparison to other individual sequenced genomes.

Those studies carried out whole genome sequencing on different platforms, from Sanger shotgun sequencing to NGS platforms (Illumina, Roche/454, Life SOLiD, Helicos and Complete Genomics). The error rate of NGS is still higher than conventional Sanger sequencing, especially elevated at the start and end of a read (Koboldt et al., 2010). It is notable that the generation and analysis of data from NGS instruments present numerous challenges, including sample contamination, library chimaeras, sample mix-ups, variable run quality and computation issues with sequence alignment (Xi et al., 2010).

The algorithm used is another variable and limitation for SV detection from whole genome sequencing studies. Even for the same individual, applying different algorithms can result in different calls, for example, the concordance of deletion calls of NA18501 from two studies each used RD or EPM method was not high (14.4% CNVs detected by EPM were also detected by RD). Moreover, an RD and EPM method each has its problems. First of all, due to the short length of sequenced bases, many reads cannot be uniquely mapped to the genome. Second, the alignment is particularly problematic at segmental duplication rich regions; read-depth methods could detect variants at those locations, but their resolution is relatively poor. Third, PEM-based methods have the advantage to detect dosage-invariant SVs, but these algorithms have limited power in detecting insertions larger than the insert size. Fourth, the G+C content throughout the genome, amplification error and uneven likelihood of fragmentation all may cause different representation of certain regions compared to others. Last but not least, many of the data sets do not have sufficient coverage to infer all SVs with statistical significance (Xi et al., 2010).

A survey of deletions from five sequenced genomes showed that the majority of deletions between 1kb and 3Mb were short in length and at the lower level for robust detection by CGH or SNP typing (Figure 3.1). This skewed distribution is also true for CGH or SNP genotyping detection methods (McCarroll et al., 2008; Redon et al., 2006), but on a different scale from NGS. For example, in a study of CNVs in three European populations from SNP genotyping data (Chapter 7 of this thesis), about 60% of all deletions between 1kb and 3Mb were larger than 10kb, while over 80% of deletions were smaller than 10kb from the sequence data of five individuals. This discrepancy of size range reflects limitation of algorithms to detect larger SVs from sequence data. Although NGS has a lot of advantages in SV detection than conventional methods, it is still not a one-stop solution for SV discovery. Improvements in the pipeline for SV detection from sequence data are needed. Some endeavors have been made by combining several complementary algorithms, to predict SVs more accurately (Hormozdiari et al., 2009). It is also advised that future studies may need to use multiple libraries with different insert sizes to discover SVs of a wider size range (Medvedev et al., 2009). One the other hand, CGH or SNP genotyping based methods are useful in detecting larger SVs, especially large insertions which are technically difficult for sequencing based methods, therefore applying these two categories of SV detection for the same individual and combining result from both platforms would enable the discovery of SVs across the full size range.

The overlap of deletions among the five whole genome sequenced individuals was low (2.1%). About half of all discovered deletions for each individual were unique to that individual. Compared with single nucleotide variants, which were 74% to 95%, concordant with the reference human sequence in dbSNP (table 4.1), structural variants were much less congruent, reflecting in part the technical limitations on robust detection and alignment. Concordance of deletions was higher between two Asian genomes but lower between two European genomes, but many more examples will be needed to draw any fair conclusions about potential variations in frequency or distribution between individuals and ethnic groups. An analysis which compared deletion variants in three ethnic groups showed that the rate of sharing was lower between African and Europeans than that between African and Asians or Asians and Europeans, which was similar to SNV sharing in five individuals (NA18507, JCV, JDW, YH, AK1) from the same three ethnic groups (Ahn et al., 2009). This result may reflect the genetic origin of structural variants, but again with such a small sample size of sequenced genomes, those between-group differences could not be studied at a population level.

In this small set of individuals, I noted that those deletions observed in a single individual overlapped with fewer genes than those shared in multiple individuals (P=0.0001314). Deletions are generally considered to be more likely to be harmful than insertions (Conrad et al., 2006), being more likely to directly disrupt gene function. Consequently, deleterious deletions would not survive purifying selection and are less likely to be fixed over generations. Neutral deletions are less sensitive to purifying selection so they would be expected to be more common than deletions which affect more genes.

In conclusion, platforms and algorithms for detecting SVs from sequence data needs to be improved and one should consider incorporate other SV detection method into the pipeline to produce most accurate calls of SVs in the whole size range, covering wider regions in the genome. The comparison of deletions detected for five sequenced individuals showed that CNVs in human genomes were ubiquitous but the understanding of them, including basic features like distribution of all CNV loci, was still far from complete. Difference of some characteristics, for example gene content was observed between CNVs restricted to one genome and those shared by multiple individuals, however more samples were needed to draw a conclusion at population level. Recently, the 1000 Genome Project (http://www.1000genomes.org/) has published their primary CNV map based on whole genome DNA sequencing data from 185 human genomes, which consist of 22,025 deletions and 6,000 additional SVs (Mills et al., 2011). They found common deletions were more often shared across populations, whereas rare alleles were frequently observed in only one population. They also pointed out that due to limitations of current SV calling methods which depend on mapping reads onto their genomic locus of origin, only a fraction of the total SVs could be detected. With the advance of projects which aimed at sequencing more individuals, including the 1000 Genome Project and The Cancer Genome Atlas, (http://cancergenome.nih.gov/) and technological development of sequencing platform and SV detection algorithms, it is believed that our knowledge of structural variants, including copy number variation, will substantially increase and the full spectrum of CNVs in human genomes will be revealed.

Chapter 4

Structured Review of SNP-based CNV studies in HapMap samples

4.1 Preface

Single nucleotide polymorphism (SNP) oligonucleotide genotyping arrays are widely accepted to analyze copy number variants (CNVs). Profiling CNVs from SNP data has become popular in established Genome Wide Association Studies because of the convenience of data acquisition and processing. However, in the absence of completed high-resolution, genome-wide maps of variation, the extent to which commercially available SNP platforms accurately capture CNVs remains unknown. Moreover, genotyping platforms and CNV calling algorithms vary largely across studies, making it difficult to assess the robustness of SNP-based CNV detection.

The probe coverage for four commonly used SNP arrays (Illumina HumanHap 300, Affymetrix 500K, Illumina Human 1M and Affymetrix 6.0) had been investigated, in nine human genomes whose variants had been systematically detected by fosmid ESP mapping and validated by orthogonal approaches (Cooper et al., 2008). With the fine-scale map of genomic structural variation determined (Kidd et al., 2008), these nine HapMap genomes provide reliable information of the locations, breakpoints and copy number status of their CNVs, therefore can serve as a good bench mark to evaluate the CNV discovery results generated by other methods. It has been found from Cooper (2008)'s study that probes in older genome-wide platforms such as Illumina HumanHap 300 and Affymetrix 500K only cover about 25% of deletions, and fewer than 20% of deletions harbour multiple markers (most detection algorithms call CNV events only if the CNVs span at least two markers). Even when newer arrays (Illumina Human 1M and Affymetrix 6.0) were considered, about 30% deletions annotated by complete fosmid sequencing were not covered by the markers on those platforms. With probe coverage of 25%-70%, one can presume that the detection rate of actual CNV calling methods using probe intensity data from any of these arrays will turn out to be even lower.

The International HapMap project (The International HapMap Consortium, 2003) provides a key resource for researchers to study human DNA sequence variation, by characterizing sequence variants, their frequencies, and correlations between them, in DNA samples from four major world populations with African, Asian and European ancestry. The HapMap resource includes genotype data on over 4 million single nucleotide polymorphisms (SNPs), gene expression data using various microarray platforms and other phenotypic data such as drug response as well as structural variation data. In recent years, many researchers who are interested in human genome structure variants have taken advantage of HapMap genotyping data to study copy number variants and to test or validate their CNV calling algorithms.(Conrad et al., 2006; Cooper et al., 2008; Kohler and Cutler, 2007; Korn et al., 2008; Lin et al., 2008; Locke et al., 2006; McCarroll et al., 2008; Redon et al., 2006; Wang et al., 2007; McCarroll et al., 2005)

With CNV detection from SNP data becoming a routine for GWAS studies, a reliable protocol is required to guarantee accurate CNV detection. But a golden standard has yet to be discovered. In this chapter, a structured review of all the SNP-based CNV studies to date using HapMap samples is conducted, and the data from each included study were extracted and analyzed. By comparing CNV calls made in those studies with the baseline CNV from the same HapMap individual, it was hoped to shed light on the evaluation of SNP-based CNV detection methods.

4.2 Structured literature search and data extraction

4.2.1 Literature search across databases

Research articles included in three major databases, ISI Web of Knowledge, Medline and Embase were searched. ISI Web of Knowledge was searched for the key words "SNP or single nucleotide polymorphism" in combination with "copy number" and yielded 864 articles. Ovid databases-Medline and Embase were searched for the key words "SNP or single nucleotide polymorphism" in combination of "copy number", with the result limited to human, non-review, published between 1995 and September 2009 and yielded 344 articles. These key word combinations were adopted to target all potential studies that detect CNVs from SNP genotyping arrays. The results of the two searches were merged. After deleting 411 duplications and 20 articles published before 1995 (from the ISI search), finally 778 potentially informative articles remained for further study.

The title and abstract of each of the 778 articles was read. Any studies focusing on non-human data or which did not utilize whole-genome SNP data to infer CNVs were excluded. At this stage 263 articles remained. Among these, 17 articles have “HapMap” in their titles or abstracts. Full texts and supplementary data of these 17 articles were obtained and examined, to identify the studies with extractable data. Some of these only used HapMap samples as a reference set; some made use of the HapMap SNP information to test CNV calling methods but the list of CNV calls are not shown. Finally, 7 studies (McCarroll et al., 2005; Conrad et al., 2006; Cooper et al., 2008; Kohler and Cutler, 2007; McCarroll et al., 2008; Redon et al., 2006; Wang et al., 2007), each with detailed information of identified CNVs or CNVRs (Copy Number Variant Regions), fulfilled the inclusion criteria and the CNV data from each study were extracted. The searching process is indicated in Figure 4.1. It was found in the search that all related articles analyze all or subsets of HapMap phase I and phase II samples from a total of 270 individuals: 30 adult-and-both-parents trios from Ibadan, Nigeria (YRI), 30 trios of U.S. residents of northern and western European ancestry (CEU), 44 unrelated individuals from Tokyo, Japan (JPT) and 45 unrelated Han Chinese individuals from Beijing, China (CHB). None of the CNV studies has used HapMap phase III data, which consists of genotypes of a total 1397 individuals and PCR resequencing results from 712 individuals (http://www.sanger.ac.uk/humgen/hapmap3/).

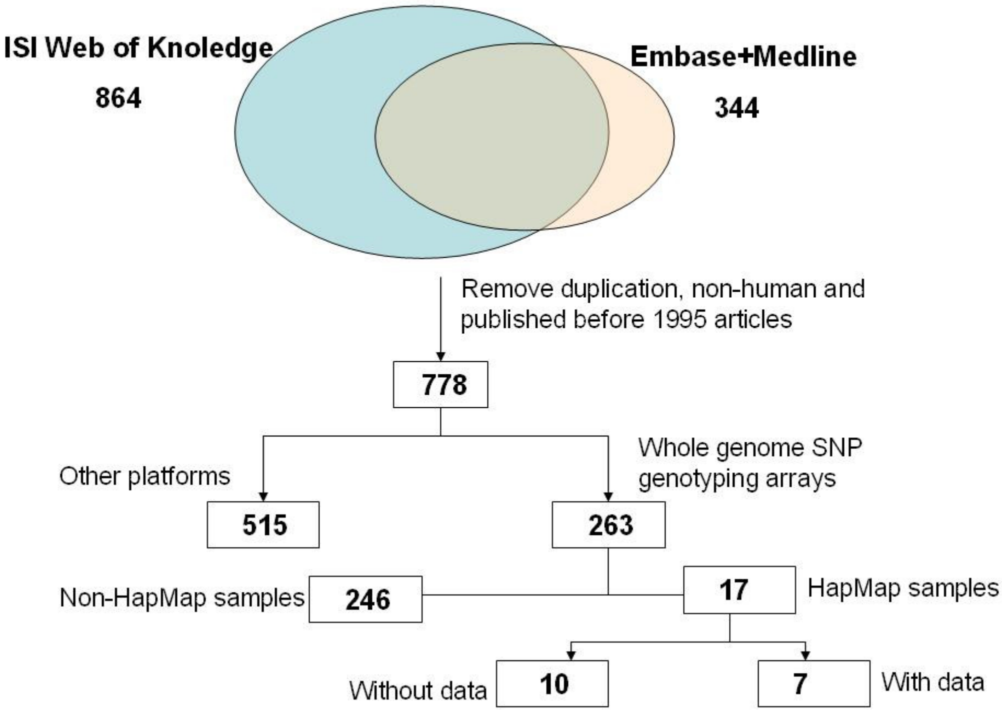


Figure 4.1 Literature search for research articles which detect CNVs from whole-genome SNP genotyping data in HapMap individuals

Table 4.1 Summary of selected information from 7 SNP-based CNV studies in HapMap samples: sample description, genotyping strategy and CNV detection methods

Author	Year	Pop	Sample Size		Genotyping platform	# Prob(K)	CNV calling				
			Total	in CNV analysis			Software	Algorithm	# final CNVs	# CNV carriers	CNVs/carrier
McCarrol et al	2005	CEU+YRI+ CHB/JPT	269	269	HapMap Project SNP genotype	~1300	in house analytical platfrom	Mendel failure, null genotypes and H-W disequilibrium	541	269	2.01
Redon et al	2006	CEU+YRI+ CHB/JPT	270	270	Affymatrix Eearly Access 500K	475	in-house analytical platform	pair-wise comparisons (SW-ARRAY), signal intensity change and SNP information	6458	270	24
Conrad et al	2006	CEU+YRI	180 (60 trios)	60(offspring s)	HapMap Project SNP genotype	1108/1086(CEU/YRI)	in-house analytical platform	Mendelian incompatibility	345/590(CEU/YRI)	30/30	11.5/ 19.7
Wang et al	2007	CEU+YRI+ CHB/JPT	112(16*3 CEU+12*3YRI+28 CHB/JPT)	112	illumina HumanHap 550	~550	PennCNV	HMM utilizing log2R and BAF	2987	112	26.7
Kohler et al	2007	CEU+YRI	90 CEU+90 YRI	180	HapMap Project SNP genotype	867/932	microdel (self-developed)	allele frequencies, genotyping-error rates, missing-data rates and deletion frequency			
McCarroll et al	2008	CEU+YRI+ CHB/JPT	270	270	Affymatrix SNP 6.0	906	Birdseye	HMM utilizing log2R and BAF	46931	270	173.8
Cooper et al	2008	CEU+YRI+ CHB/JPT	9	9	Illumina HumanHap 1M	~1000	HMMseg	HMM utilizing log2R and BAF	368	9	40.9

Pop: population formation, CEU-central European, YRI-Yoruba in Ibadan, CHB-Beijing,China, JPT-Tokoyo,Japan; #Prob: number of SNP probes in the genotyping assay (in thousands); CNV calling: name and algorithm of CNV calling software, the number of CNVs called, number of CNV carriers and average number of CNV in each carrier in each study

Table 4.1(continued) Summary of selected information from 7 SNP-based CNV studies in HapMap samples: CNVs, CNVRs and validation

Author	Year	Deletions	Amplifications	Length range (kb)	Mean/median length (kb)			CNVR		Validation		
					All	Del	Amp	Calling	# CNVRs	Methods	# tested CNVs	# validated CNVs
McCarrol et al	2005	541	0	1-745	18.3/ 7.8	18.3/ 7.8	0/0			FISH/SNP genotyping/PCR/quantitative PCR	93	81
Redon et al	2006	3454	3004		205.8/ 80.8	141.7/ 48.7	279.5/131.6	merging & summarizing all pair-wise comparisons	980	whole genome TilePath/locus-specific assay/in another incividual/in a previous study	6458 (980 CNVRs)	1789 CNVs overlap WGTP results (957 CNVRs)
Conrad et al	2006	345/590(CEU/YRI)	0/0		41.4/ 36.9	41.4/ 36.9	0/0			qPCR/380K oligonucleotide microarray	12 del/93 del	12 del/ 80del
Wang et al	2007	2060	927	0.001-7834.1	45.1/ 12.2	34.9/ 9.3	67.8/24.6					
Kohler et al	2007	CNVs were grouped into CNVRs	0		CNVR mean length 10.2	CNVR mean length 10.2	0/0	outmost boundaries of any CNV at the same loci	693(213 CEU+329 YRI)			
McCarroll et al	2008	39346	7585	0.074-1134	27.3/ 22.1 CNVR	14/5.7 del only CNVR	20.3/537.6 amp only CNVR	non-overlapping small regions containing any SNP showing CNV	1319 (877 del only+197 amp only+245 showing both)	qPCR/Fosmid ESP	27 CNPs in 30 individuals/8 individuals	99.3% concordance/76% >10kb, 64%>5kb, 27%<5 kb
Cooper et al	2008	258	110	1-1451	80.3/41.8	45.6/ 19.8	161.7/113			fosmid ESP mapping (Kidd et al 2008)	368	>2/3 of 368 CNVs validated

Number of deletions and amplifications in each study; length of each type of CNVs; CNVR: calling-the definition of a CNVR in each study, #CNVRs-number of CNVRs defined in each study; validation: methods-the way a subset of the CNVs being validated, #tested CNVs-number of CNVs selected for validation, #validated CNVs-number of CNVs validated by other methods out of the tested CNVs in each study

4.2.2 Overview of seven relevant studies

The information from these 7 studies is extracted and tabulated into spreadsheets, including author, publishing year, population composition, genotyping platform, number of probes, SNP selection criteria, CNV calling strategy, number of CNVs, CNV length and validations. The summary of information from these studies is displayed in Table 4.1.

All these studies applied certain platform/algorithm combinations to the whole (Redon 2006 and McCarroll 2008) or a subset (the other 5 studies) of the HapMap sample collection. The SNP content ranges from 475,000 (Redon) to over 1,300,000 (McCarroll 2005). Three studies (McCarroll 2005, Conrad 2006, and Kholer 2007) directly adopt SNP genotypes from the HapMap project genotype collection, and inferred copy number change via mendelian incompatibility, genotyping error and missing genotypes. Due to technology limitations, these 3 studies were only able to present deletions (copy number loss events). The other four studies used signal intensity data obtained from commercial SNP genotyping platforms (Affymetrix Early Access 500K, Affymetrix SNP 6.0, Illumina HumanHap 550 and Illumina HumanHap 1M) to detect both deletions and amplifications. Hidden Markov Model based algorithms are utilized in these studies to simultaneously indicate copy number status of the SNPs on these arrays. In all studies, each sample is detected to posses one or more CNVs. It was shown in these four studies that there were more deletions than amplifications detected in HapMap samples. Although the samples by all seven studies were drawn from the same HapMap collection, the average numbers of detected CNVs per carrier by different platform/algorithm were obviously different, from 2 events per CNV carrier to 174 events per CNV carrier. Redon, Kohler (2007) and McCarroll (2008) group CNVs into non-redundant copy number variation regions (CNVRs) each by a different CNVR definition. Instead of presenting raw CNVs with chromosomal locations, McCarroll (2008) identified common CNPs by searching for genomic regions across which copy number probes show cross-sample patterns of same status (copy number gain or loss), then assigned each individual corresponding CNVRs rather than the original CNVs. Kohler (2007) only presented CNVRs which contained any evidence of copy number variants but not original sample-wise CNVs detected in his samples.

4.2.3 Platforms and CNV detection methods

4.2.3.1 HapMap genotypes

McCarroll (2005), Conrad and Kholer obtain the SNP genotyping data from HapMap phase I data release. During phase I, genotyping of over one million SNPs was carried out by 10 centres across the world, using seven different genotyping technologies. The genotyping data generated for 270 individuals was placed in the public domain and is available for download (http://hapmap. ncbi.nlm.nih.gov/).

Redon (2006), Wang (2008), McCarroll (2008) and Cooper (2008) use probe signal intensity data from commercially available Affymetrix Early Access 500K (500K EA), Illumina HumanHap 550, Affymetrix SNP 6.0 and Illumina HumanHap 1M genotyping arrays, respectively.

4.2.3.2 Whole Genome genotyping arrays by Affymetrix and Illumina

Affymetrix's 500K EA array is a precursor version of the GeneChip® Human Mapping 500K array set, comprising 534,500 SNPs on two arrays and is used in conjunction with the whole genome sampling assay (WGSA). The median physical distance between SNPs is 2.5 kb and the average distance between SNPs is 5.8 kb on the 500K EA array. The latest version of Affymetrix genotyping array, Genome-wide Human SNP Array 6.0 contains 1.8 million genetic probes, including more than 906,600 SNPs and more than 946,000 probes for the detection of copy number variation. The inclusion of CNV probes makes Affymetrix Human SNP Array 6.0 the only platform to date with analysis tools to truly integrate copy number and association analysis simultaneously on a single array. It is claimed to have the highest physical coverage of the genome (http://www.affymetrix.com/).

Illumina's HumanHap550 Genotyping BeadChip enables whole-genome genotyping of over 550,000 tag SNPs derived from the International HapMap Project on a single BeadChip.The assay combines specific hybridization of genomic DNA to arrayed probes with allele-specific primer extension and signal amplification, therefore effectively increasing the signal-to-noise ratio in genotype calling (Wang et al., 2007). The Human 1M DNA Analysis BeadChip interrogates nearly 1.2 million SNP probes per sample, providing Illumina's most comprehensive genome-wide coverage of SNPs. The uniform genome-wide coverage results in a median spacing of 1.5 kb between markers and fewer large gaps. There are around 60,000 probes, developed in collaboration with deCODE Genetics, covering regions likely to contain undiscovered CNVs—segmental duplications, megasatellites and region lacking SNPs. This feature makes the Human 1M array of high value for CNV detection, in addition to genome-wide SNP genotyping (http://www.illumina.com/products/human1m_duo_dna_analysis_beadchip_kits.ilmn).

4.2.3.3 Deletion discovery from SNP genotypes

McCarroll et al (2005) argue that segregating deletions can cause SNP genotypes to appear 'abnormal', such as apparent deviations from mendelian inheritance, apparent deviations from Hardy-Weinberg equilibrium and null genotypes. Based on this principle, they developed a procedure to identify deletions from SNP genotypes. First of all, they detected indications of the potential presence of deletions from mendelian incompatibility (i.e. the deviation from mendelian inheritance), Hardy-Weinberg disequilibrium and null genotypes, in family trios or unrelated individuals. They then looked for regions of the genome in which the same type of “failed” profile appeared repeatedly at nearby markers in a manner that is statistically unexpected based on chance. Finally a subset of ‘failed’ SNP genotyping assays which are likely to reflect structure variants was determined.

Conrad et al (2006) adopted a similar approach for deletion detection from SNP genotypes. They examined consecutive SNPs transmitted from parent to child and flagged the regions of SNPs whose genotypes appeared to violate mendelian transmission. Then they distinguished between two classes of mendelian incompatibility which were either the mendelian incompatibility consistent with a deletion or inconsistent with a deletion. At last, they detected deletions which appear to be regions of runs of SNPs displaying mendelian errors, each in a single HapMap family trio. Due to the nature of this method, deletions only in the offspring of a family trio are represented. Compared with McCarroll (2005), this method only concentrates on Mendelian incompatibilities and doesn’t take into account other types of apparent genotyping errors (H-W disequilibrium, null genotypes) which might be resulted from deletions, and so gave results which appeared to be more stringent and conservative.

4.2.3.4 Methods to identify copy number change from signal intensity data

Redon (2006) used an algorithm described in Komura et al (2006) to identify copy number from Log$_2$R ratio of SNPs in the whole-genome SNP genotyping arrays. This approach comprised three major steps: intensity pre-processing, CNV detection and Copy-Number inference. After step one, CNV detection began with pairwise comparisons of probe intensities for all possible pairs of samples (n-1 comparisons for n samples), an adapted Smith-Waterman algorithm helped to find isolated islands of substantially higher (or lower) intensity ratios and assigned significance to each finding by a permutation test. Then the results for all comparisons were merged to extract candidate CNV regions for each sample. During step three, each CNV region was assigned a copy number and the boundaries were determined, by using a maximum clique algorithm to define the diploid samples for any given region based on the results from the large reference data.

Wang (2007), McCarroll (2008) and Cooper (2008) each facilitate a Hidden Markov Model based program (PennCNV, Birdseye and HMMseg) to detect CNV by analysing both Log2R ratio and B Allele Frequency (BAF) of the SNPs. The Hidden Markov Model is a statistical technique that models a Markov process, where the probability of observing a particular state at a particular time point only depends on the state at previous time points. This feature makes HMM widely used in the field of SNP array based CNV detection, when practice it to model status of copy numbers at nearby SNPs (Wang et al., 2007). In this model, a sequence of SNPs are each assigned a most likely copy number based on the calculated probability of copy number status, then the copy number variable fragments are assessed and determined (details of HMM in the Methods chapter).

PennCNV: this was developed by an academic group for the Illumina genotyping platform and is freely available to users to apply to their own data. The standard output from this package is a list of detected copy number variant events and brief summary statistics to be used for quality check. It can run using command line options or integrated into Illumina's BeadStudio data analysis software as a plug-in. It has a few downstream analysis options, such as using family trio data to increase accuracy of prediction(Wang et al., 2007) .

Birdseye: this is a component in the Birdsuite analytical solutions. The Birdsuite set developed by Korn et al (Korn et al., 2008) combines SNP genotyping, copy number detection and genotyping of common CNP. (The concept of CNP genotyping is similar to SNP genotyping: instead of SNP probes, probes which represent common copy number polymorphism are arrayed and hybridized with DNA samples, and then the CNP genotypes are determined.) Four different software programs are integrated into Birdsuite for Affymetrix dataset: The Canary algorithm genotypes common CNPs; Birdseed yield SNP genotyping results; Birdseye uses the HMM to identify and assess previously unkown CNVs in the data set and Fawkes merges all the results from the three previous stages.

HMMseg: this is a command line operated algorithm for segmenting continuous genomic datasets on a scale-specific basis using HMM. Scale specificity is achieved by an optional smoothing step. HMMseg uses Gaussian emission distributions to detect consecutive SNPs in the genome with aberrant signal intensities, with diagonal covariance for multiple datasets, and supports both the Viterbi and posterior decoding methods for copy number state assignments(Day et al., 2007).

4.3 Comparison of CNV findings of two individuals in six HapMap studies

Among the seven studies which are included in this review, all but one (Kohler (2007)) provide CNV call results for each of the HapMap sample in their studies. On the other hand, Kohler (2007) groups sample-wise CNVs into CNVRs and only showed the constructed map of CNVRs instead of listing the original CNV calls. The lists of CNVs for each sample in the rest six studies were downloaded from supplementary information of each of these six studies; the chromosomal coordinates for each CNV were all converted to hg17 (Build 35) using liftOver (http://genome.ucsc.edu/cgi-bin/hgLiftOver). The entries which could not be mapped to hg17 were discarded. In the remaining six studies, 2 same individuals, NA12878 and NA19240, were used by each of them. They were both in the HapMap sample collection; NA12878 is of European ancestry and NA19240 is of African ancestry. Genome-wide fosmid end-sequence-pair (ESP) maps have been developed for nine humans, which also include the same two individuals. To address the extent to which existing

SNP-based CNV detection methods accurately capture true CNVs, the CNV results of each study for NA12878 and NA19240l were compared to CNVs detected using ESP sequencing for the same individual.

4.3.1 Deletions for NA12878 and NA19240

McCarroll (2005) and Conrad et al (2006) only investigate the deletion polymorphism in their studies, therefore to construct a paralleled comparison of copy number variants across the six studies, only autosomal deletions were considered.

The number of deletions for the same two HapMap individuals from each study is shown in Table 4.2. Between 12 and 69 deletions were identified for NA12878 and between 9 and 57 for NA19240. The SNP density of each platform did not affect the number of deletions detected (P value=0.838, not significant). Two categories of CNV detection methods were used: McCarrol (2005) and Conrad (2006) chose to detect deletions based on Mendelian error/genotyping error in SNP genotypes, while the other four analyse signal intensity. Considering the rational of the two categories of methods, the genotyping analysis methods also takes into account the family trio information, therefore one may argue the results might be conservative and the number of deletions of the same individual might be underestimated. However, the difference of the number of detected deletion events between two categories of method did not reach significant threshold (P value=0.4364, two tailed).

Table 4.2 Number of deletions detected for NA12878 and NA19240 in six studies

Study	Method	SNP density	NA12878	NA19240	Both	Deletion median length(kb)
McCarroll et al(2005)	genotype	high	16(568.0Kb)	16(242.7Kb)	32	11.2
Redon et al	signal intensity	low	15(755.8Kb)	9(850.1Kb)	24	49.7
Conrad et al	genotype	high	12(537.9Kb)	20(614.8Kb)	32	20.8
Wang et al	signal intensity	low	21(1.325Mb)	13(178.2Kb)	34	11.6
McCarroll et al(2008)	signal intensity	high	69(1.564Mb)	57(1.325Mb)	126	9.7
Cooper et al	signal intensity	high	43(2.367Mb)	20(537.2Kb)	63	27.7

For NA12878 and NA19240, the number of deletions indicated in each of the six studies was listed. The total lengths of genomic regions covered by those deletions were listed in brackets.

The median length of detected deletions generated by the different platform/algorithm combination varied (from 9.7 to 49.7 kb). The deletions detected by Redon (2006) were longer than any other studies (median length 49.7kb, almost two fold of the second longest detected deletion median size in Cooper (2008)'s study). It was suspected that the deletions identified by genotype analysis might be shorter since family information provides more evidence of true deletion boundaries; but no significant difference of deletion length was found between two categories of detection methods (p=0.5745). Both number of detected deletion and the chromosomal region covered by those deletions vary greatly for the same individual across the six studies.

4.3.2 Cross comparison of deletions from six studies

The locations of each deletion for the same individual were compared across the six studies (Table 4.3). The results from two genotype analysis studies (McCarroll (2005) and Conrad (2006)) were widely validated in at least one other study (recovery rate of 64.5% and 62.5%, respectively), while most of the deletions (84.9%) from McCarroll (2008) could not be found in any other studies. Correlation tended to be better when both studies in a pair to be compared adopt the same methods, for example, deletions from McCarroll (2005) concordant with those from Conrad (2006) best, which both used genotype/family information analysis, Wang (2007) had more deletions also detected in studies in which they analyzed signal intensity rather than genotype/family information.

Table 4.3 The concordance of detected deletions in the six studies.

Study	Method	SNP density	#Total CNVs	#CNVs recovered in other studies						#CNVs not recovered (%)
				McCarroll et al(2005)	Redon et al	Conrad et al	Wang et al	McCarroll et al(2008)	Cooper et al	
McCarroll et al(2005)	genotype	high	31	-	2	13	1	8	6	11 (35.5%)
Redon et al	signal intensity	low	24	2	-	5	9	5	8	10 (41.7%)
Conrad et al	genotype	high	32	13	5	-	7	4	8	12 (37.5%)
Wang et al	signal intensity	low	34	1	9	7	-	2	12	17 (50%)
McCarroll et al(2008)	signal intensity	high	126	8	5	4	2	-	10	107 (84.9%)
Copper et al	signal intensity	high	61	6	8	8	12	10	-	36 (59.0%)

Note: The number in each cell is the number of CNVs detected in each corresponding pair of studies. For ease of comparison across multiple studies, CNVs in each of the two genomes were grouped into CNVRs (CNV loci), eg. in McCarrol(2005) study, 16 deletions of NA12878 were at 16 CNV loci, 16 deletions of NA 19240 were at 15 CNV loci, therefore those deletions were at 31 sample-wise CNV loci.

4.4. Concordance with deletions determined by ESP sequencing

4.4.1 Reference deletions detected by ESP sequencing

Kidd et al (2008) applied the clone-based Fosmid end-sequence pair (ESP) technology to determine structure variants in nine human genomes, including four individuals of Yoruba Nigerian ethnicity, three of Western and Northern European ethnicity and two of Eastern Asia ethnicity. For each individual a whole genomic library of about 1 million clones was constructed, by using fosmid subcloning strategy. Each library was arrayed and both ends of each clone insert were sequenced to generate a pair of high-quality end sequences. At the end a physical clone map for each individual human genome was generated, the regions discrepant by size or orientation on the basis of the placement of end sequence against the reference assembly were flagged. Following approaches of validation, 2725 sample-wise insertion, deletion and inversions more than 8kb in length were identified. The scale of this CNV discovery study is the finest to date compared with other platforms/methods, therefore in this chapter this data set was taken as the baseline reference to assess the robustness of CNV detection platforms/methods mentioned above. The list of all copy number variants identified in this study was downloaded from the supplementary information to this article.

In total those CNVs could be merged into 1607 non-redundant CNVRs. A CNVR was defined as the region containing any SNP showing CNV. If CNVs in multiple samples are at the same locus, they were assigned to be in the same CNVR. The boundaries of each CNVR were the locations of outmost reach of SNPs showing CNV at this locus. 1129 of the 1607 CNVRs were only observed in single individuals, among them, 475 belong to the African population, 362 to the European population and 292 to the Asian population. 478 (29.7%) CNVRs were shared by more than one individual. 133 of these CNVRs were shared by three populations; 110 only shared between European and Asian populations, 94 only shared between European and African populations and 69 only shared between African and Asian populations. (Figure 4.2.)

The same pattern of CNVR sharing among three populations was observed in Cooper (2008)'s data. The 368 CNVs from Cooper et al were grouped into 278 CNVRs, in the same manner as analysing Kidd (2008)'s results. Most of these CNVRs (243 out of 278) belong to a single population; most of the population specific CNVRs are found in African population. Only a small fraction of the CNVRs (8 CNVRs) were shared by three populations; European and African populations shared more CNVRs. (Figure 4.3.)

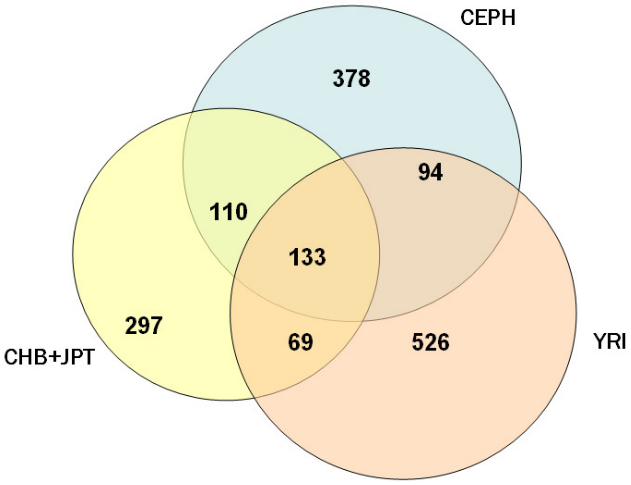


Figure 4.2 CNVR sharing in 9 HapMap individuals from Kidd et al (2008).
Note: numbers in the shaded areas indicate the numbers of shared CNVRs.

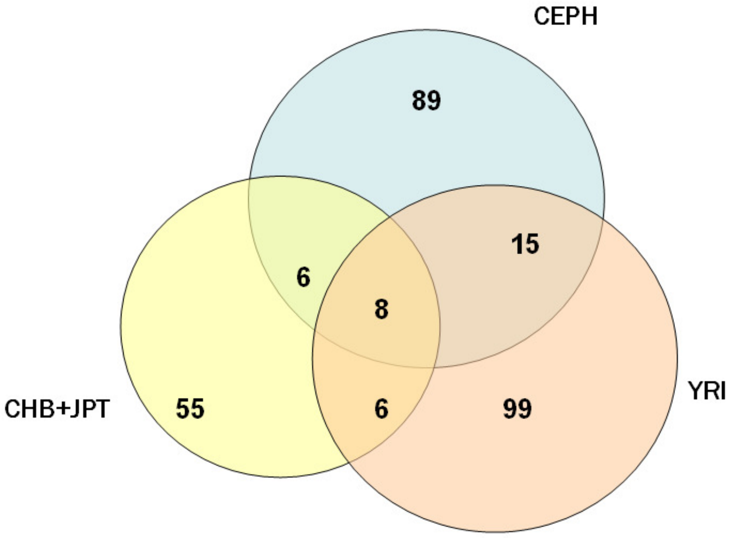


Figure 4.3 CNVR sharing in 9 HapMap individuals from Cooper (2008) et al.
Note: numbers in the shaded areas indicate the numbers of shared CNVRs.

155 autosomal deletions (7.48Mbp) were detected for NA12878 and 142 (5.76Mbp) for NA19240. Among those deletion regions, 143 for NA12878 (7.01Mbp) and 130 for NA19240 (5.29Mbp) covered multiple SNPs (the reference SNPs are from the Illumina Human 1M array SNP collection which represented one of the highest SNP densities in the six studies), which means maximally 92% of the baseline reference deletions could have a chance to be recovered by any SNP based detection methods ("detectable"), because it

always takes more than 2 consecutive SNPs which shown evidence of a shift of copy number status away from normal towards the same direction (either gain or loss).

4.4.2. Overlap with fosmid ESP map

The positions of deletions for NA12878 and NA19240 were compared to the fosmid ESP map of deletions for the same two individuals. Of the 143 detectable deletions for NA12878 on the fosmid ESP map, 42 were also detected in at least one study from the six SNP-based CNV studies (3.72Mbp). On the other hand, 26 (1.11Mbp) out of the 130 detectable deletions for NA19240 were recovered each by one or more other studies. Thus the rate for the fosmid ESP validated deletions to be recovered by any SNP-based technologies discussed in this chapter was 29.3% (or 53% by length) for NA12878 and 20% (or 21.0% by length) for NA19240. The average recovery rate for both individuals was 24.9% (or 34.2% by length).

To demonstrate the breakdown of concordant deletions by each study, the results of comparison with Kidd (2008)'s data is shown in Table 4.4. The concordance rates for the six studies range from 25.4% (McCarroll 2008) to 58.73% (Cooper). Except for McCarroll (2008), other three studies working on signal intensities tended to be more robust in identifying reference deletions (54.17%, 44.12% and 58.73% vs. 40.63% and 31.25%). Despite the low density of SNP genotyping platform chosen by Redon et al (2006), their algorithm still yielded a concordance rate of over 50%.

4.5 Discussion

Comparison of detected events in the same dataset by different methods is a good way of assessing accuracy of detection algorithms. The HapMap samples are considered the most intensively studied genetic sample collection; various studies on the whole or a subset of these samples produced genotyping results generated on various SNP genotyping array platforms, as well as fosmid ESP sequencing results. With the development of methodology to detect copy number variable events and construct maps of structural variation for HapMap samples based on those data, the HapMap collection can serve as a resource for CNV detection algorism comparison. Seven CNV discovery studies were targeted from a structured literature search, which aimed to identify all SNP-based CNV studies to date on the well characterized HapMap samples. Those were published between 2005 and 2008, either directly used HapMap genotypes (Kohler and Cutler, 2007; McCarroll et al., 2005; Redon et al., 2006; Conrad et al., 2006) or analyzed signal intensity data of SNP genotyping arrays for those samples (Cooper et al., 2008; McCarroll et al., 2008; Redon et al., 2006; Wang et al., 2007). The principles of inferring copy number variation via SNP genotypes are to identify the 'imprints' in the genome which might be caused by deletions, such as missing genotypes, deviations from Mendelian inheritance and violations of Hardy-Weinberg equilibrium. On the other hand, implying copy number status from signal intensities of a series of SNPs on a chromosome can yield results of both copy number gains and copy number losses.

Table 4.4 Deletions in concordance with Kidd et al. 2008

Study	Method	SNP density	# Detected deletions	concordance with Kidd et al			
				NA12878	NA19240	both	concordance rate
McCarroll et al(2005)	genotype	high	32	8	5	13	40.63%
Redon et al	signal intensity	low	24	10	3	13	54.17%
Conrad et al	genotype	high	32	5	5	10	31.25%
Wang et al	signal intensity	low	34	9	6	15	44.12%
McCarroll et al(2008)	signal intensity	high	126	22	10	32	25.40%
Cooper et al	signal intensity	high	63	28	9	37	58.73%

The existence of structural variance is ubiquitous in human genomes. Despite platform/algorithm choices, multiple CNVs were found for every sample in all seven studies. This fact is supported in many other human genetics studies. CNVs are also commonly found to be widely spread in other species, for example mice, chimpanzees, pigs and maize(Perry et al., 2008; Springer et al., 2009; Adams et al., 2005; Fadista et al., 2008). It is possible that for any single genome there may be regions displaying different levels of copy number polymorphisms.

The genotype-analysis based methods (in McCarroll 2005 and Conrad 2006) tend to detect less deletion events per sample (although the difference was not significant) compared with signal intensity-based methods (Table 4.1). It might be explained by the nature of these two categories of methods. The genotype analysis assessed missing genotypes, H-W disequilibrium and Mendelian incompatibilities in family trios, screened regions of genome which carried above events but possibly were caused by deletions rather than being true genotyping errors. It is presumed that incorporating family information and assess transmission of these inheritable genetic features might increase accuracy in deletion detection, but at the same time more deletions in these trios might have been missed out, for example homozygous deletions in the offspring is undetectable by this method, and if all three samples in a trio appear to be homozygous for a number of consecutive SNPs, it is hard to tell if they are truly all homozygous or if some of them are hemizygous for a deletion in this region, when no apparent Mendelian incompatibility is detected in this trio. Conrad et al (2006) performed a simulation to calculate power of their method, and claimed it has moderate power to detect deletion events in family trios. They also predicted the total number of deletions based on simulations, that the CEU children were estimated to carry around 30 deletions and the YRI children about 50 deletions of >5 kb (Conrad et al., 2006). Their method detected 14.4 deletions of >5kb, which was less than half of the estimated number of deletions in the 60 HapMap offspring. The simulations suggested that the number of events may be underestimated by genotype-analysis methods.

The density of SNPs in the seven studies was not the same; the earlier studies (Redon (2006) and Wang (2007)) used arrays containing about 500k SNPs while others analyzed above 1M SNPs. The relationship of SNP density and the number of CNVs detected (deletions were considered instead of looking at both deletions and amplifications, because three of the seven studies only studied deletions) was investigated and no significant association was found. The median size of deletions in these studies were between 5.7 kb and 48.7 kb, meanwhile the median physical distance between SNPs on one of the lower SNP density genotyping platform (Affymetrix 500K EA array) is 2.5 kb. The deletion length suggested most detected deletions spanned multiple SNPs; the sufficient length of deletions made them already detectable by both lower and higher density SNP genotyping arrays.

The shared HapMap samples, NA12878 and NA19240, are the basis of the method comparison among six studies which provided detected individual CNVs. The cross comparison showed that generally some deletions found in one study could be also found in other studies; it is assumed that the larger numbers of similar events across different platform/methods could mean higher true positive rate, therefore all the programmes were able to detect at least a proportion of true deletion events. The proportion of deletions detectable by other studies was above 40% for all studies except for McCarroll (2008). Although some studies had lower percentage of overlapping events, it is important to also consider the number of events as well as the proportion, for example about 60% of the events detected by Redon (2006) were confirmed but other algorithms had detected more events in total. In McCarroll's study, CNV identification was split into two steps: an initial detection of common copy number polymorphisms (CNPs) captured by CNP probes followed by a HMM based algorithm (Birdseye) to further detect rare events. The HMM model adapted in Birdseye was different from that by others (PennCNV and HMMseg): it took into account the probe intensities from both copy number and SNP probes to indicate true underlying hidden states (e.g. copy number status of each probe) while others analyze solely SNP probe intensities. The novelty of McCarroll's platform might result in the lower overlapping rate with results from other studies, including studies which also utilized HMM.

A difference in the size of the predicted deletions between platform/algorithms was notified for NA12878 and NA19240. This was to be expected when using different genotyping arrays as probe location and SNP density vary. On the other hand, this kind of effect can also be caused by simply altering algorithm parameters and CNV definitions.

It has been claimed by inter-population CNV investigations that more copy number variable events were to be expected in African populations rather than in European populations. A small subset of the HapMap collection which consists 9 samples from Europe (CEU), Africa (YRI) and Asia (CHB+JPT) were studied for structural variants, by Cooper (2008) et al and Kidd et al (2008) using SNP genotyping and sequencing, respectively. Similar CNV sharing patterns were found in these two studies, which confirmed the more frequent observation of CNVs in Africans and there was an median overlap of same CNV events between two distinct populations. In the comparison of deletions of NA12878 and NA19240, in most of the cases more deletions for NA12878 were detected than for NA19240, among various platform/algorithm combinations. This violation of predictions might just due to sampling.

When taking the deletions which Kidd et al (2008) detected for the same two samples by ESP sequencing as reference, the comparison of findings between each of the six studies and Kidd et al (2008) showed a variety of concordance that the detected deletions overlaps Kidd (2008)'s results. Despite SNP density difference, most of the signal intensity-based HMM algorithms outperformed genotype analysis methods. The two studies (McCarroll 2005 and Conrad (2006)), which identified deletions in family trios by analyzing SNP genotypes, not only found less events compared with reference but also had lower concordance rate. McCarroll et al 2008 which used intensity-based method had only 15% of the detected deletions in concordance with Kidd et al (2008), suggesting a possible high false-positive rate of their method. But one should also notice the number of overlapping events with Kidd (2008) in McCarroll 2008's study was one of the largest, showing their ability to detect confirmed events. Winchester et al. (Winchester et al., 2009) performed four algorithms (including cnvPartition, GADA, PennCNV and QuantiSNP) on the same set of SNP signal intensity data on one HapMap individual and they also found a significant discrepancy in results among different algorithms.

Considering all six studies together, only about a quarter of deletions from Kidd's ESP sequencing results, which were potentially detectable on SNP genotyping arrays, were also discovered in at least one of the six SNP-based studies. Although the platform and algorithms to detect CNVs from SNP genotyping arrays discussed in this chapter were not exhaustive, this finding revealed the limitation of current SNP-based CNV detection methods to discover true CNVs. Winchester et al also compared their results from multiple algorithms to Kidd (2008)'s deletion predictions and the same low level of consistency was claimed.

Several challenges remain in SNP-based CNV discovery: first of all, on commercial SNP genotyping arrays, probes are not uniformly distributed across the genome and are particularly sparse in regions of segmental duplication and complex CNV regions, for example 10% deletions identified by Kidd (2008) for two HapMap individuals do not cover at least two consecutive SNPs on any of the SNP platforms discussed in this chapter. To overcome this limitation, newer platforms, such as Affymetrix SNP 6.0 and Illumina HumanHap 1M were developed; Affymetrix SNP 6.0 revolutionarily included a huge number of CNV probes in the hope of targeting more common CNVs. However the results showed the concordance rate even for the newer platforms are low. Secondly genotype analysis methods which use genotyping error to predict CNVs are only capable of detecting deletions and require family information so these should be considered as a useful additional method to identify CNVs from genotyping data rather than a direct discovery tool (Carter, 2007). Another major concern for the detection of CNVs utilizing array technology is the definition of putative CNV when assessing shifts in relative signal intensity changes from arrays. Some define a CNV as a change in the intensity of certain number of consecutive SNPs that exceed a pre-defined threshold and some define a CNV through more complex statistical models. The robustness of CNV detection depends on an accurate algorithm which distinguishes a region in which SNPs have unusual signal intensity from the rest of the genome. Last but not least, technical issues such as signal-to-noise ratio and the choice of reference to be compared with putative CNV regions can also cause problems in CNV calling (Carter, 2007).

From the comparison of different platform/algorithm combination for the same subset of HapMap samples in this chapter, one can argue that gaps still exist in software development; it is important to improve the methods to make them more sensitive and powerful in detecting CNV from SNP data. But it is also important to take into account the information different algorithm provided and it is sensible to utilize the different advantages of each algorithm. In two studies comparing CNV calling from multiple algorithm or algorithm/platform combinations which utilized SNP genotyping data for HapMap samples, a large discrepancy in the calling results was also identified (Winchester et al., 2009; Zhang et al., 2011). It is recommended to take a second algorithm on a single dataset to assist to produce the most confident predictions (Winchester et al., 2009). At last, a range of platform/algorithms for the same dataset are presented in this chapter, however one can not definitively confirm the existence of a CNV without independent biological replication.

Chapter 5

Comparison of Methods to detect CNVs from SNP genotyping arrays

5.1 Preface

Various algorithms and programs have been developed to detect copy number variants (CNVs) from genome-wide single nucleotide polymorphism (SNP) arrays. However, their reliability and performance has remained uncertain. I compared four CNV detection methods: QuantiSNP, cnvPartition, PennCNV and DNAcopy using SNP data from 966 individuals genotyped on the Illumina Human Hap 300K array. Both QuantiSNP and PennCNV are based on the Hidden Markov Model (HMM) algorithm, while cnvPartition and DNAcopy are based on Circular Binary Segmentation (CBS) algorithm (details in Chapter1, 1.3.4.2).

It is important to attempt to assess the performance of the analytic methods when applied to data from this less than optimal genotyping platform (Illumina 300K arrays), in being able to identify true CNVs. An approach relying on the concordance rate in duplicated genotyped samples was used to estimate false positive and false negative rates for each of the four analytic methods. It was found that setting a threshold for filtering less reliable CNV calls helped to increase the accuracy and power of CNV detection by QuantiSNP and cnvPartition. It was also stated that QuantiSNP and cnvPartition outperformed other two methods in terms of false positive and false negative rates. Another independent approach was employed which assessed the ability of one of the detection methods, QuantiSNP, to recover true CNVs from genotyping data, based on a HapMap data set from eight humans who had their CNVs determined and whose SNP genotyping data was publically available. The results suggested that QuantiSNP was a conservative method which had a high specificity (low false positive rate) but low sensitivity (high false negative rate).

A discrepancy in the occurrence, length, type of CNVs detected by the different methods was observed. When the CNVRs derived from CNVs called by the four methods were compared it was found that QuantiSNP and cnvPartition had the best concordance in terms of CNVRs detected.

5.2 Materials and methods

5.2.1 Samples

965 samples from Croatia and 1 sample from Orkney were included in the analysis. To assess the false positive and false negative rates of each CNV detection method, duplicate samples were needed. A duplicate is defined here as an individual who was genotyped multiple times (twice) at the same platform. CNV results were independently obtained from each genotyping dataset and then these were compared so that false positive/false negative rates could be calculated. Initially an individual in the Croatian population, Kom388, was selected to serve as a duplicate and was genotyped twice. However, one of the two genotyping datasets did not pass the quality control step and so these data could not be used. For the above reason, an Orcadian individual, ORC2091, was genotyped twice in order to serve as a duplicate sample. All 966 samples had passed quality control (details in Methods chapter).

5.2.2 Parameter setting of the four methods to detect CNVs from SNP array data

CNVs were determined for each sample using QuantiSNP (version 1.0), cnvPartition (version 1.0.2), PennCNV and DNAcopy, respectively. These represented the four analytic methods available at the time of the analysis (between year 2007 and 2008) which had been reported in published articles. QuantiSNP and PennCNV are based on a Hidden Markov Model (HMM), which indicates copy number status of each SNP depending on information from neighbouring SNPs. CnvPartition and DNAcopy are both Circular Binary Segmentation (CBS) methods, which divide the chromosomes into segments to extract segments of aberrant signal intensities which indicate a copy number deviation from the normal copy number of such segments.

PennCNV and DNAcopy didn’t have a filter option whereas various parameter options could be individually selected in the QuantiSNP and cnvPartition software.

The QuantiSNP filter includes times of permutation (EMiters), characteristic length of CNVs (L), GC correction option (doGCcorrect) and threshold of Bayes Factor (Yau, 2007; Colella et al., 2007). QuantiSNP uses an expectation maximization (EM) algorithm to fit HMM model parameters to the data and after permutations the model was optimized and convergence was achieved. The more permutation steps taken, the better the model fits the data but longer computation times are required. The parameter EMiters determines the maximum number of optimization steps to be used. A recommended value of 25 was chosen which was believed to balance both the precision of model optimization and computational time. In CNV detection, longer CNVs may be called with more confidence as LogR ratio and B allele frequencies of more markers were taken into account. This might be expected to reduce the false positive rate of CNV calls. For this reason, QuantiSNP requires a defined characteristic length of CNVs (denoted L) for filtering possible false positives. L=3000000bp was chosen to be the maximum length in this analysis. It had been reported that correcting Log_2R ratio for local GC content would reduce noise and increase accuracy of CNV detection (Colella et al., 2007), so the GC correction option was selected. For each CNV event given by the EM algorithm, a Bayes Factor (BF) was reported to indicate the degree of confidence for the event being of significance. Greater BF values indicate more confidence in the validity of the CNV call. A BF value of above 30 was recommended to reduce false positive calls (Colella et al., 2007). In this analysis both unfiltered CNV calls and CNV calls with BF>=30 were analyzed separately and were compared to assess the possible advantage of setting a BF threshold.

For cnvPartition (version 1.0.2), alterations can be made in setting parameters including "Confidence Threshold", "Include Mitochondrial Chromosomes", "Include Sex Chromosomes" and "Probe Gap Size Threshold". For the analysis which was limited to autosomes the two parameters "Include Mitochondrial Chromosomes" and "Include Sex Chromosomes" were set to be false. The Probe Gap Size threshold is the upper limit of region length between probes and regions within probe gaps whose size is greater than this value would not be considered to be within CNV regions. Setting such a threshold would help prevent CNVs from being called across large probe gaps, such as centromeres. The default value of 1,000,000 bp was adopted in the current analysis. cnvPartition incorporated an algorithm to assign a Confidence Score to each CNV call. A higher Confidence Score value denotes greater confidence in the validity of the CNV call. The recommended threshold for this Confidence Score was 35 (Illumina Manual 3). In this analysis both unfiltered CNV calls and CNV calls with Confidence Threshold of 35 were analyzed separately and were compared to assess the possible advantage of setting a Confidence Threshold.

5.2.3 Computation

QuantiSNP, cnvPartition and PennCNV required Log$_2$R ratio and B allele frequency data for each SNP within each sample while DNAcopy required only Log$_2$R ratio data. HMM based methods such as QuantiSNP and PennCNV are fast with QuantiSNP taking only 6-7 minutes to process per sample. CBS methods are more computationally intensive with a processing time for DNAcopy of 12 to 15 minutes to process a sample. cnvPartition adopted an improved method to decrease computation time, which resulted in a processing time of less than 10 minutes per sample. QuantiSNP, cnvPartition and DNA copy were run on a desktop computer with a 2.76GHz processor and 4GB of RAM. PennCNV were performed on a computer of 1.77GHz processor and 3GB of RAM.

5.2.4 Copy number assignment for DNAcopy

Unlike the other three methods, the DNAcopy procedure didn't result in a list of CNV segments each with a copy number. DNAcopy segments a chromosome according to the Log_2R intensity ratio values and only highlights chromosome regions with abnormal values which could indicate either deletions or amplifications. I adopted a method which was developed for cnvPartition (Illumina, 2007) to determine the copy number status of each potential copy number variant region identified by DNAcopy. Due to the lack of B allele frequency information, each region was assigned either 'deletion (copy number<2)' or 'amplification (copy number>2)' instead of an exact number (0,1,3, 4 etc).

5.2.5 Assessment of sensitivity and specificity of algorithms

The computation of sensitivity and specificity of the algorithms considered in this chapter comprises two parts.

The first part is to test the false positive and false negative rate of CNV detection for each of the four methods (and with different filtering options) by using 966 samples, one of whom had been genotyped twice. In the current study design, the "truth" of individual observations is viewed as unknown (that is the true positives and true negatives and unobserved in the study sample without any further validation). Under these circumstances, a strategy of relying on concordance of replicates was employed (Jakobsson et al., 2008).

In a population of sample size n, in total M CNVs were detected at m loci. Several individuals in that population had been genotyped twice and CNVs were detected independently for each replicated pair, but only one genotyping data for each individual were added to the total population. Denote false positive rate by α and false negative rate by β and these values can be obtained by equations:

$$\alpha = \frac{\tau - \rho\tau + \tau\chi - \rho\tau\chi - \sqrt{\rho\tau(1-\rho)(1+\chi)(2\chi - \tau - \tau\chi)}}{(1-\rho)(1+\chi)} \quad (1)$$

$$\beta = \frac{\rho - \rho\tau + \rho\chi - \rho\tau\chi - \sqrt{\rho\tau(1-\rho)(1+\chi)(2\chi - \tau - \tau\chi)}}{\rho(1+\chi)} \quad (2)$$

Where τ denotes the probability that an allele is called as a CNV, χ denotes the concordance of CNV calls in replicated pair and ρ denotes the probability that a CNV is truly present for a given allele. Equations 1 and 2 provide the basis of estimating false positive and false negative rate. In these cases, α and β are each determined by 3 parameters, τ, χ and ρ. The first 2 parameters can be estimated while only the last one is unknown. ρ can be estimated as:

$$\rho_{estimated}=M/(2\times m\times n) \quad (3)$$

and χ is the number of concordant CNV loci (m_1) divided by total number of CNV loci called in a replicated pair for the same individual (m_2).

$$\chi_{estimated}=m_1/m_2 \quad (4)$$

Inserting the calculated value of ρ and χ, α and β are then presented as functions of unknown parameter ρ. If an approximate range of ρ can be estimated, the range of false positive and false negative rates can be determined.

Define τ as the probability that an allele is called as a CNV, ρ as the probability that a CNV is truly present for a given allele, an allele is called as a CNV either when 1) a true positive CNV present at a true positive loci and 2) the CNV, detected at the loci which wasn't a true location for CNV is a false positive. Therefore there is Equation 5:

$$\tau=\alpha(1-\rho)+(1-\beta)\rho \quad (5)$$

Specificity (also termed 'accuracy'), which equals to 1-α, is defined as the ratio of number of true negatives to number of total negatives (true negatives + false positives). Sensitivity (also termed 'power'), which equals 1-β, is defined as the ratio of number of true positives to number of all positives (true positives + false negatives). Therefore lower false positive rate indicates higher sensitivity and lower false negative rate indicates higher specificity.

Duplicate genotypes were needed for this assessment. An individual from Orkney, ORC2091, had been independently genotyped twice. These data were then combined with data from the 965 Croatian samples for this analysis, making the total sample size 966 individuals. Two sets of genotyping data from the same individual were assessed for CNVs when the true copy number status is unknown. Data from two genotyping panels of the same individual were each processed with QuantiSNP and cnvPartition to yield CNV results. Only one set (set A) CNVs of ORC2091 was combined with Croatian population, the other set (set B) was compared with set A but not CNVs from other samples. All CNVs of each combined data set (Croatian+ORC2091(A)) were mapped on chromosomes; CNV loci are defined as non-redundant chromosomal regions which harboured at least 1 CNV from at least one individual.

The second part is to test sensitivity and specificity of QuantiSNP to detect CNVs, using genotyping data from 8 HapMap samples (NA12156, NA12878, NA15510, NA18507, NA18517, NA18555, NA18956 and NA19129). These individuals had been end-sequenced and two array-CGH platforms were also used to validate CNVs detected by end-sequencing (Kidd et al., 2008). The CNV calls of the 8 HapMap samples were downloaded from Database of Genomic Variants (http://projects.tcag.ca/variation/) as a reference set to represent the 'true' CNVs within those samples. These individuals had also been DNA genotyped on Illumina Human1M arrays which comprises 1,072,820 SNPs SNP intensity files for these 8 humans were downloaded from the Illumina ftp site (http://www. illumina.com/ forms/ftp.ilmn). QuantiSNP was run on the intensity data to generate CNV calls. The CNVs called by QuantiSNP were then compared to those in the reference set for the same individual.

5.3 Results

5.3.1 SNP coverage of the 300K array for verified CNVs

The Illumina HumanHap 300K genotyping array (earlier version) comprised 308,330 autosomal SNPs. To assess the probe coverage of CNV events, verified CNVs from four human genomes, of J. Craig Venter (Levy et al., 2007), James Watson (Wheeler et al., 2008), NA18507 (Bentley and et al., 2008) and YH (Wang et al., 2008), whose whole-genome DNA has been sequenced were extracted to constructed a reference set (details in Chapter 3).

The 308,300 SNPs were then mapped to each of the reference CNVs and the number of overlapping SNPs for each reference CNV was recorded (Table 5.1). It was found that the HumanHap 300 platform had a poor coverage of CNVs; it lacked probes within on average 85% of the reference CNVs in the four genomes, and about 93% couldn't be tagged by multiple probes (Table 5.1).

5.3.2 Occurrence, type, length and frequency of CNVs detected by four methods

SNP data from each of the 965 Croatian samples were processed by QuantiSNP (BF>30), cnvPartition (confidence score>35), PennCNV and DNAcopy, respectively. An overview of number of events, CNV type, length, and CNV burden per sample is given in Table 5.2.

QuantiSNP detected 29964 CNV events, of which the majority were deletions. After applying a Bayes Factor threshold of 30, the number CNVs remaining was 1619. Thus the filtered data reduced the number of CNV calls by about 20 fold. Many of the small CNVs didn’t meet the threshold, which suggested a trend with the QuantiSNP algorithm to give more confidence to longer CNVs. Of the excluded CNVs, most were deletions. This could be explained by the fact that amplifications were in general of longer length than deletions.

Table 5.1 Illumina HumanHap300K genotyping array probe coverage of CNVs in four sequenced individuals.

Author	Subject	Ethic origin	Platform	# CNVs	Mean length (bp)	Length range (bp)	300K array probe coverage**			
							0 SNP	1 SNP	⩾2 SNPs	Missing*
Levy et al (2007)	J.C. Venter	Caucasian	Sequencing & arrays	382	18470	1006 to 920100	345	8	29	0.92
Wheeler et al (2007)	James. Watson	Caucasian	Sequencing& arrayCGH	625	14210	1007 to 1580000	448	84	93	0.85
Bentley et al (2008)	Anonymous	South African	Solexa	693	4072	1002 to 50000	624	57	12	0.98
Wang et al (2008)	Anonymous	Chinese	Illumina Sequencing	494	6227	1004 to 158300	443	30	21	0.96
Average				549	9952	1002 to1580000	465	45	39	0.93

* “Missing” denotes the missing coverage rate which was the proportion of CNVs that didn’t overlap at least two consecutive SNPs on the 300K platform

** “300K array probe coverage” denotes the number of CNVs in the individual genomes covered by 0,1 or ⩾2 SNPs cnvPartition identified more amplifications than deletions before filtering. The confidence score threshold of 35 limited the total number of CNVs to half of the unfiltered number. After applying the confidence score threshold an elevation in median length of both deletions and amplifications was observed, which might suggest that cnvPartition also placed confidence in longer CNVs (as with QuantiSNP).

A discrepancy in the features of CNVs detected by different methods was observed (Table 5.2). QuantiSNP (filtered) and cnvPartition (filtered) identified a similar number of deletions and amplifications; PennCNV identified 2 fold more deletions than amplifications; and DNAcopy identified mainly deletions, with only 11% events being amplifications. The median length of CNVs detected by QuantiSNP and cnvPartition was significantly longer than those detected by the other two methods; DNAcopy identified a large number of smaller deletions, which make the median length of CNVs it detected the shortest among the four methods. The lengths of the CNVs detected by the 4 methods all follow an L shape, with many small events and few large events (e.g. length of CNVs called by QuantiSNP, as in Figure 5.1). Only 1.7 and 3.4 events per sample were detected by QuantiSNP and cnvPartition, respectively (filtered data). DNAcopy detected 11.8 events per sample and PennCNV detected 30.1 events per sample. None of these numbers exceeded 39, which was the average number of true CNV events which can be captured by 2 or more SNPs on the 300K chips, based mainly on the analysis of the sequence data of 4 humans described above.

Table 5.2 Type, length and occurrence of CNVs detected by four methods.

Method		Total events	Deletions	Amplifications	Median length (kb)	Median deletion length (kb)	Median amplification length (kb)	CNV per sample
QuantiSNP	BF>30	1619	747	872	154.4	97	160.3	1.7
	unfiltered	29964	20916	9048	40.1	35.5	52.6	31.1
cnvPartition	confidence>35	3293	1367	1926	136.7	68.6	185.2	3.4
	unfiltered	6328	2707	3626	92.9	56.2	51.6	6.6
PennCNV		29062	19252	9810	50	64.7	53.7	30.1
DNAcopy		11380	10091	1289	11.5	10.5	11.7	11.8

Summary of both unfiltered and selected (Bayes Factor>30) CNVs detected by QuantiSNP are listed. Also unfiltered and selected (confidence score>30) CNVs detected by cnvPartition are listed

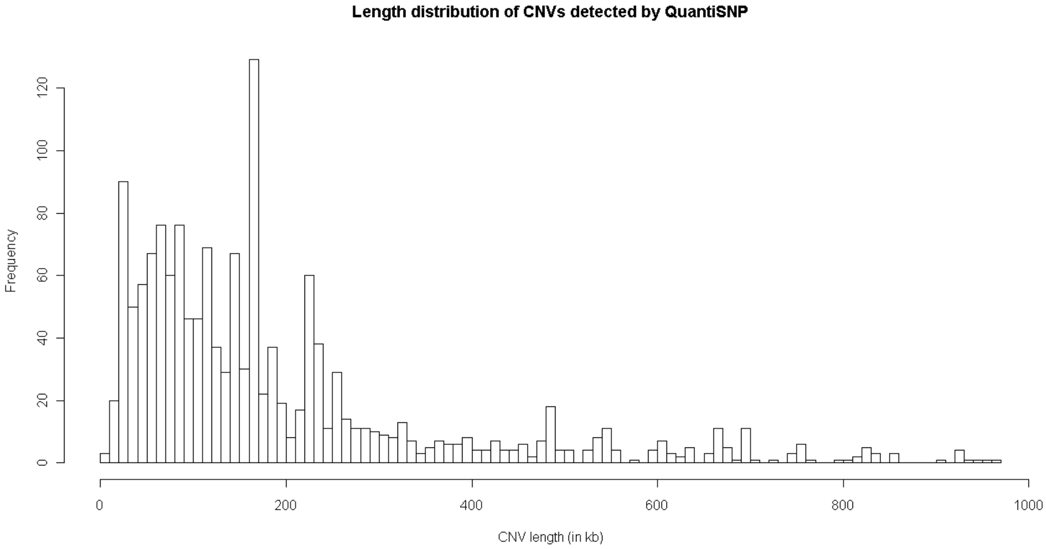


Figure 5.1 Distribution of CNV length (between 1kb and 1Mb) detected by QuantiSNP

5.3.3 Test of the validity of threshold setting parameters of QuantiSNP and cnvPartition

Both QuantiSNP and cnvPartition introduced parameters for quality control purposes, as described in 5.2.2. The unfiltered CNVs detected were far greater in number than the CNVs called after filtering the data (Table 5.2).

29982 unfiltered CNVs at 7484 loci for 966 samples were identified by QuantiSNP. 18 and 33 CNV loci were each detected for the two independent genotyping panels of ORC2091, of which 10 were concordant. The filter of BF=30 resulted in 1621 CNVs at 416 loci for total samples, 2 identical CNV loci were found in the two genotyping panels of ORC2091. The estimated probability value that an allele is called a CNV (τ) in 966 samples and the concordance rate (χ) of two observations for the same individual, ORC2091, were calculated. Details and calculations of τ and χ for cnvPartition were also listed in Table 5.3.

Table 5.3 Estimated probability value that an allele is called a CNV (τ) in 966 samples and the concordance rate (χ) of two observations for the same individual, ORC2091.

Method		965 Croatians	ORC2091 (A)	Croatians+ ORC2091 (A)	Croatians+ ORC2091 (A)	P of an allele called CNV, τ	ORC2091 (B)	ORC2091 (A) and (B)	ORC2091 (A) or (B)	Concordance χ
		#CNVs			#Loci					
QuantiSNP	BF>30	1619	2	1621	416	0.0020	2	2	2	1.00
	unfiltered	29964	18	29982	7484	0.0021	33	10	31	0.32
cnvPartition	confidence>35	3293	2	3295	776	0.0022	2	2	2	1.00
	unfiltered	6328	5	6333	1435	0.0023	5	3	7	0.43
PennCNV		29062	7	29069	5385	0.0028	7	4	10	0.40
DNAcopy		11380	4	11384	3485	0.0017	9	3	10	0.30

The CNVs were detected by four methods; both unfiltered and filtered results were included in the table. ORC2091 (A), the one of two datasets to be combined with 965 Croatians; ORC2091 (B), the other dataset of the duplicated sample. ORC2091 (A) and (B): the number of CNV loci detected in both duplicates (m_1); ORC2091 (A) or (B): the number of CNVs detected at least one duplicate (m_2). τ was calculated according to Equation 3 and c was calculated according to Equation 4

Inserting value of τ and χ into Equations 1 and 2, false positive and false negative rates could be plotted as functions of ρ. A comparison of false positive and false negative was made between the filtered and unfiltered CNV results for QuantiSNP and cnvPartition, respectively (Figure 5.2 and Figure 5.3).

The threshold of 30 notably reduced both false positive and false negative rates of CNVs detected by QuantiSNP (Figure 5.2). The false positive rates for both threshold settings were low, bounded above by 0.22% (unfiltered) and 0.17% (filtered). The false negative rates approximated to zero from ρ=0.0019 and ρ=0.0033, for unfiltered and filtered CNVs respectively.

Similar results were observed with cnvPartition. The performance improved after setting the confidence score filter at a value of 35 both in term of reducing false positive and false negative rates (Figure 5.3). However, the difference before and after applying the filter was not as great as with QuantiSNP: the departure of unfiltered and filtered curves in both Figure 5.3 (a) and (b) was not as great as in Figure 5.2 (a) and (b).

5.3.4 False positive and false negative rates of four methods

In Table 5.3, the number of CNVs, number of CNV loci in the 966 sample and a comparison of the two sets of genotyping data for the same duplicated individual were tabulated. As stated in Section 6.4.3, applying filters significantly improved power and accuracy for both QuantiSNP and cnvPartition; therefore only filtered CNVs were used in this further analysis. The average probability of an allele being a CNV (denoted by τ) and concordance rate in the pair of duplicated datasets (denoted by χ) were calculated based on the information mentioned, each for QuantiSNP (BF>30), cnvPartition (confidence score>35), PennCNV and DNAcopy.

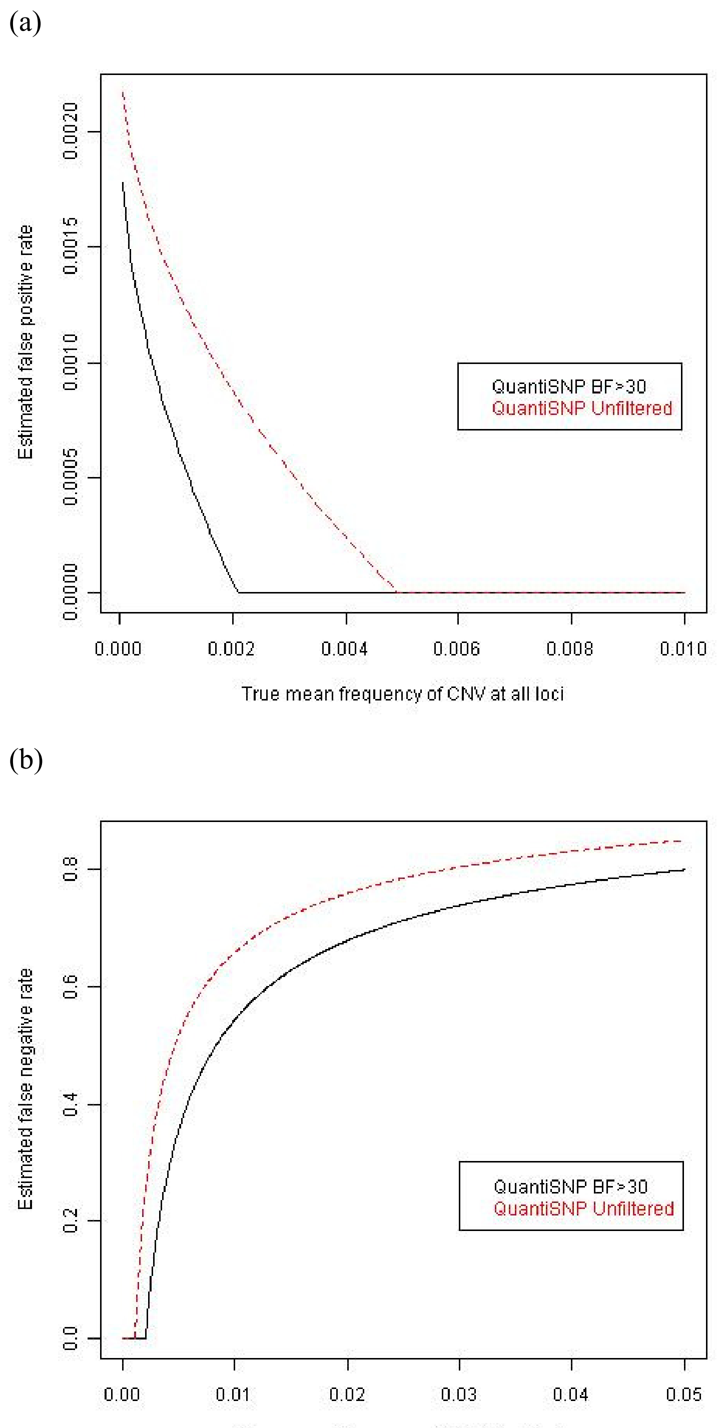


Figure 5.2 Estimated false positive and false negative rates as functions of the unknown true mean frequency of CNVs detected by QuantiSNP, with different threshold settings, across all loci in the 966 samples. (a) False positive rates, (b) False negative rates

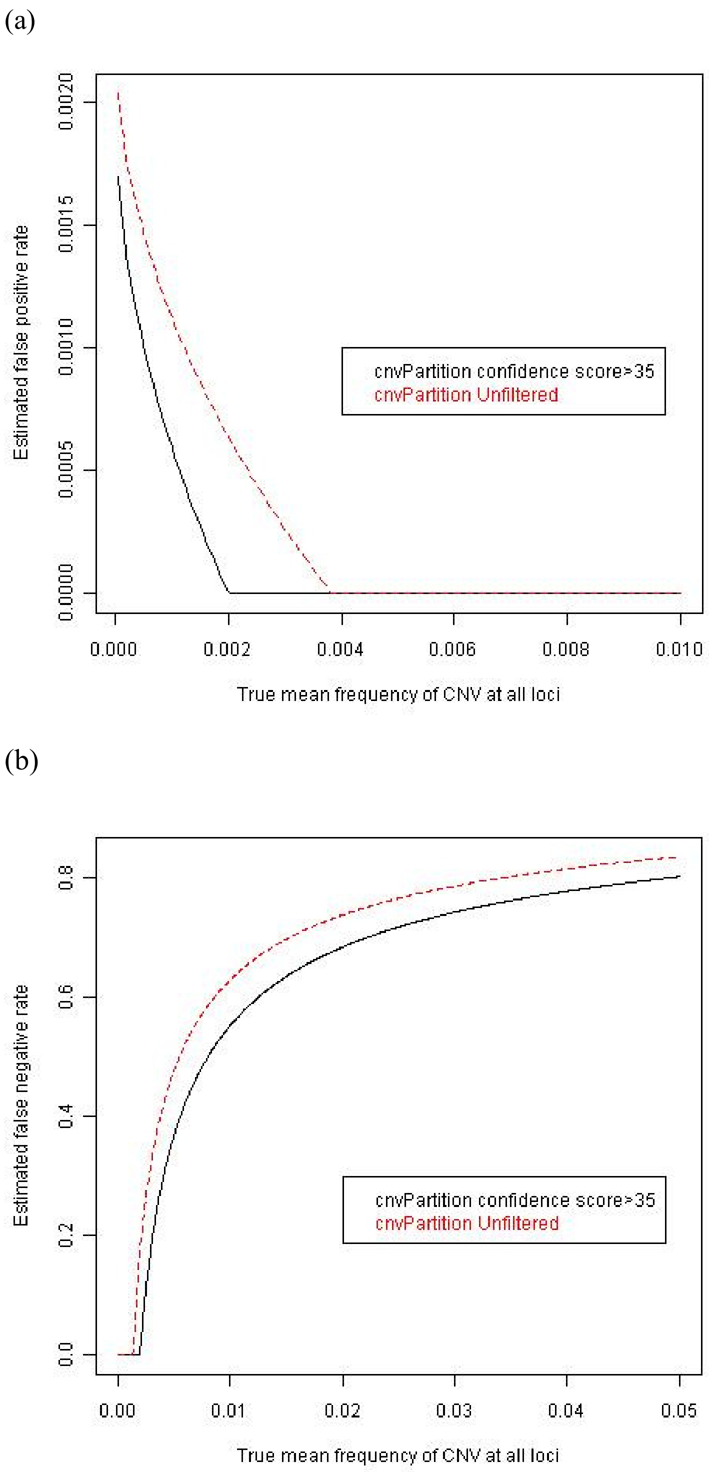


Figure 5.3 Estimated false positive and false negative rates as functions of the unknown true mean frequency of CNVs detected by cnvPartition, with different threshold settings, across all loci in the 966 samples. (a) False positive rates, (b) False negative rate.

The average probabilities of an allele being a CNV across all loci was similar for the four methods were between 0.0017 and 0.0028. QuantiSNP and cnvPartition had very close τ values (0.0020 and 0.0022) while PennCNV had the highest and DNAcopy the lowest values.

QuantiSNP and cnvPartition had the best concordance rate of 1. Both of them detected two identical CNVs from the pair of duplicated datasets of ORC2091. DNAcopy had the lowest concordance rate of 0.3.

With the estimated τ and χ values, the false positive rate α and false negative rate β could be plotted from Equations 1 and 2 as functions of the only unknown parameter, r. The plots of false positive and false negative rates for the four methods were shown in Figure 5.4 and Figure 5.5.

The intercept of the y axis on Figure 6.4 was 0.0018 for QuantiSNP, 0.0017 for cnvPartition, 0.0015 for PennCNV and 0.0025 for DNAcopy. Note that under the assumption for any useful test that the true positive rate 1-β is greater than or equal to the positive rate α, a rearrangement of Equation 5 result in $\alpha \leqslant \tau$, so the false positive rate of DNAcopy should be bounded by the probability t, which being 0.0017 for DNAcopy (Table 5.3). Thus the false positive rates for the four methods were all very low at under 0.18%.

Figure 5.5 showed β values for the 4 analytic methods with ρ ranging between 0 and 0.05. QuantiSNP and cnvPartition seemed to have the lowest false negative rate, while DNAcopy had the highest false negative rate. Even when the probability of an allele being called a CNV was as low as 0.01, DNAcopy lost over 70% of its power to detect true CNVs, while QuantiSNP and cnvPartition still had about 50% of the power.

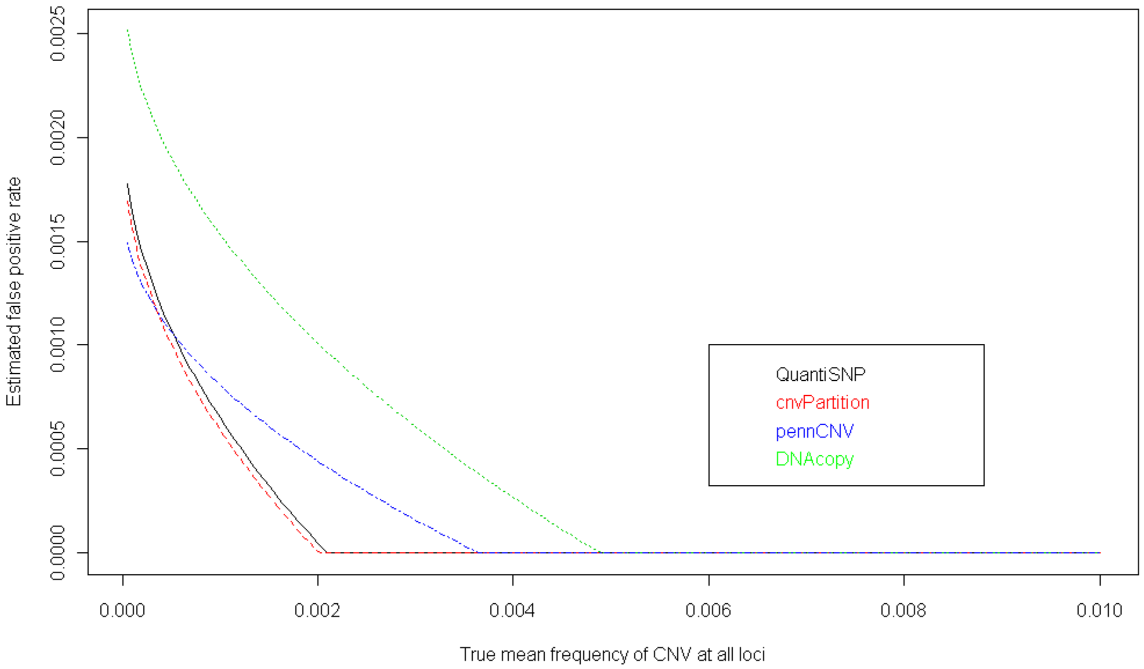


Figure 5.4 Estimated false positive rates as functions of the unknown true mean frequency of CNVs detected by QuantiSNP, cnvPartition, PennCNV and DNAcopy, across all loci in the 966 samples.

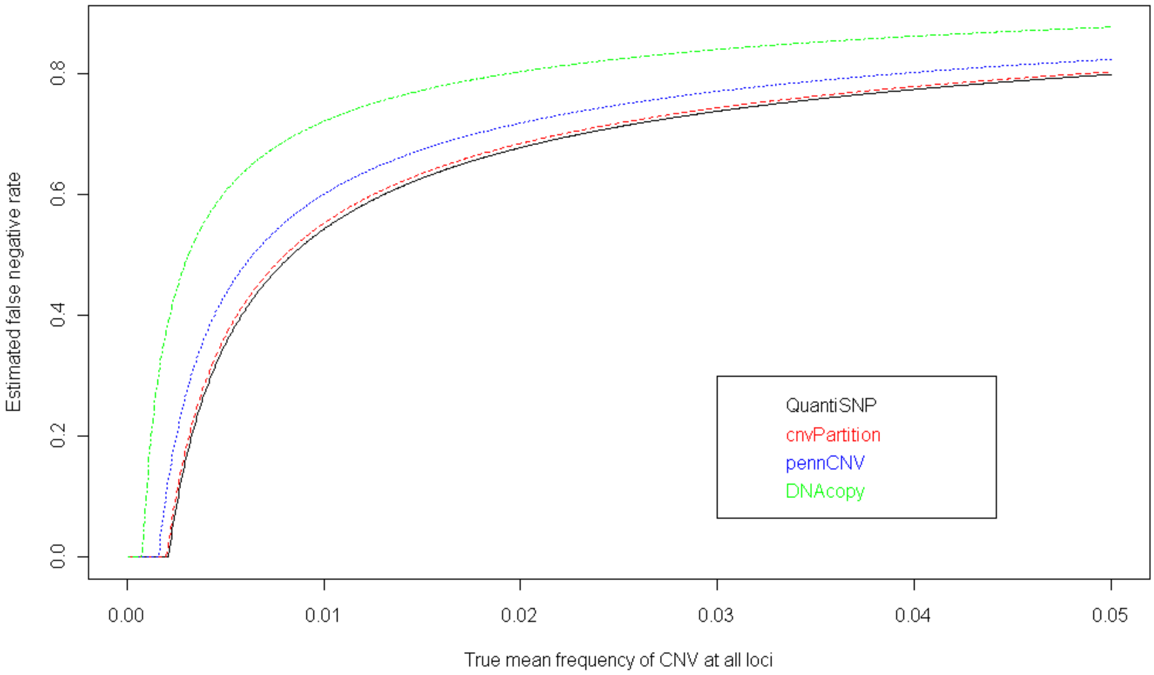


Figure 5.5 Estimated false negative rates as functions of the unknown true mean frequency of CNVs detected by QuantiSNP, cnvPartition, PennCNV and DNAcopy, across all loci in the 966 samples.

5.3.5 Estimation of false positive and false negative rate of QuantiSNP with validated data

1959 autosomal CNVs were confirmed by fosmid end-pair analysis and validated by array-CGH and/or DNA sequencing for these 8 individuals, with median length of 23.4 kb (length range: 701 bp-930kb), which had been downloaded from Database of Genomic Variants.

The SNP intensity files for these 8 humans were downloaded from the Illumina ftp site (http://www.illumina.com/forms/ftp.ilmn). Using the 1M SNP data, 1694 unfiltered autosomal CNVs were detected by QuantiSNP. 207 of them had a BF>30. The median length of these CNVs is 85 kb (length range: 942bp-4981 kb). SNPs in the Human1M array were densely distributed across the genome with the average gap between two SNPs being approximately 3000bp.

67 out of 1959 validated CNVs were recovered in the dataset of CNVs detected by QuantiSNP (the length of the overlapped part should exceed 50% of the length of the shorter sequence to be compared in a pair), while 140 CNVs detected by QuantiSNP were not true. n CNVs segmented a chromosome into n+1 regions of no CNVs, thus the total 1959 CNVs resulted in 8 human autosomes 1959+22×8=2135 non-CNVs. Therefore, the false positive rate of QuantiSNP to detect CNVs in these 8 samples was 140 / (140+2095) =6.55% and false negative rate is 67/(67+1892)=96.6% (Table 5.4).

Table 5.4 The positive, negative, false positive and false negative CNV events detected by QuantiSNP in 8 HapMap samples

		Actual condition (True CNVs)	
		Present	**Absent**
Test result (CNVs detected)	**Positive**	**67** (true CNV detected)	**140** (detected CNVs are not true) **Type I error**
	Negative	**1892** (true CNVs which are not detected) **Type II error**	**2095** (CNVs not detected are not true CNVs)

5.3.6 Concordance of CNVs detected by four methods

For 966 individuals from Vis and Orkney, the 1619 CNVs detected by QuantiSNP were grouped into 430 CNVRs; 3293 CNVs detected by cnvPartition were grouped in 857 CNVRs; 29062 CNVs detected by PennCNV were grouped in 6528 CNVRs; and 11380 CNVs detected by DNAcopy were grouped in 3624 CNVRs (Table 5.5).

All of the CNVRs detected by any of the four methods were combined and aligned together. This resulted in 8873 non-redundant CNVRs along the chromosomes, each detected by one to four methods. Two DNA segments were considered to be overlapped if they mapped to approximately the same location on the genome and the length of overlapping part exceeded 50% of the length of the shorter segments. It was found that a number of CNVRs were detected by multiple CNV detection methods, however the majority of CNVRs were only detected by only one method, especially for PennCNV (4847 private CNVRs) and DNAcopy (2129 private CNVRs) (Figure 5.6).

Table 5.5 Concordance of detected CNVRs between each pair of methods

	QuantiSNP (%)	cnvPartition (%)	PennCNV (%)	DNAcopy (%)
QuantiSNP	430	357 (41.7)	360 (5.5)	263 (7.3)
cnvPartition	357 (83.0)	857	662 (10.1)	493 (13.6)
PennCNV	360 (83.7)	662 (77.2)	6528	1368 (37.7)
DNAcopy	263 (61.1)	493 (57.5)	263 (4.0)	3624

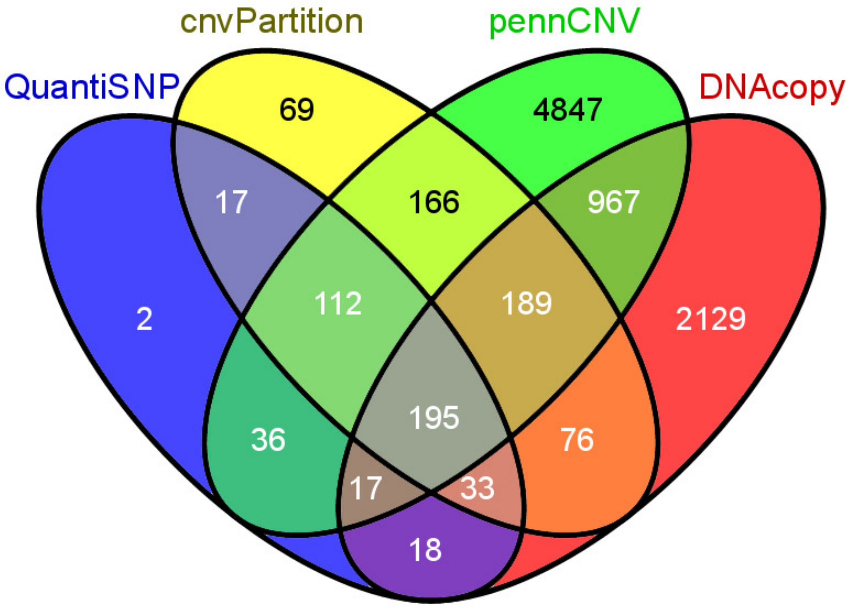


Figure 5.6 Number of overlapped CNVRs detected by QuantiSNP, cnvPartition, PennCNV and DNAcopy

The concordance of CNVRs detected by QuantiSNP was higher with both cnvPartition (357 out of 460) and PennCNV (360 out of 460). 77% of the CNVRs detected by cnvPartition were also detected by PennCNV. Only 25% of the CNVRs detected by PennCNV were also detected by other methods, with the highest concordance with CNVRs detected by DNAcopy. DNAcopy only identified CNVRs of which 40% overlapped with CNVRs detected by other methods (Table 5.5).

Length of overlapped CNVRs among QuantiSNP, cnvPartition and PennCNV are shown in Figure 5.7. DNAcopy result was excluded from this analysis, because it had poor concordance with other three algorithms and it has highest false positive and false negative rate in detecting CNVs, indicated in section 5.3.4. 74.8Mb of 357 CNVRs was detected by both QuantiSNP and cnvPartition, 185 Mb of 360 CNVRs was detected by both QuantiSNP and PennCNV, and 131.4Mb of 662 CNVRs was detected by both cnvPartition and PennCNV. 65.7Mb of 307 CNVRs was detected by all three methods, which in length was 87.8% of the overlap between QuantiSNP and cnvPartition, 50% of the overlap between cnvPartition and PennCNV, and 35.5% of the overlap between QuantiSNP and PennCNV.

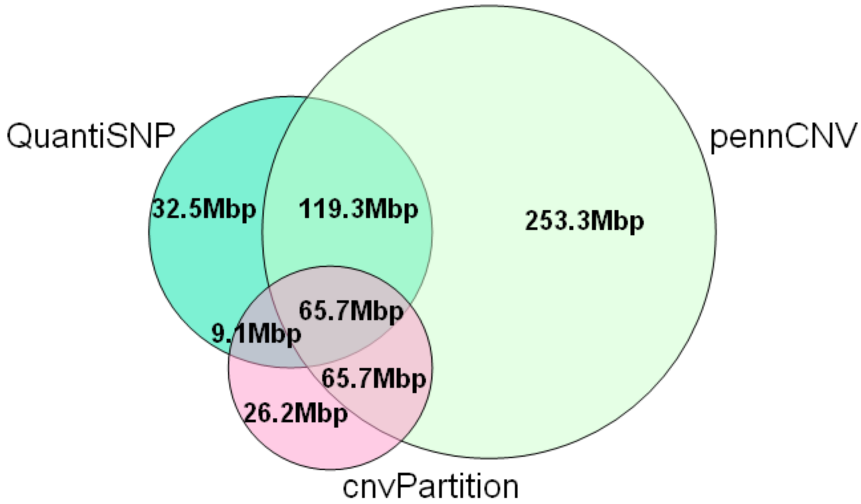


Figure 5.7 Total lengths of CNVRs detected by QuantiSNP, cnvPartition and PennCNV

5.4 Discussion

The SNPs contained in Illumina 300K arrays were mapped to genomic regions which showed evidence of CNV in four whole-genome sequenced genomes, to determine SNP coverage of the 300K arrays. On average only 39 CNVs per genome (7.1%) were covered by multiple (⩾2) SNPs. This would be the maximum number of CNVs per genome which could possibly be detected on the 300K platform. In practice, this proportion might further decrease due to limitations in the sensitivity of CNV detection algorithms to detect CNVs using SNP genotyping arrays. However, limitations of sequencing could lead to false negative observations (discussed in Chapter 3) and the CNV profiles in different individuals might vary. This may result in an overall underestimation of the true number of CNVs in these four genomes.

A statistical model, which calculates false positive and false negative rates based on concordance of replicated data with population CNV information, was utilized in the analysis of this chapter. The limitations of this analysis include: 1) the assumption was that the CNVs were common therefore might not hold well for rare CNVs; 2) there was only one duplicate for the current study, which may introduce bias in the estimation of concordance rate; 3) the range of true mean frequency of CNV at all loci for a population was uncertain, which makes the calculation of power difficult. Notwithstanding these limitations, a comparison was made for the same method, sample samples and same duplicate but only varying CNV detection algorithms and parameter setting.

It was shown that setting filtering parameters for QuantiSNP and cnvPartition resulted in lower false positive and false negative rates for CNV detection. Applying can be recommended for use with these analytic approaches.

The result of the comparison of false positive and false negatives, when the true status of an allele being a CNV at a locus is unknown, indicated that for the 966 samples studied, QuantiSNP and cnvPartition outperformed other two analytic methods and DNAcopy had the worst performance of the four methods (Figure 5.4 and Figure 5.5).

These very low false positive rates suggested the majority of errors came from false negatives. However, one should note that this estimation is based on the very low probability of that an allele is called as a CNV (Table 5), which may not always be the true case. For example, in the 8 HapMap samples described above, in total 2444 true CNVs at 1368 loci represent a probability of 0.11, which is much larger than the estimated rates in Table 5 without known true copy number status of each loci in the population. The false negative rate is hard to determine due to the uncertainty in ρ. QuantiSNP was only robust in detecting CNVs if the average frequencies of true CNVs were low: to achieve the power of 80% (β=0.2), the true frequency should below 0.0033 if applying a filter; at the same frequency, power for detecting CNV by QuantiSNP without setting a BF threshold would be only 60%. QuantiSNP without setting this threshold lost its power much faster than performing with a filter.

The approach discussed above was based on the assumption that the true probability of an allele being a CNV at a locus was unknown, which coincided with the real situation of the current study: CNVs were detected from SNP genotyping arrays without further validation, so the authenticity of those CNVs could not be assessed directly. To directly access performance of CNV detection algorithms, a dataset of published CNVs for 8 human genomes was constructed as a reference set. This reference set was used to demonstrate performance of QuantiSNP, which is one of the better algorithms of the four based on result from former analysis (section 5.2.5). The false positive rate of QuantiSNP was far higher than the estimates in section 5.2.5 (<0.2%) but it was still in an acceptable range which promised reasonable accuracy of detecting true CNVs. However, the false negative rate was very high, indicating that there was the risk that QuantiSNP would discard a high proportion of true CNVs. It is noted that all 1959 laboratory validated CNVs were included in this analysis, some of which might be unable to be detected by 1M array therefore the false negative rate could be inflated. The 1M array SNP coverage of CNVs in these 8 genomes was not assessed in the current study, but previous study suggested 60% of the deletions from the same 8 HapMap samples were covered by the SNPs on Illumina 1M array platform (Kidd et al., 2008). Based on the estimate above, the majority of errors were false negatives in which case many true CNVs were not detected by QuantiSNP. The low false positive rates and high false negative rates suggested the algorithm and filter which QuantiSNP adopted was conservative. Thus this limitation resulted in a low detection rate (sensitivity) for QuantiSNP to detect true CNVs from SNP genotyping data.

The comparison of overlapped CNVs detected by four methods showed that the overlap of results from different algorithms was low. QuantiSNP and cnvPartition had higher percentage of CNVs detected also by another algorithm (Table 5.5). Events detected by multiple methods were considered to be of more confidence that the events were true (Winchester et al., 2009); 65.7Mb of 307 CNVRs was detected by QuantiSNP, cnvPartition and PennCNV, which in length was 87.8% of the overlap between QuantiSNP and cnvPartition, compared to the percentage of those more confident CNVs to the overlap between cnvPartition and PennCNV, and the overlap between QuantiSNP and PennCNV, QuantiSNP and cnvPartition had the best concordance in CNV detection.

After the point of time of the method comparison in current study, several studies published results for comparison of CNV calling algorithms (Dellinger et al., 2010; Tsuang et al., 2010; Winchester et al., 2008; Zhang et al., 2011). Winchester et al. used Affymetrix 6.0 and Illumina 1M Duo data from a well characterized CEPH sample, NA10861, to test the performance of seven algorithms: Birdsuite, CNAT (Chromosome Copy Number Analysis Tool, Affymetrix, Inc.), cnvPartition, GADA, Nexus (Biodiscovery Inc.), PennCNV and QuantiSNP. The overlap of results from any two algorithms range from 2% to 100%, mostly below 60% with highest resemblance between data generated on the same genotyping platforms. Taking the structural variants identified by fosmid end-pair sequence (EPS) method for the same individual (Kidd et al., 2008) as reference, the false positive rate (based on lack of overlapping with the EPS result) range from 51% to 80%. Using the events detected by Kidd et al (2008). on another sample, NA15510 as reference, the result from four algorithms, cnvPartition, GADA, PennCNV and QuantiSNP, showed the false negative rates were between 77% and 96% (Winchester et al., 2009). Dellinger et al. ran each of the seven methods on their 10 samples from Myopia case-control study: CBS, CNVFinder, cnvPartition, gain and loss of DNA, PennCNV and QuantiSNP. They evaluated statistical power, false positive rates and receiver operating characteristic (ROC) curve residuals by simulation studies. They showed that QuantiSNP outperformed other methods based on ROC curve; Nexus had low specificity and high power; PennCNV detected the fewest numbers of CNVs (Dellinger et al., 2010). Tsuang et al. compared outputs from four algorithms, QuantiSNP, cnvPartition, pennCNV and HelixTree for 48 Caucasian schizophrenia cases and 48 matching controls. They found substantial discrepancy in the results from different algorithms, from the aspects of total number, size, and number per person of events (Tsuang et al., 2010). Zhang et al. evaluated the performance of four software packages, Birdsuite, Partek, HelixTree and PennCNV in two datasets, one consists of 90 HapMap CEU sample and the other of 1001 bipolar cases and 1033 controls. Birdsuite recovered the highest percentages of known HapMap CNVs of median length. It also called the most CNVs consistent with qPCR validation in one CNV region, but the accuracy in other two regions was extremely low. Birdsuite and Partek predicts more rare event than other algorithms (Zhang et al., 2011).

Despite difference in choice of algorithms to be included in each study, all method comparison studies mentioned above, together with the current study suggest nonnegligible discrepancy of results from different CNV calling algorithms for SNP data. Therefore it is advised to choose appropriate algorithms with caution in the planning stage of CNV research. The number and features of CNV called depends on algorithms utilized, and also choice of filtering settings. For example, Zhang et al. found PennCNV called fewer events than QuantiSNP, which is contrary to the finding in the current study; this discrepancy is explained that Zhang et al. used a lower filtering threshold for QuantiSNP outputs, which leads to much larger number of events detected compared to those might have been resulted from a higher filtering threshold (as in the current study). Combining results from two algorithms is recommended (Winchester et al., 2008), however, while decreasing false positive rate, it might also increase false negative rate. Without a ‘true’ gold standard (complete set of CNVs in the whole genome), the sensitivity and specificity of any particular algorithm or combination of algorithms are impossible to estimate accurately. More comprehensive and sophisticated CNV detecting method and denser SNP coverage of genotyping platform (for example Affymetrix SNP 6.0) are desired, and a reliable reference set of CNVs in well defined sample genomes is needed to assess the performance of detection algorithms.

Chapter 6

Copy Number Variation across European Populations

6.1 Preface

Copy Number Variation (CNV) is defined here as DNA segments of 1kb or longer in length and present at variable copy number in comparison with a reference genome (Redon et al., 2006). CNVs are commonly found in the genomes of human and other species (Cutler and Kassner, 2008; Dopman and Hartl, 2007; Fadista et al., 2008; Zhang et al., 2009). To date, 35% of the human genome demonstrates evidence of coverage by CNVs (Database of Genomic Variants, DGV, http://projects.tcag.ca/variation/). It is suggested that CNVs, in the form of deletions, insertions, duplications and complex multi-site variants, may contribute to human phenotypic variation, either directly by gene dosage and proportionate variation in gene expression (Stranger et al., 2007), and/ or indirectly through a) position effects on expression levels *per se* or developmental patterns of expression, or b) by affecting recombination rates and thus genome evolution (Redon et al., 2006).Indeed, several studies have reported evidence for a direct contribution of CNVs to complex disease phenotypes in human populations, such as Schizophrenia and Autism (Int Schizophrenia Consortium, 2008; Pinto et al., 2010; Sebat et al., 2007), and in other species (Garshasbi et al., 2008; Kamatani et al., 2008; Williams et al., 2010; Yang et al., 2009; Perry et al., 2007; Jackson et al., 2007; Pielberg et al., 2002; Norris and Whan, 2008).

Copy number variation can be directly assayed by quantitation of hybridisation to specialist oligonucleotide (Bailey et al., 2008; Cowell and Lo, 2009) or clone arrays (Fiegler et al., 2006) or by direct genome sequencing (Bentley and et al., 2008; Wang et al., 2008), but also conveniently extracted from single nucleotide polymorphism (SNP) array data (Jakobsson et al., 2008; Cooper et al.,2008). As well as being applied to the search for genetic contributions to disease phenotypes, several studies have provided global estimates of CNV frequency and distribution in HapMap samples ((Redon et al., 2006; Stranger et al., 2007) and large population cohorts (Jakobsson et al., 2008; Zogopoulos et al., 2007; Franke et al., 2008; McQuillan et al., 2008), but relatively little attention has been given to potential variation within major population groups. Comparisons of CNV frequency and distribution between independent studies have also been hampered by discrepancies in study design, platform choice and analytical methods between studies.

Geographical population isolates are valuable resources for the dissection of complex genetic traits and disease outcomes (Peltonen, 2000; Shifman and Darvasi, 2001; Wright et al., 1999) Genetic isolates have reduced genetic heterogeneity, as measured by fewer net mutations and numbers of polymorphic SNPs compared with outbred populations (Shifman and Darvasi, 2001). Furthermore, by virtue of population bottlenecks, genetic drift and high kinship, each isolate will have a different evolutionary history and thus different genetic makeup. For example, isolate populations have been reported to show increased linkage disequilibrium and reduced haplotype diversity relative to outbred populations, consistent with reduced effective population size and increased genetic relatedness (Vitart et al., 2006).

Here, I take the opportunity provided by the EUROSPAN project (Mascalzoni et al., 2009) which brings together several groups working on the genomic and phenotypic analysis of population isolates across Europe. Our objective was to make use of high density genome-wide genotyping data to describe and compare frequencies of each CNV and their distribution within and between these population isolates, and thus determine to what extent CNVs can be used as measures of relatedness and identifiers of population origin. Using Illumina whole genome data with more than 300,000 SNPs from each of three European population isolates, spanning from Northern to Southern Europe, 4016 CNVs in 1964 individuals were detected, which clustered into 743 copy number variable regions (CNVRs). The frequency and distribution of these

CVNRs was compared and shown to differ significantly between the Orcadian, South Tyrolean and Dalmatian populations. Consistent with the inference that this indicated population-specific CNVR identity and origin, it was also demonstrated that CNVR variation within each population can be used to measure genetic relatedness.

6.2 Materials and Methods

6.2.1 Study sample

2789 individuals with data passing quality control (QC) from the island of Vis, Croatia (the CROAS study (Vitart et al., 2006), n=965), the Orkney Isles, Scotland (The Orkney Complex Disease Study, ORCADES (McQuillan et al., 2008), n=691) and South Tyrol, Italy (The Genetic Study of Three Population Micro-isolates in South Tyrol, MICROS (Pattaro et al., 2007), n=1133) are included in the CNV analysis. These studies followed similar study procedures as part of the EU FP7 EUROSPAN study (Mascalzoni 2010) All three projects were approved by the relevant ethics committees. Data collection was carried out between 2003 and 2007 in the three locations. Informed consent and blood samples were received from all study participants. (See Chapter 2 for details).

6.2.2 Genotyping

The Dalmatian samples were genotyped on the Illumina Infinium HumanHap 300 v1 platform while the Orcadian and South Tyrolean samples were genotyped on the Human Hap 300 v2 platform (Illumina, San Diego, CA, USA). The genotyping was done in two sites: Individuals with less than 90% call rate were removed. Sex checks and IBD sharing between first- and second-degree relative pairs were performed with the PLINK program (http://pngu.mgh.harvard.edu/purcell/plink/) (Purcell et al., 2007), and individuals with discordant pedigree and genomic data or falling outside expected ranges were removed from the study. SNPs on the sex chromosomes were excluded. Finally 300,938, 309,200 and 308,396 SNPs remained in Dalmatian, Orcadian and South Tyrolean datasets, respectively.

6.2.3 CNV calling

For each individual, the Log_2R ratio and B allele frequency of each SNP were processed by QuantiSNP and cnvPartition software to generate CNV calls.

The two independent sets of CNV calls made for the same individual were then assessed. The output from QuantiSNP and cnvPartition both provide information for each CNV on the chromosome number and chromosomal coordinates of the start and end of each CNV (breakpoints). One sample possessing >35 CNVs detected by cnvPartition was excluded from the further analysis. Genomic coordinates of each CNV detected in each person were mapped to hg18 sequence assembly using LiftOver (http://genome.ucsc.edu/ cgi-bin/hgLiftOver).

SNP coverage in centromeric regions is very low, thus CNVs called in these regions are likely to be false positive. For this reason all the CNVs spanning centromeres were excluded from the analysis (according to the coordinates of centromeres on each chromosome). CNVs smaller than 1kb or larger than 3Mb were excluded.

QuantiSNP and cnvPartition outputs were combined to produce a list of sample wise CNVs. A confirmed CNV call was made if 1) the CNV was identified by both methods at the same locus and the overlap indicated by both methods exceeds 50% in length; 2) the type of a copy number change event (copy number loss or copy number gains) called by both methods was consistent and 3) overlap length was between 1000 bp and 3Mbp. The boundaries of a CNV were taken as the beginning and end of the overlapped section.

To locate CNVs on chromosomes, individual-wise CNVs were merged into Copy Number Variable Regions (CNVRs). A CNVR is the maximum region shared among all individuals carrying a CNV at the same locus.

6.2.4 Haplotype and SNP tagging

9 and 22 CNVRs from Vis and Orkney, respectively, each with a population frequency of >1%, were analyzed with Plink (http://pngu.mgh.harvard.edu/~purcell/ plink/) (Purcell et al., 2007). SNP genotyping data were exported from BeadStudio and merged with CNV genotypes of the same individuals. Tagging SNPs were investigated with a window size of 3Mb spanning each CNVR. For each CNVR, the adjacent SNPs 1Mb upstream and downstream to the genomic location of each CNVR were selected in haplotype analysis.

6.2.5 Genetic clustering analysis

Genetic clusters of a selected set of CNVRs, in which each CNVR was shared by two or more individuals, were inferred by the software Structure [35], under assumptions of admixture, correlated allele frequencies and no prior population information. For each number of clusters (K) from 2 to 4, a Burnin length of 10,000 iterations followed by 10,000 Markov Chain Monte Carlo iterations was used. The second order rate of change of logarithmic probability of data between subsequent K values was estimated to identify the optimal number of clusters in the data.

6.2.6 Analysis of CNV kinship correlation

The kinship coefficient is a measure of overall genetic similarity relative to some base population in two diploid organisms.

For each population, P, with T individuals in total, suppose there are N CNVRs: $CNVR_1$, $CNVR_2$, …, $CNVR_N$, each with M_1, M_2,…,M_N CNV carriers ({M}>=2 and {M}<T). For the nth CNVR ($1 \leq n \leq N$), $CNVR_n$, there are Mn people carrying the same CNVR.

Extract a sub kinship matrix from the population kinship matrix with those carriers C_1, C_2, …, C_{Mn} for CNVRn:

$$
\begin{array}{cccccc}
 & C_1, & C_2, & C_3, & \ldots, & C_{Mn} \\
C_1 & 0.5 & - & - & \ldots & - \\
C_2 & k_{12} & 0.5 & - & \ldots & - \\
C_3 & k_{13} & k_{23} & 0.5 & \ldots & - \\
: & : & : & : & & : \\
C_{Mn} & k_{1Mn} & k_{2Mn} & k_{3Mn} & \ldots & 0.5
\end{array}
$$

This is a Mn*Mn matrix, which is symmetrical around the diagonal line. Let k_{ij} denote the pairwise kinship coefficient between individuals C_i and C_j (i={1,2,3,…Mn}, j={1,2,3,…Mn}). At the diagonal line of this matrix, $k_{ij}|i=j$ =0.5, because when considering the probability of a random chosen allele to be IBD between two identical genomes, the same allele can be drawn twice.

In this sub-matrix for CNVRn, let *Kn* denote the non-redundant collection of all pair-wise kinship coefficients between any two individuals out of all Mn carriers.

$Kn=\{(k_{12}), (k_{13}, k_{23}), (k_{13}, k_{23}, k_{33}), \ldots(k_{1Mn,} k_{2Mn}, k_{3Mn,\ldots}, k_{(Mn-1)Mn})\}$

Let *Kpop* denote the non-redundant collection of all pair-wise kinship coefficients between any two individuals out of all T individuals in the population

$Kpop=\{(k_{12}), (k_{13}, k_{23}), (k_{13}, k_{23}, k_{33}), \ldots(k_{1T,} k_{2T}, k_{3T,\ldots}, k_{(T-1)T})\}$

Therefore *Kn* has (Mn-1)! elements and *Kpop* has (T-1)! elements.

Then a t-test is performed to test the difference of means between *Kn* and *Kpop*. The probability, p_n is calculated to indicate significance of this difference. A permutation procedure is taken to adjust p_n: another Mn*Mn matrix is randomly drawn from population kinship matrix, with the pair-wise kinship coefficients

$Krandom=\{(k_{12}), (k_{13}, k_{23}), (k_{13}, k_{23}, k_{33}), \ldots(k_{1Mn,} k_{2Mn}, k_{3Mn,\ldots}, k_{(Mn-1)Mn})\}$

A p value, p_{perm} is obtained from a t-test of comparing means of *Krandom* and *Kpop.* The same random process repeats 1000 times, result in 1000 P_{perm} values. p_n is then ranked among the permutated p values, the adjusted p_n, $p_{nadjust}$ is the number of permutated p values which do not exceed p_n, divided by the number of permutations.

6.2.7 Statistical analysis

The reference CNV list was downloaded from DGV. The record of known genes and recombination rates in the human genome was downloaded from the UCSC genome browser. Intra- and inter-chromosomal segmental duplications (SDs) of >90 identity and >1kb in length, which cover 150.8Mbp of human genome (5.3%) (She et al., 2004; Bailey et al., 2002) were downloaded from the public Segmental Duplications Database (http://humanparalogy.gs.washington.edu/, build 36).

All calculations and alignments were performed with the R 2.10.1 software package, with scripts compiled by myself. The test of difference in means was conducted using student's t-test for normalized data or the non-parametric Mann-Whitney U test, significant threshold set to 0.05.

6.3 Results

6.3.1 Overview of copy number variation in Dalmatian, Orcadian and South Tyrolean populations

The study samples were recruited from three populations across Europe, namely the Island of Vis, Croatia, Orkney Islands, Scotland and South Tyrol, Italy (Figure 2.1). 2789 individuals who passed quality control were included in the analysis. To generate more informative results (Winchester et al., 2009), two algorithms, QuantiSNP (Colella et al., 2007) and cnvPartition (Illumina Manual 2) were utilized to detect CNV events from SNP genotyping data (see chapter 5 for details). The combined analysis of CNV calling by QuantiSNP and cnvPartition software (see Methods) identified 4016 autosomal CNVs in 1964 individuals, out of the total 2789 samples, which makes 70.4% of them CNV carriers, with an average number of 2.05 detectable CNVs per carrier. 7.8% of the all autosomal SNPs were covered by CNVs. A correlation of SNP density and CNV length was observed, with higher SNP density in shorter CNVs and lower SNP density in longer CNVs ($p<2.2*10^{-16}$).

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). Equal numbers of amplification and deletion events were detected in each of the populations (Table 6.1). The overall length distributions of observed CNVs were also very similar between the three population isolates (Figure 6.1). Most CNVs were small in length (94.1% of the CNVs were between 1kb to 300kb, mean length was 205.1kb, Table 1 and Figure 2).The lengths of amplifications (259kb) were significantly greater (Mann-Whitney U test, P<2.2*10-16) than those of deletions (142.4kb) (Table 6.1). 3778 out of 4016 CNVs (94.1%) overlapped with CNVs reported in the Database of Genomic Variants.

The 4016 CNVs (Appendix 1) were clustered into 743 non redundant CNVRs (Appendix 2) which covered a total of 187.95 Mb (6.6%) of the 22 autosomes. 649 CNVRs (87.3%) overlap reported CNVs in DGV. Most of the CNVRs contained either only deletions or only amplifications, but 59 regions harbored both types of variants (Table 6.2). In these 'gain-and-loss' CNVRs, all of them contained at least one pair of CNVs whose boundaries were not equivalent from two individuals.

Table 1. Characteristics of Copy Number Variants (CNVs) in Dalmatian, Orcadian and South Tyrolean populations

Population	Sample size	CNV carriers (percentage of carriers in population)	Number of CNVs	CNVs per person	Amplifications	Deletions	CNV mean length (kb)
Vis	965	702 (72.7%)	1384	1.43	803	581	216
Orkney	691	367 (53.1%)	630	0.91	324	306	192.6
South Tyrol	1133	895 (79.0%)	2002	1.77	1033	969	201.6
Combined	2789	1964(70.4%)	4016	1.44	2160	1856	205.1

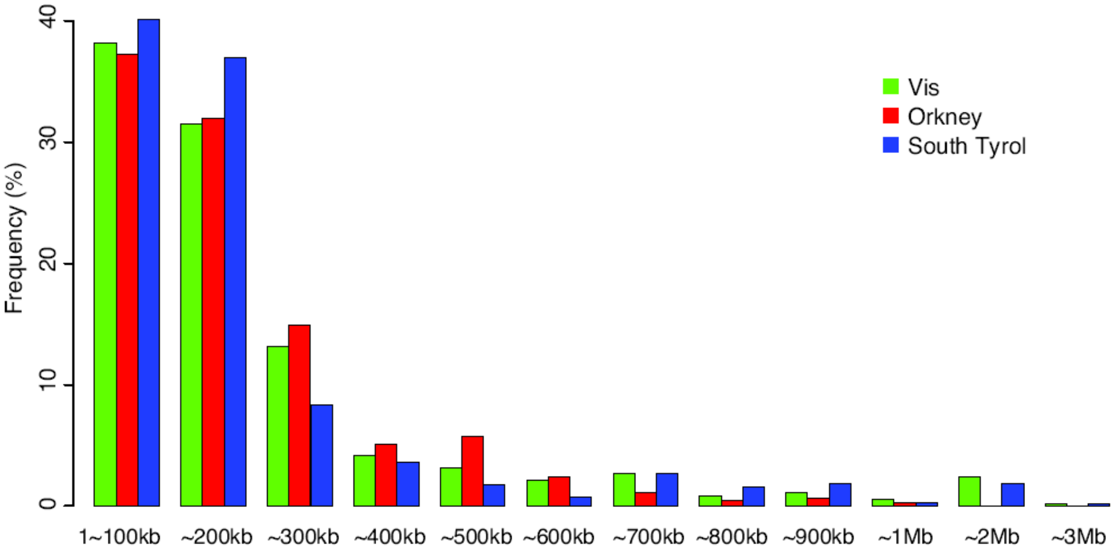


Figure 6.1 Distribution of CNV lengths in the three genetic isolate populations.

Table 6.2 Copy Number Variable Regions (CNVRs) in the three genetic isolate populations

Population	**Number of CNVRs**	**CNVRs overlapping reported regions**	**Number of deletion only CNVRs**	**Number of amplification only CNVRs**	**CNVRs of both deletion and amplification**	**CNVR mean length (kb)**
Vis	365	332	184	164	17	304.5
Orkney	210	193	93	105	12	281.8
South Tyrol	380	334	156	207	17	256.9
Combined	743	649	323	361	59	253.0

6.3.2 CNV frequency and CNV sharing among populations

Each CNVR was found in from 1 to 253 individuals, which made the overall frequency range of CNVRs to be from 0.00051 to 0.12882 (median=0.00102). The CNVs identified were generally of low frequency. 337 CNVRs (45.4%) were detected in only one individual and 321 (43.2%) were shared by between 2 and 10 individuals. Only 37 CNVRs (5%) were present at a frequency >1% in all three population isolates.

Different patterns of CNV frequency were observed in different populations (Figure 6.2); 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; 96 CNVRs were shared by two of the three populations (57 between South Tyroleans and Dalmatians, 25 between South Tyroleans and Orcadians, and 14 between Dalmatians and Orcadians); and 59 were present in all three populations, none of which were novo. Less than half of these population-specific CNVRs (279 out of 588) were reported previously, according to DGV. Rare CNVs were found to be mostly restricted to a single population, while more frequent CNVs were often shared by two or three populations (Figure 6.3a). A gradual increase of population mixture was observed as the frequency of CNVRs increased: more common CNVRs were often shared in more than one population whereas lower frequency CNVRs were more likely to present in a single population (Figure 6.3b). The more frequent CNVRs in one population (population frequency>1%) were often observed to be also frequent in other populations. In South Tyrol, the frequencies of more common CNVs closely correlated with those of Dalmatian and Orcadian CNVs (Pearson's r=0.73, P=7.5*10^{-18} and r=0.43, P=0.005, respectively); the frequent Dalmatian CNVs also correlated with the frequent Orcadian and South Tyrolean CNVs (Pearson's r=0.62, P=0.001 and r=0.65, P=5.2*10^{-4}, respectively), but there was no significant correlation between Orcadian and either Dalmatian or South Tyrolean CNVs of frequency>1% (Pearson's r=0.38, P=0.1347 and r=0.22, P=0.4046, respectively).

Of the 588 population specific CNVRs, more than half (337 CNVRs) contained only one CNV event. The mean length of CNVs in those population specific CNVRs was 250.3kb, 205.5kb and 195.6kb in length, for Vis, Orkney and South Tyrol, respectively, which were on average longer than the ones for shared CNVRs (mean length 198.4kb) (P=0.04).

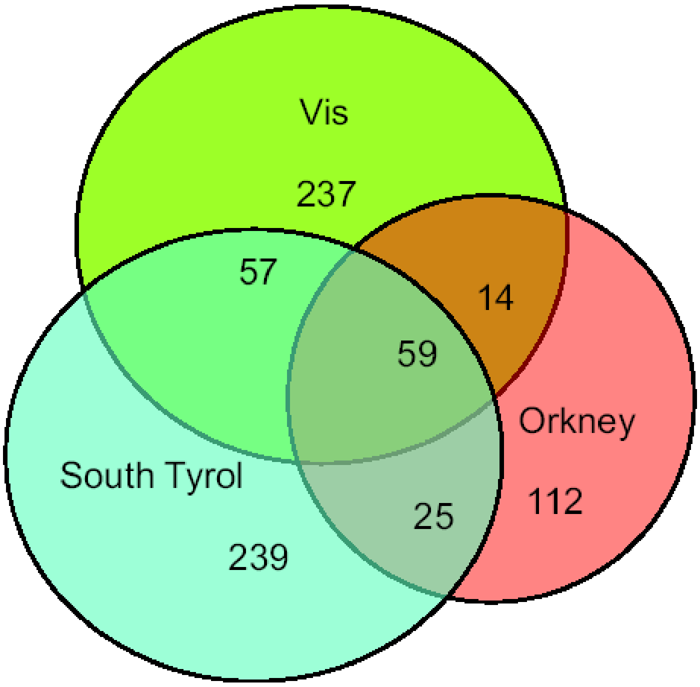


Figure 6.2 Venn diagram showing the number of CNVR shared between the three European genetic isolate populations.

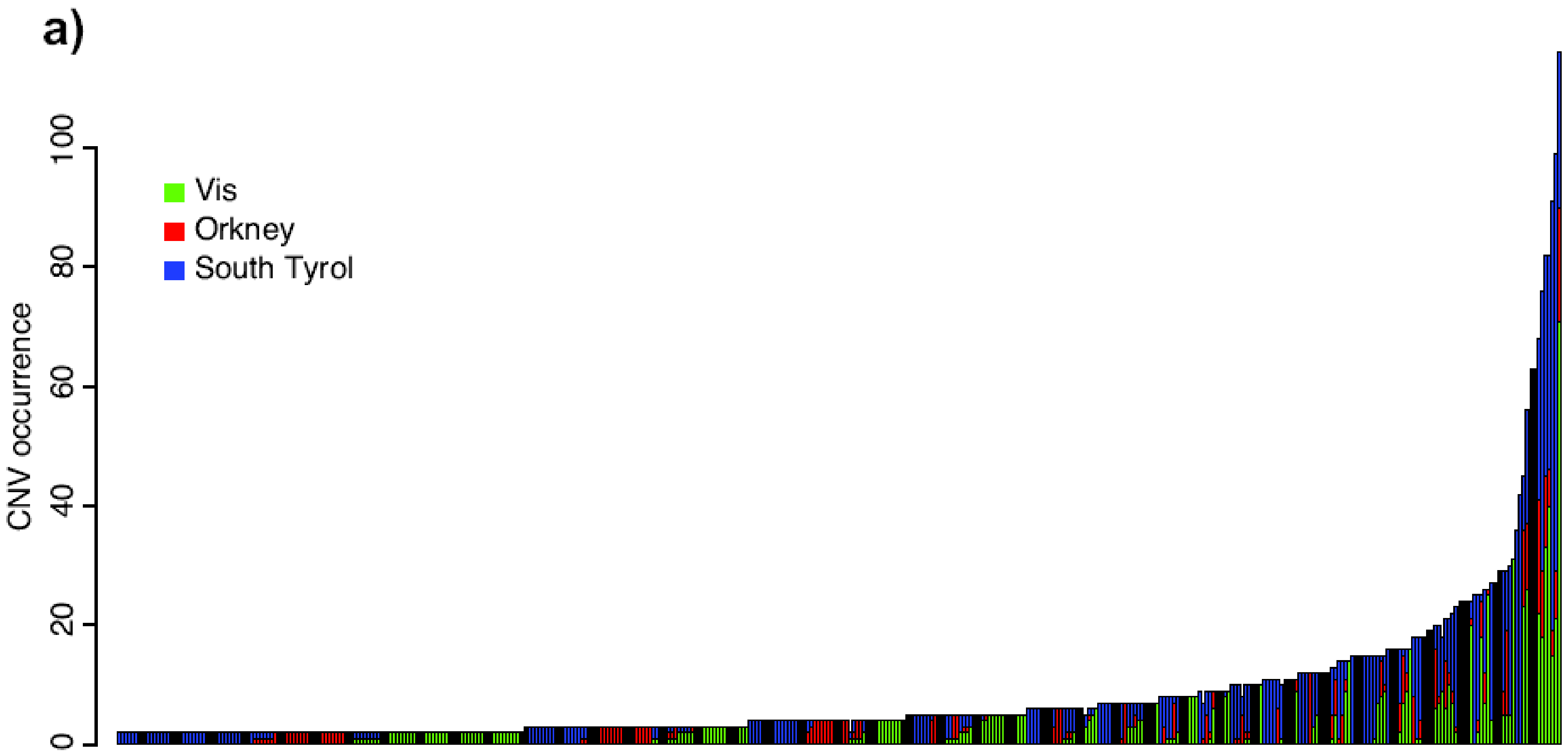


Figure 6.3 CNVR sharing in Dalmatian, Orcadian and South Tyrolean populations. (a) The population make up for each shared CNVR (shared by at least two individuals): each vertical bar represents for a CNVR, the height of each bar is the number of CNV carriers for each CNVR; colour blocks depict the proportions of CNV carriers from each of the three populations, green=Vis, red=Orkney, blue=South Tyrol. (b) Summary of population presentations for CNVRs of different frequencies: each bar represents a group of CNVRs of a certain frequency (from occurring twice to more than 10 times), different colours indicate the proportion of CNVRs private to only one population (in dark grey), CNVRs present in 2 populations (in grey) and CNVRs present in all 3 populations (in light grey).

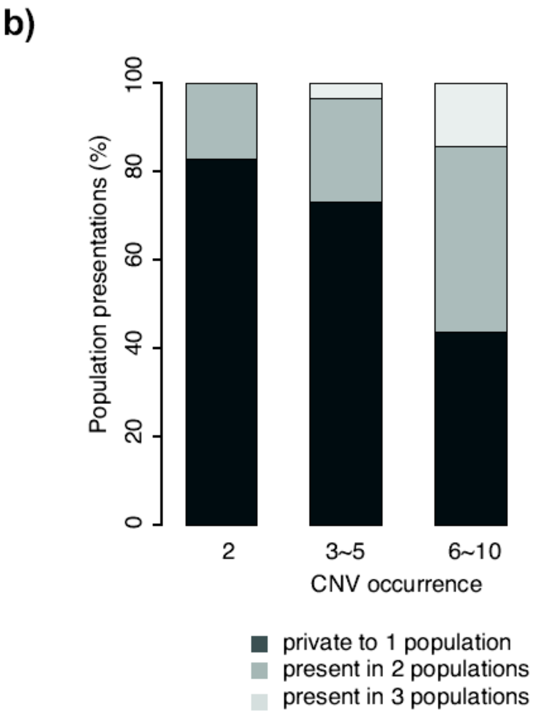


Figure 6.3 CNVR sharing in Dalmatian, Orcadian and South Tyrolean populations (Continued). CNV occurrence is the number of individuals carrying CNVs at a certain loci.

6.3.3 Haplotype and SNP tagging for CNVs

To determine if the CNVs in our study sample were tagged by SNPs and to explore haplotype structure around CNVs, a correlation analysis was carried out on the common CNVRs in Vis and Orkney samples (population frequency>1%): 2 of the 7 CNVRs in Vis, 1 of the 17 in Orkney and 15 of the 47 in South Tyrol were population specific, respectively. No tagging SNPs were found for any of these CNVRs with r^2>0.8. 36 of these CNVRs overlapped CNVRs discovered in a large scale survey of tagging SNP for CNVs in UK samples (WTCCC, 2010). Tagging SNPs were found in only 8 of these 36 regions. Haplotype block detection was performed for the 7 Vis and 17 Orkney CNVRs with SNPs 3Mb upstream and downstream of each CNVR boundary. One CNVR (CNVR271, Chr6:67058287-67111682), could be placed in a haplotype block with 5 adjacent SNPs in all three populations. In addition, two CNVRs (CNVR367, Chr8:15987084-16065839 and CNVR386, Chr8:106005821-106293050) formed two haplotype blocks with nearby SNPs in the South Tyroleans.

6.3.4 Genetic Clustering of individuals according to CNV genotypes

406 CNVR loci were observed multiple times in 1893 individuals (664 Dalmatians, 354 Orcadians and 875 South Tyroleans). Each of those loci were coded for these individuals as "CNV locus" or "non-CNV locus", then software programme Structure (Jakobsson et al., 2008; McQuillan et al., 2008; Pritchard et al., 2000) was used to determine how the individual clustered according to their possession of CNV. Graphical representation of membership in clusters for K=2, 3 and 4 is shown in Figure 6.4. The distribution of the probability of the data between successive values of K showed a peak at K=3 (Ln probability of data=-16991.8 for K=2, -16962.6 for K=3 and -17449.1 for K=4), therefore it is inferred that the most likely number of genetic clusters for these individuals was three, with clusters roughly corresponding to the three geographical locations. 284 of 875 South Tyroleans (32.4%) were assigned to Cluster 1, 259 of 663 (39.1%) Dalmatians assigned to Cluster 2 and 136 of 354 (38.4%) Orcadians assigned to Cluster 3, with membership coefficients >=50%.

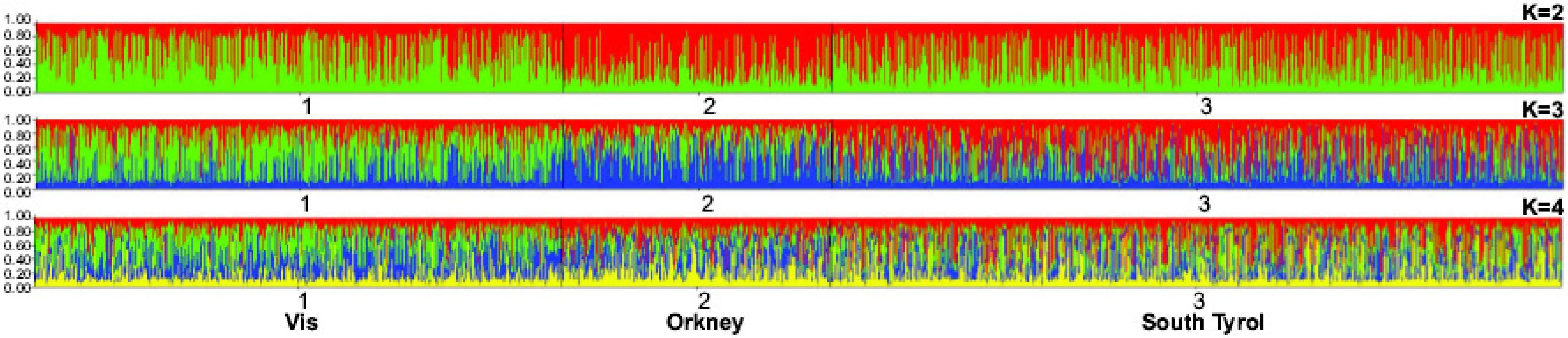


Figure 6.4 Genetic Clustering of individuals according to CNV genotypes. Cluster membership according to analyses of genotypes at 406 CNVR loci in 1893 individuals, for K=2, 3 and 4. Each inferred cluster is represented by a different color. Cluster 1, Cluster 2 and Cluster 3 refers to Vis, Orkney and South Tyrol, respectively.

6.3.5 Gene content

To test whether the detected CNVs were biased in any way towards genetic regions or were evenly distributed across the genome, the gene content of CNVs in the data set were investigated. 2211 CNVs in 441 CNVRs overlapped UCSC known genes. The mean number of genes covered by a CNV was 4.8, which was greater than the average gene content on autosomes (P=0.00574). After introducing SNP density as a covariate into this regression model, the significance still remains (P=0.00042). This result suggested a higher concentration of genes in CNVs. It was also found that the population specific CNVs overlapped more genes (on average 3.1) compared with common CNVs which were shared in more than one population (on average 2.3. $p=3.097*10^{-5}$). No elevated G+C content was detected (on average 40.41% in CNVRs) compared with the autosomal average G+C content (40.35%).

6.3.6 Distribution along chromosomes

To test whether there was any bias in the overall chromosomal distribution of CNVs, CNV density was compared in pre-specified chromosomal regions (i.e. peri-telometric regions, defined as the 10Mb region from the two most distal SNP on both chromosome ends and sub-centromeric regions, defined as the 10Mb region from the two SNPs which were most close to centromere) to that in the rest of the chromosome. A trend was observed towards enrichment in peri-telomeric and/or sub-centromeric regions (define 1Mb from both telomeres on a chromosome as peri-telomeric regions and 1Mb from centromere as sub-centromeric regions, the difference of CNV density in those regions compared to the rest of genome was significant at $p<2.2*10^{-16}$)(Figure 6.5).

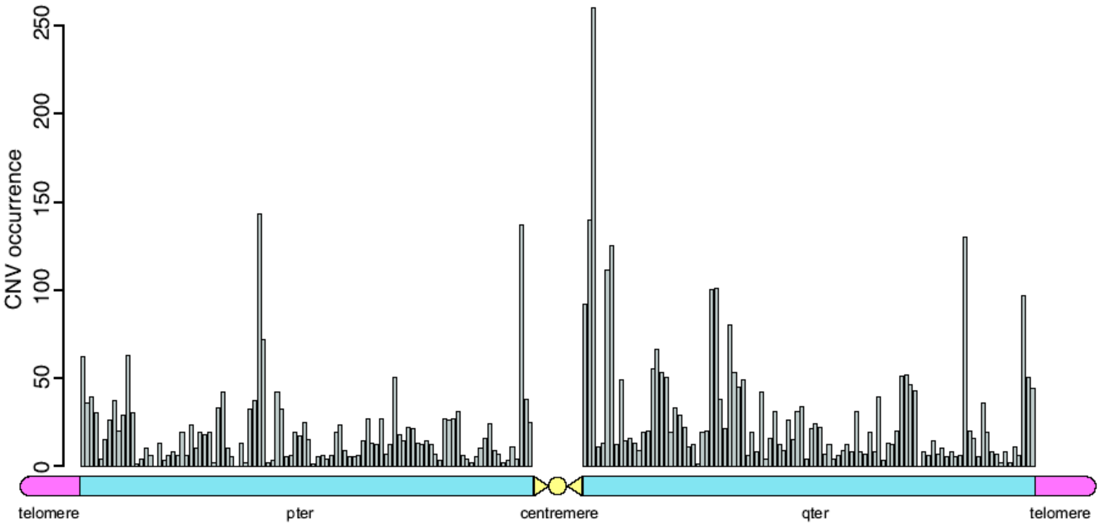


Figure 6.5 The schematic distribution of CNVs on all autosomes, in a physical map. The length of each chromosome arm is adjusted to be 100Mb. Each bar comprises CNVs in a 1Mbp bin on the chromosomes.

6.3.7 Segmental duplications and CNVRs

Of the 743 CNVRs, 222 (98.1Mb, 3.4% of all autosomes) overlap reported segmental duplications (SDs) or putative rearrangement hotspots: 102 CNVRs (41.3Mb) overlap SDs but did not expand into the intervening regions between two SDs on the same chromosome; 153 CNVRs (68.5Mb) were located in between two SDs of known rearrangement hotspots; the remaining 488 CNVRs (89.9Mb) were not in SD regions or known rearrangement hotspot regions; of these 488, 409 (62.2Mb) were population-specific.

Though no difference in G+C content was detected in CNVRs in general, a small increase of G+C content (41.79%) was found in CNVRs outside SDs, compared with that of CNVRs which overlap SDs (39.76%) ($P=1.78*10^{-7}$).

The proportion of CNVRs overlapping SDs was significantly lower for population-specific CNVRs (154 out of 588, 26.2%) than for shared CNVRs (68 of 155, 43.8%) (chi squared test, $P<2.06*10^{-16}$),

6.3.8 Kinship correlation of CNVs

We were interested to test whether carriers of shared CNVs showed more than average relatedness and developed a method to do so by incorporating a kinship coefficient, *k*, into the analysis (see Methods). The kinship coefficient is a parameter not dependent on population frequencies that measures the overall genetic similarity relative to some base population between a pair of individuals. For each CNVR with at least two carriers, the pair-wise kinship coefficients were calculated for all carrier pairs, then the value of those kinship coefficients were compared to the population mean of pair-wise kinship coefficients of all pairs of individuals in the corresponding population. It was observed that for most CNVRs (63.4% in Vis, 76.8% in Orkney and 83.4% in South Tyrol), CNV carriers had higher values of kinship coefficients compared to the population mean, indicating that carriers of shared CNVs are indeed more related to each other. (Table 6.3)

Many CNVs with higher mean k_n could be found to segregate in known families. Two examples were presented to illustrate the segregation of CNVs in pedigrees (Figure 6.6). CNVR686, an amplification on chromosome 19, was detected in 6 individuals who all turned out to have come from the same family (Figure 6.6 a) and b)). The inheritance pattern of this CNVR appeared to be autosomal dominant. CNVR54, an amplification on chromosome 2, was detected in 8 individuals. 4 of them were from the same known family, 2 of them were parent-offspring from another family while the other two were singletons (Figure 6.6 c) and d)).

Table 6.3 Mean kinship coefficients of CNV carriers for CNVRs in three populations.

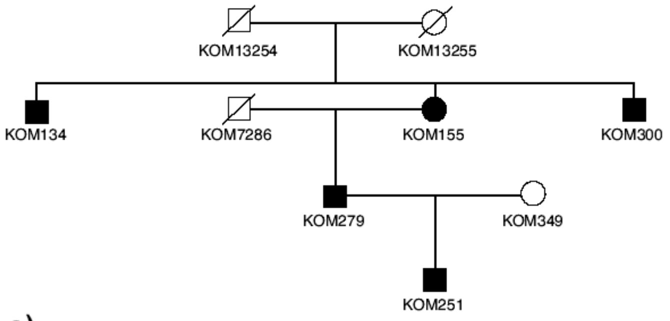
Population	**Vis**	**Orkney**	**South Tyrol**
Mean k_{pop}(±s.d)	0.000402±0.008027	0.001061±0.013336	0.001291±0.0137502
Range of Mean k_n	0 to 0.3125	0 to 0.3125	0 to 0.3125
Total CNVRs (of more than one carrier)	172	112	205
No. CNVRs with p_{nadj}<0.05 (%)	109(63.4%)	86(76.8%)	171(83.4%)

k_{pop}, pair-wise kinship coefficients in one population. k_n, pair-wise kinship coefficients of CNV carriers for the nth CNVR. p_{nadj} is the adjusted p value to describe significance of the differences of kinship coefficients among CNV carriers compared to the population mean pair-wise coefficients.

a)

	KOM251	KOM134	KOM155	KOM279	KOM300	KOM349
KOM251	-					
KOM134	0.0625	-				
KOM155	0.0625	0.25	-			
KOM279	0.25	0.125	0.25	-		
KOM300	0.0625	0.3125	0.25	0.125	-	
KOM349	0.25	0	0	0	0	-

b)



c)

	KOM141	KOM220	KOM283	KOM292	KOM6	KOM331	KOM246	KOM160
KOM141	-							
KOM220	0.125	-						
KOM283	0.25	0.25	-					
KOM292	0.125	0.0625	0.125	-				
KOM6	0.125	0.0625	0.0625	0.25	-			
KOM331	0.125	0.25	0.25	0.0625	0.0625	-		
KOM246	0	0	0	0	0	0	-	
KOM160	0	0	0	0	0	0	0	-

d)

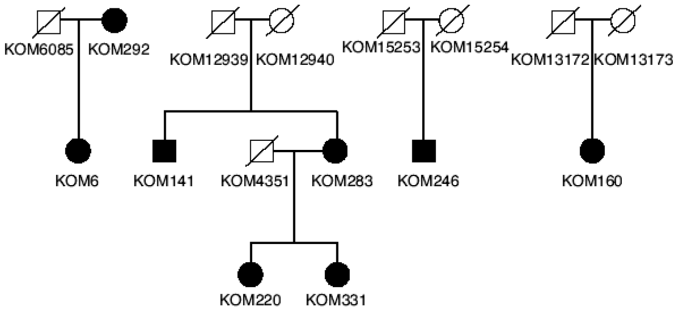


Figure 6.6 Two examples of segregation of CNVs in pedigrees: CNVR686 and CNVR54. (a) The kinship matrix of 6 carriers for CNVR686. They are all from the same population. The mean kinship coefficient of any pair of these 6 carriers is k_{691}=0.175, which is significantly higher than the population mean (adjusted p value<0.001) (b) The carriers for CNV686 placed in pedigree. Squares indicate male sex, circles indicate female sex. Filled squares or circles indicate CNV carriers. A cross through a square or a circle indicates the individual is either deceased or ungenotyped. (c) The kinship matrix of 8 carriers for CNVR54. They are all from the same population. The mean kinship coefficient of any pair of these 8 carriers is k_{55}=0.078, which is significantly higher than the population mean (adjusted p value<0.001) (d) The inheritance of CNV54. The key to the pedigree presentation is the same as for section (b).

6.4 Discussion

Copy Number Variation was profiled in three population isolates from UK, Italy and Croatia and representing a North-South, West-East geographical cline and components of the genetic diversity across Europe. This comparison of CNV characteristics was made possible by virtue of common choice of genotyping platform and copy number detection methods.

In common with previous reports from various populations and cohorts, the great majority of individuals (70%) were found to be carrying at least one CNV. CNVs were also widespread in the genome: 6.6% in length of all autosomal regions showed evidence of CNV in one or more samples. The proportion of SNPs covered by CNVs was 7.8%. The density of SNPs in CNVRs was 175.3 SNPs per Mb, while that in non-CNVRs was 117.1 SNPs per Mb ($p<2.2*10^{-16}$). The lower density of SNPs in regions outside of detected CNVRs indicates that CNVs which reside in the SNP-sparse regions might not be captured on the commercial SNP genotyping platforms which lack coverage in certain chromosomal regions. The SNPs distribute more sparsely in longer CNV regions compared to those in shorter regions, therefore the boundaries determined for longer CNVs were less certain, which reflects the limitation of the HumanHap 300K arrays in terms of SNP coverage. A number of detected CNVRs were represented by both gains and losses. These ‘gain-and-loss’ CNVRs could reflect cases where the reference genome contains both CNV alleles, but individual genomes are homozygous for one or other allele. If true, then gains and losses within the same CNVRs should have equivalent boundaries. However, in all observed cases the gain-and-loss CNVRs in fact contained at least one pair of CNVs from two individuals whose boundaries are not equivalent. Although precise boundary determinations were subject to some technical uncertainty, it does appear that these gain-and-loss CNVRs most likely reflect recurrent CNV changes at the same locus, which are initiated and/or resolved at slightly different points.

Similar to other genetic polymorphisms such as microsatellites and SNPs, it is show here that CNVs differ greatly among different populations. Indeed, the majority of CNVRs (588 out of 743 CNVRs) were restricted to one population and were often of very low frequency, their non-sharing across populations could be due to sampling variances or the fact that they were recent and/or possibly deleterious events. On the other hand, only the most frequently occurring CNVs, which were likely of more ancient origin, were shared between the three population isolates, consistent with a more ancient and neutral evolutionary histories, and also their geographic separation. The longer length and higher gene content of the population-specific CNVRs compared to those of the common CNVRs also supported the hypothesis that they may be more deleterious and therefore kept to low frequencies, or, those are more recent mutations that have had insufficient time to experience disruptive recombination events.

Wether SNPs can serve as a good proxy for CNVs has long been debated (Redon et al., 2006; McCarroll and Altshuler, 2007). Some studies suggested that deletion polymorphisms are generally in strong linkage disequilibrium and segregate on ancestral SNP haplotypes (WTCCC, 2010; McCarroll et al., 2005; Hinds et al., 2006) while some others argue that although a number of CNVs are in strong linkage disequilibrium with nearby markers, accurate genotypes can only be captured for a small proportion of the tested CNVs (Redon et al., 2006). I attempted to investigate LD between SNPs and CNVs, but due to the general low frequencies of the CNVRs in our populations, only a small number were available for testing. No tagging SNPs were found for 7 CNVRs in Vis, 17 CNVRs in Orkney and 47 CNVRs in South Tyrol.

These CNVRs were also found to be poorly tagged by SNPs in the WTCCC samples (supplementary information, WTCCC, 2010). Haplotype analysis revealed only three tagged CNVR, of which one CNVR (CNVR271, Chr6:67058287-67111682) was notable for being shared by all three populations. It was argued in a survey of LD between CNV and SNP that most (77%) highly frequent (MAF>5%) CNVs could be well tagged by SNPs, whereas only 23% of the rare CNVs could be similarly tagged (Conrad et al., 2010). The CNVs selected in the current study for LD analysis were generally of low frequency. Analysis of an expanded set of CNVRs is warranted before firm conclusions on this issue can be drawn.

The CNV profiles in Vis and South Tyrol were more similar to each other compared to that of Orkney, in terms of number of shared CNVRs, correlation of CNV lengths and frequency. This may reflect their relative close geographical distances: Orkney is at 59 degrees north, whereas Vis and South Tyrol are both in Southern Europe.

Genetic clustering analysis formally demonstrated that CNVs can be used to classify the three population groups studied here and one can predict that the same will be true for other human populations, providing a potentially useful and applicable genomic tool for ancestry and evolutionary studies.

Consistent with other recent studies (Nguyen et al., 2006; McCarroll et al., 2008), it was found that CNVs tended to cluster in peri-telomeric /sub-centromeric regions, and commonly overlapped with segmental duplications and recombination hotspots, again consistent with the idea that they may serve well as ancestry markers.

As in many other studies (Kim et al., 2008; Perry et al., 2008; Nguyen et al., 2008), a higher gene content was discovered in CNVRs. It is argued that there is a high G+C content in gene rich regions (Nguyen et al., 2008), which are more frequently subject to copy number change. However, no elevated G+C content was detected in the observed CNVRs in this study. Although high gene content could be due to the bias of SNP choice in commercial genotyping arrays, after correcting for SNP density, the significance still remained. Some have argued that most of these genes are under negligible selective constraint; the CNVs influencing disease genes might have been eliminated by purifying selection. It is also noted a significantly higher gene content within recent, population specific CNVRs. Further studies are warranted to test whether these are due to length of population specific CNVs being longer or they are under positive selection or can be linked (or elevated / diminished) to quantitative traits specifically in population isolates.

Finally, it is shown by the application of kinship coefficients that the majority of rare CNVs are passing through germ-lines rather than being *de novo* variants, and therefore are heritable and provide an index of relatedness. The inheritance of CNVs could be observed in actual pedigrees, which confirmed the increased relatedness between CNV carriers. The similar relationship between genetic variants and kinship was observed in a study of the same population in Vis, which found kinship inferred from pedigree information was consistent with segregation of SNPs in the population (Vitart et al., 2010).

Illumina HumanHap300 SNP genotyping platforms were used to determine copy number variant events in our analysis. Despite the relatively lower SNP content of the 300K microarray compared with products such as Illumina Human 1M and Affymetrix snp 6.0, the power of our method to detect CNVs from the 300K platform was adequate, and it was able to detect a large number of CNV events in the three isolated populations and draw conclusion of the differences between individuals from distinct communities in the context of CNV. However, it is argued that due to insufficient coverage of informative probes in certain chromosome regions (eg. gene sparse and segmental duplication regions) and the inability to discriminate higher number of copies (copy number>4) of a duplicated region for most CNV calling algorithms for SNP arrays, it is hard to accurately quantify the true extent of human copy number variation (Cooper et al., 2008). In light of whole genome sequencing project such as the 1000 Genome Project (http://www.1000genomes.org/), which provides a resource of whole genome sequences of multiple individuals (Sudmant et al., 2010), it is believed that we can benefit from high quality CNV detection directly from sequence data of samples, to better understand the diversity of CNVs within and between populations. In the meantime, mining the widely available SNP arrays coupled with family data of CNV calling represents a useful way of validating CNV calling and studying evolutionary history of CNVs.

Chapter 7

Copy Number Variation and Quantitative Traits

7.1 Preface

The rationale behind genome wide association studies (GWAS) is the common disease, common variant hypothesis, which assumes that the heritability in common diseases can be captured by relatively few common genetic variants in the form of single nucleotide polymorphisms (Wang et al., 2005). However, GWAS based upon SNPs have discovered that SNPs only account for a modest proportion of the total genetic variation, while a substantial proportion of the heritability of many diseases examined in GWAS remain unexplained. It is argued that other genetic variants, for example CNVs, may be a potential source of this so-called missing heritability (Manolio et al., 2009). On the other hand, rare variants of moderate to large effect sizes can also contribute to disease outcome (Wright et al., 2003), and studies which sequence a large fraction of the genome in people with extreme phenotypes (those at the extremes of trait distributions) may be particularly informative in identifying rare as well as common variants associated with common disease (Wang et al., 2005).

The incorporation of the study of CNVs, as well as SNPs, in genetic association studies in becoming increasingly common. There are a growing number of reports of the impact of common and rare CNVs in various diseases, including in AIDS (Gonzalez et al., 2005), autism (Pinto et al., 2010; Wang L.et al., 2010), schizophrenia (Glessner et al., 2010a; The International Schizophrenia Consortium, 2008), bipolar disorder (Zhang et al., 2009; Chen et al., 2010), and obesity (Wang K. et al., 2010; Glessner et al., 2010b). However, the association between CNVs and quantitative traits has rarely been studied. Only a few have been reported to date including those with body mass index (Wineinger et al., 2011; Sha et al., 2009) and aortic root diameter (Wineinger et al., 2011).

Metabolic syndrome comprises a combination of several risk factors for cardiovascular disease and is related to disorders such as type 2 diabetes (T2D), obesity, dyslipidemia, and hypertension (Lanktree and Hegele, 2008). Measurement of a combination of metabolic-related traits such as body mass index (BMI), fasting serum concentrations of lipids, indicators of glucose homeostasis (glucose and insulin) and blood pressure is used to identify individuals with metabolic syndrome (http://www.metabolicsyndromeinstitute.com). The study of genetic components for these traits can shed light on the etiology of metabolic disorders.

In this chapter, the association of CNVs and seven metabolic-related quantitative traits (body mass index, waist circumference, hip circumference, subscapular skinfold thickness, suprailiac skinfold thickness, glucose and insulin) were investigated in 978 individuals from two European populations. Association analysis was performed between common CNVs and measures of these metabolic traits. The role of rare CNVs was also investigated. The results suggested that CNVs, (both common and rare) might contribute to variation in common disease risk and the level of disease-related quantitative traits.

7.2 Methods

7.2.1 Study sample, genotyping and phenotyping

Study participants were enrolled in the CROAS study and ORCADES study, from Island of Vis, Croatia and Orkney Isles, Scotland, respectively. Informed consent was given by all participants and Ethical approval by the relevant Research Ethics Committees.

The Dalmatian samples were genotyped on the Illumina Infinium HumanHap 300 v1 platform while the Orcadian samples were genotyped on the Human Hap 300 v2 platform (Illumina, San Diego, CA, USA). Individuals with a call rate less than 90% were removed. Quality checks of recording of gender and IBD sharing between first- and second-degree relative pairs were performed with the PLINK program (http://pngu.mgh.harvard.edu/purcell/plink/), and individuals with discordant pedigree and genomic data or data values falling outside expected ranges were removed from the study. 1656 individuals (965 from Vis and 691 from Orkney Isles) passed quality control and were included in the CNV investigation.

DNA copy number gain/loss was determined by a joint analysis using QuantiSNP and cnvPartition (see details in Chapter 6) which utilize signal intensity data from SNP probes and implement Hidden Markov Model and Circular Binary Segmentation algorithms respectively to identify abnormal copy numbers,.. A list of CNVs was generated in each population, defined by those CNVs which were identified by both approaches. After prediction of CNV intervals in each individual, overlapping CNVs were merged into CNV regions (CNVRs). A CNVR is a region spanning the boundaries of all CNVs at this locus; i.e., it represents a union of overlapping CNVs (Figure 7.1).

Among the individuals for whom the CNVs had been determined, 1005 unrelated individuals were selected for association analysis. Those comprised 914 singletons (individuals who had no genotyped relatives), with the remainder being probands (the eldest genotyped member) and their genotyped spouse (if applicable) from families of size >=2.

978 of the 1005 unrelated individuals genotyped took part in the biometrical examinations. Measurements were recorded on age, gender, height, weight, body mass index, waist circumference, hip circumference, subscapular skinfold thickness, suprailiac skinfold thickness,

and fasting glucose level and insulin level. Some individuals had one or more missing values for one or more of the above measurements. The descriptive statistics of the final set of participants in the association analysis is shown in Table 7.1.

7.2.2 Construction of a list of candidate genes for metabolic phenotypes

The CNVs which showed evidence for association in the analysis of metabolic phenotypes were compared with a list of candidate genes related to the seven metabolic traits. Only the core regions of each CNV were considered. The core region of a CNV was configured as in Figure 7.1.

The candidate genes for metabolic phenotypes came from four sources: 1) a literature review on candidate genes identified from association analysis on metabolic-related quantitative traits (BMI, glucose, insulin, low density lipoprotein cholesterol, high density lipoprotein, triglycerides); 2) a review of candidate genes for type 2 diabetes disease risk based on data from association studies (including meta-analyses), animal models of the disease, pharmacological and physiological studies, or studies of the Mendelian forms of the disease (Hancock et al., 2008); 3) a review of candidate genes for obesity risk from the same author as in 2) (Hancock et al., 2008); 4) a study of CNVs categorised in DGV which overlapped candidate genes for metabolic syndromes (Lanktree and Hegele, 2008). The gene names from each source were extracted and assembled as a list of all candidate genes for metabolic phenotypes which were related to the seven quantitative traits considered in this chapter.

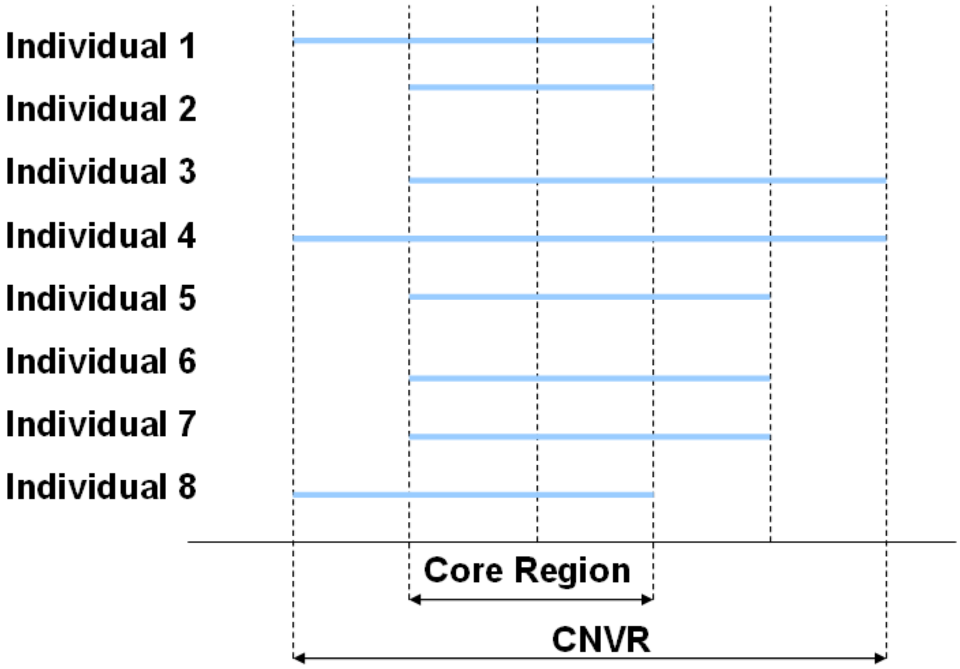


Figure 7.1 Copy number variant regions (CNVR) defined in the study sample and configuration of the “core region” of a CNVR. The core region of a CNVR is the maximum region shared by all the individuals carrying CNV within the same CNVR.

.

7.2.3 Statistical analysis

Statistical analysis was performed using R version 2.8.1.

For common CNVs (frequency >1% in total sample defined by the above procedure) linear regressions were performed to identify associations between CNVs and the seven quantitative traits: body mass index, waist circumference, hip circumference, subscapular skinfold thickness, suprailiac skinfold thickness, glucose and insulin. A linear regression analysis was performed for each trait to evaluate the effects of possible covariates (age, sex, BMI, cohort), with only significant ($P<0.05$) covariates corrected for (Table 7.1). The residues of trait values (adjusted for significant covariates) were then rank normalized with the R package GenABEL (Aulchenko et al., 2007) and were taken as trait values in the following analysis. A linear regression of individual copy number at each CNV locus upon trait values was performed for each of the traits, to test if the copy numbers were associated with metabolic phenotypes. Nominal significance was taken as P<0.05, and a Bonferroni correction was further performed to account for multiple testing (Blauw et al., 2008). The genes covered by CNVs which showed evidence for association were then compared with a list of candidate genes for metabolic syndromes.

For rare CNVs (population frequency<1%), it is hypothesized that multiple rare CNVs may collectively contribute to phenotypic variation of the seven metabolic traits. First of all, the general burden of rare CNVs in individuals with moderate and extreme trait values was tested. Two methods were used: 1) a regression of number of rare CNVs carried on trait values to determine if the overall number of rare CNVs had an effect on the trait values; 2) a regression of rare CNV status (carrying no rare CNV or carrying one or more CNVs) on trait values, to find out if being a carrier of rare CNVs has any effect on the trait values. Secondly, a pathway analysis was conducted to find out if there was an enrichment of genes involved in metabolic pathways, in individuals with extreme trait values. For each trait, the samples were divided into two groups: a "moderate group" with trait values ranked in the 25%-75% range of distribution of all the values for this trait and an "extreme group" with trait values distributed in the upper 25% and lower 25% of the spectrum of all values. The number of rare CNVs in each group was counted, and also the number of genes covered by those CNVs in the two groups. The rare CNVs which only belonged to the "extreme group" were selected and analyzed using both Gene Ontology (GO) and KEGG pathways, via a Web-based Gene Set Analysis Toolkit (WebGestalt http://bioinfo.vanderbilt.edu/ webgestalt/).

7.3 Results

7.3.1 Basic Characteristics of the study sample

The basic characteristics of the study sample, including age, sex, height, weight, BMI, waist and hip circumference, subscapular and suprailiac skinfold thickness, fasting glucose and insulin concentration are summarized in Table 7.1. The significant covariates (which were adjusted for in the following analysis) for each trait are also listed.

7.3.2 Construction of candidate genes for metabolic phenotypes

A literature review of association studies on quantitative traits (body mass index, low density lipoprotein cholesterol, high density lipoprotein cholesterol, and glucose, insulin, and triglycerides concentrations) for metabolic syndromes identified 46 candidate genes in 12 studies (Table 7.2). 177 candidate genes were identified in a literature review of genes which contributed to type 2 diabetes disease risk and 374 candidate genes for obesity disease risk (Hancock et al., 2008). 19 candidate genes for metabolic syndromes overlapped with reported CNVs in DGV (Lanktree and Hegele, 2008).

The genes identified from the above sources were combined to generate a list of candidate genes for metabolic phenotypes. Overall, the list contained 613 genes.

Table 7.1 Descriptive statistics of participants and significant covariates for each trait

	n	mean±s.d.	significant covariants
Age (years)	978	58.3±13.9	
Females (%)		58.2	
Height (cm)	973	167.0±9.6	
Weight (kg)		78.1±14.9	
BMI (kg/m^2)		28.0µU/ml4.6	sex
Waist circumference (cm)	972	959.8±128.1	age, sex, BMI, cohort
Hip circumference (cm)	971	1035±100.6	age, sex, BMI
Subscapular skinfold thickness (cm)	971	249.9±111.5	age, sex, BMI
Suprailiac skinfold thicknesss (cm)	972	276.7±140.6	age, BMI, cohort
Glucose (mmol/L)	931	5.6±1.3	age, sex, BMI, cohort
Insulin (µU/ml)	926	8.3±17.1	age, sex, BMI, cohort

Table 7.2 Genes associated with body mass index, low density lipoprotein cholesterol, high density lipoprotein cholesterol, and glucose, insulin, and triglycerides concentrations in 12 studies

Gene Symbol	Trait	Method	Reference
ABCA1	HDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
ABCB11	GLU	GWAS	Chen et al.(2008)
ACAA2	HDL	meta-analysis	Kathiresan et al.(2008)
ADCY5	GLU	meta-analysis	Dupuis et al.(2008)
ADRA2A	GLU	meta-analysis	Dupuis et al.(2008)
ANGPTL3	TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
APOA1-C3-A4-A5	HDL,TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
APOB	LDL,TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
APOE-C1-C4-C2	LDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
ATG4C	TC	meta-analysis	Kathiresan et al.(2008)
BCL7B	TC	meta-analysis	Kathiresan et al.(2008)
BUD13	HDL,TC	meta-analysis	Kathiresan et al.(2008)
C2CD4B	GLU	meta-analysis	Dupuis et al.(2008)
CELSR2	LDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
CETP	HDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
CILP2	LDL,TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
DGKB	GLU	meta-analysis	Dupuis et al.(2008)
DOCK7	TC	meta-analysis	Kathiresan et al.(2008)
FADS1	GLU	meta-analysis	Dupuis et al.(2008)
FTO	BMI	candidate gene	Frayling et al.(2007),Loos et al.(2008)
G6PC2	GLU	GWAS,meta-analysis	Bouatia-Naji et al.(2008,2009), Dupuis et al.(2008),Prokopenko et al.(2009)
G6PC3	GLU	GWAS	Chen et al.(2008)
GALNT2	HDL,TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
GCK	GLU	GWAS,meta-analysis	Dupuis et al.(2008),Prokopenko et al.(2009),Weedon et al.(2006)
GCKR	GLU,INS, TC	candidate gene,GWAS, meta-analysis	Dupuis et al.(2008),Kathiresan et al.(2008), Orho-Melander et al.(2008), Sparso et al.(2008),Willer et al.(2008)
GLIS3A	GLU	meta-analysis	Dupuis et al.(2008)
HMGCR	LDL	meta-analysis	Kathiresan et al.(2008)
LDLR	LDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
LIPC	HDL,TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
LIPG	HDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
LPL	HDL,TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
MADD	GLU	meta-analysis	Dupuis et al.(2008)
MC4R	BMI	candidate gene	Loos et al.(2008)
MLXIPL	TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
MTNR1B	GLU	candidate gene, meta-analysis	Bouatia-Naji et al.(2008),Dupuis et al.(2008), Lyssenko et al.(2009), Prokopenko et al.(2009)
MVK/MMAB	HDL	GWAS	Willer et al.(2008)
NCAN	LDL,TC	GWAS	Willer et al.(2008)
PBX4	LDL,TC	meta-analysis	Kathiresan et al.(2008)
PCSK9	LDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
PROX1	GLU	meta-analysis	Dupuis et al.(2008)
PSRC1	LDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
SLC2A2	GLU	meta-analysis	Dupuis et al.(2008)
SORT1	LDL	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
TBL2	TC	meta-analysis	Kathiresan et al.(2008)
TRIB1	TC	GWAS,meta-analysis	Kathiresan et al.(2008),Willer et al.(2008)
ZNF259	TC, HDL	meta-analysis	Kathiresan et al.(2008)

BMI: body mass index, LDL: low density lipoprotein cholesterol, HDL: high density lipoprotein cholesterol, GLU: glucose concentration, INS: insulin concentration, TC: triglycerides concentration

7.3.3 Common CNV association

A total of 1164 individual CNVs were identified. All these CNVs could be merged into 407 non-redundant CNVRs. 19 of these CNVRs (Table 7.3) had frequencies of more than 1% and were selected for association analysis with the seven quantitative traits. The CNVs in the selected 19 CNVRs covered 9.2 Mb with a mean length of 190.9 kb. Seven of the 19 CNVRs contained both copy number gains and copy number losses.

Three common CNVs were associated with three metabolic traits with nominal levels of statistical significance (uncorrected p<0.05) (Table 7.4). CNVR729 was associated with BMI (p=0.0235), CNVR122 with waist circumference (p=0.0366) and CNVR447 with both waist circumference (p=0.0217) and insulin concentration (p=0.0226). None of these associations remained statistically significant after Bonferroni correction.

The core region of CNVR122 overlapped two genes, GPR128 (G protein-coupled receptor 128) and TFG (TRK-fused gene). The core region of CNVR447 overlapped no known genes. The core region of CNVR729 overlapped with four genes, DGCR2, DGCR5, DGCR6 (DiGeorge syndrome critical region gene 2, 5 and 6) and PRODH (proline dehydrogenase (oxidase) 1). None of these six genes were known genes related to metabolic phenotypes.

Table 7.3 Characteristics of 19 common CNVs (frequency>1%) for association analysis with metabolic traits in the current study

ID	Chr	Region	Start (bp)	End (bp)	Length (bp)	# Carriers			Frequency
						Total	Gain	Loss	(%)
CNVR5	1	5	9223195	9310031	86837	16	16	0	1.636
CNVR122	3	122	101804733	101955538	150806	12	12	0	1.227
CNVR271	6	271	67058287	67111682	53396	28	0	28	2.863
CNVR297	6	297	168078929	168352184	273256	23	23	0	2.352
CNVR320	7	320	61075979	62372905	1296927	13	13	0	1.329
CNVR322	7	322	64144339	64593616	449278	12	3	9	1.227
CNVR390	8	390	137747933	137932941	185009	11	0	11	1.125
CNVR410	9	410	11430880	12227417	796538	18	0	18	1.840
CNVR447	10	447	47013328	47173619	160292	62	60	2	6.339
CNVR464	10	464	134913018	135284293	371276	26	23	3	2.658
CNVR491	11	491	133749532	134225383	475852	11	11	0	1.125
CNVR494	12	494	7876208	8121428	245221	28	19	9	2.863
CNVR501	12	501	31101381	31311573	210193	45	45	0	4.601
CNVR569	15	569	29704566	30721385	1016820	19	19	0	1.943
CNVR648	18	648	1917798	1970668	52871	14	0	14	1.431
CNVR727	22	727	15412698	15674251	261554	11	11	0	1.125
CNVR729	22	729	17257787	19792353	2534567	16	10	6	1.636
CNVR730	22	730	20659747	20897762	238016	11	9	2	1.125
CNVR734	22	734	23970628	24324013	353386	20	15	5	2.045

Table 7.4 CNVRs showed nominal significance (uncorrected p value of <0.05) in association with metabolic traits.

Trait	CNVR	p value
BMI	CNVR729	0.0226
Waist circumference	CNVR122	0.0366
	CNVR447	0.0217
Insulin	CNVR 447	0.0226

7.3.4 Rare CNVRs and metabolic traits

388 CNVRs had a population frequency of <0.01. The mean length of CNVs in these rare CNVRs was 197.9 kb, which in total covered 106.4 Mb of the genomic DNA. No correlation was found between the number of rare CNVs and trait values in any of the seven traits. Neither was there any observed correlation between the number of overall CNVs (both common and rare) and trait values. When comparing rare CNVs in “moderate” and “extreme” groups for each trait, no difference was found in the number of rare CNVs. However, there was a statistically significant enrichment of unique genes (genes only covered by rare CNVs from moderate group and genes only covered by rare CNVs from the extreme group) overlapped by CNVs from the extreme group (Table 7.5).

For each of the seven traits, unique rare CNV-overlapped genes in moderate and extreme groups were each tested for enrichment in metabolic pathways. In GO and KEGG pathway enrichment analysis, no statistically significant enrichment (p>0.05) was found in either moderate groups or extreme groups, for any one of the seven traits.

Table 7.5 Number of rare CNVs and unique genes in individuals of moderate and extreme values for each of the seven traits

Trait	Number of rare CNVs			Number of unique genes		
	Moderate	Extreme	*p*	Moderate	Extreme	*p*
BMI	239	229	0.6439	77	153	$1.651*10^{-5}$
Waist circumstance	240	238	0.9271	70	147	$1.495*10^{-5}$
Hip circumstance	238	237	0.9634	71	144	$4.194*10^{-5}$
Subscapular skinfold thickness	229	245	0.4624	71	152	$5.938*10^{-5}$
Suprailiac skinfold thickness	230	233	0.8891	66	152	$2.843*10^{-6}$
Glucose	233	236	0.8898	66	148	$6.182*10^{-6}$
Insulin	233	246	0.2882	56	142	$1.696*10^{-6}$

Moderate: CNVs or unique genes in individuals whose trait values were in the middle 50% of the trait value distribution. Extreme: CNVs or unique genes in the remainder of individuals from the total sample.

7.4 Discussion

Copy number variants, which are known to account for a significant proportion of human genetic polymorphism, have been predicted to play a role in the genetic susceptibility to common disease and disease-related quantitative traits. In this chapter, both common and rare CNVs were investigated in metabolic phenotypes.

Type 2 diabetes (T2D) and obesity are two metabolic disorders characterized by a high glucose level in the context of insulin resistance and relative insulin deficiency and high value of body mass index (BMI), respectively. Other quantitative traits, such body fat mass, waist and hip circumference, subscapular and suprailiac skinfold thickness and blood triglycerides levels, are also important indicators of risk for metabolic syndromes. Despite successes in identifying genetic contributions to metabolic phenotypes, only a small part of the heritable component of these traits has so for been explained, mainly by common SNP variants. It has been reported that many of the identified CNVs overlapped genes with important functions in metabolic pathways (Lanktree and Hegele, 2008), therefore one may hypothesise that copy number change in these genes could lead to functional alteration at the expression level, and thus affect susceptibility to metabolic syndrome and metabolic quantitative trait values. In the last few years, some studies have attempted to extend the investigation of effects of SNPs to CNVs on these metabolic phenotypes.

Shtir et al. studied genome-wide association between CNVs and T2D in 194 Caucasian patients from the Framingham Heart Study, but found little evidence of such an association (Shtir et al., 2009). In a genome-wide CNV association study of body mass index (BMI) in the Chinese population, 3 CNVs were found to show a suggestive association with BMI; one of the genes covered by these CNVs was PPRR1 (pancreatic polypeptide receptor 1) which was a known gene related to obesity (Sha et al., 2009). The Wellcome Trust Case Control Consortium (WTCCC) carried out a genome-wide CNV association analysis using large samples and array comparative genomic hybridization and found a weak ($p=3.9*10^{-5}$) association between a CNV and T2D. This CNV overlapped TSPAN8 which was reported to be associated with T2D in previous SNP studies (WTCCC, 2010). In a recent study of obesity with a focus on 39 CNVs in the Prader-Willi syndrome (PWS) critical region in 1000 unrelated Caucasians, 3 CNVs were found to be associated with increased body mass at nominal statistical significance. The known genes for PWS and obesity which were close to these 3 CNVs included NDN (necdin homolog), C15orf2 (Chromosome 15 open reading frame 2) and PWRN1 (Prader-Willi region nonprotein-coding RNA1). None of these genes showed evidence of association in the genome-wide SNP study with the same sample (Chen et al., 2011). None of the above associations remained statistically significant after adjustment for multiple testing.

In the present association analysis of CNVs for seven metabolic traits (body mass index, low density lipoprotein cholesterol, high density lipoprotein cholesterol, and glucose, insulin, and triglycerides concentrations), the effect of 19 genome-wide common CNVs was evaluated.

Three common CNVs were associated with three metabolic traits with nominal statistical significance: CNVR729 with BMI (p=0.0235), CNVR122 with waist circumference (p=0.0366) and CNVR447 with both waist circumferences (p=0.0217) and insulin concentration (p=0.0226). The three CNVs overlapped six genes: GPR128 (G protein-coupled receptor 128), TFG (TRK-fused gene), DGCR2, DGCR5, DGCR6 (DiGeorge syndrome critical region gene 2, 5 and 6) and PRODH (proline dehydrogenase (oxidase) 1). GPR128 is involved in the G-protein coupled receptor protein signaling pathway. TFG encodes several fusion oncoproteins and participates in several oncogenic rearrangements. DGCR2, DGCR5, DGCR6 all reside in chromosome 22q11.2 region. Deletions of the 22q11.2 have been associated with a wide range of developmental defects (notably DiGeorge syndrome, velocardiofacial syndrome, conotruncal anomaly face syndrome and isolated conotruncal cardiac defects) classified under the acronym CATCH 22. The DGCR2 gene encodes a novel putative adhesion receptor protein which could play a role in neural crest cells migration, a process which has been proposed to be altered in DiGeorge syndrome. DGCR6 is a candidate for involvement in DiGeorge syndrome pathology and in schizophrenia. However, the six genes' molecular function and the relationship with metabolic traits and the onset of metabolic syndromes are still unknown. The above associations were also weak; none of them retained statistical significance after correction for multiple testing. Further analysis, both replication of the association analysis findings utilizing a larger sample size and functional studies, are needed to identify their potential role on metabolic phenotypes.

The association analysis is based on the common disease-common variant hypothesis, therefore only common CNVs were considered. However rare CNVs, on the other hand, could also play an important part in explaining genetic variation in disease risk or levels of quantitative traits. Several studies found significant enrichment of rare CNVs in cases of mental disorders compared to those in control individuals, and many of the rare CNVs overlapped genes which had functional relevance to those diseases (Pinto et al., 2010; The International Schizophrenia Consortium, 2008; Walsh et al., 2008; Williams et al., 2010). In order to identify rare variants related to metabolic traits, Cohen et al sequenced candidate regions from individuals with extreme values of high density lipoprotein cholesterol (HDL-C), and found that nonsynonymous sequence variants were significantly more common in individuals with low HDL-C values than in those with high HDL-C values (Cohen et al., 2004). These findings suggest that rare variants may collectively contribute to variation in metabolic phenotypes. However, little attention has been paid to rare CNVs in genome wide CNV studies for metabolic syndromes. One of the few examples was an observation of enrichment of multiple large and rare CNVs in obesity cases, which disrupt several obesity candidate genes (Wang K. et al., 2010).

In this chapter, the overall burden and gene content of rare CNVs for seven metabolic traits were examined. No excess burden of rare CNVs (either the sum rare CNVs possessed or the presence/absence status of any rare CNVs for an individual) was observed in individuals who had more extreme trait values for any of the seven traits. However, a significant enrichment of unique genes overlapped by rare CNVs in individuals with extreme trait values was found for all the traits. This result could suggest a functional difference of the rare CNV covered genes in individuals with moderate and extreme metabolic trait values. An enrichment analysis was then performed for the unique genes in the moderate and extreme groups, to test if there was an enrichment of genes involved in metabolic pathways in individuals with more extreme trait values. However, no statistically significant enrichment was found by neither GO nor KEGG pathway analysis.

The results in this chapter should be interpreted cautiously due to several study limitations recognised. First, the HumanHap 300K genotyping array might not be ideal to profile all CNVs in an individual genome. The SNP density of 300K genotyping array is comparatively low, and this platform is not primarily designed to capture copy number variants. SNPs in some regions such as segmental duplications are sparse. Therefore a proportion of CNVs might have gone undetected on the 300K platform. Second, the frequencies of common CNVs selected in the current study is generally low (only one of the 19 common CNVs had a frequency of >5%), which might lead to reduced power in association analysis. Additionally, at the time of this study the methods of analyzing rare CNVs for disease and disease-related QTs were still limited. The pathway enrichment analysis is only a primary approach to illustrate the general picture of all rare variants and to categorize those in grouped pathways for different biological functions. However, even if enrichment was found for genes in a particular pathway, it is still hard to determine which ones of all the rare variants were causative and which ones were not relevant. Finally, as SNP association has only identified a fraction of the loci contributing to phenotypic variation, it follows that CNVs may have an impact on the unrecognized risk loci. The present study is underpowered to conclude whether CNVs contribute to metabolic trait variance, but is will be of interest to investigate whether the impact of CNVs indicated in the current study is substantial, by combining data from multiple large studies.

In conclusion, there was some suggestive evidence of association between several common CNVRs and metabolic phenotypes, although none overlapped known candidate genes for metabolic phenotypes. No association with overall burden of rare CNVRs was observed, but significant enrichment of unique genes was found in individuals with extreme values of metabolic quantitative traits. Those genes, however, failed to show enrichment in metabolic pathways. These results suggest that CNVs may be potentially important for metabolic phenotype variation. Further research is required to confirm or reject these initial findings. This should include studies employing improved (more sensitive) methods of identifying CNVs, (much) larger sample sizes and study populations in several global regions. In addition these should be complemented by molecular biological experiments to investigate functions of the CNVs, and improved and more robust pathway analysis methods to study rare variants.

Chapter 8

Summary of thesis and future directions

Copy number variation is a type of genetic variation which has been extensively studied in recent years. With advances in methods/platform development for CNV detection which have enabled CNVs to be identified in multiple individuals in an efficient manner, numerous studies have endeavored to reveal features of CNVs in human population and their relevance to disease and disease-related phenotypes. However, our understanding of this kind of genetic variation is still limited.

One very basic but still unanswered question is the scale of individual CNVs. In Chapter 3, using CNV data gained from five sequenced human genomes, I surveyed 12 studies which sequenced 33 human genomes, and obtained information of copy number loss (deletions) from five individuals, from three different ethnic groups. Generally, the deletions detected from sequencing data were short in length, compared to those detected from SNP or CGH arrays. The overlap of deletions between the five individuals was low. The two Asian genomes had more deletions in common but the concordance of deletions was low for the two European genomes. The CNVs shared by multiple individuals covered fewer genes than those private to only one individual. These differences could have true biological relevance, but could also be due to differences in platforms/algorithms choice and the small number of individual complete genomes available for analysis.

Mining of genome wide SNP data can be used to extract CNV calls in large number of samples, which enables CNV studies at the population level. However, the choice of a reliable protocol to call CNVs from SNP data is an important issue to consider. In Chapter 4, to find out comparable CNV data at the population level, I selected seven studies which reported CNVs in HapMap samples, from a structured literature search including 778 articles which detected CNVs from SNP genotyping data. Large discrepancies were observed for CNVs identified in terms of total occurrence and length. For the two HapMap samples common to six studies, concordance in CNV calling was low and results showed dependence on the genotyping platform and/or calling algorithm employed. Moreover, for two individuals in whom a direct physical mapping method was used, it was clear that only a small portion of CNV calls were detected from SNP genotyping data. This Chapter demonstrated that platform/algorithm choice could greatly influence the results of CNV calling and limited the use of SNP genotyping platforms to detect CNVs.

In the following chapter I used Illumina 300K SNP intensity data from 965 individuals together with 8 HapMap samples to assess the performance of four CNV detection algorithms: QuantiSNP, cnvPartition, PennCNV and DNAcopy. Based upon concordance rates in duplicates, QuantiSNP and cnvPartition outperformed the other two algorithms on both sensitivity and specificity. However, it was also noted from the comparison of CNVs called from QuantiSNP and those validated from a previous study (Kidd et al., 2008) for the same 8 HapMap samples, that the algorithm only could recover a small portion of CNVs validated by direct physical methods, such as ESP mapping and array-CGH.

In Chapter 6 I used a combination of QuantiSNP and cnvPartition to profile CNVs in 2789 individuals from three European population isolates (Vis, Orkney and South Tyrol) who had been genotyped on the Illumina HumanHap 300K platforms. 4016 CNVs in 1964 individuals were detected, which clustered into 743 copy number variable regions (CNVRs). The frequency and distribution of these CVNRs was compared and shown to differ significantly between the Orcadian, South Tyrolean and Dalmatian populations. Consistent with the inference that this indicated population-specific CNVR identity and origin, I also demonstrated that CNVR variation within each population can be used to measure genetic relatedness.

In the last section, I looked for evidence of association between CNVs and seven metabolic-related quantitative traits: body mass index, waist circumstance, hip circumference, subscapular skinfold thickness, suprailiac skinfold thickness, glucose and insulin, in 978 individuals from Vis and Orkney. Three out of 19 common CNVs tested showed nominal significance for association with one or more traits, but the significance didn't remain after multiple-testing correction. None of them overlapped with known candidate genes for metabolic phenotypes. No excess burden of rare CNVs was observed in individuals with extreme trait values for any of the traits analyzed, but I did find that more genes were affected by rare CNVs in the individuals with extreme trait values. However, pathway analysis showed no significant enrichment of those genes in metabolic pathways.

In summary, this thesis investigated current CNV detection methods, conducted a discovery study of CNVs in three European populations and attempted to test association of those CNVs to metabolic-related quantitative traits. This thesis made some contribution to the understanding of copy number variation, but future work is needed to further clarify the features and impact of CNVs on phenotypic outcomes in human populations.

It was noted that the genotyping data gained from Illumina 300K arrays was not ideal for CNV detection, and there were various problems with current CNV calling algorithms. Therefore, improvements in genomic technologies are needed for more accurate CNV detection in the future. These include higher-resolution and higher-throughput platforms (SNP/CGH arrays), advances in next-generation sequencing technologies and more robust algorithms for CNV detection on those platforms. Molecular biology experiments are needed to validate CNVs detected in the current samples from SNP genotyping data. CNVs at the population level merit further study. Studies such as the 1000 Genome Project (http://www.1000genomes.org/), which aim to accurately detect and genotype CNVs in multiple individuals, will shed more light on individual by individual variation, and on the origins and evolution of CNVs. Preliminary knowledge of genome-wide association between common CNVs and common diseases was gained, benefitting from large study design (WTCCC, 2010), however for complex traits and diseases, the 'hidden' heritability void left by GWAS would not always be accounted for by common CNVs (Conrad et al., 2010). Therefore one should consider and assess the impact of both common and rare CNVs on phenotypic outcomes. Last but not least, genetic analysis on CNVs alone is not sufficient to unravel all contributing factors for complex human traits. The combination of genetic approaches combining SNP and CNV variants in association studies, better algorithms to assess association between CNVs and disease or disease related phenotypes, functional studies of putative genes influenced by CNVs, refined bioinformatics tools for pathway analysis, systems biology and animal models need to be integrated and combined in order to provide a complete picture of the origins, structure and functional consequences of copy number variation.

References

Abecasis,G.R., Cookson W.O.C., and Cardon,L.R. (2001). The Power to Detect Linkage Disequilibrium with Quantitative Traits in Selected Samples. Am J Hum Genet. 68, 1463-1472.

Adams,D.J., Dermitzakis,E.T., Cox,T., Smith,J., Davies,R., Banerjee,R., Bonfield,J., Mullikin,J.C., Chung,Y.J., Rogers,J., and Bradley,A. (2005). Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. Nature Genetics 37, 532-536.

Ahn,S.M., Kim,T.H., Lee,S., Kim,D., Ghang,H., Kim,D., Kim,B.C., Kim,S.Y., Kim,W.Y., Kim,C., Park,D., Lee,Y.S., Kim,S., Reja,R., Jho,S., Kim,C.G., Cha,J.Y., Kim,K.H., Lee,B., Bhak,J., and Kim,S.J. (2009). The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. Genome Res, 19(9): 1622–1629.

Aulchenko,Y.S., Ripke,S., Isaacs,A., and van Duijn,C.M. (2007). GenABEL: an R library for genome-wide association analysis. Bioinformatics 23, 1294-1296.

Autism Genome Project Consortium, Szatmari,P., Paterson,A.D., Zwaigenbaum,L., Roberts,W., Brian,J., Liu,X.Q., Vincent,J.B., Skaug,J.L., Thompson,A.P., Senman,L., Feuk,L., Qian,C., Bryson,S.E., Jones,M.B., Marshall,C.R., Scherer,S.W., Vieland,V.J., Bartlett,C., Mangin,L.V., Goedken,R., Segre,A., Pericak-Vance,M.A., Cuccaro,M.L., Gilbert,J.R., Wright,H.H., Abramson,R.K., Betancur,C., Bourgeron,T., Gillberg,C., Leboyer,M., Buxbaum,J.D., Davis,K.L., Hollander,E., Silverman,J.M., Hallmayer,J., Lotspeich,L., Sutcliffe,J.S., Haines,J.L., Folstein,S.E., Piven,J., Wassink,T.H., Sheffield,V., Geschwind,D.H., Bucan,M., Brown,W.T., Cantor,R.M., Constantino,J.N., Gilliam,T.C., Herbert,M., LaJonchere,C., Ledbetter,D.H., Lese-Martin,C., Miller,J., Nelson,S., Samango-Sprouse,C.A., Spence,S., State,M., Tanzi,R.E., Coon,H., Dawson,G., Devlin,B., Estes,A., Flodman,P., Klei,L., McMahon,W.M., Minshew,N., Munson,J., Korvatska,E., Rodier,P.M., Schellenberg,G.D., Smith,M., Spence,M.A., Stodgell,C., Tepper,P.G., Wijsman,E.M., Yu,C.E., Roge,B., Mantoulan,C., Wittemeyer,K., Poustka,A., Felder,B., Klauck,S.M., Schuster,C., Poustka,F., Bolte,S., Feineis-Matthews,S., Herbrecht,E., Schmotzer,G., Tsiantis,J., Papanikolaou,K., Maestrini,E., Bacchelli,E., Blasi,F., Carone,S., Toma,C., Van,E.H., de,J.M., Kemner,C., Koop,F., Langemeijer,M., Hijmans,C., Staal,W.G., Baird,G., Bolton,P.F., Rutter,M.L., Weisblatt,E., Green,J., Aldred,C., Wilkinson,J.A., Pickles,A., Le,C.A., Berney,T., McConachie,H., Bailey,A.J., Francis,K., Honeyman,G., Hutchinson,A., Parr,J.R., Wallace,S., Monaco,A.P., Barnby,G., Kobayashi,K., Lamb,J.A., Sousa,I., Sykes,N., Cook,E.H., Guter,S.J., Leventhal,B.L., Salt,J., Lord,C., Corsello,C., Hus,V., Weeks,D.E., Volkmar,F., Tauber,M., Fombonne,E., Shih,A., and Meyer,K.J. (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nature Genetics. 39(3):319-28.

Bademci,G., Edwards,T.L., Torres,A.L., Scott,W.K., Zuchner,S., Martin,E.R., Vance,J.M., and Wang,L. (2010). A rare novel deletion of the tyrosine hydroxylase gene in Parkinson disease. Hum. Mutat. 31, E1767-E1771.

Bae,J.S., Cheong,H.S., Kim,J.O., Lee,S.O., Kim,E.M., Lee,H.W., Kim,S., Kim,J.W., Cui,T., Inoue,I., and Shin,H.D. (2008). Identification of SNP markers for common CNV regions and association analysis of risk of subarachnoid aneurysmal hemorrhage in Japanese population. Biochemical & Biophysical Research Communications. 373(4):593-6.

Bailey,J.A., Gu,Z., Clark,R.A., Reinert,K., Samonte,R.V., Schwartz,S., Adams,M.D., Myers,E.W., Li,P.W., and Eichler,E.E. (2002). Recent Segmental Duplications in the Human Genome. Science 297, 1003-1007.

Bailey,J.A., Kidd,J.M., and Eichler,E.E. (2008). Human copy number polymorphic genes. Cytogenetic and Genome Research 123, 234-243.

Bassett,A.S., Marshall,C.R., Lionel,A.C., Chow,E.W.C., and Scherer,S.W. (2008). Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. Hum. Mol. Genet. 17, 4045-4053.

Bauman,J.G.J., Wiegant,J., Borst,P., and van Duijn,P. (1980). A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. Experimental Cell Research 128, 485-490.

Bea,S., Salaverria,I., Armengol,L., Pinyol,M., Fernandez,V., Hartmann,E.M., Jares,P., Amador,V., Hernandez,L., Navarro,A., Ott,G., Rosenwald,A., Estivill,X., and Campo,E. (2009). Uniparental disomies, homozygous deletions, amplifications, and target genes in mantle cell lymphoma revealed by integrative high-resolution whole-genome profiling. Blood 113, 3059-3069.

Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R., Boutell,J.M., Bryant,J., Carter,R.J., Keira Cheetham,R., Cox,A.J., Ellis,D.J., Flatbush,M.R., Gormley,N.A., Humphray,S.J., Irving,L.J., Karbelashvili,M.S., Kirk,S.M., Li,H., Liu,X., Maisinger,K.S., Murray,L.J., Obradovic,B., Ost,T., Parkinson,M.L., Pratt,M.R., Rasolonjatovo,I.M.J., Reed,M.T., Rigatti,R., Rodighiero,C., Ross,M.T., Sabot,A., Sankar,S.V., Scally,A., Schroth,G.P., Smith,M.E., Smith,V.P., Spiridou,A., Torrance,P.E., Tzonev,S.S., Vermaas,E.H., Walter,K., Wu,X., Zhang,L., Alam,M.D., Anastasi,C., Aniebo,I.C., Bailey,D.M.D., Bancarz,I.R., Banerjee,S., Barbour,S.G., Baybayan,P.A., Benoit,V.A., Benson,K.F., Bevis,C., Black,P.J., Boodhun,A., Brennan,J.S., Bridgham,J.A., Brown,R.C., Brown,A.A., Buermann,D.H., Bundu,A.A., Burrows,J.C., Carter,N.P., Castillo,N., Chiara E.Catenazzi,M., Chang,S., Neil Cooley,R., Crake,N.R., Dada,O.O., Diakoumakos,K.D., Dominguez-Fernandez,B., Earnshaw,D.J., Egbujor,U.C., Elmore,D.W., Etchin,S.S., Ewan,M.R., Fedurco,M., Fraser,L.J., Fuentes Fajardo,K.V., Scott Furey,W., George,D., Gietzen,K.J., Goddard,C.P., Golda,G.S., Granieri,P.A., Green,D.E., Gustafson,D.L., Hansen,N.F., Harnish,K., Haudenschild,C.D., Heyer,N.I., Hims,M.M., Ho,J.T., Horgan,A.M., Hoschler,K., Hurwitz,S., Ivanov,D.V., Johnson,M.Q., James,T., Huw Jones,T.A., Kang,G.D., Kerelska,T.H., Kersey,A.D., Khrebtukova,I., Kindwall,A.P., Kingsbury,Z., Kokko-Gonzales,P.I., Kumar,A., Laurent,M.A., Lawley,C.T., Lee,S.E., Lee,X., Liao,A.K., Loch,J.A., Lok,M., Luo,S., Mammen,R.M., Martin,J.W., McCauley,P.G., McNitt,P., Mehta,P., Moon,K.W., Mullens,J.W., Newington,T., Ning,Z., Ling Ng,B., Novo,S.M., 'Neill,M.J., Osborne,M.A., Osnowski,A., Ostadan,O., Paraschos,L.L., Pickering,L., Pike,A.C., Pike,A.C., Chris Pinkard,D., Pliskin,D.P., Podhasky,J., Quijano,V.J., Raczy,C., Rae,V.H., Rawlings,S.R., Chiva Rodriguez,A., Roe,P.M., Rogers,J., Rogert Bacigalupo,M.C., Romanov,N., Romieu,A., Roth,R.K., Rourke,N.J., Ruediger,S.T., Rusman,E., Sanches-Kuiper,R.M., Schenker,M.R., Seoane,J.M., Shaw,R.J., Shiver,M.K., Short,S.W., Sizto,N.L., Sluis,J.P., Smith,M.A., Ernest Sohna Sohna,J., Spence,E.J., Stevens,K., Sutton,N., Szajkowski,L., Tregidgo,C.L., Turcatti,G., vandeVondele,S., Verhovsky,Y., Virk,S.M., Wakelin,S., Walcott,G.C., Wang,J., Worsley,G.J., Yan,J., Yau,L., Zuerlein,M., Rogers,J., Mullikin,J.C., Hurles,M.E., McCooke,N.J., West,J.S., Oaks,F.L., Lundberg,P.L., Klenerman,D., Durbin,R., and Smith,A.J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53-59.

Blangero,J. (2004). Localization and identification of human quantitative trait loci: King Harvest has surely come. Current Opinion in Genetics & Development 14, 233-240.

Blauw,H.M., Veldink,J.H., van Es,M.A., van Vught,P.W., Saris,C.G.J., van der Zwaag,B., Franke,L., Burbach,J.P., Wokke,J.H., Ophoff,R.A., and van den Berg,L.H. (2008). Copy-number variation in sporadic amyotrophic lateral sclerosis: a genome-wide screen. Lancet Neurol 7, 319-326.

Carter,N.P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. Nature Genetics 39, S16-S21.

Chen,W.K., Swartz,J.D., Rush,L.J., and Alvarez,C.E. (2009). Mapping DNA structural variation in dogs. Genome Res. 19, 500-509.

Chen,X., Li,X., Wang,P., Liu,Y., Zhang,Z., Zhao,G., Xu,H., Zhu,J., Qin,X., Chen,S., Hu,L., and Kong,X. (2010). Novel association strategy with copy number variation for identifying new risk Loci of human diseases. Plos One 5, e12185.

Chen,Y., Liu,Y.J., Pei,Y.F., Yang,T.L., Deng,F.Y., Liu,X.G., Li,D.Y., and Deng,H.W. (2011). Copy Number Variations at the Prader-Willi Syndrome Region on Chromosome 15 and associations with Obesity in Whites. Obesity.

Cifola,I., Spinelli,R., Beltrame,L., Peano,C., Fasoli,E., Ferrero,S., Bosari,S., Signorini,S., Rocco,F., Perego,R., Proserpio,V., Raimondo,F., Mocarelli,P., and Battaglia,C. (2008). Genome-wide screening of copy number alterations and LOH events in renal cell carcinomas and integration with gene expression profile. Molecular Cancer 7.

Cohen,J.C., Kiss,R.S., Pertsemlidis,A., Marcel,Y.L., McPherson,R., and Hobbs,H.H. (2004). Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol. Science 305, 869-872.

Cohen,J. (2007). GENOMICS: DNA Duplications and Deletions Help Determine Health. Science 317, 1315-1317.

Colella,S., Yau,C., Taylor,J.M., Mirza,G., Butler,H., Clouston,P., Bassett,A.S., Seller,A., Holmes,C.C., and Ragoussis,J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucl. Acids Res. 35, 2013-2025.

Conrad,D.F., Andrews,T.D., Carter,N.P., Hurles,M.E., and Pritchard,J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. Nature Genetics. 38(1):75-81.

Conrad,D.F., Pinto,D., Redon,R., Feuk,L., Gokcumen,O., Zhang,Y., Aerts,J., Andrews,T.D., Barnes,C., Campbell,P., Fitzgerald,T., Hu,M., Ihm,C.H., Kristiansson,K., MacArthur,D.G., MacDonald,J.R., Onyiah,I., Pang,A.W.C., Robson,S., Stirrups,K., Valsesia,A., Walter,K., Wei,J., Tyler-Smith,C., Carter,N.P., Lee,C., Scherer,S.W., and Hurles,M.E. (2010). Origins and functional impact of copy number variation in the human genome. Nature 464, 704-712.

Cooper,G.M., Nickerson,D.A., and Eichler,E.E. (2007). Mutational and selective effects on copy-number variants in the human genome. Nat Genet.

Cooper,G.M., Zerr,T., Kidd,J.M., Eichler,E.E., and Nickerson,D.A. (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nature Genetics. 40(10):1199-203.

Cowell,J.K. and Lo,K.C. (2009). Application of oligonucleotides arrays for coincident comparative genomic hybridization, ploidy status and loss of heterozygosity studies in human cancers. Methods Mol Biol 556, 47-65.

Cronin,S., Blauw,H.M., Veldink,J.H., van Es,M.A., Ophoff,R.A., Bradley,D.G., van den Berg,L.H., and Hardiman,O. (2008). Analysis of genome-wide copy number variation in Irish and Dutch ALS populations. Hum. Mol. Genet. 17, 3392-3398.

Cutler,G. and Kassner,P.D. (2008). Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. Cytogenetic and Genome Research 123, 297-306.

Day,N., Hemmaplardh,A., Thurman,R.E., Stamatoyannopoulos,J.A., and Noble,W.S. (2007). Unsupervised segmentation of continuous genomic data. Bioinformatics 23, 1424-1426.

Dellinger,A.E., Saw,S.M., Goh,L.K., Seielstad,M., Young,T.L., and Li,Y.J. (2010). Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucl. Acids Res. 38, e105.

Dopman,E.B. and Hartl,D.L. (2007). A portrait of copy-number polymorphism in Drosophila melanogaster. PNAS 104, 19920-19925.

Drmanac,R., Sparks,A.B., Callow,M.J., Halpern,A.L., Burns,N.L., Kermani,B.G., Carnevali,P., Nazarenko,I., Nilsen,G.B., Yeung,G., Dahl,F., Fernandez,A., Staker,B., Pant,K.P., Baccash,J., Borcherding,A.P., Brownley,A., Cedeno,R., Chen,L., Chernikoff,D., Cheung,A., Chirita,R., Curson,B., Ebert,J.C., Hacker,C.R., Hartlage,R., Hauser,B., Huang,S., Jiang,Y., Karpinchyk,V., Koenig,M., Kong,C., Landers,T., Le,C., Liu,J., McBride,C.E., Morenzoni,M., Morey,R.E., Mutch,K., Perazich,H., Perry,K., Peters,B.A., Peterson,J., Pethiyagoda,C.L., Pothuraju,K., Richter,C., Rosenbaum,A.M., Roy,S., Shafto,J., Sharanhovich,U., Shannon,K.W., Sheppy,C.G., Sun,M., Thakuria,J.V., Tran,A., Vu,D., Zaranek,A.W., Wu,X., Drmanac,S., Oliphant,A.R., Banyai,W.C., Martin,B., Ballinger,D.G., Church,G.M., and Reid,C.A. (2010). Human Genome

Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. Science 327, 78-81.

Dunbar,A.J., Gondek,L.P., O'Keefe,C.L., Makishima,H., Rataul,M.S., Szpurka,H., Sekeres,M.A., Wang,X.F., McDevitt,M.A., and Maciejewski,J.P. (2008). 250K Single Nucleotide Polymorphism Array Karyotyping Identifies Acquired Uniparental Disomy and Homozygous Mutations, Including Novel Missense Substitutions of c-Cbl, in Myeloid Malignancies. Cancer Res 68, 10349-10357.

Fadista,J., Nygaard,M., Holm,L.E., Thomsen,B., and Bendixen,C. (2008). A Snapshot of CNVs in the Pig Genome. Plos One 3, e3916.

Fiegler,H., Redon,R., Andrews,D., Scott,C., Andrews,R., Carder,C., Clark,R., Dovey,O., Ellis,P., Feuk,L., French,L., Hunt,P., Kalaitzopoulos,D., Larkin,J., Montgomery,L., Perry,G.H., Plumb,B.W., Porter,K., Rigby,R.E., Rigler,D., Valsesia,A., Langford,C., Humphray,S.J., Scherer,S.W., Lee,C., Hurles,M.E., and Carter,N.P. (2006). Accurate and reliable high-throughput detection of copy number variation in the human genome. Genome Res. 16, 1566-1574.

Fix,A., Lucchesi,C., Ribeiro,A., Lequin,D., Pierron,G., Schleiermacher,G., Delattre,O., and Janoueix-Lerosey,I. (2008). Characterization of amplicons in neuroblastoma: High-resolution mapping using DNA microarrays, relationship with outcome, and identification of overexpressed genes. Genes Chromosomes & Cancer 47, 819-834.

Franke,L., de Kovel,C.G.E., Aulchenko,Y.S., Trynka,G., Zhernakova,A., Hunt,K.A., Blauw,H.M., van den Berg,L.H., Ophoff,R., Deloukas,P., van Heel,D.A., and Wijmenga,C. (2008). Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. American Journal of Human Genetics 82, 1316-1333.

Freedman,M.L., Reich,D., Penney,K.L., McDonald,G.J., Mignault,A.A., Patterson,N., Gabriel,S.B., Topol,E.J., Smoller,J.W., Pato,C.N., Pato,M.T., Petryshen,T.L., Kolonel,L.N., Lander,E.S., Sklar,P., Henderson,B., Hirschhorn,J.N., and Altshuler,D. (2004). Assessing the impact of population stratification on genetic association studies. Nat Genet 36, 388-393.

Garshasbi,M., Hadavi,V., Habibi,H., Kahrizi,K., Kariminejad,R., Behjati,F., Tzschach,A., Najmabadi,H., Ropers,H.H., and Kuss,A.W. (2008). A defect in the TUSC3 gene is associated with autosomal recessive mental retardation. American Journal of Human Genetics. 82(5):1158-64.

Gijsbers,A.C.J., D'haene,B., Hilhorst-Hofstee,Y., Mannens,M., Albrecht,B., Seidel,J., Witt,D.R., Maisenbacher,M.K., Loeys,B., van Essen,T., Bakker,E., Hennekam,R., Breuning,M.H., De Baere,E., and Ruivenkamp,C.A.L. (2008). Identification of copy number variants associated with BPES-like phenotypes. Human Genetics 124, 489-498.

Glessner,J.T., Wang,K., Cai,G., Korvatska,O., Kim,C.E., Wood,S., Zhang,H., Estes,A., Brune,C.W., Bradfield,J.P., Imielinski,M., Frackelton,E.C., Reichert,J., Crawford,E.L., Munson,J., Sleiman,P.M., Chiavacci,R., Annaiah,K., Thomas,K., Hou,C., Glaberson,W., Flory,J., Otieno,F., Garris,M., Soorya,L., Klei,L., Piven,J., Meyer,K.J., Anagnostou,E., Sakurai,T., Game,R.M., Rudd,D.S., Zurawiecki,D., McDougle,C.J., Davis,L.K., Miller,J., Posey,D.J., Michaels,S., Kolevzon,A., Silverman,J.M., Bernier,R., Levy,S.E., Schultz,R.T., Dawson,G., Owley,T., McMahon,W.M., Wassink,T.H., Sweeney,J.A., Nurnberger,J.I., Coon,H., Sutcliffe,J.S., Minshew,N.J., Grant,S.F., Bucan,M., Cook,E.H., Buxbaum,J.D., Devlin,B., Schellenberg,G.D., and Hakonarson,H. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 459(7246):569-73.

Glessner,J.T., Reilly,M.P., Kim,C.E., Takahashi,N., Albano,A., Hou,C.P., Bradfield,J.P., Zhang,H.T., Sleiman,P.M.A., Flory,J.H., Imielinski,M., Frackelton,E.C., Chiavacci,R., Thomas,K.A., Garris,M., Otieno,F.G., Davidson,M., Weiser,M., Reichenberg,A., Davis,K.L., Friedman,J.I., Cappola,T.P., Margulies,K.B., Rader,D.J., Grant,S.F.A., Buxbaum,J.D., Gur,R.E., and Hakonarson,H. (2010a). Strong synaptic transmission impact by copy number variations in schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 107, 10584-10589.

Glessner,J.T., Bradfield,J.P., Wang,K., Takahashi,N., Zhang,H.T., Sleiman,P.M., Mentch,F.D., Kim,C.E., Hou,C.P., Thomas,K.A., Garris,M.L., Deliard,S., Frackelton,E.C., Otieno,F.G., Zhao,J.H., Chiavacci,R.M., Li,M.Y., Buxbaum,J.D., Berkowitz,R.I., Hakonarson,H., and Grant,S.F.A. (2010b). A Genome-wide Study Reveals Copy Number Variants Exclusive to Childhood Obesity Cases. American Journal of Human Genetics 87, 661-666.

Gondek,L.P., Tiu,R., O'Keefe,C.L., Sekeres,M.A., Theil,K.S., and Maciejewski,J.R. (2008). Chromosomal lesions and uniparental disomy detected by SNP arrays in MDS, MDS/MPD, and MDS-derived AML. Blood 111, 1534-1542.

Gonzalez,E., Kulkarni,H., Bolivar,H., Mangano,A., Sanchez,R., Catano,G., Nibbs,R.J., Freedman,B.I., Quinones,M.P., Bamshad,M.J., Murthy,K.K., Rovin,B.H., Bradley,W., Clark,R.A., Anderson,S.A., O'Connell,R.J., Agan,B.K., Ahuja,S.S., Bologna,R., Sen,L., Dolan,M.J., and Ahuja,S.K. (2005). The Influence of CCL3L1 Gene-Containing Segmental Duplications on HIV-1/AIDS Susceptibility. Science 307, 1434-1440.

Gunderson,K.L. and Peiffer,D.A. (2006). LOH and DNA copy number changes - Using SNP-CGH to profile for amplifications, duplications, and deletions. Genetic Engineering News 26, 32-33.

Hancock,A.M., Witonsky,D.B., Gordon,A.S., Eshel,G., Pritchard,J.K., Coop,G., and Di Rienzo,A. (2008). Adaptations to Climate in Candidate Genes for Common Metabolic Disorders. PLoS Genet 4, e32.

Herr,A., Grutzmann,R., Matthaei,A., Artelt,J., Schrock,E., Rump,A., and Pilarsky,C. (2005). High-resolution analysis of chromosomal imbalances using the Affymetrix 10K SNP genotyping chip. Genomics 85, 392-400.

Hinds,D.A., Kloek,A.P., Jen,M., Chen,X.Y., and Frazer,K.A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. Nature Genetics 38, 82-85.

Hormozdiari,F., Alkan,C., Eichler,E.E., and Sahinalp,S.C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res. 19, 1270-1278.

Hoyer,J., Dreweke,A., Becker,C., Gohring,I., Thiel,C.T., Peippo,M.M., Rauch,R., Hofbeck,M., Trautmann,U., Zweier,C., Zenker,M., Huffmeier,U., Kraus,C., Ekici,A.B., Ruschendorf,F., Nurnberg,P., Reis,A., and Rauch,A. (2007). Molecular karyotyping in patients with mental retardation using 100K single-nucleotide polymorphism arrays. Journal of Medical Genetics. 44(10):629-36.

Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.

Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W., and Lee,C. (2004). Detection of large-scale variation in the human genome. Nat Genet 36, 949-951.

Illumina Manual 1:BeadStudio Genotyping Module User Guide. Illumina® Systems and Software. Illumina Inc. 2007

Illumina Manual 2: DNA Copy Number Analysis Algorithms. Technical Note: Illumina® Systems and Software. Illumina Inc. 2008

Illumina Manual 3: cnvPartition CNV Analysis Plug-in v1.0.2 for BeadStudio. Technical Note: Illumina® Systems and Software. Illumina Inc. 2008

International Schizophrenia Consortium (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455, 237-241.

Jacobs,S., Thompson,E.R., Nannya,Y., Yamamoto,G., Pillai,R., Ogawa,S., Bailey,D.K., and Campbell,I.G. (2007). Genome-wide, high-resolution detection of copy number, loss of heterozygosity, and genotypes from formalin-fixed, paraffin-embedded tumor tissue using microarrays. Cancer Res 67, 2544-2551.

Jakobsson,M., Scholz,S.W., Scheet,P., Gibbs,J.R., VanLiere,J.M., Fung,H.C., Szpiech,Z.A., Degnan,J.H., Wang,K., Guerreiro,R., Bras,J.M., Schymick,J.C., Hernandez,D.G., Traynor,B.J., Simon-Sanchez,J., Matarin,M., Britton,A., van de Leemput,J., Rafferty,I., Bucan,M., Cann,H.M., Hardy,J.A., Rosenberg,N.A., and Singleton,A.B. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451, 998-1003.

James,J., Merz,T., Janower,M.L., and Dorst,J.P. (1971). Radiological features of the most common autosomal disorders: Trisomy 21-22 (Mongolism or Down's syndrome), trisomy 18, trisomy 13-15, and the cri du chat syndrome. Clinical Radiology 22, 417-433.

Kallioniemi,A., Kallioniemi,O.P., Sudar,D., Rutovitz,D., Gray,J.W., Waldman,F., and Pinkel,D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258, 818-821.

Kamatani,Y., Matsuda,K., Ohishi,T., Ohtsubo,S., Yamazaki,K., Iida,A., Hosono,N., Kubo,M., Yumura,W., Nitta,K., Katagiri,T., Kawaguchi,Y., Kamatani,N., and Nakamura,Y. (2008). Identification of a significant association of a single nucleotide polymorphism in TNXB with systemic lupus erythematosus in a Japanese population. Journal of Human Genetics. 53(1):64-73.

Kawamata,N., Ogawa,S., Yamamoto,G., Lehmann,S., Levine,R.L., Pikman,Y., Nannya,Y., Sanada,M., Miller,C.W., Gilliland,D.G., and Koeffler,H.P. (2008). Genetic profiling of myeloproliferative disorders by single-nucleotide polymorphism oligonucleotide microarray. Experimental Hematology 36, 1471-1479.

Kidd,J.M., Cooper,G.M., Donahue,W.F., Hayden,H.S., Sampas,N., Graves,T., Hansen,N., Teague,B., Alkan,C., Antonacci,F., Haugen,E., Zerr,T., Yamada,N.A., Tsang,P., Newman,T.L., Tuzun,E., Cheng,Z., Ebling,H.M., Tusneem,N., David,R., Gillett,W., Phelps,K.A., Weaver,M., Saranga,D., Brand,A., Tao,W., Gustafson,E., McKernan,K., Chen,L., Malig,M., Smith,J.D., Korn,J.M., McCarroll,S.A., Altshuler,D.A., Peiffer,D.A., Dorschner,M., Stamatoyannopoulos,J., Schwartz,D., Nickerson,D.A., Mullikin,J.C., Wilson,R.K., Bruhn,L., Olson,M.V., Kaul,R., Smith,D.R., and Eichler,E.E. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56-64.

Kim,P.M., Lam,H.Y.K., Urban,A.E., Korbel,J.O., Affourtit,J., Grubert,F., Chen,X., Weissman,S., Snyder,M., and Gerstein,M.B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. Genome Res. 18, 1865-1874.

Kim,J.I., Ju,Y.S., Park,H., Kim,S., Lee,S., Yi,J.H., Mudge,J., Miller,N.A., Hong,D., Bell,C.J., Kim,H.S., Chung,I.S., Lee,W.C., Lee,J.S., Seo,S.H., Yun,J.Y., Woo,H.N., Lee,H., Suh,D., Lee,S., Kim,H.J., Yavartanoo,M., Kwak,M., Zheng,Y., Lee,M.K., Park,H., Kim,J.Y., Gokcumen,O., Mills,R.E., Zaranek,A.W., Thakuria,J., Wu,X., Kim,R.W., Huntley,J.J., Luo,S., Schroth,G.P., Wu,T.D., Kim,H., Yang,K.S., Park,W.Y., Kim,H., Church,G.M., Lee,C., Kingsmore,S.F., and Seo,J.S. (2009). A highly annotated whole-genome sequence of a Korean individual. Nature 460, 1011-1U96.

Kirov,G., Grozeva,D., Norton,N., Ivanov,D., Mantripragada,K.K., Holmans,P., Craddock,N., Owen,M.J., O'Donovan,M.C., Int Schizophrenia Consortium, and Wellcome Trust Case Control Consor (2009). Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. Hum. Mol. Genet. 18, 1497-1503.

Koboldt,D.C., Ding,L., Mardis,E.R., and Wilson,R.K. (2010). Challenges of sequencing human genomes. Briefings in Bioinformatics.

Kohler,J.R. and Cutler,D.J. (2007). Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies. American Journal of Human Genetics 81, 684-699.

Komura,D., Shen,F., Ishikawa,S., Fitch,K.R., Chen,W.W., Zhang,J., Liu,G.Y., Ihara,S., Nakamura,H., Hurles,M.E., Lee,C., Scherer,S.W., Jones,K.W., Shapero,M.H., Huang,J., and Aburatani,H. (2006). Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Research 16, 1575-1584.

Korbel,J.O., Kim,P.M., Chen,X., Urban,A.E., Weissman,S., Snyder,M., and Gerstein,M.B. (2008). The current excitement about copy-number variation: how it relates to gene duplications and protein families. Current Opinion in Structural Biology 18, 366-374.

Korn,J.M., Kuruvilla,F.G., McCarroll,S.A., Wysoker,A., Nemesh,J., Cawley,S., Hubbell,E., Veitch,J., Collins,P.J., Darvishi,K., Lee,C., Nizzari,M.M., Gabriel,S.B., Purcell,S., Daly,M.J., and Altshuler,D. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nature Genetics 40, 1253-1260.

Kurotaki,N., Shen,J.J., Touyama,M., Kondoh,T., Visser,R., Ozaki,T., Nishimoto,J., Shiihara,T., Uetake,K., Makita,Y., Harada,N., Raskin,S., Brown,C.W., H 枚 glund,P., Okamoto,N., and Lupski,J.R. (2005). Phenotypic consequences of genetic variation at hemizygous alleles: Sotos syndrome is a contiguous gene syndrome incorporating coagulation factor twelve (FXII) deficiency. Genetics in Medicine 7.

LaFramboise,T., Weir,B.A., Zhao,X., Beroukhim,R., Li,C., Harrington,D., Sellers,W.R., and Meyerson,M. (2005). Allele-specific amplification in cancer revealed by SNP array analysis. PLoS Comput Biol 1, e65.

Langemeijer,S.M.C., Kuiper,R.P., Berends,M., Knops,R., Aslanyan,M.G., Massop,M., Stevens-Linders,E., van Hoogen,P., Van Kessel,A.G., Raymakers,R.A.P., Kamping,E.J., Verhoef,G.E., Verburgh,E., Hagemeijer,A., Vandenberghe,P., de Witte,T., van der Reijden,B.A., and Jansen,J.H. (2009). Acquired mutations in TET2 are common in myelodysplastic syndromes. Nature Genetics 41, 838-U102.

Lanktree,M. and Hegele,R.A. (2008). Copy number variation in metabolic phenotypes. Cytogenetic and Genome Research 123, 169-175.

Lee,J.A., Carvalho,C.M.B., and Lupski,J.R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. Cell 131, 1235-1247.

Lee,J.H. and Jeon,J.T. (2008). Methods to detect and analyze copy number variations at the genome-wide and locus-specific levels. Cytogenetic and Genome Research 123, 333-342.

Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G., Lin,Y., MacDonald,J.R., Pang,A.W.C., Shago,M., Stockwell,T.B., Tsiamouri,A., Bafna,V., Bansal,V., Kravitz,S.A., Busam,D.A., Beeson,K.Y., McIntosh,T.C., Remington,K.A., Abril,J.F., Gill,J., Borman,J., Rogers,Y.H., Frazier,M.E., Scherer,S.W., Strausberg,R.L., and Venter,J.C. (2007). The Diploid Genome Sequence of an Individual Human. PLoS Biol 5, e254.

Li,C. and Wong,W.H. (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. PNAS 98, 31-36.

Li,C., Beroukhim,R., Weir,B.A., Winckler,W., Garraway,L.A., Sellers,W.R., and Meyerson,M. (2008). Major copy proportion analysis of tumor samples using SNP arrays. Bmc Bioinformatics

9, Article.

Li,J., Jiang,T., Mao,J.H., Balmain,A., Peterson,L., Harris,C., Rao,P.H., Havlak,P., Gibbs,R., and Cai,W.W. (2004). Genomic segmental polymorphisms in inbred mouse strains. Nat Genet 36, 952-954.

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,C.J., Drees,B.M., Hamilton,J.J., Recker,R.R., Cheng,J., and Deng,H.W. (2009). Whole Genome Distribution and Ethnic Differentiation of Copy Number Variation in Caucasian and Asian Populations. Plos One 4, e7958.

Lifton,R.P., Dluhy,R.G., Powers,M., Rich,G.M., Cook,S., Ulick,S., and Lalouel,J.M. (1992). A chimaeric ll[beta]-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. Nature 355, 262-265.

Lin,C.H., Huang,M.C., Li,L.H., Wu,J.Y., Chen,Y.T., and Fann,C.S.J. (2008). Genome-wide copy number analysis using copy number inferring tool (CNIT) and DNA pooling. Hum. Mutat. 29, 1055-1062.

Locke,D.P., Segraves,R., Carbone,L., Archidiacono,N., Albertson,D.G., Pinkel,D., and Eichler,E.E. (2003). Large-Scale Variation Among Human and Great Ape Genomes Determined by Array Comparative Genomic Hybridization. Genome Res. 13, 347-357.

Locke,D.P., Sharp,A.J., McCarroll,S.A., McGrath,S.D., Newman,T.L., Cheng,Z., Schwartz,S., Albertson,D.G., Pinkel,D., Altshuler,D.M., and Eichler,E.E. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. American Journal of Human Genetics. 79(2):275-90.

Lucito,R., Healy,J., Alexander,J., Reiner,A., Esposito,D., Chi,M., Rodgers,L., Brady,A., Sebat,J., Troge,J., West,J.A., Rostan,S., Nguyen,K.C.Q., Powers,S., Ye,K.Q., Olshen,A., Venkatraman,E., Norton,L., and Wigler,M. (2003). Representational Oligonucleotide Microarray Analysis: A High-Resolution Method to Detect Genome Copy Number Variation. Genome Res. 13, 2291-2305.

Lupski,J.R., de Oca-Luna,R.M., Slaugenhaupt,S., Pentao,L., Guzzetta,V., Trask,B.J., Saucedo-Cardenas,O., Barker,D.F., Killian,J.M., Garcia,C.A., Chakravarti,A., and Patel,P.I. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66, 219-232.

Lupski,J.R., Reid,J.G., Gonzaga-Jauregui,C., Rio Deiros,D., Chen,D.C.Y., Nazareth,L., Bainbridge,M., Dinh,H., Jing,C., Wheeler,D.A., McGuire,A.L., Zhang,F., Stankiewicz,P., Halperin,J.J., Yang,C., Gehman,C., Guo,D., Irikat,R.K., Tom,W., Fantin,N.J., Muzny,D.M., and Gibbs,R.A. (2010). Whole-Genome Sequencing in a Patient with Charcot-Marie-Tooth Neuropathy. New England Journal of Medicine 362, 1181-1191.

MacDonald, IL and Walter,Z. Hidden Markov and Other Models for Discrete-Valued Time Series. 1st edition .Chapman and Hall/CRC, 2007

Manolio,T.A., Collins,F.S., Cox,N.J., Goldstein,D.B., Hindorff,L.A., Hunter,D.J., McCarthy,M.I., Ramos,E.M., Cardon,L.R., Chakravarti,A., Cho,J.H., Guttmacher,A.E., Kong,A., Kruglyak,L., Mardis,E., Rotimi,C.N., Slatkin,M., Valle,D., Whittemore,A.S., Boehnke,M., Clark,A.G., Eichler,E.E., Gibson,G., Haines,J.L., Mackay,T.F.C., McCarroll,S.A., and Visscher,P.M. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.

Marroni,F., Pichler,I., De Grandi,A., Beu Volpato,C., Vogl,F.D., Pinggera,G.K., Bailey-Wilson,J.E., and Pramstaller,P.P. (2006). Population Isolates in South Tyrol and Their Value for Genetic Dissection of Complex Diseases. Annals of Human Genetics 70, 812-821.

Marshall,C.R., Noor,A., Vincent,J.B., Lionel,A.C., Feuk,L., Skaug,J., Shago,M., Moessner,R., Pinto,D., Ren,Y., Thiruvahindrapduram,B., Fiebig,A., Schreiber,S., Friedman,J., Ketelaars,C.E., Vos,Y.J., Ficicioglu,C., Kirkpatrick,S., Nicolson,R., Sloman,L., Summers,A., Gibbons,C.A., Teebi,A., Chitayat,D., Weksberg,R., Thompson,A., Vardy,C., Crosbie,V., Luscombe,S., Baatjes,R., Zwaigenbaum,L., Roberts,W., Fernandez,B., Szatmari,P., and Scherer,S.W. (2008). Structural variation of chromosomes in autism spectrum disorder. American Journal of Human Genetics. 82(2):477-88.

Mascalzoni,D., Janssens,A.C.J., Stewart,A., Pramstaller,P., Gyllensten,U., Rudan,I., van Duijn,C.M., Wilson,J.F., Campbell,H., and Quillan,R.M. (2009). Comparison of participant information and informed consent forms of five European studies in genetic isolated populations. Eur J Hum Genet 18, 296-302.

McCarroll,S.A. and Altshuler,D.M. (2007). Copy-number variation and association studies of human disease. Nature Genetics 39, S37-S42.

McCarroll,S.A., Hadnott,T.N., Perry,G.H., Sabeti,P.C., Zody,M.C., Barrett,J.C., Dallaire,S., Gabriel,S.B., Lee,C., Daly,M.J., Altshuler,D.M., and The International HapMap Consortium (2005). Common deletion polymorphisms in the human genome. Nat Genet 38, 86-92.

McCarroll,S.A., Kuruvilla,F.G., Korn,J.M., Cawley,S., Nemesh,J., Wysoker,A., Shapero,M.H., de Bakker,P.I.W., Maller,J.B., Kirby,A., Elliott,A.L., Parkin,M., Hubbell,E., Webster,T., Mei,R., Veitch,J., Collins,P.J., Handsaker,R., Lincoln,S., Nizzari,M., Blume,J., Jones,K.W., Rava,R., Daly,M.J., Gabriel,S.B., and Altshuler,D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature Genetics 40, 1166-1174.

McKernan,K.J., Peckham,H.E., Costa,G.L., McLaughlin,S.F., Fu,Y.T., Tsung,E.F., Clouser,C.R., Duncan,C., Ichikawa,J.K., Lee,C.C., Zhang,Z., Ranade,S.S., Dimalanta,E.T., Hyland,F.C., Sokolsky,T.D., Zhang,L., Sheridan,A., Fu,H.N., Hendrickson,C.L., Li,B., Kotler,L., Stuart,J.R., Malek,J.A., Manning,J.M., Antipova,A.A., Perez,D.S., Moore,M.P., Hayashibara,K.C., Lyons,M.R., Beaudoin,R.E., Coleman,B.E., Laptewicz,M.W., Sannicandro,A.E., Rhodes,M.D., Gottimukkala,R.K., Yang,S., Bafna,V., Bashir,A., MacBride,A., Alkan,C., Kidd,J.M., Eichler,E.E., Reese,M.G., De la Vega,F.M., and Blanchard,A.P. (2009). 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.

McQuillan,R., Leutenegger,A.L., bdel-Rahman,R., Franklin,C.S., Pericic,M., Barac-Lauc,L., Smolej-Narancic,N., Janicijevic,B., Polasek,O., Tenesa,A., MacLeod,A.K., Farrington,S.M., Rudan,P., Hayward,C., Vitart,V., Rudan,I., Wild,S.H., Dunlop,M.G., Wright,A.F., Campbell,H., and Wilson,J.F. (2008). Runs of homozygosity in European populations. American Journal of Human Genetics 83, 359-372.

Medvedev,P., Stanciu,M., and Brudno,M. (2009). Computational methods for discovering structural variation with next-generation sequencing. Nat Meth 6, S13-S20.

Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen,K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K., Cheetham,R.K., Chinwalla,A., Conrad,D.F., Fu,Y., Grubert,F., Hajirasouliha,I., Hormozdiari,F., Iakoucheva,L.M., Iqbal,Z., Kang,S., Kidd,J.M., Konkel,M.K., Korn,J., Khurana,E., Kural,D., Lam,H.Y.K., Leng,J., Li,R., Li,Y., Lin,C.Y., Luo,R., Mu,X.J., Nemesh,J., Peckham,H.E., Rausch,T., Scally,A., Shi,X., Stromberg,M.P., Stutz,A.M., Urban,A.E., Walker,J.A., Wu,J., Zhang,Y., Zhang,Z.D., Batzer,M.A., Ding,L., Marth,G.T., McVean,G., Sebat,J., Snyder,M., Wang,J., Ye,K., Eichler,E.E., Gerstein,M.B., Hurles,M.E., Lee,C., McCarroll,S.A., and Korbel,J.O. (2011). Mapping copy number variation by population-scale genome sequencing. Nature 470, 59-65.

Moreno-De-Luca,D., Mulle,J.G., Kaminsky,E.B., Sanders,S.J., Myers,S.M., Adam,M.P., Pakula,A.T., Eisenhauer,N.J., Uhas,K., Weik,L., Guy,L., Care,M.E., Morel,C.F., Boni,C., Salbert,B.A., Chandrareddy,A., Demmer,L.A., Chow,E.W.C., Surti,U., Aradhya,S., Pickering,D.L., Golden,D.M., Sanger,W.G., Aston,E., Brothman,A.R., Gliem,T.J., Thorland,E.C., Ackley,T., Iyer,R., Huang,S., Barber,J.C., Crolla,J.A., Warren,S.T., Martin,C.L., and Ledbetter,D.H. (2010). Deletion 17q12 Is a Recurrent Copy Number Variant that Confers High Risk of Autism and Schizophrenia. Am J Hum Genet 87, 618-630.

Nannya,Y., Sanada,M., Nakazaki,K., Hosoya,N., Wang,L., Hangaishi,A., Kurokawa,M., Chiba,S., Bailey,D.K., Kennedy,G.C., and Ogawa,S. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Research. 65(14):6071-9.

Need,A.C., Ge,D., Weale,M.E., Maia,J., Feng,S., Heinzen,E.L., Shianna,K.V., Yoon,W., Kasperaviciute,D., Gennarelli,M., Strittmatter,W.J., Bonvicini,C., Rossi,G., Jayathilake,K., Cola,P.A., McEvoy,J.P., Keefe,R.S.E., Fisher,E.M.C., St Jean,P.L., Giegling,I., Hartmann,A.M., Moller,H.J., Ruppert,A., Fraser,G., Crombie,C., Middleton,L.T., St Clair,D., Roses,A.D., Muglia,P., Francks,C., Rujescu,D., Meltzer,H.Y., and Goldstein,D.B. (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet 5, e1000373.

Nguyen,D.Q., Webber,C., and Ponting,C.P. (2006). Bias of Selection on Human Copy-Number Variants. PLoS Genet 2, e20.

Nguyen,D.Q., Webber,C., Hehir-Kwa,J., Pfundt,R., Veltman,J., and Ponting,C.P. (2008). Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Res. 18, 1711-1723.

Norris,B.J. and Whan,V.A. (2008). A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. Genome Res. 18, 1282-1293.

Olshen,A.B., Venkatraman,E.S., Lucito,R., and Wigler,M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. Biostat 5, 557-572.

Ozaki,K., Ohnishi,Y., Iida,A., Sekine,A., Yamada,R., Tsunoda,T., Sato,H., Sato,H., Hori,M., Nakamura,Y., and Tanaka,T. (2002). Functional SNPs in the lymphotoxin-[alpha] gene that are associated with susceptibility to myocardial infarction. Nat Genet 32, 650-654.

Pattaro,C., Marroni,F., Riegler,A., Mascalzoni,D., Pichler,I., Volpato,C., Dal Cero,U., De Grandi,A., Egger,C., Eisendle,A., Fuchsberger,C., Gogele,M., Pedrotti,S., Pinggera,G., Stefanov,S., Vogl,F., Wiedermann,C., Meitinger,T., and Pramstaller,P. (2007). The genetic study of three population microisolates in South Tyrol (MICROS): study design and epidemiological perspectives. BMC Medical Genetics 8, 29.

Pelak,K., Shianna,K.V., Ge,D., Maia,J.M., Zhu,M., Smith,J.P., Cirulli,E.T., Fellay,J., Dickson,S.P., Gumbs,C.E., Heinzen,E.L., Need,A.C., Ruzzo,E.K., Singh,A., Campbell,C.R., Hong,L.K., Lornsen,K.A., McKenzie,A.M., Sobreira,N.L.M., Hoover-Fong,J.E., Milner,J.D., Ottman,R., Haynes,B.F., Goedert,J.J., and Goldstein,D.B. (2010). The Characterization of Twenty Sequenced Human Genomes. PLoS Genet 6, e1001111.

Peltonen,L. (2000). Positional cloning of disease genes: advantages of genetic isolates. Hum Hered 50, 66-75.

PennCNV Tutorial. http://www.openbioinformatics.org/penncnv/penncnv_beadstudio_tutorial.html. Accessed 2008-2010

Perry,G.H., Dominy,N.J., Claw,K.G., Lee,A.S., Fiegler,H., Redon,R., Werner,J., Villanea,F.A., Mountain,J.L., Misra,R., Carter,N.P., Lee,C., and Stone,A.C. (2007). Diet and the evolution of human amylase gene copy number variation. Nat Genet 39, 1256-1260.

Perry,G.H., Yang,F., Marques-Bonet,T., Murphy,C., Fitzgerald,T., Lee,A.S., Hyland,C., Stone,A.C., Hurles,M.E., Tyler-Smith,C., Eichler,E.E., Carter,N.P., Lee,C., and Redon,R. (2008). Copy number variation and evolution in humans and chimpanzees. Genome Res. 18, 1698-1710.

Pielberg,G., Olsson,C., Syvanen,A.C., and Andersson,L. (2002). Unexpectedly High Allelic Diversity at the KIT Locus Causing Dominant White Color in the Domestic Pig. Genetics 160, 305-311.

Pinkel,D. and Albertson,D.G. (2005). COMPARATIVE GENOMIC HYBRIDIZATION. Annu. Rev. Genom. Human Genet. 6, 331-354.

Pinto,D., Pagnamenta,A.T., Klei,L., Anney,R., Merico,D., Regan,R., Conroy,J., Magalhaes,T.R., Correia,C., Abrahams,B.S., Almeida,J., Bacchelli,E., Bader,G.D., Bailey,A.J., Baird,G., Battaglia,A., Berney,T., Bolshakova,N., Bolte,S., Bolton,P.F., Bourgeron,T., Brennan,S., Brian,J., Bryson,S.E., Carson,A.R., Casallo,G., Casey,J., Chung,B.H.Y., Cochrane,L., Corsello,C., Crawford,E.L., Crossett,A., Cytrynbaum,C., Dawson,G., de Jonge,M., Delorme,R., Drmic,I., Duketis,E., Duque,F., Estes,A., Farrar,P., Fernandez,B.A., Folstein,S.E., Fombonne,E., Freitag,C.M., Gilbert,J., Gillberg,C., Glessner,J.T., Goldberg,J., Green,A., Green,J., Guter,S.J., Hakonarson,H., Heron,E.A., Hill,M., Holt,R., Howe,J.L., Hughes,G., Hus,V., Igliozzi,R., Kim,C., Klauck,S.M., Kolevzon,A., Korvatska,O., Kustanovich,V., Lajonchere,C.M., Lamb,J.A., Laskawiec,M., Leboyer,M., Le Couteur,A., Leventhal,B.L., Lionel,A.C., Liu,X.Q., Lord,C., Lotspeich,L., Lund,S.C., Maestrini,E., Mahoney,W., Mantoulan,C., Marshall,C.R., McConachie,H., McDougle,C.J., McGrath,J., McMahon,W.M., Merikangas,A., Migita,O., Minshew,N.J., Mirza,G.K., Munson,J., Nelson,S.F., Noakes,C., Noor,A., Nygren,G., Oliveira,G., Papanikolaou,K., Parr,J.R., Parrini,B., Paton,T., Pickles,A., Pilorge,M., Piven,J., Ponting,C.P., Posey,D.J., Poustka,A., Poustka,F., Prasad,A., Ragoussis,J., Renshaw,K., Rickaby,J., Roberts,W., Roeder,K., Roge,B., Rutter,M.L., Bierut,L.J., Rice,J.P., Salt,J., Sansom,K., Sato,D., Segurado,R., Sequeira,A.F., Senman,L., Shah,N., Sheffield,V.C., Soorya,L., Sousa,I., Stein,O., Sykes,N., Stoppioni,V., Strawbridge,C., Tancredi,R., Tansey,K., Thiruvahindrapduram,B., Thompson,A.P., Thomson,S., Tryfon,A., Tsiantis,J., Van Engeland,H., Vincent,J.B., Volkmar,F., Wallace,S., Wang,K., Wang,Z., Wassink,T.H., Webber,C., Weksberg,R., Wing,K., Wittemeyer,K., Wood,S., Wu,J., Yaspan,B.L., Zurawiecki,D., Zwaigenbaum,L., Buxbaum,J.D., Cantor,R.M., Cook,E.H., Coon,H., Cuccaro,M.L., Devlin,B., Ennis,S., Gallagher,L., Geschwind,D.H., Gill,M., Haines,J.L., Hallmayer,J., Miller,J., Monaco,A.P., Nurnberger Jr,J.I., Paterson,A.D., Pericak-Vance,M.A., Schellenberg,G.D., Szatmari,P., Vicente,A.M., Vieland,V.J., Wijsman,E.M., Scherer,S.W., Sutcliffe,J.S., and Betancur,C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466, 368-372.

Pique-Regi,R., Monso-Varona,J., Ortega,A., Seeger,R.C., Triche,T.J., and Asgharzadeh,S. (2008). Sparse representation and Bayesian detection of genome copy number alterations from microarray data. Bioinformatics 24, 309-318.

Pique-Regi,R., Ortega,A., and Asgharzadeh,S. (2009). Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. Bioinformatics 25, 1223-1230.

Pritchard,J.K., Stephens,M., and Donnelly,P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945-959.

Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., de Bakker,P.I.W., Daly,M.J., and Sham,P.C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81, 559-575.

Pushkarev,D., Neff,N.F., and Quake,S.R. (2009). Single-molecule sequencing of an individual human genome. Nature Biotechnology 27, 847-U101.

Rabiner,L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceddings of the IEEE 77.

Redon,R., Ishikawa,S., Fitch,K.R., Feuk,L., Perry,G.H., Andrews,T.D., Fiegler,H., Shapero,M.H., Carson,A.R., Chen,W., Cho,E.K., Dallaire,S., Freeman,J.L., Gonzalez,J.R., Gratacos,M., Huang,J., Kalaitzopoulos,D., Komura,D., MacDonald,J.R., Marshall,C.R., Mei,R., Montgomery,L., Nishimura,K., Okamura,K., Shen,F., Somerville,M.J., Tchinda,J., Valsesia,A., Woodwark,C., Yang,F., Zhang,J., Zerjal,T., Zhang,J., Armengol,L., Conrad,D.F., Estivill,X., Tyler-Smith,C., Carter,N.P., Aburatani,H., Lee,C., Jones,K.W., Scherer,S.W., and Hurles,M.E. (2006). Global variation in copy number in the human genome. Nature 444, 444-454.

Rigaill,G., Hupe,P., Almeida,A., La Rosa,P., Meyniel,J.P., Decraene,C., and Barillot,E. (2008). ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. Bioinformatics 24, 768-774.

Romeo,S., Pennacchio,L.A., Fu,Y., Boerwinkle,E., Tybjaerg-Hansen,A., Hobbs,H.H., and Cohen,J.C. (2007). Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39, 513-516.

Sanders,M.A., Verhaak,R.G., Mc Geertsma-Kleinekoort,W., Abbas,S., Horsman,S., van der Spek,P.J., Lowenberg,B., and Valk,P.J. (2008). SNPExpress: integrated visualization of genome-wide genotypes, copy numbers and gene expression levels. Bmc Genomics 9, Article.

Schuster,S.C., Miller,W., Ratan,A., Tomsho,L.P., Giardine,B., Kasson,L.R., Harris,R.S., Petersen,D.C., Zhao,F., Qi,J., Alkan,C., Kidd,J.M., Sun,Y., Drautz,D.I., Bouffard,P., Muzny,D.M., Reid,J.G., Nazareth,L.V., Wang,Q., Burhans,R., Riemer,C., Wittekindt,N.E., Moorjani,P., Tindall,E.A., Danko,C.G., Teo,W.S., Buboltz,A.M., Zhang,Z., Ma,Q., Oosthuysen,A., Steenkamp,A.W., Oostuisen,H., Venter,P., Gajewski,J., Zhang,Y., Pugh,B.F., Makova,K.D., Nekrutenko,A., Mardis,E.R., Patterson,N., Pringle,T.H., Chiaromonte,F., Mullikin,J.C., Eichler,E.E., Hardison,R.C., Gibbs,R.A., Harkins,T.T., and Hayes,V.M. (2010). Complete Khoisan and Bantu genomes from southern Africa. Nature 463, 943-947.

Sebat,J., Lakshmi,B., Troge,J., Alexander,J., Young,J., Lundin,P., Maner,S., Massa,H., Walker,M., Chi,M., Navin,N., Lucito,R., Healy,J., Hicks,J., Ye,K., Reiner,A., Gilliam,T.C., Trask,B., Patterson,N., Zetterberg,A., and Wigler,M. (2004). Large-Scale Copy Number Polymorphism in the Human Genome. Science 305, 525-528.

Sebat,J., Lakshmi,B., Malhotra,D., Troge,J., Lese-Martin,C., Walsh,T., Yamrom,B., Yoon,S., Krasnitz,A., Kendall,J., Leotta,A., Pai,D., Zhang,R., Lee,Y.H., Hicks,J., Spence,S.J., Lee,A.T., Puura,K., Lehtimaki,T., Ledbetter,D., Gregersen,P.K., Bregman,J., Sutcliffe,J.S., Jobanputra,V., Chung,W., Warburton,D., King,M.C., Skuse,D., Geschwind,D.H., Gilliam,T.C., Ye,K., and Wigler,M. (2007). Strong Association of De Novo Copy Number Mutations with Autism. Science 316, 445-449.

Sebat,J. (2007). Major changes in our DNA lead to major changes in our thinking. Nat GenetNat Genet.39 (7 Suppl):S3-5

Sha,B.Y., Yang,T.L., Zhao,L.J., Chen,X.D., Guo,Y., Chen,Y., Pan,F., Zhang,Z.X., Dong,S.S., Xu,X.H., and Deng,H.W. (2009b). Genome-wide association study suggested copy number variation may be associated with body mass index in the Chinese population. Journal of Human Genetics 54, 199-202.

Sharp,A.J., Cheng,Z., and Eichler,E.E. (2006). Structural Variation of the Human Genome. Annu. Rev. Genom. Human Genet. 7, 407-442.

Sharp,A.J., Locke,D.P., McGrath,S.D., Cheng,Z., Bailey,J.A., Vallente,R.U., Pertz,L.M., Clark,R.A., Schwartz,S., Segraves,R., Oseroff,V.V., Albertson,D.G., Pinkel,D., and Eichler,E.E. (2005). Segmental Duplications and Copy-Number Variation in the Human Genome. Am J Hum Genet 77, 78-88.

She,X., Jiang,Z., Clark,R.A., Liu,G., Cheng,Z., Tuzun,E., Church,D.M., Sutton,G., Halpern,A.L., and Eichler,E.E. (2004). Shotgun sequence assembly and recent segmental duplications within the human genome. Nature 431, 927-930.

Shelling,A.N. and Ferguson,L.R. (2007). Genetic variation in human disease and a new role for copy number variants. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis In Press, Corrected Proof.

Shen,F., Huang,J., Fitch,K.R., Truong,V.B., Kirby,A., Chen,W., Zhang,J., Liu,G., McCarroll,S.A., Jones,K.W., and Shapero,M.H. (2008). Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes. Bmc Genetics 9, Article.

Shifman,S. and Darvasi,A. (2001). The value of isolated populations. Nat Genet 28, 309-310.

Shtir,C., Pique-Regi,R., Siegmund,K., Morrison,J., Schumacher,F., and Marjoram,P. (2009). Copy number variation in the Framingham Heart Study. BMC Proceedings 3, S133.

Smith,A.J., Tsalenko,A., Sampas,N., Scheffer,A., Yamada,N.A., Tsang,P., Ben-Dor,A., Yakhini,Z., Ellis,R.J., Bruhn,L., Laderman,S., Froguel,P., and Blakemore,A.I. (2007). 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. ddm208.

Smith,B., Campbell,H., Blackwood,D., Connell,J., Connor,M., Deary,I., Dominiczak,A., Fitzpatrick,B., Ford,I., Jackson,C., Haddow,G., Kerr,S., Lindsay,R., McGilchrist,M., Morton,R., Murray,G., Palmer,C., Pell,J., Ralston,S., St Clair,D., Sullivan,F., Watt,G., Wolf,R., Wright,A., Porteous,D., and Morris,A. (2006). Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. BMC Medical Genetics 7, 74.

Snijders,A.M., Nowak,N.J., Huey,B., Fridlyand,J., Law,S., Conroy,J., Tokuyasu,T., Demir,K., Chiu,R., Mao,J.H., Jain,A.N., Jones,S.J.M., Balmain,A., Pinkel,D., and Albertson,D.G. (2005). Mapping segmental and sequence variations among laboratory mice using BAC array CGH. Genome Res. 15, 302-311.

Solinas-Toldo,S., Lampel,S., Stilgenbauer,S., Nickolenko,J., Benner,A., D 枚 hner,H., Cremer,T., and Lichter,P. (1997). Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. Genes Chromosom. Cancer 20, 399-407.

Springer,N.M., Ying,K., Fu,Y., Ji,T., Yeh,C.T., Jia,Y., Wu,W., Richmond,T., Kitzman,J., Rosenbaum,H., Iniguez,A.L., Barbazuk,W.B., Jeddeloh,J.A., Nettleton,D., and Schnable,P.S. (2009). Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. PLoS Genet 5, e1000734.

Stankiewicz,P. and Lupski,J.R. (2006). The genomic basis of disease, mechanisms and assays for genomic disorders. Genome and Disease 1, 1-16.

Stefansson,H., Rujescu,D., Cichon,S., Pietilainen,O.P.H., Ingason,A., Steinberg,S., Fossdal,R., Sigurdsson,E., Sigmundsson,T., Buizer-Voskamp,J.E., Hansen,T., Jakobsen,K.D., Muglia,P., Francks,C., Matthews,P.M., Gylfason,A., Halldorsson,B.V., Gudbjartsson,D., Thorgeirsson,T.E., Sigurdsson,A., Jonasdottir,A., Jonasdottir,A., Bjornsson,A., Mattiasdottir,S., Blondal,T., Haraldsson,M., Magnusdottir,B.B., Giegling,I., Moller,H.J., Hartmann,A., Shianna,K.V., Ge,D., Need,A.C., Crombie,C., Fraser,G., Walker,N., Lonnqvist,J., Suvisaari,J., Tuulio-Henriksson,A., Paunio,T., Toulopoulou,T., Bramon,E., Di Forti,M., Murray,R., Ruggeri,M., Vassos,E., Tosato,S., Walshe,M., Li,T., Vasilescu,C., Muhleisen,T.W., Wang,A.G., Ullum,H., Djurovic,S., Melle,I., Olesen,J., Kiemeney,L.A., Franke,B., Sabatti,C., Freimer,N.B., Gulcher,J.R., Thorsteinsdottir,U., Kong,A., Andreassen,O.A., Ophoff,R.A., Georgi,A., Rietschel,M., Werge,T., Petursson,H., Goldstein,D.B., Nothen,M.M., Peltonen,L., Collier,D.A., St Clair,D., and Stefansson,K. (2008). Large recurrent microdeletions associated with schizophrenia. Nature 455, 232-236.

Stranger,B.E., Forrest,M.S., Dunning,M., Ingle,C.E., Beazley,C., Thorne,N., Redon,R., Bird,C.P., de Grassi,A., Lee,C., Tyler-Smith,C., Carter,N., Scherer,S.W., Tavare,S., Deloukas,P., Hurles,M.E., and Dermitzakis,E.T. (2007). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science 315, 848-853.

Sudmant,P.H., Kitzman,J.O., Antonacci,F., Alkan,C., Malig,M., Tsalenko,A., Sampas,N., Bruhn,L., Shendure,J., Genomes,P., and Eichler,E.E. (2010). Diversity of Human Copy Number Variation and Multicopy Genes. Science 330, 641-646.

Szentirmay,Z. (2003). The effect about the knowledge of the human genome on the development of pathology. Orvosi Hetilap 144, 2499-2508.

The International HapMap Consortium (2003). The International HapMap Project. Nature 426, 789-796.

The International HapMap Consortium (2005). A haplotype map of the human genome. Nature 437, 1299-1320.

The International Schizophrenia Consortium (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455, 237-241.

Ting,J.C., Roberson,E.D., Miller,N.D., Lysholm-Bernacchi,A., Stephan,D.A., Capone,G.T., Ruczinski,I., Thomas,G.H., and Pevsner,J. (2007). Visualization of uniparental inheritance, Mendelian inconsistencies, deletions, and parent of origin effects in single nucleotide polymorphism trio data with SNPtrio. Human Mutation. 28(12):1225-35.

Toujani,S., Dessen,P., Ithzar,N., Danglot,G.l., Richon,C., Vassetzky,Y., Robert,T., Lazar,V., Bosq,J., Da Costa,L., P茅rot,C., Ribrag,V., Patte,C., Wiels,J., and Bernheim,A. (2009). High Resolution Genome-Wide Analysis of Chromosomal Alterations in Burkitt's Lymphoma. Plos One 4, e7089.

Tsuang,D.W., Millard,S.P., Ely,B., Chi,P., Wang,K., Raskind,W.H., Kim,S., Brkanac,Z., and Yu,C.E. (2010). The Effect of Algorithms on Copy Number Variant Detection. Plos One 5, e14456.

Tuzun,E., Sharp,A.J., Bailey,J.A., Kaul,R., Morrison,V.A., Pertz,L.M., Haugen,E., Hayden,H., Albertson,D., Pinkel,D., Olson,M.V., and Eichler,E.E. (2005). Fine-scale structural variation of the human genome. Nat Genet 37, 727-732.

Urban,A.E., Korbel,J.O., Selzer,R., Richmond,T., Hacker,A., Popescu,G.V., Cubells,J.F., Green,R., Emanuel,B.S., Gerstein,M.B., Weissman,S.M., and Snyder,M. (2006). High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America 103, 4534-4539.

Velagaleti, Gopalrao V. N., Bien-Willner, Gabriel A., Northup, Jill K., Lockhart, Lillian H., Hawkins, Judy C., Jalal, Syed M., Withers, Marjorie, Lupski, James R., and Stankiewicz, Pawel. Position Effects Due to Chromosome Breakpoints that Map <900 Kb Upstream and <1.3 Mb Downstream of SOX9 in Two Patients with Campomelic Dysplasia. American Journal of Human Genetics 76[4], 652-662. 1-4-2005.

Venkatraman, E.S. and Olshen, A.B. (2007b) DNAcopy: A Package for Analyzing DNA Copy Data. Department of Epidemiology and Biostatistics. Memorial Sloan-Kettering Cancer Center

Venkatraman,E.S. and Olshen,A.B. (2007a). A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23, 657-663.

Vitart,V., Bencic,G., Hayward,C., Herman,J.S., Huffman,J., Campbell,S., Bucan,K., Zgaga,L., Kolcic,I., Polasek,O., Campbell,H., Wright,A., Vatavuk,Z., and Rudan,I. (2010). Heritabilities of Ocular Biometrical Traits in Two Croatian Isolates with Extended Pedigrees. Investigative Ophthalmology & Visual Science 51, 737-743.

Vitart,V., Biloglav,Z., Hayward,C., Janicijevic,B., Smolej-Narancic,N., Barac,L., Pericic,M., Klaric,I.M., Skaric-Juric,T., Barbalic,M., Polasek,O., Kolcic,I., Carothers,A., Rudan,P., Hastie,N., Wright,A., Campbell,H., and Rudan,I. (2006). 3000 years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia. Eur J Hum Genet 14, 478-487.

Vrijenhoek,T., Buizer-Voskamp,J.E., van,d.S., I, Strengman,E., Genetic Risk and Outcome in Psychosis (GROUP) Consortium, Sabatti,C., Geurts van,K.A., Brunner,H.G., Ophoff,R.A., and Veltman,J.A. (2008). Recurrent CNVs disrupt three candidate genes in schizophrenia patients. American Journal of Human Genetics. 83(4):504-10.

Walsh,T., McClellan,J.M., McCarthy,S.E., Addington,A.M., Pierce,S.B., Cooper,G.M., Nord,A.S., Kusenda,M., Malhotra,D., Bhandari,A., Stray,S.M., Rippey,C.F., Roccanova,P., Makarov,V., Lakshmi,B., Findling,R.L., Sikich,L., Stromberg,T., Merriman,B., Gogtay,N., Butler,P., Eckstrand,K., Noory,L., Gochman,P., Long,R., Chen,Z., Davis,S., Baker,C., Eichler,E.E., Meltzer,P.S., Nelson,S.F., Singleton,A.B., Lee,M.K., Rapoport,J.L., King,M.C., and Sebat,J. (2008). Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. Science 320, 539-543.

Wang,J., Wang,W., Li,R., Li,Y., Tian,G., Goodman,L., Fan,W., Zhang,J., Li,J., Zhang,J., Guo,Y., Feng,B., Li,H., Lu,Y., Fang,X., Liang,H., Du,Z., Li,D., Zhao,Y., Hu,Y., Yang,Z., Zheng,H., Hellmann,I., Inouye,M., Pool,J., Yi,X., Zhao,J., Duan,J., Zhou,Y., Qin,J., Ma,L., Li,G., Yang,Z., Zhang,G., Yang,B., Yu,C., Liang,F., Li,W., Li,S., Li,D., Ni,P., Ruan,J., Li,Q., Zhu,H., Liu,D., Lu,Z., Li,N., Guo,G., Zhang,J., Ye,J., Fang,L., Hao,Q., Chen,Q., Liang,Y., Su,Y., san,A., Ping,C., Yang,S., Chen,F., Li,L., Zhou,K., Zheng,H., Ren,Y., Yang,L., Gao,Y., Yang,G., Li,Z., Feng,X., Kristiansen,K., Wong,G.K.-S., Nielsen,R., Durbin,R., Bolund,L., Zhang,X., Li,S., Yang,H., and Wang,J. (2008). The diploid genome sequence of an Asian individual. Nature 456, 60-65.

Wang,K., Li,M., Hadley,D., Liu,R., Glessner,J., Grant,S.F., Hakonarson,H., and Bucan,M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research. 17(11):1665-74.

Wang,K., Wei-Dong Li, Joseph T.Glessner, Struan F.Al Grant, Hakon Hakonarson, and R.Arlen Price (2010). Large Copy-Number Variations Are Enriched in Cases With Moderate to Extreme Obesity. Diabetes 59, 2690-2694.

Wang,L., Fidler,C., Nadig,N., Giagounidis,A., la Porta,M.G., Malcovati,L., Killick,S., Gattermann,N., Aul,C., Boultwood,J., and Wainscoat,J.S. (2008). Genome-wide analysis of copy number changes and loss of heterozygosity in myelodysplastic syndrome with del(5q) using high-density single nucleotide polymorphism arrays. Haematologica. 93(7):994-1000.

Wang,L.S., Hranilovic,D., Wang,K., Lindquist,I., Yurcaba,L., Petkovic,Z.B., Gidaya,N., Jernej,B., Hakonarson,H., and Bucan,M. (2010). Population-based study of genetic variation in individuals with autism spectrum disorders from Croatia. BMC Medical Genetics 11, 134.

Wang,W.Y.S., Barratt,B.J., Clayton,D.G., and Todd,J.A. (2005). Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6, 109-118.

Wheeler,D.A., Srinivasan,M., Egholm,M., Shen,Y., Chen,L., McGuire,A., He,W., Chen,Y.J., Makhijani,V., Roth,G.T., Gomes,X., Tartaro,K., Niazi,F., Turcotte,C.L., Irzyk,G.P., Lupski,J.R., Chinault,C., Song,X.Z., Liu,Y., Yuan,Y., Nazareth,L., Qin,X., Muzny,D.M., Margulies,M., Weinstock,G.M., Gibbs,R.A., and Rothberg,J.M. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872-8U5.

Williams,N.M., Zaharieva,I., Martin,A., Langley,K., Mantripragada,K., Fossdal,R., Stefansson,H., Stefansson,K., Magnusson,P., Gudmundsson,O.O., Gustafsson,O., Holmans,P., Owen,M.J., O'Donovan,M., and Thapar,A. (2010). Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. The Lancet 376, 1401-1408.

Winchester,L., Newbury,D.F., Monaco,A.P., and Ragoussis,J. (2008). Detection, breakpoint identification and detailed characterisation of a CNV at the FRA16D site using SNP assays. Cytogenetic and Genome Research 123, 322-332.

Winchester,L., Yau,C., and Ragoussis,J. (2009). Comparing CNV detection methods for SNP arrays. Briefings in Functional Genomics and Proteomics 8, 353-366.

Wineinger,N.E., Patki,A., Meyers,K.J., Broeckel,U., Gu,C.C., Rao,D., Devereux,R.B., Arnett,D.K., and Tiwari,H.K. (2011). Genome-wide joint SNP and CNV analysis of aortic root diameter in African Americans: the HyperGEN study. BMC Medical Genomics 4.

Wright,A., Charlesworth,B., Rudan,I., Carothers,A., and Campbell,H. (2003). A polygenic basis for late-onset disease. Trends in Genetics 19, 97-106.

Wright,A.F., Carothers,A.D., and Pirastu,M. (1999). Population choice in mapping genes for complex diseases. Nat Genet 23, 397-404.

WTCCC (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713-720.

Xi,R., Kim,T.M., and Park,P.J. (2010). Detecting structural variations in the human genome using next generation sequencing. Briefings in Functional Genomics and Proteomics 9, 405-415.

Xing,J., Zhang,Y., Han,K., Salem,A.H., Sen,S.K., Huff,C.D., Zhou,Q., Kirkness,E.F., Levy,S., Batzer,M.A., and Jorde,L.B. (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res.

Xu,B., Roos,J.L., Levy,S., van Rensburg,E.J., Gogos,J.A., and Karayiorgou,M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. Nat Genet advanced online publication.

Yang,H., Ding,Y.M., Hutchins,L.N., Szatkiewicz,J., Bell,T.A., Paigen,B.J., Graber,J.H., de Villena,F.P.M., and Churchill,G.A. (2009). A customized and versatile high-density genotyping array for the mouse. Nature Methods 6, 663-U55.

Yang,S., Wang,K., Gregory,B., Berrettini,W., Wang,L.S., Hakonarson,H., and Bucan,M. (2009). Genomic landscape of a three-generation pedigree segregating affective disorder. Plos One 4, e4474.

Yang,T.L., Chen,X.D., Guo,Y., Lei,S.F., Wang,J.T., Zhou,Q., Pan,F., Chen,Y., Zhang,Z.X., Dong,S.S., Xu,X.H., Yan,H., Liu,X., Qiu,C., Zhu,X.Z., Chen,T., Li,M., Zhang,H., Zhang,L., Drees,B.M., Hamilton,J.J., Papasian,C.J., Recker,R.R., Song,X.P., Cheng,J., and Deng,H.W. (2008). Genome-wide Copy-Number-Variation Study Identified a Susceptibility Gene, UGT2B17, for Osteoporosis. American Journal of Human Genetics 83, 663-674.

Yau, C. (2007) QuantiSNP Pre-Release Version User Documentation. University of Oxford.

Ylstra,B., van den IJssel,P., Carvalho,B., Brakenhoff,R.H., and Meijer,G.A. (2006). BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucl. Acids Res. 34, 445-450.

Yoon,S., Xuan,Z., Makarov,V., Ye,K., and Sebat,J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 19, 1586-1592.

Zhang, J. (2008). Markov Chains and Hidden Markov Model. Lecture notes. Purdue University, US. http://www.stat.purdue.edu/~jianzhan/notes/HMM.pdf (downloaded in March 2010)

Zhang,D., Cheng,L., Qian,Y., liey-Rodriguez,N., Kelsoe,J.R., Greenwood,T., Nievergelt,C., Barrett,T.B., McKinney,R., Schork,N., Smith,E.N., Bloss,C., Nurnberger,J., Edenberg,H.J., Foroud,T., Sheftner,W., Lawson,W.B., Nwulia,E.A., Hipolito,M., Coryell,W., Rice,J., Byerley,W., McMahon,F., Schulze,T.G., Berrettini,W., Potash,J.B., Belmonte,P.L., Zandi,P.P., McInnis,M.G., Zollner,S., Craig,D., Szelinger,S., Koller,D., Christian,S.L., Liu,C., and Gershon,E.S. (2009). Singleton deletions throughout the genome increase risk of bipolar disorder. Molecular Psychiatry 14, 376-380.

Zhang,D., Qian,Y., Akula,N., liey-Rodriguez,N., Tang,J., Gershon,E.S., Liu,C., and The Bipolar,G.S. (2011). Accuracy of CNV Detection from GWAS Data. Plos One 6, e14511.

Zhang,F., Gu,W., Hurles,M.E., and Lupski,J.R. (2009). Copy number variation in human health, disease, and evolution. Annu Rev Genomics Hum Genet 10, 451-481.

Zogopoulos,G., Ha,K.C., Naqib,F., Moore,S., Kim,H., Montpetit,A., Robidoux,F., Laflamme,P., Cotterchio,M., Greenwood,C., Scherer,S.W., Zanke,B., Hudson,T.J., Bader,G.D., and Gallinger,S. (2007). Germ-line DNA copy number variation frequencies in a large North American population. Human Genetics. 122(3-4):345-53.

List of Appendices

Appendix 1: Table of Individual CNVs in Dalmatian, Orcadian and South Tyrolean Populations

Appendix 2: Table of CNVRs in Dalmatian, Orcadian and South Tyrolean Populations

The content of Appendices can be viewed in electronic form, which is contained in a CD in the case attached to the back cover of the thesis.