The human genome; you gain some, you lose some

Kriek, M.

Chapter I

Introduction

I-1. The plasticity of the human genome

Many authors have discussed the significance of gene and whole genome duplication in evolution (these publications are reviewed in (Taylor and Raes 2004)). Indeed, Ohno (1970) (in Evolution by gene duplication. New York: Springler-Verlag) stated that duplications of the genetic material were the most important factor driving evolution. Recently, projects using genome sequencing have shown that large scale gene duplications have contributed to the creation and expansion of gene families. Whether a duplication is passed onto future generations depends on whether the change is beneficial for survival. One example is the olfactory gene family. These (pseudo)genes create a redundancy of sequences contributing to the ability to smell, which appears to be beneficial for mammalian survival. A more recent example was published by Perry et al. (2007). They found that the copy number of the AMY1 gene is positively correlated with the amount of starch in a diet. We have also learned that the susceptibility of developing a disease is influenced by changes in CNVs. It has been shown that altered copy number of the CCL3L1 and FCGR3B genes influence susceptibility to HIV infection and systemic lupus erythematosus (SLE), respectively (Gonzalez et al. 2005; Aitman et al. 2006). These examples indicate that selection may operate on copy number variants containing sequences that are coding or regulating functions involved in survival.

A substantial proportion of (partial) gene duplications are gathered in segmental duplications (**chapter II-1**). Segmental duplications presumably originated from the duplication and subsequent transposition (and / or inversion) of genomic blocks (Eichler 2001a) from one chromosomal region to another some tens of million years ago (Bailey *et al.* 2002b; Armengol *et al.* 2003). It appears that these segmental duplications are often present at (breakpoint) loci where the human genome differs from that of the great apes (Samonte and Eichler 2002a) (Stankiewicz *et al.* 2001; Locke *et al.* 2003) and other species, such as mice (Armengol *et al.* 2003).

Besides duplications of existing sequences, another frequent form of variation in the human genome is deletion of unique sequences. In fact, it has been shown that these deletions are quite common in the human genome, with each individual having at least 30-50 deletions larger than 5 kb (Conrad *et al.* 2006). Van Ommen (2005) estimated that one in eight live births may have a *de novo* deletion. Some of these may enhance adaptation to environmental changes and might therefore be beneficial for survival. It is assumed that these deletion polymorphisms are exposed to more strict selection than Single Nucleotide Polymorphisms (SNPs), based on the fact that the X-chromosome contains less deletion polymorphisms compared to SNPs (Conrad *et al.* 2006).

In contrast to their potentially positive role in evolution, duplications and deletions (e.g. copy number variations = CNVs) (figure 1 A&B) in the human genome can also be related to inherited disease, mental retardation (MR), and congenital malformations (CM). For decades, it has been clear that numerical chromosome aberrations (e.g. trisomy 13, 18 and 21) and large CNVs have enormous influence on embryonic development and can lead to malformation syndromes or intra-uterine death. More recently, a systematic search for submicroscopic CNVs leading to MR and CM was initiated by Flint *et al.* (1995). These authors focused on the chromosome ends (also called the subtelomeres) and they found the percentage of alterations in their MR study population to be around 6%. Since that time, many different screening tools have been successfully implemented to find such cryptic (subtelomeric) CNVs (table 1). Detecting small CNVs on a genome-wide scale has only recently become possible with the development of micro-arrays. First results indicate that many CNVs are detected in patients with MR and CM (CNVs with phenotypic trait) as well as in healthy individuals (CNVs without an obvious phenotypic trait). In the most comprehensive CNV study to date no less than 12% of the human genome showed variations among healthy individuals (Redon *et al.* 2006). Consequently, our main challenge is currently to determine whether a variation is related to a phenotypic trait or not. This will remain so in the near future until the complete plasticity of the human genome has been fully mapped.

In short, copy number variations (CNVs) in the human genome are inherent in both evolutionary progression as well as the etiology of disease. The introduction of this thesis will review CNVs that appear to be neutral as well as CNVs that appear to be related to a phenotypic trait. This will be followed by a review of the many different technical approaches that can be used for detecting genomic rearrangements.

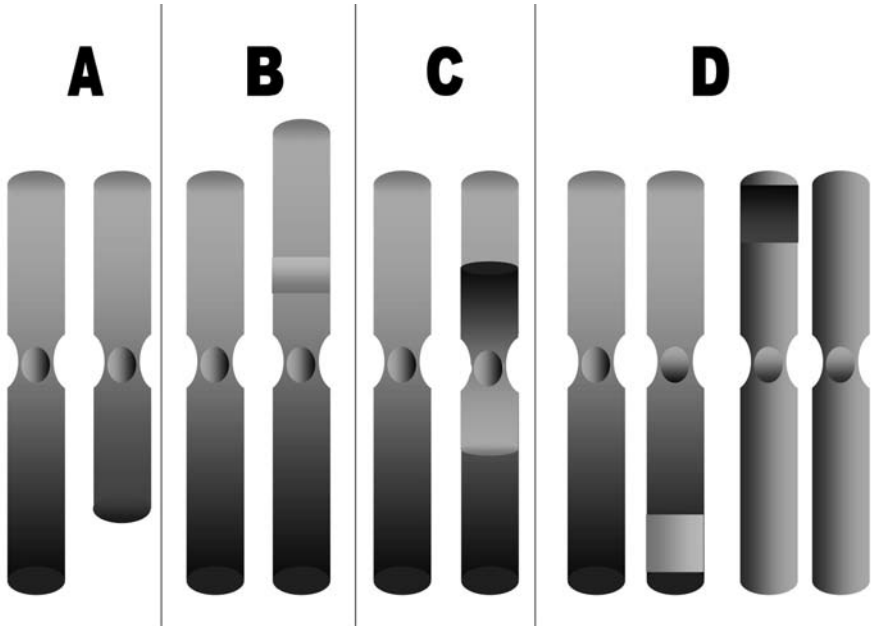
The articles **(chapter II & III)** describe several studies that have applied the rapidly evolving techniques for CNV detection to the clinical problem of unexplained MR and CM. The availability of the new diagnostic tools will greatly increase our understanding of the genetic causes of MR and CM, and might one day lead to therapeutic interventions in some cases.

I-2. CNVs WITH NO OBVIOUS PHENOTYPIC TRAIT

2.1. Neutral CNVs

Copy number variants have been identified since the start of the cloning era, however, the full extent of the variability and plasticity of the human genome has only recently

Figure 1. Deletion, duplication, inversion and balanced translocation.

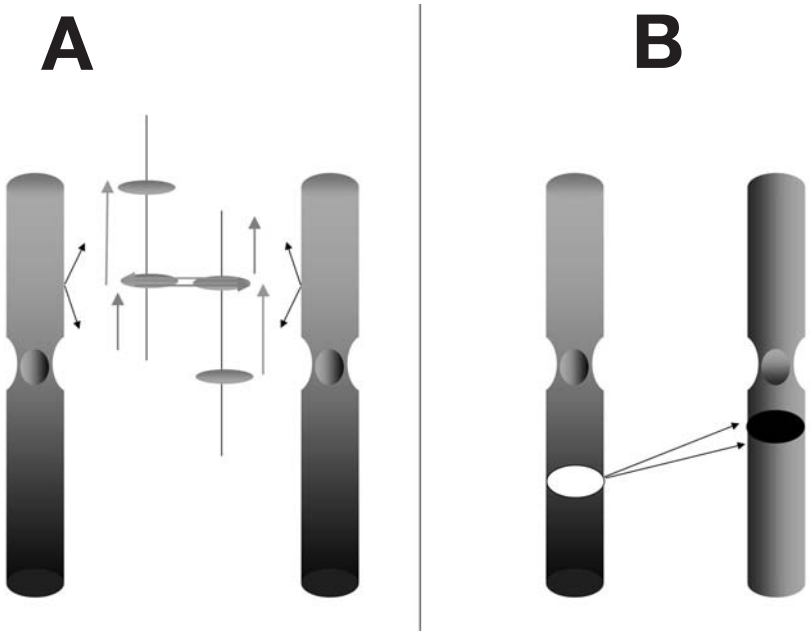


A. Part of the long arm of the right chromosome is missing. The loss of genomic material is called a deletion.
B. A part of the short arm of the chromosome is present twice (right). This extra material is called a duplication. As the duplicated region is localised within the chromosome, this duplication is called an interstitial duplication.
C. The amount of genetic material in part C of this picture is similar to the unaffected left chromosome. However, a part of the chromosome is inverted. As the centromere is localised within the invertion, this situation is called a pericentromeric inversion.
D. Again the amount of genetic material is normal, however, a part of the information of the dark grey chromosome has been transported to the light grey chromosome and vice versa. This is called a balanced translocation.

[See appendix: colour figures.]

been appreciated (Iafrate *et al.* 2004; Sebat *et al.* 2004; Fredman *et al.* 2004). Sebat *et al.* (2004) presented the first study assessing the frequency of CNVs in the healthy population using genome-wide screening tools. CNVs were shown to be frequent and, although they are present all over the human genome, loci enriched for structural rearrangements are not randomly distributed. Regions within or flanked by segmental duplications show a higher frequency of CNVs compared to regions outside these duplications. Furthermore, the genes that show enrichment in CNVs are also not random. Genes associated with immunity-, defence, cancer susceptibility, drug detoxification, signal transduction and sex hormone metabolism frequently show variations (Eichler 2006), including null-alleles. McCarroll *et al.* (2006) showed these variations to result in expression level differences, indicating that these variants are related to adaptation. On the other hand, the

Figure 2. Non-allelic homologous recombination and insertions.



A. Non allelic homologous recombination. The two alleles of a chromosome contain regions that are highly homologous (e.g. segmental duplications, low copy repeats or duplicons). The presence of these segmental duplications can result in misalignment of these regions and subsequently in non allelic homologous recombination. The green arrow shows the origin of a duplication of the region present between two highly homologous regions, whereas the red arrow indicates the origin of a deletion.
B. In this situation a part of the left chromosome is inserted in another chromosome. This is called an insertion.
[See appendix: colour figures.]

majority of deletions found thus far were located in so called gene-deserts (Conrad *et al.* 2006) and may therefore be neutral variants or have modest regulatory effects due to the presence of microRNA, noncoding RNA and other highly conserved regions.

Nearly half of all CNVs seem to be complex events, formed by more than one event (for example an inversion (figure 1C) and a deletion, or a deletion combined with a duplication) (Eichler unpublished data).

2.2 Segmental duplications

2.2.1. Characteristics of segmental duplications

Segmental duplications have been defined as sequences of DNA greater than 1 Kb in size sharing a homology of at least 90 % (She *et al.* 2006). Previous studies

indicate that at least 5% (154 Mb) of the human genome is composed of such duplications (Bailey *et al*. 2002a; Cheung *et al*. 2003b; She *et al*. 2004; Zhang *et al*. 2005), also called Low Copy Repeats (LCRs) or duplicons. Duplicons can have either a simple or a complex structure (Ji *et al*. 2000) and contain genes, pseudogenes, gene fragments, repeat gene clusters (Ford and Fried 1986) and other chromosomal segments (Eichler *et al*. 1996; Samonte and Eichler 2002b; Horvath, Schwartz, and Eichler 2000). Especially the pericentromeric regions consist of a mosaic of different genomic segments (Horvath, Schwartz, and Eichler 2000). Compared to the chimpanzee and baboon, the human genome is particularly enriched for the number and the length of mainly *Alu* repeats (Liu *et al*. 2003). Also, the degree of genome sequence identity is higher in humans compared to other vertebrates (She *et al*. 2006).

Misalignment between segmental duplications followed by Non Allelic Homologous Recombination can result in a duplication and reciprocal deletion of the sequence flanked by these duplicons (figure 2A). However, the high degree of sequence homology between segmental duplications alone is not sufficient for providing 'repetitive breakpoints events', and therefore additional conditions are needed before recombination occurs. These include minimum length of 100% homology required for recombination in human mitosis and meiosis (minimal region of homology was estimated to be 220 – 300bp and 300 – 500 bp, respectively) (Lupski *et al*. 1992; Waldman and Liskay 1988), AT-rich sequences (Peoples *et al*. 2000), for example those present on both sites of a recombination hotspot in Smith Magenis Syndrome (Bi *et al*. 2003) and enrichment of *Alu* repeats near or within the junctions present in segmental duplications (Stoppa-Lyonnet *et al*. 1990; Potocki *et al*. 2000; Bailey, Liu, and Eichler 2003).

Segmental duplications are also largely responsible for the fact that a part of the human genome sequence working draft contains gaps or is misassembled. The higher the sequence similarity the more difficult it is to distinguish and correctly assemble LCRs (Eichler 2001b).

2.2.2. Intra- and interchromosomal segmental duplications

Segmental duplications can be divided in two categories, interchromosomal and intrachromosomal. Interchromosomal segmental duplications are based on the transposition of DNA sequences towards other chromosomes, whereas intrachromosomal segmental duplications originated from a sequence that is transported to another region within the same chromosome. The prevalence of intrachromosomal segmental duplications in humans is higher than interchromosomal segmental duplications (3.97%,

113.66 Mb versus 2.37 %, 67.86 Mb)(Samonte and Eichler 2002b; Cheung *et al*. 2003a; She *et al*. 2006).

Interchromosomal segmental duplications are frequently found at pericentromeric and subtelomeric sites (Cheung *et al*. 2001). An example is the pericentromeric region of the short arm of chromosome 16, which contains four different segmental duplications that were duplicated and subsequently transposed from Xq28, 15q13, 2p11 and 14q32 (Ji *et al*. 2000) towards 16p11.

While studying the olfactory gene family, which is spread over several chromosomes, (Trask *et al*. 1998) found that there are differences in subtelomeric segmental duplications between different ethnic groups, suggesting that such rearrangements are still ongoing.

I-3. CNVs with phenotypic trait: genomic disorders

3.1. Genomic disorders

Genomic disorders were defined in 1998 (Lupski 1998) as the clinical condition, all types of phenotypic features included, that result from the dosage alteration of gene(s) located within a rearranged segment of the genome. It was estimated that about 0.7-1 / 1000 live births suffer from a genomic disorder (Ji *et al*. 2000). Different types of CNV are involved in genomic disorders, e.g whole, and partial chromosome alterations (see **section 4**). These alterations include deletions, duplications, inversions, insertions and translocations (see figure 1 and figure 2). Three clinical conditions frequently arising from such CNVs are discussed below.

3.2. Mental retardation (MR)

MR or developmental delay (DD) is defined as a significant impairment of cognitive and adaptive functions (Battaglia and Carey 2003). It is a clinically important condition as it affects about 1:30 – 1:50 people. MR can be categorised into four degrees of severity (WHO 1980, International classification of Impairments, disabilities and handicaps. Geneve: World Health Organisation, 1980): Mild MR (intelligent quotient (IQ) between 50 and 70), moderate MR (IQ between 35 and 50), severe MR (IQ between 20 and 35) and profound MR (IQ below 20).

Both genetic - and environmental factors can contribute to the origin of mental retardation. Environmental factors can involve pre- peri- and postnatal events, such as oxygen deprivation (perinatal event), infection (prenatal, postnatal), teratogenic

influences (prenatal) (Hamel 1999. X-linked MR. A clinical and molecular study (Alkmaar: Dekave)).

Genetic causes for mental retardation include (1) chromosomal causes such as aneuploidies, chromosome end rearrangements, rearrangements in regions related to microdeletion syndromes and other interstitial rearrangements, (2) complex disorders (caused by mutations in multiple genes) and (3) monogenic disorders (**section 4.2**.). A substantial number of point mutations have been identified in isolated genes that play an important role in early development (Petrij *et al*. 1995), such as mutations in the *RAI1* (Slager *et al*. 2003) causing Smith Magenis syndrome, mutations in the *CREBBP* gene (responsible for Rubinstein Taybi syndrome) and the CTG expansion of the *FMR*-gene which accounts for about 1:4000 – 1:6000 male cases of mental retardation (Fragile X syndrome) (Murray *et al*. 1996; Turner *et al*. 1996; De Vries *et al*. 1997) (**section 4.2.**).

It is known that the causes of mental retardation vary with the severity of the condition. Large CNVs are more frequently associated with severe cases. Chromosomal and genetic disorders account for 30%- 50% of moderate to severe mental retardation (I.Q.< 50); environmental insults explain a further 10%-30% (Gustavson, Holmgren, and Blomquist 1987; McDonald 1973; Elwood and Darragh 1981; Flint and Wilkie 1996). In mild mental retardation cases (I.Q. between 50 and 70), approximately equal proportions of genetic and environmental causes are diagnosed, about 10-30% each (Lamont and Dennis 1988; Bundey, Thake, and Todd 1989; Einfeld 1984).

The cause of MR remains unclear in about 40-50% of cases, indicating that, despite its high prevalence, the pathogenesis of MR is poorly understood. It is expected, however, that this rather high percentage will decline with the use of recently developed high-resolution genome analysis (see **section 6.2. and 6.3.**).

3.3. Congenital Malformation (CM)

Along with mental retardation, CNVs in the human genome may also result in a wide range of congenital malformations, such as organ and skeletal defects. These clinical features are already present at birth, before the mental retardation becomes apparent, so these entities can be the first indication of a genetic defect. The presence of more than one CM in a newborn that lacks a characteristic pattern of a specific microdeletion syndrome is an indication for genome-wide screening for CNV.

I-4. CNVs with phenotypic trait: Different types of variations

4.1. Whole chromosome variations

Since it was shown that an extra chromosome 21 causes Down syndrome (LEJEUNE, TURPIN, and GAUTIER 1959; Jacobs *et al.* 1959), it became clear that aneuploidy has significant influence on early development as well as on the intellectual capacities of an individual. Moreover, the severity of congenital malformations associated with trisomy 13 or 18 is such that only a small percentage of these fetuses will be viable with a drastically reduced life expectancy. Complete aneusomies of the remaining autosomal chromosomes have not been reported among live births, indicating that these are not compatible with life. Studies on material from spontaneous abortions support this statement (Carr 1971; Lauritsen *et al.* 1972; Boue and Boue 1977).

The fact that cells use one copy of the X chromosome while inactivating extra copies, combined with the small number of genes on the Y chromosome results in the less severe impact of sex chromosomes aneuploidies on the development of the embryo. Karyotypes such as 45,X, 47,XXX, 47,XXY, 47,XYY constitute the most common class of chromosome abnormality in humans (Hall, Hunt, and Hassold 2006).

Incomplete aneusomies of autosomal and sex chromosomes (chromosomal mosaicisms) are also known to be present in both affected and healthy individuals. The phenotypic consequence of a chromosomal mosaicism depends on the chromosome involved, the percentage of abnormal cells and the tissue(s) that contain cells with an abnormal chromosomal constitution.

Some of the whole chromosome variations originate from Robertsonian translocations in one of the parent of the affected fetuses / newborn. The frequency of Robertsonian translocations is 1:1000 (Shaffer and Lupski 2000).

4.1. Partial chromosome variations

4.1.1. Subtelomeric CNVs

The subtelomeric regions are localized proximal to the telomere proper, which consists of short repetitive sequences that cap the end of the chromosome. The subtelomeric regions from different chromosomes are highly variable, with some having a simple pattern and little similarity to other chromosome ends, whereas others contain complex and extensive patterns of homology. A good example regarding similarity of two subtelomeric regions is 4q and 10q, both encompassing repeats that share >98% sequence homology (van Overveld *et al.* 2000; van Geel *et al.* 2002). The subtelomeres are particularly dynamic regions, due to repeat-rich sequences that have a high frequency

Table 1. Overview of subtelomeric screening studies in chronological order. Based on Rooms *et al.* (2004a) with addition of more recent publications.

Reference	Method of analysis	Number of cases	Detection rate
Flint *et al.* (1995)	VNTR marker analysis	99	3%
Knight *et al.* (1999)	Multiprobe FISH	284 moderate/severe	7.4%
		182 mild	0.5%
Slavotinek *et al.* (1999)	Microsatellitemarker analysis	27	7.5%
Bonifacio *et al.* (2001)	PRINS	65	3.1%
Borgione *et al.* (2001)	Microsatellitemarker analysis	60	6.6%
Colleaux *et al.* (2001)	Microsatellitemarker analysis	29	6.9%
Fan *et al.* (2001)	Multiprobe FISH	150	4%
Riegel *et al.* (2001)	Multiprobe FISH	254	5%
Rosenberg *et al.* (2001)	Microsatellitemarker analysis	120	4.1%
Rossi *et al.* (2001)	Multiprobe FISH	200	6%
Sismani *et al.* (2001)	Multiprobe FISH / MAPH	70	1.4%
Anderlid *et al.* (2002)	Multiprobe FISH	111	9%
Baker *et al.* (2002)	Multiprobe FISH	53 isolated MR	1.9%
		197 MR and dysmorphic features/malformations	4.1%
Clarkson *et al.* (2002)	Multiprobe FISH/ SKY	50	6%
Dawson *et al.* (2002)	Multiprobe FISH	40	10%
Hélias-Rodzewicz *et al.* (2002)	Multiprobe FISH	33	9%
Hollox *et al.* (2002)	MAPH	37	13.5%
Popp *et al.* (2002)	M-TEL	30	13.3%
Rio *et al.* (2002)	Microsatellitemarker analysis	150	10%
Van Karnebeek *et al.* (2002)	Multiprobe FISH	184	0.5%
Hulley *et al.* (2003)	Multiprobe FISH	13	7.7%
Jalal *et al.* (2003)	Multiprobe FISH	372	6.8%
Bocian *et al.* (2004)	Multiprobe FISH	59 moderate-severe	10%
		24 mild	12.5%
Harada *et al.* (2004)	Array CGH	69	5.8%
Koolen *et al.* (2004)	MLPA	210	6.7%
Kriek *et al.* (2004)	MAPH	184	4.3%
Pickard *et al.* (2004)	MAPH / FISH	69 mild	1.5%
Rodriguez-Revenga *et al.* (2004)	Multiprobe FISH	8 moderate-severe	12.5%
		22 mild	4.5%
Rooms *et al.* (2004b)	Microsatellitemarker analysis	70	-
Rooms *et al.* (2004a)	MLPA	75	5.2%
Walter *et al.* (2004)	Multiprobe FISH	50	10%
Novelli *et al.* (2004)	Multiprobe FISH	92	16.3%
Li and Zhao (2004)	Multiprobe FISH	46	4.4%
Rooms *et al.* (2006)	MLPA	275	4.4%
Lam *et al.* (2006)	MLPA / multprobe FISH	20	15%
Palomares *et al.* (2006)	MLPA	50	10%
	Multiprobe FISH	50	10%

of recombination. They are also gene- rich, and the plasticity of these chromosomal regions may be one of the factors responsible for phenotypic diversity (Mefford and Trask 2002).

CNVs near the chromosome ends are a significant cause of idiopathic mental retardation (Flint *et al*. 1995; Knight *et al*. 1999; Flint and Knight 2003). Flint *et al*. (1995) demonstrated that ~6% of the patients with idiopathic mental retardation have a rearrangement in a subtelomeric region. These findings were verified by observations in many other studies. Biesecker (2002) and later Rooms *et al*. (2004a) summarized subtelomeric aneusomy screening studies using various detection methods (table 1). In our study, **(chapter II-1)** 4.3% subtelomeric alterations were found among 184 idiopathic mild to severe MR patients.

The percentage of aberrations detected varies considerably between different studies. This is due to the different criteria for the selection of patients, different techniques used, and, in smaller patient groups, by stochastic factors. It seems that the number of CNVs detected goes up with increasing complexity and severity of the clinical problems of the patients.

A proportion of the subtelomeric imbalances originate from reciprocal translocations in one of the parents. The frequency of reciprocal translocations is 1:625 (Shaffer and Lupski 2000). All chromosomes seem to participate in reciprocal translocations and most of the breakpoints are family-specific, however some breakpoints are recurrent, such as t(11;22)(q23-q11.2) and t(4;8)(p16;p23) (Giglio *et al*. 2002). These common and recurrent breakpoints originate from misalignment between interchromosomal duplicons, which can lead to crossing over between non homologous chromosomes (Kurahashi *et al*. 2000; Kurahashi *et al*. 2003).

Gribble *et al.* (2005) studied a group of patients with a phenotypic trait and who had initially been diagnosed to have a balanced translocation based on the outcome of karyotyping. The majority of these apparent balanced translocations appeared to consist of several complex rearrangements often combined with the presence of one or more imbalances. To gain more insight in different 'balanced' translocations and their consequences, Danish investigators started to collect and characterize large numbers of balanced chromosomal rearrangements (Bugge *et al.* 2000).

4.1.2. CNVs in microdeletion syndromes regions

Microdeletion syndromes result from the loss of several genes (contiguous gene syndrome) or may result from the loss of a single gene. The majority of the microdeletion related regions are localised between intrachromosomal segmental duplications. These

Table 2. Characteristics of syndromes flanked by duplicons (recombination hotspots) of which the reciprocal alteration has also been identified to have clinical consequences.

Localisation	CNV	Genomic	Size of duplicon (kb)	Size of CNV (Mb)	Freq.	References
17p12	Del	Hereditary Neuropathy with liability to Pressure Palsy	24	1.5	1:20000	Reiter *et al*. (1996); Reiter *et al*. (1998); Inoue *et al*. (2001)
	Dup	Charcot-Marie-Tooth syndrome			1:2500	Valentijn *et al.* (1992); Pentao *et al.* (1992); Lupski *et al.* 1992; Lupski *et al.* (1991)
22q11	Del	DiGeorge - / Velo-CardioFacial Syndrome	200	3	1: 4000	Shaikh *et al.* (2000); Edelmann, Pandita, and Morrow (1999)
	Dup	22q11 duplication syndrome			Probably equal	Yobb *et al.* (2005) Ensenauer *et al.* (2003)
7p11.2	Del	Smith Magenis syndrome	250 - 400	5.0	1:25000	Bi *et al.* (2003); Slager *et al.* (2003) Shaw, Bi, and Lupski (2002)
	Dup	Potocki-Lupski syndrome			Probably equal	Chen *et al.* (1997) Potocki *et al.* (2000); Bi *et al.* 2003; Potocki *et al.* (2007)
7q11.23	Del	Williams syndrome	320	1.6	1:20000-50000	Bayes *et al.* (2003); Peoples *et al.* (2000) Urban *et al.* (1996); Francke (1999)
	Dup	Duplication of the Williams Critical region			Probably equal	Somerville *et al.* (2005); Kriek *et al.* (2006)

As reciprocal duplications have only been discovered recently, the frequency cannot be determined based on literature. Based on Non Allelic Homologous Recombination one can assume that the frequency of reciprocal duplication is equal to that of the corresponding deletion, although there is no reason to assume that the consequence of a deletion or duplication would be the same. Nevertheless, it seems that the frequency of HNPP is an underestimation. In addition to the duplication of the region involved in DiGeorge/VCF syndrome, tetrasomy of this 22q11 region has also been described in Cat eye syndrome. Del = deletion, dup =duplication, Freq. = frequency, CNV = Copy Number Variation.
This table was based on table 3 of Shaffer and Lupski (2000).

homologous regions facilitate unequal crossing over, resulting in deletions as well as duplications (Chance *et al*. 1994). This indicates that the frequency of reciprocal duplications of such regions is in principle equal to that of the corresponding deletions. In general, clinical phenotypes of these duplications are milder compared to the deletion of the same region (for references see right column of table 2), and some of these

duplications might not even result in MR. In addition, duplications used to be more difficult to detect compared to deletions. This explains the lower frequency of publications regarding micro- duplications within such regions. Examples of microdeletion syndromes that are flanked by duplicons include Hereditary Neuropathy with liability to Pressure Palsy (HNPP), Williams-Beuren syndrome, DiGeorge- / Velocardiofacial syndrome, Smith Magenis syndrome (see table 2), Angelman - /Prader Willi syndrome (Miller, Dykes, and Polesky 1988; Amos-Landgraf *et al.* 1999) (see table 2). Up to now microdeletion syndromes have been recognised by their distinctive clinical phenotypes, using targeted fluorescence in situ hybridisation (FISH) to detect the deletion in patients selected by a dysmorphologist. Recently, the genome-wide array-CGH method revealed additional microdeletions among MR patients that at first sight appeared to lack salient and distinct features. A recent example of such a microdeletion is the 17q21.31 microdeletion syndrome that is associated with parental inversion of this region (Shaw-Smith *et al.* 2006; Koolen *et al.* 2006; Sharp *et al.* 2006). After identification of the deletion, dysmorphologists do see common features in a series of patients, possibly enabling the recognition of these patients in the clinic.

4.1.3. Other interstitial CNVs

Several CNVs localised outside the subtelomeres and microdeletion related regions have been identified as being involved in the etiology of MR/CM.

Bailey *et al.* (2002) described a bioinformatic approach to analyse the human genome sequence, and identified nearly two hundred potential hotspots for CNVs, e.g. regions flanked by segmental duplications (Bailey *et al.* 2002a). Some of these regions appear to be related to genomic disorders. 130 of these regions were subsequently tested for rearrangements among 47 healthy individuals using a segmental duplicon BAC microarray (Sharp *et al.* 2005). 79 of the 130 potential CNV hotspots showed no alteration among this study population, supporting the hypothesis that alterations within these regions could be related to disease. **Chapter II-2** summarizes our results of screening for CNVs of regions flanked by intrachromosomal duplicons among 105 MR/CM patients. As expected, the rearrangement frequency per unit of DNA is much higher in regions flanked by duplicons compared to regions without known duplicons nearby, supporting the statement that regions flanked by duplicons are enriched for copy number variations. Of course, pathogenic CNVs outside duplicon-flanked_regions have also been identified, for example the interstitial deletion of chromosome band 2p16p21 (Sanders *et al.* 2003; Lucci-Cordisco *et al.* 2005) (see **chapter III-4**) and the *DMD* gene (Blonden *et al.* 1991; Nobile *et al.* 2002).

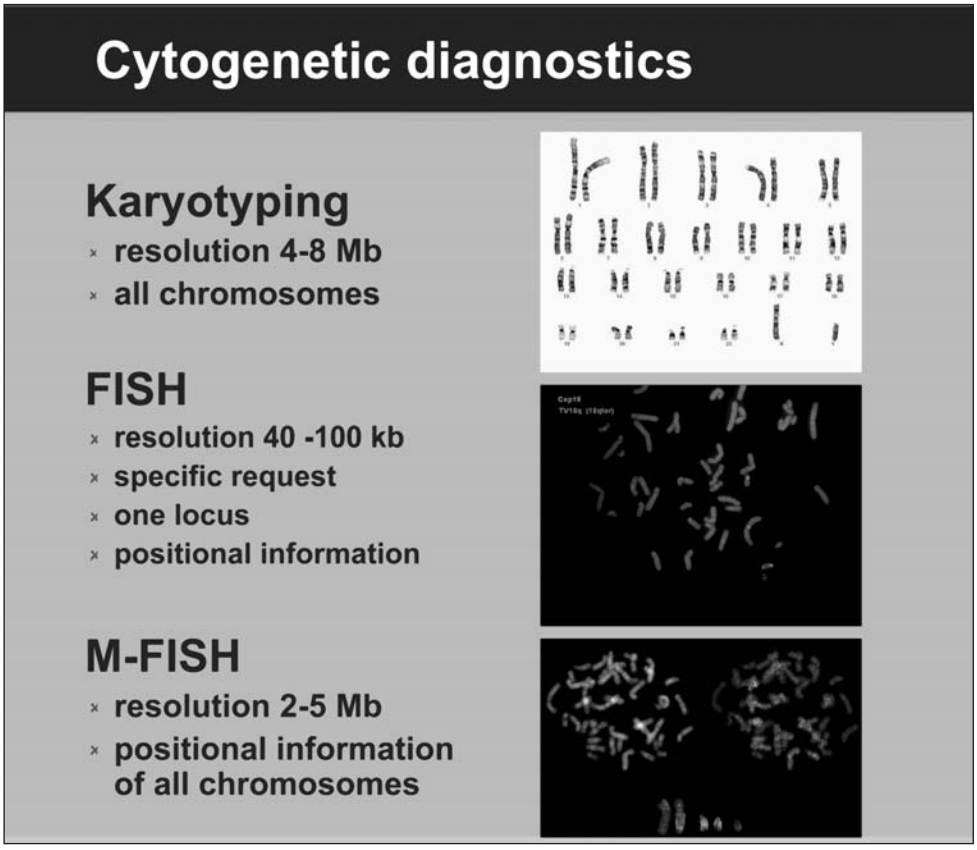
4.2. Other variations

Several microdeletion syndromes are in fact caused by the inactivation of a single gene. An example is the Rubinstein Taybi Syndrome (RTS). After two reciprocal translocations with a breakpoint in the short arm of chromosome 16 had been described in RTS patients, submicroscopic deletions were detected in six of a series of 25 patients with the syndrome (Breuning *et al*. 1993). Subsequent mutation detection using the protein truncation test identified two point mutations in the CREBBP gene in 16p (Petrij *et al*. 1995), indicating that RTS was not, as previously thought, a contiguous gene syndrome, but due to haplo-insufficiency of a single gene. Similarly, Smith Magenis syndrome was initially found to be caused by a microdeletion of chromosome band 17p11.2. Subsequently mutations in the *RAI1* gene were shown to be responsible for the vast majority of the clinical features associated with the syndrome (Slager *et al*. 2003). More recent examples of variants within a single gene that are found to related to a syndrome or a sequence include the gene for CHARGE sequence (Vissers *et al*. 2004) and the gene involved in Cornelia de Lange syndrome (Krantz *et al*. 2004). In 2006, the gene linked to Peters Plus syndrome was identified after finding two splice donor site mutations within the *B3GALTL* gene (**chapter III-2**). This year, Zweier *et al*. revealed that haplo-insufficiency of *TCF4* is responsible for the Pitt Hopkins syndrome (Zweier *et al*. 2007).

I-5. Considerations regarding pathogenicity of CNVs

The vast majority of the large CNVs related to genomic disorders are thought to be *de novo* (except for CNVs with an X-linked or autosomal recessive inheritance), as affected patients often have a severe phenotype and are unable to have offspring. However, for some microdeletion syndromes an autosomal dominant transmission has been documented (Leana-Cox *et al*. 1996; Morris, Thomas, and Greenberg 1993), emphasizing that even CNVs that are known to cause genomic disorders can demonstrate phenotypic variability. The pathogenicity of familial CNVs is often hard to interpret, as variable expression of the remaining allele and incomplete penetrance can influence the clinical consequences in different family members. An example is the phenotypic variability associated with a duplication of the DiGeorge- / Velocardiofacial syndrome region. Edelmann *et al*. (1999) described an individual with this duplication who was affected by failure to thrive, marked hypotonia, sleep apnoea and seizure-like episodes. The healthy mother and grandmother however also carried the same duplication. Ad-

Figure 3. Current standard cytogenetic diagnostic tools and their characteristics.



[See appendix: colour figures.]

ditional reports verified that this specific alteration, despite showing a very wide range of clinical features, is not a benign genomic variant (Ensenauer *et al.* 2003; Yobb *et al.* 2005). A second example includes the 1.5 Mb duplication of chromosome band 16p13.1 that has been recently found among four severe autistic male patients. The same duplication was detected among less affected and unaffected family members (Ullmann *et al.* 2007).

In general, the presence of a particular CNV in a patient as well as in family members does not exclude a causal relation with the clinical problem, since autosomal recessive, digenic, complex or multifactorial inheritance can apply. The identification of the gene responsible for Peters' plus syndrome (**chapter III-2**) is the perfect example to

underline the presence of an autosomal recessive inherited disorder. This syndrome was suspected to be an autosomal recessive disorder, although cryptic unbalanced translocations could not be excluded based on the presence of multiple spontaneous miscarriages in several families. We identified an interstitial deletion in two affected brothers that was also present in the mother and the maternal grandmother. The latest two were both suffering from breastcancer. Additional investigation of the brothers identified a mutation in the *B3GLTL* gene from the same region on the paternal allele.

A *de novo* variant is often assumed to be causative, however, since many CNVs are (neutral) polymorphisms, *de novo* variations can also be inconsequential. Van Ommen (2005) discussed the frequency of *de novo* deletions and duplications. He estimated a frequency of 1 in 8 for deletions, and 1 in 50 for duplications comprising random events in human newborns. It was noted that these are likely to be underestimates as, in addition, segmental duplicons cause recurrent non-random variations. Given, therefore, that *de novo* CNV is relatively frequent and not in all cases linked to genomic disorders, the finding of a *de novo* variation in a patient is not sufficient to conclude that this CNV is causally related to the clinical phenotype.

Recent initiatives, such as those of the Sanger Institute (www.sanger.ac.uk/PostGenomics/decipher/) and Ecaruca, to create platforms for collecting and comparing molecular cytogenetic data from many clinical genetic centers in relation to the human genome sequence, will assist in giving a better understanding of the role of CNVs in MR, CM and other genetic diseases.

I-6. Detection of CNVs

6.1. (Standard) Cytogenetic tools (figure 3)

6.1.1. Karyotyping

Analysis of chromosomes using the light microscope has been the gold standard for chromosome analysis during the past five decades. The banding technique, developed in the 1970s, enables the identification of specific chromosomes and large rearrangements (Caspersson, Lomakka, and Zech 1972; Yunis 1976). Using this technique, it became clear that chromosomes from healthy individuals are not completely similar. For each and every chromosome, microscopically visible variations not related to any phenotypic trait have been identified (Wyandt HE, Tonk VS (eds), 2004. Atlas of human chromosome heteromorphisms, Kluwer). These variants are called heteromorphisms.

Karyotyping has been implemented worldwide in a diagnostic setting, as it is very specific and reproducible in detecting large chromosomal variations among different groups of patients.

Even with optimal quality, however, it is not possible to identify structural imbalances smaller that 3-5 Mb (figure 3).

The implementation of the high-resolution banding (more than 800-band level) may not always resolve the resolution problem, as it can result in both false positive and false negative results (Kuwano *et al*. 1992; Delach *et al*. 1994; Butler 1995). An example of this was published by Francke *et al*. (1985). They described a patient suffering from Duchenne muscular dystrophy, chronic granulomatous disease associated with cytochrome b deficiency and with the McLeod phenotype in the Kell red cell antigen system and retinitis pigmentosa due to an interstitial deletion of part of band Xp21. This deletion could be identified by standard resolution chromosome banding. However, using higher resolution chromosomes, the loss of genetic material was very hard to appreciate. Flint and Knight (2003) also found a negative correlation between the resolution of the banding and the number of chromosomal alterations found. This phenomenon may be explained by the fact that high resolution banding uses chromosomes that are in the prometaphase stage. At this stage the condensation of the chromatids is incomplete, resulting in elongated chromosomes. Since the condensation process is ongoing and variable during pro-metaphase, apparent differences in length may be due to unequal condensation instead of a "real" difference caused by a gain or loss of genetic material.

6.1.2. Fluorescent in Situ Hybridisation (FISH) analysis

FISH analysis (Prooijen-Knegt *et al.* 1982; Landegent *et al.* 1985; Ried *et al.* 1990) (figure 3) is based on the hybridisation of a fluorescently labelled probe containing a sequence of several tens (cosmids) to hundreds of kilobases (Bacterial Artificial Chromosomes (BACs)/ P1 derived Artificial Chromosomes (PACs)) that is complementary to the region of interest. The fluorescently labelled sequences will bind to the genomic DNA, which is subsequently visualised under a microscope. The two types of FISH analysis commonly used in diagnostic procedures are (1) metaphase FISH, that uses cultured cells for analysis, and (2) interphase FISH, that does not require culturing of cells. The advantage of interphase FISH analysis is that it has a higher resolution, allowing the detection of small tandem duplications, whereas FISH using metaphase cells will often miss such duplications as the extra signal is overlapping the original signal. Furthermore, interphase FISH can be used for the detection of low-level mosaics as large numbers of cells can be scored. On the other hand, the advantage of metaphase

FISH analysis is that individual chromosomes are visible, providing positional information of the CNV.

Detecting CNVs using FISH analysis is only possible if the following criteria are fulfilled: (1) The CNV must be characterized by a specific phenotype, (2) this phenotype must be recognized by a specialist (for example clinical geneticist) and (3) a specific diagnostic FISH test must be available.

6.1.3. Fiber FISH

Fiber FISH refers to the analysis of extended chromatin fibers. It provides a higher resolution than conventional FISH, because the chromosomes are analysed as distinct single threads under the microscope. Fiber FISH can also be used to resolve complex rearrangements. The principal drawback of this approach is that it is technically challenging and time consuming (Wiegant *et al.* 1992; Florijn *et al.* 1995; Rosenberg *et al.* 1995; Giles *et al.* 1997; Raap *et al.* 1996).

6.1.4. Multi-probe FISH (M-FISH) and SKY (Spectral Karyotyping)

Multiple color FISH was first described in the late eighties (Nederlof *et al.* 1989; Nederlof *et al.* 1990; Dauwerse *et al.* 1992). In general, Multiprobe FISH and SKY (Schrock *et al.* 1997) provide recognition of many chromosomes simultaneously by labelling them with a distinct combination of fluorochromes (Fan *et al.* 2000; Speicher, Gwyn, and Ward 1996). By pooling cloned DNA fragments of a particular (part of a) chromosome, the FISH probe can 'paint' the chromosome or a region of interest. By combining different fluorophores in different proportions, chromosome specific colors can be generated (Tanke *et al.* 1999; Raap and Tanke 2006). This COmBined RAtio labelling or COBRA–FISH is particularly useful for the detection of balanced translocations or to determine the content of a marker chromosome. As shown in figure 3, the resolution of tools is better than that of karyotyping. COBRA-FISH was used for the screening of subtelomeres (Engels *et al.* 2003). By applying the subtelomeric COBRA-FISH method, it was possible to screen 41 subtelomeres (except for the p-arms of the acrocentric chromosomes), with BACs/PACs localised approximately 230 Kb from the telomeres, using only two hybridisations and four fluorochromes.

Knight *et al.* (1997) developed a multi-hybridisation protocol, using a slide divided into 24 small hybridisation chambers. By applying different dyes to label each chromosome arm, the slide can be used to perform FISH analysis for all subtelomeres in one assay (Flint and Knight 2003). As this approach is quite laborious and consequently the throughput is very limited, it is currently not used on a wide scale.

By applying karyotyping and (different applications of) FISH analysis, a significant number of chromosomal anomalies remain undetected. Therefore, there is a strong need for screening techniques with a higher resolution.

6.2. High resolution tools (not genome-wide)

6.2.1. History

As stated previously, the phenomenon of copy number variation has been recognised since the earliest days of human gene cloning. The first gene clusters cloned, those coding for the alpha and beta chain of haemoglobin were found to frequently undergo gross rearrangements, showing deletions as well as duplications. Some, but certainly not all, of the deletions appear to be related to crossing-over between repeat elements as described by Higgs *et al.* (1984). Herrmann, Barlow, and Lehrach (1987) were the first to identify a molecular basis for recombination across a large inverted duplication that resulted in duplicated and deleted regions. For their study, which was published in 1987, restriction fragment length polymorphisms of cloned regions combined with pulse field gel electrophoresis were applied.

Studying another gene cluster, using hybridisation analysis of labelled cosmid clone fragments, Groot *et al.* (1990) hypothesized that unequal intrachromosomal crossing-over might be a frequent event leading to multiple and variable copies of the amylase genes. This model was recently confirmed using array and Fiber FISH analysis (Iafrate *et al.* 2004).

This section will briefly describe several techniques used for the detection of CNVs.

6.2.2. Restriction fragment length polymorphisms

Restriction fragment length polymorphisms (RFLP) are detected by digestion of (amplified) DNA using endonucleases, which only cut in the presence of specific DNA sequences (the restriction sites). The restriction fragments are then separated according to length by agarose gel electrophoresis. Depending on changes within these sequences, the length of the fragments and thus the position of the corresponding gel bands differ between individuals. The result of RFLP may be enhanced by Southern blotting (see 6.2.3). Using RFLP analysis, it was possible to identify duplications or deletions of a certain region of the genome. For example, RFLP analysis was applied within the first series of randomly cloned DNA fragments for the detection of probes showing non-Mendelian segregation. Both missing and extra alleles were identified (E. Bakker, personal communications, 1983).

6.2.3. Southern blotting

For many years, Southern blot analysis followed by densitometry was the main assay that was utilized for the detection of CNVs in clinical molecular genetic laboratories. It was the first technique to analyse human DNA on a wider scale. The Southern blotting procedure (Southern 1975) could show differences in length of restriction fragments and was used to study single copy, as well as low copy repeat sequences. Quantitative analysis was also possible on a very limited scale. Presence or absence of a sequence was of course no problem, but even the difference between one or two copies of a fragment with similar length required optimal experimentation. In some cases a rearrangement within a gene could be visualised by finding a new junction fragment. Since the technique required the use of radioactive labels and is very laborious, it has become less popular and has been largely replaced by quantitative PCR- based techniques, such as Q-PCR and Multiplex Ligation dependent Probe Amplification (MLPA) (Schouten *et al.* 2002).

6.2.4. Pulse field gel electrophoresis (PFGE)

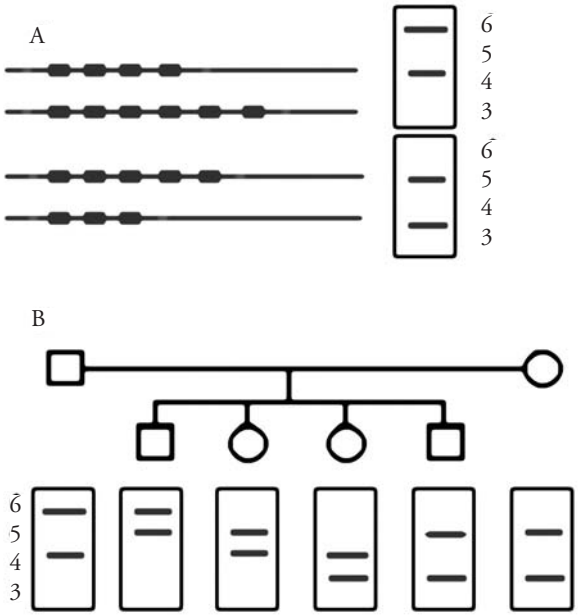
This technique (van Ommen *et al.* 1986; Den Dunnen *et al.* 1987) extends Southern blotting to include detection of very large DNA molecules (20 kb to several Mb in length) that are too large to be separated using normal agarose gel electrophoresis. It can be used to detect a rearrangement-specific junction fragment. Shearing of the genomic DNA is prevented by preservation and enzymatic digestion in solid agarose. The agarose-embedded DNA is cut by a rare-cutting restriction endonuclease and subsequently separated by an electrical current. During electrophoresis, the relative orientation of the electric field is periodically altered (Strachan and Read, Human Molecular Genetics, third edition, chapter 6.2). Fragments of different sizes will migrate at different speeds through the gel, and consequently PFGE is capable of detecting structural rearrangements.

Despite being technically challenging, is still used to study large repeat arrays e.g. FSHD (Buzhov *et al.* 2005).

6.2.5. Microsatellites for detecting CNVs

Microsatellites are sequences containing variable number of tandem repeats (hence are also known as variable number of tandem repeat markers (VNTRs). The number of repeat units for a given locus may differ between individuals, resulting in alleles of varying lengths. The differences in repeat length can be visualised either by using a nearby single copy probe on a Southern blot or by PCR-based methods. Allelic variation, the number of repeats, and allelic frequencies are available for thousands of markers across numerous

Figure 4. Identification of the parental origin of an allele.



A. Different VNTR lengths in both parents present on a specific region in the human genome.
B. One of the children has the identical combination of VNTR lengths as one of its parents. Uniparental disomy (of genetic material from the parent with identical VNTR lengths) or a deletion present at the allele inherited from the 'other' parent should be considered. Picture derived from www.geninfo.no.

[See appendix: colour figures.]

organisms. These polymorphisms can be used for the identification of CNVs by observing abnormal inheritance of parental alleles (figure 4), such as uniparental disomy. The limitation of this type of genetic marker for the detection of imbalances is that its success depends on the availability of parental DNA (Wilke, Duman, and Horst 2000).

All techniques described above have major disadvantages. They are either technically demanding, expensive, slow, require fresh samples, or have a low throughput (Heath, Day, and Humphries 2000). The major limitation is the small number of loci that can be tested in one experiment. The development of PCR based techniques, such as Multiplex Amplifiable Probe Hybridisation (MAPH) and Multiplex Ligation-dependent Probe Amplification (MLPA) allowed more widespread analysis of gene dosage.

6.2.6. Quantitative real-time Polymerase Chain Reaction (Q-PCR)

This method is independent of the availability of informative markers in the region of interest. Quantitation of input DNA is achieved by using dyes or dual-labelled probes, and a fluorescence scanner to monitor the amount of product generated during the amplification process. The method was originally designed to facilitate quantification of RNA, but it can also be used to quantify the copy number of a genomic sequence. The combination of real-time PCR and TaqMan TM fluorescent probes for the detection of CNVs has been described by Wilke, Duman, and Horst (2000) and Laurendeau *et al*. (1999). In this case, one only needs the amplification of one reference locus to measure the copy number of the test loci, instead of using different diluted DNA fragments for standardisation.

6.2.7. Towards MAPH and MLPA

In 1995, a PCR method was described which simplifies quantitative multiplex PCR (Shuber, Grondin, and Klinger 1995) where gene specific primers were tagged at the 5'end with an unrelated 20 nucleotide universal primer binding site. Based on this method, new applications of multiplex-PCR were designed such as quantitative fluorescent multiplex PCR (QFM-PCR) (Heath, Day, and Humphries 2000) that was published in the same year as Armour published another application, called Multiplex Amplifiable Probe Hybridisation, MAPH (see below). QFM-PCR, MAPH (**section 6.2.6**), MLPA (**section 6.2.7**.) are all useful, effective and reliable methods for the detection of both deletions and duplications in the same assay.

6.2.8. MAPH

MAPH was first described by Armour *et al*. (2000). MAPH is a PCR-based method for simultaneously determining the copy number of a set of up to 50 different chromosomal loci (White *et al*. 2002). The probes, usually exons from candidate genes, are individually cloned such that all can be amplified using one pair of primers. To detect copy number changes, the probes are hybridised to denatured genomic DNA that has been immobilised and cross-linked on numbered nylon filters. After stringent washing, only the probes that hybridise specifically to the complementary sequence on the genomic DNA will remain bound. These hybridised probes are recovered off the filters, quantitatively amplified using PCR and analysed. The initial publication used a radioactively labelled primer followed by separation on a slab gel. This was then exposed to a film, with the resulting bands being measured using densitometry. White *et al*. (2002) simplified the procedure by using a fluorescently labelled primer followed

by analysis using a 96 capillary sequencer. The yield, represented by peak height and area, is determined for each probe. Changes in probe yield correspond to changes in copy number of the sequence analysed, i.e. a deletion or duplication.

The first report of subtelomere screening in patients with MR using MAPH was from Sismani *et al.* (2001). In their study, a group of 70 mentally retarded individuals was screened, using multiprobe telomeric FISH assay and MAPH. One subtelomeric deletion was found and confirmed with an independent technique. It has to be mentioned, however, that not all the subtelomeric probes were informative.

It has been calculated previously (Hollox *et al.* 2002), that about 0.12% of the mentally retarded patients were reported to have false positive results (that is, MAPH analysis detected an alteration that could not be verified using an independent technique), using MAPH based screening of subtelomeres, suggesting that this technique is reliable for the detection of CNVs. Obviously, the percentage depends highly on thresholds applied in a certain study.

6.2.9. MLPA

MLPA is based on the ligation of two adjacently annealing oligonucleotides, followed by the quantitative PCR amplification of the ligated products (Schouten *et al.* 2002). The left half-probe is chemically synthesised. It consists of a unique sequence complementary to the locus of interest along with a sequence containing the primer-binding site common to all probes. The other half-probes consist of three parts. In addition to the parts present in the left half-probe, this right half-probe also contains a spacer sequence, responsible for the difference in length of the MLPA probes. As the size of the right-sided half-probe initially was designed up to 440 nt, it was not possible to synthesize this oligonucleotide. Therefore, M13 vectors were used carrying the spacer sequences. However, generating a right half-probe with a spacer requires a laborious and time consuming cloning step. Therefore, a modified protocol for designing probes was implemented (White *et al.* 2004). Using this protocol, the right half probe is also chemically synthesised followed by 5'phosphorylation. Each probe was designed to be of unique size, enabling easy differentiation. This alternative MLPA protocol significantly reduces the time necessary for MLPA probe design, however, the number of loci that can be tested by MLPA using one fluorescent dye is limited. A second (and even a third) dye can be used by designing probes with another primer binding sequence (White *et al.* 2004; Harteveld *et al.* 2005). In this way, it is possible to screen up to 60 loci in nearly 100 patients in one assay.

6.2.10. Data analysis of MLPA and MAPH

Several methods for data analysis have been described (Hollox *et al*. 2002; White *et al*. 2002) and analysis protocols are available at www.mlpa.com.

Besides analysing the result of MLPA and MAPH using either a polyacrylamide gel or through polymer-filled capillaries, both techniques can be adapted for an array- or bead based read out. This will increase the number of loci than can be tested simultaneously in one patient (Gibbons *et al.* 2006). To detect the amplified fragments, universal arrays can be designed using specific zip codes. These are spotted on the array, with the complementary sequences being incorporated into the probes. An added advantage of this approach is that the half probes used can have identical sizes, facilitating uniform amplification. Using the 3-Dimensional, Flow-Through Microarray Platform from PamGene, hybridisation time of the amplified fragments to their target sequences can be reduced to minutes. This technique has been used for the rapid detection of aneusomies, resulting in a gain in time of more than 60 hours compared to karyotyping (Kalf *et al.* in preparation).

The advantage of MAPH and MLPA compared to other techniques, including (multi-probe) FISH and array-CGH, is that the resolution of detection is limited only by the size of the probes used (100-500 bp). In addition, using specific probe design, it is even possible to detect point mutations using MLPA analysis.
Both MAPH and MLPA facilitate the parallel screening of large numbers of patients at many different loci in one experiment with rather cheap consumables.
A disadvantage of these methods is that they are not suitable for genome-wide screening.

6.3 Whole genome (high resolution) screening tools; recent genome approaches

6.3.1. Overview

Affordable, high-resolution, genome-wide approaches for DNA copy number analysis have been available for less than five years. In contrast to FISH, where small fragments of DNA are labelled and hybridised to genomic DNA (in the form of chromosome spreads), array-based approaches label the genomic DNA, which is then hybridised to small fragments of DNA.

Currently, there are two main formats, array-CGH and SNP-based arrays. Both are discussed in more detail below. For array-CGH, the probes used are (3K – 30K) genomic clones or up to 400K 60-mer oligonucleotides, with the size and number determining the resolution of analysis.

SNP arrays, containing 10K–1000K loci have recently proven to facilitate, in addition to genome-wide association studies, the detection of deletions and duplications (see section 6.3.4.). The resolution of the SNP arrays depends on the number of SNP loci present and on their coverage across the genome.

The coverage of the genome of all genome-wide mapping platforms is rapidly improving.

It should be noted that these tools can not be used to detect copy-neutral rearrangements like translocations, insertions and inversions.

6.3.2. Array-CGH using BAC clones

High-resolution comparative genomic hybridisation (CGH)-based micro-arrays (Solinas-Toldo *et al.* 1997; Pinkel *et al.* 1998; Snijders *et al.* 2001) were developed to increase the resolution of chromosome studies. The technique is based on immobilised DNA isolated from Bacterial Artificial Chromosome (BAC) clones that were amplified by either DOP-PCR (Telenius *et al.* 1992) or ligation-mediated PCR (Snijders *et al.* 2001). The amplified DNA, spotted on coated microscope slides by an arrayer, is usually present in triplicate enabling internal standardisation. Test and reference DNA are differently labelled by random priming to incorporate fluorescently labelled nucleotides, and subsequently mixed with Cot –1 DNA to block repetitive DNA sequences. After hybridisation for 16-24 hours, images of hybridised fluorochromes can be obtained. The resolution obtained with BAC-arrays depends on the genomic distance between the BACs spotted on the array and the size of the BACs (Snijders, Pinkel, and Albertson 2003).

Clinical applications of array-CGH using different subsets of the human genome have been published by several groups (Veltman *et al.* 2002; Rauen *et al.* 2002; Bruder *et al.* 2001; Rosenberg *et al.* 2006). Veltman *et al.* (2002) estimated, based on their results obtained by screening 20 patients with known cytogenetic abnormalities, that the incorrect positive result of the 3500 BAC-array is approximately 0.4%, whereas no abnormality was missed. Many papers have been published regarding findings of screening MR patients using BAC-array of ~3500 BAC DNA probes spaced at ~1 Mb density over the full genome (3K array) (table 3). De Vries *et al.* (2005), Vissers *et al.* (2005) and Koolen *et al.* (2006) presented the results of screening using a BAC array with 10 fold higher resolution (33000 BACs). BAC arrays are also widely used in cancer diagnostics (Snijders *et al.* 2003; Weiss *et al.* 2003). The genomic variation among 55 healthy individuals was also tested using array-CGH (Iafrate *et al.* 2004). This study found as many as 255 alterations that were suspected to be neutral variants.

BAC-based array-CGH has been very important for the initiation of genome-wide screening at high resolution. It has proven to be a reliable and reproducible technique. Recently, oligonucleotide-based arrays have become available. These arrays come in two types, 60-mer oligos (see section 6.3.3.) for the detection of small CNVs and shorter 25-mer oligos for SNP (see section 6.3.4.) detection.
In their latest versions, these arrays have an effective resolution below 10 kilobases. A disadvantage of array-based methods is that they are currently still rather expensive.

6.3.3. Array-CGH using long oligos

Examples of these arrays include Nimblegen and Agilent. The 60 nucleotide is longer than the sequence that is spotted on the SNP array. As a result, these oligo based arrays are not suitable for SNP analysis, however, they do give stronger signal intensity. Therefore, CNVs can be detected using solely the signal intensity.

In addition, as the location of the oligos is not limited to known SNPs, it is possible to analyse regions of the genome where no validated SNPs are available. This can be particularly important when looking at duplicated regions. The most recent Agilent micro array contains ~244.000 spots on the array.

6.3.4. SNP based arrays

The 25-mer probe arrays were originally designed to detect SNPs to be used in genome wide linkage and association studies. However, they were quickly used to estimate copy number changes by using both signal strength and allele scoring. Initial studies used the Affymetrix 10K array, which demonstrated the principle that the arrays could provide quantitative data (Herr *et al*. 2005). Subsequent work has taken advantage of higher resolution chips, currently up to 500-1000K (Komura *et al*. 2006). In practice, these arrays have an effective resolution below 10 kilobases, meaning that much smaller rearrangements can be detected compared to previous genome-wide technologies.

6.3.5. Comparing cross platform

Currently, there is no golden standard available to determine which platform, CGH-based or SNP-based, is the most accurate. It might be argued that high density SNP genotyping would be the most appropriate to implement for screening for copy number alterations, as this tool offers the simultaneous measurement of copy number changes and copy-neutral loss of heterozygosity (i.e uniparental disomy). On the other hand, SNP arrays have been selected based on criteria such as heterozygosity, being in Hardy-

Weinberg equilibrium. Although these features are important for association studies, where SNPs need to be informative, they are less critical for copy number analysis where even spacing is more important. Indeed, many regions prone to rearrangements (e.g. duplicons) are lacking or underrepresented on these arrays, as the associated SNPs did not meet the required quality criteria. This is in contrast to array-CGH in which the location of the oligonucleotides is not limited to known SNPs, and, therefore, it is possible to analyse regions of the genome where no validated SNPs are available. Indeed, the study of Redon *et al.* (2006) shows that in addition to the SNP-arrays, arrayCGH analysis is required to cover all CNV regions in the human genome, otherwise at least one third of the CNVs will be missed. New arrays of both Affymetrix and Illumina now close this gap by combining SNP- and non-SNP probes on one array.

Chapter III-4 attempts to compare different whole genome screening tools by applying them to four unrelated patients suffering from overlapping interstitial 2p deletions. Comparing cross-platform, we found that the localisation of both proximal and distal breakpoints was largely in agreement.

There have been few studies published screening MR patients with the new oligo-array platforms (table 3). Most studies described to date looked at either CNVs in healthy individuals (table 4) or the validation of techniques for detecting CNVs in patient populations. Using the 10K genechip of Affymetrix, seven known alterations with a size between 0.2-3.7Mb were not detectable due to insufficient SNP density in the regions involved (Rauch *et al.* 2004). Slater *et al.* (2005) were able to find all known alterations previously found by karyotyping, FISH or MLPA analysis using a ten-fold higher density (>110 K) SNP chip of Affymetrix, except for one duplication at the end of chromosome 9q. The same mapping tool was successfully validated by another group (Ting *et al.* 2006). The utility of the beadchip (SNP) array of Illumina, assaying 109,000 and 317,000 SNP loci, to detect chromosomal aberrations in samples bearing constitutional aberrations as well tumor samples at sub-100 kb effective resolution has also been described (Peiffer *et al.* 2006). In addition, summaries of different whole genome high resolution mapping tools have been published recently (Veltman 2006; Coe *et al.* 2007).

I-7. Scope of this thesis

The main aim of this thesis was to assess several new techniques for the detection of genomic rearrangements in patients with MR and / or CMs. In quick succession,

Table 3. A selection of studies using genome-wide screening tools to screen for CNVs in MR patients.

References	Methods of Analysis	Genome Coverage	Sample size	No. of dels. (*de novo*)	No. of duplications (*de novo*)	U.T	% Alterations (% *de novo*)
Vissers *et al.* (2003)	BAC arrays	3,500 BACs	20 MR patients	3 (2)	2 (1)	0	25% (15%)
Schoumans *et al.* (2005)	BAC array	2,600 BACs	41 MR patients + dysm. features	4 (4)	0	0	9.8% (9.8%)
Tyson *et al.* (2005)	BAC array	3,000 BACs	22 MR patients	1 (1)	2 (1)	0	14% (9%)
De Vries *et al.* (2005)	BAC array	33,000 BACs	100 MR patients	Many (7)	Many (3)	0	10% (10%)
Menten *et al.* (2006)	BAC array	3,500 BACs	140 MR patients	18 (11)	7 (3)	3	20% (10%)
Miyake *et al* (2006)	BAC array	2,173 BACs	30 MR patients	3 (1)	1 (1)	1(1*)	17% (10%)
Rosenberg *et al.* (2006)	BAC array	3,500 BACs	80 MR patients	12 (5)	6 (2)	2 (1*)	25% (10%)
Shaw-Smith *et al.* (2006)	BAC array	3,500 BACs	50 MR patients + dysm. features	7 (6)	5 (1)	0	24% (14%)
Ming *et al.* (2006)	Affymetrix gene chip	100K SNPs	10 MCA patients	2(2)	0	0	20% (20%)
Friedman *et al.* (2006)	Affymetrix gene chip	100 K SNPs	100 MR patients	8 (8)	(3) (1 was a mosaic)	0	11 (11%)
Sebat *et al.* (2007)	ROMA	85,000 oligos	195 autistic patients	12 (12)	3 (3)	0	7,7% (7,7%)

This table summarizes the eight studies screening MR patients using BAC arrays, and three studies screening a MR or autistic study population using oligo based arrays. Based on the data presented in this table, it shows that, independent of the sample size tested, the number of *de novo* alterations detected using whole genome screening tools is around 10%. It is noteworthy that although the number of loci tested using a BAC-array is increased significantly compared to the initial BAC-arrays, the number of *de novo* alterations detected remains 10%. The same holds true for the implementation of the 100K SNP array.
*: one of the parents is a carrier of a balanced translocation. Affy: SNP array designed by Affymetrix, ROMA: representational oligonucleotide microarray analysis, dels: deletion, U.T.: unbalanced translocation, dysm.: dysmorphic

MAPH, followed by MLPA, and MLPA in combination with array-CGH, have been implemented to expand the possibilities for diagnostic screening for deletions and duplications. By applying these high-resolution techniques, new regions and genes involved in the etiology of MR/CM were identified, resulting in an increased number of patients with a known cause for their developmental disorders. Currently, using the new genome-wide high(er) resolution techniques, such as the oligo based array, the number of variations detected in the human genome will increase even further. At this

Table 4. The results of screening for CNVs among healthy individuals using different whole genome screening tools.

References	Methods of Analysis	Genome Coverage	Sample size	Total No of CNVs
Iafrate *et al.* (2004)	BAC array	5,264 BACs	55 healthy individuals	255
Sebat *et al.* (2004)	Oligo based array (ROMA)	85,000 oligo nt	20 healthy individuals	221
Conrad *et al* (2006)	Mendelian errors	1,3 million genotyping assays	180 healthy individuals (3* 60)	586
Mc Carrol *et al.* (2006)	Clustered genotype & Mendelian errors (Hapmap data)	1,3 million genotyping assays	269 healthy individuals	541
Komura *et al.* (2006)	Affymetrix gene chip	500 K	270 healthy individuals	1,203
Redon *et al.* (2006)	Array-CGH & affymetrix gene chip	26,574 clones 500 K	270 healthy individuals	1,447

moment, the consequence of the detection of a CNV in an affected individual is not always clear. Therefore, the main challenge will be determining whether a variation is related to disease or one of the many neutral polymorphisms.

I-8. In summary

The following two chapters contain seven papers. Chapter II includes three studies where groups of patients were tested for CNVs. The frequency of subtelomeric alterations as well as interstitial variations in and outside duplicons were determined among different groups of mentally retarded patients. We were able to report the second patient with the reciprocal duplication of the Williams syndrome critical region and a previously undescribed duplication within the 16p13.1 region. In addition, based on our findings using parallel testing of both MLPA- and array based analysis, an alternative, cost effective approach is recommended for screening mentally retarded patients. Chapter III is comprised of four studies using small numbers of patients and a case report. The first report describes a complex rearrangement on both copies of chromosome 22. Different characteristics of the rearrangements were defined using different diagnostic tools. We found that haplo-insufficiency of the Cat eye critical region is probably not related to a clinical phenotype. The phenotypic variability in relation

to the size of the deletion of patients having the ATR-16 (α-thalassemia retardation-16) syndrome was explored in the next paper. It was concluded that in MR patients showing microcytic (= small cell) hypochromatic anemia, the presence of ATR-16 syndrome should be excluded.

Thirdly, we were able to unravel the etiology of the Peters Plus syndrome, an autosomal recessive inheritable disorder, using a genome-wide screening tool. Finally, four high resolution genome-wide mapping tools were compared using four patients with an overlapping interstitial 2p deletion.