Chapter 1
Introduction

1.1 Copy Number Variation (CNV) as a Source of Genetic Diversity

Copy number variation (CNV) has long been studied as gene copy number difference among individuals at specific loci, for example in the α- and β- globin gene locus; or as genomic imbalance resulting in diseases such as Prader-Willi and Angelman Syndrome. Until recently, such rearrangements in the human genome were assumed to be limited in scale. Since 2004, however, research has revealed CNVs as a major source of human genetic variation (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005; Conrad et al. 2006; McCarroll et al. 2006; Redon et al. 2006). The unprecedented level of genetic diversity conferred by CNVs has opened a new chapter in our understanding of phenotypic variation, human evolution and disease susceptibility.

Copy number variation is defined as deletion and/or duplication of a DNA sequence larger than 1 kb in length (Feuk et al. 2006a; Freeman et al. 2006). It belongs to a spectrum of genetic variations ranging from the ubiquitous single nucleotide polymorphisms (SNPs), to fine-scale copy number changes such as small insertions and deletions (INDELs), microsatellite and minisatellite repeats, to larger scale structural variations such as inversions, translocations and CNVs (Bowcock et al. 1994; Armour et al. 1996; IHMC 2005; Feuk et al. 2006a; Mills et al. 2006; Conrad and Hurles 2007) (Figure 1.1). It is these variations and polymorphisms that constitute the dynamic human genome architecture and underlie the differences between individuals.

With remarkable advance in microarray technologies, and the availability of a complete human genome sequence, it became possible to obtain genome-wide maps of the locations and frequencies of CNVs. The first reports of wide-spread copy number variations or copy number polymorphisms (CNPs) (CNVs at >1% frequency) appeared in 2004. Iafrate *et al.* and Sebat *et al.* independently surveyed the human genome for copy

number changes, revealing the extent of this source of human genetic variation at a previously unanticipated level (Iafrate et al. 2004; Sebat et al. 2004). The two microarray-based CNV studies detected a total of 476 CNV loci from 75 individuals. Some of these loci affect genes that play important biological roles such as neurological functions and metabolism (Sebat et al. 2004). Since then, more copy number changes were mined from SNP genotyping data and clone paired-end sequencing data (Tuzun et al. 2005; Conrad et al. 2006; McCarroll et al. 2006).

In 2006, a comprehensive map of copy number variations was released based on 270 individuals from four different ethnic populations of European, Asian or African descents, who were originally included in the International HapMap Project. Redon *et al.* reported a total of 1447 copy number variable regions (CNVRs), corresponding to 360 Mb of human DNA sequences or 12% of the human genome (Redon et al. 2006). The CNVRs are enriched in functional categories such as cell adhesion, sensory perception of smell and of chemical stimulus, as well as neurophysiological processes. Distribution of CNVRs is non-uniform along the genome, with copy number changes preferentially clustered near segmental duplications (SDs), defined as duplicated sequences of >1 kb with 90% or more sequence identity in the reference human genome assembly (Bailey et al. 2002). Multi-allelic and complex CNVs are especially enriched in these SD regions. A number of following reports targeting specifically at segmental duplication regions also defined these loci as hotspots for chromosomal rearrangement and copy number variations (Sharp et al. 2005; Locke et al. 2006).

In the past years, reports of more comprehensive CNV discovery and characterization have come into view (de Smith et al. 2007; Pinto et al. 2007; Simon-Sanchez et al. 2007; Wang et al. 2007; Wong et al. 2007; Zogopoulos et al. 2007; Cooper et al. 2008). In

particular, recent studies on fine-scale CNV detection based on novel sequencing technologies and genome comparisons provide us with new perspectives on CNVs and structural variations (Khaja et al. 2006; Korbel et al. 2007; Levy et al. 2007; Kidd et al. 2008). These studies have revealed that individual human genomes are at least 0.5% different, with the order of 600-900 CNVs between any two individuals, contrary to our traditional view of human sharing 99.9% similarity (Korbel et al. 2008).

The phenomenon of copy number variations is not unique to the human genome. CNVs have been characterized in a number of mammalian species from the great apes (Locke et al. 2003; Perry et al. 2006; Wilson et al. 2006b; Perry et al. 2008) to rat and mouse (Li et al. 2004; Adams et al. 2005; Snijders et al. 2005; Egan et al. 2007; Guryev et al. 2008; She et al. 2008). CNV maps for the Drosophila genome was also recently published (Dopman and Hartl 2007; Emerson et al. 2008; Zhou et al. 2008), cataloguing differences in Drosophila's gene copy numbers, a concept that was developed in Drosophila genetics as early as the 1930s (Bridges 1936).

Whilst early CNV detection platforms mature and new technologies emerge, the discovery and characterization of CNVs has surfaced as an exciting branch of human genetics. Copy number changes were increasingly explored in the context of population demographics and human evolution (Conrad and Hurles 2007), with studies based on inter-population comparisons (Redon et al. 2006; Jakobsson et al. 2008) and inter-species evaluations (Locke et al. 2003; Perry et al. 2008). Population bias in copy numbers at specific loci have been demonstrated to play an important role in environmental adaptations, for example the evolution of CNV in the salivary amylase gene locus has been influenced by human diet (Perry et al. 2007); and the HIV-1 susceptibility CNV locus *CCL3L1* showed variable distribution in different populations

(Gonzalez et al. 2005). Of equal importance, CNV has brought a new dimension to disease genetics. Analogous to SNPs, an instrumental tool in discovering disease genes through linkage and association analysis, CNVs is a major source of genetic variation with potential clinical relevance in a number of diseases (see Chapter 1.4).

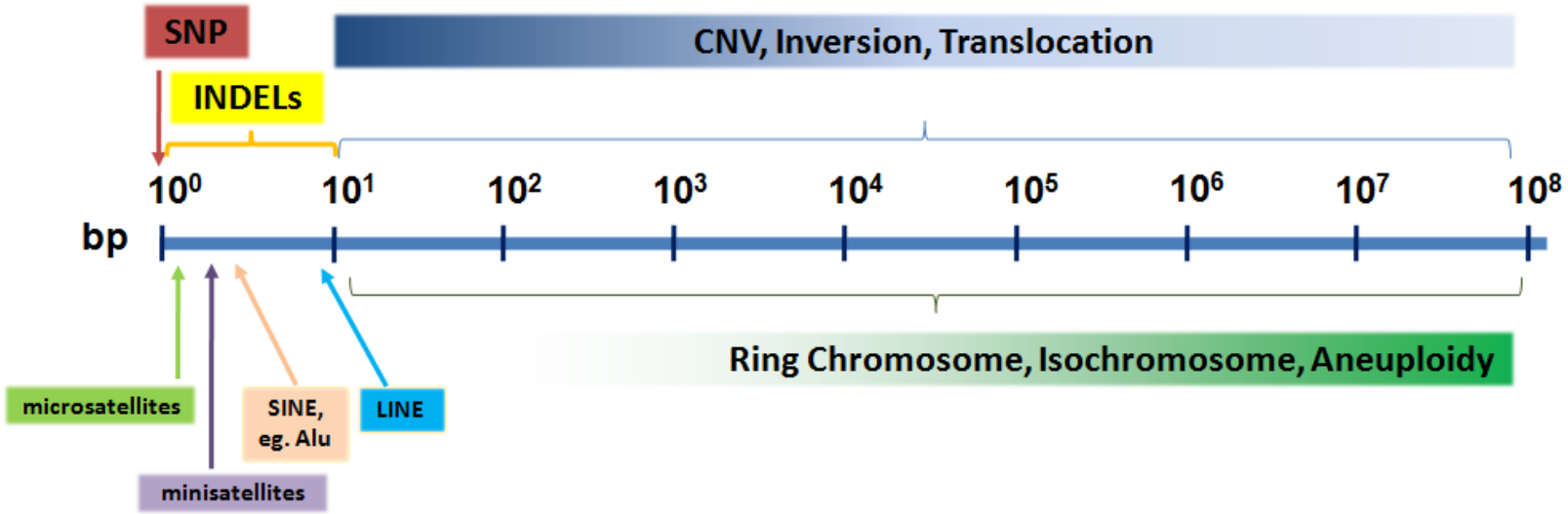

Figure 1.1 Types of genetic variants and their relative sizes. (SNP: Single Nucleotide Polymorphism; CNV: Copy Number Variation; LINE: Long Interspersed Nuclear Element; SINE: Short Interspersed Nuclear Element; Alu: a family of repeat elements named after the *Alu*I restriction site; bp: base pair)

1.2 Detection of Copy Number Variation

1.2.1 Classical Cytogenetic Techniques for the Detection of Structural Variations

Classical cytogenetics and traditional karyotype techniques such as Giemsa-banding (G-banding) (Craig and Bickmore 1993) have revealed gross structural variations as microscopically visible alterations. Cases of large-scale chromosomal rearrangements leading to Down syndrome (trisomy 21) and cri-du-chat syndrome (terminal 5p deletion) were discovered in the 1950-60s, many decades before the introduction of high-resolution techniques for rearrangement detection.

Fluorescence in situ hybridisation (FISH) further increased the resolution of cytogenetic techniques to 104–106 bp, extending the reach of traditional karyotyping to the detection of submicroscopic copy number changes. FISH (Bauman et al. 1980) is an in situ hybridisation technique in which a labelled probe of specific DNA sequences (e.g. a Bacterial Artificial Chromosome (BAC)/fosmid clone) is hybridized to a preparation of metaphase chromosomes or interphase DNA, usually attached to a glass slide. The chromosomal DNA and probe mixture was then denatured, allowing the single-stranded probe and single stranded DNA to re-anneal, with the probe hybridizing to the complementary sequences on the DNA and reformed a double stranded molecule. Following hybridization, unbound probes were washed away, and the hybridized probes were visualized directly if they were tagged with fluorochromes (e.g. Cyanine (Cy) or Alexa Fluors dyes), or detected by antibodies against hapten-tagged probes, or affinity agents such as avidin or streptavidin if probes were labelled with the biotin and digoxigenin systems (Langer, 1981).

FISH techniques have evolved throughout the years. For instance, improvement on probe labelling has led to the development of multi-probe FISH and spectral karyotype (SKY) (Schrock et al. 1996). These multi-colour hybridizations allow visualizing of rearrangements involving different chromosomes. On the other hand, the DNA preparation method has also progressed such that extended chromatin fibres (as a replacement for interphase or metaphase preparation) were fixed onto glass slides for high-resolution detection of submicroscopic changes, a technique known as Fiber-FISH (Florijn et al. 1995). This allows visualization of small copy number changes, for example deletion gaps or tandem duplications down to the size of a fosmid clone (40 kb) or even smaller. The technique can also be used to solve more complex rearrangements, as in the case of the expansion of CCL3L1 and related segmental duplication in human and chimpanzees.

Along with the development of FISH, complementary techniques such as comparative genome hybridization (CGH) (Kallioniemi et al. 1992) were developed for the detection of structural rearrangement. In CGH, differentially labelled test and reference DNA were simultaneously hybridized to chromosome spreads. Unlabelled Cot-I DNA was also co-hybridized to block DNA repeats. The hybridisation was then detected with two fluorochromes, and genomic regions of gains or losses would be detected as changes in the ratio of intensities of the two fluorophores along the chromosome. The technique was widely used in the analysis of tumor malignancies and constitutional chromosomal aberrations (Forozan et al. 1997).

The major advantage of CGH is that it allowed whole-chromosome or whole-genome surveys of chromosomal rearrangement and aberrations, compared to previous target-specific approaches. Figure 1.2 shows the image of a CGH experiment applied on a

cancer cell line when the technique first developed. High-level amplification of the *myc* locus at 8q24 was revealed as an elevated ratio of test/reference signal intensity. Nevertheless, such CGH approaches remained limited in resolution by the use of metaphase chromosomes as DNA hybridization targets. Therefore, higher resolution detection would still require laborious locus-by-locus techniques, such as Southern blot analysis (Southern 1975) or pulse field gel electrophoresis (PFGE) (Schwartz et al. 1983; Herschleb et al. 2007) (see section 1.2.5).

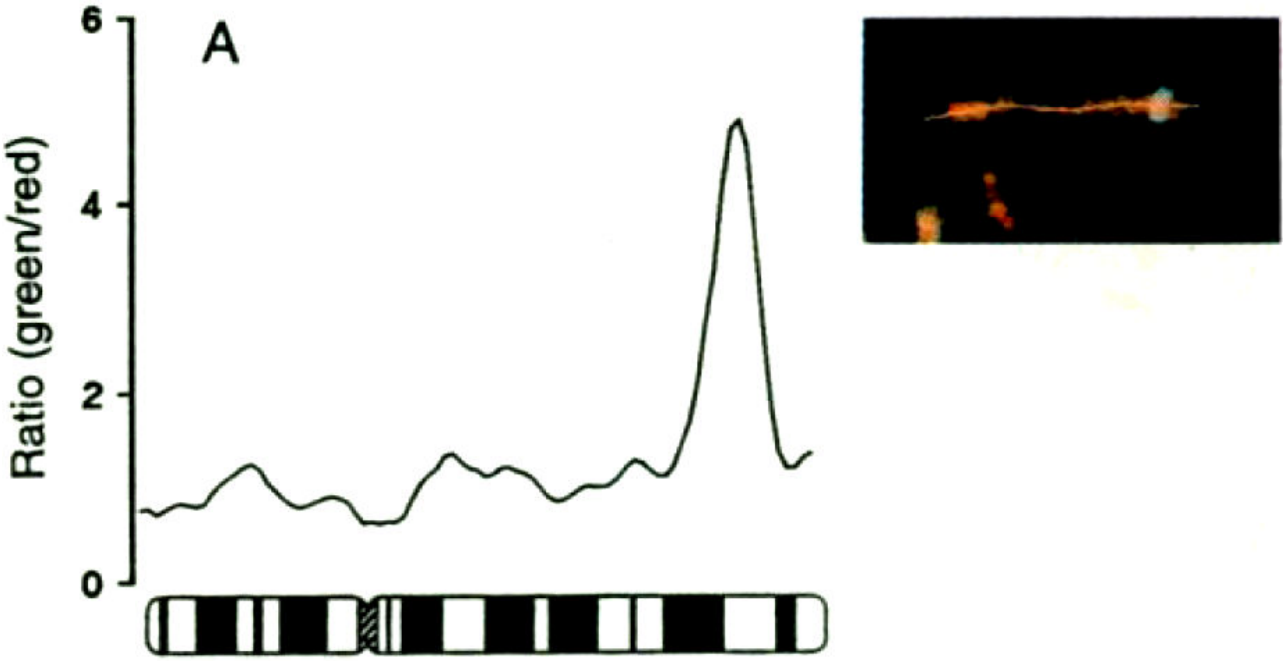

Figure 1.2 Comparative Genome Hybridization applied on a cancer cell line revealed amplification of the *myc* locus. (Diagram reproduced from Kallioniemi *et al.* 1992, *Science* **258**(5083))
Kallioniemi *et al.* performed a CGH experiment on cell line COLO 320HSR (cancer cell line) (labelled in green) versus a normal reference cell line (labelled in red). Signal from the two fluorophores was captured and analyzed by digital image analysis system, which estimated the ratio of intensities of the two fluorophores along chromosome 8 (left). The *myc* locus at chromosome 8q24 showed a high green-to-red ratio (right), consistent with known high-level amplification of *myc* in the cancer cell line.

1.2.2 Array-based Comparative Genome Hybridization

Array-based comparative genome hybridisation (array CGH), a modification to the traditional CGH technique, greatly enhanced the resolution and dynamic range for copy number detection. By substituting chromosome targets in a traditional CGH experiment by a matrix of defined nucleic acid target sequences spotted on glass chips, Solinas-Toldo *et al.* developed the prototype of a CGH array (Solinas-Toldo et al. 1997). Pinkel *et al.* further implemented the platform to screen for copy number changes in human breast cancer, revealing previously undetected aberrations (Pinkel et al. 1998). The array CGH platform has since opened many new opportunities to assess copy number rearrangements associated with human disease and genetic diversity.

A typical CGH array (refer to Fig. 1.3) consists of mapped DNA sequences mechanically spotted or directly synthesized onto microscope glass slides. Various array approaches employed different sources of DNA sequences, which could be broadly classified as genomic inserts (such as BAC, cosmid or fosmid clones (Solinas-Toldo et al. 1997; Pinkel et al. 1998; Snijders et al. 2001), cDNA clones (Pollack et al. 1999), genomic Polymerase Chain Reaction (PCR) products (Dhami et al. 2005) or oligonucleotides (Urban et al. 2006). For each hybridization experiment, uniquely labelled subject and control DNA are co-hybridised with Cot-1 blocking agent (to suppress signal from common repetitive sequences). The test and reference DNA signal intensity was recorded for all probes on the array. Significant deviation from the 1:1 test/reference for a probe (or a series of consecutive probes) would be interpreted as DNA copy number changes (Pinkel and Albertson 2005).

1.2.2.1 BAC Array CGH

The first generation of CGH arrays, and indeed the first genome-wide human CNV screens, was predominantly based on Bacterial Artificial Clones immobilized onto coated glass slides (Solinas-Toldo et al. 1997; Pinkel et al. 1998; Snijders et al. 2001; Fiegler et al. 2003; Iafrate et al. 2004; Redon et al. 2006). BAC arrays started out targeting specific regions of the genome, or tiling the genome at an average resolution of ~1 Mb. With the availability of the overlapping sequencing clone contigs generated for the public domain of the Human Genome Project (Lander et al. 2001; IHGSC 2004), BAC arrays can now be synthesized with over 30,000 features at a tiling path resolution of ~80-150 kb (Ishkanian et al. 2004; Fiegler et al. 2006). In terms of DNA for array spotting, traditionally, substantial effort was required in bacterial culture handling to extract enough DNA from the BAC clones (Solinas-Toldo et al. 1997; Pinkel et al. 1998; Albertson and Pinkel 2003). The problem was later circumvented by applying DNA amplification methods such as rolling circle replication (Buckley et al. 2002), linker adaptor PCR (Klein et al. 1999; Snijders et al. 2001) or degenerate oligonucleotide primers PCR (DOP-PCR) (Fiegler et al. 2003). The whole-genome tiling path (WGTP) array platform was generated using the DOP-PCR strategy (Telenius et al. 1992), with BAC DNA amplified using three different, specifically designed degenerate oligonucleotide primers. Complete amplification of the clone DNA was achieved, and contamination from *E. coli* host vector DNA was minimized. In summary, whole genome tiling path BAC arrays provide a rapid, sensitive and reliable platform for the detection of genomic imbalances (Fiegler et al. 2006).

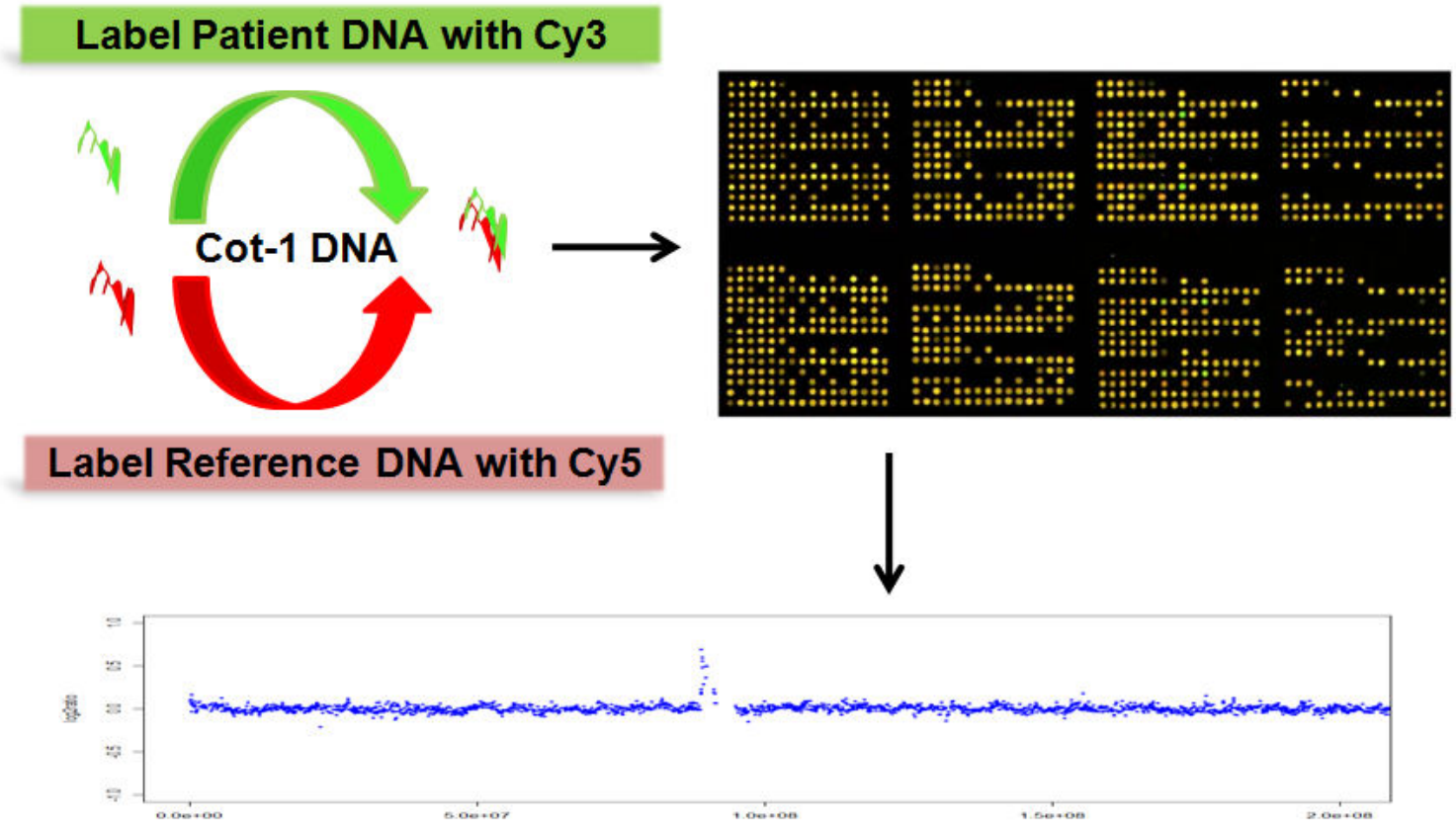

Figure 1.3 Schematics of an array comparative genome hybridization experiment.
The test (green) and reference (red) DNAs are differentially labelled and hybridised to a microarray slide in the presence of Cot-1 DNA. Duplications and deletions are identified as deviations in the signal intensity ratios from the two fluorochromes.

1.2.2.2 Oligonucleotide Array CGH

An alternative to BAC arrays are oligonucleotide-based arrays, which are mostly supplied commercially (e.g. from Roche Nimblegen and Agilent Technologies). These arrays are synthesized in-situ with ~20-80-mer oligonucleotides which form the probes or features for CNV detection. Array designs can be optimized to avoid highly repetitive regions, but with the option of covering low copy repeats such as segmental duplications where CNVs are abundant. Designs can also be customized to target clinically relevant regions for disease studies (Urban et al. 2006; Baldwin et al. 2008).

Early oligo arrays were generally of poorer signal-to-noise ratio compared to BAC arrays, thus resulting in more variable reported signals for CNV detection (Carter 2007). Furthermore, due to the higher costs involved in commercial arrays and reagent purchase, oligo arrays were initially applied mostly for validation (Conrad et al. 2006; Locke et al. 2006; Wong et al. 2007) or breakpoint mapping (Sharp et al. 2005; Gribble et al. 2007). Advancement in technologies, such as the use of digital mask photolithography (Nuwaysir et al. 2002), allow oligo arrays to be constructed at much higher density (Urban et al. 2006), providing considerably better resolution and precision for CNV detection . With improved signal-to-noise ratio, enhanced reproducibility and quality control and decreasing cost per feature, oligo arrays are now recognized as a tool for accurate, high-resolution CNV detection (Ylstra et al. 2006). Examples of the latest whole-genome oligo arrays for human CNV detection include the Agilent 244A (with ~244.000 features) and Nimblegen HG18 WG Tiling CGH 2.1M (with 2.1 million features).

Other related array CGH platforms include representational oligonucleotide microarray analysis (ROMA). The platform consists of oligonucleotide probes from the human genome sequence, hybridized with "representations" of the test and reference genome,

which is prepared by cleaving the genomes with restriction enzymes, followed by differential PCR amplifications (Lucito et al. 2003). This technique was used in an early CNV discovery study (Sebat et al. 2004) as well as a few recent reports of CNV association with psychiatric diseases (Sebat et al. 2007; Walsh et al. 2008).

1.2.2.3 Choice of Reference DNA for Array Hybridization

In array CGH studies, co-hybridization of test and reference genome allows reported signal ratios to be subjective and independent of the probe concentration of each spots, or of variations within array (Carter 2007). The choice of reference, however, is inconsistent within the research community, making direct comparison of array CGH results difficult to interpret. In the early days, pools of reference DNA were used to dilute the effect of CNVs unique to individuals (Iafrate et al. 2004; Sebat et al. 2004). However, this gives imprecise signal intensity ratio especially at genomic regions that harbour complex and multi-allelic copy number changes. The current recommended practice is to perform array hybridization to a single, well-characterized reference genome (Scherer et al. 2007). In our laboratory we routinely use a well-characterized trio-offspring DNA from the HapMap collection, NA10851 (Conrad et al. 2006; McCarroll et al. 2006; Redon et al. 2006), for all array CGH studies.

1.2.3 Genotyping CNVs using Single Nucleotide Polymorphisms

The analysis of single nucleotide polymorphism is a well-established tool for genetic studies for the following reasons. First, over recent decades, a large amount of effort has been put into the discovery, validation and characterization of SNPs in the human genome (Altshuler et al. 2000; Reich et al. 2003; Hinds et al. 2005). Secondly, the HapMap Project provides a catalogue of well-characterized SNPs in four major ethnic populations (IHMC 2005), an important resource for assessing genetic variants in the context of population association studies and evolution. Thirdly, high-throughput array technologies for SNP genotyping have been developed (of particular success are commercial arrays from Affymetrix and Illumina Inc.).

1.2.3.1 Linkage Disequilibrium Based Tag-SNP Approach

One way of utilizing SNPs for CNV investigation is to "tag" common copy number changes, or CNPs, with surrounding SNPs based on linkage disequilibrium. A number of studies have investigated the linkage-disequilibrium properties of CNPs. Early studies indicated that deletion polymorphisms are generally in strong linkage disequilibrium and segregate on ancestral SNP haplotypes (Hinds et al. 2005; McCarroll et al. 2006). Later studies suggest that although some CNVs are in appreciable linkage disequilibrium with nearby markers, accurate genotypes can only be captured for a minority of CNPs tested (Redon et al. 2006). Results were far less robust for CNPs in complex regions such as segmental duplications, which may partially be due to the scarcity of SNP markers surrounding those regions (Locke et al. 2006). Other reasons such as high recombination rate in regions of CNV or high rate of spontaneous recurrence of CNVs may also be possible (Lee and Lupski 2006). In conclusion, current genome-wide technologies are limited to tag CNVs using nearby polymorphic SNP markers (Eichler et al. 2007; McCarroll and Altshuler 2007).

1.2.3.2 CNV Genotyping Using SNP Arrays

Arrays originally designed to genotype SNPs in genome-wide linkage association studies can be used to reliably estimate copy number changes. Affymetrix arrays, for example, were synthesized with 25-mer of matched and mismatched probe pairs to target each SNP under investigation. For array hybridization, a single test DNA is digested with restriction enzyme(s), which was then ligated with adaptors for universal DNA amplification. Hybridization signals of probes at each locus were then compared to those from a single or a group of references hybridized on the same array type, from which CNV calls were generated. Furthermore, CNVs (in particular deletion polymorphisms) 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. 2006).

The first SNP-array based CNV studies started off using the Affymetrix GeneChip Human 10K (Herr et al. 2005), slowly advancing to high-resolution and more robust analysis as SNP array chips and associated bioinformatics tools matured. The Affymetrix GeneChip 500K array, for instance, was used for CNV detection as a complementary platform to BAC array in the HapMap CNV study published in 2006 (Komura et al. 2006; Redon et al. 2006). An apparent advantage of such an array platform is its versatility- SNPs, CNVs and other clinically useful data such as uniparental disomy (copy number neutral LOH) (Friedman et al. 2006; Peiffer et al. 2006) can all be mined from data generated from the same chip. A major drawback of the early SNP arrays (e.g. Affymetrix 250K/500K or Illumina HumanHap300) was uneven probe spacing, with particularly sparse probes at regions near segmental duplication, and at repeat-rich centromeres and telomeres, perhaps due to problems they create against robust SNP genotyping (Carter 2007). To enhance the power of CNV detection, the most recent SNP genotyping platforms have

included non-polymorphic probes specifically selected for their genomic positions and for linear response to copy number changes. Examples of the latest platforms are Affymetrix Genome-wide Human SNP array 5.0 and 6.0, the latter with ~906,600 SNPs and ~946,000 CNV probes; and Illumina Human1M-Duo Bead Chip, with ~1.2 million markers and 1.5 kb median marker spacing, with probes designed to target known CNV regions and gaps between HapMap SNPs. These platforms have generated a new paradigm for genome-wide disease association studies, integrating both SNP and CNV assessments in the work-flow of such investigations (Korn et al. 2008).

1.2.4 Validation and Detection of CNV at Targeted Loci

Quantitative or semi-quantitative measurements of copy number variations at targeted loci are required for a number of reasons: i) as independent platforms to validate array-based CNV discoveries; ii) to convert analog signal-intensity ratios from array-based CNV detection to discrete CNV genotypes; iii) to precisely map breakpoints of CNV regions; and/or iv) to develop low-cost, reliable assays for large-scale genotyping in a large number of samples, for example in disease association studies or in clinical diagnostic settings. Traditional methods such as FISH, Southern Blot, PFGE and long range PCR serve some of these purposes, but they remain laborious, technically demanding, and low-throughput therefore unsuitable for larger scale studies. A number of complimentary techniques have evolved along with the progress of CNV discoveries.

1.2.4.1 Quantitative Fluorescent Real-time PCR (qPCR)

QPCR was first developed for quantification of RNA (Higuchi et al. 1992; Heid et al. 1996). In qPCR assays, input DNA quantity was monitored with dyes (e.g. SYBR-green) or dual-labelled probes (e.g. Taqman^{TM} probes), and fluorescence was recorded in real-time during the PCR amplification reaction. For instance, Taqman^{TM} assays draw on the 5'-3' exonuclease activity of Taq polymerase, which cleaves the dual-labeled probes attached to DNA, releasing a reporter dye (e.g. FAM) which emits a fluorescence signal during PCR amplification. The technique was first used in the context of gene copy number changes assessing the hemizygosity associated with tumor, and of the gene dosage at hereditary neuropathy with liability to pressure palsies and Charcot Marie Tooth disease (Laurendeau et al. 1999; Wilke et al. 2000). In the absence of post-PCR manipulation, qPCR has the advantage of being fast, semi-automated with accurate quantification. Furthermore, a minimal amount of input DNA is needed and therefore the method is

scalable to high-throughput screening (Hoebeeck et al. 2007), although multiplex assays are still difficult to optimize (Conrad and Hurles 2007). It is now widely used as CNV validation strategy and for clinical diagnosis (Weksberg et al. 2005; Qiao et al. 2007; Malakho et al. 2008).

1.2.4.2 Multiplex Quantitative Fluorescent Real-time PCR

More quantitative PCR methods have been developed to increase throughput and accuracy of traditional qPCR. For example multiplex ligation-dependent probe amplification (MLPA) (Schouten et al. 2002), Multiplex Amplicon Quantification (MAQ) and multiplex amplifiable probe hybridization (MAPH) (Armour et al. 2000; White et al. 2002) are routinely used for copy number quantification in research as well as diagnostic settings. These multiplex approaches allow PCR amplification to be performed using universal primer sets, with different probes specifically bound for different target amplicons generating specific fluorescent peaks for semi-quantitative detection of gene dosage. All these methods facilitate parallel screening of a large number of samples across a large number of putative CNV loci at affordable costs and high reliability.

1.2.4.3 Other Methods

Other methods for CNV dosage quantification include novel techniques based on competitive PCR. In competitive PCR, a control target of known concentration is co-amplified with an unknown test competitor sequence, and the concentration of the test sample is inferred by comparison with the control. The difference between the competitive sequences could be as small as a single nucleotide mismatch. Interspecies PCR, for example, exploits the large number of SNPs or dispersed repeat element differences between human and chimpanzee as competitive sequences for CNV dosage detection (Armour et al. 2007; Williams et al. 2008). These techniques have demonstrated robust

and inexpensive measurement of copy numbers at the 22q11 loci (Williams et al. 2008) and beta-defensin locus (Hollox et al. 2008) .

1.2.5 Genome Sequencing and CNV Detection

The availability of the human genome sequence (Lander et al. 2001; Venter et al. 2001; IHGSC 2004) was a pre-requisite for the success of CNV discoveries. Genomic insert clones (e.g. BACs) or oligonucleotides on whole-genome arrays are tiling representations of the reference genome sequence (Shendure and Ji 2008). SNP genotyping arrays, another major CNV detection platform, were developed with the International HapMap Project (IHMC 2005) playing an instrumental role. Nevertheless, both the human genome assembly and the International HapMap Project are not perfect, and the reference genome sequence tends to conceal or collapse regions of the genome most likely to harbor copy number variants, for example, regions at or near segmental duplications. Technologies for deep sequencing and comparison of more individual genomes are therefore likely to reveal more about copy number variations.

1.2.5.1 Clone End Mapping and Sequencing

In clone end mapping, a genomic sequence is fragmented and subcloned into circular clone (e.g. fosmid) vectors to create a genomic library. The ends of these genomic insert are sequenced from universal primers in the vector, and these are mapped to a reference genome. End sequence pairs that are discordant in terms of length (suggesting insertions/deletions) or orientation (suggesting inversion) are recorded (Eichler et al. 2007). In 2005, Tuzun *et al.* reported 297 structural variant events between two individual genomes based on fosmid end sequence mapping of a DNA sample NA15510 against the reference genome sequence (Tuzun et al. 2005). Two years later, Korbel *et al.* reported a study based on higher-resolution method known as paired-end mapping (PEM) on two individual genomes (NA15510 and NA18505), detected over 1000 insertions/deletions (Korbel et al. 2007). More recently, clone-end sequence-pair (ESP)

maps of eight human genomes were released, revealing even more novel CNVs and refining previously known structural variations (Kidd et al. 2008).

1.2.5.2 Novel Sequencing Technology and Genome Comparison

Computational comparisons between published genome assemblies have provided us more insights on copy number variations (Khaja et al. 2006; Levy et al. 2007). With the advancement in technologies, a number of novel sequencing methods were developed on top of the traditional Sanger sequencing technique (Sanger 1981). In particular, "second generation" sequencing methods based on cyclic-array sequencing have already been realized as commercial products, for instance in 454 sequencing (Roche Applied Science), Solexa technology (Illumina Inc) and the SOLiD platform (Applied Biosystems) (Shendure and Ji 2008). These massively parallel DNA sequencing technologies will allow complete genome sequencing and re-sequencing of more individuals at an affordable cost (Service 2006; Eichler et al. 2007), unveiling the true extent of CNVs in the human genome.

Figure 1.4 summarizes the sensitivity and the throughput of various CNV detection techniques we have discussed so far.

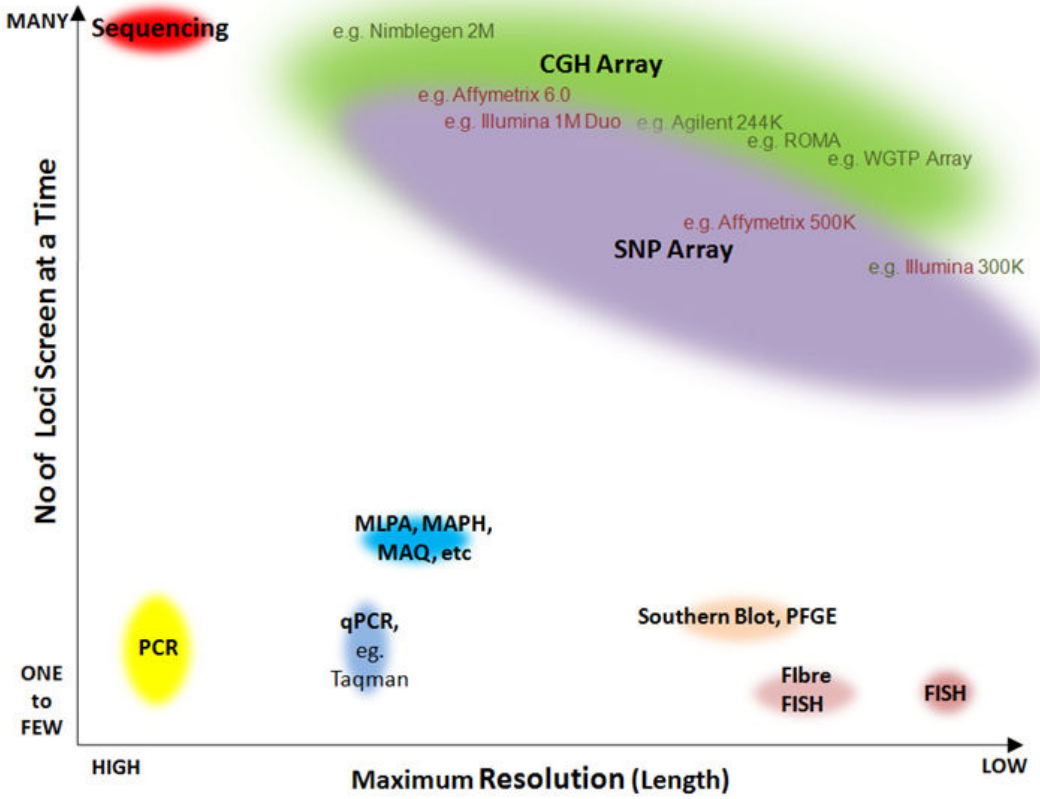

Figure 1.4 Sensitivity and throughput of various CNV detection techniques. x-axis: maximum resolution of the technique (indicate sensitivity); y-axis: number of loci that can be screened at a time (indicate throughput and scalability).
PCR: polymerase chain reaction; qPCR: quantitative PCR; MLPA: multiplex ligation-dependent probe amplification; MAPH: multiplex amplifiable probe hybridization; MAQ: Multiplex Amplicon Quantification; PFGE: pulse field gel electrophoresis; FISH: fluorescence in-situ hybridization; SNP Array: single nucleotide polymorphism genotyping array; CGH Array: Comparative Genome Hybridization array

1.3 Mechanisms of Copy Number Variation Generation

Understanding the underlying mutational processes for CNV will provide information on how this type of genetic variant emerges, and should yield important insights into the genomic distribution, evolution and frequency of CNVs in the population. There are two major mechanisms for CNV generation, namely non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ). Both are cellular mechanisms intended for maintaining the integrity of DNA by repairing DNA double stranded breaks (DSBs).

1.3.1 Non-allelic Homologous Recombination

The evidence of NAHR being a major mechanism of CNV generation came from observations that copy number changes frequently fall in close proximity to repeat sequences, for instance segmental duplications (Sharp et al. 2005) or Alu repeats (Lupski 2006; de Smith et al. 2008). Early examples came from genomic disorders, a class of diseases resulting from alteration of dosage-sensitive genes, usually by large-scale microdeletion or microduplication of the genome (Stankiewicz et al. 2003; Lee and Lupski 2006). These genomic disorders, such as Velocardiofacial Syndrome, Williams Buren Syndrome and Charcot Marie Tooth Disease, frequently have breakpoints clustering at highly homologous segmental duplicons or low copy repeats (LCRs). It was suggested that the LCRs act as substrates for NAHR, mediating the rearrangements (Stankiewicz and Lupski 2002). It was also hypothesized that recurrent CNVs were most frequently results of NAHR (Lee and Lupski 2006).

Copy number changes mediated by NAHR may occur at homologous sequences either on the same or the non-homologous chromosome (inter-chromosomal or

intrachromosomal NAHR). Apart from LCRs, more divergent repeats such as SINEs (eg. Alu), LINEs and human endogenous retroviruses (HERVs) can all act as substrates for NAHR (Roy et al. 2000; Fredman et al. 2004; Hurles 2005), although such events are less likely to be recurrent (Cooper et al. 2007).

During meiotic NAHR, misalignments or unequal cross-over of the homologous sequences generate germline rearrangements such as duplications, deletions and inversions (Figure 1.5), and be detected as structural variations using techniques as discussed in section 1.2. Somatic rearrangement may also be generated in a similar manner (Cook and Scherer 2008), although this class of CNV is less well-studied (Piotrowski et al. 2008).

Depending on the distribution and participation of the homologous sequences, copy number changes generated by NAHR could be simple deletion or duplication (Figure 1.5a & b); or more complex rearrangement, such as tandemly duplicated arrays (Hurles 2005), as in the case of the opsin locus for red-green color vision (Neitz and Neitz 1995) (Figure 1.5d); or other complicated structural variations involving multiple homologous duplicons.

1.3.2 Non-homologous End Joining

NHEJ is an alternative mechanism for repairing DSBs in cells. When random DSB occur at regions with no extensive homologous sequences to act as the repair template for NAHR, the broken ends of the DSB may be rejoined by nucleases removing the nucleotides, and the Pol X family of DNA polymerases filling in the missing nucleotides. NHEJ is an error-prone repair mechanism, with frequent gains and losses of nucleotides

at the junctions (Lieber et al. 2003). Much less is known about NHEJ compared to NAHR (Lee et al. 2007). The mechanism is suggested to be mediated by microhomologies (< 25 bp homology) (Lieber et al. 2003). It is also known that NHEJ is more prevalent in unstable (or fragile) regions of the genome, for example in the subtelomeric regions (Nguyen et al. 2006; Kim et al. 2008b). The mechanism has been implicated in a number of genomic disorders (Inoue et al. 2002; Shaw and Lupski 2005)

1.3.3 Other Mechanisms

Recently, a few novel mechanisms of CNV generation have been suggested, which may generate interest for further characterization. One example is a mechanism involving DNA stalling and template switching during mitotic replication, as observed in Pelizaeus-Merzbacher disease, an X-linked genomic disorder (Lee et al. 2007). Another mechanism involves both homologous and non-homologous recombination, as seen in the methyl-CpG-binding protein 2 (*MECP2*) duplication on chromosome Xq28 (Bauters et al. 2008).

1.3.4 Insights from Breakpoint Mapping

A recent paper systematically investigating the breakpoints of CNVs (Kim et al. 2008b) provided new insights into CNV generation mechanisms. In the study, 534 (fosmid-paired end sequenced) CNV breakpoints were subjected to bioinformatics analysis to look for co-localization with various classes of repeats. In particular, they reported that 28% of CNVs were colocalized with segmental duplication, a result consistent with earlier studies (Sharp et al. 2005; Redon et al. 2006; Cooper et al. 2007) albeit at a lower rate. In terms of signatures with other NAHR substrates, CNV breakpoints were not shown to be significantly associated with Alu elements, contrary to a previous study

showing significant association (Cooper et al. 2007). On the other hand, CNV breakpoints were shown to co-localize with LINE elements (>20% of breakpoints) and microsatellite repeats (3% of breakpoints) (Kim et al. 2008b).

Moreover, Kim *et al.* showed evidence for NHEJ as a major mechanism of CNV generation. 40% of the breakpoints showed microhomologies, implicating NHEJ involvement. Another 14% demonstrates micro-insertions (base pairs of gains/losses) at breakpoint junctions, which is also suggestive of NHEJ mediation (Kim et al. 2008b).

In summary, the non-random distribution of CNV, its co-localization with segmental duplications and its clustering at telomeric and centromeric regions, probably reflect the underlying mechanism of CNV generation (Stankiewicz and Lupski 2002; Hurles 2005; Lupski 2007). Whether such mutational bias plays a particular role in human evolution and diseases remain to be investigated.

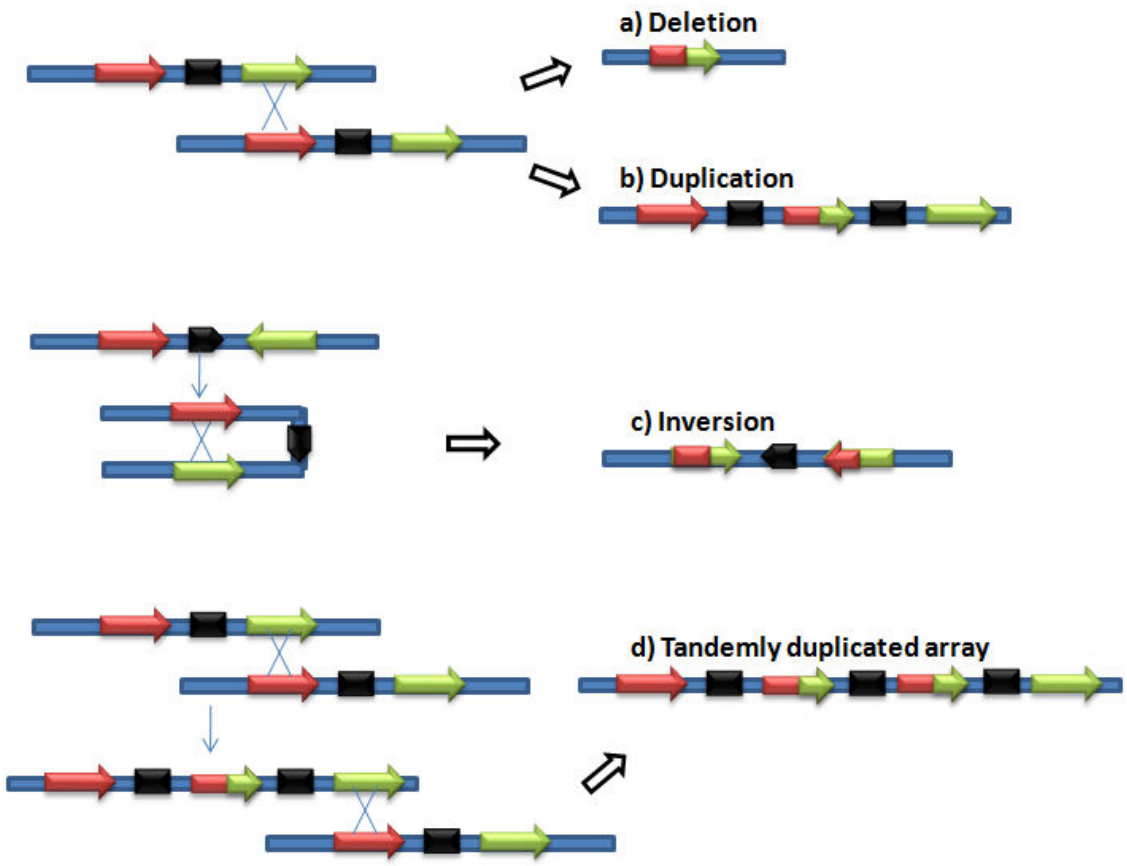

Figure 1.5 Non-allelic homologous recombination by low copy repeat. Examples of mechanisms generating different structural variants: **a)** deletion; **b)** duplication; **c)** inversion; **d)** tandemly duplicated array

1.4 Biological Impact of Copy Number Variation

1.4.1 Phenotypic Effect of CNVs

Genetic variants of all sizes and types could confer phenotypic effects on individuals via two broad mechanisms: a) perturbing gene expression, by acting on transcription, splicing, or translation and stability; or b) modify the function of a protein, for example by altering the protein structure. Based on these two mechanisms, below is a non-exhaustive list of how copy number variations could confer functional consequences (Lupski and Stankiewicz 2005; Feuk et al. 2006b; Eichler et al. 2007; Shelling and Ferguson 2007):

- Deletion and duplication may act on dosage-sensitive genes, with gene copy number changes resulting in alteration of gene expression;
- Rearrangement may alter the regulatory elements for a nearby gene, with possible long-range effect (Kleinjan and van Heyningen 2005) which changes gene expression at a distance;
- CNVs may disrupt a gene, e.g. by interrupting protein coding sequences, causing functional loss or modification;
- Such rearrangements may also generate novel fusion products, for example by fusing different protein domains, by inserting protein coding sequences to the proximity of regulatory elements, or by deleting exons generating novel splice variants;
- The change of expression of a gene involved in CNV may upset the stoichiometry, or the balance, of a macromolecular complex or network, translating the CNV effect to other genes and proteins (Korbel et al. 2008);
- Deletions may unmask mutations or functional SNPs in the remaining allele

- CNVs may affect microRNAs in the genome, which in turn may lead to alterations in gene expression

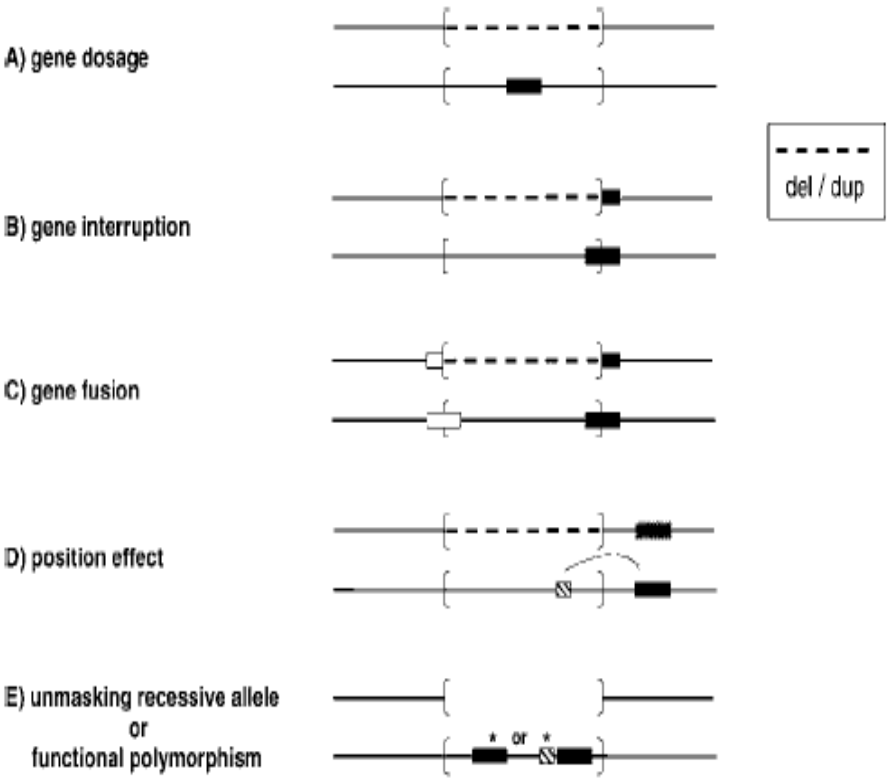

Figure 1.6 phenotypic effects of CNVs (diagram adapted from Lupski and Stankiewicz, 2005). Examples of CNVs affecting phenotypes through: **a)** gene dosage; **b)** gene interruption; **c)** gene fusion; **d)** position effect; **e)** unmasking recessive allele

1.4.1.1 CNV and Expression

The effect of CNV on gene dosage is most prominent in genomic disorders. In the example of the Charcot Marie Tooth locus (CTM-1A) on chromosome 17, for example, haploinsufficiency of the dosage-sensitive gene peripheral myelin protein 22 (*PMP22*) causes Hereditary Neuropathy with Liability to Pressure Palsies (HNPP), whereas reciprocal duplication results in trisomic expression of *PMP22*, leading to another neuropathy, the Charcot Marie Tooth Disease (Chance and Fischbeck 1994).

Compared to simple deletions or duplications, multi-allelic CNVs may give subtler yet more variable effects on the range of gene expression by altering dosage levels for individuals in the population. In the case of the human salivary amylase gene, *AMYI*, the diploid copy number of this starch hydrolysis gene varies from 2 to >15 (Iafrate et al. 2004; Redon et al. 2006), with populations having higher starch diet generally found to have higher *AMY1* copy number and vice versa (Perry et al. 2007). It was demonstrated that this CNV locus affects gene expression at both the transcriptional and translational levels, with copy numbers shown to correlate with mRNA and *AMY1* protein level in human saliva samples (Perry et al. 2007).

As a genome-wide survey of the impact of CNV on expression patterns, Stranger *et al.* examined mRNA levels in lymphoblastoid cell lines from 210 unrelated HapMap individuals from 4 different ethnic groups, in combination with CNV data generated from the same HapMap samples (Redon et al. 2006; Stranger et al. 2007). 8.75% of variation in expression levels of 972 genes could be attributed to copy number changes (Stranger et al. 2007). In some cases, the effects of CNVs on gene expression were exhibited across all 4 ethnic groups, while in other cases they remain population-specific. Distant regulation or positional effect were also reported, with > 50% of the expression probes

associated with CNVs being away from the CGH clone encompassing the CNV, some of them as far as 2 Mb apart (Stranger et al. 2007). One example detected was the changes in mRNA expression of UDP-glucuronosyltransferase 2B17 (*UGT2B17*), which vary with copy number (McCarroll et al. 2006; Stranger et al. 2007). Variability of the gene was previously shown to associate with testosterone level, male insulin sensitivity, fat mass and prostate cancer (Xue et al. 2008).

A simple, direct model for the effect of CNV on gene expression suggests proportional relationship between the two, with expression increasing as gene dosage increases. This probably explains the effects of many CNV loci, for instance the nearly linear relationship of protein expression with salivary amylase gene copy number (Perry et al. 2007). Negative relationship has also been suggested, in which higher copy number decreases expression level. For instance a small duplication downstream of the proteolipid protein gene *PLP1* was proposed to cause gene silencing by positional effect, thus lowering gene expression (Lee et al. 2006). However, considering the intricate and convoluted human biological networks and pathways, it is possible that some CNVs confer more complex effects on gene expression. Recently, Mileyko et al. examined this by modelling the effect of four common gene regulatory motifs: “positive feedback”, “bistable feedback”, “toggle switch” and “repressilator” (Mileyko et al. 2008). They proposed that a small change in gene copy number within these motifs could completely switch a biological network from one alternative steady state to another, and this change in equilibrium could result in large and unanticipated changes in gene expression.

Nonetheless, some CNVs are phenotypic benign variants (as marked by the numerous CNVs in apparently normal populations) (Cooper et al. 2007). This suggests that changes in copy number may not necessarily translate into gene expression,

phenotypes or human traits. In many cases, CNVs are neither disease-causing variants nor variants with adverse phenotype. These apparently "neutral" CNVs may be explained by i) the genomic location of CNV (e.g. being located in gene-poor regions); ii) the insensitivity of the CNV genes to dosage; or iii) the effect of dosage compensation (Birchler et al. 2005), a mechanism which would balance out gene expression changes resulting from CNVs. In conclusion, the severity of phenotypes caused by a given CNV via gene expression is a combination of the dosage sensitivities of genes affected by the CNV, plus their interaction with other genetic and environmental variants.

1.4.1.2 CNV and Gene Disruption

Since not all genes are dosage sensitive, an alternative mechanism that CNV might confer phenotypes and diseases is by disrupting genes and modifying protein function. In a recent genome-wide CNV screen in schizophrenia patients, Walsh *et al*. suggests that CNV breakpoints in the case cohort disrupt many more genes that are involved in neurodevelopmental pathways than in the control cohort (Walsh et al. 2008). One example was a ~400 kb deletion disrupting the *ERBB4* gene, which encodes a type I transmembrane tyrosine kinase receptor. Using 3' rapid amplification of cDNA ends (RACE), the authors cloned the mutant transcript, which was found to lack the intracellular kinase domain of the receptor. This mutant allele could potentially have a dominant negative effect, with evidence from a previous mouse model. CNVs that disrupt genes can potentially have major phenotypic effects, forming a class of variants with high penetrating effect (Walsh et al. 2008). With more CNV breakpoints being mapped by high-resolution techniques and the resulting allelic sequences and structures characterized, we will gain a more complete understanding of the effects of copy number variants.

1.4.2 Phenotypic Variations and Evolution

1.4.2.1 CNV and Human Traits

Over the years, through a combination of genome-wide association or single locus approaches, single nucleotide polymorphisms have been shown to influence a diverse range of normal human traits, from the ability in speech and language (e.g. *FOXP2*) (Lai et al. 2001), to variations in height (Gudbjartsson et al. 2008; Lettre et al. 2008), even to sprinting ability in athletic performance (e.g. *ACTN3*) (North et al. 1999; Yang et al. 2003). Human copy number variations were demonstrated to encompass more base pairs of changes between individuals than all SNPs combined together (Redon et al. 2006). With growing numbers of CNV discovery studies, a major question arising is what impact does this ubiquitous class of genetic variant have on normal human phenotypic trait (Nguyen et al. 2006). A clear example comes from the aforementioned *AMY1* locus, in which copy number was shown to correlate with the amount of starch-digesting amylase protein in human saliva (Perry et al. 2007). The geographical variation of copy number at this locus was proposed to be shaped by the evolution of diet in high-starch eating populations (such as Japanese) compared to low-starch eating populations (such as Biaka) (Perry et al. 2007).

Another argument comes from the X-linked opsin gene locus dictating red-green color vision (Deeb 2005). The locus comprises red (L-opsin) and green (M-opsin) photopigment genes and a “locus control region” regulating gene expression. Variable copy numbers were detected in the L-opsin gene (0-4 haploid copy), M-opsin gene (0-7 haploid copy) and the locus control region unit (0 or 1 haploid copy). It is a combination of the copy number of these three units, sequence variation and unequal cross-over between homologous L-opsin and M-opsin that determines the phenotype of color vision, even to the extreme of color blindness (Cooper et al. 2007).

In addition, copy number at the cytochrome P450 gene *CYP22A6* has been implicated in altered nicotine metabolism and smoking behaviour (Rao et al. 2000), and a deletion polymorphism at UDP-glucuronosyltransferase 2B17 (UGT2B17) was demonstrated to correlate with level of testosterone excretion (Jakobsson et al. 2006). The *UGT2B17* CNV was also shown to have major geographic variation, with evidence of positive selection in the East Asia population (Xue et al. 2008).

1.4.2.2 Positive and Negative Selections on CNV

CNVs at specific loci (e.g. AMY1) have been suggested as being involved in human adaptive evolution (Perry et al. 2007). On a genome-wide level, several studies have proposed that natural selection processes may have influenced and shaped CNV genomic distribution and gene content (Nguyen et al. 2006; Cooper et al. 2007; Nguyen et al. 2008).

Duplications have long been suggested as substrates on which selection processes can act. A duplicated gene and its regulatory elements could subsequently be modified for new functions, facilitating a species' diversification and evolution (Lynch and Conery 2000; Dermitzakis and Clark 2001; Korbel et al. 2008). Some of these innovations may have been favoured in positive selection processes. In contrast, negative selection could also have shaped the characteristics of CNVs, and can easily be envisaged in deletion scenarios where deleterious loss-of-function alleles could be generated. In support of this, we see an enrichment of recent common CNVs (those that do not overlap with SDs) in gene-poor regions (Cooper et al. 2007). Array CGH studies further suggested that deletions in the genome are more biased away from the morbid OMIM (Online Mendelian Inheritance in Man) genes than duplications are (Redon et al. 2006).

One could hypothesize that the current distribution of CNVs in the human genome has been driven by past positive selection and adaptive evolution. Alternatively, from a more neutralist's point of view, CNVs could be under reduced purifying selection (Nguyen et al. 2008). This model suggests that variants are enriched in nonessential genes, with fixation of slightly deleterious substitutions, while the "neutral" CNVs were under genetic drift.

Examining the functional categories of CNV gene content may provide a glimpse into the underlying selection processes. A number of studies have suggested that CNVs significantly co-localize with genes that have environmentally responsive functions, such as sensory perception and immunity (Nguyen et al. 2006; Redon et al. 2006). Such enrichment could reflect these functional classes being less deleterious, therefore remaining in the genome by reduced purifying selection (Nguyen et al. 2008). On the other hand, proteins within these biological categories were well-documented in mammalian adaptive evolutions (Cooper et al. 2007), and may have acted as substrates for natural selection. To this end, CNVs affecting environmental response genes have recently been demonstrated having involvement in a number of infectious and autoimmune diseases, such as HIV-1 susceptibility, glomerulonephritis and Crohn's Disease (Gonzalez et al. 2005; Aitman et al. 2006; Fellermann et al. 2006).

In summary, natural selection may differentially act on a spectrum of CNVs with variable phenotypic consequences, from beneficial, neutral, benign to deleterious (e.g. disease-causing CNVs). Further characterization of CNVs in a range of phenotypic and disease contexts will expand on our rudimentary knowledge about this source of genetic variants.

1.4.3 CNV and Disease

One of the major challenges in biology is to unravel mechanisms underlying human diseases. Studies in copy number variations have indicated its role in a number of sporadic diseases, Mendelian disorders as well as some multifactorial and complex disease phenotypes. The earliest evidence of such came from a group of diseases known as 'genomic disorders'.

1.4.3.1 Genome Disorder

Genomic Disorders are defined as diseases caused by genomic rearrangements affecting dosage sensitive genes (Stankiewicz and Lupski 2002; Lupski 2007). They are classic models of DNA sequences being altered as copy number variations resulting in adverse human phenotypes (Lupski and Stankiewicz 2005). Well-known examples are Williams Beuren Syndrome (WBS) [del(7)(q11.23q11.23)], Velocardiofacial/ DiGeorge Syndrome (VCFS/DGS) [del(22)(q11.2q11.2)], Prader–Willi (PWS) [pat del(15)(q11.2q13)] or Angelman syndrome (AS) [mat del(15)(q11.2q13)], and Charcot Marie Tooth Disease (CMT) with reciprocal Hereditary Neuropathy with Liability to Pressure Palsies (HNPP) (Lupski 2006). These classical examples are typically large microdeletions and duplications (of several Mb) (Inoue and Lupski 2003), as they were discovered via traditional cytogenetic technique (e.g. karyotyping). Recent advance in CNV detection techniques have revealed a growing number of novel genomic disorders, such as the 15q13.3 microdeletion syndrome (~1.5 Mb) associated with mental retardation and seizures (Sharp et al. 2008), and the 17q21.31 microdeletion syndrome (~500 kb) associated with mental retardation and developmental delay (Koolen et al. 2006; Sharp et al. 2006; Shaw-Smith et al. 2006).

Genomic disorders have provided us with important insights regarding CNVs and human diseases: First, these syndromes have revealed the plasticity of the human genome: alterations of megabases of DNA sequences affecting genes with important biological functions can apparently be tolerated in human. In some cases, such as Charcot Marie Tooth Disease, phenotypes of the disorder could be narrowed down to a single gene (*PMP22*) within the large rearrangement (Lupski et al. 1991; Patel et al. 1992). Furthermore, several genomic disorders are the result of reciprocal deletions/duplications of the same loci, demonstrating that varying dosage of the same gene(s) could result in diverse phenotypes (Lupski and Stankiewicz 2005).

Secondly, genomic disorders provide us insights into the underlying mutational mechanism of copy number changes. Some cases of genomic disorders may cluster to breakpoints of LCRs or segmental duplications (Lupski and Stankiewicz 2005), suggesting NAHR as one of the disease mechanisms. Realizing genomic disorders may reflect the underlying genomic architecture, several research groups have interrogated specifically the SD-rich regions of the genome, facilitating discovery of further genomic disorders and diseases loci (Sharp et al. 2006). Other higher-order genomic architecture, such as nearby inversion polymorphism, may also predispose the disease locus to rearrangement. This has been demonstrated in 17q12.13 microdeletion syndrome, where a parental H2 inversion is necessary for disease transmission (Koolen et al. 2006; Shaw-Smith et al. 2006; Koolen et al. 2008), and has been suggested in Sotos Syndrome (Kurotaki et al. 2005), another genomic disorder.

Furthermore, these disorders demonstrated the interaction of CNVs with other genetic factors in diseases. PWS and AS, for example, are derived from the paternal and maternal microdeletions of the same genomic locus, resulting in distinct disease

phenotypes (Lupski 2006). This reveals an interconnection between CNVs and epigenetic factors (imprinting mechanism).

Finally, the majority of genomic disorders involve nervous system disorders and neuropathies, providing us invaluable insights into diseases of the brain. For example, these mental retardation syndromes were found to be comorbid with various psychiatric symptoms (eg. PW/AS with autism (Veltman et al. 2005), VCFS/DGS with schizophrenia (Murphy and Owen 2001)), suggesting genes altered in these loci may play multiple roles resulting in overlapping behavioral and cognitive phenotypes, or vice versa the diverse phenotypes could be the consequences of the perturbation of a common biological pathway via the alteration of common gene(s).

1.4.3.2 Rare CNVs in Mendelian Disease Traits

In contrast to genomic disorders, which in many cases are sporadic diseases resulting from *de novo* CNVs (Lupski 2007), the involvement of CNVs in diseases can also be illustrated with rare inherited copy number changes. These inherited CNVs act in the same way as Mendelian mutations do in diseases (Estivill and Armengol 2007). Below are examples originated from two neurodegenerative disorders.

Triplication or duplication of the alpha-synuclein (*SNCA*) locus was shown to cause Parkison's diseases (Singleton et al. 2003; Chartier-Harlin et al. 2004). Singleton *et al.* first demonstrated triplication of the SNCA to hereditary early-onset parkinsonism with dementia. Compared to the triplication cases, the duplication kindred identified by Chartier-Harlin *et al.* seemed to show milder disease progression closely resembling idiopathic Parkinson's, with late onset and no cognitive decline or dementia, suggestive of a dosage effect in disease progression. Alpha-synuclein protein is a major component of Lewy bodies, the "pathological hallmark of Parkinson's Disease" (Singleton et al.

2003). Examination of brain tissue showed increased copy number correlates with the level of soluble alpha-synucelin, and an even more prominent increase in the form of aggregated, insoluble deposition (Miller et al. 2004) .

In the second example, duplication of the amyloid precursor protein (*APP)* on chromosome 21 was demonstrated to cause autosomal dominant form of early-onset Alzheimer disease (ADEOAD) with cerebral amyloid angiopathy (CAA) (Rovelet-Lecrux et al. 2006). From their study, Rovelet-Lecrux *et al.* reported that 5 out of 65 families of ADEOAD with CAA showed duplication at the *APP* locus, estimating the frequency of this duplication to be 8% (5 of 65) in the disease. This is also consistent with previous reports suggesting that *APP* duplication coincide with patients with Alzheimer disease (Delabar et al. 1987; Sleegers et al. 2006). Alterations in the *APP* gene probably lead to accumulation of amyloid precursor protein which results in neurodegeneration (Rovelet-Lecrux et al. 2006), highlighting *APP* as a dosage sensitive gene. The copy number change at the *APP* locus was also put into context of SNP mutations in the same gene, where the CNV is estimated to correspond to half of SNP missense mutations leading to disease (Rovelet-Lecrux et al. 2006).

1.4.3.3 Common CNVs in Multifactorial or Complex Disease

Common diseases involving multiple genetic and environmental factors are in principle more susceptible to copy number variations, which could alter gene dosage without disrupting protein functions and therefore result in high phenotypic plasticity (McCarroll and Altshuler 2007). Furthermore, many multifactorial diseases, such as a number of complex psychiatric disorders, were shown to be heritable with a substantial genetic contribution, whilst the underlying genetic factors remain largely elusive (Burmeister et al.

2008). The involvement of CNVs in complex diseases is therefore an area under intense investigation.

There is now considerable evidence of complex diseases under the influence of copy number variants. These include a number of infectious and autoimmune diseases, cancer, as well as psychiatric diseases (discuss in section 1.6). Below I discuss three examples in detail, namely the human chemokine gene *CCL3L1* in HIV-1 susceptibility and progression (Gonzalez et al. 2005); the Fc-receptor gene *FGCR3B* in lupus and related autoimmune diseases (Aitman et al. 2006; Fanciulli et al. 2007), and the β-defensin gene cluster (including *DEFBA3 & DEFB2*) in Psorasis and Colonic Crohn's Disease (Fellermann et al. 2006; Hollox et al. 2008).

In 2005, Gonzalez *et al.* reported probably the first evidence of copy number variants stably transmitted to contribute to complex diseases (Gonzalex, 2005). The authors studied the distribution of chemokine gene-containing segmental duplications on chromosome 17, which include genes *CCL3L1*, *CCL4L1* and *TBC1D3*, in 1064 humans from 57 populations and 83 chimpanzees. The copy number of the SD has previously been demonstrated to regulate the production of human chemokine *CCL3L1* (Townson et al. 2002). Gonzalez *et al.* described a geographic variation of the *CCL3L1*-containing SD, with African populations possessed a significantly higher copy number than non-Africans. Furthermore, within populations, a lower gene dose relative to the average copy number in the population was demonstrated to confer HIV-1 susceptibility, as well as the increased risk of progression to AIDS or death. *CCL3L1* encodes a ligand for the receptor CCR5, which in turn interacts with the HIV-1 virus glycoprotein (gp) 120 (Gonzalez et al. 2005).

For the role of CNV in autoimmune disease, Aitman *et al.* demonstrated that copy number changes of the Fc receptor *FGCR3B* and its ortholog in rat *Fcgr3-rs* could play a role in immunologically mediated glomerulonephritis (Aitman et al. 2006). In the rat, a 226 bp deletion of the *Fcgr3-rs* allele in the Wistar Kyoto rat strain was shown to increase susceptibility to crescentic glomerulonephritis. The same study also reported a lower copy number of *FGCR3B* in human patients with lupus glomerulonephritis or lupus erythematosus. In a subsequent study *FGCR3B* was reported to contribute to another two systemic autoimmune diseases: microscopic polyangiitis and Wegener's granulomatosis (Fanciulli et al. 2007). In the immune system, Fc receptors play important roles in the activation and modulation of immune responses. Furthermore, in a separate study, low copy number of the complement component C4, a different effector protein in the immune system, was also shown to involve in systemic lupus erythematosus (Yang et al. 2007).

Research also suggested that copy number of the beta-defensin segmental duplication on chromosome 8 is involved in infectious and inflammatory diseases. The segmental duplication containing *DEFB4* (encoding the protein human beta-defensin 2, hBD-2) and a number of other DEFB genes at the 8p23 β-defensin locus shows variable copy number from 2-12 copies per diploid genome (Hollox et al. 2008). Lower copy number was suggested to predispose to Crohn's disease of the colon (Fellermann et al. 2006), while higher copy number was associated with the risk to psoriasis, a skin inflammatory disease (Hollox et al. 2008). β-defensins are small antimicrobial peptides which play a proinflammatory role (Hollox et al. 2008).

The above studies provided important insights regarding the role of CNVs in complex disease genetics: first, in these complex diseases, copy number variations, and

frequently multi-allelic CNVs (ranging from 2 to >10, e.g. in *CCL3L1* and *DEFB4)*, could play a role in disease susceptibility; secondly, in particular diseases (e.g. *CCL3L1* in HIV-1 susceptibility) it was the copy number in relation to the population average, rather than the copy number *per se*, that confer disease susceptibility (Eichler et al. 2007). This suggests the context of population and ethnicity matters, and that other genetic or environmental factors could act in conjunction with CNVs to confer disease susceptibility; thirdly, the same CNV loci were involved in multiple complex diseases, indicating common disease mechanisms and pathways (e.g. *FGCR3B* in multiple autoimmune diseases); and finally, such a role of multi-allelic CNVs in complex diseases may not be easily captured even in large linkage analysis, as modelled bioinformatically by Hollox *et al.* in the case of psoriasis (Hollox et al. 2008).

In summary, it is now evident that copy number variations in genes could be direct risk factors for a number of human diseases. This list of CNV disease loci is likely to expand tremendously, as results from further genome-wide association studies or analysis of family pedigrees provide systematic analysis of the role of copy number variations.

1.5 Schizophrenia

1.5.1 The Concept of Schizophrenia

In 1896, the German physician Emil Kraepelin first described schizophrenia as a mental illness with a discrete disease entity, which he referred to as “dementia praecox” (dementia of early life) (Kraepelin 1919). The term “schizophrenia”, was later introduced in 1911 by the Swiss psychiatrist Eugen Bleuler (Bleuler 1911). The word derived from its Greek roots “schizo” (split) and “phrene” (mind). It was a refinement to “dementia praecox” (Kahn and Pokorny 1964), since it was recognized that the disease did not necessarily include a dementia process, and could sometimes occur late as well as early in life .

Schizophrenia (OMIM181500) is a debilitating psychiatric illness with a prevalence of 1% worldwide (Jablensky et al. 1992). The disease is characterized by psychotic symptoms such as delusion, hallucination, and disorganized thinking, together with cognitive and social impairment (Andreasen 1995). Schizophrenia usually has its onset in the second or third decade of the patient’s life (Hafner et al. 1998; Hafner and an der Heiden 1999), and males were described to have earlier onset than females (Stromgren 1987). Once the psychotic symptoms emerge, the mental illness could persist for a lifetime with recurrent patterns, leading to severe impairment in social and occupational functioning, as well as massive societal costs and consequences. In particular, mortality rate is elevated among schizophrenia patients (Harris and Barraclough 1998), partially due to a high suicidal rate (Fenton et al. 1997). The society cost of the illness for one year in United States alone was estimated to be in excess of 62.7 billion USD (Wu et al. 2005).

The first treatments for the illness, discovered in the 1930s, were based on electroconvulsive shock (ECS) therapy (by Ugo Cerletti) (Cerletti 1954), insulin coma

therapy (by Manfred Sakel) (Sakel 1954) and prefrontal lobotomy (Jenkins et al. 1954). The latter two were later abandoned due to low efficacy. Drug treatments were developed from the 1950s, and were broadly defined into conventional neuroleptics (which in greek means "to clasp the neuron") (e.g. chlorpromazine); and the atypical antipsychotics (e.g. clozapine). Side effects from these drugs are common, including extra pyramidal symptoms (Kandel 2000) such as Parkinsonism, dystonia and tardive dyskinesia[1] (for conventional antipsychotics), and gain of weight and diabetes (for atypical antipsychotics). Current treatments mainly comprise a combination of medication, psychotherapy and social adjustment (Bustillo et al. 2001). ECS, one of the most controversial procedures in modern psychiatry, is still being used to attenuate symptoms of schizophrenia (Brindley 2008).

[1] Patients with parkinsonism fail to walk normally and usually develop tremor; Patients with dystonia have painful muscle spasms of the head, tongue, or neck; Patients with tardive dyskinesia, a chronic EPS, show features of slow, rhythmic and automatic movements.

1.5.2 Phenotypes and Diagnosis

1.5.2.1 Positive and Negative Symptoms

Both Kraepelin and Bleuler recognized that schizophrenia symptoms tend to cluster into distinct categories, and Bleuler was the first to describe schizophrenia symptoms as "positive" or "negative" (Sass 1989). "Positive symptoms" are symptoms of thoughts, emotion or perceptions that are beyond normal experience. They include hallucinations (false sensory perception, typically auditory but can also be visual, olfactory, tactile etc.), delusions (false beliefs based on incorrect inference of reality), as well as thought and speech disorganization. In contrast, "negative symptoms" are thoughts, emotion or perceptions common in normal people but absent in patients. They include social withdrawal, blunted effect (lack of emotional reactivity), alogia (poverty of speech), and avolition (lack of motivation) (Andreasen 1995; Gelder 1996). Another core feature of schizophrenia is cognitive deficits, specifically the impairment of memory, attention and executive function (Kandel 2000).

1.5.2.2 Endophenotypes

In addition to the aforementioned clinical phenotypes, the concept of endophenotype is also of relevance to the understanding of schizophrenia. Endophenotypes, as defined by Gottesman and Gould, are "measurable components unseen by the unaided eye along the pathway between disease and distal genotype" (Gottesman and Gould 2003). These phenotypes are quantitative and heritable, and in contrast to clinical phenotypes, they usually require the use of special processes or instruments to detect or measure, as they may not be readily observable (Owen et al. 2005a). In schizophrenia, examples of endophenotypes include auditory and sensory gating deficit (e.g. P50 prepulse inhibition, investigating acoustic startle reflex), oculomotor functioning (e.g. antisaccade

performance, concerning the rapid redirection of gaze to locations of interest), and cognitive phenotypes such as working memory and verbal memory (Braff et al. 2008).

1.5.2.3 Course of Illness

Schizophrenia is a chronic mental illness. The course of illness can be recognized as three phases (American Psychiatric Association. 1994): (a) the prodromal phase, usually occur in teenage years preceding onset, when patients exhibit mild phenotypes such as social isolation and movement disorder; (b) the acute phase characterized mainly by psychotic symptoms such as delusion and hallucination; and (c) the residual phase, similar to prodromal phase with less severe symptoms than during actue phase, sometimes negative features such as blunted effect are common. Over the duration of disease, patients typically enter numerous acute phases of psychotic episodes, intermittent with numerous residual phases with variable degree of remission. Scales such as the Brief Psychiatric Rating Scale, first published by Overall and Gorham in 1960s, have been used by clinicians to estimate the course of illness (Leucht et al. 2005).

1.5.2.4 Standardised Diagnostic Methods

Current diagnostic methods for schizophrenia evolved from the work of Kraepelin as well as the descriptions of first rank symptoms by Kurt Schneider in 1959 (Bertelsen 2002). The most influential diagnostic guidelines are DSM-IV (Diagnostic and Statistic Manual of Mental Disorder- 4th Edition) (American Psychiatric Association. 1994); and ICD-10 (Classification of Mental and Behavioural Disorders) from the World Health Organization (WHO 1992). The two diagnostic methods are partially concordant (Bertelsen 2002), and a comparison is presented in Table 1.1.

Table 1.1 Two most influential diagnostic guidelines for schizophrenia: the DSM-IV and ICD-10. *(table adapted from Gelder et al., 1996)*

DSM-IV	ICD-10
At least two of the symptoms in (a) must be present for a minimum of one month. Additionally, (b) must be present for six months, and criteria (c)-(e) must be fulfilled	Minimum of one clear symptoms, or at least two if less clear cut, as listed in (a)-(d), or at least two of the symptoms in (e)-(i) should be present for one month of more
(a) delusion, hallucinations, bizarre behavior and negative symptoms	**(a)** thought echo, thought insertion or withdrawal, and thought broadcasting
(b) occupational or social dysfunction	**(b)** delusion of control or passivity
(c) schizoaffective or mood disorder exclusion	**(c)** hallucinatory voices
(d) disturbance must not be due to medication or drug abuse	**(d)** persistent delusions
(e) if a patient has a pervasive development disorder, prominent delusions or hallucinations must be present for one month	**(e)** significant and consistent change of personal behavior
	(f) negative symptoms that are not due to depression or neuroleptic treatment
	(g) persistent hallucinations accompanied by deletions or by persistent over-valued ideas consistently occurring for weeks or months
	(h) incoherent or irrelvant speech
	(i) catatonic behavior

1.5.3 Aetiology and Neurobiology

Schizophrenia is a mental illness with a complex aetiology. A complete picture of the neurobiology of the disease is yet to be determined, but studies regarding the neurochemistry, neurodevelopment and neuropathology aspects of the illness have provided some insights into schizophrenia.

1.5.3.1 Dysfunction of Neurotransmitter Systems

The theories of neurochemical imbalances in schizophrenia originated from pharmacological studies. The dopamine hypothesis was first proposed to explain the aetiology of schizophrenia (van Rossum 1966). The hypothesis was based on two pharmacological observations: (1) First, schizophrenia symptoms improved in response to conventional antipsychotic medications, such as chlorpromazine, which worked through the dopamine receptors. It was established that such traditional neuroleptics worked by competitive blockade of the D2 subtype of dopamine receptor (Creese et al. 1976; Seeman et al. 1976). (2) Secondly, administrations of drugs such as amphetamines increase dopamine levels, concurrently exacerbate psychotic symptoms of schizophrenia (Abi-Dargham et al. 1998). Subsequently, it was suggested that the simple hyperactive dopamine theory cannot explain the complicated aetiology of schizophrenia, including the negative symptoms of the illness, and the depression of dopamine subtypes in the prefrontal cortex. The dopamine hypothesis was later reformulated, proposing increased dopaminergic transmission in the basal ganglia to underlie positive symptoms and psychosis, whereas dopamine deficits in the cortical area were associated with cognitive impairment (Owen et al. 2005b).

The new generation of schizophrenia drug treatment (atypical antipsychotics), however, generally had low affinity for dopamine receptors, yet they were effective in reducing

positive and negative symptoms (Kandel 2000). This implicated the involvement of another neurotransmission system, and led to an alternative hypothesis concerning glutamate transmission, in particular NMDA (N-methyl-D-aspartate) receptor signalling, the major excitatory glutamate transmitter pathway. The first evidence of the theory came from two drugs, phencyclidine (PCP, also known as “angel dust”) and ketamine (also known as “Special K”). Administration of either leads to both positive symptoms (such as hallucinations) and negative symptoms reminiscent of schizophrenia. Furthermore, the drugs were demonstrated to exacerbate psychotic symptoms in chronic schizophrenia patients (Krystal et al. 1994; Lahti et al. 1995). Both drugs act by blocking NMDA receptor subtypes. In contrast, NMDA-receptor agonists such as glycine were found to alleviate schizophrenia symptoms and provided promising clinical results in some studies when used in combination of other drug treatments (Goff et al. 1995; Evins et al. 2002). Finally, mice engineered with a hypomorphic allele (with reduced expression) of NMDA-receptor subunit *NR1* (glutamate receptor, ionotropic, N-methyl D-aspartate 1) were shown to display sensory gating deficits and behavioural phenotypes, including increased motor activity and deficits in social interaction, reminiscent of the human disease (Duncan et al. 2002; Duncan et al. 2004; Morris et al. 2004; Fradley et al. 2005). Similarly, glycine binding site mice mutants also displayed behavioural and cognitive phenotypes, for example hyperactivity, increased startle response and LTP deficits (Ballard et al. 2002).

The dopamine and glutamate hypothesis were both well-established with independent support and evidence. Moreover, the two neurochemical pathways may interact to result in pathogenesis of schizophrenia. One suggestion was that dopaminergic was secondary to glutamatergic transmission (Laruelle et al. 2003). Alternatively, they could act synergistically in the neurocircuit to result in disease phenotypes. In addition, a

number of other neurotransmission systems such as serontonin transmission and GABAergic cortical dysfunction have been implicated in schizophrenia pathogenesis (Carlsson et al. 1999).

1.5.3.2 Neurodevelopment and Neuropathology

Apart from the neurochemical imbalance models, a neurodevelopment theory was also proposed, suggesting schizophrenia could result from abnormal brain development, with pathology originated in the middle stage of intrauterine life (Murray and Lewis 1987; Dazzan and Murray 1999). One support for the neurodevelopment model was the observation that 'soft' neurological signs, such as involuntary movement and dyskinesias, could extend into early ages such as childhood or infancy (Murray and Lewis 1988). Other evidence came from neuropathology, which demonstrated abnormalities of the central nervous system during development. Changes in the cytoarchitecture of the brain, for example increase in ventriclular size in patients, have been identified (Harrison 1999; Harrison and Weinberger 2005). Advance in imaging techniques such as magnetic resonance imaging (MRI) and poisitron emission tomography (PET) scan, as well as improvement in neuroanatomic markers, further enhanced the discoveries in neuropathology. Although some of the findings in the field were inconsistent and the effects of changes might be small, reasonable evidence had fostered the view that such neuropathological abnormality could relate to schizophrenia (Harrison and Weinberger 2005; Owen et al. 2005b). Recent evidence from mouse models further suggested that adult neurogenesis, in particularly in the brain region of dentate gyrus, could be involved in the pathogenesis of schizophrenia (Duan et al. 2007; Yamasaki et al. 2008).

1.5.4 The Genetics of Schizophrenia

1.5.4.1 Evidence of Genetic Contribution

While the aetiology of schizophrenia remains ambiguous, there is consistent evidence for a strong genetic component and inherited factors in schizophrenia (McGuffin et al. 1995), with support from twin, adoption and family studies (Sullivan 2005). Heritability estimates were demonstrated to be ~0.80 (Cardno and Gottesman 2000; Sullivan et al. 2003). Twin studies demonstrated increased risk in monozygotic (~41-65% lifetime prevalence) versus dizygotic twins (up to 28%) (Cardno and Gottesman 2000). Adoption studies established a higher risk in the adoption cluster (offspring of one set of parents raised by unrelated strangers from early in life), and that adoptees have increased risk with regard to biological versus adoptive affected parents (Heston 1966; Kety et al. 1994). Finally, family studies showed a higher risk in first degree relatives (Sullivan 2005). Figure 1.7 summarizes some of these findings.

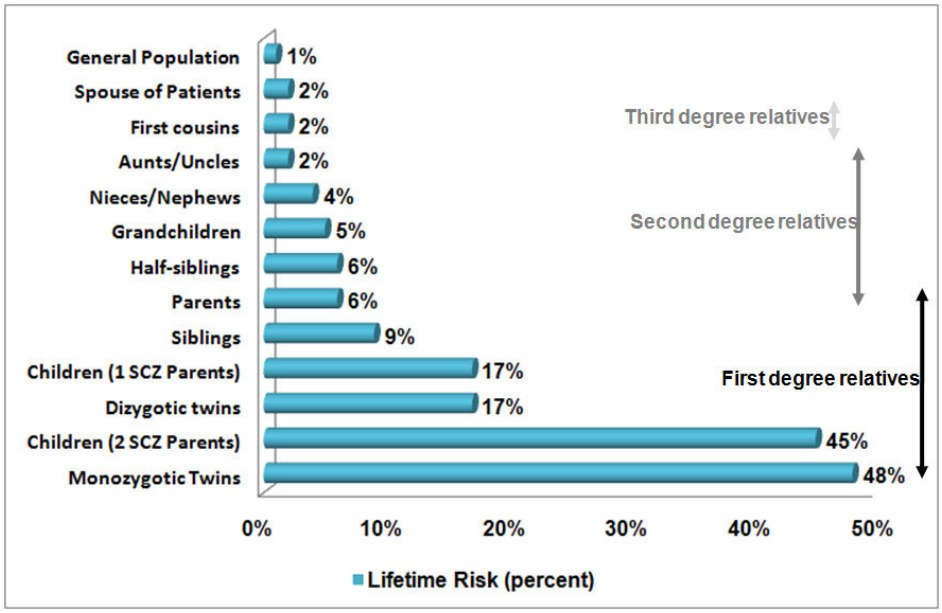

Figure 1.7 Lifetime risk of developing schizophrenia based on relationship. *(diagram adapted from Gottesman 1991)* (Gottesman 1991) The lifetime prevalence of schizophrenia in the general population is 1%. Lifetime risk increases as the degree of relatedness increases. (*SCZ*: schizophrenia)

1.5.4.2 A Search for Candidate Genes

Schizophrenia is a complex trait (Sullivan et al. 2003), probably involving a number of causative genes, with the complication of environmental and incidental factors (Norton et al. 2006). The evidence above suggested the existence of a strong genetic component, however the search for such genetic susceptibility factors has remained a challenge in molecular genetics. A description of the history of schizophrenia genetics would start with early linkage analysis, progressing to association studies of biological or positional candidates, and finally to recent genome-wide association studies (GWAS). These molecular genetic studies were also complemented by early identification of cytogenetic abnormalities, and recently whole-genome discovery and association of copy number variants in schizophrenia (to be discussed in section 1.6).

Standard linkage analysis used large pedigrees (multiplex families) to discover regions of the genome in linkage disequilibrium with the susceptibility locus/ disease allele. Parametric linkage analysis requires specification of the model of segregation, and has been useful in deciphering Mendelian traits. Non-parametric linkage analysis does not require knowledge on the mode of transmission, and was particularly useful for complex diseases such as schizophrenia (Burmeister et al. 2008). Early linkage results were, however, disappointing, probably due to inadequate sample sizes, and that schizophrenia presumably has a large number of susceptibility genes with small effect (Owen et al. 2005a). Meta-analysis only confirmed the inconsistencies of findings and further supported the view of multiple susceptibility loci (Badner and Gershon 2002; Lewis et al. 2003).

Association studies make use of large number of unrelated cases and controls, or family groupings such as sib pairs or trios (parent-child set) to identify genetic loci (Shelling and

Ferguson 2007). Association tests rely on polymorphic DNA markers, for example microsatellites, VNTR (variable number of tandem repeat) and SNP (single nucleotide polymorphism). The aim is to identify markers in linkage disequilibrium with disease variants by association tests. Such studies do not require specification of mode of transmission, can detect genes of small effect (Risch and Merikangas 1996), and improved the resolution of linkage analysis. Early association studies were mostly single-gene efforts, testing association of candidates from positional studies or from biological pathways associated with schizophrenia.

Although not all linkage and association studies in schizophrenia generated reproducible loci, some susceptibility genes or genetic region did receive multiple strands of evidence, and encompassed biologically plausible candidates. The candidates that received the most attention include (a) neuregulin 1 (*NRG1*), which plays a role in both neurodevelopment and glutamatergic synaptic transmission; (b) dysbindin (*DTNBP1*), part of the dystrophin protein complex and is involved in the uptake of glutamate into synaptic vesicles; (c) Disrupted in Schizophrenia 1 (*DISC1*), which plays a neurodevelopment role and was first identified by a translocation segregating in a schizophrenia family; (d) D-amino acid oxidase activator DAOA (G30/G72), which is involved in the oxidisation of d-serine, an endogenous modulator of NMDA receptors; and (e) regulator of G-protein signalling (*RGS4*), which modulates intracellular signalling for many G-protein-coupled receptors (Owen et al. 2005a; Owen et al. 2005b; Sullivan 2005; Burmeister et al. 2008) (Table 1.2).

Table 1.2 Schizophrenia candidate genes with evidence from linkage and association analysis. *(table modified from Burmeister et al., 2008; Sullivan et al., 2005)*

Gene	Gene Description	OMIM	Cytogenetic Band	Cyotegnetic Abnormality	Linkage Evidence	Association Study Support
NRG1	Neuregulin 1	142445	8p12	No	Yes	Yes
DTNBP1	Dystrobrevin binding protein 1	607145	6p22.3	No	Yes	Yes
DISC1	Disrupted in schizophrenia 1	605210	1q42.2	Yes	No	Yes
DAOA	D-amino acid oxidase and activator	607415	13q33.2	No	Yes	Yes
RGS4	Regulator of G-protein signaling 4	602516	1q23.3	No	Yes	Yes

Recent advance in molecular genetics included the realization of GWAS (Wollstein et al. 2007). These whole-genome association studies were made possible due to the increase in density and characterization of single nucleotide polymorphisms (SNPs) (Hinds et al. 2005; IHMC 2005), together with the development of genotyping chips.

Two schizophrenia GWAS with modest sample sizes (500–800 cases) first published in 2007 and early 2008 generated one candidate locus (pseudo-autosomal, near cytokine genes) of genome-wide significance (Lencz et al. 2007; Sullivan et al. 2008). The most recent GWA study, however, used a multistage association study design with an initial 479 samples and replication cohorts of 16,726 subjects, provided genome-wide significance at 3 candidate loci, and the strongest evidence at 2q32.1 near *ZNF804A* (a zing finger protein) (O'Donovan et al. 2008).

Finally, the latest SNP genotyping platforms used for GWAS also allowed identification of copy number variants, resulted in some promising findings of schizophrenia susceptibility loci (see section 1.6).

1.5.4.3 Genes and Environment

For full manifestation of the illness, the genetic susceptibility factors which have increasingly received evidence probably act in concert with non-genetic elements such as environmental, epigenetic or other stochastic factors. Evidence of environmental factors came from monozygotic twin studies, which suggested that although ~40-60% of the disease risks can be explained by genetic vulnerability, there is a considerable non-genetic contribution (Sullivan et al. 2003). Some environmental factors which demonstrated association with schizophrenia include obstetric complications, place and time of birth (e.g. urban areas, famine, winter), advanced paternal age, and exposure to infectious agents (e.g. influenza, rubella) (Sullivan 2005).

In summary, schizophrenia is a debilitating illness exhibiting tremendous genetic and phenotypic complexity. Like other complex diseases, schizophrenia is most likely to be multifactorial, with contributions from both genes and the environment. Pharmacological studies, brain imaging, and genetic research have offered us a glimpse of the disease aetiology, with more underlying factors waiting to be unveiled.

1.6 CNV in Schizophrenia and other Psychiatric Diseases

1.6.1 Early Studies on Chromosomal Abnormalities in Schizophrenia

With the involvement of CNVs being established for a number of complex diseases and human traits, it is not a major leap to hypothesize that CNVs could also account for the genetic basis of psychiatric disease, which in essence is a combination of behavioural, psychological and cognitive variations among individuals.

Structural variants have long been identified as mutations causing schizophrenia in familial studies. Several translocations, e.g. at *DISC1* (Disrupted in Schizophrenia 1) and *PDE4B* (Phosphodiesterase 4B), as well as the microdeletion on 22q11 in patients with DiGeorge/Velocardiofacial Syndrome, were all identified as high-penetrant schizophrenia variants through traditional cytogenetics and karyotyping (MacIntyre et al. 2003). These chromosomal abnormalities provided some classical examples endorsing the involvement of rare chromosomal rearrangements in the pathogenesis of schizophrenia.

1.6.1.1 DISC1 and Breakpoint Study in Schizophrenia

Disrupted in Schizophrenia 1 (*DISC1*) is a schizophrenia candidate gene with perhaps the strongest evidence so far, from human genetics, molecular and cell biology to insights from animal models. *DISC1* was first identified as a balanced reciprocal translocation t(1;11) (q42;q14) which cosegregated in a large Scottish family with schizophrenia and other psychiatric diseases (St Clair et al. 1990; Blackwood et al. 2001). The initial pedigree included 21 individuals with schizophrenia, bipolar disorder, recurrent depression and conduct disorder, of which 16 were identified with the 1q42 translocation, giving a highly significant linkage signal with LOD (logarithm of odds) score of >7.0. Two brain-expressed genes were revealed at the breakpoint: Disrupted in Schizophrenia 1 and 2 (*DISC1 and DISC2*) (Millar et al. 2000; Blackwood et al. 2001).

The 1q24 linkage result was replicated in a subsequent study in a Finnish cohort (Ekelund et al. 2001), and the involvement of *DISC1* in schizophrenia was later confirmed in a number of association studies (Hodgkinson et al. 2004; Kockelkorn et al. 2004). A number of mouse models (with missense mutations, truncated *DISC1* or inducible expression) have further provided evidence of *DISC1* being involved in behaviour and learning and memory (Koike et al. 2006; Clapcote et al. 2007; Pletnikov et al. 2008; Shen et al. 2008).

DISC1 plays important roles in neurodevelopment, cytoskeletal function and cAMP signalling (Chubb et al. 2008). The study of *DISC1* and its interacting proteins have provided us with important clues into disease pathogenesis (Brandon 2007). In particular, phosphodiesterase 4B (*PDE4B*), an interacting partner of *DISC1*, was also revealed as a disease-causing chromosomal translocation (Millar et al. 2005). Both *DISC1* and *PDE4B* are involved in cAMP signalling, an important pathway in neuronal signal network and the synapse (Millar et al. 2005; Bradshaw et al. 2008). *PDE4B* in flies (the dunce fly) was demonstrated to have a role in learning and memory, and was targeted by the antidepressant rolipram (Davis and Dauwalder 1991). These results strengthened the role of *DISC1* and its interactors as genetic predisposition factors of schizophrenia.

DISC1 translocation and the related examples suggest how disruption of genes through chromosomal rearrangement could reveal penetrating variants and biological pathways, a basis for further investigation in case-control cohorts and molecular genetics. Although gene disruption by balanced translocation does not involve copy number change, it is highly analogous to copy number variation, where the breakpoints of the rearrangement could lead to disruption of genes associated with behaviour and cognitive phenotypes.

1.6.1.1 22q11 Microdeletion and Schizophrenia

The 22q11 microdeletion syndrome (also known as DiGeorge syndrome, Velocardiofacial Syndrome, or CATCH22), is a complex disorder with patients displaying learning disability, cardiac defect, craniofacial abnormalities and palatal defect (Karayiorgou et al. 1995; Murphy 2002; Paylor and Lindsay 2006). Among patients there is an unusually high prevalence (~30%) of behavioral abnormalities, including schizophrenia, bipolar disorder, autism and other psychosis (Murphy et al. 1999; Murphy and Owen 2001; Fine et al. 2005). A number of linkage studies have also indicated involvement of this region in schizophrenia as well as bipolar disorder (Sklar 2002), with the first evidence coming from Lachman *et al.* (Lachman et al. 1997). The microdeletion was said to be one of the highest known risk factors for schizophrenia (Murphy and Owen 2001), second to family inheritance such as disease concordance among monozygotic twins.

The increased risk of schizophrenia results from hemizygosity of a 3 Mb region (or in some cases a smaller 1.5 Mb region) (Edelmann et al. 1999), and is expected to act through haploinsufficiency of one or more genes, or through unmasking deleterious polymorphism(s) in the region (Wilson et al. 2006a). To this end, a number of mouse models, including mice deleted with a 1.1 Mb region of the 22q11 locus generated from chromosome engineered deficiency, showed deficit in sensorimotor gating and cognitive impairment (Paylor et al. 2001; Paylor and Lindsay 2006).

The exact gene(s) causing the predisposition is unknown, and there are some functional candidates within the deleted region including catechol-O- methyltransferase (*COMT)*, proline dehydrogenase (*PRODH)* and the micro-RNA processing gene DiGeorge Syndrome Critical region 8 (*DGCR8)* (Bray 2008). *COMT* is one of the two principal

enzymes that degrade catecholamines (e.g. dopamine) (Meyer-Lindenberg et al. 2005). In 1996, Lachman *et al.* described a common polymorphic SNP leading to a valine-to-methionine codon substitution in *COMT* (Lachman et al. 1996), which was later demonstrated to alter dopamine in the pre-frontal cortex (Meyer-Lindenberg et al. 2005) and affect performance in working memory tasks in normal human subjects (Egan et al. 2001; Malhotra et al. 2002). Knockout mice of *Comt* also exhibited elevated dopamine levels in the prefrontal cortex and altered emotional and social behavior (Gogos et al. 1998).

The *PRODH/DGCR8* region represented another susceptibility locus within the 22q11 deletion (Liu et al. 2002). Mouse models with homozygous deletion of *PRODH* (Gogos et al. 1999) showed sensorimotor gating deficits, but with different effect. More recently, Stark and colleagues (Stark et al. 2008) generated another mouse model with a deletion of the segment syntenic with the 1.5 Mb deleted interval in human, removing a number of genes including *Comt*, *Prodh2* and *Dgcr8*. Transcriptional profiling of the mice revealed dysregulation of genes outside the deletion region, particularly those involved in synaptic transmission. The mouse model also revealed a rather unexpected role of *Dgcr8* and micro-RNA biogenesis in some of the sensorimotor phenotypes and abnormalities in neuronal morphology (Stark et al. 2008).

Finally, the breakpoints of the deletion were also of interest both in terms of deletion mechanism and gene involvement. The deletion was thought to cluster to several known homologous low copy repeats or segmental duplications, mediating the recurrent rearrangement of the 3 Mb or 1.5 Mb regions (Edelmann et al. 1999; Emanuel and Shaikh 2001; Shaikh et al. 2001; Sharp et al. 2005). Recent studies from Urban *et al.* used Nimblegen oligonucleotide array CGH and high-resolution PEM to detect

breakpoints with much higher accuracy. They revealed differences in patients whose chromosomal abnormalities had been indistinguishable using more conventional cytogenetic methods (Urban et al. 2006). These differences (of up to 14 genes and 200 kb either side of the breakpoint) may play a role in the variability of phenotypes in patients of hitherto identical breakpoints and deletion region.

In short, 22q11 microdeletion demonstrates a model of highly complex gene interaction and genotype-phenotype relationship, resulting in a variety of behavioral phenotypes including schizophrenia. Research into the region has also provided important insights into subsequently identified recurrent loci (ISC 2008; Walsh et al. 2008), many of which are mapped to mutational hotspots of segmental duplication in the genome similar to the 22q11 locus.

1.6.2 Large Scale CNV Screen in Schizophrenia Patients

Chromosomal abnormalities in schizophrenia patients provided the first evidence of copy number variation altering complex behavioural traits and maintaining disease-associated genotypes in the population. Recent genome-wide CNV studies provided emerging evidence of additional rare causative copy number variants in schizophrenia.

1.6.1.1 Summary of CNV Findings

Since 2006, a number of pilot whole-genome schizophrenia copy number screens have been published (Moon et al. 2006; Wilson et al. 2006a). Interpretation of the early results remained difficult, since the CNV techniques were at their infancy and inconsistent results were reported (Sutrala et al. 2007). A later CNV screen targeting 15 candidate genes for schizophrenia, including neuregulin (*NRG1*) and *DISC1*, also revealed no fruitful results (Sutrala et al. 2008).

The year 2008 was a turning point in terms of schizophrenia genetics, with promising results from a number of large scale whole-genome schizophrenia association studies including a number of reports on copy number variations (ISC 2008; Kirov et al. 2008; Rujescu et al. 2008; Stefansson et al. 2008; Vrijenhoek et al. 2008; Walsh et al. 2008; Xu et al. 2008).

The report from Walsh *et al.* was one of the first studies revealing the genomic architecture of schizophrenia based on CNV screening (Walsh et al. 2008). They screened a case cohort of 150 patients, and a second cohort of 83 childhood-onset schizophrenia cases, using both ROMA and SNP arrays. A number of rare variants were revealed; with breakpoints perturbing genes involved a number of key pathways during neurodevelopment. Of particular significance, Walsh *et al.* showed a higher proportion of

rare variants in cases compared to controls, suggesting rare CNVs are schizophrenia-predisposition factors.

Another two reports described results of large-scale genome-wide screens on thousands of schizophrenia patients using SNP arrays (ISC 2008; Stefansson et al. 2008). Collections of patient and control DNA samples were pooled from various clinical and research groups around the world. The first study by the International Schizophrenia Consortium (ISC) reported their CNV analysis on >3000 Caucasian patient samples with ethnically matched controls based on Affymetrix SNP arrays (ISC 2008). The second study by Steffanson *et al.* reported CNV findings on samples mainly from the SGENE consortium in addition to a number of participating groups (herein refer as SGENE+), using Illumina arrays as CNV screening platforms (Stefansson et al. 2008). The two studies combined revealed three novel schizophrenia-associated recurrent CNV loci. The role of CNV mutational burden was also confirmed in the ISC study.

A fourth study came from Xu *et al.* addressing the role of *de novo* copy number variations in schizophrenia (Xu et al. 2008). By screening 200 trios with an affected child (152 sporadic cases and 48 familial cases) and a similar number of non-affected control trios using Affymetrix SNP arrays, Xu and colleagues discovered a higher burden of *de novo* copy number mutations in sporadic disease cases.

And finally reports from Kirov *et al.* and Vrijenhoek *et al.* revealed a number of rare variants with relevance to synaptic development (Kirov et al. 2008; Vrijenhoek et al. 2008). Of particular interest was the gene Neurexin1 (*NRXN1*), which was identified as a CNV locus in both studies, and was later confirmed as a schizophrenia-associated recurrent CNV (Rujescu et al. 2008).

1.6.2.1 Identification of Large Recurrent Schizophrenia Loci

Among the most promising outcome from the various CNV screens was the identification of at least three recurrent loci of statistically significant association with schizophrenia, as seen in the two large-scale case-control studies from ISC and SGENE+ (ISC 2008; Stefansson et al. 2008). The three loci were at chromosome 1q21.1, 15q13.3 and 15q11.2, with the former two replicated in both studies (Figure 1.8). The regions at 1q21 and 15q13 were extremely rare events, occurring in ~0.1-0.3% in cases, and about ten times less frequent in controls (according to SGENE+). Estimated odds ratios were high (>10 in SGENE+ and 6 to 18 in ISC). The region at 15q11 was relatively more frequent among the three, but still at a rare 0.55% in cases (and 5 times less frequent in controls), with odds ratio at 2.7. Table 1.3 summarizes the details of the three schizophrenia-associated CNV loci.

Of the genes affected in the three regions, a number of highly plausible schizophrenia candidates are particularly worth mentioning. At 1q21, the gene connexin-50 (*GJA8)* encodes a gap junction subunit and was previously reported as associated with schizophrenia in case-control and family studies (Ni et al. 2007). This CNV region has also been shown as associated with schizophrenia by linkage analysis (Stefansson et al. 2008). The 15q11.2 CNV partially overlaps with the Prader-Willi Syndrome locus (Christian et al. 1995) and maps to the breakpoints as identified in some of the Prader-Willi patients (Murthy et al. 2007). Within the ~500 kb region there is the cytoplasmic FMR1 interacting protein 1 (*CYFIP1*) gene*,* which encodes a protein that interacts with both the fragile X mental retardation protein FMRP and the translation initiation factor eIF4E (Napoli et al. 2008). The CYFIP1/FMRP interaction regulated mRNA translation in neuronal dendrites, demonstrating its role in synaptic plasticity and brain development. It was noted that both Prader-Willi Syndrome patients and fragile X patients frequently

have autistic features (Rogers et al. 2001; Dimitropoulos and Schultz 2007). Finally at the third locus 15q13.3, the α7 nicotinic receptor gene *CHRNA7* is a schizophrenia candidate gene (Freedman et al. 1997), and was shown to interact with neuregulin1-erbB signalling (Hancock et al. 2008).

A noteworthy characteristic of the three recurrent loci was that they were all flanked by LCRs, which, analogous to the genomic disorders as discussed before, reflects the underlying architecture of the genome (ISC 2008). NAHR, for example, could be mediated by these LCRs, generating recurrent deletions and reciprocal duplications (Lupski 2006). Consistent with this, reciprocal duplications have been observed for these recurrent loci (ISC 2008). The mutational rate for this type of genomic rearrangement hotspot is estimated to be 10^{-6} to 10^{-4}, many folds higher than for SNP (10^{-8}) (Lupski 2007). This may partially explain the recurrent feature of schizophrenia at a similar rate in all populations, and the maintenance of schizophrenia as a disease in the population even though the disease confers reduced fecundity (Stefansson et al. 2008). Analyzing the 3 deletions in the Icelandic population (part of SGENE Consortium) with accurate genealogy provided further evidence for negative selection in these 3 deletions (St Clair 2008).

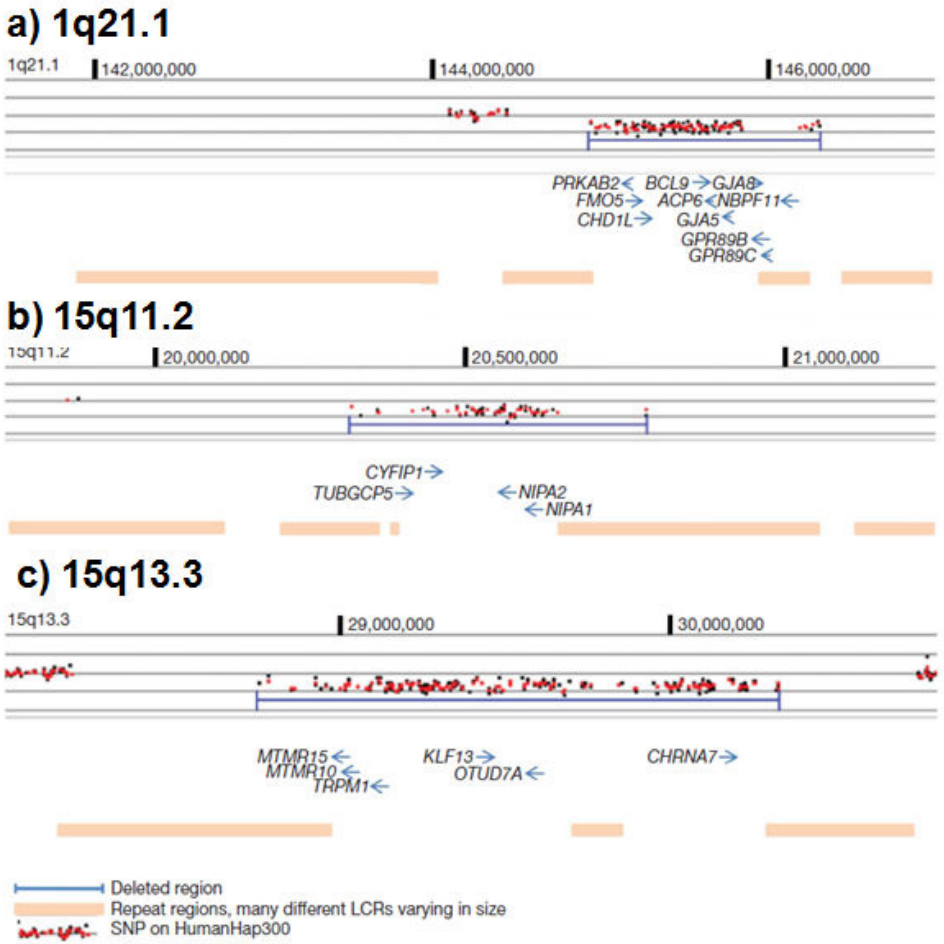

Figure 1.8 Genomic loci of the three novel recurrent deletions associated with schizophrenia (Diagram adopted from Stefansson *et al.*) a) 1q21.1 region; b) 15q11.2 region; c) 15q13.3 region

Table 1.3 Three novel recurrent deletions associated with schizophrenia.

				1q21.1	15q11.2	15q13.3
Estimated Chr Coordinates				chr1:142-146.3 (Mb)	chr15:20.31–20.78 (Mb)	chr15:28 – 31 (Mb)
Estimated Size				~1.3 Mb (small)/ ~3.5 Mb(large)	470kb	~3 Mb
Prevalence	SGENE+	Case	(n=4718)	0.23%	0.55%	0.17%
		Control	(n=41199)	0.02%	0.19%	0.02%
	ISC	Case	(n= 3391)	0.2949%	NA	0.2654%
		Control	(n= 3181)	0.0314%	NA	0.0000%
Odds Ratio (p-value)	SGENE+			14.83 *(p=0.024)*	2.73 *(p=0.007)*	11.54 *(p=0.040)*
	ISC			6.6 *(p=0.023)*	NA	17.9 *(p=0.0023)*
Affected Genes				*SEC22B, NOTCH2NL, HFE2, TXNIP, POLR3GL, LIX1L, RBM8A, PEX11B, ITGA10, ANKRD35, PIAS3, POLR3C, ZNF364,CD160, PDZK1, NBPF11, FAM108A3, PRKAB2, FMO5, CHD1L, BCL9, ACP6, GJA5, GJA8, GPR89A, NM_207400, NBPF15*	*TUBGCP5, CYFIP1, NIPA2, NIPA1, BC044583, WHDC1L1*	*CHRFAM7A, MTMR10, TRPM1, has-mir-211, KLF13, OTUD7A, CHRNA7*

SGENE+: Report from Stefansson et al. 2008; ISC: Report from ISC, 2008

1.6.2.2 Increased Mutation Burden of CNV in Schizophrenia Patients

Another major finding from the CNV screens was the excess of mutational burden as conferred by rare CNVs in cases versus control. This provides important insight into the genetic model of schizophrenia.

The effect of mutational burden by CNVs was evaluated by Walsh *et al.*, the ISC and Xu *et al.* (ISC 2008; Walsh et al. 2008; Xu et al. 2008). All three studies detected a large number of rare variants in the cases, and demonstrated the causal role of these rare variants in increasing disease risks. In particular, Walsh *et al.* reported a three-fold increase of the proportion of rare, genic variants (those that delete or duplicate genes) compared to total CNVs in cases (15%) versus control (5%). The effect was magnified when only childhood-onset cases were taken into account (20%), reflecting a possible increased genetic burden of childhood psychiatric disorders (Walsh et al. 2008).

The ISC study also demonstrated a higher CNV burden of rare CNV (as defined as CNV <1% frequency in all samples screened). The extent of the CNV burden was smaller compared to the report from Walsh *et al.* (from 1.15 to 1.4 times increase). The signal was intensified (to 1.45 to 1.67) when only the extremely rare CNV events – the singletons occurring once in all samples screened - were considered (ISC 2008).

Xu *et al.* added another level of genetic complexity by studying affected trios (sets of parents and affected child) and demonstrated the role of *de novo* rare CNVs in disease. Their results revealed an eight-times increased rate of *de novo* CNVs in sporadic cases compared to controls. The effect was much smaller in familial cases, showing a distinct difference in the genetic determinations of sporadic versus familial cases (Xu et al. 2008).

All these studies came to the conclusion that rare variants (in particular the rare genic CNVs) were collectively significant risk factors for schizophrenia. In sporadic cases in particular, *de novo* CNVs were significantly involved in disease predisposition.

1.6.2.3 Rare Variants Converging into Neurodevelopmental Pathways

Although the role of rare CNVs in schizophrenia pathogenesis has been collectively demonstrated, the bulk of the published schizophrenia CNV regions are composed of unique events, with minimal overlap between studies, apart from the three large recurrent loci. It is therefore extremely difficult to pinpoint which CNV(s) or gene(s) is/are schizophrenia risk factors. As a method to decipher the molecular basis of the disease, one could look for convergence of molecular components into biological pathways. This is possible with the increasing amount of accumulated CNV data on schizophrenia.

One common feature that may start to emerge from these rare events is the involvement of proteins related to synaptic development and functions (Walsh et al. 2008), in particular those that contribute to neuregulin and neurexin signalling in the synapse (Figure 1.9). The strongest evidence comes from the 2p16.3 locus with the gene neurexin1 (*NRXN1*), one of the very few recurrent schizophrenia loci that has been replicated so far in multiple studies. The 2p16.3 locus has been previously shown as associated with autism (Feng et al. 2006; Szatmari et al. 2007; Kim et al. 2008a; Marshall et al. 2008; Zahir et al. 2008). Evidence suggesting its role in schizophrenia first came from Kirov *et al.* (Kirov et al. 2008) who screened 93 patients using array CGH and identified a segregating *NRXN1* CNV in a pair of affected schizophrenia siblings and their asymptomatic mother. Walsh *et al.* also detected a neurexin CNV in a pair of identical twins diagnosed with childhood-onset (COS) schizophrenia, among a cohort of 83 COS patients (Walsh et al. 2008). Subsequently, Vrijenhoek *et al.* identified one CNV

in the region by screening an initial cohort of 54 patients by array CGH, and revealed another 4 cases when screening this locus in more patients (Vrijenhoek et al. 2008). Furthermore, the ISC detected a number of neurexin CNVs with different breakpoints in cases (as well as in controls) (ISC 2008).

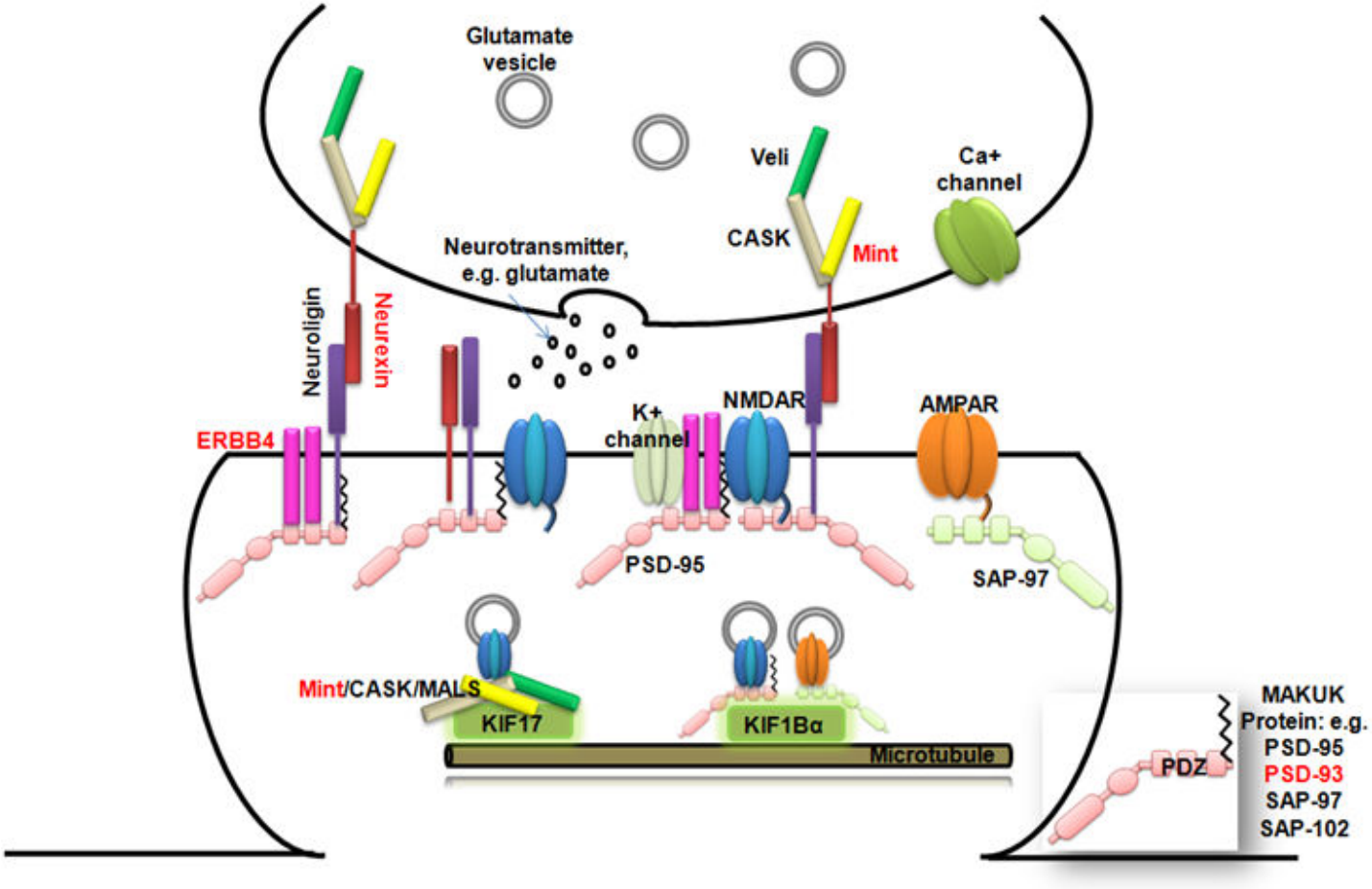

Figure 1.9 Schematic diagram of the synapse displaying known involvement of synaptic proteins in schizophrenia CNV loci (Scannevin and Huganir 2000; Kim and Sheng 2004; Keith and El-Husseini 2008).
Red: proteins identified as schizophrenia CNV loci. (legend to be continued on next page)

*ERBB4, v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 4; Mint, amyloid beta A4 precursor protein-binding; CASK, calcium/calmodulin-dependent serine protein; MALS, lin-7 homolog; K+ channel, potassium channel; NMDAR, NMDA (*N*-methyl-D-aspartate) receptor; AMPAR, AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptor; Ca+ channel,calcium channel; KIF17 kinesin family member 17; KIF1Bα, kinesin family member 1B α; PSD-95, postsynaptic density protein 95; PSD-93, postsynaptic density protein 93; SAP-97, synapse-associated protein 97; SAP-102 synapse-associated protein 10*

Apart from the above genome-wide studies, Rejuscu *et al.* recently targeted 3 neurexin genes (*NXRN1*, *NXRN2* and *NXRN3*) using a candidate CNV approach to look for structural rearrangement in ~3000 patients and >30,000 controls (Rujescu et al. 2008). They identified 61 *NRXN1* deletions (one *de novo*) and 5 duplications throughout the locus. By restricting the analysis to CNVs that disrupt exons, the authors revealed a ten-fold increase in mutational burden, with 0.17% in cases harbouring such CNVs compared to 0.020% in controls (odds ratio ~10). This recurrent locus distinguishes itself from the three aforementioned large recurrent deletions in two ways: first the breakpoints of the neurexin CNVs were not flanked by LCRs. There was no common breakpoint (consistent with previous findings), and the CNVs vary in size (18 kb to 420 kb). Long range phasing analysis on the *de novo* case provided evidence against NAHR as a rearrangement-generating mechanism (Rujescu et al. 2008). Secondly, rather than deleting or duplicating whole gene(s) as described previously, CNVs in this locus most likely acted by disrupting the gene *NRXN1*. It was concluded that NRXN1 deletions (in particular the ones that affect exons) confer risk of schizophrenia (Rujescu et al. 2008).

Neurexin 1 belongs to a large family of proteins that act as neuronal cell-surface receptors. Neurexin is concentrated at the pre-synaptic membrane, where it functions as a neuroligin receptor (Reissner et al. 2008). Interestingly, neuroligins were shown to be sufficient to trigger presynaptic differentiation through neurexin (Dean et al. 2003), and neurexins could in turn trigger postsynaptic differentiation (Craig and Kang 2007). This transynaptic neurexin-neuroligin complex is important for excitatory glutamatergic and inhibitory GABAergic synapses in the mammalian brain (Craig and Kang 2007), playing a central role in synapse formation and neurotransmission.

Of direct relevance to the *NXRN1* CNV was a schizophrenia study identifying 3 patients carrying genomic deletions of the *CNTNAP2* gene, whilst no CNV was detected in 512 controls (Friedman et al. 2008). *CNTNAP2* encodes Caspr2 contactin-associated protein 2), a member of the neurexin superfamily (Poliak et al. 1999). Furthermore, the same study which first detected the *NXRN1* disruption in schizophrenia revealed in another patient a CNV involving *APBA2* (*Mint2*) (Kirov et al. 2008), a neuronal adaptor protein that binds neurexins as part of a *CASK*-containing protein complex. It is likely that the complex functions in neurotransmitter synaptic vesicle docking/fusion through recruitment of the vesicle fusion protein Munc-18 to the sites of exocytosis (Biederer and Sudhof 2000).

The study by Walsh *et al.* further revealed a number of CNVs disrupting genes involved in synaptic transmission or neurodevelopment (Walsh et al. 2008). Of particular interest is the neuregulin (*NRG*) signalling pathway, which has diverse roles from neuronal migration, axon guidance, glial cell development, axon myelination, neurotransmitter receptor expression and synapse formation (Mei and Xiong 2008). These important functions position neuregulin as a key signalling protein for synaptic plasticity and neuronal survival.

Genes involved in the *NRG* pathway and disrupted by CNVs included two interacting partners *ERBB4* and *MAGI2* (Walsh et al. 2008). *ERBB4* is a type I transmembrane tyrosine kinase receptor for neuregulin-1 (*NRG1*). It was detected as a deletion in a patient. The neuregulin-erbB complex was suggested to interact with PDZ-containing proteins at neuronal synapses (Garcia et al. 2000). The complex interacts with *MAGI2* at neuronal synapses, detected as duplicated in another patient (Walsh et al. 2008). In addition, *ERBB4* was suggested to be recruited by *PSD95* (a PDZ-containg protein) to

the neurexin-neuroligin complex for synapse formation (Lin et al. 2000), providing a link between the CNVs discussed in this section so far. These evidences provide a glimpse into the complexity of the neuronal network, suggesting how the apparent genetic heterogeneity revealed by the schizophrenia CNV screens could be reconciled.

Lastly, although the three novel recurrent loci further suggested heterogeneity, they also confirmed the convergence of CNV loci to candidate neuronal and synaptic pathways. The α7 nicotinic acetylcholine receptor *CHRNA7A* at 15q13.3 was recently shown to be targeted by neuregulin1-erbB signalling to axons for surface expression in sensory neurons (Hancock et al. 2008). The *CYBIP1* protein deleted at 15q11.2 - together with its interactor, the fragile X mental retardation protein FMRP - regulates mRNA translation at the synapse (Napoli et al. 2008). FMRP was shown to regulate mRNA stability for PSD95 (Todd et al. 2003; Zalfa et al. 2007). Furthermore, in a study comparing the gene expression profiles of three lymphoblastoid cell lines - from (a) an autistic patient with the 15q11-13 duplication, (b) a Fragile X mental retardation patient with autism and (c) a normal control - Neuregulin-2 (*NRG2*) was detected as one of the 68 dsyregulated genes common in both patient cell lines (Nishimura et al. 2007).

In conclusion, investigating copy number variations as schizophrenia genetic risk factors has elucidated not only the genomic architecture of the disorder, but also the biology of the disease. Further screening for schizophrenia CNVs, and integrating them into biological pathways, will offer better understanding of the genetics of schizophrenia as well as other psychiatric diseases.

1.7 Scope of Thesis

The aim of this dissertation is to identify copy number variations which could play a role in schizophrenia pathogenesis. The thesis consists of three main sections:

In the first section (Chapter 3), I aim to assess the role of copy number variations in familial cases of schizophrenia with other psychiatric disorders. Using whole-genome tiling path (WGTP) array CGH, I will look for segregation of CNVs with diseases in 3 families, each with a schizophrenia proband and at least another close affected family member. Rare, family-specific variants that are not identified in normal controls will be further investigated as putative disease-causing variants. I will also use a candidate gene approach to study CNV in an extended pedigree with a high resolution oligonucleotide array, and to look for segregation with disease.

The second section (Chapter 4 and 5) will describe a population-based CNV screen on 91 schizophrenia patients and 92 matched controls using the WGTP platform. Following recent success of similar CNV studies, and their implication of rare variants in disease aetiology (ISC 2008; Stefansson et al. 2008; Walsh et al. 2008), I will focus on the study of rare, schizophrenia-specific CNVs and their role in psychiatric phenotypes. Replication of previous findings will be reported, as well as novel putative disease CNV loci in our dataset, with an emphasis on the assessment of gene content. In addition, I will examine the role of common copy number variations, an approach that has been less explored in previous schizophrenia CNV studies. I will attempt two methods to assess the contribution of frequent CNVs in disease. Candidates generated from these approaches will be tested in an extended cohort using target-specific PCR-based assays.

The last section (Chapter 6) will comprise of a CNV screen on variants overlapping with a set of genes playing crucial roles in behavior and cognition- the NMDA (N-methyl-D-aspartate) receptor complex. This complex consists of synaptic molecules for learning and memory functions. Components of the complex are also associated with various psychiatric disorders including schizophrenia (Grant et al. 2005). I will assess the contribution of CNVs in these genes in normal HapMap individuals. One particular genomic locus, the N-ethylmaleimide-sensitive factor at 17q21, will targeted for further investigation.