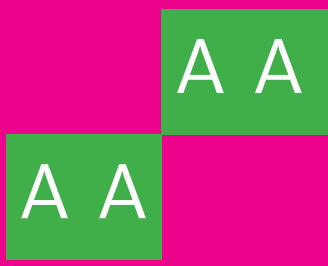
New techniques to detect genomic variation



Emmelien Aten

New techniques to detect genomic variation

Emmelien Aten

ISBN: 978-94-6191-547-4

Cover design & lay-out: Esther Beekman (www.estherontwerpt.nl)
Printed by: Ipskamp Drukkers BV, Enschede, The Netherlands

New techniques to detect genomic variation

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 7 februari 2013
klokke 16.15 uur

door

Emmelien Aten
geboren te Amersfoort
in 1978

Promotiecommissie

Promotores:	Prof. dr. M.H. Breuning Prof. dr. J.T. den Dunnen
Overige leden:	Prof. dr. M.H. Vermeer Prof. dr. R.C.M. Hennekam (Academic Medical Center, Amsterdam) Prof. dr. J.H.L.M. van Bokhoven (Radboud University Medical Center, Nijmegen)

The research presented in this thesis was performed at the Department of Human and Clinical Genetics, Leiden University Medical Center.


T C C G A G T T C C C T G G A
T C C G A G T T C C C T G G A
G T T C C T T C C G A G T T C
A A T G A G A A T C C G C C G
G T G A G A G C C C C G T C T
G T A C C T A C T G A G T T C
G A G T T A T G G T T C C T T
T C C G A G T T G T A A T T T
A A A T T T G A A A T C T G G
T G C C T A C T G A G T T C C
G T G C C T A C T G A G T T C
G A G C C C C G T C T G G T A
G T T C C T T C C G A G T T C
G G T T C C T T C C G A G T T
T T C C T T C C G A C T T C C

Table of contents


1

General Introduction

General Introduction

The human genome consists of ~ 6 billion base pairs, which is divided over 23 pairs of chromosomes. It is estimated that there are 20000-25000 protein coding genes. The DNA sequence in our genome is on average 99.9% identical to any other human being [1]. The more closely related two people are, the more similar their genomes. However, every human being is genetically unique and variations in composition and structure of the DNA are found throughout the genome. Human genetic variation refers to genetic differences between individuals and is important for diversity in a population. Without genetic variability a population cannot adapt to changes in the environment. The differences in genotype can cause differences in phenotype with subtle or sometimes quite obvious effects. Despite the large amount of variation in the human genome, most sequence variants have no obvious functional consequences. Determining if a specific variant has an effect is an important part of genetic analysis. The identification of genomic variation in large numbers of individuals helps to distinguish neutral variants (not involved in disease, or 'non-pathogenic' variants) from variants disrupting gene function (involved in disease, or 'pathogenic variants'). DNA analysis of patients with Mendelian disorders has resulted in the identification of a broad range of variants in genes, from definitely pathogenic mutations, to unclassified variants, to neutral polymorphisms. The relation between genomic variation and complex and quantitative human traits (i.e. obesity, height, multifactorial disease) has also been studied extensively.

As new methods for studying DNA are developed, different types of genomic variation have been discovered. Detection of large numbers of unclassified variants (UVs), with unclear significance for disease, have further emphasized the importance of cataloging genomic variation and studying the functional effects. The major challenge for molecular geneticists, cytogeneticists, clinical geneticists, and genetic counsellors is assessing the impact of all types of genomic variation on monogenic and complex disorders, along with the effective communication of findings to counselees.

Genomic variation; definition of sequence variation and structural variation

Genomic variation is a general term that covers all types of possible DNA variants, ranging from alterations affecting entire chromosomes to single nucleotide changes.

Differences in classification and terminology of genomic variation often cause confusion, therefore recommendations for the description of DNA variants have been proposed by the Human Genome Variation Society (HGVS).

Two main groups of variation can be identified; 1) sequence variation and 2) structural variation. An overview of the different types of variation, definitions and the spectrum in

which they operate are given in Table 1a+b. Examples are shown in Figure 1a+b.

Sequence variants can be classified as single or multi-nucleotide changes and range from single nucleotide differences to 1 kilobase (kb)-sized changes to full chromosome or even full genome changes of a segment of DNA [2]. Single nucleotide changes affect only one base pair, whereas multi-nucleotide changes are changes in a stretch of nucleotides. Variation in sequence include substitutions, insertions, deletions, duplications, and inversions. Indels are more complex changes and can be regarded as a combination of single nucleotide changes and multi-nucleotide changes. Sequence variations include "mutations" and "polymorphisms" and are usually referred to as single nucleotide polymorphisms (SNPs) or single nucleotide variants (SNVs). In some disciplines the term **"**mutation**"** is used to indicate "a change", while in other disciplines it is used to indicate "a disease-causing change". Similarly, the term **"**polymorphism**"** is used both to indicate "a non-disease-causing change", a change involved in human traits or a change found at a frequency of 1% or higher in the population [3]. To prevent this confusion, neutral terms such as **"** variant**"** or "alteration" are preferred as they are less ambiguous [4].

Structural variation is the genetic variation in structure of an organism's chromosome. It is generally defined as a region of DNA 1 kb and larger in size. As the resolving power of genetic analysis has increased, the focus of structural variation has shifted from entire chromosomes (in the prebanding chromosome era), parts of chromosomes (in the G banding era) to kilobases (using restriction enzymes, DNA probes and the Southern blotting). Sequencing techniques have shown that structural variants also include much smaller events, and overlap with the spectrum of sequence variation. Structural variation includes cytogenetically detectable and submicroscopic types of variation, such as deletions, duplications, insertions, inversions, translocations, indels and transpositions. Deletions and duplications, collectively referred to as copy-number variation (CNV), are a subset of structural variation and result in variable copy numbers of copies of specific DNA sequences [5-11]. CNVs are a major source of human genetic variation. Over 10,000 distinct CNVs have been described, ranging in size from kilobases to megabases. Most CNVs in humans are <50 Kb in size [12]. CNVs can be defined as recurrent or nonrecurrent, depending on their mechanism of formation [13]. Many recurrent CNVs are flanked by segmental duplications (also called low copy repeats; regions of DNA >1 kb present more than once in the genome with copies which are >90% identical) and are of a fixed size [9, 14]. Because these repeated sequences tend to misalign during meiosis, the resultant rearrangements tend to recur, creating clusters of variants with common endpoints. CNVs at these loci arise, by a mechanism named nonallelic homologous recombination (NAHR). Nonrecurrent CNVs, in contrast, which are not flanked by low copy repeats but other DNA elements (ALU elements or other repetitive elements), are of variable size and are thought to arise via

mechanisms like nonhomologous end joining (NHEJ) and replication-based mechanisms such as fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication models (MMRDR) [15]. It is becoming clear that most disease-causing CNVs are nonrecurrent, and generally arise via replication-based mechanisms.
Structural variants are associated with repetitive DNA, making accurate characterization more difficult [16]. Systematic assessment of structural variation in our genome has been difficult, primarily due to lack of appropriate methods for analysis. As such, the nucleotide resolution architecture of most structural variants remains unknown.

Table 1a: Description of variation types based on HGVS definitions

Types of variation	Description
Substitution	One nucleotide is replaced by another nucleotide
Deletion	One or more nucleotides are removed
Duplication	A copy of one or more nucleotides is inserted elsewhere in the genome
Tandem duplication	A copy of one or more nucleotides, directly following the original sequence
Insertion	One or more nucleotides are inserted between two original nucleotides but the insertion is not a tandem duplication
Insertion-deletion *(Indel)*	One nucleotide is replaced by more than one other nucleotide or More than one nucleotide is replaced by one or more other nucleotides
Inversion	More than one nucleotide is the reverse complement of the original sequence and replaces the original sequence
Translocation	The sequence of one chromosome interchanges with the sequence of another chromosome
Transposition *(interchromosomal insertion)*	The sequence of one chromosome inserts into another chromosome
Copy Number Variation	Submicroscopic deletions and duplications, which are losses or gains of DNA segments
Conversion	A range of nucleotides replacing the original sequence and are a copy of a sequence elsewhere in the genome

Table 1b: The spectrum of variation in the human genome

Variation	Type	Size Range
Single base pair change	Substitution, deletion, insertion, duplication	1 bp
Multiple base pair changes	Insertions, deletions, duplications, inversions, indels	Up to 1 kb
Full chromosome changes	aneuploidy	Entire chromosome/genome
Fine and intermediate scale structural variation	Insertions, deletions, duplications, inversions, indels, transpositions	1 kb-50 kb
Large scale structural variation	Insertions, deletions, duplications, inversions, indels, transpositions	50 kb-5 Mb
Chromosomal variation	translocations	~ ≥5 Mb

Light grey = sequence variation, black= structural variation. The operational spectrum partially overlaps with respect to size range.

Figure 1a: Types of variation in the human genome (sequence view). These can be single base changes (substitution, deletions, and insertions) or involve larger segments of DNA (deletions, insertions, inversions, indels, and duplications). Adapted from Frazer et al [17]

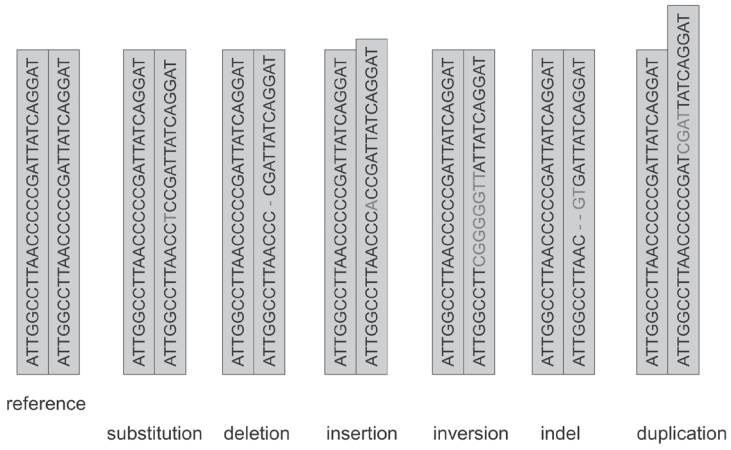
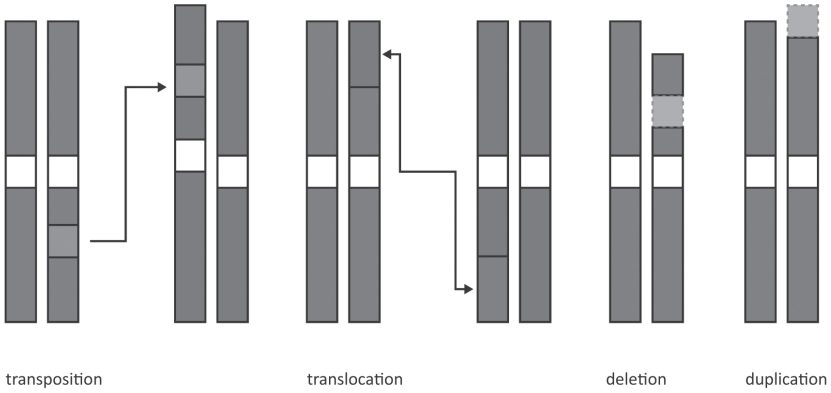


Figure 1b: Types of variation in the human genome (chromosome view). Examples of structural variation. 1) transposition-transfer of segment of DNA to a new position. 2) translocation-balanced event where two DNA segments are interchanged. 3) Copy number variation: deletion and duplication-loss or gain of DNA segments.



Genomic variation in the general population

Genomic variation refers to alterations at the DNA-level. Variation in DNA occurs due to malfunction of DNA replication during cell division or if DNA repair mechanisms fail after DNA damage induced by chemicals or radiation. These are random and spontaneous changes. Recent parent to offspring studies determined the human mutation rate (the rate at which variation occurs) at 1×10^{-8} per base per generation [18].

Population genetics is based on the study of genomic variation in natural selection. If a variant increases fitness it will undergo positive selection, and eventually be conserved in the genome. In contrast, a variant that has a negative effect will be selected against, and eventually lost. Our genome mutates spontaneously and randomly; mostly neutral, sometimes detrimental and, very rarely, beneficial. Ultimately, a gene pool arises which is best adapted to the environment. Most common genetic variants arose once in human history, and are shared by many individuals today through descent from common ancestors. Most analysis estimate that SNVs occur 1 in 1000 base pairs, although they do not occur at a uniform density. On average, every human being has 3 million nucleotide differences (SNVs) with any other human. Everyone is genetically unique. Even monozygotic twins, derived from the same fertilization event, have infrequent genetic differences due to early somatic changes (somatic variants) [19, 20]. There is also significant genetic difference between individuals from different ethnic backgrounds and numbers may increase if indels and structural variation is taken into account [21]. It has become apparent that human genomes

differ more as a consequence of structural variation than single base-pair differences [9]. In 2010 the human genome contains an estimated 15 million SNVs, 1 million short insertions and deletions and 20000 structural variants [1].

CNVs have been recognized as a common form of genomic structural variation. High resolution microarrays and sequencing approaches are able to identify 600–900 CNVs in a single individual [22, 23]. Approximately 65% to 80% of individuals carry a CNV that is at least 100 kbp in size, 5 to 10 % of individuals harbour a CNV at least 500 kbp, and 1% of individuals carry a large CNV of at least 1 Mbp in size [24]. This means that larger CNVs are skewed toward rare variants. As the full extent of structural variation in our genome has been revealed, it has been estimated that CNVs account for ~ 13% of the human genome [7, 25]. The *de novo* rate of large (> 100 kb) CNV formation in humans was estimated at 1.2×10^{-2} CNVs per genome per transmission against a high selection pressure, suggesting that each of these *de novo* CNVs persists in the population for only a few generations [26, 27]

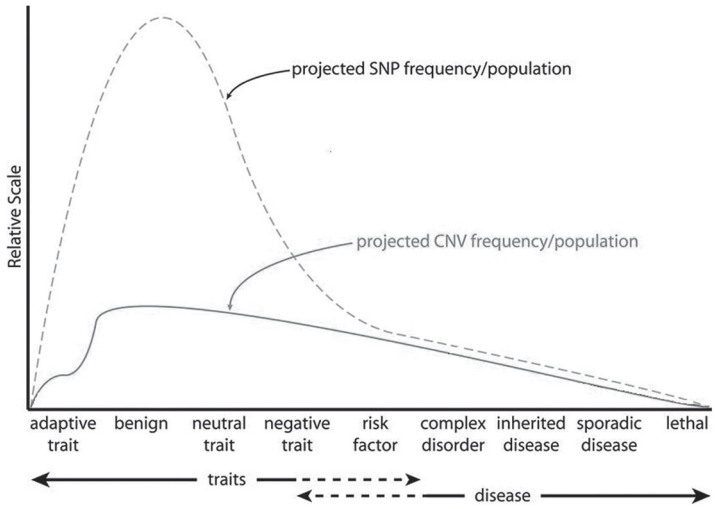
The full extent to which CNVs are likely to contribute to the diversity of human phenotypes is still under assessment. It is clear that the phenotypic impact of CNVs occurs as a continuum from 'neutral' to pathogenic, and can act in more complex and sometimes unexpected ways. Figure 2 shows conceptual curves of projected frequencies of SNVs and CNVs in the population associated with their phenotypic impact. In general, SNVs and CNVs with a high population frequency are annotated as benign or neutral in their effect, while rare variants are more likely to be pathogenic. In between, CNVs and SNVs can be associated with degrees of function described as 'traits', 'risk factors' and, beyond a certain threshold, 'disease'. After extensive phenotype-genotype studies, some of the SNVs and CNVs previously annotated as benign or neutral in their effect will be reclassified as predisposing risk factors.

Uncovering the genetic basis of human phenotypic differences requires a comprehensive understanding of all forms of genetic variation, both at fine scale (sequence variants) and large scale (structural variants). Different genomic technologies are required to detect structural variation at different levels. Next generation sequencing technologies will be used to generate comprehensive maps of human genetic variation.

Effects of genomic variation

The decoding of information from DNA to protein begins with transcription, as messenger RNA (mRNA) is created from a DNA template, followed by translation. Translation is the process of protein synthesis from mRNA, with specific amino acids encoded by three nucleotide combinations (codons). Synthesis of the protein takes place in a specific reading frame according to these codons. As such, each type of variant may have consequences at different levels (DNA, RNA, protein), also depending on the surrounding genomic context

Figure 2: Conceptual curves of SNV and CNV characteristics. Projected frequencies in the population for SNVs (dashed grey) and CNVs (blue) show the relation between genomic variation and a spectrum of phenotypic impact. Adapted from Buchanan et al [28].



(coding or non-coding). Changes in coding or regulatory sequences are most likely to affect gene expression or affect the function of protein products [29]. Any phenotypic effect is highly dependent on the location (type of tissue) and the developmental stage in which the sequence variant is expressed.

Classification of genomic variation can also be based on the effect at different levels (DNA, RNA, protein). In diploid organisms (such as humans), changes in the DNA may occur on one (heterozygous) or both (homozygous) alleles. Single nucleotide changes that lead to amino acid changes in the protein coding region of a gene may be classified into three types (silent, missense and nonsense), depending upon what the erroneous codon codes for. Synonymous variants (also known as silent variants) do not lead to a different amino acid being encoded (due to redundancy in the codon code). Non synonymous variants do alter the amino acid sequence of a protein, and can be sub-classified such as missense variants (encoding a different amino acid) and nonsense variants (creating a stop codon, leading to premature protein truncation). Frameshift variants (variants disrupting the reading frame such as deletions and duplications) lead to an altered, usually non-functional, protein product.

On RNA level any step of gene expression, post-transcriptional modification (capping,

polyadenylation, splicing) may be modulated. Most human genes can be transcribed into different mRNAs composed of different exons, leading to alternative splicing and the expression of functionally diverse protein isoforms [30]. A significant fraction of exonic variants (including silent and missense variants) and intronic variants cause disease by disrupting normal splicing [31]. Silent changes may affect translational efficiency, resulting in different protein levels or affect protein function via folding [32, 33].
Variation can also be classified by effect on function. Variants can have no effect (neutral), or lead to a change in function. Loss-of-function changes result in non-functional proteins. When the dose of a gene product is affected and not enough for a normal function this is called haploinsufficiency or dosage effect. Gain-of-function changes yield a protein that has a new function. Dominant negative changes have an altered gene product that has a negative impact on normal function. Variants that prevent viability are termed lethal.

Changes not to the DNA sequence itself but modifications or addition of chemical groups to individual nucleotides (e.g. methyl groups or proteins introduce another type of variation called epigenetics. This type of variation may have an effect on phenotype by altering gene expression. Comparable to genomic variation, epigenetic differences between individuals and monozygotic twins have been described [34, 35]. However, the epigenome is not the subject of this thesis and will not be discussed further.

Genomic variation in relation to disease: intellectual disability and/or congenital anomalies

The number of known pathogenic variants in human genes that underlie or are associated with human inherited disease is > 110000, in more than 4000 genes [36]. Currently, causal variants for ~ 3000 Mendelian disorders have been reported in the online database for Mendelian disorders in man (OMIM May 2012, # entries for phenotype description, molecular basis known).
Mental retardation (MR) or intellectual disability (ID) is defined as a significant impairment of cognitive and adaptive functions [37] and affects around 1-3% of individuals [38, 39]. The proportion of intellectual disability that can be at least partially accounted for by genetic factors (chromosomal, monogenic or multifactorial) is difficult to estimate, but may account for ~ 30% of all cases. In approximately 50% of cases, the cause remains unknown [40]. G-banded karyotyping supplemented by (sub)telomere screening or fluorescence in-situ hybridization (FISH) detects significant abnormalities in up to 10% of intellectual disability cases [41]. ~ 5% of patients have a monogenic disorder, where a causative variant in a single gene can be identified [42]. CNVs (> 500 kb) detected by array-based technologies (aCGH, SNP array) explain another ~ 15% of cases [43]. *De novo* CNVs and point pathogenic variants

of large effect may explain the majority of all intellectual disability cases in the population, and could thereby explain why such a disorder with reduced fecundity remains present in the population [44].

In recent years, several new monogenic disease genes and numerous submicroscopic deletion and duplication syndromes ('genomic disorders') have been identified. It has been estimated that about 0.7-1 per 1000 live births has a genomic disorder [45] . In general, CNVs implicated in genomic disorders are *de novo* and large in size (> 50 kb). However, it seems likely that additional genomic disorders due to much smaller *de novo* CNVs remain to be discovered [46].

In addition to microdeletion/microduplication syndromes, CNV's are involved in many common complex traits, including autism and schizophrenia [47-51]. Several genomic disorders involving inherited CNVs have been described (1q11.2, 1q21.1, 15q11, 16p13.11, 16p11.2, 22q11). Microdeletions and microduplications at these loci show variable penetrance and expression, and are thus not necessarily recognised as clinically distinctive syndromes. Inherited CNVs may be involved in disease, by acting as modifiers in milder phenotypes, either in combination with other CNVs or with sequence variants. This is illustrated by the fact that 25% of ID children carry a 2nd CNV, in addition to an inherited CNV [52]. Some authors have coined the term 'second site' or 'two hit' model for this phenomenon, which is confusing since those terms are widely used in the cancer field to indicate somatic mutations disturbing the regulation of cell growth. We prefer to stick to 'multifactorial' a term formerly used to indicate any number of unknown factors above one, but can be used as well for known factors. Recently, compound inheritance of a null allele (deletion 1q21.1) together with the presence of a SNV was proven to be associated with TAR syndrome [53]. The effect of different combinations of (multiple) inherited CNVs and/or sequence variants on phenotype urgently requires further study, and identifying additional genetic and/or environmental factors is the challenge of our time.

Genetic inheritance of genomic variation

Mendelian inheritance patterns can demonstrate a greater level of complexity than simple dominant, recessive or X-linked inheritance, through a number of non-Mendelian processes including imprinting, X-inactivation, mosaicism and variation in penetrance and variable expressivity. When searching for the exact genetic cause of a disease, and calculating risk of transmission to offspring, it is important to take this level of complexity into account. Examples of complex inheritance includes disorders caused by more than one gene (di/polygenic disorders such as in Axenfeld-Rieger syndrome [54]), trinucleotide repeat expansion disorders (such as Huntington's disease [55]) where the number of CAG repeats correlate with the severity of disease and the age of onset in combination with

the characteristic of anticipation (the tendency for progressively earlier or more severe expression of the disease in successive generations), genes whose expression is governed by parent of origin (Angelman/Prader Willi syndrome [56], Beckwith-Wiedemann/Silver Russell syndrome [57]), triallelic inheritance (Bardet-Biedl syndrome [58]). Complex inheritance involving co-inheritance of CNVs [52] or a CNV in combination with sequence variants [59, 53] have also been demonstrated.

Sequence variants and CNVs can be inherited from a parent, or occur *de novo*. From the analysis of complete genome sequences of two parent-offspring ('trios') studies, it has been estimated that each child inherits about 30 to 50 new variants [9, 18]. These new variants can either be germline variants that have arisen during the production of gametes in the parental generation, or be present in a subset of cells from either parent, and represent a germ line mosaic with recurrence risk to subsequent offspring [60]. A significantly greater proportion of new variants is of paternal origin, as a result of the larger number of divisions during spermatogenesis [61-63], a phenomenon already described by population geneticists more than half a century ago [61]. A similar finding was reported for *de novo* CNVs [64].

Techniques to detect genomic variation

History

Within the past 40 years, a variety of experimental methods have emerged; typically each focuses on a particular class of genomic variation limited by the size range of the events. Early cytogeneticists studying chromosome variation can be legitimately regarded as the first genome pioneers [65]. The establishment of the human chromosome count [66] and the discovery of trisomy 21 in Down syndrome [67] have been the groundwork for studying disease-related genomic variation. Since the 1970's, G-banding and later high resolution chromosome banding techniques, enabled detection of microscopic variation (> 5 Mb). In the 1980's, FISH (Fluorescent *in situ* hybridisation) was developed for microscopic detection of specific structural chromosome abnormalities [68, 69]. At the same time, with the development of molecular markers and application of recombinant DNA techniques, genomic variation at DNA level (sequence variation) was discovered [70]. The relative easy and reliable chain-termination method developed by Sanger soon became the method of choice for sequencing [71]. In the early 90's, Comparative Genome Hybridization (CGH) allowed the detection of submicroscopic structural variation [72]. In the last decade, array-based methods were developed (array-CGH, SNP-arrays) and contributed greatly to our knowledge of structural variation.

Overview of techniques

Quantitative and qualitative changes require different detection methods. At this moment it is not possible to detect all types of genomic variation with just one technique. Each technique has advantages and disadvantages concerning not only specificity, sensitivity, throughput and resolution, but also costs and feasibility in the laboratory. In the field of clinical genetics, many techniques have been used to study our genome. The techniques discussed here are those used in the research comprising this thesis. Table 4 gives an overview of the techniques used for detection of genomic variation. In general, the resolution of a technique depends on its design and is rarely a fixed number. Improvement in resolution mainly depends on probe size and distances between probes.

Cytogenetic and molecular techniques

Karyotyping, using the light microscope to visualize G-banding patterns, enables rapid identification of all chromosomes in a metaphase spread in one view. This enables clonal analysis, but the resolution of a light microscope is limited. Large (> 5 Mb) chromosomal rearrangements such as deletions, duplications, translocations, inversions and insertions can usually be detected. FISH using probes covering the regions affected enables microscopic detection of structural chromosomal abnormalities directly on metaphase chromosomes and interphase nuclei using fluorescently labelled DNA probes. As with karyotyping, routine FISH analysis has a limited resolution (50 kb-2 Mb). Both techniques do not detect single nucleotide variants or small structural variations. The ultimate resolution of the microscopic approach is obtained by Fiber FISH [75] allowing the detection of small deletions and duplications (5 kb-500 kb).

In addition to FISH, array-based methods have been developed. Array comparative genome hybridisation (Array-CGH) platforms are based on the principle of combined hybridisation of two differentially labelled samples to hybridisation targets. As targets, DNA isolated from Bacterial Artificial Chromosomes (BAC clones), or PCR products there of, or synthetically produced long oligonucleotides are spotted onto a glass slide (array). Subsequently, labelled genomic DNA from a test and a reference sample are hybridised to the array and fluorescence intensities are measured. Relative intensity ratios are calculated, with imbalances indicating the presence of copy number variation. The resolution of array-CGH is unlimited but dependent on the spacing of the clones and their insert size. Arrays with 3500 BAC clones (resolution ~ 1Mb) or tiling arrays (33000 BAC clones) with a 10 fold higher resolution are usually used. This resolution does not allow the accurate determination of breakpoints. Custom high-density oligonucleotide arrays are available on demand, allowing the discovery of CNVs down to 500 bp and more precise breakpoint mapping [9]. For high resolution, genome wide analysis of copy number changes,

Table 4: Techniques to detect genomic variation. Table adapted from Schoumans and Ruivenkamp [73], Gijsbers [74].
-= not possible, ±= less suitable, + = sufficient, ++ = appropriate

Genomic variation	Conventional karyotyping (> 5Mb)	locus specific metaphase FISH (50 Kb- 2 Mb)	locus specific interphase FISH (50 Kb- 2 Mb)	array CGH (~ 0,1-1 Mb)	snp array (3-5 snp probes)	MLPA (45-70 bp)	QF-PCR	HRMA (≥ 1 bp)	WES (≥ 1 bp)
balanced translocation	+	++	-	-	-	-	-	-	±
unbalanced translocation	+	++	-	++	++	±	-	-	±
inversion	+	++	-	-	-	-	-	-	±
insertion	-	++	-	-	-	-	-	-	+
complex rearrangement	+	++	-	±	±	±	-	-	-
deletion	±	+	+	++	++	++	++	-	±
duplication	±	-	±	++	++	++	++	-	±
triplication	±	-	±	++	++	+	++	-	±
trisomy	++	++	++	+	+	++	++	-	±
triploidy	++	++	++	-	±	±	++	-	-
monosomy	++	++	++	+	+	++	++	-	±
uniparental disomy	-	-	-	-	+	-	-	-	±
methylation defect	-	-	-	-	-	++	++	+	+
copy neutral LOH	-	-	-	-	+	-	-	-	±
single base pair changes	-	-	-	-	-	±	±	++	++

SNP-based arrays can be used. There are differences in experimental design by different manufacturers, which will not be discussed here. In general, they are single colour assays, designed to be used for SNP genotyping and copy number analysis in one experiment. One DNA sample is hybridised per array, and copy number ratios are determined by clustering the intensities of each probe across many samples. A major advantage is the possibility to detect low level mosaicism (~>15%), uniparental disomies (UPD) and identity-by-descent (IBD) [76-78]. An important limitation of microarrays are the difficulties of probe design in repeat-rich and duplicated regions, with structural variants being strongly correlated to these regions. Smaller CNVs (< 10kb) are also more difficult to detect routinely, which also leads to under-representation in CNV databases. All array-based techniques cannot detect very small structural deletions or duplications and balanced rearrangements (e.g. translocations). Although it is technically possible, SNP arrays are not routinely used to detect single nucleotide variants.

Several PCR-based techniques have been developed to investigate targeted regions for the presence of CNVs. Multiplex ligation-dependent probe amplification (MLPA) [79] is based upon the ligation of two adjacent-annealing oligonucleotides, followed by a quantitative PCR of the ligated products. This technique detects copy number variations of test DNA, where the relative amount of each product is correlated to the copy number of the locus being tested. Normalization using control probes, and normalization against control samples, determines the relative copy number in the test sample. Detection of deletions (ratio threshold under 0.75) and simple duplications (ratio threshold above 1.25) is relatively straightforward, but this technique is less accurate in detecting higher numbers of duplicates or mosaic changes. In addition, using specific probe design at the ligation site it is possible to detect point single nucleotide variants. At the same time, this also means detection of a deletion may be false positive if a SNV is present at, or close to, the ligation site. The advantage of this technique is the flexibility in experimental design, high resolution of CNV detection (45-70 bp) and the possibility to screen large numbers of patients for different loci in a single experiment.

QF-PCR (quantitative PCR) or QMPSF (Quantitative Multiplex PCR of Short Fluorescent Fragments), like MLPA, is a quantitative assay based on PCR amplification of genomic DNA using fluorescently labelled primers. It monitors the amount of product generated during the amplification process compared to a reference. Quantitative differences are used to estimate a relative copy number.

In addition to the previously described techniques that allow detection of quantitative differences, several methods have been developed to detect qualitative differences.

High resolution melting curve analysis (HRMA) is a method used for genotyping, single nucleotide variant scanning and sequence matching [80, 81] and is also suitable for detecting

mosaicism. It is based on the dissociation-characteristics of double-stranded DNA during heating. Following PCR amplification of a target region with a DNA binding fluorescent dye (LC-green), the amplicon is melted out by increasing the temperature in the solution. Double stranded DNA will become single stranded and the dye will be released. The specific sequence of the amplicon (primarily GC content and length) determines the melting behaviour. Each amplicon has a unique melting pattern, based on its sequence. DNA with a higher G-C content, whether because of its source or because of SNVs, will have a higher melting temperature than DNA with a higher A-T content. Amplicons can be compared by plotting the change in fluorescent signal against the melting temperature (Tm).

DNA sequencing technology has developed at an unprecedented rate in recent years. Sanger sequencing is regarded as the gold standard technique to study alterations at the single nucleotide level. It allows easy detection of SNVs, small insertions, deletions and a moderate level of mosaicism (15-50%) [82, 83]. There is a risk of false negative results when there is allelic dropout due to variants in the primer binding sites, or when the target (exon/gene) sequence is deleted [84]. When the amplified alleles have significant size differences (e.g. due to insertions or deletions) preferential amplification of the shorter allele may mask the second allele. Current sequencing techniques ('Next generation sequencing' or NGS) allow screening of the whole genome in one experiment. Many different sequencing technologies and systems have emerged the past three years, utilising different methodologies. Ideally, complete genome sequencing followed by *de novo* assembly and comparison to a high-quality reference could identify thousands of sequence variants. In practice, NGS still faces technical, but more importantly, substantial computational and bioinformatic challenges. Read lengths are <100 bp for most approaches, significantly less than the 500-1000 bp routinely obtainable from Sanger sequencing. Base calling error rates are dependent on the platform that is used, but ranges from 0.01% to 16% [85]. Increasing the coverage (depth) can minimize false calls, allowing for accurate detection of SNVs [86].

Targeted sequencing approaches have the advantage of isolating multiple genomic regions of interest in a single experiment in an efficient and cost-effective manner. Whole Exome Sequencing (WES) captures only coding regions (~ 180000 exons) of the genome, and has been successfully applied in finding causative variants in many Mendelian disorders [87-91]. The total size of the human exome is approximately 30 Mb and comprises ~ 1% of the human genome. Therefore, when the region of interest is only protein coding sequences, exome sequencing is an efficient approach for obtaining the desired coverage for variant detection [92]. For WES, publicly available programs can be used for variant calling and annotation. Sequenced individuals have typically had 5000-10000 variant calls, representing non-synonymous substitutions in exonic regions, splicing alterations or small

indels [88, 89, 93, 94]. Advanced filtering based on predicted effect (silent, nonsense, missense, effect on splicing, frameshift) or position (intronic, UTR, exonic, intergenic) can be used to find pathogenic variants. Two important assumptions underlie filtering strategies. The first is that causal variants for Mendelian disorders are rare, and therefore unlikely to be present in public databases or control sequencing data. The second is that synonymous variants are unlikely to be causative. Both assumptions are not always true [32, 95] . Mendelian disorders with milder phenotypes may have been overlooked in the general population, and may thus be present in control databases. Also, variants involved in recessive disorders with a high carrier frequency can be reported in control databases. For example, the most common variant in cystic fibrosis has an allele frequency of about 3% in Western European populations. Filtering such variants might erroneously exclude those pathogenic variants from consideration. As synonymous variants may have functional effects, and can be targeted by natural selection [96, 97] it is not always appropriate to filter these.

There are a several strategies that can be applied for the detection of structural variants. Various computational algorithms for identifying and characterizing variants have been developed. Read-pair methods assess the span and orientation of paired-end reads, and can be used to detect all types of structural variation (indels, inversions, tandem duplications and translocations) [22, 98]. Read-depth methods are simply based on the theory that CNV regions show differences in the number of reads, and therefore assess mapping depth in the sequenced sample. This will detect CNVs (including aneuploidy) and large insertions, but not inversions and translocations. Split-read approaches are able to detect all breakpoints (deletions, tandem duplications, translocations and inversions) [98, 99]. However, none of these strategies is comprehensive, and each will require substantial further development. Recently, an integrated computational pipeline for whole genome sequencing was published, enabling detection of all types of genetic variations (single nucleotide variants, short insertions or deletions (indels) and larger structural variations (SVs) [100] .

Interpretation of genomic variation

The major challenge in the analysis of genomic variation is to distinguish benign variants from variants that have clinical consequences. Increasingly, clinical molecular laboratories are detecting novel CNVs and sequence variations of unclear significance (UVs) in the course of testing patients. Guidelines on the interpretation of such variants have been developed [76, 101-105]. Sequence variants can consequently be classified on their phenotypic consequences within a spectrum of interpretations, ranging from those in which the variation is almost certainly of clinical significance, to those in which it is almost certainly not [106, 107].

When the impact of a sequence variant is undetermined, follow-up activities may be

useful to clarify this relationship, and assist with risk assessment. There are several lines of evidence suggested for designating a variant as either phenotype-modifying or neutral. (1) Occurrence in patients or in a control population as listed in a database, (2) Variant type (missense, nonsense, silent), (3) *In silico* prediction of effect on protein structure or RNA splicing, (4) Conservation among species, and or presence in a known functional domain, (5) Co-segregation with a phenotype in the family.

The presence or absence of a variant in established databases, ideally containing all the lines of evidence listed above, is of great importance. It is helpful to establish whether a variant has been reported before, and if there is an associated phenotype. Distinction should be made between variants directly linked to a phenotype or variants found by genome-wide association studies (GWAS) which have identified many variants with a very small contribution to an associated trait, in which additional genes and/or environmental factors also influence phenotypic outcome (multifactorial traits) [108].

Currently, the numbers of submicroscopic imbalances that have been reported are increasing, but the delineation of associated clinical features remains difficult and incomplete. When large stretches of DNA encompassing several genes are deleted, complex phenotypes may emerge (so called 'contiguous gene syndromes'). It may often be unclear which gene in the interval is responsible for a given part of the phenotype (e.g. heart malformations, limb malformations), which is another reason why CNVs are collected and compared in patient databases.

As described previously, the capacity of a given variant (silent, missense, nonsense, frameshift, splicing) to affect gene expression or the function of its protein products must be determined. *In silico* prediction of pathogenic effects using information on interspecies sequence conservation can be helpful. The observation that pathogenic variants are more likely to occur at positions that are conserved through evolution suggested that prediction could be based on sequence homology [109]. Variants that alter conserved residues by replacing them with amino acids with different physical characteristics are likely to affect polypeptide structure and function. It was also observed that disease-causing amino acid substitutions have a common structural feature (for instance sites with low solvent accessibility, sites involved in disulphide bonds, sites involved in folding) that distinguishes them from neutral substitutions, suggesting that structure can also be used for predicting functional consequences [110, 111].

Databases collecting information on prevalence and frequency of sequence and structural variants in the human genome have been established, and provide information on the frequency of previously reported variants in disease and control cohorts. For rare recessive disorders it is generally assumed that a variant is unique, or at least has a low carrier frequency in the general population, whereas a variant that is found at higher frequency

may be neutral. However, some recessive diseases are relatively frequent in subpopulations due to survival benefit for heterozygotes or genetic drift (for example sickle cell trait gives resistance to malarial infection), stressing the importance of registration of ethnicity in databases.
In silico prediction of an effect of a DNA variant on mRNA splicing should be verified by RNA studies. Such studies are not always feasible since not all genes are transcribed in easily accessible tissues such as blood.
Identification of a variant in a patient should be followed by testing of the parents. *De novo* variants are more likely to be pathogenic, but since we know each child may have between 30 to 50 new variants [9,18], this is by no means certain. In rare cases co-segregation of the variant with the disease in a family may corroborate the suspicion of pathogenicity. Finally, testing the gene product in a functional assay or constructing animal models (e.g. mouse, rat or fruit fly) carrying the same variant may provide definitive assessment of the phenotypic effect of the variant *in vivo*. A reliable functional assay is generally regarded as one of the best means of confirming pathogenicity, however this is rarely available as part of a routine diagnostic service.

For structural variation such as CNV, it is generally considered that if a phenotypically normal family member carries the same chromosomal anomaly, the anomaly is of no phenotypic relevance [112]. However, caution remains necessary since the identification of a CNV in a patient that is also present in an apparently healthy parent does not rule out pathogenicity as there could be a pathogenic variant on the other allele. Also, some phenotypes may be mild and therefore overlooked in a parent. As described earlier for inherited CNVs, the CNV may in fact be a risk factor that only reveals a phenotype when combined with one or more other variants in the genome. In addition, this also means that even if a CNV is relatively frequent in a population this does not rule it out as a risk factor as long as it has not been reported as homozygously deleted in a healthy population.

Tools and databases

Databases allow researchers to share knowledge and retrieve information about genomic sites of variation under study. However, at present there is no database summarizing all that is known about a certain variant.
There are several human cytogenetic databases that link submicroscopic chromosomal imbalances (microdeletions/duplications/insertions, translocations and inversions) with clinical phenotype. These include DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources)[113, 114] and ECARUCA (European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations) [115]. Since

these databases contain patient information, complete access is limited to clinicians or cytogeneticists. Some information is publicly available through specified tracks in genome browsers. The Database of Genomic Variants (DGV) is a comprehensive database for the deposition, retrieval and visualization of human structural variation [116]. The database currently contains 179450 CNVs (April 2012) and reports CNVs, inversions and Indels (100 bp-1 Kb) in apparently healthy human cases. However, interpretation of the variants should be performed carefully, as different platforms have been used to detect CNVs, and population and medical data of the cohorts are poorly defined or absent. DbVar is a public repository that accepts direct submissions and provides archiving, accessioning and distribution of publicly available genomic structural variants [117]. It accepts data from all species, and includes clinical data. It marks and allows searching for variants that are known to be pathogenic.

The dbSNP database serves as a central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms [118]. More than 17 million SNPs have been documented in this database, with a false-positive rate estimated at 15-17% [92]. However, the database does contain validated SNPs (allele frequencies provided) and can be queried accordingly, increasing the utility of the database. Allele and genotype frequencies from the HapMap project [119] have also been submitted to dbSNP. In the HapMap project, four large populations of African, Asian, and European ancestry have been studied extensively to catalogue population specific variation.

The 1000 Genomes Project [1] generated the most comprehensive map of human genetic variation yet, using next-generation sequencing technologies. It contains three pilot studies with a range of coverage. The exome variant server (EVS) [120] contains exome data of 6500 samples. A recent initiative is the generation of the genome of the Netherlands (GoNL), where whole genome sequences of 250 healthy Dutch parent-child trios are collected and stored in a biobank [121].

Some databases only contain information on genetic variation causing genetic disorders or traits. The Human Gene Mutation Database (HGMD) includes the first example of all exonic and +1,+2, -1,-2 splice-site variants causing or associated with human inherited disease, plus disease-associated polymorphisms reported in the literature [36]. Locus Specific DataBases (LSDB) are databases recording all variation within a gene. The databases contain accurate (curated), clearly referenced data naming variants at the DNA, RNA and protein level, and include all relevant comments relating to the clinical interpretation of the variant. Existing LSDBs can be checked by the URL "GeneSymbol.LOVD.nl" (e.g.MBTPS2.LOVD.nl) [122]. The Human Genome Variation Society (HGVS) keeps a list of locus specific variant databases [123]. Several tools have been developed to predict the effect of non-synonymous variants on

a protein. The Grantham score (GMS) [124] represents one of the first attempts to assess the effect of amino acid substitutions on protein structure based on chemical properties, including the side-chain composition, polarity and molecular volume. The GMS is a measure of dissimilarity between a human amino acid and the residues seen at the same site in homologs. Several studies used a GMS score less than 60 to define neutral variants, whereas a GMS score significantly larger than 60 indicates that the amino-acid change is evolutionarily intolerant [125, 126].

Polyphen [29, 127, 128] utilizes a combination of 3D structural parameters and sequence homology to make a prediction about a functional effect. This prediction is based on a number of features comprising the sequence, phylogenetic and structural information characterizing the variant. It returns predictions of "probably damaging, "possibly damaging," benign and "unknown." 'Sorting Intolerant From Tolerant' (SIFT) [129] uses sequence homology and the physical properties of amino acids to predict effect on proteins. Next to SNVs, it can classify coding indels. It returns predictions of "affect protein function" and "tolerated" for each SNV. Due to differences in algorithms used for the predictions, Polyphen2 and SIFT may present contradictory results.

Tools that calculate conservation scores can aid in variant interpretation. A variant that leads to a nonconservative substitution of an evolutionarily conserved amino acid is more likely to be causative of the disorder than a variant that leads to a conservative substitution or alters an amino acid that is not evolutionarily conserved. Both phastCons [130] and phyloP [131] calculate conservation scores for three groups of organisms (primates, placental mammals and vertebrates). The two conservation scores are informative in different ways. Phastcons estimates the probability that a nucleotide belongs to a conserved element, taking neighbouring bases into account, while PhyloP predicts conservation purely at the base level.

UMD predictor [132] provides a combinatorial approach that associates several data such as localization within the protein, conservation, biochemical properties of the variant and wild-type residues, splice-site predictions and the potential impact of the variant on mRNA. Alamut [133] is a commercial package designed to help interpret variants quickly and uses different splice site prediction algorithms, PolyPhen, SIFT and calculates theoretical consequences of substitutions, insertions and deletions (effects on protein sequence, frameshifts, splicing effects, miRNA targets, nonsense-mediated mRNA decay).

Computational prediction of pathogenicity may also give false-negative results [134]. Most prediction methods for protein alterations do not take DNA sequence context into account. As a result, they can miss changes that alter splice sites [135]. Experimental verification by functional analysis of possible pathogenicity remains the golden standard.

Outline and scope of this thesis

Intellectual disability (ID) with or without multiple congenital malformations (MCA) is one of the main reasons for referral to a clinical geneticist. Causes of this ID/MCA are extremely heterogeneous and range from point mutation in one single specific gene to loss or gain of an entire chromosome. Despite enormous progress in diagnostic techniques in the past 50 years the cause for ID remains unknown in approximately 50% of the cases.

Establishing a diagnosis and understanding the cause of a genetic disorder is of great benefit for the patient and his/her family. This may provide information on prognosis, clinical management options, and anticipation on associated health issues for the patient, and may even be of therapeutic relevance in the future. Family members can be informed about recurrence risk and may be provided with options for prenatal (PND) and pre-implantation genetic diagnosis (PGD).

The main objective of the research in this thesis was to develop and apply novel molecular techniques to study the genetic basis of patients with intellectual disability (ID), multiple congenital anomalies (MCA), or other inherited disorders, and ultimately to gain insight into genetic disease mechanisms.

This thesis is a mirror image of the rapid evolution of techniques for analysis of DNA between 2006 and 2011. Many techniques described in this thesis (karyotyping, fluorescence in situ hybridization (FISH), array-comparative genome hybridization (CGH), single nucleotide polymorphism (SNP)-arrays, multiplex ligation dependent probe amplification (MLPA), high resolution melting curve analysis (HRMA), Sanger sequencing, and whole exome sequencing (WES) have been applied to find causative genome variants in our patient cohort. All of these techniques are now routinely used in clinical diagnostic laboratories, except whole exome and whole genome sequencing, implementation of which is being worked on.

More detailed data on human genetic variation are now rapidly accumulating, as new techniques for determining the primary structure of genomes have been developed at an unprecedented rate.

Although new genomic variants can arise in both the germline cells (whose DNA will be passed to offspring) and in somatic cells (the majority of cells in the human body), this thesis has a focus on the genetic characterization of germline variation.

At the outset of this work five years ago, genomic microarrays were introduced and have been applied in this thesis to identify structural variation in genetic disorders. **Chapter 2** describes methods for detecting CNV in the human genome. The focus lies on the development of a targeted array used to study patients with intellectual disability of unknown aetiology. **Chapter 3** shows the application of microarrays and MLPA in the

elucidation and delineation of a new microdeletion syndrome. Because microarrays increase the resolution of chromosome analysis by 100-1000 fold and because small deletions and duplications are a major cause of ID/MCA, conventional karyotyping was replaced by microarray analysis in routine diagnostics.

Because of its speed, simplicity, and low cost, HRMA has become a popular technique for detection of sequence variants. The two major applications are targeted genotyping and gene scanning. **Chapter 4** describes the versatility of HRMA as a molecular technique and illustrates several applications. In **Chapter 5,** SNP arrays were used to define the breakpoints of a previously defined locus for Keratosis Follicularis Spinulosa Decalvans (KFSD) and show the application of HRMA as a presequencing tool in the identification of the genetic basis of this disorder.

The introduction of NGS techniques allowed genome wide detection of sequence variants in disorders of unknown aetiology. **Chapter 6 and 7** describe the success of next generation sequencing in the identification of the genetic basis in two disorders (TOD and Aarskog-Scott syndrome), and shows the value of exome sequencing in detection of variants outside coding regions. Besides identification of the genetic cause of Coffin Siris syndrome**. Chapter 8** outlines recent findings on the mode of inheritance and frequency of this monogenic disorder using exome sequencing.

Finally, in **Chapter 9** the rapid evolution of techniques for genome analysis is discussed and alterations in the diagnostic approach to ID/MCA patients are proposed.

1. The 1000 Genomes Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467 (7319):1061-1073

2. Rahim NG, Harismendy O, Topol EJ, and Frazer KA (2008) Genetic determinants of phenotypic diversity in humans. Genome Biol 9 (4):215

3. Cummings M (1999) Human Heredity: Principles and Issues . 5th ed.

4. Cotton RG (2002) Communicating "mutation:" Modern meanings and connotations. Hum Mutat 19 (1):2-3

5. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453 (7191):56-64

6. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, and Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36 (9):949-951

7. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H et al (2006) Global variation in copy number in the human genome. Nature 444 (7118):444-454

8. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, and Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37 (7):727-732

9. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J et al (2010) Origins and functional impact of copy number variation in the human genome. Nature 464 (7289):704-712

10. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, and Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305 (5683):525-528

11. Feuk L, Carson AR, and Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7 (2):85-97

12. Sharp AJ (2009) Emerging themes and new challenges in defining the role of structural variation in human disease. Hum Mutat 30 (2):135-144

13. van Binsbergen (2011) Origins and breakpoint analyses of copy number variations: up close and personal. Cytogenet Genome Res 135 (3-4):271-276

14. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, and Eichler EE (2002) Recent segmental duplications in the human genome. Science 297 (5583):1003-1007

15. Lee JA, Carvalho CM, and Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131 (7):1235-1247

16. Alkan C, Coe BP, and Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12 (5):363-376

17. Frazer KA, Murray SS, Schork NJ, and Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10 (4):241-251

18. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, and Galas DJ (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328 (5978):636-639

19. Reumers J, De RP, Zhao H, Liekens A, Smeets D, Cleary J, Van LP, Van Den Bossche M, Catthoor K, Sabbe B, Despierre E, Vergote I, Hilbush B, Lambrechts D, and Del-Favero J (2012) Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat Biotechnol 30 (1):61-68

20. Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de ST, Menzel U, Sandgren J, von TD, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, van Ommen GJ, Boomsma DI, Pedersen NL, den Dunnen JT, Wirdefeldt K, and Dumanski JP (2008) Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82 (3):763-771

21. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha JY, Kim KH, Lee B, Bhak J, and Kim SJ (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res 19 (9):1622-1629

22. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, and Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318 (5849):420-426

23. Hehir-Kwa JY, Wieskamp N, Webber C, Pfundt R, Brunner HG, Gilissen C, de Vries BB, Ponting CP, and Veltman JA (2010) Accurate distinction of pathogenic from benign CNVs in mental retardation. PLoS Comput Biol 6 (4):e1000752

24. Girirajan S, Campbell CD, and Eichler EE (2011) Human copy number variation and complex genetic disease. Annu Rev Genet 45:203-226

25. Li J, Yang T, Wang L, Yan H, Zhang Y, Guo Y, Pan F, Zhang Z, Peng Y, Zhou Q, He L, Zhu X, Deng H, Levy S, Papasian CJ, Drees BM, Hamilton JJ, Recker RR, Cheng J, and Deng HW (2009) Whole genome distribution and ethnic differentiation of copy number variation in Caucasian and Asian populations. PLoS One 4 (11):e7958

26. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, and Eichler EE (2010) *De novo* rates and selection of large copy number variation. Genome Res 20 (11):1469-1481

27. Rees E, Moskvina V, Owen MJ, O'Donovan MC, and Kirov G (2011) *De novo* rates and selection of schizophrenia-associated copy number variants. Biol Psychiatry 70 (12):1109-1114

28. Buchanan JA and Scherer SW (2008) Contemplating effects of genomic structural variation. Genet Med 10 (9):639-647

29. Ramensky V, Bork P, and Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30 (17):3894-3900

30. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, and Shoemaker DD (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302 (5653):2141-2144

31. Wang GS and Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8 (10):749-761

32. Waldman YY, Tuller T, Keinan A, and Ruppin E (2011) Selection for translation efficiency on synonymous polymorphisms in recent human evolution. Genome Biol Evol 3:749-761

33. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, and Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315 (5811):525-528

34. Gringras P and Chen W (2001) Mechanisms for differences in monozygous twins. Early Hum Dev 64 (2):105-117

35. Petronis A (2006) Epigenetics and twins: three variations on the theme. Trends Genet 22 (7):347-350

36. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, and Cooper DN (2009) The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Hum Genomics 4 (2):69-72

37. Battaglia A and Carey JC (2003) Diagnostic evaluation of developmental delay/mental retardation: An overview. Am J Med Genet C Semin Med Genet 117C (1):3-14

38. World Health Organization. Assessment of People with Mental Retardation. 1992.

39. Maulik PK, Mascarenhas MN, Mathers CD, Dua T, and Saxena S (2011) Prevalence of intellectual disability: a meta-analysis of population-based studies. Res Dev Disabil 32 (2):419-436

40. Stevenson RE, Procopio-Allen AM, Schroer RJ, and Collins JS (2003) Genetic syndromes among individuals with mental retardation. Am J Med Genet A 123A (1):29-32

41. Baker K, Raymond FL, and Bass N (2012) Genetic investigation for adults with intellectual disability: opportunities and challenges. Curr Opin Neurol 25 (2):150-158

42. Rauch A, Hoyer J, Guth S, Zweier C, Kraus C, Becker C, Zenker M, Huffmeier U, Thiel C, Ruschendorf F, Nurnberg P, Reis A, and Trautmann U (2006) Diagnostic yield of various genetic approaches in patients with unexplained developmental delay or mental retardation. Am J Med Genet A 140 (19):2063-2074

43. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, Williams C et al (2011) A copy number variation morbidity map of developmental delay. Nat Genet 43 (9):838-846

44. Vissers LE, de LJ, Gilissen C, Janssen I, Steehouwer M, de VP, van LB, Arts P, Wieskamp N, del RM, van Bon BW, Hoischen A, de Vries BB, Brunner HG, and Veltman JA (2010) A *de novo* paradigm for mental retardation. Nat Genet 42 (12):1109-1112

45. Ji Y, Eichler EE, Schwartz S, and Nicholls RD (2000) Structure of chromosomal duplicons and their role in mediating human genomic disorders. Genome Res 10 (5):597-610

46. McCarroll SA (2008) Extending genome-wide association studies to copy-number variation. Hum Mol Genet 17 (R2):R135-R142

47. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH et al (2011) Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70 (5):863-885

48. Stankiewicz P and Lupski JR (2010) Structural variation in the human genome and its role in disease. Annu Rev Med 61:437-455

49. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B et al (2007) Strong association of *de novo* copy number mutations with autism. Science 316 (5823):445-449

50. Cook EH, Jr. and Scherer SW (2008) Copy-number variations associated with neuropsychiatric conditions. Nature 455 (7215):919-923

51. Sharp AJ, Hansen S, Selzer RR, Cheng Z, Regan R, Hurst JA, Stewart H, Price SM, Blair E, Hennekam RC, Fitzpatrick CA, Segraves R, Richmond TA, Guiver C, Albertson DG, Pinkel D, Eis PS, Schwartz S, Knight SJ, and Eichler EE (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. Nat Genet 38 (9):1038-1042

52. Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, Vives L et al (2010) A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet 42 (3):203-209

53. Albers CA, Paul DS, Schulze H, Freson K, Stephens JC, Smethurst PA, Jolley JD et al (2012) Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet 44 (4):435-2

54. Kelberman D, Islam L, Holder SE, Jacques TS, Calvas P, Hennekam RC, Nischal KK, and Sowden JC (2011) Digenic inheritance of mutations in FOXC1 and PITX2 : correlating transcription factor function and Axenfeld-Rieger disease severity. Hum Mutat 32 (10):1144-1152

55. The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72 (6):971-983

56. Knoll JH, Nicholls RD, Magenis RE, Graham JM, Jr., Lalande M, and Latt SA (1989) Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. Am J Med Genet 32 (2):285-290

57. Eggermann T (2009) Silver-Russell and Beckwith-Wiedemann syndromes: opposite (epi) mutations in 11p15 result in opposite clinical pictures. Horm Res 71 Suppl 2:30-35

58. Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, Scambler PJ, Davidson WS, Beales PL, and Lupski JR (2001) Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder. Science 293 (5538):2256-2259

59. Lerer I, Sagi M, Ben-Neriah Z, Wang T, Levi H, and Abeliovich D (2001) A deletion mutation in GJB6 cooperating with a GJB2 mutation in trans in non-syndromic deafness: A novel founder mutation in Ashkenazi Jews. Hum Mutat 18 (5):460

60. Helderman-van den Enden AT, de JR, den Dunnen JT, Houwing-Duistermaat JJ, Kneppers AL, Ginjaar HB, Breuning MH, and Bakker E (2009) Recurrence risk due to germ line mosaicism: Duchenne and Becker muscular dystrophy. Clin Genet 75 (5):465-472

61. Haldane J.B (1947) The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. Ann Eugen 13 (4):262-271

62. Hurst LD and Ellegren H (1998) Sex biases in the mutation rate. Trends Genet 14 (11):446-452

63. Makova KD and Li WH (2002) Strong male-driven evolution of DNA sequences in humans and apes. Nature 416 (6881):624-626

64. Hehir-Kwa JY, Rodriguez-Santiago B, Vissers LE, de LN, Pfundt R, Buitelaar JK, Perez-Jurado LA, and Veltman JA (2011) *De novo* copy number variants associated with intellectual disability have a paternal origin and age bias. J Med Genet 48 (11):776-778

65. Pearson PL (2006) Historical development of analysing large-scale changes in the human genome. Cytogenet Genome Res 115 (3-4):198-204

66. Tjio JH and Levan a (1956) The chromosome number in man. Hereditas 42:1-6

67. Lejeune J, Gautier M, and Turpin MR (1959) Etude des chromosomes somatiques de neuf enfants mongoliens. CR Acad Sci III (248):1721-1722

68. Gerhard DS, Kawasaki ES, Bancroft FC, and Szabo P (1981) Localization of a unique gene by direct hybridization in situ. Proc Natl Acad Sci U S A 78 (6):3755-3759

69. Van Prooijen-Knegt AC, Van Hoek JF, Bauman JG, Van DP, Wool IG, and Van der Ploeg M (1982) In situ hybridization of DNA sequences in human metaphase chromosomes visualized by an indirect fluorescent immunocytochemical procedure. Exp Cell Res 141 (2):397-407

70. Francomano CA and Kazazian HH, Jr. (1986) DNA analysis in genetic disorders. Annu Rev Med 37:377-395

71. Sanger F, Nicklen S, and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74 (12):5463-5467

72. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, and Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258 (5083):818-821

73. Schoumans J and Ruivenkamp C (2010) Laboratory methods for the detection of chromosomal abnormalities. Methods Mol Biol 628:53-73

74. Gijsbers AC (2010) high-resolution karyotyping by oligonucleotide microarrays: the next revolution in cytogenetics.

75. Florijn RJ, Bonden LA, Vrolijk H, Wiegant J, Vaandrager JW, Baas F, den Dunnen JT, Tanke HJ, van Ommen GJ, and Raap AK (1995) High-resolution DNA Fiber-FISH for genomic DNA mapping and colour bar-coding of large genes. Hum Mol Genet 4 (5):831-836

76. Gijsbers AC, Lew JY, Bosch CA, Schuurs-Hoeijmakers JH, van HA, den Hollander NS, Kant SG, Bijlsma EK, Breuning MH, Bakker E, and Ruivenkamp CA (2009) A new diagnostic workflow for patients with mental retardation and/or multiple congenital abnormalities: test arrays first. Eur J Hum Genet 17 (11):1394-1402

77. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, and Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16 (9):1136-1148

78. Rodriguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, Kogevinas M, Villa O, Hutchinson A, Earl J, Marenne G, Jacobs K, Rico D, Tardon A, Carrato A, Thomas G, Valencia A, Silverman D, Real FX, Chanock SJ, and Perez-Jurado LA (2010) Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. Am J Hum Genet 87 (1):129-138

79. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, and Pals G (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30 (12):e57

80. Reed GH and Wittwer CT (2004) Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin Chem 50 (10):1748-1754

81. Reed GH, Kent JO, and Wittwer CT (2007) High-resolution DNA melting analysis for simple and efficient molecular diagnostics. Pharmacogenomics 8 (6):597-608

82. Rohlin A, Wernersson J, Engwall Y, Wiklund L, Bjork J, and Nordling M (2009) Parallel sequencing used in detection of mosaic mutations: comparison with four diagnostic DNA screening techniques. Hum Mutat 30 (6):1012-1020

83. Necker J, Kovac M, Attenhofer M, Reichlin B, and Heinimann K (2011) Detection of APC germ line mosaicism in patients with *de novo* familial adenomatous polyposis: a plea for the protein truncation test. J Med Genet 48 (8):526-529

84. Sian Ellard, Ruth Charlton, Michael Yau, David Gokhale, Graham R Taylor, and AndrewWallace6 and Simon C Ramsden. CMGS: Practice guidelines for Sanger Sequencing Analysis and Interpretation. 2009.

85. Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11 (5):759-769

86. Koboldt DC, Ding L, Mardis ER, and Wilson RK (2010) Challenges of sequencing human genomes. Brief Bioinform 11 (5):484-498

87. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, and Shendure J (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42 (9):790-793

88. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, and Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42 (1):30-35

89. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van LB, Steehouwer M, van RJ, Kant SG, Roepman R, Knoers NV, Veltman JA, and Brunner HG (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet 87 (3):418-423

90. Hoischen A, van Bon BW, Gilissen C, Arts P, van LB, Steehouwer M, de VP, de RR, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, and Veltman JA (2010) *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 42 (6):483-485

91. Wang JL, Yang X, Xia K, Hu ZM, Weng L, Jin X, Jiang H, Zhang P, Shen L, Guo JF, Li N, Li YR, Lei LF, Zhou J, Du J, Zhou YF, Pan Q, Wang J, Wang J, Li RQ, and Tang BS (2010) TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 133 (Pt 12):3510-3518

92. Ku CS, Naidoo N, and Pawitan Y (2011) Revisiting Mendelian disorders through exome sequencing. Hum Genet 129 (4):351-370

93. Rodelsperger C, Krawitz P, Bauer S, Hecht J, Bigham AW, Bamshad M, de Condor BJ, Schweiger MR, and Robinson PN (2011) Identity-by-descent filtering of exome sequence data for disease-gene identification in autosomal recessive disorders. Bioinformatics 27 (6):829-836

94. Rios J, Stein E, Shendure J, Hobbs HH, and Cohen JC (2010) Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum Mol Genet 19 (22):4313-4318

95. Gilissen C, Hoischen A, Brunner HG, and Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20 (5):490-497

96. Chamary JV, Parmley JL, and Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7 (2):98-108

97. Sun Y, Almomani R, Aten E, Celli J, van der HJ, Venselaar H, Robertson SP, Baroncini A, Franco B, Basel-Vanagaite L, Horii E, Drut R, Ariyurek Y, den Dunnen JT, and Breuning MH (2010) Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. Am J Hum Genet 87 (1):146-153

98. Albers CA, Lunter G, Macarthur DG, McVean G, Ouwehand WH, and Durbin R (2011) Dindel: accurate indel calls from short-read data. Genome Res 21 (6):961-973

99. Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25 (21):2865-2871

100. Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, and Snyder M (2012) Detecting and annotating genetic variations using the HugeSeq pipeline. Nat Biotechnol 30 (3):226-229

101. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, and Ward BE (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. Genet Med 10 (4):294-300

102. Cotton RG and Scriver CR (1998) Proof of "disease causing" mutation. Hum Mutat 12 (1):1-3

103. Bell J, Danielle Bodmer Erik Sistermans and Simon C Ramsden. Practice guidelines for the Interpretation and Reporting of Unclassified Variants (UVs) in Clinical Molecular Genetics. 2007.

104. Lee C, Iafrate AJ, and Brothman AR (2007) Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. Nat Genet 39 (7 Suppl):S48-S54

105. Koolen DA, Pfundt R, de LN, Hehir-Kwa JY, Nillesen WM, Neefs I, Scheltinga I, Sistermans E, Smeets D, Brunner HG, van Kessel AG, Veltman JA, and de Vries BB (2009) Genomic microarrays in mental retardation: a practical workflow for diagnostic applications. Hum Mutat 30 (3):283-292

106. American college of medical genetics. Practice guidelines for Sanger Sequencing Analysis and Interpretation . 2006. The Standards and Guidelines for Clinical Genetics Laboratories.

107. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, and Tavtigian SV (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat 29 (11):1282-1291

108. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, and Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9 (5):356-369

109. Miller MP and Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10 (21):2319-2328

110. Wang Z and Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17 (4):263-270

111. Sunyaev S, Ramensky V, and Bork P (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet 16 (5):198-200

112. Barber JC, Maloney V, Hollox EJ, Stuke-Sontheimer A, du BG, Daumiller E, Klein-Vogler U, Dufke A, Armour JA, and Liehr T (2005) Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. Eur J Hum Genet 13 (10):1131-1136

113. DECIPHER. http://decipher.sanger.uk.

114. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van VS, Moreau Y, Pettett RM, and Carter NP (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84 (4):524-533

115. ECARUCA. http://www.ecaruca.net.

116. Database of Genomic Variants. http://projects.tcag.ca/variation/.

117. Dbvar. http://www.ncbi.nlm.nih.gov/dbvar.

118. dbSNP. http://www.ncbi.nlm.nih.gov/projects/SNP.

119. HapMap. http://hapmap.ncbi.nlm.nih.gov.

120. Exome Variant Server. http://evs.gs.washington.edu/EVS/.

121. GoNL. http://www.nlgenome.nl.

122. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, and den Dunnen JT (2011) LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32 (5):557-563

123. Human Genome Variation Society. http://www.hgvs.org/dblist/glsdb.html.

124. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185 (4154):862-864

125. Abkevich V, Zharkikh A, Deffenbaugh AM, Frank D, Chen Y, Shattuck D, Skolnick MH, Gutin A, and Tavtigian SV (2004) Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. J Med Genet 41 (7):492-507

126. Lee E, McKean-Cowdin R, Ma H, Chen Z, Van Den Berg D, Henderson BE, Bernstein L, and Ursin G (2008) Evaluation of unclassified variants in the breast cancer susceptibility genes BRCA1 and BRCA2 using five methods: results from a population-based study of young breast cancer patients. Breast Cancer Res 10 (1):R19

127. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, and Kuznetsov EN (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12 (5):387-394

128. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7 (4):248-249

129. Ng PC and Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11 (5):863-874

130. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, and Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15 (8):1034-1050

131. Siepel A, Pollard KS and Haussler D. New methods for detecting lineage-specific selection. 190-205. 2006. In Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006).

132. Frederic MY, Lalande M, Boileau C, Hamroun D, Claustres M, Beroud C, and Collod-Beroud G (2009) UMD-predictor, a new prediction tool for nucleotide substitution pathogenicity -- application to four genes: FBN1, FBN2, TGFBR1, and TGFBR2. Hum Mutat 30 (6):952-959

133. Alamut. http://www.interactive-biosoftware.com/alamut.html

134. Raymond FL, Whibley A, Stratton MR, and Gecz J (2009) Lessons learnt from large-scale exon re-sequencing of the X chromosome. Hum Mol Genet 18 (R1):R60-R64

135. Ng PC and Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61-80


2

Methods to detect CNVs in the human genome

Emmelien Aten, Stefan J. White, Margot E. Kalf, Rolf H.A.M. Vossen, Helene H. Thygesen, Claudia A. Ruivenkamp, Marjolein Kriek, Martijn H.B. Breuning, Johan T. den Dunnen

Cytogenet Genome Res. 2008;123(1-4):313-21.

Abstract

The detection of quantitative changes in genomic DNA, i.e. deletions and duplications or Copy Number Variants (CNV), recently gained a considerable interest. First, detailed analysis of the human genome showed a surprising amount of CNV, involving thousands of genes. Second, it was realised that the detection of CNVs as a cause of genetic disease was often neglected, but should be an essential part of a complete screening strategy. In both cases new efficient CNV screening methods, covering the entire range from specific loci to genome-wide, were behind these developments. This paper will briefly review the methods that are available to detect CNVs, discuss their strong and weak points, show some new developments and look ahead. Methods covered include microscopy, fluorescence in situ hybridisation (incl. fiber-FISH), Southern blotting, PCR-based methods (incl. MLPA), array technology and massive parallel sequencing. In addition we will show some new developments, incl. a 1400-plex CNV bead assay, fast-MLPA (from DNA to result in ~ 6h) and a simple Melting Curve Analysis assay to confirm potential CNVs. Using the 1400-plex CNV bead assay, targeting selected chromosomal regions only, we detected confirmed rearrangements in 9% of 320 mental retardation patients studied.

Introduction - Methods to detect CNVs

Microscopy

The first report describing a CNV in the human genome was that by Leujeune et al. (1959); using a simple microscope he discovered that children with Down's syndrome have an extra copy of chromosome 21 ("trisomy 21"). Later specific methods were developed to generate chromosome-banding patterns. This simplified the discrimination of individual chromosomes as well as resolution inside a chromosome and facilitated the identification of inter and intra-chromosomal rearrangements. Together these tools were used to identify many other rearrangements in relation to genetic disease and cancer, especially numeric changes (aneuploidies), Structural Variants (SVs; translocations, inversions, insertions, transpositions) and large CNVs (>5-10 Mb deletions and duplications).

Resolution of microscopic chromosome analysis was further improved using in-situ hybridisation, initially using radioactive (ISH) and later fluorescent labelled probes (FISH, Landegent et al. (1985)). Nowadays FISH, especially using multiple probes labelled in different colours, is widely used in clinical diagnostics as a screening technology to confirm the presence of CNVs and SVs in either patients or carriers. Using probes close to the telomere of human chromosomes, Flint and co-workers were able to detect the presence

of CNVs in a significant fraction of patients with mental retardation and an apparently normal karyotype (Knight et al. (2000)).

Depending on the question, metaphase chromosome spreads are most commonly used. However, when a higher resolution is required (closely spaced probes) or when larger numbers of cells need to be counted (cancer), FISH can also be performed on interphase nuclei. Ultimate resolution is obtained by fiber-FISH, where probes are visualised on stretched single DNA fibers (Florijn et al. (1995)). Fiber-FISH is currently the method preferred to precisely determine the genomic structure of complex CNVs (Perry et al. (2007)).

Southern blotting

Although Southern blotting is a versatile tool for the detection of CNV, it has not been widely used. The main problems associated are the workload involved (DNA digestion, electrophoresis, blotting, hybridisation and exposure) and the fact that quantitative analysis is not simple; it demands technical experience and high-quality results. The exception has been the analysis of male samples for X-linked diseases where CNVs are frequent, e.g. in Duchenne and Becker muscular dystrophy (D/BMD) where hundreds of deletions have been revealed using Southern blotting (White and Den Dunnen (2006)). Interestingly, although duplications were described to be present in 5-10% of D/BMD patients early on (Den Dunnen et al. (1989)), only recently, using newly developed techniques, do we see that all screening studies report duplications (White and Den Dunnen (2006)). Similarly for autosomal diseases, e.g. breast cancer (Petrij-Bosch et al. (1997)) and Alzheimer (Rovelet-Lecrux et al. (2006)), it often took many years after identification of the disease gene before CNVs were first reported. These examples indicate that, although Southern blotting is a good tool for CNV detection, it is probably not the simplest tool to apply.

In specific cases, Southern blotting has been applied more widely. This includes diseases where specific CNVs are recurrent, facilitating the design of an assay targeting the unique breakpoint fragments, e.g. in alpha- and beta-thalassemia (Craig et al. (1994)). Another example is Southern blotting in combination with Pulsed-Field Gel-Electrophoresis (PFGE), allowing CNV detection at great distance from a specific probe (Den Dunnen et al. (1987)) or to determine the number of repeat sequences inside a large DNA fragment like in facioscapulohumeral muscular dystrophy (FSHD, Van Der Maarel et al. (1999)).

PCR-based methods

Because of its ease of use, PCR has been the most widely applied method to screen for CNVs. Since quantitative PCR is not that simple, PCR methods initially were used to confirm CNVs indirectly, i.e. by amplifying across unique deletion or duplication breakpoints, e.g. in delta/beta-thalassemia and hereditary persistence of fetal hemoglobin (Craig et al. (1994)).

Since the introduction of real-time PCR, facilitating simplified quantification, it has been applied as quantitative-PCR (qPCR) to directly screen for CNVs, discriminating 1 versus 2 (deletion) or 2 versus 3 copies (duplication). The main disadvantage of qPCR approaches is that only one fragment can be analysed and this fragment should be inside the CNV to get a positive result. Since CNVs in disease genes are rarely identical but can cover any part of the gene, several qPCRs will be required to perform a complete CNV screen. To bypass this problem, multiplex-qPCR approaches have been designed, for each locus tested yielding a fragment with either a unique length and/or colour. Except for again X-linked diseases to detect patients or carrier females (Beggs et al. (1990), Yau et al. (1996)), these assays have not found widespread application.

Only recently more powerful PCR-based methods have been developed, including Multiplex Amplifiable Probe Hybridization (MAPH, Armour et al. (2000)), Multiplex Ligation-dependent Probe Amplification (MLPA, Schouten et al. (2002)) and Multiplex Amplicon Quantification (MAQ, Suls et al. (2006)). These methods largely circumvent the inherent problems of multiplex-PCR, i.e. each fragment having different amplification characteristics with overall smaller fragments give higher yields. These methods achieve this by combining the advantages of an optimised primer design (incl. universal amplification primers), carefully determined primer concentrations and specific primer-mixes to achieve uniform amplification yields. Of these methods, MLPA has been by far the most successful, at the time of writing (May 2008) giving over 260 hits in PubMed. The winning features undoubtedly include its ease of use, the fact that standard laboratory equipment can be used (PCR and capillary electrophoresis) and that many disease-specific kits are readily available through commercial suppliers. More, recently array technology was applied to increase the performance of MLPA, e.g. allowing the accurate analysis of the entire *DMD* gene using 128 probes spotted in duplicate (Zeng et al. (2008)).

Array technology

The main advantage of array-based methods is their multiplexability, i.e. the number of loci that can be screened simultaneously, ultimately covering the entire human genome. Array-CGH (Comparative Genomic Hybridisation) was first developed, using arrays that contained ~ 3,000 probes (100-200 Kb genomic insert PAC/BAC clones), covering the genome at a 1 Mb resolution (Pinkel et al. (1998)). Soon these arrays contained >30,000 clones, covering the entire genome (Ishkanian et al. (2004)). In addition, arrays for genome-wide linkage and association studies determining alleles based on Single Nucleotide Polymorphism (SNP), were used early on for CNV detection (e.g. Zhou et al. (2004)). Since these arrays were not designed for this purpose, specific software had to be developed to allow quantitative analysis of the data. While the first arrays contained only 10,000 SNPs

(Zhou et al. (2004)), the latest arrays contain up to 1 million SNPs. Next to the SNP probes, these latter arrays contain yet another 1 million non-SNP probes, filling in the gaps and yielding the best genome coverage possible. These genome-wide SNP-arrays revealed an unexpectedly large and complex variability in the human genome (Redon et al. (2006)), a variability we still have not completely mapped (Kidd et al. (2008)). In addition these arrays have been instrumental in the recent discovery of many new genes and gene regions involved in genetic disease (Beckmann et al. (2007)).

New methods

Although many powerful methods are available, there is always room to improve. Array-based approaches are very powerful but overall, although prices steadily drop, they are relatively costly and not ideal when thousands of samples need to be screened. Furthermore, for diagnostic applications genome-wide screens can often not be used. Using a genome-wide tool to determine whether there is a pathogenic CNV in the *DMD* gene of a suspected carrier female for Duchenne muscular dystrophy might reveal CNVs elsewhere in the genome, giving the diagnostic lab unwanted dilemmas. Similarly, CNVs might be found for which the phenotypic consequences are unclear, yielding an inconclusive diagnosis. With these considerations in mind, we set out to develop a new assay to bridge the gap between genome-wide (array-based) and locus-specific methods. It should facilitate the cost-effective screening of 1000 or more loci, with a flexible choice of loci to include (custom design) and the possibility of automated analysis of many samples.

1400-plex CNV bead assay

The assay developed was based on Illumina's GoldenGate assay (Fan et al. (2003)) with two important changes; targeting non-SNP DNA sequences and adapted to a single colour assay. The assay was designed to screen patients with mental retardation (MR) of unknown aetiology. It should detect all trisomies, telomere-end rearrangements (incl. unbalanced translocations), known micro-deletion/-duplication syndromes and perform a rough whole genome CNV scan (see M&M probe design). Initial assessment of the array performance was made by analysing 44 samples containing known rearrangements (see M&M). All known rearrangements were detected and all CNV breakpoints matched those known.

Detailed analysis of the first 80 MR-patient samples detected one or more CNVs in 69 cases (Fig.1), in total 103 losses and 255 gains. Ten cases carried likely pathogenic autosomal CNVs. Five of these were selected and all were confirmed using a second technique (MLPA, FISH or SNP array - Fig.1B). Three cases could be proven to contain *de novo* aberrations. 59 of the 80 patients showed CNVs in regions known to be polymorphic (Toronto Database

of Genomic Variants - http://projects.tcag.ca/variation), incl the *CDKN1c, TERT, SMN1* and *NSF* genes. In addition we detected polymorphic CNVs in the *FOXD3, SOX12, TBX4, HOXD1* and *TBX21* genes, not reported in the Toronto database. The results from 240 additional patients are currently under study. So far, likely pathogenic autosomal CNVs could be confirmed in 29 cases (Table 1); 1 trisomy, 7 unbalanced translocations, 7 telomeric deletions (incl.1 ring chromosome), 10 deletions and 4 duplications (incl. a partial tetrasomy). Overall the assay thus detected a pathogenic CNV in 9% of the cases analysed.

Table 1: 1400-plex CNV bead assay. In total 320 mental retardation patients were screened using the 1400-plex CNV bead assay. Data from the first set of 80 patients has been analysed extensively, that of the second set of 240 additional patients is still in progress. Shown is an overview of all confirmed CNVs and where known the diagnosis of the phenotype in addition to mental retardation. DGS = DiGeorge syndrome, MCA = multiple congenital anomalies, dup22q11 = 22q11 microduplication syndrome, MR = mental retardation, MWS = Mowat Wilson syndrome, PLS = Potocki-Lupski syndrome, SMS = Smith Magenis syndrome, WBS = Williams Beuren syndrome.

CNV	Number of patients	Phenotype diagnosis
Trisomy (21q)	1	Down's syndrome
Unbalanced translocations (telomeric deletion + duplication)	7	MCA
Ring chromosome (11) (11q telomeric deletions)	1	MCA
Deletion, telomeric	6	2 alpha thalessemia, 4 MCA
9p tetrasomy (9p triplication)	1	MCA
Micro-deletion	10	6 DGS, MWS, SMS, WBS
Micro-duplication	2	dup22q11, PLS
1 probe duplication	1	-

Fast-MLPA

Most CNV detection methods typically take one to several days to get from DNA to result. Although time is usually not a critical aspect, there are diagnostic cases where the presence of a CNV needs to be confirmed as soon as possible. For instance newborns with multiple

congenital malformations and complex heart defects may require extensive surgery and intensive care to survive. In cases where trisomies for chromosomes 13 or 18 are present, prognosis is particularly unfavourable and one would prefer to refrain from operating to minimize the child's suffering. Therefore we developed a test that could be performed within 1 day (<8 h).

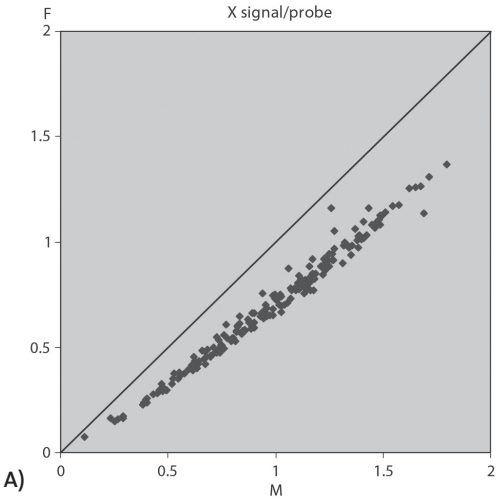
In the ultimate set up, a column-based DNA isolation method was used yielding purified DNA from a blood sample in 30 min. The time required for the MLPA could be reduced to ~ 4.5h. This was achieved by reducing the over-night hybridisation step to 2.5h and by redesigning the probe set, selecting for those probes that allowed short MLPA reaction times. Time for MLPA readout, usually acquired through capillary electrophoresis taking several hours, was reduced to 20 min. using flow-through micro-array technology (Wu et al. (2004); Zeng et al. (2008)). Overall, starting from a blood sample, the time to get to a CNV result was ~ 6 h.

The assay was designed to detect aneusomies for chromosomes 13, 18, 21 and X. First the assay was tested using samples from 4 individuals with a known aberrant karyotype, incl. trisomy-13, -18, -21. The results obtained confirmed previous findings (Fig.2). Next the assay was tested using 23 blinded DNA samples of which a number had an abnormal karyotype. All sexes and all rearrangements were correctly scored. Two samples initially gave a discrepant result, one due to a mislabelling and the other because the individual involved had received treatment for sex-reversal, with the gender reported being her new sex.

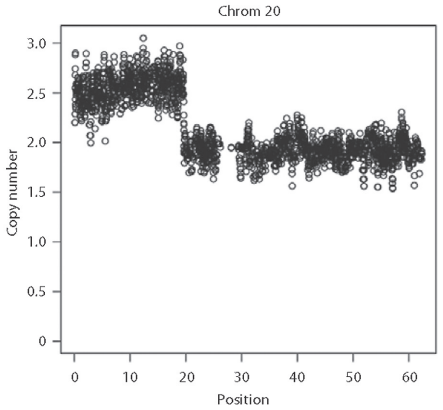
As a proof-of-principle and to check the time required we started with freshly taken blood samples of four healthy individuals. DNA was isolated, MLPA performed, array signals measured and scored. Based on the signal of the Y-chromosome probe and the altered ratios of the X-probes compared to those of the other chromosomes, we were able to correctly score the sex for all samples in about 6.5h. Thus far the assay has been applied once in a clinical setting, confirming a suspected trisomy-18 case directly after birth.

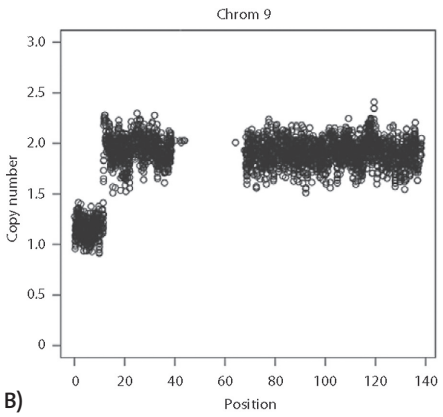
hrMCA-CNV

Now that genome-wide tools are applied in diagnostic labs worldwide, hundreds of new CNVs are being detected and there is an urgent need for a quick, reliable and cheap method to confirm the initial findings. Most CNV methodologies can not be easily used in a custom design setting or are labour and/or time intensive to develop and thus costly. While screening disease genes for sequence variants using high-resolution Melting Curve Analysis (BRCA - van der Stoep et al, DMD - Al-Momani et al, submitted) we were impressed by its resolution and sensitivity. We therefore tested its performance to confirm CNVs, simply by comparing the melting curves of different pre-defined mixes of the opposite alleles and the test sample.

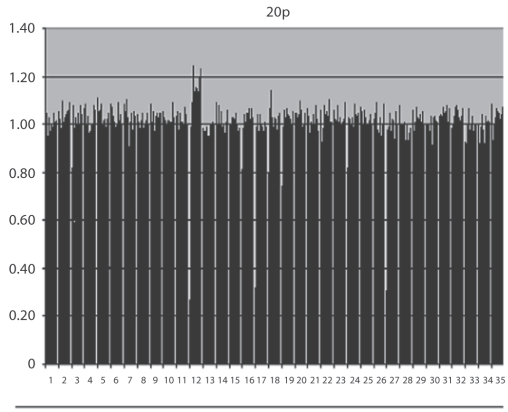


A)





B)



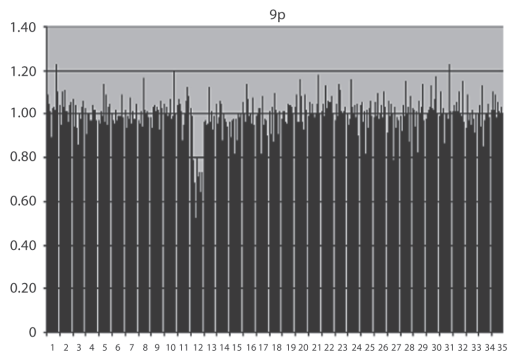


Figure 1: 1400-plex CNV bead assay. **A)** Strength of signal for probes on the X-chromosome obtained in males (M, X-axis) versus females (F, Y-axis). Signals are clearly different yet the difference in F:M signal obtained is not 1:0.5 as expected but 1:~0.7 (see also panel 1B bottom left and **1C**). **B)** Unbalanced translocation between chromosomes 20p (top, 1 extra copy) and 9p (bottom, 1 missing copy) detected in MR-patient 12 using the telomeric ruler probes (left panels). The rearrangement was confirmed using a SNP-array (right panels, Affymetrix 250K-StyI). Besides MR the patient had multiple congenital anomalies. **C)** Deletion spanning at least the COPS3 (top) and DRG2 (bottom) genes at chromosomes 17p11.2 detected in MR-patient 60 (confirmed to have Smith-Magenis syndrome). Note that in some cases only 1 of the probe pairs indicates a CNV (e.g. patient 18 for COPS3); such cases have not yet been considered.

C)

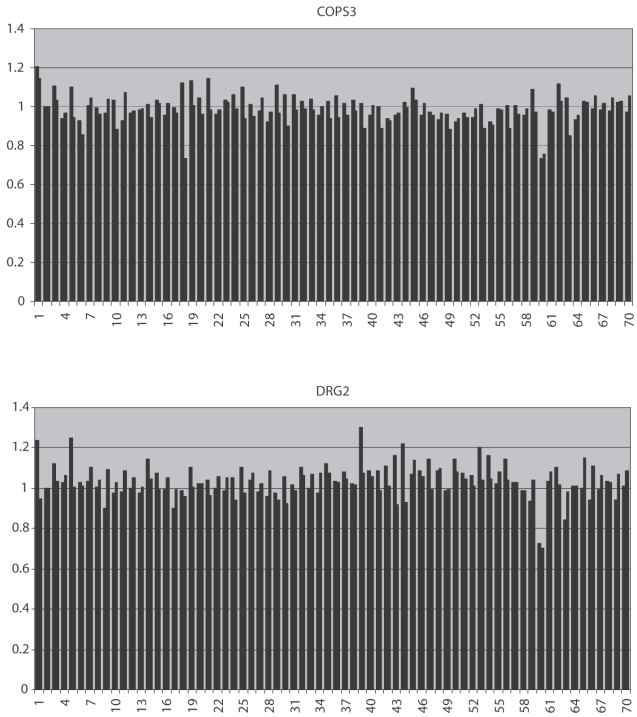


Figure 2: Fast-MLPA. Detection of a trisomy 21 (top left), Turner (45, Xo; top right) and 49, XXXXY case (bottom) using fast-MLPA.

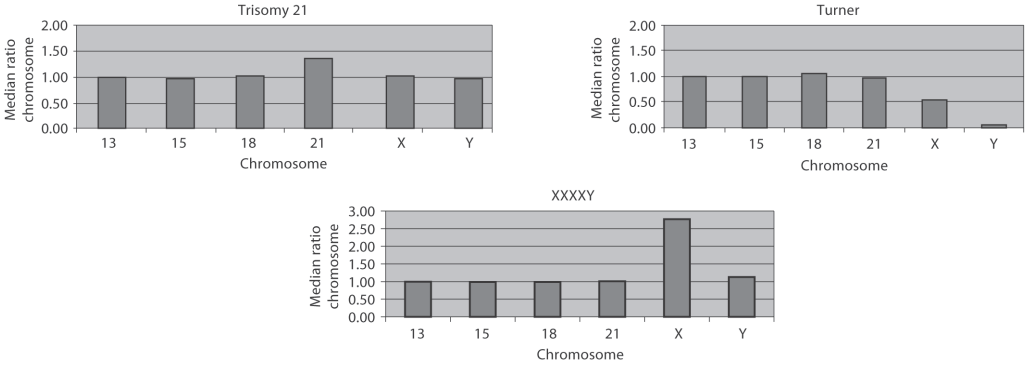


Table 2: Confirmation of CNVs using high-resolution Melting Curve Analysis (hrMCA)
Depending on the CNV to confirm the test sample, control samples (homozygote AA, AB and heterozygote AB) and specific mixes of test and control samples have to be prepared. Shared melting curves indicate the allelic composition of the test sample.

CNV	test sample	CNV confirmed when
Deletion	Ao?	Ao?/BB mix = 1:2 AA/BB control mix and differs from 1:1 AA/BB control mix and AB heterozygote
	Bo?	AA/Bo? mix = 2:1 AA/BB control mix and differs from 1:1 AA/BB control mix and AB heterozygote
Duplication	AAA?	AAA?/BB mix = 3:2 AA/BB control mix and differs from 1:1 AA/BB control mix and AB heterozygote
	AA?B	AA?B/BB mix = 2:3 AA/BB control mix and differs from 1:1 AA/BB control mix and AB heterozygote
	ABB?	AA/ABB? mix = 3:2 AA/BB control mix and differs from 1:1 AA/BB control mix and AB heterozygote
	BBB?	AA/BBB? mix = 2:3 AA/BB control mix and differs from 1:1 AA/BB control mix and AB heterozygote

The mix corresponding best with the melting curve of the test sample, resolves its allelic composition and confirms the CNV or not (Table 2).

An example of the sensitivity of the assay is shown in Fig.3. In the example alleles can be discriminated in steps of 12.5%, suggesting, besides confirming CNVs (~ 25% difference) it can be applied to type tetraploid and maybe octaploid organisms (plants). This is well within the sensitivity required to confirm CNVs where in mixed samples 3-5 alleles need to be discriminated. We have applied the hrMCA-CNV successfully to confirm both deletions and duplications detected using SNP-arrays as well as to determine the level of somatic mosaicism for CNVs in samples from monozygotic twins (Bruder et al. (2008)).

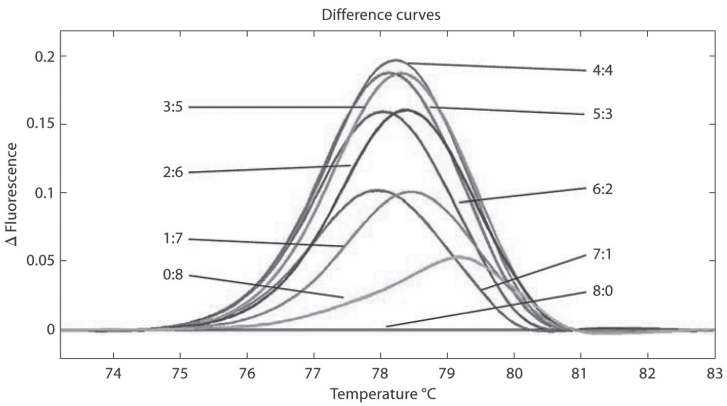
General considerations

There are many techniques and platforms available for detecting copy number changes in genomic DNA. The choice of the best method largely depends on project specific factors and the question to answer. Usually compromises have to be made with regard to the sample available, number of CNVs to analyse, resolution, cost and throughput. Not all CNV methods available can be performed with any sample. Microscopy/FISH methods require intact cells and/or nuclei. Southern blot analysis, especially when combined with Pulsed-

Field Gel-Electrophoresis, can only be performed when DNA of sufficient length is available (Van Der Maarel et al. (1999)). PCR-based methods are usually least demanding, although not all methods will work with low quality DNA (i.e. fragmented and/or contaminated (Kessler et al. (2004)). Overall methods that amplify short stretches of DNA perform better on low quality DNA compared to methods amplifying larger fragments. E.g. for whole genome array-based CNV studies, we obtained better results with SNP-arrays from Illumina then Affymetrix when low quality DNA samples were used. In rare cases the amount of DNA available might be limiting, making PCR-based methods most appropriate. However, even when only 1 or 2 cells are available, techniques like FISH are successfully applied to analyse CNVs in preimplantation diagnosis.

Multiplexability, i.e. the number of CNVs that can be studied simultaneously, often plays a decisive role. In most cases, sample throughput is inversely related with multiplexability; available methods do allow high-throughput analysis of a few loci but genome-wide CNV studies are difficult to perform for thousands of samples. In most cases the ease with which the assay can be automated determines the throughput that can be achieved. PCR-based methods are generally performed in single-locus mode and can be performed to analyse thousands of samples per week. MLPA, MAPH and MAQ facilitate the analysis of 20-50 loci using capillary electrophoresis and can be used to analyse hundreds of samples per week.

Figure 3: hrMCA CNV. High-resolution Melting Curve Analysis of two DNA samples being homozygous for the opposite SNP-alleles (rs213950:G>A) mixed in different proportions (from 8:0 to 0:8). The difference plots of all mixes can be easily discriminated, giving a sensitivity of at least 12.5%, suggesting that octaploid alleles could be typed.



To analyse more loci per sample (50-1000), either arrays (Zhou et al. (2004)) or bead-approaches (e.g. xMAP - Luminex, BeadArray - Illumina) can be applied. Such assays are ideal for diagnostic application, testing a range of targeted loci with known consequences only. The 1400-plex CNV bead assay described here is such an assay. The assay has the intrinsic possibility to score CNVs as well as specific SNPs or pathogenic mutations in one assay. Another design, which is also possible, is to use two different colours and either analyse two samples on one array (labelled in different colours, yielding improved data quality) or to double the number of loci analysed. When flexible custom-design is the most important aspect of the assay, the VeraCode technology (BeadXpress - Illumina) might have additional advantages, facilitating custom mixing of barcode-labelled probes.

Fast-MLPA tackles another, sometimes important aspect an assay, time-to-result. It allows a CNV-assay, from blood sample to result, to be performed within a working day. The PamChip technology used also facilitates real-time monitoring of hybridisation signals (Anthony et al. (2003)) and could be applied to analyse samples under very specific and highly stringent hybridisation conditions, e.g. to determine copy number of multi-copy sequences containing one or a few point mutations.

With the rapidly increasing popularity of whole genome CNV analysis using high-density SNP-arrays there is a great need for validation of the results obtained. While measuring 5-20 consecutive SNPs, Wagenstaller et al. (2007), analysing patients with mental retardation for CNV using 100K SNP-arrays, reported a false positive rate of 30%. Smaller aberrations, covering less then 5 probes, although true, are even more problematic. FISH, powerful yet laborious, is often not suitable for confirmation; cells/nuclei might not be available, probes difficult to get or spanning a region bigger then the CNV. MLPA is suitable, even for smaller aberrations, but developed to screen 20-40 loci simultaneously and taking 2-4 weeks to design (White et al. (2004)). The hrMCA-CNV method presented here is easy to perform and simple to design, requiring only SNPs from the region of interest. Assuming the CNVs were detected using a SNP-array, one often has a large and easy choice.

Future developments

Looking further ahead, the near future will undoubtedly see application of the new massive parallel sequencing technologies. Although simple and available for decades sequencing-based counting methods seem, with few exceptions (Bailey et al. (2002)), not to have been applied for CNV detection. The new massive parallel sequencing technologies seem to provide a new, very powerful tool for the detection of CNVs (e.g. using SAGE-like approaches); the resolution required determining the number of total sequences needed. In addition, applying paired-end sequencing, these new technologies will allow detection of all structural variation (Kidd et al. (2008)); CNVs as well as insertions, inversions and

translocations. In addition sequencing will be instrumental to determine the structure of the CNV as well as breakpoint sequences, the latter facilitating direct screening by breakpoint PCR.

One methodological hurdle remains to be solved; determining the exact copy number of multi-copy CNVs (>6-8 copies) in hundreds of samples and testing their possible association with specific diseases or phenotypic traits. For the latter to work, the assay should be exact and without errors, a demand where most current CNV methods already fail for copy numbers of 3-5, even when specific precautions are taken (Armour et al. (2007)). Massive deep sequencing technology might be applicable here but is still costly and not high-throughput. New innovative methods are under development, incl. automated fiber-FISH methods (BioNanomatrix) and NanoString's nCounter technology (Geiss et al. (2008)), but it is still too early to say whether they will be able to resolve this issue.

Materials and methods

Patient samples

All DNA samples were obtained from the department of Clinical Genetics (LUMC, Leiden). To check the performance of newly developed assays we used 44 control samples; 8 healthy individuals and 36 patients with known genomic rearrangements (incl. 23 DB/MD patients and carriers). In total 320 patients with mental retardation of unknown aetiology were analysed using the 1400-plex CNV bead assay, using gender information as an internal control. All subjects, or their representatives, gave informed consent for DNA studies.

Genomic DNA was isolated from blood samples using standard methods. DNA concentrations were measured using PicoGreen (Invitrogen-Molecular probes) and diluted to a concentration of 50ng/ul.

DNA isolation for fast-MLPA was performed using the Perfect gDNA Blood Mini kit (Qiagen), taking ~ 30 min; proteinase-K sample lysis (13 min), binding DNA to column (3 min), washing (4 min), elution (4min). MLPA required ~ 4.5 h; hybridization 2.5 h, ligation 20 min, PCR 1.5 h. PamChip®analysis ~ 20 min; hybridization 15 min, washing 3 min. Data analysis using ArrayPro; 10 min. Total assay time < 6 h.

1400-plex CNV bead assay

We wanted to develop a CNV assay that would facilitate the cost-effective screening of 1000 or more loci, with a flexible choice of loci to include (custom design) and the possibility of automated analysis of many samples. When we used an array-based readout (Zeng et al. (2008)) it clearly showed the advantage of using probes of equal length to obtain uniform

yields of all fragments amplified in a highly complex multiplex PCR. Illumina's Golden Gate assay uses a hybridisation-extension-ligation approach with similarities to MLPA (Fan et al. (2003)), allowing largely automated high-throughput screening of ~ 1600 loci using a bead-based micro-titer plate read-out format. Initial experiments confirmed the potential of this assay to detect CNVs, but also suggested that a non-SNP assay, i.e. using loci not covering SNPs, might have additional advantages (i.e. improved signal, simplified data analysis, doubling the number of loci that can be analysed).

For the 1400-plex CNV bead assay we designed in total 1324 non-SNP probes using standard design rules (Fan et al. (2003)). All probes for specific loci were designed in duplicate, i.e. separate for two closely spaced sequences, in unique sequences and where possible inside a gene (exonic). A telomeric ruler was created by designing probes at 0.5, 1.0, 1.5, 2.0, 3.0, and 4.0 Mb from the end of all chromosomes, except the p arm of the five acrocentric chromosomes (in total 482 probes). Regions known to be involved in micro-deletion / duplication syndromes were targeted with duplicate probes in at least one gene in every region selected. To determine the accuracy of the assay, control probes were designed for 19 exons in the *DMD* gene. 773 probes were designed for other loci of interest, incl. loci known to be copy number variable in a normal population (e.g. *CCL3L1, NSF*), containing other disease associated loci, potential regulatory regions and loci randomly spaced across the genome to provide a rough whole genome scan.

Per analysis we used 250 ng total genomic DNA following the manufacturer's recommendations. Labelled products were purified and hybridised to a Sentrix Array Matrix (SAM). After hybridisation the SAM was washed and imaged on the Illumina BeadArray Reader. The 44 control samples were used to evaluate the assay and probe performance. Both probes for one locus should give the same copy number; loci for which 85% or less of the samples gave a concordant outcome were omitted for further analysis. Probes showing an unexpected CNV in >5% of the controls were studied more carefully. These CNVs are either false positives (e.g. due to low probe quality) or true CNVs, i.e. located in hitherto unknown CNV polymorphic regions.

Intensity signals were extracted, imported to MS-Excel and analysed as described (White et al. (2004)). Ratios between 0.75 and 1.25 are regarded as a normal (i.e. two copies), below 0.75 as a CNV-loss and above 1.25 as a CNV-gain. Samples with poor DNA quality (defined as >10% of the control probes showing copy number variation within one sample) were excluded. The number of CNVs was calculated per locus, when three or more patient samples showed CNV for a locus it was regarded as polymorphic, when only one or two samples showed a CNV is was considered as a possible pathogenic variant.

Initial assessment of the array performance was made by analysing 44 samples containing known rearrangements. Overall signals were stable and the intensities of

the duplicate probes showed little variation. Comparing female and male samples for X chromosome probe signals (Fig 1) clearly separated both sexes, yet the difference in signal was not 1:0.5 as expected but only ~ 1:0.7 (Fig.1A), an as yet unexplained but known phenomenon (Pollack et al. (1999)). The possibility that the amount of hybridising material was saturating the available target was ruled out by performing the hybridisation with less material, which resulted in lower signal but no alteration in the derived ratios.

All known rearrangements were detected at exon resolution and all deletion / duplication breakpoints matched those known. The DMD deletions and duplications in 23 control samples were all detected, proving that CNVs involving a single probe could be ascertained. Although an increase in ratio is clearly seen when there are more then two copies present, it was not possible to distinguish between three and four copies of the *DMD* gene. This is not a significant problem however, as the primary purpose is to detect gains or losses per se. Where necessary to determine the precise copy number of a given locus, other methods can be applied.

Independent confirmation of CNVs was performed using several methods; custom design MLPA (White et al. (2004)), Fluorescence in situ hybridization (Dauwerse et al. (1992)) and whole genome SNP arrays (Illumina-317K and Affymetrix-5.0, performed according to the manufacturer's protocols and analysed using Beadstudio, CNAG, dChip). PAC/BAC clones for FISH were obtained from the Welcome Trust Sanger Institute (Cambridge, UK).

Fast-MLPA

Fast-MLPA probe design was basically performed as described (White et al. (2004)). Several rounds of probe optimisation were performed, selecting probe combinations that allowed short MLPA hybridisation, ligation and amplification times. The final MLPA probe set consisted of six probes per chromosome (13, 18, 21 and X) and one probe for chromosomes 15 and Y. Of the MLPA probe pair, each left-hand oligonucleotide contained a 20-nucleotide zip-sequence, facilitating selective hybridisation to a PamChip (Wu et al. (2004)), PamGene, Den Bosch, Nederland). All zip detector probes were spotted in duplicate on the array.

MLPA was performed as described (White et al. (2004)) with the exception of the hybridisation step (2.5 hours instead of overnight). PCR was performed for 33 cycles with either a Cy5- (control) or Cy3- (patient) labelled forward primer.

Array experiments were performd in the PamStation-4 or -FD10 (Pamgene). Before hybridisation, arrays were washed with 20µl 1x PBS-Tween (1 cycle of 1 min.) and 20µl 5x SSPE (1 cycle). Pre-hybridisation, 10 min. at 55°C, was carried out using 2µl tRNA (10µg/µl), 5µl 20x SSPE and13µl H_2O. Hybridization was performed using a mix containing 6µl patient DNA sample (Cy3-labelled), 6µl control DNA sample (Cy5-labelled), 1µl MAPHF and 1µl MAPHR primer (2 µM/µl each), 5µl 20x SSPE, 1µl H_2O. Hybridisation was for 10 min. at

55°C, the array was washed three times (one cycle) with fresh 20µl 5x SSPE-buffer followed by image capture. For experiments performed in the PamGene-FD10 all volumes were doubled.

Images were analysed with ArrayPro (Media Cybernetics) and the intensity data exported to MS-Excel. The median ratio was taken for normalization (White et al. (2004)). The average ratio of the duplicate spots was calculated and then the median ratio per chromosome. This gives tight ratios close to 1.0 with the advantage that lower thresholds can be set, allowing e.g. detection of mosaic cases, where an aneusomy is present in only a percentage of the cells, as we could prove by mixing DNA from a normal and a trisomy 21 case in different proportions.

hrMCA-CNV

High-resolution Melting Curve Analysis (hrMCA) was performed using the LightScanner (Idaho Technologies) and LightCycler-480 (Roche Diagnostics). PCR amplicons were designed using Primer-3 and standard design parameters, targeting fragments of 100-200 bp covering the SNP of interest. To confirm a CNV, two SNPs in the CNV region are selected. The resolution of this assay depends on the shift in melting curves between a homozygous and heterozygous sample, with larger shifts giving higher sensitivity. Therefore, we selected those SNPs from the candidate CNV region that are expected to give a large shift (i.e. A>G changes). Next to the sample containing the potential CNV, samples carrying the two opposite alleles homozygously (AA and BB) are required (Table 2).

Depending on the CNV to confirm, a deletion or duplication, a set of samples mixes are made. For deletions a 1:1 and 1:2 mix of the AA and BB and a 1:1 mix of the test "A0?" and opposite homozygous BB sample. When for a deletion the mixed A0?/BB sample shares its melting curve with that of the 1:2 AA/BB mix ("ABB") and not with the 1:1 AA/BB mix ("AB") nor with that of the heterozygote AB control, the deletion is confirmed. For the samples and mixes needed to confirm duplications see Table 2.

Acknowledgements

We like to acknowledge the invaluable help and advise of Jian-Bing Fan, Marina Bibikova, Lixin Zhou, Jing Chen and Eliza Wickham-Garcia (Illumina Inc., San Diego, USA) for developing the 1400-plex MLPA and Ying Wu and Rini van Beuningen (PamGene, Den Bosch, Nederland) while developing the fast-MLPA assays.

2

Anthony RM, Schuitema AR, Chan AB, Boender PJ, Klatser PR, Oskam L: Effect of secondary structure on single nucleotide polymorphism detection with a porous microarray matrix; implications for probe selection. BioTechniques 34:1082-1089 (2003).

Armour JA, Palla R, Zeeuwen PL, den HM, Schalkwijk J, Hollox EJ: Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic Acids Res 35:e19.1-19.8 (2007).

Armour JA, Sismani C, Patsalis PC, Cross G: Measurement of locus copy number by hybridisation with amplifiable probes. Nucl Acids Res 28:605-609 (20-1-2000).

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: Recent segmental duplications in the human genome. Science 297:1003-1007 (2002).

Beckmann JS, Estivill X, Antonarakis SE: Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8:639-646 (2007).

Beggs AH, Koenig M, Boyce FM, Kunkel LM: Detection of 98% of DMD/BMD gene deletions by polymerase chain reaction. Hum Genet 86:45-48 (1990).

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, de Stahl TD, Menzel U, Sandgren J, von Tell D, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, Van Ommen GJB, Boomsma DI, Pedersen NL, Den Dunnen JT, Wirdefeldt K, Dumanski JP: Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82:763-771 (2008).

Craig JE, Barnetson RA, Prior J, Raven JL, Thein SL: Rapid detection of deletions causing delta beta thalassemia and hereditary persistence of fetal hemoglobin by enzymatic amplification. Blood 83:1673-1682 (1994).

Dauwerse JG, Wiegant JCAG, Raap AK, Breuning MH, Van Ommen GJB: Multiple colors by fluorescence in situ hybridization using ratio-labelled DNA probes create a molecular karyotype. Hum Mol Genet 1:593-598 (1992).

Den Dunnen JT, Bakker E, Klein-Breteler EG, Pearson PL, Van Ommen GJB: Direct detection of more than 50% Duchenne muscular dystrophy mutations by field-inversion gels. Nature 329:640-642 (1987).

Den Dunnen JT, Grootscholten PM, Bakker E, Blonden LAJ, Ginjaar HB, Wapenaar MC, Van Paassen HMB, Van Broeckhoven C, Pearson PL, Van Ommen GJB: Topography of the DMD gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. Am J Hum Genet 45:835-847 (1989).

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS: Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69-78 (2003).

Florijn RJ, Blonden LAJ, Vrolijk H, Wiegant J, Vaandrager JW, Baas F, Den Dunnen JT, Tanke HJ, Van Ommen GJB, Raap AK: High-resolution FISH for genomic DNA mapping and colour bar-coding of large genes. Hum Mol Genet 4:831-836 (1995).

Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L: Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 26:317-325 (2008).

Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA, Ling V, MacAulay C, Lam WL: A tiling resolution DNA microarray with complete coverage of the human genome. Nat Genet 36:299-303 (2004).

Kessler N, Ferraris O, Palmer K, Marsh W, Steel A: Use of the DNA flow-thru chip, a three-dimensional biochip, for typing and subtyping of influenza viruses. J Clin Microbiol 42:2173-2185 (2004).

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE: Mapping and sequencing of structural variation from eight human genomes. Nature 453:56-64 (2008).

Knight SJ, Lese CM, Precht KS, Kuc J, Ning Y, Lucas S, Regan R, Brenan M, Nicod A, Lawrie NM, Cardy DL, Nguyen H, Hudson TJ, Riethman HC, Ledbetter DH, Flint J: An optimized set of human telomere clones for studying telomere integrity and architecture. Am J Hum Genet 67:320-332 (2000).

Landegent JE, Jansen in de Wal N, Van Ommen GJB, Baas F, De Vijlder JJM, Van Duijn P, Van der ploeg M: Chromosomal localization of a unique gene by non- autoradiographic in situ hybridization. Nature 317:175-177 (1985).

Leujeune J, Gautier M, Trupin R: Study of somatic chromosomes from 9 mongoloid children. C R Hebd Seances Acad Sci 248:1721-1722 (1959).

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC: Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256-1260 (2007).

Petrij-Bosch A, Peelen T, Van Vliet M, Van Eijk R, Olmer R, Drusedau M, Hogervorst FBL, Hageman S, Arts PJ, Ligtenberg MJ, Meijers-Heijboer H, Klijn JG, Vasen HF, Cornelisse CJ, Van 't Veer LJ, Bakker E, Van Ommen GJB, Devilee P: BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. Nat Genet 17:341-345 (1997).

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207-211 (1998).

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 23:41-46 (1999).

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

Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerriere A, Vital A, Dumanchin C, Feuillette S, Brice A, Vercelletto M, Dubas F, Frebourg T, Campion D: APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet 38:24-26 (2006).

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G: Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30:e57 (2002).

Suls A, Claeys KG, Goossens D, Harding B, Van Luijk R, Scheers S, Deprez L, Audenaert D, Van Dyck T, Beeckmans S, Smouts I, Ceulemans B, Lagae L, Buyse G, Barisic N, Misson JP, Wauters J, Del Favero J, De Jonghe P, Claes LR: Microdeletions involving the SCN1A gene may be common in SCN1A-mutation-negative SMEI patients. Hum Mutat 27:914-920 (2006).

Van Der Maarel SM, Deidda G, Lemmers RJ, Bakker E, van der Wielen MJ, Sandkuijl L, Hewitt JE, Padberg GW, Frants RR: A new dosage test for subtelomeric 4;10 translocations improves conventional diagnosis of facioscapulohumeral muscular dystrophy (FSHD). J Med Genet 36:823-828 (1999).

Wagenstaller J, Spranger S, Lorenz-Depiereux B, Kazmierczak B, Nathrath M, Wahl D, Heye B, Glaser D, Liebscher V, Meitinger T, Strom TM: Copy-number variations measured by single-nucleotide-polymorphism oligonucleotide arrays in patients with mental retardation. Am J Hum Genet 81:768-779 (2007).

White SJ, Den Dunnen JT: Copy number variation in the genome; the human DMD gene as an example. Cytogenetic and Genome Research 115:240-246 (2006).

White SJ, Vink GR, Kriek M, Wuyts W, Schouten JP, Bakker E, Breuning MH, Den Dunnen JT: Two-colour MLPA; detecting genomic rearrangements in hereditary multiple exostoses. Hum Mutat 24:86-92 (2004).

Wu Y, De Kievit P, Vahlkamp L, Pijnenburg D, Smit M, Dankers M, Melchers D, Stax M, Boender PJ, Ingham C, Bastiaensen N, de Wijn R, van Alewijk D, Van Damme H, Raap AK, Chan AB, Van Beuningen R: Quantitative assessment of a novel flow-through porous microarray for the rapid analysis of gene expression profiles. Nucleic Acids Res 32:e123.1-e123.7 (2004).

Yau SC, Bobrow M, Mathew CG, Abbs SJ: Accurate diagnosis of carriers of deletions and duplications in Duchenne/Becker muscular dystrophy by fluorescent dosage analysis. J Med Genet 33:550-558 (1996).

Zeng F, Ren ZR, Huang SZ, Kalf M, Mommersteeg M, Smit M, White S, Jin CL, Xu M, Zhou DW, Yan JB, Chen MJ, van BR, Huang SZ, Den Dunnen JT, Zeng YT, Wu Y: Array-MLPA: comprehensive detection of deletions and duplications and its application to DMD patients. Hum Mutat 29:190-197 (2008).

Zhou X, Mok SC, Chen Z, Li Y, Wong DT: Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. Hum Genet 115:327-330 (2004).


3

Split hand-foot malformation, tetralogy of Fallot, mental retardation and a 1 Mb 19p deletion-evidence for further heterogeneity?

Emmelien Aten, Nicolette den Hollander, Claudia Ruivenkamp, Jeroen Knijnenburg, Hans van Bokhoven, Johan den Dunnen, Martijn Breuning

Am J Med Genet A. 2009;149A(5):975-81.

Abstract

Congenital limb malformations are the second most common birth defects observed in infants. Split hand foot malformation (SHFM), also known as central ray deficiency, ectrodactyly and cleft hand/foot, occurs isolated or in combination with other malformations.

We report a male patient with SHFM, tetralogy of Fallot and a clinical phenotype of Angelman syndrome. Using array based genome analysis (3K BACs and 500K SNPs), we identified a *de novo* deletion of chromosome 19p13.11, confirmed by Fluorescent *In Situ* Hybridization analysis. The deletion is 0.99 Mb in size and contains 28 genes. The proximal breakpoint of the deletion is in *EPS15L1*, which may be involved in vertebrate limb development.

Subsequent screening of 21 syndromic and non-syndromic SHFM patients (*TP73L* mutation negative) for rearrangements using Multiplex Ligation-dependent Probe Amplification did not detect other deletions or duplications in chromosome 19. These findings suggest that our patient may have a new contiguous gene syndrome and indicates that SHFM is genetically more heterogeneous than currently known.

Key words

Mental retardation - Split Hand Foot Malformation - Tetralogy of Fallot- -19p deletion

Introduction

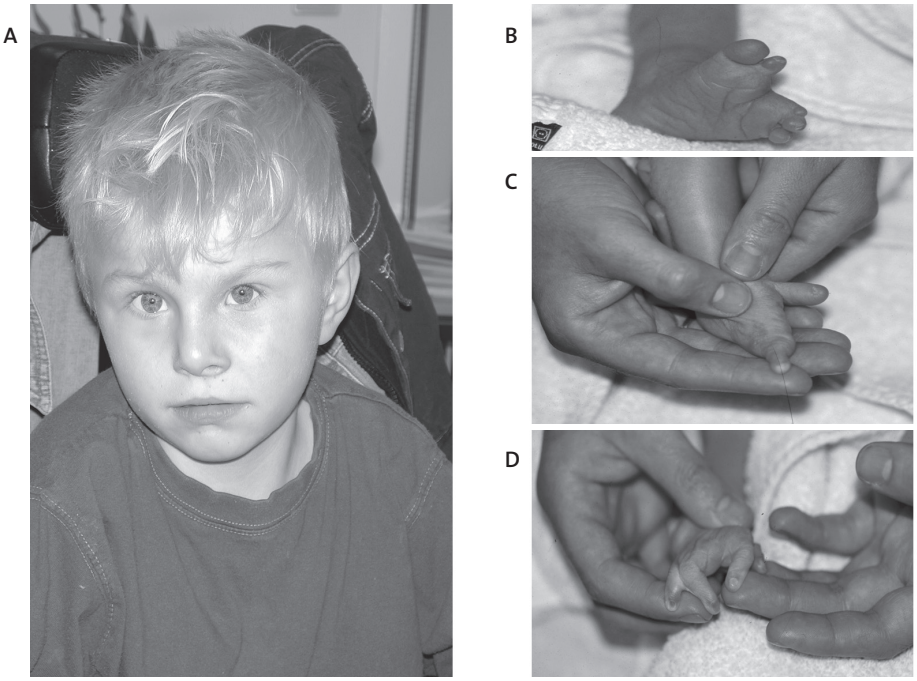
The variability of Split Hand Foot Malformation (SHFM) causes the classification to be very difficult. The hand and foot abnormalities range greatly in their phenotypic spectrum from mild abnormalities (i.e. digital shortness) to more severe abnormalities (i.e. adactyly), occurring in a single or in all four limbs. SHFM occurs either as an isolated malformation ('non-syndromic') or in association with other malformations ('syndromic'). To date several syndromic forms have clinically been identified (EEC OMIM#129900, EEM OMIM#225280, LMS OMIM#603543, MMEP OMIM#601349, KNS OMI M#183800 and ADULT OMIM #103285). Identification of genes involved in defects in human limb patterning has not yet been very successful, in part, by the enormous spectrum of phenotypes.

For the development of the limb, specialized cell clusters are important. The SHFM phenotype is believed to be caused by an underlying defect of one of the cell clusters named the apical ectodermal ridge (AER). Both genetic defects and environmental factors may cause SHFM by interfering with the AER [Duijf et al., 2003; Seto et al., 1997]. Because of expression in the AER, candidate genes in the different loci have been postulated but mutations have not yet been identified. Currently five genomic loci, based on case and family studies, have been implicated. They are known as SHFM1 (chromosome region 7q21-q22), SHFM2 (Xq26), SHFM3 (10q24), SHFM4 (3q27), and SHFM5 (2q31). Molecular testing is only available for SFHM4 since most EEC syndrome patients and 5-10% of non-syndromic SHFM patients have been shown to carry mutations in the *TP73L* (*p63*) gene [Duijf et al., 2003; de Mollerat et al., 2003; Brunner et al., 2002].

So far copy number variation (CNV, i.e. the presence of deletions or duplications) in patients with SHFM has only been described in SHFM1, SHFM3 and SHFM5. For SHFM1 only deletions have been described, mostly resulting from complex rearrangements and unbalanced translocations [Scherer et al., 1994; Bernardini et al., 2008; Sharland et al., 1991]. To explain inconsistent translocation/inversion breakpoints and failure to identify mutations in the candidate genes, a position effect has been postulated. For SHFM3 only duplications of 10q24 have been described. Extensive study of the candidate gene Dactylin did not reveal how the duplication leads to the SHFM3 phenotype [Everman et al., 2006; de Mollerat et al., 2003]. SHFM5 has been associated with interstitial deletions suggesting removal of a 'digit enhancer' or genes involved in the AER [Boles et al., 1995; Goodman et al., 2002; Bijlsma et al., 2005]. For SHFM, no large deletion or duplication events on other chromosomes have been described.

Here, we report a patient who manifested SHFM and who was found to have a 0.99 Mb deletion on chromosome 19.

Figure 1: Clinical features of the index patient. (A) Phenotype of the patient (here six years old) with subtle facial features including a wide mouth, fair hair and light skin. Central ray deficiencies as detected at birth. Left foot (B) with a cleft, hypoplasia of the second digit and a nubbin representing the third digit. Right foot (C) with a proximally placed first digit and presence of digit V. Digits II-IV are absent. Right hand (D) with a central cleft leading to absence of the third digit. Digit IV and V show cutaneous syndactyly.



Clinical report

The male patient (Fig. 1A) was born at term after an uneventful pregnancy. He is the second child of healthy unrelated parents. Family history for SHFM or mental retardation was negative.

At birth, congenital malformations of all extremities were conspicuous. Examination revealed a central ray deficiency of both hands and ectrodactyly of both feet. (Fig. 1B and 1C) The left-hand showed preaxial polydactyly with a bifid thumbnail, absence of the third digital ray and cutaneous syndactyly of the first and second digit and the fourth and fifth digit. The right-hand was a split hand with central ray deficiency and a cutaneous

syndactyly of the fourth and fifth digit (Fig.1D). The left foot showed a split foot with a hypoplastic second digital ray and a rudimentary third digital ray. There was cutaneous syndactyly of the third and fourth digital ray. The right foot showed ectrodactyly where only the first and fifth digital rays where present. The first digital ray was positioned far too proximally. A Ventricular Septum Defect (VSD) and stenosis of the pulmonal artery was suspected, which was later diagnosed as a tetralogy of Fallot. At the time of re-evaluation in the clinical genetics department the patient was two years old. His length was 88 cm (-0.3 SD) and head circumference 47 cm (-1 SD). Physical examination showed few dysmorphic features, brachycephaly with a flat occiput, slight upslant of his eyes, and broad mouth with a thin upper lip. He had very fair hair. His psychomotor development was slightly behind but this was considered to be normal given his physical problems early in life. Surgical correction of both the SHFM and the tetralogy of Fallot had taken place. Corrective surgery for strabismus was planned. MRI of the cranium and ultrasound of the kidneys was performed but showed no aberrations. EEG did not show epileptiform activity. The patient was re-evaluated at the age of six. By this time he had a severe delay in psychomotor development. His speech was severely impaired (4-5 words). He showed characteristic behavior with frequent laughing. Together with the previously described dysmorphic features this strongly suggested to the diagnosis of Angelman syndrome. Karyotyping did not reveal any chromosomal abnormalities (46, XY). M-FISH was normal and mutation screening for the *TP73L* gene as well as the *UBE3A* gene (causing Angelman syndrome) were negative. Also, deletions or methylation defects of *UBE3A* or UPD of chromosome 15 could not be detected. Evaluation of the patient at the current age of 10 years did not show any significant changes in phenotype.

Materials and methods

Subjects

The index patient was referred to our department for genetic counseling. His phenotype was established by examination of a clinical geneticist. Informed consent for genetic analyses was obtained from both his parents. Molecular testing was performed on whole blood genomic DNA.

Mutation analysis of the *UBE3A* gene and the *TP73L* gene were performed in diagnostic laboratories in the Netherlands (Erasmus MC Rotterdam, Radboud MC Nijmegen).

21 syndromic and non-syndromic SHFM patients from the Department of Clinical Genetics (Radboud MC Nijmegen) were ascertained for a MLPA based study and collected on the basis of clinical (and radiographic) findings. Patients were known to be negative for

TP73L mutations and had normal karyotypes. Array based genome analysis has not been performed on these patients. Informed consent was obtained from all participants.

DNA analysis

DNA was isolated from blood samples using standard methods. DNA concentrations were measured using a nanodrop.

Fish

Fluorescence in situ hybridization (FISH) was performed following standard protocols [Dauwerse et al., 1992]. The PAC/BAC clones RP11-413M18, CTD-2231E14 and CTD-3149D2, located at 19p13.11, were used for confirmation of the deletion. Clones were obtained from the Welcome Trust Sanger Institute, Cambridge, UK.

Array platforms

For array-comparative genomic hybridization (CGH), 1Mb spaced large insert clone arrays were used [Knijnenburg et al., 2005]. This platform contains ~3500 large insert clones spotted in triplicate at approximately 1Mb density over the full genome. The clone set was distributed by the Wellcome Trust Sanger Institute. All laboratory processing and hybridizations were performed according to published protocols [Fiegler et al., 2003].

The Affymetrix 500K oligonucleotide array was used according to manufacturer's protocols (http://www.affymetrix.com). Regions of copy number gain and loss were detected using the hidden markov model output of CNAG 2.0 [Nannya et al., 2005]. Deletions detected were described according to the recommendations of the HGVS (http://www.hgvs.org/mutnomen/).

Multiplex Ligation-dependent Probe Amplification

In total, 22 MLPA probe pairs were designed for genes of interest in and outside the deleted 19p region (Table 1).The probes were divided between two probe sets. Within each probe set, control probes for unlinked loci were included as a reference. Sequences are available on request. Peak ratios between 0.75 and 1.25 are considered normal (i.e. two copies).

Probe design and the MLPA reaction and analysis were performed as described [White et al., 2004; Schouten et al., 2002] Genescan (Applied Biosystems) and Genemarker 1.51 (Softgenetics) were used for data analysis.

Table 1: Genes on Chromosome 19p13.11

RefSeq DNA ID	Gene Name	Mlpa probe	Gene Start (bp)	Gene End (bp)
NM_032493	AP1M1	X	16169731	16207149
NM_016270	KLF2	X	16296637	16299683
NM_021235	EPS15L1	X	16333408	16443762
NM_145046	CALR3	X	16450888	16468003
NM_032207	C19orf44	X	16468205	16493163
NM_006387	CHERP	X	16489700	16514263
NM_024881	SLC35E1	X	16523584	16544193
NM_004831	MED26	X	16546718	16600015
NM_024104	C19orf42	X	16617967	16631968
NM_024074	TMEM38A	X	16632938	16660814
NM_001007525	NWD1		16703001	16789756
NM_015260	SIN3B	X	16801218	16852164
NM_003950	F2RL3	X	16860826	16863830
NM_015692	CPAMD8	X	16864765	16998625
NM_033417	AC020908.7		17021573	17047343
NM_004145	MYO9B	X	17073466	17185093
NM_018467	USE1		17187155	17191638
NM_024578	OCEL1	X	17198055	17201027
NM_005234	NR2F6	X	17203694	17217151
NM_031941	USHBP1	X	17221838	17236573
NM_014173	C19orf62		17239232	17251162
NM_152363	ANKLE1	X	17253454	17259455
NM_024527	ABHD8	X	17263941	17275282
NM_023937	MRPL34	X	17277477	17278652
NM_024050	DDA1		17281350	17292098
NM_020959	TMEM16H		17295034	17306638
NM_133644	GTPBP3		17309379	17314530
NM_031310	PLVAP		17323264	17349159
NM_004335	BST2	X	17374750	17377401
NM_138401	FAM125A		17391856	17397140
NM_138454	NXNL1		17427236	17432725
NM_198580	SLC27A1	X	17442300	17477977

Genes inside the deleted 19p11 region (indicated in yellow) and outside the deleted region (indicated in white). MLPA probes were designed for 22 genes (human reference sequence NCBI Build 36.1)

Results

Index Patient

Array-CGH revealed a deletion of 1 BAC clone RP11-413M18. Flanking clones (CTD-3149D2, CTD-2231E14) on chromosome19p13.11 were not deleted, giving an estimated maximal deleted size of 1.44 Mb (Fig. 2A).

FISH analysis using RP11-14M18 confirmed the deletion in chr.19p13.11, with flanking probes CTD-3149D2 centromeric and CTD-2231E14 telomeric giving normal signals (data not shown).

To determine the extent of the deletion more precisely, a 500 K oligonucleotide array (Affymetrix) was used. This array showed a deletion (rs12460131_rs10411936)_(rs4808641_rs6512194). Rs10411936, the first deleted SNP, is positioned in an intron of *EPS15L1*. Rs4808641, the last deleted SNP, is a SNP localized between two genes (*FAM125A* and a novel gene). The deletion spans ~ 0.99 Mb and contains 28 genes (Fig. 2B).

Figure 2: (A) Array-CGH showed one deleted BAC clone (RP11-413M18) on 19p13.11 at 16.8 Mb with a maximal deleted size of 1.44 Mb. **(B)** SNP array refined the deletion from rs10411936 (19: 16,409,375) - rs4808641 (19:17,408,292) with a maximal deleted size of 0.99 Mb

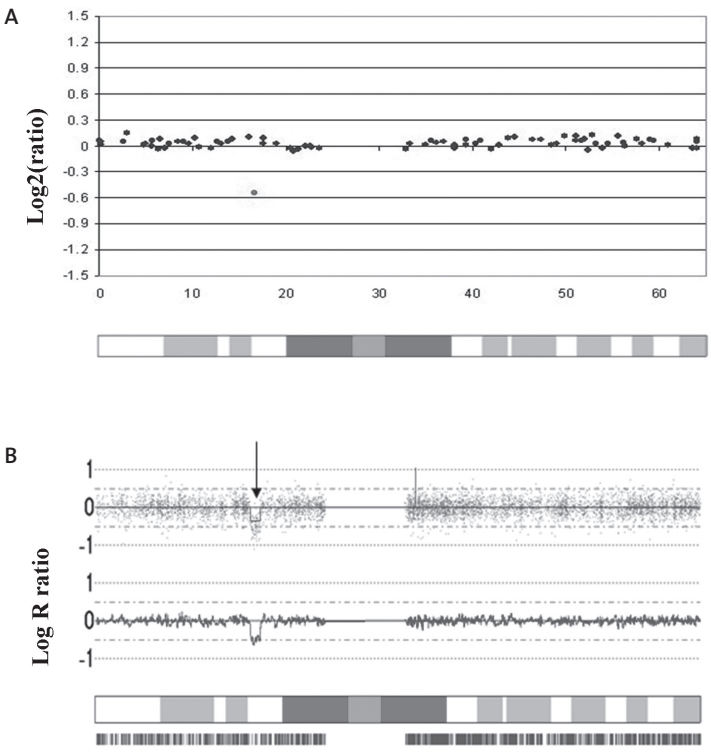
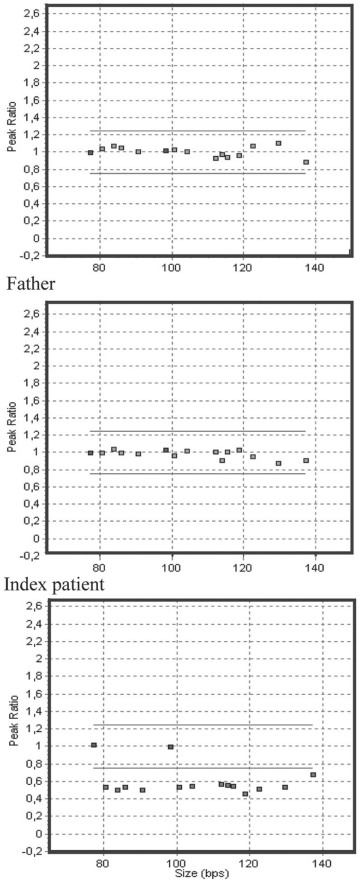
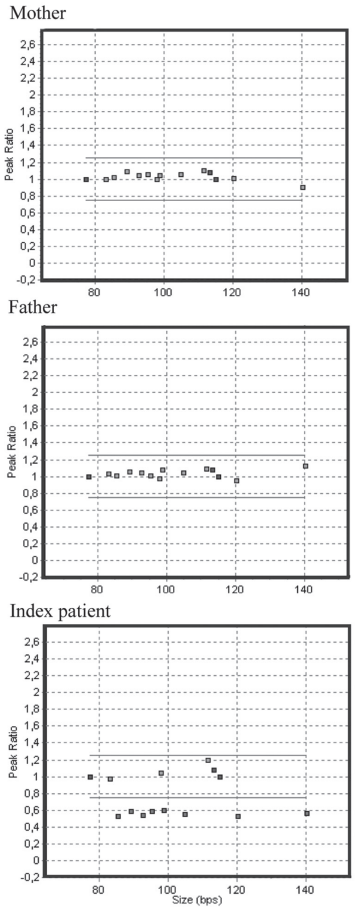


Figure 3: MLPA results for the index patient and his parents. **(A)** Peak ratio plots for the two probe sets, confirming the deletion to be *De Novo*. Control probes are represented by blue squares. **(B)** Peak ratio results for the mother (1), the father (2) and the index patient (3). Peak ratios in the index patient are normal for probes located in KLF2, AP1M1 and SLC27A1, coinciding with the proximal and distal deletion breakpoints.

A



B



		REF -	1	2	3
1	*control1	77.2	0.996	0.989	1.016
2	ANKLE1_1	80.4	1.014	0.970	0.520
3	ANKLE1_2	83.8	1.066	1.037	0.501
4	CPAMD8	86.0	1.044	0.990	0.526
5	EPS15L1	90.6	1.000	0.979	0.500
6	*control2	98.5	1.018	1.026	0.995
7	MYO9B	100.7	1.030	0.962	0.534
8	MED26	104.2	1.000	1.018	0.541
9	USHBP1	112.1	0.923	0.999	0.570
10	ABHD8	114.1	0.969	0.904	0.559
11	CALR3	115.6	0.917	0.979	0.538
12	BST2	118.9	0.952	1.012	0.456
13	TMEM38A	122.6	1.071	0.943	0.509
14	C19ORF44	129.8	1.088	0.866	0.529
15	SLC35E1	137.6	0.881	0.898	0.671

		REF -	1	2	3
1	*control1	77.4	1.000	1.000	1.000
2	KLF2	83.4	1.016	1.051	0.991
3	F2RL3	85.5	1.097	1.083	0.566
4	ANKLE1	89.5	1.102	1.069	0.590
5	SIN3B	92.9	1.088	1.089	0.567
6	CHERP	95.3	1.086	1.041	0.605
7	SLC27A1	97.9	1.016	0.988	1.053
8	OCEL1	99.0	1.053	1.089	0.601
9	NR2F6	105.0	1.070	1.059	0.563
10	AP1M1	111.6	1.106	1.094	1.196
11	*control2	113.2	1.076	1.072	1.072
12	*control3	115.2	1.000	1.000	1.000
13	MRPL34	120.3	1.024	0.970	0.534
14	C19ORF42	140.3	0.898	1.113	0.560

To study the deletion on chromosome 19p13.11 in more detail and to facilitate the screening of a larger set of patient samples, we designed a MLPA kit for twenty-two genes in the 19p13.11 region. MLPA analysis confirmed the deletion on chr. 19p13.11 in our patient. The deletion was not present in either parent, so proven to be *de novo* (Fig. 3A). Genes located proximal (*KLF2, AP1M1*) and distal (*SLC27A1*) from the deleted region were present in two copies in the index patient (Fig 3B).

SHFM patients

Twenty-one SHFM (TP73L mutation negative) patients were collected to asses whether copy number variation on chromosome 19p13.11 occurred more frequent in this patient group. The MLPA kit was used for additional screening of rearrangements in the 19p13.11 region. This did not reveal copy number changes in any of the tested patients.

Discussion

We present a case study in which a *de novo* deletion of chromosome 19p13.11 was found. The patient we report has, apart from SHFM, a congenital heart defect and mental retardation with a clinical phenotype suspected of Angelman syndrome. None of the genes within this deletion could obviously be related to the phenotype. Since the SHFM was such a salient feature in our patient, we decided to focus on limb malformations. This is the first case where a deletion on chromosome 19 has been associated with SHFM. The proximal breakpoint of the *de novo* deletion is in *EPS15L1*, a gene coding for epidermal growth factor receptor pathway substrate 15-like 1. The encoded protein seems to be a constitutive component of clathrin-coated pits that is required for receptor-mediated endocytosis. EPS15L1, like its homologue Eps15 on chromosome 1, functions as a substrate for the tyrosine kinase activity of the epidermal growth factor receptor (EGFR) [Wong et al, 1995].The epidermal growth factor receptor signaling pathway is one of the most important pathways that regulate growth, survival, proliferation and differentiation. It has been implicated in vertebrate and invertebrate limb development, including the AER [Simcox et al., 1997; Omi et al., 2005; Clifford et al., 1989]. Moreover, there is evidence that p63 regulates EGFR expression [Nishi et al., 2001], suggesting *EPS15L1* to be a serious candidate gene for SHFM. However, no copy number changes in *EPS15L1* were found in any of the other SHFM patients suggesting that a deletion or duplication of this gene is not frequently involved in the pathogenicity of SHFM. Nevertheless, point mutations as a possible mechanism can not be excluded.

To exclude Copy Number Variations in the SHFM1-SHFM5 loci as a possible cause for the ectrodactyly, both our index patient and the 21 SHFM patients were tested with MLPA. A candidate gene in each SHFM locus (*DSS1, FGF13, FBXW4, TP73L, DLX2*) was tested. Our patient did not show any dosage abnormalities in these genes.

A previous linkage study on EEC syndrome in a large Dutch family originally mapped a gene to a region of chromosome 19 comprising 19p13.11 [O'Quinn et al., 1998]. However, a causative mutation in the *p63* gene on chromosome 3q27 was later identified in this family, which refuted the presence of an SHFM locus on chromosome 19 [van Bokhoven et al., 2001]. However, the possibility of a chromosome 19 linked modifier effect in this family cannot be excluded. Copy number changes in chromosome 19p13 have been described by Thienpont et al., et al. using array-CGH to study patients with congenital heart defects [Thienpont et al., 2007]. They describe a patient (DECIPHER CHG00001031) with a *de novo* duplication (19) (p13.12-p13.11) which has recently been shown to be a more complex intrachromosomal rearrangement including a microduplication and two microtriplications spanning 3.2 Mb [Thienpont et al., 2008]. A small part of the first triplication and the major part of the duplication overlaps the region deleted in our patient, thus containing the same genes. Interestingly, the clinical phenotype of mental retardation, microcephaly, muscular VSD, short stature, strabismus and speech delay partly resembles the phenotype of our patient. However, the lack of overlap in facial dysmorphism and the absence of SHFM are under dispute. In general in microdeletion/duplication syndromes, phenotypes corresponding to the duplication are different and milder than those found in deletion cases. In OMIM none of the 19p13.11 deleted genes in our patient have been implicated with disease. Only a common variant located in the *MYO9B* gene has been associated with celiac disease [Monsuur et al., 2005].

At present, the interpretation of detected copy number changes is not straightforward, given the fact that as much as 11% of our genome shows copy number variations that do not appear to be clearly disease-related [Redon et al., 2006]. To distinguish pathogenic from innocent variants not only data from patients should be collected, but especially those found in 'normal' individuals. The Database of Genomic Variants (http://projects.tcag.ca; version April 2008), in which variants not known to cause disease are collected, reports one case with a duplication on chromosome 19 (19:16,998,070-16,999,171) that lies within our deleted region [de Smith et al., 2007]. The duplicated gene *CPAMD8* encodes the C3 and PZP-like, alpha-2-macroglobulin domain containing 8. This protein belongs to a family of the complement component-3, involved in innate immunity and damage control. No other genes within our deletion have been reported in the Database of Genomic Variants.

Elucidating the underlying cause of known and new syndromes remains highly challenging.

Copy Number Variants can impact gene function in several ways. The formation of a fusion gene as a possible pathogenic mechanism is not plausible because the distal border of our deletion is intergenic. The phenotype of our patient may be due to a dosage effect on one or multiple genes but it is still very likely that other mechanisms underlie pathogenicity such as disruption of a regulatory element that influences flanking genes. Notably, no miRNA's are encoded by genes in the deleted area.

Based on the phenotype of our patient, the size and high gene content of the deletion, the fact that it is *de novo* and its absence in the Database of Genomic Variants, we believe that this deletion is causative and represents a new contiguous gene syndrome. Moreover, our finding suggests SHFM to be even more genetically heterogeneous than previously suggested. Collecting larger sets of patient data in databases such as DECIPHER and ECARUCA may confirm this assumption, and should ultimately help to narrow down the critical region, and eventually pinpoint relevant genes or regulatory sequences.

Acknowledgements

We gratefully acknowledge the patient and his parents for giving their collaboration to publish clinical data and photos. We would like to thank Cathy Bosch for technical assistance and data analysis. This work was supported by ZonMw 912-04-0417.

Bernardini, L., C. Palka, C. Ceccarini, A. Capalbo, I. Bottillo, R. Mingarelli, A. Novelli, and B. Dallapiccola. 2008. Complex rearrangement of chromosomes 7q21.13-q22.1 confirms the ectrodactyly-deafness locus and suggests new candidate genes. Am J Med Genet Part A 146A:238-244.

Bijlsma, E. K., A. C. Knegt, C. M. Bilardo, and F. R. Goodman. 2005. Increased nuchal translucency and split-hand/foot malformation in a fetus with an interstitial deletion of chromosome 2q that removes the SHFM5 locus. Prenatal Diagnosis 25:39-44.

Boles, R. G., B. R. Pober, L. H. Gibson, C. R. Willis, J. Mcgrath, D. J. Roberts, and T. L. Yangfeng. 1995. Deletion of Chromosome 2Q24-Q31 Causes Characteristic Digital Anomalies - Case-Report and Review. American Journal of Medical Genetics 55:155-160.

Brunner, H. G., B. C. J. Hamel, and H. van Bokhoven. 2002. P63 gene mutations and human developmental syndromes. Am J Med Genet 112:284-290.

Clifford, R. J. and T. Schupbach. 1989. Coordinately and differentially mutable activities of torpedo, the Drosophila melanogaster homolog of the vertebrate EGF receptor gene. Genetics 123:771-787.

Dauwerse, J. G., E. A. Jumelet, J. W. Wessels, J. J. Saris, A. Hagemeijer, G. C. Beverstock, G. J. B. Vanommen, and M. H. Breuning. 1992. Extensive Cross-Homology Between the Long and the Short Arm of Chromosome 16 May Explain Leukemic Inversions and Translocations. Blood 79:1299-1304.

de Mollerat, X. J., F. Gurrieri, C. T. Morgan, E. Sangiorgi, D. B. Everman, P. Gaspari, J. Amiel, M. J. Bamshad, R. Lyle, J. L. Blouin, J. E. Allanson, B. Le Marec, M. Wilson, N. E. Braverman, U. Radhakrishna, C. Delozier-Blanchet, A. Abbott, V. Elghouzzi, S. Antonarakis, R. E. Stevenson, A. Munnich, G. Neri, and C. E. Schwartz. 2003. A genomic rearrangement resulting in a tandem duplication is associated with split hand-split foot malformation 3 (SHFM3) at 10q24. Hum Mol Genet 12:1959-1971.

de Smith, A. J., A. Tsalenko, N. Sampas, A. Scheffer, N. A. Yamada, P. Tsang, A. Ben Dor, Z. Yakhini, R. J. Ellis, L. Bruhn, S. Laderman, P. Froguel, and A. I. F. Blakemore. 2007. Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Hum Mol Genet 16:2783-2794.

Duijf, P. H. G., H. van Bokhoven, and H. G. Brunner. 2003. Pathogenesis of split-hand/split-foot malformation. Hum Mol Genet 12:R51-R60.

Everman, D. B., C. T. Morgan, R. Lyle, M. E. Laughridge, M. J. Bamshad, K. B. Clarkson, R. Colby, F. Gurrieri, A. M. Limes, J. Roberson, C. Schrander-Stumpel, H. van Bokhoven, S. E. Antonarakis, and C. E. Schwartz. 2006. Frequency of genomic rearrangements involving the SHFM3 locus at chromosome 10q24 in syndromic and non-syndromic split-hand/foot malformation. Am J Med Genet Part A 140A:1375-1383.

Fiegler, H., P. Carr, E. J. Douglas, D. C. Burford, S. Hunt, C. E. Scott, J. Smith, D. Vetrie, P. Gorman, I. P. Tomlinson, and N. P. Carter. 2003. DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. Genes Chromosomes. Cancer 36:361-374.

Goodman, F. R., F. Majewski, A. L. Collins, and P. J. Scambler. 2002. A 117-kb microdeletion removing HOXD9-HOXD13 and EVX2 causes synpolydactyly. Am J Hum Genet 70:547-555.

Knijnenburg, J., K. Szuhai, J. Giltay, L. Molenaar, W. Sloos, M. Poot, H. J. Tanke, and C. Rosenberg. 2005. Insights from genomic microarrays into structural chromosome rearrangements. Am J Med Genet Part A 132A:36-40.

Monsuur, A. J., P. I. de Bakker, B. Z. Alizadeh, A. Zhernakova, M. R. Bevova, E. Strengman, L. Franke, R. van't Slot, M. J. van Belzen, I. C. Lavrijsen, B. Diosdado, M. J. Daly, C. J. Mulder, M. L. Mearin, J. W. Meijer, G. A. Meijer, E. van Oort, M. C. Wapenaar, B. P. Koeleman, and C. Wijmenga. 2005. Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. Nat. Genet. 37:1341-1344.

Nannya, Y., M. Sanada, K. Nakazaki, N. Hosoya, L. Wang, A. Hangaishi, M. Kurokawa, S. Chiba, D. K. Bailey, G. C. Kennedy, and S. Ogawa. 2005. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. 65:6071-6079.

Nishi, H., M. Senoo, K. H. Nishi, B. Murphy, T. Rikiyama, Y. Matsumura, S. Habu, and A. C. Johnson. 2001. p53 Homologue p63 represses epidermal growth factor receptor expression. J. Biol. Chem. 276:41717-41724.

O'Quinn, J. R., R. C. M. Hennekam, L. B. Jorde, and M. Bamshad. 1998. Syndromic ectrodactyly with severe limb, ectodermal, urogenital, and palatal defects maps to chromosome 19. Am J Hum Genet 62:130-135.

Omi, M., M. Fisher, N. J. Maihle, and C. N. Dealy. 2005. Studies on epidermal growth factor receptor signaling in vertebrate limb patterning. Dev. Dyn. 233:288-300.

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

Scherer, S. W., P. Poorkaj, H. Massa, S. Soder, T. Allen, M. Nunes, D. Geshuri, E. Wong, E. Belloni, S. Little, L. M. Zhou, D. Becker, J. Kere, J. Ignatius, N. Niikawa, Y. Fukushima, T. Hasegawa, J. Weissenbach, E. Boncinelli, B. Trask, L. C. Tsui, and J. P. Evans. 1994. Physical Mapping of the Split Hand Split Foot Locus on Chromosome-7 and Implication in Syndromic Ectrodactyly. Hum Mol Genet 3:1345-1354.

Schouten, J. P., C. J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, and G. Pals. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Research 30.

Seto, M. L., M. E. Nunes, C. A. Macarthur, and M. L. Cunningham. 1997. Pathogenesis of ectrodactyly in the Dactylaplasia mouse: Aberrant cell death of the apical ectodermal ridge. Teratology 56:262-270.

Sharland, M., M. A. Patton, and L. Hill. 1991. Ectrodactyly of Hands and Feet in A Child with A Complex Translocation Including 7Q21.2. Am J Med Genet 39:413-414.

Simcox, A. 1997. Differential requirement for EGF-like ligands in Drosophila wing development. Mech. Dev. 62:41-50.

Thienpont, B., J. Breckpot, J. R. Vermeesch, M. Gewillig, and K. Devriendt. 2008. A complex submicroscopic chromosomal imbalance in 19p13.11 with one microduplication and two microtriplications. Eur J Med Genet 51:219-225.

Thienpont, B., L. Mertens, T. Ravel, B. Eyskens, D. Boshoff, N. Maas, J. P. Fryns, M. Gewillig, J. R. Vermeesch, and K. Devriendt. 2007. Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. Eur Heart Journal 28:2778-2784.

van Bokhoven, H., B. C. J. Hamel, M. Bamshad, E. Sangiorgi, F. Gurrieri, P. H. G. Duijf, K. R. J. Vanmolkot, E. van Beusekom, S. E. C. van Beersum, J. Celli, G. F. M. Merkx, R. Tenconi, J. P. Fryns, A. Verloes, R. A. Newbury-Ecob, A. Raas-Rotschild, F. Majewski, F. A. Beemer, A. Janecke, D. Chitayat, G. Crisponi, H. Kayserili, J. R. W. Yates, G. Neri, and H. G. Brunner. 2001. p63 gene mutations in EEC syndrome, limb-mammary syndrome, and isolated split hand-split foot malformation suggest a genotype-phenotype correlation. Am J Hum Genet 69:481-492.

White, S. J., G. R. Vink, M. Kriek, W. Wuyts, J. Schouten, B. Bakker, M. H. Breuning, and J. T. den Dunnen. 2004. Two-color multiplex ligation-dependent probe amplification: Detecting genomic rearrangements in hereditary multiple exostoses. Hum Mutation 24:86-92.

Wong, W. T., Schumacher, C., Salcini, A. E., Romano, A., Castagnino, P., Pelicci, P. G., and Di Fiore, P. P. 1995. A Protein-Binding Domain, EH, Identified in the Receptor Tyrosine Kinase Substrate Eps15 and Conserved in Evolution. Proc.Natl.Acad.Sci.U.S.A 92(21):9530-4.


4

High-Resolution Melting Analysis (HRMA)—More Than Just Sequence Variant Screening

Rolf H.A.M. Vossen, Emmelien Aten, Anja Roos, and Johan T. den Dunnen

Human Mutation. 2009;30(6):860-6.

Abstract

Transition of the double-stranded DNA molecule to its two single strands, DNA denaturation or melting, has been used for many years to study DNA structure and composition. Recent technological advances have improved the potential of this technology, especially to detect variants in the DNA sequence. Sensitivity and specificity were increased significantly by the development of so-called saturating DNA dyes and by improvements in the instrumentation to measure the melting behavior (improved temperature precision combined with increased measurements per time unit and drop in temperature).

Melt analysis using these new instruments has been designated high-resolution melting curve analysis (HRM or HRMA). Based on its ease of use, simplicity, flexibility, low cost, nondestructive nature, superb sensitivity, and specificity, HRMA is quickly becoming the tool of choice to screen patients for pathogenic variants. Here we will briefly discuss the latest developments in HRMA and review in particular other applications that have thus far received less attention, including presequence screening, single nucleotide polymorphism (SNP) typing, methylation analysis, quantification (copy number variants and mosaicism), an alternative to gel-electrophoresis and clone characterization. Together, these diverse applications make HRMA a multipurpose technology and a standard tool that should be present in any laboratory studying nucleic acids.

Key words

mutation detection; methodology; melt curve analysis; HRM; HRMA

The Hardware

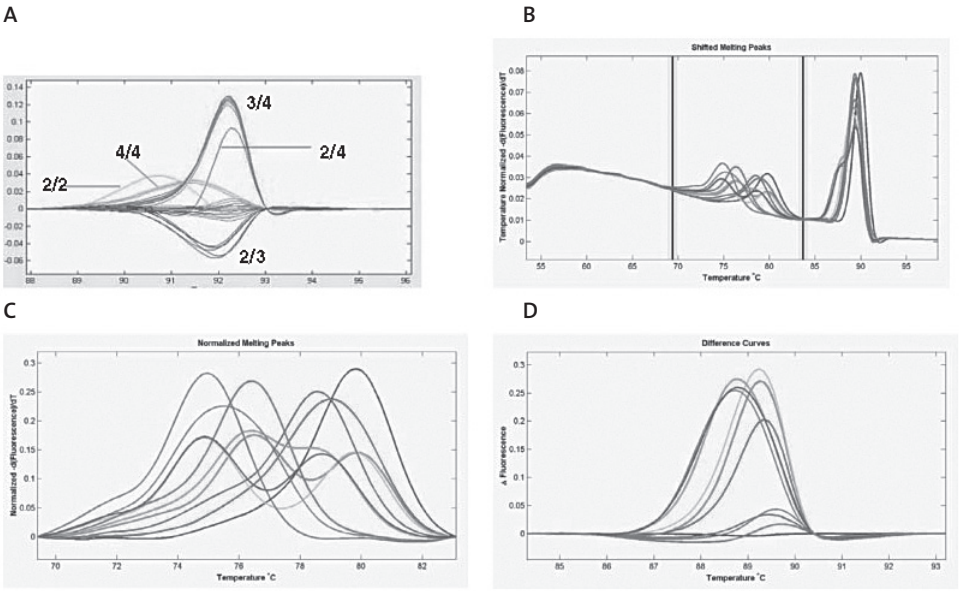
Several suppliers produce systems for high-resolution melting curve analysis (HRM or HRMA). Overall, these systems perform as can be expected but resolution may differ significantly [Herrmann et al., 2006, 2007]. The most sensitive system currently available is the HR-1 (Idaho Technology Inc., Salt Lake City, UT) generating fluorescence data from 55–95°C at a temperature transition rate of 0.1°C/sec and 200 data points/°C. Some assays, for example, multiplex SNP-typing [Seipp et al., 2008], critically depend on this resolution. The HR-1 uses capillaries and can analyze only one sample at a time; other capillary systems analyze up to 32 samples. Microtiter plate systems are more popular, facilitating the analysis of 96 or 384 samples simultaneously. Another distinction is whether the system is used for HRMA only or whether it is a combined (real-time) polymerase chain reaction (PCR) and HRMA instrument. As nicely demonstrated by Rouleau et al. [2009], combined qPCR and HRMA instruments allow detection of both quantitative (deletion/duplication) and qualitative (nucleotide) changes in one assay.

In most systems small well-to-well temperature differences exist that negatively influence sensitivity. To bypass this problem, Seipp et al. [2007] added control temperature calibration probes: one melting at low and one at high temperature. The analysis software can use the melt positions of these probes to compensate for temperature differences between wells, decreasing T_m SD by 38% [Seipp et al., 2007] notably increasing sensitivity. Note that the software available on HRMA instruments is an important element determining the ultimate sensitivity achieved [Herrmann et al., 2007], and not all packages yet facilitate the use of temperature calibration probes.

LC-Green was the first saturating dye available [Wittwer et al., 2003]; now there are many more, including LCGreens®+ (Idaho Technology Inc.), Syto9® (Invitrogen, Carlsbad, CA), EvaGreen® (Biotum) and LightCycler® 480 ResoLight Dye (Roche, Indianapolis, IN). Pricing of the dyes differs significantly; all have slightly different characteristics, and they often demand slightly different PCR buffers and conditions.

Implementation of HRMA is quite straightforward. In the most simple setting, requiring no modifications of existing conditions, dye is added post-PCR directly prior to melting. This simple approach allowed us, for example, to immediately discriminate all possible apoE alleles (Fig. 1A). ApoE typing with other techniques is often problematic, mostly requiring typing of each allele separately. Ultimately, performing HRMA in a closed-tube assay by adding the dye pre-PCR is more attractive. For existing assays this demands modification of PCR conditions, usually increasing Mg^{2+} concentration 2–3 µM and the annealing temperature by 1–5°C will be sufficient. When Mg^{2+} addition raises the T_m above the maximum instrument temperature, additives like DMSO (10%) or betaine (0.5 M) can

Figure 1: Sequence variant detection using HRMA. A: detection of all six possible ApoE-alleles combinations; LC-Green+ was added post- PCR, before HRMA (the PCR contained 10% DMSO). The ApoE 3/3 allele was set as standard (horizontal gray line). B–D: SNP typing using an unlabeled amino-blocked melt probe covering three independent variants in the first exon of the MBL2 gene (Roos et al., in preparation). B: overall derivative plot; C: enlargement of the low (melt probe) and D: high (PCR fragment) Tm peaks. Using the combined melt profiles all 10 possible alleles can be clearly discriminated.



be added to lower the T_m. For difficult cases, a gradient PCR-cycler can be used to quickly determine the most optimal annealing temperature.

HRMA demands no big changes in the laboratory and does not require specific skills; it is a simple PCR performed under slightly modified conditions and in the presence of a specific dye. The most expensive element is acquisition of the instrument. However, with € 10; 000; 50; 000, depending on the system, even this compares favorably to some other technologies. Because HRMA, unlike many other techniques (e.g., SSCA, dHPLC, DGGE, or capillary electrophoresis), does not require post-PCR separation, significant cost savings are achieved. Furthermore, HRMA is a nondestructive method, and subsequent analysis by, for example, gel-electrophoresis or sequencing, can still be performed after melt analysis.

Applications

Mutation Detection

HRMA has been developed for the detection of DNA sequence variants and it was applied first for genotyping [Wittwer et al., 2003]. Simplicity, low cost, ease of use, and a high sensitivity/ specificity have been the most prominent features, making HRMA an attractive new tool for genotyping and application in diagnostic labs. In Supp. Table S1 we give a comprehensive list of all (human) gene-based assays we could find. HRMA systems come with powerful software tools facilitating automated scoring that is very robust, although a quick manual check remains advisable. Using HRMA for mutation detection has been reviewed recently [Erali et al., 2008]; therefore, we will make general comments only. Furthermore, this issue of Human Mutation contains several papers describing application of HRMA for sequence variant detection, elegantly demonstrating its current state of the art (see also Supp. Table S1).

Tindall et al. [2009] compared two instruments and fluorescent dyes in particular regarding the detection of combinations of DNA variants present in GC-rich fragments. They demonstrated the current limitations of HRMA and called for caution when using it as the sole method to make a clinical diagnosis. Nguyen-Dumont et al. [2009] showed the power of including a melt probe to improve identification of rare variants in combination with a known SNP as well as to dramatically reduce sequencing effort. Van Der Stoep et al. [2009] followed a similar approach designing a sequence variant screen covering the BRCA1 gene, including melt probes against known frequent SNPs. In the setting of the EuroGenTest consortium, the authors went through the effort to perform an elaborate interlaboratory evaluation and validation of HRMA and generated guidelines for setting up and implementing it as a scanning technique for new genes. In a blind study on 28 patient samples the protocol resulted in a 100% detection sensitivity at a specificity of 98%, indicating a low incidence of false positives. Rouleau et al. [2009] used the possibility provided by some instruments to perform quantitative PCR and HRMA in one instrument to scan for both quantitative (deletions/duplications) and qualitative nucleotide changes in one assay. Finally, Dobrowolski et al. [2009] described the use of HRMA to scan the entire 16.6 kb human mitochondrial genome (mtDNA) for sequence variants in less then 2 hr. Identification of mtDNA variants is complicated, as many are heteroplasmic, with the variant allele present at highly variable percentages. The fact that the authors successfully identified variants present at levels ranging from 1–100% heteroplasmy nicely shows the sensitivity of the assay as well as the power of HRMA to detect quantitative changes (see Quantification).

Although HRMA of fragments up to 600 bp and more has been reported, our experience

is that the technology is more sensitive for smaller fragments. For fragment screening, fragments of 150–250 bp are used, that is, in general, one fragment per exon. When assays are designed to type specific variants (SNP typing) we target fragment sizies of 80–100 bp. We find it critical for high sensitivity that the melt profile contains not more then one to two melt domains. When fragments contain more melt domains chances increase that not all variants are detected. Today the design of new assays is simplified by the availability of powerful design programs, often delivered together with the instrument.

In early HRMA experiments we have seen that a second and sometimes even a third melt of the PCR products may improve results. Similarly, especially when the concentration of the DNA fragments to analyze differ considerably, results can be improved by adding 1 µl high salt buffer (1.0M KCL, 0.5M Tris-HCl [pH58.0]) followed by a new melt (Fig. 2). Salt addition may increase resolution and improve clustering, but success of the method is unpredictable and depending on the samples, fragments, and variants analyzed

When a few simple rules are taken into account (i.e., avoid long fragments and multiple melt domains), designing a HRMA screen for a gene is rather straightforward. We have successfully designed assays to screen a range of genes of which the DMD gene with 90 fragments covering 79 exons was the largest (Al-Momani et al., in press). As Nguyen-Dumont et al. [2009] and Van Der Stoep et al. [2009] show, it is time and cost saving to include unlabeled melt probes to positively identify known nonpathogenic variants. This prevents unnecessary sequencing of such fragments, while it at the same time safeguards against overlooking other variants in the same fragment (see Presequence Screening).

It should be noted that the shape of an HRMA curve in itself is usually not sufficient to type a specific variant [Tindall et al., 2009]; to achieve this either a melt probe should be added or the fragment should be sequenced. The power of adding an unlabeled melt probe to discriminate specific variants or multiple alleles is astonishing. An example is shown in Figure 1B–D (Roos et al., in preparation). Using a melt probe containing the wild-type sequence we were able to discriminate all 10 possible alleles deriving from a series of three closely spaced variants in the MBL2 gene.

The cheapest block available to prevent extension during PCR of the melt probe is a 30 phosphate. Unfortunately, this block is unstable, and after time undesired additional melt peaks may emerge. Other blocks are more stable but also more expensive. Zhou et al. [2008] elegantly solved this issue by using a so-called snapback primer, that is, a 50 tailed primer including a loop region and a sequence complementary to its extension product, covering the variant to scan. A potentially weak point of HRMA is the detection of homozygous variants. Although recent developments have further improved resolution [Gundry et al., 2008], the difference for some variants (e.g., A–T to T–A changes) are so subtle that they can easily be missed. Especially when samples from different sources have

to be analyzed, sample-to-sample variation and thus experimental noise increases and subtle changes might go undetected. Therefore, in sequence variant scanning applications (clinical diagnostics) we consistently use sample mixing to generate hetero-duplexes. First, samples are melted to obtain a standard melt profile. Next, using the scheme shown (Fig. 3, designed for DMD, an X-linked disease), samples are mixed and then a second melt curve is generated. The simple mixing scheme results in two heteroduplexes for each sample (note that sample mixing can be easily automated). Homozygotes will result in two wells with heteroduplexes, greatly improving their detection. Heterozygotes will be detected from the standard premixing melt curves, although they will usually be obvious from the mixed samples as well (1:3 ratio, see Quantification). The scheme shown will only fail when samples with identical variants are mixed but this can be simply prevented by adding more controls or by mixing samples from unrelated individuals only. It should be noted that the scheme presented (Fig. 3) has the potential to discriminate hemizygous from true homozygous alleles as well.

It should be noted that the sensitivity of HRMA to detect heterozygotes is much better than that of DNA sequencing (see Quantification). Consequently, when HRMA indicates the presence of a variant that cannot be confirmed using sequencing it might well be that this variant is present in a relatively low fraction of the sample (somatic mosaicism). Other techniques, for example, cloning1sequencing or single molecule dilution1PCR and HRMA, might be required to confirm these variants.

Pressequence Screening

Using HRMA for presequence screening, in particular for larger genes, may yield significant cost savings; Provaznikova et al. [2008] reported avoiding unnecessary sequencing of more than 85% for the MYH9 gene. In the example of the DMD gene, where 79 exons need to be analyzed (Al-Momani et al., in press), assuming bidirectional sequencing costs h10 per exon, screening a patient amounts to ~ €800. PCR per exon (€1) and HRMA (€0:10 dye) would cost ~ €90 per patient. Assuming 5 exons show a melt shift requiring verification by sequence analysis (including one pathogenicvariant, three nonpathogenic variants and a false positive) would add €50, making a total of €140. A cost saving of more then €650 per patient that can be further reduced when melt probes are included to confirm the presence of the most frequent nonpathogenic variants. In fact, based on these figures, HRMA prescreening of a five-exon gene is already cost-effective. In addition, to detect somatic changes or heteroplasmy [Dobrowolski et al., 2009], HRMA seems more sensitive then sequencing (see Quantification).

Figure 2: Effect of high-salt. A: HRMA analysis of a series of samples. B: The same samples as in A analyzed after addition of 1 ml 1.0MKCl/ 0.5M Tris-HCl (pH58.0) and remelting. Note the improved separation and sharpening of the three groups.

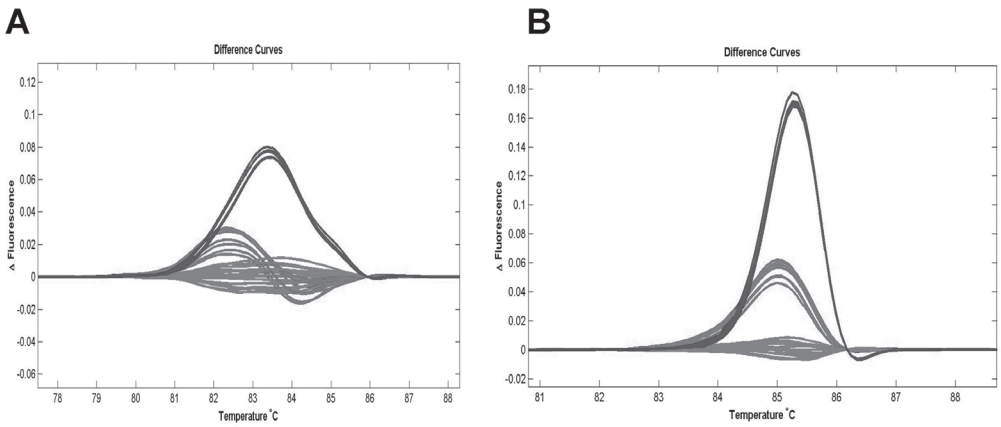
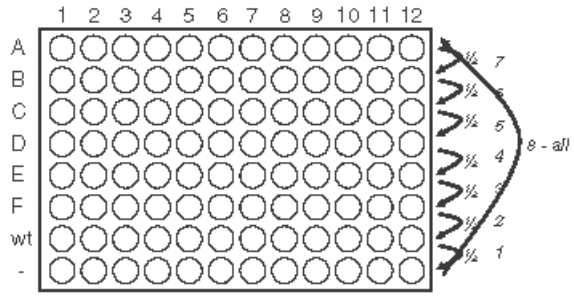


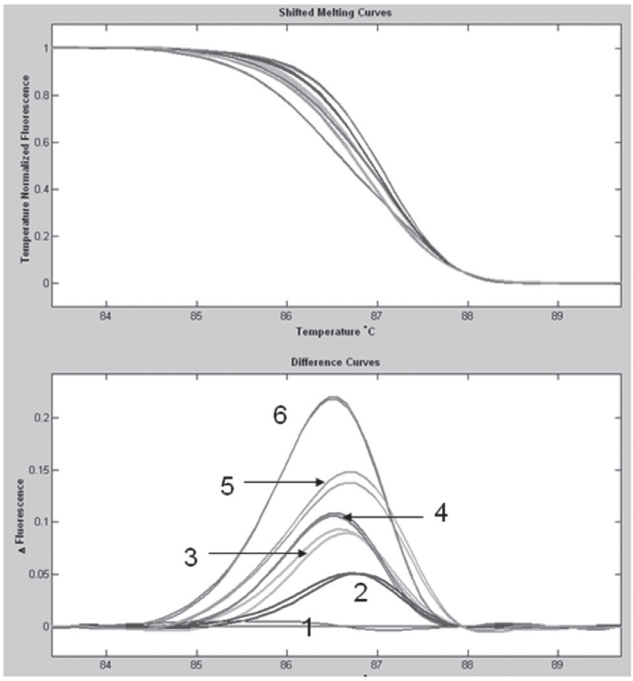
Figure 3: Generating heteroduplexes. To ensure the detection of all homozygous variants the scheme shown can be used. Wells 1–12 contain PCR products from different fragments (exons) amplified for each sample (patients A–F and wt5wild-type control). Row 8 is empty. After a first HRMA mixing starts with taking half of the sample from row 7 (wt, 5–10 ml of the PCR product) and transferring this to row 8. Subsequently, half the volume of from row 6 is transferred to row 7, half of row 5, to row 6, etc. Finally, the content of row 8 is transferred to row 1 and the fragments are melted again.



SNP typing

HRMA for SNP-typing can be very attractive. Assay design is cheap, simple, and fast. When in a specific region one or two PCRs covering a SNP are designed, at least one and usually both will give a good assay. To increase sensitivity, the fragment should preferably

Figure 4: CA repeat analysis using HRMA. Six different DNA samples were analyzed in duplicate, all heterozygotes. Top panel: normalized temperature shifted melting curves. Bottom panel: derivative plot. Allele lengths were 1=18/21, 2=17/21, 3=14/17, 4=14/18, 5=15/21, 6=14/21.



be small (80–100 bp) and when a choice is available (e.g., in a haplotype block) one should select a G to A variant (predicted to give the largest melt shift) [Reed andWittwer, 2004]. Addition of an unlabeled melt probe [Nguyen-Dumont et al., 2009; Zhou et al., 2008] is recommended, giving a double check in the assay and improved scoring for homozygote variants. When ordered primers arrive, it should not take more then one or two PCRs to test amplification conditions and the assay is ready. Assay costs are then just PCR, and cost for assay design (an often largely underestimated factor) are negligible. Unless thousands of samples need to be typed, dye cost should compare favorably to assays including specific labeled probes (e.g., TaqMan). HRMA sample throughput can be increased using robotic plate loading, available as an extension on some systems or by installing an additional loader.

Based on the positive results described above one wonders whether variable number

tandum repeats (VNTRs), especially CA repeats, could be typed using HRMA (Fig. 4). Although in the example shown all curves can be clearly discriminated, we found that all possible allele combinations are so many that heterozygous CA repeat typing cannot be performed with certainty. It is possible though in small families to detect differences between family members and to use HRMA to study loss of heterozygosity (LOH) in heterozygous samples. As indicated by the experiments of Intemann et al. [2009], it can be speculated that addition of a melt probe spanning the smallest and/or largest allele might increase resolution further.

A weak point of HRMA is that it cannot easily be applied in multiplex mode, that is, to type variants in several different fragments at the same time. Theoretically, multiplexing can be achieved by exploiting color differences, temperature differences, or both. Seipp et al. [2008] used the simplest approach, that is,

T_m-differences to successfully design a quadruplex genotyping assay. However, such design is time consuming and its success critically depends on high sample DNA quality and the sensitivity of the HRMA system used.

Methylation

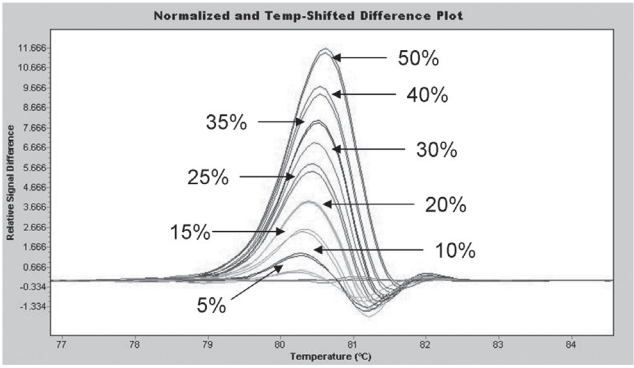
Lately, genomic studies frequently involve the analysis of epigenetic marks, especially methylation at CpG dinucleotides in relation to gene expression, imprinting, and cancer (Ehrich et al., 2006). A critical step in these procedures is bisufite treatment of the genomic DNA, changing unmethylated C nucleotides (but not methylated Cs) to Us. HRMA can be used in such studies in two stages. First, success of the bisulfite treatment can be checked by PCR of a control genomic segment devoid of methylated CpGs and comparison of its melt profile with that of a 100% converted fragment (cloned) and an untreated sample. The closer the melt profile resembles that of the 100% converted fragment, the better the bisulfite treatment worked [Worm et al., 2001]. Second, HRMA can be used to determine the percentage of C to U conversion, either directly, in combination with a melt probe [Maat et al., 2007], or after cloning (see Clone Characterization).

Quantification—Mosaicism and Copy Number Variant (CNV) Confirmation

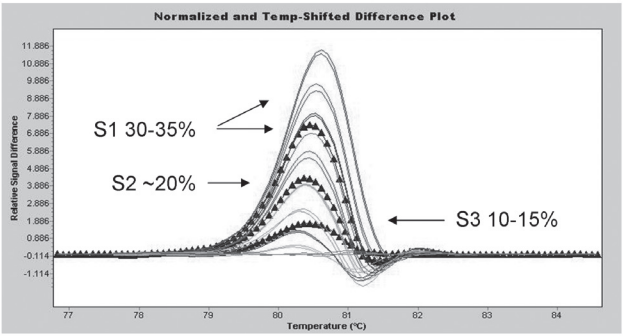
Depending on the melt shift obtained, HRMA can also be used for quantitative analysis; the larger the shift, the smaller the quantitative differences that can be detected. To quantify the fraction of variant molecules frequently used technologies are dideoxy sequencing (resolution limit down to 20–30%) and pyro-sequencing (down to 1–10%). Although powerful, pyro-sequencing is laborintensive, costly, and demands specific equipment. We have successfully applied HRMA for the quantification of different alleles [Aten et al., 2009; Bruder et al., 2008], and shown that detection in steps of 12.5% (1 in 8)

Figure 5: Quantification using HRMA. **A**: Dilution series for a pathogenic A4G variant in APC exon 8. **B**: Superposition of the HRMA profiles from three mosaic samples (S1–S3) and the estimation of the level of somatic mosaicism. Pyro-sequencing of the samples gave estimates of 30, 19, and 7%, respectively.

A



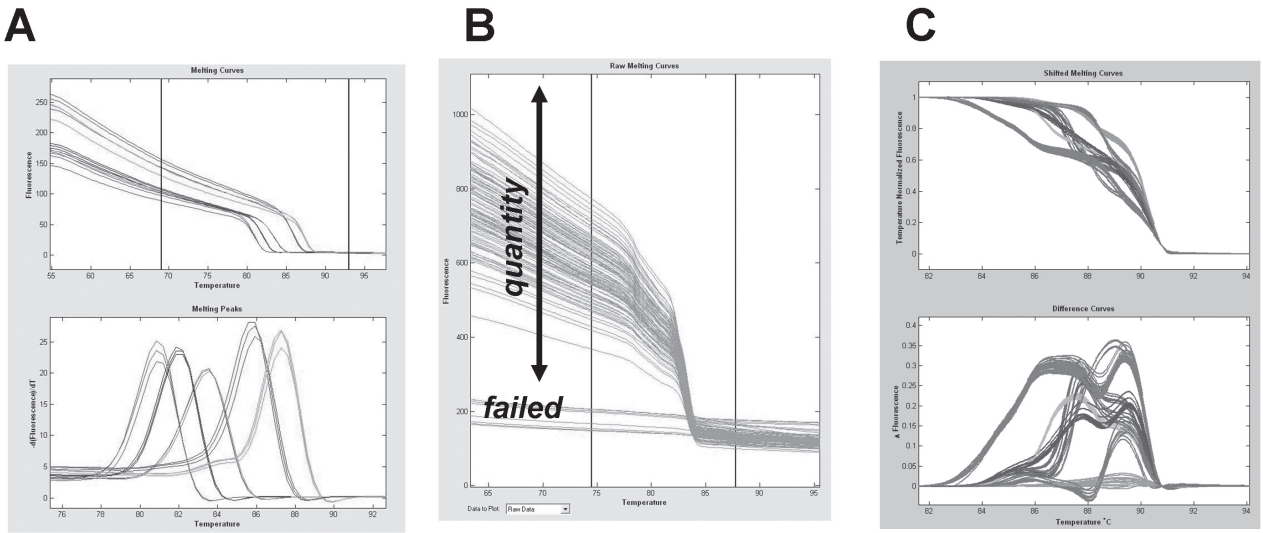
B



is usually possible [Aten et al., 2009]. Application for the detection of the level of somatic mosaicism in colon cancer is shown in Figure 5. Using a dilution series as a ruler we could readily determine the level of somatic mosaicism in three patients suspected to carry a specific pathogenic variant down to a level of 5–10%. The data obtained matched perfectly with those obtained using pyro-sequencing [Hes et al., 2008], yet the assay was less expensive and simpler to perform. It should be noted that when melt probes are included resolution can be improved to below 5%. Other qHRMA applications are the determination of differences in allelic expression, based on the presence of a variant in the mRNA and the detection of heteroplasmy in mtDNA [Dobrowolski et al., 2009].

Recently, we used the same approach to estimate the fraction of cells carrying a somatic deletion identified in one individual from an identical twin pair [Bruder et al.,

Figure 6: HRMA as alternative for gel electrophoresis. A: Analysis of a series of five different PCR fragment; because the fragments have clearly different melt profiles all five fragments can be analyzed in one analysis. When the melt profiles partly overlap, analysis can be done per fragment. B: Analysis of a PCR performed on 96 different samples. Some PCRs failed (only background) fluorescence, yield of the others can be estimated from the level of fluorescence. Purity, including absence of primer dimers, can be checked by analysis of the HRMA difference plots (not shown). C: HRMA after insert PCR of 384 phage display clones after second round selection. Several clear groups of melt profiles are identified, an indication that the clones contain identical inserts. Note that to identify all groups present, clones recognized after a first analysis need to be removed and software grouping must be repeated. This procedure has to be repeated until no further groups are recognized.



2008]. The approach can also be used to confirm CNVs (both deletions and duplications) detected using whole genome SNP arrays while screening patients of diverse diseases for genomic rearrangements. Depending on the setting, techniques like FISH, MAPH [Armour et al., 2000], MLPA
[Schouten et al., 2002], MAQ [Suls et al., 2006], and qPCR can be applied to confirm the array findings. However these techniques are either costly or demand considerable time to develop. For confirmation with HRMA one can use any SNP from the suspected region. Samples homozygous for the two opposite alleles (AA and BB) of these SNPs are used to generate a reference ruler as well as a 1:1 mix with the sample potentially carrying the rearrangement. Assume the test sample carries the A allele and is either Ao or AA. When mixed 1:1 with a homozygous BB sample the deletion is confirmed when the melt profile comigrates with the 1:2 AA:BB sample mix (AoBB). The deletion is absent when the melt profile comigrates with the 1:1 AA:BB sample mix (AABB). A duplication will be confirmed when the 1:1 mix with a homozygous BB sample comigrates with the 3:2 AA:BB sample mix (AAABB).

Alternative for Gel Electrophoresis

The current standard to check the result of a PCR or digestion is analysis of the product using agarose gel electrophoresis and ethidium–bromide staining. Identity of the fragment is characterized by its length, purity by the absence of other fragments, and yield by the strength of the fluorescence of the band. HRMA is an attractive alternative; identity characterized by the melting profile, purity by the absence of distortions from the control melt curve (and absence of additional melt peaks), and yield by the amount of fluorescence signal (Fig. 6). The advantages of using HRMA are clear; one does not have to pour gels and use hazardous chemicals (ethidium–bromide), melting is faster then electrophoresis, and data analysis can be performed automatically. Furthermore, because it is a nondesctructive method, when HRMA would not give clear results fragments can still be analyzed on gel. In our laboratory HRMA is quickly replacing gel electrophoresis for the characterization of PCR products. Dye is added post- PCR (1 µl LC-Green1[10 stock] per 10 µl sample), sample is 5 min incubated at 95°C, cooled down to room temperature, and melted. Figure 5A shows an example where five different fragments were analyzed, all clearly discernable by their individual melt behavior. Primer–dimer formation would be recognized by the presence of a melt peak at low T_m. Out et al. (manuscript submitted) used HRMA instead of gel electrophoresis to check amplification as well as to determine long-range PCR yield guiding equimolar pooling before sequencing of the MUTYH gene. The authors could successfully show detection of nearly all variants in the expected frequencies down to 0.5% (1/200 chromosomes).

Clone Characterization

Several studies generate large series of clones that need to be sequenced to determine their identity. These include in vitro mutagenesis experiments, methylation studies, cDNA cloning to determine levels of differential splicing and/or allelic expression, and phage display selections. HRMA provides an attractive tool to prescreen the clones to detect those that share the same insert and those that differ, generating considerable savings for subsequent sequencing. An example is shown in Figure 6C, showing the result of a second round phage display selection. The experiment resulted in several groups of clones with identical melt curves, indicating that the experiment was successful in positively selecting several different phage display clones. Subsequent sequencing of representative clones per group confirmed the HRMA results; sequence differences between groups and sequence identity within groups [Pepers et al., 2009]. Previously, clone inserts were fingerprinted using restriction digestion and gel electrophoresis, a less sensitive and much more laborious method [Verheesen et al., 2006].

Conclusion

The advantageous characteristics of HRMA make it a technology that quickly attracts a range of new users. Its ease of use, simplicity, flexibility, low cost, nondestructive nature, superb sensitivity, and specificity, make HRMA the method of choice to screen patients for pathogenic variants. As reviewed above, HRMA has several attractive additional applications, making it a versatile multipurpose analytical tool to analyze nucleic acids in general. Because HRMA is still a rather young technology, one can only expect exciting further developments. The company Fluidigm markets a nanoliter qPCR system [Spurgeon et al., 2008], today facilitating the analysis of 96 samples 96 PCR assays (i.e., 9,216 assays simultaneously); imagine the power of such a system when it would facilitate HRMA.

Acknowledgements

We gratefully acknowledge our colleagues from Human & Clinical Genetics (LUMC, Leiden) who allowed us to discuss their unpublished results using HRMA, in particular, Astrid Out, Rowida Al Momani, Willeke van Roon- Mom, Elsa Bik, and Nienke van der Stoep.

Al Momani R, Van Der Stoep N, Bakker E, Den Dunnen JT, Breuning MH, Ginjaar HB. 2009. Rapid and cost effective detection of small mutations in the DMD gene by high resolution melting curve analysis. Neuromuscul Disord (in press).

Armour JA, Sismani C, Patsalis PC, Cross G. 2000. Measurement of locus copy number by hybridisation with amplifiable probes. Nucleic Acids Res 28:605–609.

Aten E, White SJ, Kalf ME, Vossen RHAM, Thygesen HH, Kriek M, Breuning MH, Den Dunnen JT. 2008. Methods to detect CNVs in the human genome. Cytogenet Genome Res 123:313–321.

Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, de Stahl TD, Menzel U, Sandgren J, von Tell D, Poplawski A, Crowley M, Crasto C, Partridge EC, Tiwari H, Allison DB, Komorowski J, Van Ommen GJB, Boomsma DI, Pedersen NL, Den Dunnen JT, Wirdefeldt K, Dumanski JP. 2008. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. Am J Hum Genet 82:763–771.

Dobrowolski S, Gray J, Miller T, Sears M. 2009. Identifying sequence variants in the human mitochondrial genome using high-resolution melt (HRM) profiling. Hum Mutat 30:891–898.

Ehrich M, Field JK, Liloglou T, Xinarianos G, Oeth P, Nelson MR, Cantor CR, van den Boom D. 2006. Cytosine methylation profiles as a molecular marker in non-small cell lung cancer. Cancer Res 66:10911–10918.

Erali M, Palais R, Wittwer C. 2008. SNP genotyping by unlabeled probe melting analysis. Methods Mol Biol 429:199–206.

Gundry CN, Dobrowolski SF, Martin YR, Robbins TC, Nay LM, Boyd N, Coyne T, Wall MD, Wittwer CT, Teng DH. 2008. Base-pair neutral homozygotes can be discriminated by calibrated high-resolution melting of small amplicons. Nucleic Acids Res 36:3401–3408.

Herrmann MG, Durtschi JD, Bromley LK, Wittwer CT, Voelkerding KV. 2006. Amplicon DNA melting analysis for mutation scanning and genotyping: crossplatform comparison of instruments and dyes. Clin Chem 52:494–503.

Herrmann MG, Durtschi JD, Wittwer CT, Voelkerding KV. 2007. Expanded instrument comparison of amplicon DNA melting analysis for mutation scanning and genotyping. Clin Chem 53:1544–1548.

Hes FJ, Nielsen M, Bik EC, Konvalinka D, Wijnen JT, Bakker E, Vasen HF, Breuning MH, Tops CM. 2008. Somatic APC mosaicism: an underestimated cause of polyposis coli. Gut 57:71–76.

Intemann CD, Thye T, Sievertsen J, Owusu-Dabo E, Horstmann RD, Meyer CG. 2009. Genotyping of IRGM tetranucleotide promoter oligorepeats by fluorescence resonance energy transfer. Biotechniques 46:58–60.

Maat W, van der Velden PA, Out-Luiting C, Plug M, Dirks-Mulder A, Jager MJ, Gruis NA. 2007. Epigenetic inactivation of RASSF1a in uveal melanoma. Invest Ophthalmol Vis Sci 48:486–490.

Nguyen-Dumont T, Le Calvez-Kelm F, Forey N, McKay-Chopin S, Garritano S, Gioia-Patricola L, De Silva D, Weigel R, Breast Cancer Family Registries (BCFR), Kathleen Cuningham Foundation Consortium for research into Familial Breast cancer (kConFab), Sangrajrang S, Lesueur F, Tavtigian SV. 2009. Description and validation of high-throughput simultaneous genotyping and mutation scanning by high-resolution melting curve analysis. Hum Mutat 30:884–890.

Pepers BA, Schut MH, Vossen RHAM, Van Ommen GJB, Den Dunnen JT, Van Roon- Mom WMC. 2009. Cost-effective HRMA pre-sequence typing of clone libraries: application to phage display selection. BMC Biotechnology, in press.

Provaznikova D, Kumstyrova T, Kotlin R, Salaj P, Matoska V, Hrachovinova I, Rittich S. 2008. High-resolution melting analysis for detection of MYH9 mutations. Platelets 19:471–475.

Reed GH, Wittwer CT. 2004. Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin Chem 50:1748–1754.

Rouleau E, Lefol C, Bourdon V, Coulet F, Noguchi T, Soubrier F, Bieche I, Olschwang S, Sobol H, Lidereau R. 2009. Quantitative PCR high-resolution melting (qPCR–HRM) curve analysis, a new approach to simultaneously screen point mutations and large rearrangements: application to MLH1 germline mutations in Lynch Syndrome. Hum Mutat 30:867–875.

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligationdependent probe amplification. Nucleic Acids Res 30:e57.

Seipp MT, Durtschi JD, Liew MA, Williams J, Damjanovich K, Pont-Kingdon G, Lyon E, Voelkerding KV, Wittwer CT. 2007. Unlabeled oligonucleotides as internal temperature controls for genotyping by amplicon melting. J Mol Diagn 9:284–289.

Seipp MT, Pattison D, Durtschi JD, Jama M, Voelkerding KV, Wittwer CT. 2008. Quadruplex genotyping of F5, F2, and MTHFR variants in a single closed tube by high-resolution amplicon melting. Clin Chem 54:108–115.

Spurgeon SL, Jones RC, Ramakrishnan R. 2008. High throughput gene expression measurement with real time PCR in a microfluidic dynamic array. PLoS ONE 3:e1662.

Suls A, Claeys KG, Goossens D, Harding B, Van Luijk R, Scheers S, Deprez L, Audenaert D, Van Dyck T, Beeckmans S, Smouts I, Ceulemans B, Lagae L, Buyse G, Barisic N, Misson JP, Wauters J, Del Favero J, De Jonghe P, Claes LR. 2006. Microdeletions involving the SCN1A gene may be common in SCN1A-mutation-negative SMEI patients. Hum Mutat 27:914–920.

Tindall EA, Petersen DC, Woodbridge P, Schipany K, Hayes VM. 2009. Assessing high-resolution melt curve analysis for accurate detection of gene variants in complex DNA fragments. Hum Mutat 30:876–883.

van der Stoep N, van Paridon CDM, Janssens T, Krenkova P, Stambergova A, Macek M, Matthijs G, Bakker E. 2009. Diagnostic guidelines for highresolution melting curve (HRM) analysis: an interlaboratory validation of BRCA1 mutation scanning using the 96-well LightScanner™. Hum Mutat 30:889–909.

Verheesen P, Roussis A, de Haard HJ, Groot AJ, Stam JC, Den Dunnen JT, Frants RR, Verkleij AJ, Verrips CT, Van Der Maarel SM. 2006. Reliable and controllable antibody fragment selections from Camelid non-immune libraries for target validation. Biochim Biophys Acta 1764:1307–1319.

Worm J, Aggerholm A, Guldberg P. 2001. In-tube DNA methylation profiling by fluorescence melting curve analysis. Clin Chem 47:1183–1189.

Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ. 2003. High-resolution genotyping by amplicon melting analysis using LCGreen. Clin Chem
49:853–860.

Zhou L, Errigo RJ, Lu H, Poritz MA, Seipp MT, Wittwer CT. 2008. Snapback primer genotyping with saturating DNA dye and melting analysis. Clin Chem 54:1648–1656..

Supporting Information

Supp. Table S1: Published assays using high-resolution melting analysis (HRMA)

Gene	Assay	Reference
ABL1	variants	Polakova KM et al. 2008 Leuk Res 32:1236-1243
ACADM	gene	McKinney JT et al. 2004 Mol Genet Metab 82:112-120
ACVRL1	gene	Vandersteen JG et al. 2007, Clin Chem 53:1191-1198
AKT1	variants	Do H et al. 2008 BMC Res Notes 1:14
APOB	variants	Liyanage KE et al. 2008 Ann Clin Biochem 45:170-176
APOC2	gene	Chang YT et al. 2008 J Clin Gastroenterol 2008 Epub Nov 20
APOE	variants	Poulson MD & Wittwer CT 2007 Biotechniques 43:87-91
ARG1	gene	Mitchell S et al. 2009 Hum Mutat 30:56-60
ASL	gene	Mitchell S et al. 2009 Hum Mutat 30:56-60
ASS1	gene	Mitchell S et al. 2009 Hum Mutat 30:56-60
ATP7B	variants	Zhao X et a. 2008 Zhonghua Yi Xue Yi Chuan Xue Za Zhi 25:515-519
BRAF	variants	Willmore-Payne C et al. 2005 Hum Pathol 36:486-493
BRCA1	gene	Van Der Stoep et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
BRCA1/2	variants	Takano EA et al. 2008 BMC Cancer 8:59
	gene	De Leeneer K et al. 2008 Clin Chem 54:982-989
	gene	de Juan I et al. 2008 Breast Cancer Res Treat. Jun 5 - E-pub ahead
CFTR	gene	Montgomery J et al. 2007 Clin Chem 53:1891-1898
CPS1	gene	Mitchell S et al. 2009 Hum Mutat 30:56-60
CXCR7	variant	Tindall EA et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
CYP2C9	variants	Hill CE et al. 2006 Am J Clin Pathol 125:584-591
EGFR	variants	Do H et al. 2008 BMC Cancer 8:142
	variants	Smith GD et al. 2008 J Clin Pathol. 61:4874-93
ENG	gene	Vandersteen JG et al. 2007, Clin Chem 53:1191-1198
ERBB2	variants	Willmore-Payne C et al. 2006 Mod Pathol 19:634-640
EXT1/2	gene	Lonie L et al. 2006 Hum Mutat 27:1160
F2	variants	Seipp MT et al. 2008 Clin Chem 54:108-115
F5	variants	Seipp MT et al. 2008 Clin Chem 54:108-115
F8	gene	Lin SY et al. 2008 BMC Med Genet 9:53
FGFR3	variants	Hung CC et al. 2008 Clin Biochem 41:162-166
FLT3	variants	Tan AY et al. 2008 J Hematol Oncol 1:10
GALT	variants	Dobrowolski SF et al. 2003 J Mol Diagn 5:42-47

GJB1	gene	Kennerson ML et al. 2007 Clin Chem 53:349-352
HBB	variants	Herrmann MG et al. 2000 Clin Chem 46:425-428
HRAS	variant	Tindall EA et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
IGF1	gene	Palles C et al. 2008 Hum Mol Genet 17:1457-1464
IL10	variants	Tedde A et al. 2008 Scand J Gastroenterol 43:712-718
JAK2	variants	Rapado I et al. 2009 J Mol Diagn 11:155-161
KIT	variants	Holden JA et al. 2007 Am J Clin Pathol 128:230-238
KRAS	variants	Do H et al. 2008 BMC Cancer 8:142
	variants	Krypuy M et al. 2006 BMC Cancer 6:295
LDLR	gene	Laurie AD et al. 2008 Clin Biochem Epub Dec 11
		Vossen et al., Human Mutation 2
LPL	gene	Chang YT et al. 2008 J Clin Gastroenterol 2008 Epub Nov 20
LRRK2	variants	Tedde A et al. 2007 Cell Mol Neurobiol 27:877-881
MBL2	variants	Vossen RHAM et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
	variant	Tindall EA et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
MLH1	gene	Rouleau E et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
MTHFR	variants	Sinthuwiwat T et al. 2008 Mol Cell Probes 22:329-332
MYH	gene	rovaznikova D et al. 2008 Platelets 19:471-475
NAGS	gene	Mitchell S et al. 2009 Hum Mutat 30:56-60
NF2	gene	Sestini R et al. 2008 Genet Test 12:311-318
NKX3-1	variant	Tindall EA et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
NPM1	variants	Tan AY et al. 2008 J Hematol Oncol 1:10
NPY	variant	Tindall EA et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
OTC	gene	Dobrowolski SF et al. 2007 Hum Mutat 28:1133-1140
PAH	gene	Dobrowolski SF et al. 2007 Mol Genet Metab 91:218-227
PDGFRA	variants	Holden JA et al. 2007 Am J Clin Pathol 128:230-238
PHEX	gene	Gaucher C et al. 2009 Hum Genet Feb 15
PIK3CA	variants	Simi L et al. 2008 Am J Clin Pathol 130:247-253
RASEF	gene	Maat W et al. 2008 Invest Ophthalmol Vis Sci 49:1291-1298
RET	(gene)	Margraf RL et al. 2006 Clin Chem 52:138-141
RUNX2	variants	Pal T et al. 2007 Clin Genet 71:589-591
RYR1	variants	Grievink H & Stowell KM 2008 Anal Biochem 374:396-404
SCN5A	gene	Millat G et al. 2008 Clin Biochem Epub Nov 6
SERPINA1	variant	Tindall EA et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
SLC22A5	gene	Dobrowolski SF et al. 2005 Hum Mutat 25:306-313
SLC25A13	gene	Mitchell S et al. 2009 Hum Mutat 30:56-60

SLC25A15	gene	Mitchell S et al. 2009 Hum Mutat 30:56-60
SNRPN	variants	White HE et al. 2007 Clin Chem 53:1960-1962
TGM1	gene	Herman ML et al. 2008 Hum Mutat E-pub Dec 18
TP53	gene	Bastien R et al. 2008 Hum Mutat 29:757-764
	gene	Garritano S et al. 2009 BMC Genet 10:5
VDR	variant	Tindall EA et al. 2009 Hum Mutat 30 (this issue, e-pub pending)
mtDNA	genome	Dobrowolski SF et al. 2009 Hum Mutat 30 (this issue, e-pub pending)

Legend

Indicated are the gene (official HGNC gene symbol), whether the assay was designed to scan the entire **gene** or only specific variants, and the reference(s). When papers to both screen-specific **variants** and complete gene scans were published, we list only the latter. A regularly updated table can be found at http://www.LGTC.nl/HRMA including links to the primer and melt probe sequences used (when known).


5

Keratosis Follicularis Spinulosa Decalvans is caused by mutations in *MBTPS2*

Emmelien Aten, Lisa Brasz, Dorothea Bornholdt, Ingeborg B. Hooijkaas, Mary E. Porteous, Virginia P. Sybert, Maarten H. Vermeer, Rolf H.A.M. Vossen, Michiel J.R. van der Wielen, Egbert Bakker, Martijn H.Breuning, Karl-Heinz Grzeschik, Jan C. Oosterwijk, Johan T. den Dunnen

Human Mutation. 2010;31(10):1125-33

Abstract

Keratosis Follicularis Spinulosa Decalvans, (KFSD) is a rare genetic disorder characterized by development of hyperkeratotic follicular papules on the scalp followed by progressive alopecia of the scalp, eyelashes and eyebrows. Associated eye findings include photophobia in childhood and corneal dystrophy. Due to the genetic and clinical heterogeneity of similar disorders, a definitive diagnosis of KFSD is often challenging. Towards identification of the causative gene we re-analysed a large Dutch KFSD family. 1M SNP arrays redefined the locus to a 2.9 Mb region at Xp22.12-Xp22.11. Screening of all 14 genes in the candidate region identified *MBTPS2* as the candidate gene carrying a c.1523A>G (p.Asn508Ser) missense mutation. The variant was also identified in two unrelated X-linked KFSD families and cosegregated with KFSD in all families. In symptomatic female carriers, skewed X-inactivation of the normal allele matched with increased severity of symptoms. *MBTPS2* is required for cleavage of sterol regulatory element-binding proteins (SREBPs). In vitro functional expression studies of the c.1523A>G mutation showed that sterol responsiveness was reduced by half. Other missense mutations in *MBTPS2* have recently been identified in patients with IFAP syndrome. We postulate that both phenotypes are in the spectrum of one genetic disorder with a partially overlapping phenotype.

Key words
Keratosis Follicularis Spinulosa Decalvans, MBTPS2, IFAP, Ichthyosis Follicularis

Introduction

Keratosis Follicularis Spinulosa Decalvans (KFSD; MIM# 308800) is a rare genetic disorder showing variable expression with women usually less severely affected than men. In 1926, Siemens was the first to describe KFSD in a large German family and several Dutch cases (Siemens HW, 1926). These Dutch cases belonged to a large family that was described by Lameris (Lameris, 1905) under the name 'Ichthyosis Follicularis'. In the two large pedigrees described by Lameris and Siemens (Lameris, 1905; Siemens HW, 1926), inheritance is clearly X-linked. However, generally, 50% of carrier females show symptoms of KFSD, which made Siemens postulate that inheritance in KFSD was of the X-linked intermediate type. More X-linked pedigrees have been described (Porteous et al., 1998). However inheritance is not always clear and isolated cases and autosomal dominant inheritance (i.e. male-to male transmission) has been described as well (Baden and Byers, 1994; Castori et al., 2009; Kuokkanen, 1971; Oosterwijk et al., 1997). KFSD manifests in infancy or early childhood with thorny keratotic follicular papules, progressive alopecia of the scalp, eyelashes and mainly lateral parts of the eyebrows with variable degrees of inflammatory change. Ocular abnormalities such as photophobia in childhood, punctate defects of the cornea, corneal dystrophy and blepharitis are common findings. Hyperkeratosis of elbows, knees, palms and soles as well as nail dystrophy may occur (Oosterwijk et al., 1997; Rand and Baden, 1983; Siemens HW, 1926; van Osch et al., 1992)

Due to the clinical and genetic heterogeneity of KFSD, a definite diagnosis is often challenging. KFSD resembles other dermatological entities such as keratosis pilaris and ulerythema ophrygenes (MIM# 604093), keratitis ichthyosis deafness syndrome (KID; MIM# 148210), ichthyosis follicularis atrichia photophobia (IFAP; MIM# 308205), keratosis pilaris atrophicans and atrophoderma vermiculatum.(Oranje et al., 1994). The question remains whether these syndromes are in actual fact variations of the same entity or truly independent disorders.

Using linkage analysis in the Dutch KFSD family, the locus was mapped to Xp21.2-22.2 (Oosterwijk et al., 1992b). Subsequently, the disease location was narrowed down to Xp22.13-p22.2 (Oosterwijk et al., 1995) and later refined to Xp22.13-p22.11 (Oosterwijk et al., 1997). This locus was confirmed in an X-linked family from the UK (Porteous et al., 1998), but lack of informative crossovers prevented detection of the KFSD gene. In 2002, spermidine/spermine N(1)-acetyltransferase (SSAT; MIM# 313020) was postulated as the causative gene for KFSD (Gimelli et al., 2002) . However, the *SAT1* gene is not included in the Dutch KFSD interval.

In order to identify the causative gene, we studied the large Dutch family and some other families with a clinical diagnosis of KFSD using new molecular tools. 1M SNP arrays were

used to refine the locus and to exclude the involvement of large deletions and duplications. Subsequently, genes in the candidate gene interval were screened for possible pathogenic variants using High Resolution Melting curve Analysis (HRMA).
Here, we show that KFSD patients carry mutations in the *MBTPS2* gene and that this affects the normal function of the protein by lowering MBTPS2 activity. While this work was in progress, deficiencies in the *MBTPS2* gene were shown to also cause IFAP syndrome (Oeffner et al., 2009). Together this sheds new light on the genetic and clinical heterogeneity of these related disorders.

Materials and methods

Study subjects

An extended multigenerational Dutch pedigree with 21 affected males, 15 unaffected males and 12 female carriers was available for molecular analysis (Figure 1a). Female carriers showed a variable phenotype ranging from severely affected to total absence of KFSD symptoms (van Osch et al., 1992).

Other families and cases with a clinical diagnosis of KFSD were available for *MBTPS2* analysis. Two families showed a clear X-linked mode of inheritance, confirmed by microsatellite marker analysis: a family from the USA as seen by one of us (unpublished data V.P.Sybert), and a family from the UK that was published previously (Herd and Benton, 1996; Porteous et al., 1998). The American family (Figure 1b) consisted of five affected males, five unaffected males and three carrier females. Four individuals were available for DNA analysis. Seven family members from the UK family (pedigree available in Herd and Benton et al, 1996) (Herd and Benton, 1996) were available for DNA analysis. Clinical features and *MBTPS2* genotypes are summarized in Table 1.

DNA and RNA analysis

DNA was extracted from whole blood following standard protocols. Total RNA was extracted from whole blood lymphocytes and from cultured fibroblasts using RNA-bee ^{Tm} (Bioconnect). cDNA was made using 1ug of total RNA according to standard protocols. Fibroblasts were cultured in DMEM (Gibco) with 10% FCS, 1% Penicilin + Streptomycin, 1% glutamine and 1% glucose at 37° C and 5%CO_2. Cells were harvested at 90% confluency prior to RNA isolation.

Array platforms

For SNP typing and CNV analysis, the Illumina Human1M BeadChip (Illumina Inc.,

San Diego, CA, USA) was used and a total of 750 ng DNA was processed according to the manufacturer's instructions. SNP copy number (log R ratio) and B-allele frequency were assessed using the software programs BeadStudio version 3.2 (Illumina).

High Resolution Melting Curve Analysis

High Resolution Melting curve Analysis (HRMA) was used as a pre-mutation screening in the Dutch KFSD family. Primers were designed to cover all exons and intron/exon boundaries of the 14 candidate genes (Fig 2) using Lightscanner primer design (Idaho). Known SNPs were avoided as much as possible. Sequences are available on request. Fragments were amplified using a touchdown PCR (65°-59°) directly followed by melting (LightCycler 480 Roche). To detect possible heteroduplex formation, patient DNA was mixed with DNA from unaffected male individuals in a 1:1 ratio. Melting curves of mixed samples were compared to those of unmixed samples and healthy controls. Samples showing aberrant melting curves were selected for direct sequence analysis.

Sequence analysis

PCR products were purified using spin columns according to the manufacturer's instructions (Qiagen, The Netherlands) and directly sequenced using Sanger sequencing (ABI 3730). High-throughput PCR products were purified using magnetic beads (Ampure) and prepared for sequencing with Sephadex plates (Edge Biosystems). Sequencing analysis was performed using the Mutation Surveyor program (SoftGenetics, USA). The human assembly (GRCh37/Hg 19) was used as a reference sequence. Nucleotide numbering follows HGVS recommendations and is based on a coding DNA reference sequence with nucleotide 1 corresponding to the A of the ATG translation initiation codon (www.hgvs.org/mutnomen). For *MBTPS2* exon 11 all available subjects from the Dutch pedigree were sequenced to study the segregation of the *MBTPS2* c.1523A>G mutation in this family. Cases from two other X-linked KFSD families from the USA and UK were also tested for this variant. The remaining 6 families were screened for all exons of the *MBTPS2* gene. 50 healthy unrelated controls were sequenced and data from pilot study 1 of the 1000 genome project (www.1000genomes.org, version October 2009) was used to study the variant frequency in the healthy population.

To confirm our findings on RNA level and to study possible differences in expression, *MBTPS2* mutation screening was performed at both RNA and genomic DNA levels. Possible damaging effects of missense mutations were assessed by PolyPhen and SIFT software. A web based *MBTPS2* gene variant database using the LOVD platform (Fokkema et al., 2005) was initiated to store and share all data collected (see www.LOVD.nl/MBTPS2).

X-chromosome inactivation

The methylation status of the X-chromosome was determined using the Androgen Receptor (AR) locus (Kubota et al., 1999). The maternal and paternal X chromosomes are distinguished by polymorphisms of a CAG repeat element in the AR gene while X-inactivation level is determined by the methylation of CpG dinucleotides in the *AR* gene.. Four carrier females and one affected male of the Dutch pedigree were genotyped for their CAG repeat length in the AR gene (Xq12) using blood derived DNA. The sequence of the forward primer with the fluorescent dye FAM was 5'-ACCGAGGAGCTTTCCAGAAT-3'. The sequence of the reverse primer was 5'-CTCATCCAGGACCAGGTAGC-3'. To determine the methylation status of the AR gene alleles in the carrier females, DNA was treated with bisulphite followed by a methylation specific PCR (according to the manufacturer's protocol EZ DNA methylation-Gold kit ™, Zymo research). Methylation differences were estimated based on peak heights in the PCR fragment analysis (Genemarker V1.70, Softgenetics).

Complementation assay

Growth of stably transfected CHO-K1-M19 cells was measured in cholesterol rich medium compared to cholesterol deficient medium. CHO-K1-M19 cells (*MBTPS2* deleted) were grown in a 1:1 mixture of Ham's F-12 medium and DMEM containing Glutamax (Invitrogen), nonessential amino acids, penicillin, streptomycin (Medium A), supplemented with 10% FCS and maintained at 37° C and 5%CO_2. Cells were stably transfected with 1.5 ug of a plasmid (pcDNA3.1 vector expressing a neomycin resistance gene, Invitrogen) expressing WT human *MBTPS2* or the N508S mutant *MBTPS2*. Stable transfection of CHO-K1-M19 cells with WT human *MBTPS2* restores the enzyme defect, thereby allowing growth in sterol deficient media. Cell growth of the N508S mutant transfected CHO-K1-M19 and WT transfected cells was measured in medium with- (Medium A with 5% FCS, 5µg/ml water-soluble cholesterol, 1 mM sodium mevalonate and 20 µM sodium oleate) and without sterols (Medium A with 5% lipoprotein deficient FCS). Growth of the stable transfected CHO-K1-M19 cells was measured after 6 days of culturing. Cells were harvested at day 6 and cells were counted in a hemocytometer. The complementation assay was performed as described by Oeffner et al (Oeffner et al., 2009).

Luciferase Reporter Assays

This assay, an indirect measure of the ability of *MBTPS2* cDNAs to restore sterol-regulated transcriptional activity in transfected M19 cells with a firefly pSRE-Luciferase reporter plasmid, was performed as described by Oeffner et al (Oeffner et al., 2009; Zelenski et al., 1999). In short, CHO-K1-M19 cells were transfected with expression plasmids without *MBTPS2* cDNA insert, wild-type or mutant *MBTPS2* and a reported plasmid (*pSRE-GL4.23)*

in which the luciferase reporter gene is under transcriptional control of the human LDL receptor promotor (Sterol Regulatory Element, SRE) and a Renilla luciferase plasmid (*pRL-SV40)* as a transfection control. Cells were cultured for 16h in Medium A and then switched to medium B. Firefly- and Renilla-luciferase activity were measured in a luminometer Auto Lumat LB953 (Berthold Technologies).

Figure 1: A. Dutch KFSD pedigree with twenty-one affected males and twelve carrier females. The family shows a clear X-linked pattern of inheritance, as proven by microsatellite marker analysis. Key recombinants (VII:12 and VI:29) determine the KFSD locus. **B.** KFSD family from the USA with five affected males and three carrier females, suggestive of X-linked inheritance.

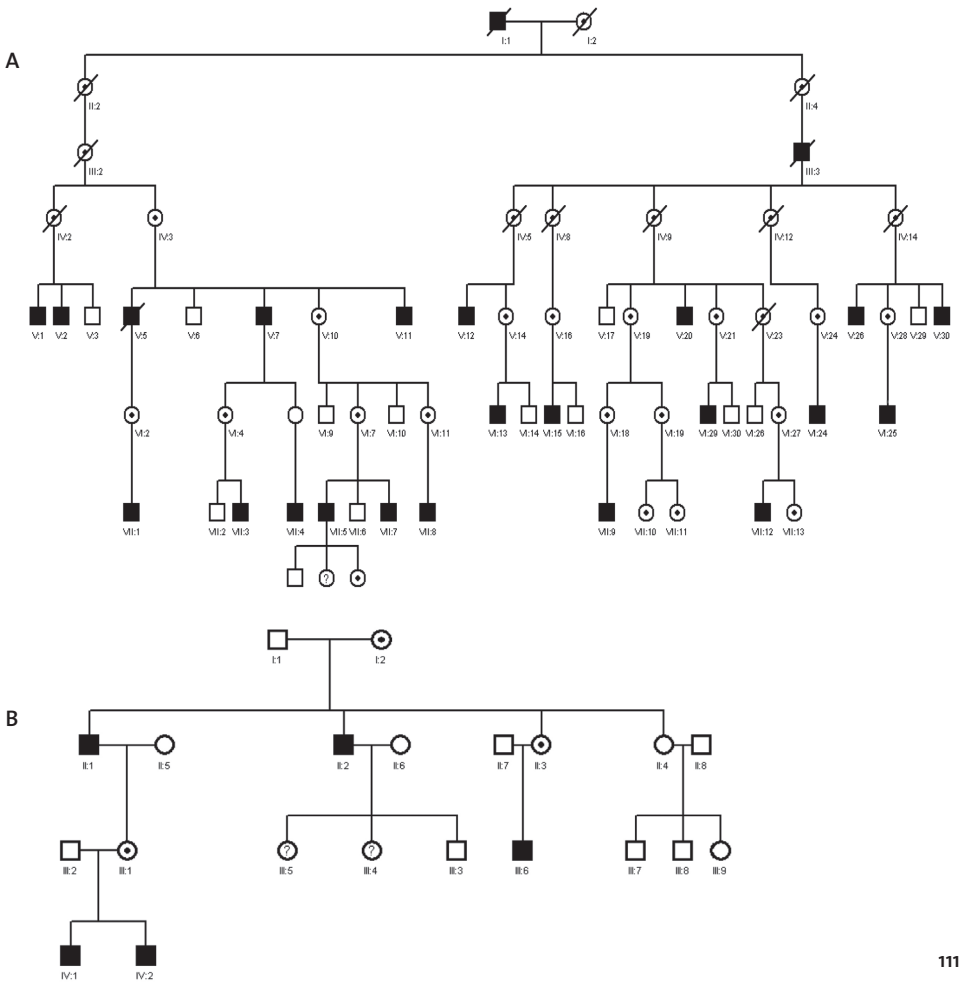


Table 1: Clinical features and *MBTPS2* genotypes of affected individuals and carrier females in 9 KFSD families.

Family/Proband	Clinical Features	MBTPS2 Genotype affected individuals
1 Dutch	van Osch et al., 1992	c.[1523A>G]+[0] in males c.[1524A>G]+[=] in females
2 UK	Herd and Benton et al., 1996	c.[1523A>G]+[0] in males c.[1524A>G]+[=] in females
3 USA/ II:1	dry bumpy skin, facial erythema, lack of eyebrows, absence body hair, photophobia, cracked feet with hyperkeratosis, unspecified alopecia	c.[1523A>G]+[0]
3 USA/III:1	dry skin, slightly thinned eyebrows with full eyelashes. sparse body hair	c.[1524A>G]+[=]
3 USA/IV:1	dry bumpy skin, lack of eyebrows, intact eyelashes, facial erythema, hyperkeratotic heels, severe photophobia, recurrent episodes of blepharitis, no hair loss on scalp, no scarring. Normal tooth development, Normal fingernails and cuticles.	c.[1523A>G]+[0]
3 USA/IV:2	dry bumpy skin, lack of eyebrows, lack of lower eyelashes, photophobia, facial erythema. Normal tooth development.	c.[1523A>G]+[0]
4-9 miscellaneous	Diagnosed by clinician as 'KFSD'	c.[=]+[0] in males c.[=]+[=] in females

All sequence information is based on GenBank reference sequence NM_015884.2.
Nucleotide numbering follows HGVS recommendations and is based on a coding DNA reference sequence with nucleotide 1 corresponding to the A of the ATG translation initiation codon (www.hgvs.org/mutnomen). c.[=] denotes a normal wild type sequence, c.[0] indicates that no paternal allele is present.

Immunohistochemistry

Skin biopsies of four carrier females and two affected males were obtained and stored at -80°C. Cryosections were cut at 0,05µm. Sections were fixed for 5' in cold acetone and blocked for 20' 0,12% H_2O_2 in demi water. After washing, sections were blocked in 1% NGS in

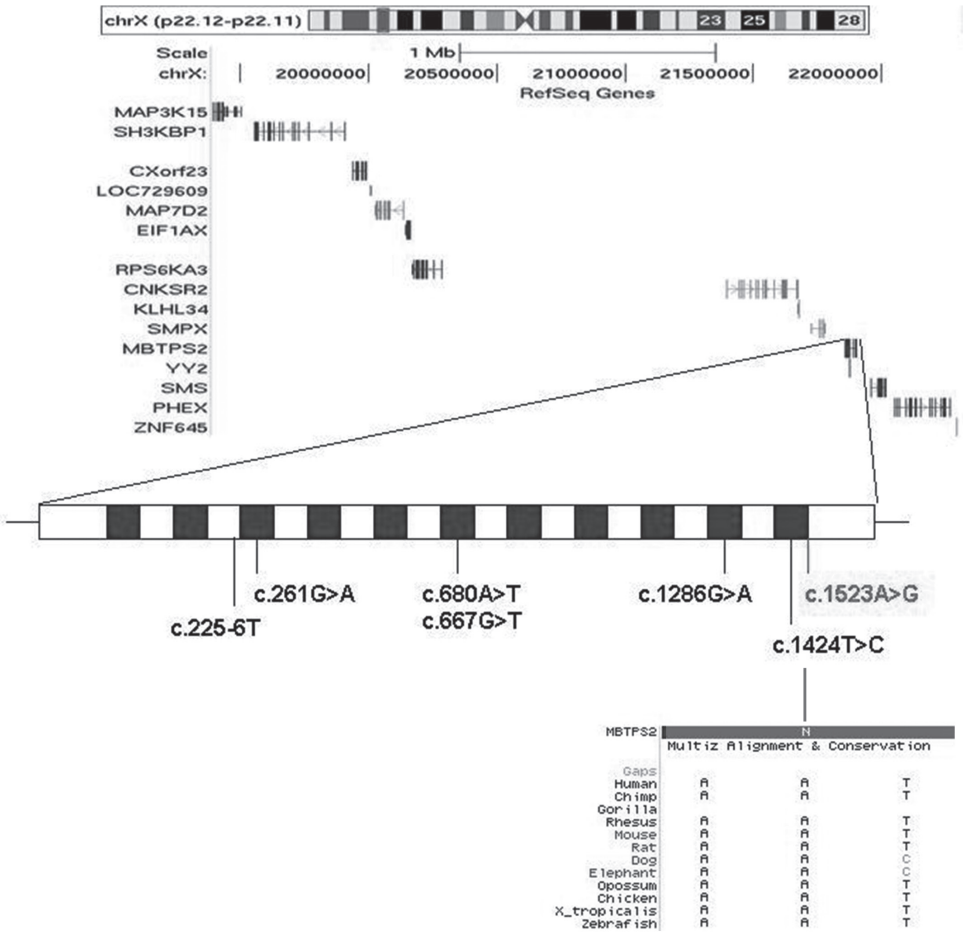
PBS and incubated with primary antibody overnight. Rabbit αMBTPS2 (Cell Signaling) (1:200 dilution in 1%NGS in PBS) and Mouse αKeratin 10 as a positive control (NeoMarkers Inc Fremont, California, USA.) (1:50 dilution in 1% NGS in PBS) were used as primary antibodies. Secondary antibody (Invitrogen) incubation in a 1: 1000 dilution for 1 hour was followed by incubation with Streptavidine-HRP antibody (Southern Biotechnology Associates) for one hour (1:2000). DAP+ was used to develop staining and sections were counterstained using Haematoxylin.

Results

Descendants of an extended Dutch pedigree (Oosterwijk et al., 1992a; Oosterwijk et al., 1992b; Oosterwijk et al., 1995; Oosterwijk et al., 1997) were subject to clinical and molecular analysis. First, to map the locus more accurately we analysed several members of each family using 1M SNP arrays. Analysis included the critical recombinants VII-12 and VI-29 (Figure 1a Dutch Pedigree) which defined the borders of the KFSD locus. Disease segregation fully matched with published linkage data (Oosterwijk et al., 1997). In addition we were able to redefine the region more precisely to a ~ 2.9 Mb region with the proximal breakpoint between chrX:19390769G>C and rs5955562 and the distal breakpoint between rs6528097 and rs4408025. Copy number variations (>5 consecutive SNPs) were not detected in this region or in any other genomic region, thereby excluding deletions or duplications as a possible genetic cause.

This refined region in Xp22.12-Xp22.11 contains 14 genes (Figure 2). High-Resolution Melting curve Analysis (HRMA) was used to screen these genes for sequence variants (Figure 2) in patients versus controls. Aberrant melt curves in one or more samples and thereby sequence variants were detected in 4 genes (Supp. Table S1). The variant in the *MBTPS2* gene was considered as the most promising disease-causing variant since it occurred in all affected males and changes a highly conserved amino acid (p.Asn508Ser) The c.1523A>G variant showed full co-segregation with the disease in the Dutch family (Figure 3a). We analysed the *MBTPS2* gene in two other KFSD families with an established X-linked inheritance, an UK family reported by Porteous et al (Herd and Benton, 1996; Porteous et al., 1998) and a USA family ascertained and investigated by Sybert (personal communication, unpublished data). In both families exactly the same c.1523A>G (p.Asn508Ser) variant was identified. The mutation was not detected by direct Sanger sequencing of 86 control chromosomes. Haplotype analysis around *MBTPS2* made a close relationship between these families very unlikely, the maximum region of overlap being at most 50.1 Kb (Supp. Table S2).

Figure 2: The KFSD locus. The KFSD linkage locus was redefined at Xp22.12-Xp22.11. High Resolution Melting curve Analysis (HRMA) identified *MBTPS2* as a candidate gene in the Dutch KFSD family. The detected variant in *MBTPS2* is indicated in red/gray. Previously identified mutations in IFAP syndrome are indicated in black.



Skin biopsies and whole blood samples were obtained from 7 individuals from the Dutch family. RNA analysis showed the expected variant in the RNA (c.1523A>G). Since the variant lies in the last exon we analysed a potential effect on RNA splicing using 3'RACE, but no effect could be detected. Carrier females showed variability in the expression of the mutated and the wild type allele (Figure 3b). To exclude RNA stability differences, the allelic expression was correlated to the level of X-inactivation (Xi) using a methylation assay. Imbalances in allelic expression perfectly matched with skewed levels of X-inactivation and

more interestingly with the clinical phenotype. This was most striking in a mother (VI-19) and her two daughters (VII-10 and VII-11). The mother preferentially expresses the mutant allele and has a moderate phenotype. One of her two daughters without any symptoms of KSFD only expresses the wild-type allele, while the second, mildly to moderately affected daughter expresses both the disease-and wild-type allele (Table 2).

Expression of *MBTPS2* mRNA in fibroblasts was studied by qRT-PCR using RNA isolated from two male KFSD patients and five carrier females. *MBTPS2* has two transcripts. The shorter transcript ranges from exon 1 up to exon 7, the full length transcript ranges from exon 1 up to exon 11. Using transcript specific primers, no significant differences in expression of the shorter or longer transcripts were found in fibroblast derived mRNA from carrier females (data not shown). Significantly higher expression of the full length *MBTPS2* transcript was found in carrier females (n=3) compared to controls (n=5), while there are no significant differences in expression in affected males (n=2) compared to controls (n=3) (Supp. Figure S1).

The amino acid Asn508 is located in an evolutionary conserved hydrophobic transmembrane region of the protein. Online prediction tools (Polyphen and Sift) indicate that the Asn508Ser amino acid substitution, which shows the same polarity, is benign. To test the effect on protein function, in particular the effect on sterol responsiveness, we used the in vitro tests developed by Oeffner et al. (Oeffner et al., 2009). Wild-type *MBTPS2* stably transfected into CHO-M19 cells were compared to Asn508Ser *MBTPS2* transfected cells and examined using complementation analysis. The number of cells transfected with the mutant *MBTPS2* able to grow in absence of sterols was lower compared to cells transfected with wild-type *MBTPS2* (Figure 4a). The luciferase reporter assay showed a clear reduction in sterol responsiveness in Asn508Ser *MBTPS2* transfected CHO-M19 cells, as compared to wild-type *MBTPS2* transfected cells (Figure 4b). The effects on the normal function of MBTPS2 were in the same range as those found in IFAP syndrome cases (Oeffner et al., 2009).

To study whether a functional effect of the c.1523A>G *MBTPS2* mutation could be detected; we have made an *in vivo* plasma lipid profile of an affected male. No abnormal lipid levels as compared to standard reference levels were detected (Cholesterol: 4.9 refs: 3.9-7.3 mmol/L, Triglycerides:1.7 refs: 0.8-2.3 mmol/L, HDL: 0.95 refs: 0.9-1.41 mmol/L, LDL: 3.18 refs: <3.0 mmol/L). Ultracentrifuge lipid profiling did not show abnormalities in VLDL-triglyceride value (0.95 mmol/L) or VLDL cholesterol/ratio (0.53). Immunohistochemical staining of MBTPS2 in skin biopsies (Supp. Figure S2 and S3) was performed for four carrier females and two affected males and three controls. The results did not show any clear differences between patients and controls and indicate normal expression *MBTPS2* levels for affected males, carrier females and controls.

Figure 3: *MBTPS2* mutation analysis in the Dutch KFSD family. a) Sanger sequencing of exon 11 in the *MBTPS2* gene identified a c.1523A>G (p.Asn508Ser) in all affected males of the Dutch KFSD family, not present in WT controls. The variant co-segregates with the disease in this family; affected male (VI-20) c.1523A>G , carrier female (VI-27) c.1523AG and unaffected male (VII-6) c.1523G. b) RNA analysis in fibroblasts in the Dutch KFSD family. cDNA sequencing in fibroblasts confirmed the presence of the c.1523A>G variant in an affected male. Carrier females (VI-19, VII-11, VI-27, and VII-10) showed variability in the expression of the mutated and the wild type allele.

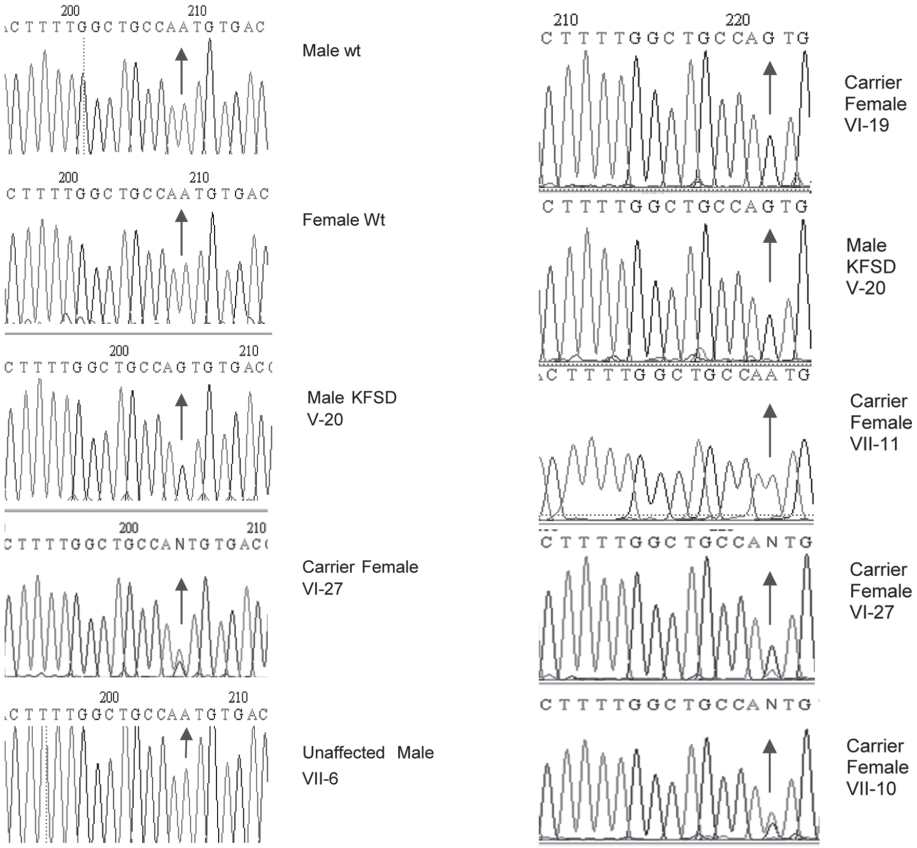


Table 2. X-inactivation patterns in the Dutch KFSD family.

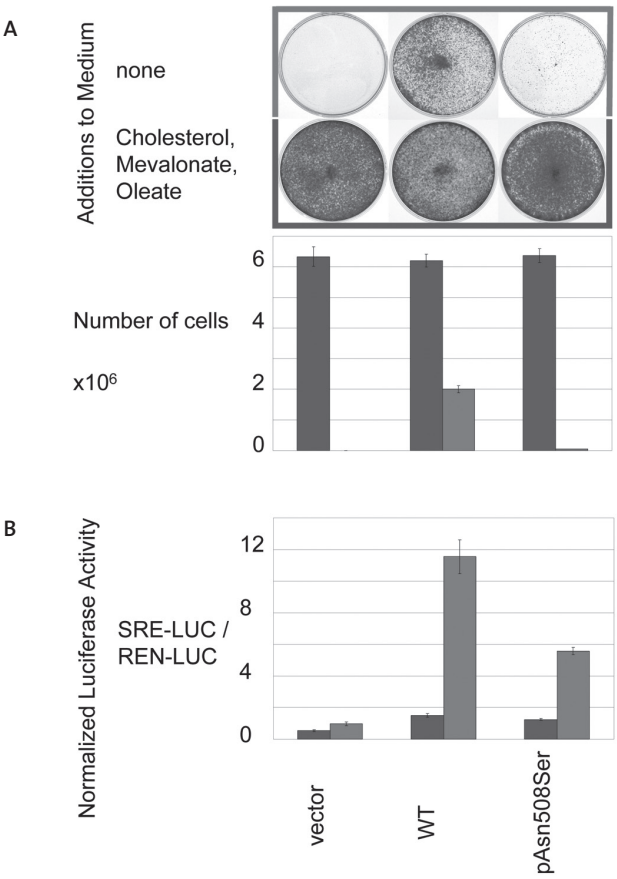
Pedigree/sex (m/f)	DNA *L	RNA *F	Phenotype	Xi pattern
V-20/m	G	G	full	-
VII-10/f	AG	AG	mild-moderate	50%-50%
VII-11/f	AG	A	none	100%-0%
VI-19/f	AG	G	moderate	100%-0%
VI-27/f	AG	a*G	mild	100%-0%

Levels of X-inactivation in carrier females were compared to the differences in RNA expression in fibroblasts. a* = less expression of the wild type allele compared to other carrier females.
The clinical phenotype of carrier females matched with differences in X-inactivation.

Discussion

Based on the criteria for designating a mutation as phenotype-modifying (Cotton and Scriver, 1998), the results from the current study provide enough evidence to establish that the c.1523A>G variant in *MBTPS2* cause X-linked KFSD. Screening of all genes in the candidate disease gene region, identified only one potential disease-causing variant, a missense change c.1523A>G (p.Asn508Ser) in *MBTPS2*. The variant was identified in two other unrelated X-linked KFSD families and perfectly co-segregates with the disease in all three families. The c.1523A>G variant was not detected by direct Sanger sequencing of 86 control chromosomes, it is not listed in the results of pilot study 1 of the 1000 genomes (www.1000genomes.org, checked October 2009) nor has it been reported in dbSNP. To facilitate future studies, in particular clinical diagnostic studies of skin defects, we established a gene variant database for *MBTPS2* (http://www.LOVD.nl/MBTPS2) following HGVS reccomendations (www.HGVS.org). The database currently contains 30 different variants of which 9 have been associated with pathogenicity (including KFSD and IFAP syndrome). The c.1523A>G variant affects a well conserved amino acid in a highly conserved gene. Differences in allelic expression of *MBTPS2* could be correlated to the clinical phenotype of carrier females. In fibroblasts and skin biopsies, qRT-PCR data and

Figure 4: Complementation assay and luciferase reporter assay. Functional studies of the c.1523A>G (p.Asn508Ser) variant using an in vitro assay testing sterol responsiveness. A. Complementation assay: Growth of stably transfected CHO-K1-M19 cells (lacking hamster MBTPS2) was measured in cholesterol rich medium (blue) and cholesterol deficient medium (red). The proportion of cells capable of growth is documented as framed photographs of the cultures and, graphically in bars by counts of growing stably transfected cells. Colours indicate absence (red) or presence (blue) of sterols. B. Luciferase reporter assay: luciferase activity functions as an indirect measure of the ability of MBTPS2 mutants to restore sterol-regulated transcriptional activity in transfected CHO-K1-M19 cells. Cells transfected with the c.1523A>G (p.Asn508Ser) variant are less able to restore sterol-regulated transcription compared to wild type when transferred from a cholesterol rich (blue bars) to a cholesterol deficient medium (red bars).



immunohistochemical staining of *MBTPS2* show normal expression levels in affected males. Western blot analysis and immunofluorescence experiments in fibroblasts (Supp. Figure S4) are in agreement with these findings and show normal expression levels and protein localization. Functional analysis of the mutant gene however, showed significant reduction in sterol responsiveness, indicating loss of proteolytic activity of the MBTPS2 protein.

In total, we have studied nine families with a clinical diagnosis of KFSD and identified the same mutation in the *MBTPS2* gene in 26 cases (three families). In six small families, lacking a clear X-linked mode of inheritance, the molecular basis of the disease is as yet unidentified. Although detailed clinical and histological data of these cases/families are not available, it is possible that they were misdiagnosed due to the similarity of KFSD to several other dermatological entities. Another possibility is that KFSD is a locus heterogeneous disorder and that there are other causative gene mutations. A critical review of these and other mutation negative cases may lead to other diagnosis or other loci. *MBTPS2* mutations have recently been identified in IFAP patients (Oeffner et al., 2009). In IFAP syndrome, the clinical severity varies to a large degree, while in KFSD most affected individuals have a relatively mild phenotype. This could be due to a genotype-phenotype effect, depending on the position of the mutation in *MBTPS2*. Elucidation of the MBTPS2 crystallography and its biochemical function is necessary to make clear correlations.

Until now, the diagnosis of KFSD has been based on clinical findings and its differential diagnosis has long been under debate (Oranje et al., 1994). Siemens described clinical findings of non-inflammatory, flesh-colored spinous 'thorns' leaving follicular scars where the skin subsequently becomes atrophic. It affects hair growth on the scalp, lateral eyebrows and eyelashes accompanied with the onset of alopecia during early adolescence. Patients can also have punctate defects on the cornea, palmoplantar hyperkeratosis with normal nails. IFAP (Macleod and Collins, 1908) was described coincidently also as 'Ichthyosis Follicularis' with noninflammatory spiny excrescences, hyperkeratosis and noncicatricial alopecia and photophobia. IFAP syndrome shares several features with KFSD, but differentiation is believed to be made on a congenital nonscarring nature of the alopecia. Until now, the small number of published cases, the phenotypic variability in both IFAP and KFSD with clinical overlap and the absence of a molecular cause made it impossible to determine if these syndromes were actual variations of the same entity or truly independent disorders. Much confusion has been generated in literature because of erroneous reporting of cases with lack of clinical details and reports of autosomal modes of inheritance. For instance the case report in which KFSD-like signs were present in a boy due to an unbalanced X-chromosomal duplication led to the notion that *SAT1* (Gimelli et al., 2002) was the plausible gene for KFSD. But putriscine levels and SSAT activity were

normal in the Dutch KFSD family, which makes SSAT overexpression a very unlikely cause for isolated (i.e. nonsyndromic) KFSD.

The pathogenesis of KFSD is largely unknown. An epidermal defect is highly likely since cells from the skin, hair and the cornea are involved. The epidermis has a protective function as the outer layer of the skin. The cornification of the skin's surface is compensated for by renewal of the epidermis, controlled by proliferation and differentiation of keratinocytes. As they move towards the upper layers, the keratinocytes become flatter and produce a mixture of lipids, cholesterol, free saturated fatty acids and ceramides into the intercellular spaces and thereby contributes to making the epidermis an effective barrier. Decreased cholesterol content in the stratum corneum could be attributable to the barrier function abnormality (Elias et al., 2008; Williams and Elias, 1981).

Several human genodermatoses have been described with mutations of genes involved in various aspects of lipid metabolism (Elias et al., 2008), such as Ichthyosis prematurity syndrome (Klar et al., 2009) , lamellar ichthyosis (Lefevre et al., 2003) , harlequin ichthyosis (Kelsell et al., 2005), Conradi-Hünermann-Happle syndrome, (Braverman et al., 1999), CHILD syndrome (Konig et al., 2000) and X-linked Ichthyosis (Basler et al., 1992), to which *MBTPS2* is now added, causing both IFAP and KFSD.

MBTPS2 functions as a metalloprotease, required for cleavage of sterol regulatory element-binding proteins (SREBPs). Within a feedback mechanism the active domain of SREBPs is cleaved by S2P and transported to the nucleus to function as a transcription factor of several targets genes, amongst which the LDL receptor gene. For both men and women, residual protease activity of MBTPS2 will most likely determine the severity of disease. However, measurements of proteolytic activity of MBTPS2 are not straightforward. Changes in proteolytic activity may have their effect elsewhere in the sterol regulated pathway which may result in general lipid abnormalities. Our study of serum lipids in one KFSD patient does support this hypothesis but rigourous investigation in a case-control study is needed given all the factors which affect this pathway.

When the X-activation is skewed towards expression of the mutated allele, an X-linked recessive disorder may affect female carriers, and the residual enzyme activity of MBTPS2 determines their phenotype. Therefore, skewed X-inactivation has been postulated as an explanation for heterogeneity in KFSD carrier females given the fact that the number of symptomatic carrier females seems larger then expected on the basis of random lyonisation. We have shown that differences in Xi-pattern expressions indeed may be correlated to the severity of KFSD symptoms in females. It is remarkable that while in some X-linked recessive diseases (Duchenne Muscular Dystrophy; MIM# 310200, Red-Green Colour blindness MIM#s 303800, 303900) females rarely show symptoms, in others, like KFSD and OTC (MIM# 311250), this is frequent. Interestingly, Siemens used KFSD

as the first disorder in which he recognised X-linked intermediate inheritance (Siemens HW, 1926). However, the distinction between different modes of X-linked inheritance is a mere quantative aspect of several underlying mechanisms, one of them being skewed Lyonisation. The standard concepts of dominance or recessiveness do often not apply to X-linked diseases (Dobyns et al., 2004) and KFSD should therefore also be simply described as X-linked inheritance.

Conclusion

This study supports a new approach towards patients who are referred to a medical clinic with KFSD or IFAP. First, critical assessment of all the symptoms together with a thorough family history is important for establishing a differential diagnosis. In patients that show the triad of follicular ichthyosis (follicular hyperkeratosis), total or subtotal atrichia of scalp, eyebrows or eyelashes and photophobia at a young age, IFAP/KFSD should be considered. When X-linked inheritance cannot be excluded, mutation analysis for *MBTPS2* is warranted and will be crucial to confirm this diagnosis. To prevent more confusion in the nosology of this disorder, we propose a new name should be chosen for KFSD/IFAP syndrome and suggest this name will be used when a *MBTPS2* mutation has been detected. If *MBTPS2* mutation analysis is negative, the extension '-like syndrome' should be added. This new nomenclature should be introduced when DNA analysis has been performed in enough KFSD/IFAP and KFSD/IFAP-like families and all clinical features have been reviewed.

Acknowledgements

We gratefully acknowledge all patients for their kind collaboration to publish clinical data. We thank M. Pietila, Kuopio, Finland for putriscine levels and SSAT activity measurements. B. Loeys is thanked for his cooperation in completing the genealogy records of the Dutch family. Technicians from the Department of Cytogenetics and LDGA are kindly thanked for technical assistance. We would like to thank Y. Ariyurek for assistance with the 1M arrays and Dr. A.H.M Smelt for his help and expertise in lipid profiling. Drs H. de Kort is thanked for her expertise in immunohistochemistry. Drs S. Commandeur and Dr. R. van Doorn are kindly thanked for supplying the Mouse αKeratin 10 antibody and useful suggestions. Dr. W.M.C van Roon-Mom, T. Messemaker, E. de Meier, J. Celli and Y. Lai are acknowledged for all their help and support.

Baden HP, Byers HR. 1994. Clinical findings, cutaneous pathology, and response to therapy in 21 patients with keratosis pilaris atrophicans. Arch Dermatol 130:469-475.

Basler E, Grompe M, Parenti G, Yates J, Ballabio A. 1992. Identification of point mutations in the steroid sulfatase gene of three patients with X-linked ichthyosis. Am J Hum Genet 50:483-491.

Braverman N, Lin P, Moebius FF, Obie C, Moser A, Glossmann H, Wilcox WR, Rimoin DL, Smith M, Kratz L, Kelley RI, Valle D. 1999. Mutations in the gene encoding 3 beta-hydroxysteroid-delta 8, delta 7-isomerase cause X-linked dominant Conradi-Hunermann syndrome. Nat Genet 22:291-294.

Castori M, Covaciu C, Paradisi M, Zambruno G. 2009. Clinical and genetic heterogeneity in keratosis follicularis spinulosa decalvans. Eur J Med Genet 52:53-58.

Cotton RG, Scriver CR. 1998. Proof of "disease causing" mutation. Hum Mutat 12:1-3.

Dobyns WB, Filauro A, Tomson BN, Chan AS, Ho AW, Ting NT, Oosterwijk JC, Ober C. 2004. Inheritance of most X-linked traits is not dominant or recessive, just X-linked. Am J Med Genet A 129A:136-143.

Elias PM, Williams ML, Holleran WM, Jiang YJ, Schmuth M. 2008. Pathogenesis of permeability barrier abnormalities in the ichthyoses: inherited disorders of lipid metabolism. J Lipid Res 49:697-714.

Fokkema IF, Den Dunnen JT, Taschner PE. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26:63-68.

Gimelli G, Giglio S, Zuffardi O, Alhonen L, Suppola S, Cusano R, Lo NC, Gatti R, Ravazzolo R, Seri M. 2002. Gene dosage of the spermidine/spermine N(1)-acetyltransferase (SSAT) gene with putrescine accumulation in a patient with a Xp21.1p22.12 duplication and keratosis follicularis spinulosa decalvans (KFSD). Hum Genet 111:235-241.

Herd RM, Benton EC. 1996. Keratosis follicularis spinulosa decalvans: report of a new pedigree. Br J Dermatol 134:138-142.

Kelsell DP, Norgett EE, Unsworth H, Teh MT, Cullup T, Mein CA, Dopping-Hepenstal PJ, Dale BA, Tadini G, Fleckman P, Stephens KG, Sybert VP, Mallory SB, North BV, Witt DR, Sprecher E, Taylor AE, Ilchyshyn A, Kennedy CT, Goodyear H, Moss C, Paige D, Harper JI, Young BD, Leigh IM, Eady RA, O'Toole EA. 2005. Mutations in ABCA12 underlie the severe congenital skin disease harlequin ichthyosis. Am J Hum Genet 76:794-803.

Klar J, Schweiger M, Zimmerman R, Zechner R, Li H, Torma H, Vahlquist A, Bouadjar B, Dahl N, Fischer J. 2009. Mutations in the fatty acid transport protein 4 gene cause the ichthyosis prematurity syndrome. Am J Hum Genet 85:248-253.

Konig A, Happle R, Bornholdt D, Engel H, Grzeschik KH. 2000. Mutations in the NSDHL gene, encoding a 3beta-hydroxysteroid dehydrogenase, cause CHILD syndrome. Am J Med Genet 90:339-346.

Kubota T, Nonoyama S, Tonoki H, Masuno M, Imaizumi K, Kojima M, Wakui K, Shimadzu M, Fukushima Y. 1999. A new assay for the analysis of X-chromosome inactivation based on methylation-specific PCR. Hum Genet 104:49-55.

Kuokkanen K. 1971. Keratosis follicularis spinulosa decalvans in a family from northern Finland. Acta Derm Venereol 51:146-150.

Lameris. 1905. Ichthyosis Follicularis. Nederlands Tijdschrift Voor Geneeskunde 41:1524.

Lefevre C, Audebert S, Jobard F, Bouadjar B, Lakhdar H, Boughdene-Stambouli O, Blanchet-Bardon C, Heilig R, Foglio M, Weissenbach J, Lathrop M, Prud'homme JF, Fischer J. 2003. Mutations in the transporter ABCA12 are associated with lamellar ichthyosis type 2. Hum Mol Genet 12:2369-2378.

Macleod JM, Collins ET. 1908. Two Cases of Advanced "Keratosis follicularis," associated with Baldness. Proc R Soc Med 1:27-31.

Oeffner F, Fischer G, Happle R, Konig A, Betz RC, Bornholdt D, Neidel U, Boente MC, Redler S, Romero-Gomez J, Salhi A, Vera-Casano A, Weirich C, Grzeschik KH. 2009. IFAP syndrome is caused by deficiency in MBTPS2, an intramembrane zinc metalloprotease essential for cholesterol homeostasis and ER stress response. Am J Hum Genet 84:459-467.

Oosterwijk JC, Nelen M, van Zandvoort PM, van Osch LD, Oranje AP, Wittebol-Post D, van Oost BA. 1992a. Confirmation of X-linked inheritance and provisional mapping of the keratosis follicularis spinulosa decalvans gene on Xp in a large Dutch family. Ophthalmic Paediatr Genet 13:27-30.

Oosterwijk JC, Nelen M, van Zandvoort PM, van Osch LD, Oranje AP, Wittebol-Post D, van Oost BA. 1992b. Linkage analysis of keratosis follicularis spinulosa decalvans, and regional assignment to human chromosome Xp21.2-p22.2. Am J Hum Genet 50:801-807.

Oosterwijk JC, Richard G, van der Wielen MJ, van d, V, Harth W, Sandkuijl LA, Bakker E, van Ommen GJ. 1997. Molecular genetic analysis of two families with keratosis follicularis spinulosa decalvans: refinement of gene localization and evidence for genetic heterogeneity. Hum Genet 100:520-524.

Oosterwijk JC, van der Wielen MJ, van d, V, Voorhoeve E, Bakker E. 1995. Refinement of the localisation of the X linked keratosis follicularis spinulosa decalvans (KFSD) gene in Xp22.13-p22.2. J Med Genet 32:736-739.

Oranje AP, van Osch LD, Oosterwijk JC. 1994. Keratosis pilaris atrophicans. One heterogeneous disease or a symptom in different clinical entities? Arch Dermatol 130:500-502.

Porteous ME, Strain L, Logie LJ, Herd RM, Benton EC. 1998. Keratosis follicularis spinulosa decalvans: confirmation of linkage to Xp22.13-p22.2. J Med Genet 35:336-337.

Rand R, Baden HP. 1983. Keratosis follicularis spinulosa decalvans. Report of two cases and literature review. Arch Dermatol 119:22-26.

Siemens HW. 1926. Keratosis Follicularis Spinulosa Decalvans. Arch Dermat Syphilil 151:384-387.

van Osch LD, Oranje AP, Keukens FM, Voorst Vader PC, Veldman E. 1992. Keratosis follicularis spinulosa decalvans: a family study of seven male cases and six female carriers. J Med Genet 29:36-40.

Williams ML, Elias PM. 1981. Stratum corneum lipids in disorders of cornification: increased cholesterol sulfate content of stratum corneum in recessive x-linked ichthyosis. J Clin Invest 68:1404-1410.

Zelenski NG, Rawson RB, Brown MS, Goldstein JL. 1999. Membrane topology of S2P, a protein required for intramembranous cleavage of sterol regulatory element-binding proteins. J Biol Chem 274:21973-21980.

Supplementary information

Figure S1: MBTPS2 levels in KFSD males compared to control males.

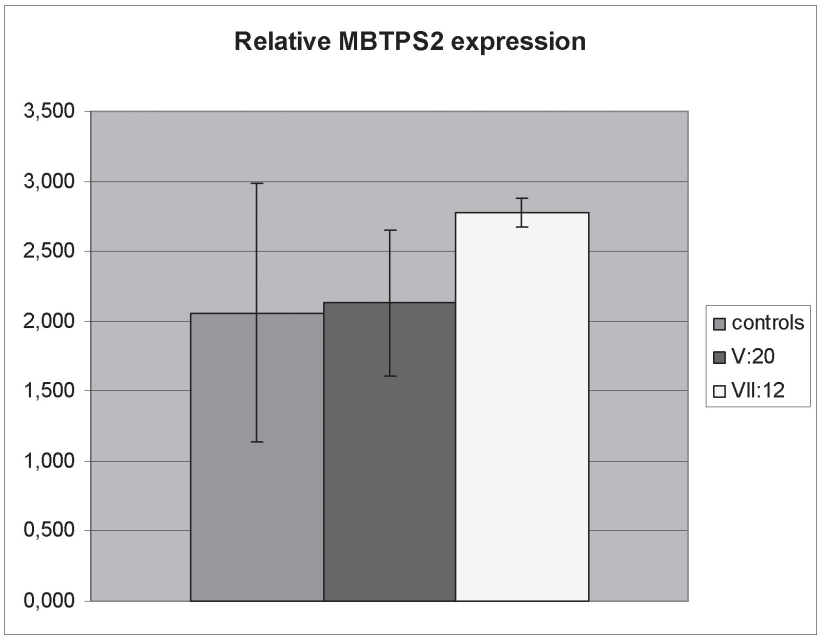
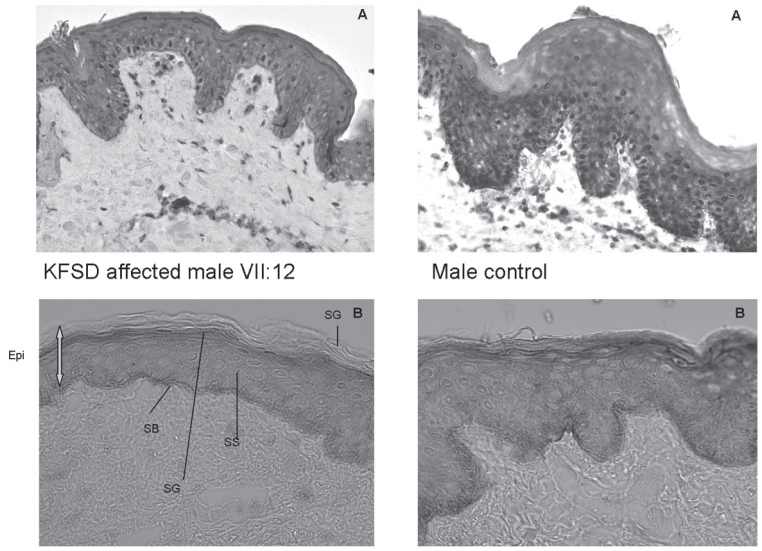


Figure S2: Skin biopsies stained with αM*BTPS2* (brown) with **(A)** and without Hematoxylin **(B)**. 40x and 20x magnification.
Epi=epidermis, SC=stratum corneum, SG=stratum granulosum, SS=stratum spinulosum, SB=stratum basale.



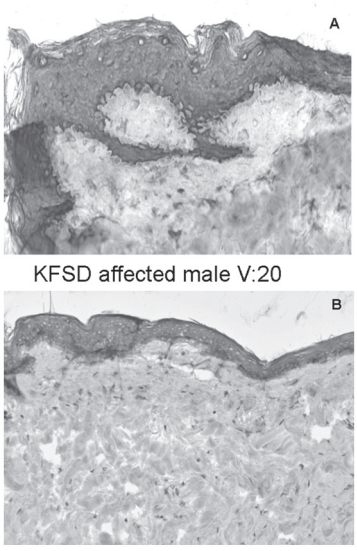


Figure S3: Skin biopsy stained with αKeratin-10 **(A)** and only counterstaining with Hematoxylin **(B)**. 20x and 10x magnification. Keratin-10 is expressed in the epidermis, except the stratum basale.

Figure S4: Cellular localisation of MBTPS2 in fibroblasts. There is no difference between a control male (a) and a male KFSD patient VII-12 (b). *MBTPS2* (green fluorescence) is found both in the cytoplasm and nucleus. Nuclei are stained in blue with DAPI. Red is a control staining of β-actin.

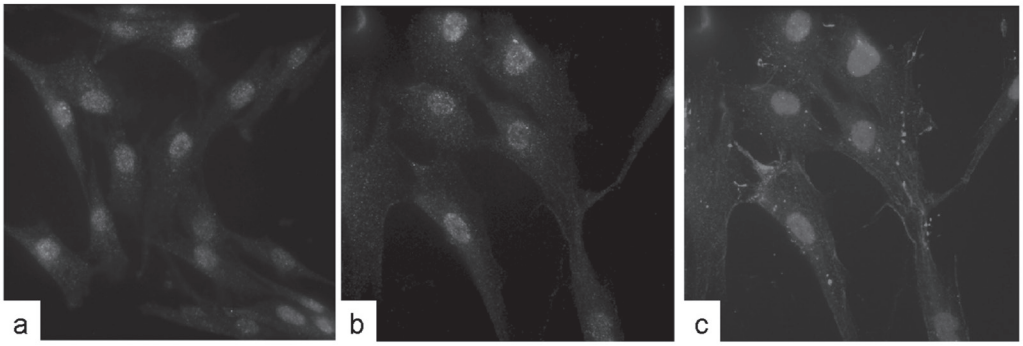


Table S1:. Variants detected in the Dutch KFSD family using High Resolution Melting curve Analysis.

Gene	Variant	SNP
FLJ14503	NM_152780.2:c.545A>G	*rs34519770*
MAP3K15	NM_001001671.2:c.1533+19G>G	*rs56058646*
MBTPS2	NM_015884.2:c.1523A>G	-
ZNF645	NM_152577.2:c.498C>A	*rs5951426*

All sequence information is based on GenBank reference sequences. Nucleotide numbering follows HGVS recommendations and is based on a coding DNA reference sequence with nucleotide 1 corresponding to the A of the ATG translation initiation codon (www.hgvs.org/mutnomen).

Table S2: Haplotype comparison around *MBTPS2*.

SNP	UK III-4 genotype	USA IV-I genotype	Dutch VII-12 genotype	Allele frequency HAPMAP-CEU
rs5951636: [T]	T	T	T	0.750
rs16981675:[T]	T	T	T	1.00
rs5951640:[C]	C	C	C	0.983
rs4446856:[T]	T	T	T	0.992
rs6653655:[C]	C	C	C	1.00

Genotype analysis around the c.1523A>G variant in *MBTPS2* of affected KFSD males from three families (UK, USA, Dutch) showed a common haplotype between SNP *rs5951636* and *rs6653655* with a maximum region of overlap of 50.1 Kb. Allele frequencies of SNPs in this interval indicate a close relationship of these families is not plausible. Polymorphism nomenclature follows dbSNP identifiers and HGVS recommendations (www.hgvs.org/mutnomen) with [..] indicating the major allele.


6

Terminal Osseous Dysplasia is caused by a single recurrent mutation in the FLNA Gene

Yu Sun,* Rowida Almomani,* Emmelien Aten, Jacopo Celli, Jaap van der Heijden, Hanka Venselaar, Stephen P. Robertson, Anna Baroncini, Brunella Franco, Lina Basel-Vanagaite, Emiko Horii, Ricardo Drut, Yavuz Ariyurek, Johan T. den Dunnen, Martijn H. Breuning

*These authors contributed equally to this work

Am J Hum Genet. 2010;87(1):146-53.

Summary

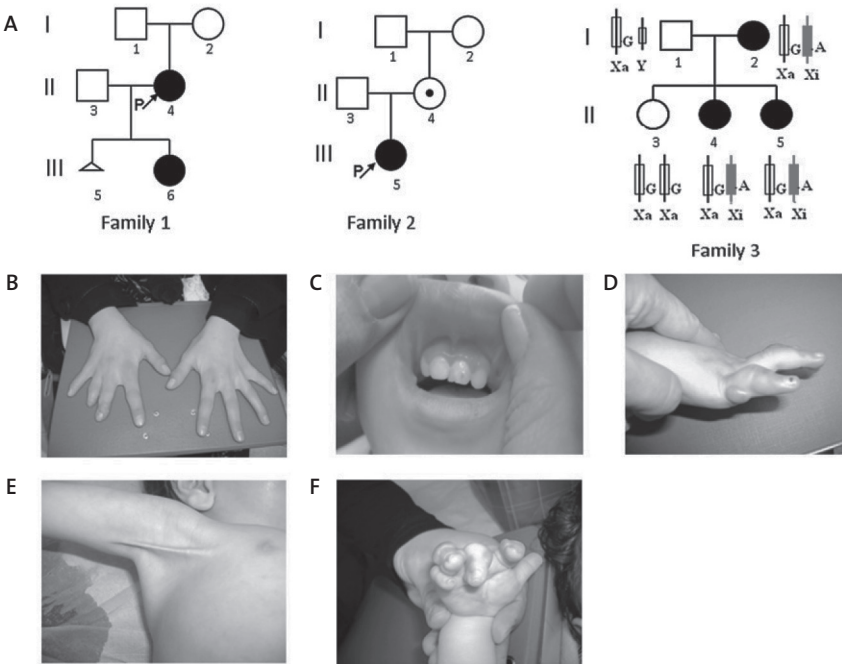
Terminal Osseous Dysplasia (TOD) is an X-linked dominant male-lethal disease characterized by skeletal dysplasia of the limbs, pigmentary defects of the skin and recurrent digital fibroma with onset in female infancy. After performing X-exome capture and sequencing we identified a mutation at the last nucleotide of exon 31 of the *FLNA* gene as the most likely cause of the disease. The variant c.5217G>A, was found in 6 unrelated cases (three families and three sporadic cases) and not in 400 control X chromosomes, nor pilot data from the 1000 Genomes Project or the *FLNA* gene variant database. In the families the variant segregated with the disease and it was transmitted four times from a mildly affected mother to a more seriously affected daughter. We show that, due to non-random X-inactivation, the mutant allele was not expressed in patient fibroblasts. RNA expression of the mutant allele was only detected in cultured fibroma cells obtained from 15-year old surgically removed material. The variant activates a cryptic splice site, removing the last 48 nucleotides from exon 31. At the protein level this results in a loss of 16 amino acids (p.Val1724_Thr1739del), predicted to remove a sequence at the surface of filamin repeat 15. Our data show that TOD is caused by this single recurrent mutation in the *FLNA* gene.

Terminal Osseous Dysplasia (OMIM 300244) is a rare condition, characterized by terminal skeletal dysplasia, pigmentary defects of the skin, and recurrent digital fibromata during infancy. It has been described as a male-lethal X-linked dominant disease in the previously reported families and cases[1]. Linkage studies mapped the mutation to Xq27.3-q28[2]. However, no disease-causing gene was discovered yet.

In the present study, we examined TOD in three families and three sporadic cases (patients 1, 2 and 3 described by Horii[3], Drut[4] and Breuning[5]). The Dutch (Figure 1A, family 1) and Italian families (Figure 1A, family 2) have been described before (Breuning[5] and Baroncini[6]). The third family (Figure 1A, family 3) has not been reported before, it is non-consanguineous and of Israeli-Arab origin. All patients, a mother and her two daughters have normal cognitive development. The mother (3I:2) suffers from chronic mild obstructive lung disease and vitamin B12 deficiency. Since her childhood she had multiple minor surgeries to remove small skin lesions from her hands and legs. On clinical examination at age 25 years her head circumference was 54 cm (25-50%), height 170 cm (75-90%), arm span 171 cm. Her right hand showed brachydactyly of digit III-V, a short fingernail on digitus IV and lateral deviation of the 5th digit. On her left hand there was lateral deviation of the 4th digit with a small lesion on the lateral aspect of the distal phalanx and clinodactyly of the 5th digit (Figure 1B). Her right foot showed a short and highly implanted 4th digit. There was bilateral widening of the distal portion of the 2nd-5th digit. She had no gingival extra frenulum and no pterygium. A skeletal X-ray survey revealed unilateral flattening of her vertebral bodies at L1-L3, secondary right scoliosis and wedging of her L1 vertebral body. Her daughter (3II:4) underwent surgery at 2 months of age to remove small skin lesions from her hands, feet and gingiva. On clinical examination at age three she had a head circumference of 48 cm (25-50%), a height of 85 cm (<3%) and a of weight 11.1 kg (<3%). She showed hypertelorism - interpupillary distance of 5.4cm (>97%), a right epicanthal fold, a normal palate, an upper and lower accessory frenulum (Figure 1C), a short neck and a short thorax. Despite earlier surgery, she had bilateral skin lesions on her 2nd and 5th digits and bilateral clinodactyly of the 5th digit (Figure 1D). Her feet showed a lesion in her 3rd toes and thickening of the nail of the 5th toes bilaterally. A skeletal X-ray survey revealed bilateral lytic lesions in the proximal humerus, the proximal femur, and multiple soft tissue lesions in her feet and hands. The youngest daughter (3II:5) was born with multiple lesions on her hands and feet including bilateral camptodactyly of the 3rd digit, and bilateral overriding of the 4th toe. On the echocardiogram she had persistent foramen ovale at birth. On clinical examination at age six months her head circumference was 42 cm (25-50%), height 58.8 cm (<3%) and weight 5.1 kg (<3%). She has mild hypertelorism, three brownish pigmented spots of different size (3 mm to 1.5 cm) in her right temporal groove, mild retrognathia, a right upper accessory frenulum and a cleft palate, a short neck and a

short thorax. She has a bilateral axillary pterygium (Figure 1E), which is more severe on the right side. Bilaterally, there is limited extension of her elbows with normal supination and pronation of her hands. In her right hand (Figure 1F) she had multiple skin lesions on her 2nd - 5th digits, clinodactyly and lateral deviation of her 2nd and 3rd digits and a narrow 5th digit with an absent distal crease. Her left hand showed skin lesions on her 2nd - 4th digits. Her 2nd digit was narrow and laterally deviated. There was camptodactyly of the 3rd-5th digits, brachydactyly and clinodactyly of the 5th digit and absence of a distal crease. In her feet she had bilateral plantar pits. The right foot has distal broadening of 2nd-5th toe, brachydactyly of the 2nd and 3rd toe accompanied by syndactyly. There was overriding of the 3rd and 4th toe. On her left foot the 2nd-5th toes were distally broad. She had a overlapping of the 2nd and 4th toes over her 3rd toe, brachydactyly of the 3rd

Figure 1: The pedigrees and the phenotype of family 3. (A) The pedigrees investigated in this study, in family 3 X inactivation patterns show the silencing of the X chromosome which carries the mutant allele. (B) shows the hands of 3I:2. (C) Multiple frenula of 3II:4. (D) The right hands of 3II:4. She has clinodactyly and digital fibroma. (E) shows the right upper accessory frenulum of 3II:5. (F) The right hand of 3II:5.



toe that was a high implanted. A skeletal X-ray survey revealed bilateral lytic lesions of the proximal humerus, lytic lesions of the left proximal femur, and multiple soft tissue lesions. She had underdeveloped tarsal bones in her feet. The phenotypes from different patients are summarized in Table 1.

DNA of patients and family members were extracted from peripheral blood (family 1, 2, 3), buccal cells (patient 1) or paraffin embedded tissue (patient 2, 3). Two probands (1II:4, 2III:5) of the Dutch and the Italian families were tested using the X-exome target enrichment methodology (SureSelect, Agilent) and next generation sequencing (Illumina Genome Analyzer II). The methods used for sequence capture, enrichment and elution followed instructions and protocols provided by the manufacturers (SureSelect, Agilent) with a little modification. In brief, 500ng of DNA was fragmented (Bioruptor, Diagenode) according to manufacturer's instructions to yield fragments from 200-300 bp. Paired-end adaptor oligonucleotides from Illumina were added to both ends. The DNA-adaptor ligated fragments were then hybridized to 250ng of SureSelect X-chromosome Oligo capture library (SureSelect, Agilent) for 14 hours. After Hybridization, washing and elution, the elute was amplified to create sufficient DNA template for downstream applications. The eluted-enriched DNA fragments were sequenced using the Illumina technology platform. We prepared the paired–end flow-cell on the supplied cluster station following the instructions of the manufacturer.

The reads were aligned to the reference human genome (hg 18, NCBI Build 36.2) by Bowtie[7] (Supplementary Table S1). Substitution variant calling was performed by searching for positions where a variant nucleotide was present in more than 30% of the reads. After removing substitutions present with high frequency in dbSNP, the variants located in the previously identified TOD linkage interval, Xq27.3-q28, were listed in Table 2. From these variants, c.5217G>A, the only variant shared by the two patients, in the *FLNA* gene were selected for further study because; (i) c.5217G>A, the last nucleotide of exon 31, was predicted to affect splicing by Human Splicing Finder[8]. The score of splicing donor site dropped from 91.2 to 80.63, indicating the wild type site may not function as usual, (ii) mutations in *FLNA* have been reported to be involved in diseases showing a partial phenotypic overlap with TOD[9].

Sanger sequencing results confirmed the presence of c.5217G>A (Figure 2A) and c.5850T>C (Figure 2B) in all affected cases (1II:4, 1III:6) in family 1 as well as c.5686+84A>G found in an intron, but not in an unaffected individual (1I:2). Further evidence came from the analysis of the Italian family where affected cases (2II:4, 2III:5) carry exactly the same variant; c.5217G>A, here together with another exonic variant c.5814C>T. Unfortunately, we did not have access to material from both parents and therefore we could not determine whether the mutations occurred de novo.

Table 1. Clinical features of the patients studied in this report.

	1II:4	1III:6	2II:4	2III:5	3I:2	3II:4	3 II:5	Patient 1	Patient 2	Patient 3
Origin	Dutch	Dutch	Italian	Italian	Israeli-Arab	Israeli-Arab	Israeli-Arab	Japanese	Argentina	Dutch
Age of onset	1 mo	3 mo		7 mo		2 mo	birth	3 mo		4 mo
Pigmentary skin anomalies										
Face	+	+	-	+			+	+	+	+
Scalp	-							-	+	-
Fibromatosis										
Digital fibromas	+	+	-	+	+	+	+	+	+	+
Limbs and skeletal system										
Synadactyly	-	-	-	+		+	+	-	-	-
Brachydactyly	+		-	+	+		+	+		
Clinodactyly			-	+	+	+	+			
Camptodactyly				+			+			
Metacarpal disorganization	+	+	-	+				+	+	+
Metatarsal disorganization	+	+	-	+			+	+	+	+
Limb long bones anomalies	-	+	-	+	-	+	+	+	+	+
Articular abnormalities	+	+	-	+				+	+	+
Facial and mouth Features										
Cleft palate	-	-	-	-		-	+	-	-	
Upslanting palpebral fissures	-		-	+				+		
Hypertelorism/Telecanthus	+		-	-		+	+	+		
Epicanthic folds	-		-	+		+		+		
Coloboma of Iris	-	+	-	-				-	-	-
Flat/depressed nasal tip	-	+	-	-				+	-	
Thick lips/Prominent	+		-	+						
Lower lip										
Papillomata	-	-	-							-
Multiple frenula			+	+	-					
Preauricular pits and tags	+							-		

Table 2: List of all exonic variants with low frequency in the European population in Xq27.3-Xq28.

HGVS name	Gene	Predicted Function	Predicted Protein Change	1II:4	2III:5	
NM_002025.2:c.1653A>G	*AFF2*	Silent	p.(=)	-	+	
NM_001183.4:c.*461A>C	*ATP6AP1*	3' UTR	p.(=)	-	+	
NM_001009932.1:c.364G>A	*DNASE1L1*	Silent	p.(=)	+	-	
NM_001110556.1:c.5217G>A	*FLNA*	Silent	p.(=)	+	+	
NM_001110556.1:c.5814C>T	*FLNA*	Silent	p.(=)	-	+	rs2070825, high frequency in a group of multiple population
NM_001110556.1:c.5850T>C	*FLNA*	Silent	p.(=)	+	-	Doesn't segragate with phenotype
NM_004961.3:c.186G>A	*GABRE*	Silent	p.(=)	+	-	
NM_005342.2:c.166G>C	*HMGB3*	Missense	p.(Glu56Gln)	-	+	
NM_005367.4:c.888A>G	*MAGEA12*	Silent	p.(=)	-	+	
NM_005362.3:c.455G>T	*MAGEA6*	Missense	p.(Ser152Ile)	+	-	Repetitive region
NM_005365.4:c.92C>A	*MAGEA9*	missense	p.(Pro31His)	+	-	Repetitive region
NM_001170944.1:c.468C>T	*PNMA6B*	Silent	p.(=)	+	-	
NM_005629.3:c.324A>G	*SLC6A8*	Silent	p.(=)	-	+	
NM_032539.2:c.1002T>C	*SLITRK2*	Silent	p.(=)	-	+	
NM_032539.2:c.309G>A	*SLITRK2*	Silent	p.(=)	-	+	
NM_001009615.1:c.240C>A	*SPANXN2*	Silent	p.(=)	+	-	
NM_014370.2:c.1014G>A	*SRPK3*	Silent	p.(=)	+	-	
NM_006280.1:c.430G>A	*SSR4*	Missense	p.(Gly144Arg)	+	-	

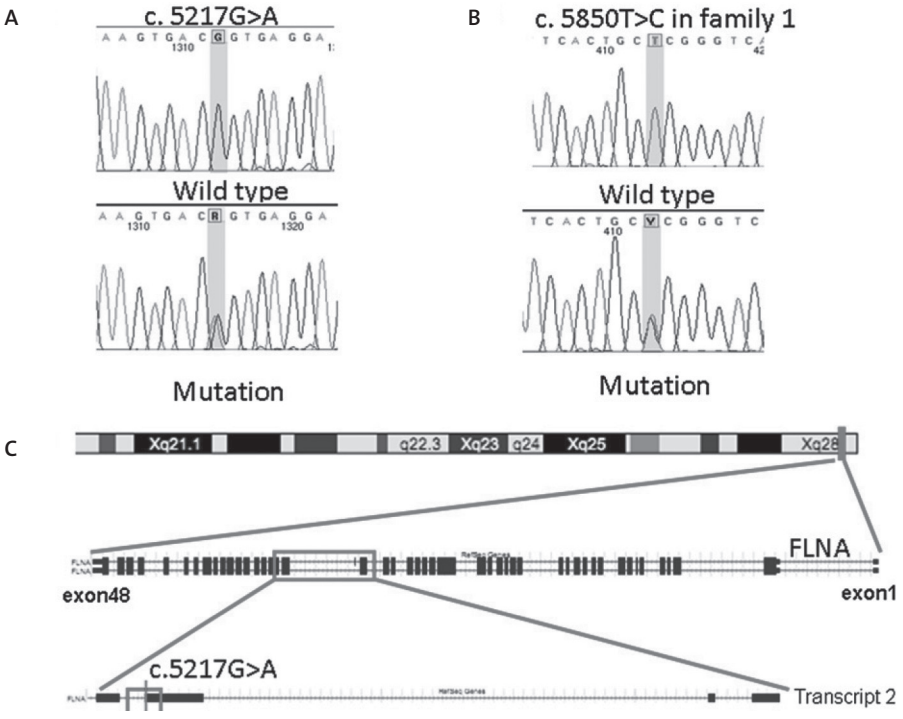
All of the HGVS numbers were generated with the use of the longest isoforms if multiple transcripts existed.

Notably, families 1 and 2 had two distinct variants adjacent to the c.5217G>A mutation, making a close and common ancestor highly unlikely. Finally, upon analysis of a third TOD family and three unrelated sporadic cases, we identified exactly the same c.5217G>A variant again in all patients, but not unaffected family members (1I:2, 3I:1 and 3II:3).

Variant c.5217G>A affects the last nucleotide of exon 31 of the *FLNA* gene (Figure 2C). At protein level, it is not predicted to change the encoded amino acid but as the last nucleotide of an exon, it may affect splicing[10-12]. RNA was isolated from cultured fibroblasts of arm skin from 1III:6, removed during a recent orthopaedic procedure under general anaesthesia with informed consent. Cells were cultured in standard medium for human fibroblasts (DMEM with 10%FBS, 1%Pen/Strep, 1%glucose, 1%glutamax) with 5%CO_2 in 37°C. RNA was extracted by RNeasy Mini Kit (QIAGEN). cDNA was synthesized from 500ng of total RNA by RevertAid RNaseH- M-MuLV reverse transcriptase in a total volume of 20 µL according to the protocol provided by the supplier (MBI-Fermentas). Target regions were amplified by RT-PCR using the primers listed in Supplementary data S2. The products were evaluated using Bioanalyzer 2100 DNA chip 1000 (Agilent) according to the manufacturers instruction. RNA from patient fibroblasts showed only normal transcripts, both of transcripts 1 (NM_01456) and 2 (NM_001110556), differing by insertion of the 24 bp exon 30 in transcript 2. Although transcript 1 has been reported as the predominant transcript in controls[13], we detected about equal expression levels in controls (Figure 3B lane 2-4, 8) and higher expression of transcript 2 in patient fibroblasts (Figure 3B lane 1). Both bands were isolated from the agarose gel by Qiaquick gel extraction kit (QIAGEN) and analysed by Sanger sequencing. Interestingly, we detected no expression of the mutant allele. This could be due to nonsense mediated decay (NMD)[14] and/or skewed X chromosome inactivation (XCI). To test the first possibility, the fibroblasts were treated with cycloheximide[15] for 4.5 hours followed by RNA analysis using the same procedures as for RNA from untreated cells. The mutant allele was still absent in RNA from cycloheximide treated cells. X-inactivation was analysed using the Androgen Receptor (AR) Assay[16]. The assay showed random X-inactivation in 1I:2 versus 100% X-inactivation of the mutant chromosome in patient 1II:4 (patient 1III:6 was uninformative), indicating that the mutant allele was inactivated.

Fifteen years ago, at the age of 1 year, fibroma tissue was surgically removed and stored in liquid nitrogen from the 5th digits of both hands and the 5th toe of the left foot of patient 1III:6. We cultured these cells, and analysed RNA. In the fibroma cells we observed 2 sets of 2 bands (Figure 3B lane 5-7), indicating altered splicing. One set had the same length as observed in normal fibroblasts (Figure 3A transcript 1 and 2), the other set was shorter (Figure 3A transcript 3 and 4, faint from RNA of tumor in left fifth finger and toe, Figure 3B lane 6 and 7 respectively). Note that the fibroma always contains a mixture of tumor and normal stroma cells.

Figure 2: Genomic Structure and mutation analysis of *FLNA*. (A) c.5217G>A was confirmed by Sanger sequencing in all the patients. The unaffected family members and controls carry the homozygous normal allele. (B) shows the sequence of c.5850T>C in family 1. (C) FLNA is located in Xq28, the target region of linkage analysis. C.5217G>A alters the last nucleotide of exon 31 of FLNA.



Sequence analysis showed a deletion removing the last 48 nucleotides of exon 31 (Figure 3C), resulting in a deletion of 16 amino acids.

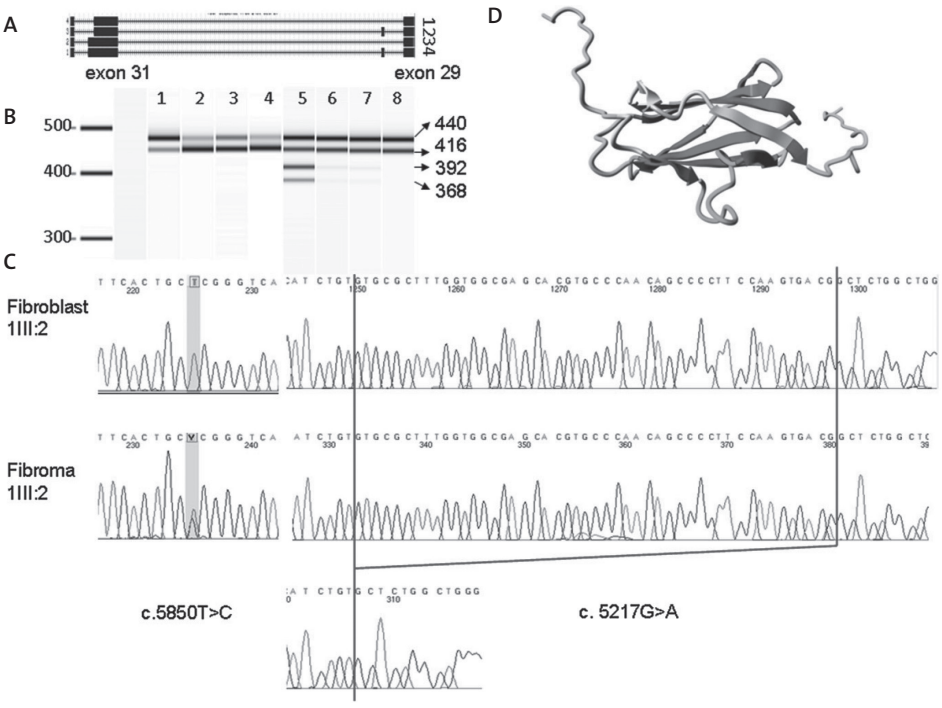
To facilitate clinical diagnostics of *FLNA* gene mutations we have established a web-based *FLNA* gene variant database using the LOVD software[17]. In this publicly available database we have collected all variants reported in literature thus far (83 in total, see *FLNA* mutation database), incl. the new variants described here. The c.5217G>A variant detected in TOD patients has not been described before; it is neither listed in dbSNP nor in the pilot study 1 of the 1000 Genomes Project. Finally, >400 chromosomes have been sequenced and the mutant allele was not found (data not shown).

Mutations in *FLNA* have been reported to cause a wide range of developmental malformations in brain, bone, limbs, heart[18] and other organs[19] in human[9], including periventricular heterotopia (PVNH) (OMIM 300049)[20-24] and OPD spectrum disorders[25] which include otopalatodigital syndrome type 1 (OMIM 311300)[26-28] and 2 (OMIM 304120)[26; 29], frontometaphyseal dysplasia (OMIM 305620) [26; 30-31] and Melnick-Needles syndrome (OMIM 309350)[26-27]. Although each of OPD spectrum disorders are characterized by specific clinical symptoms, there clearly is a clinical overlap with TOD including a generalized bone dysplasia including craniofacial anomalies, digits and long bones[9; 32]. Interestingly, the most conspicuous symptoms of TOD patients are skeletal dysplasia of the limbs and recurrent digital fibroma, suggesting a significant role of the *FLNA* mutation in the TOD phenotype.

The *FLNA* gene encodes a cytoskeletal protein, filamin A, which crosslinks actin filaments into an orthogonal network and links these to the cell membrane. Within the cytoskeleton filamin A also mediates functions relating to cell signaling, transcription and development[33]. Filamin A consists of two calponin homology sequences (CH1 and CH2) at the N-terminal and connects with 24 immunoglobin like filamin repeats, divided by 2 hinges between repeat 15 and 16, repeat 23 and 24. To check the stability of filamin A in patient cells, protein was extracted from both fibroblast and fibroma cells. Western blot was performed using mouse anti-human filamin A monoclonal antibody, MAB1680, from MILLIPORE. No difference was observed on molecular weight or quantity. Likely the difference of 18 amino acids was too small to distinguish by western blot. The c.5217G>A mutation is located in a highly conserved position at the DNA level across a wide range of vertebrate and invertebrate species except rodent, found in all ten affected patients from six different unrelated families. In addition, the mutation introduced abnormal splicing in fibroma cells. At protein level, c.5217G>A encodes the second last amino acid of repeat 15, which is immediately adjacent to hinge 1. Recent studies demonstrated repeat 9-15 contain an F-actin binding domain necessary for high avidity F-actin binding[34]. Hinge 1 plays an important part in maintaining the viscoelastic properties of actin networks[35]. Moreover, this region interacts with many binding partners like TRAF1, TRAF2[36], CaR extracellular Ca^{2+} receptor[37], and FAP52[38]. As no crystal structure has yet been described for this region, the crystal structure of repeat 15 in filamin B(PDB file 2dmb), that shows the highest identity (58%) with this region of interest, was used as a template to build a 3D model (Figure 3D). The model was built using the WHAT IF &Yasara twinset[39]. Repeat 15 consists of 2 beta-sheets. The in-frame deletion causes the removal of the top of a beta-strand in the middle of one beta-sheet, and of two beta-strands at the side of that sheet (grey part of Figure 3D). These residues are likely to form some kind of beta-strand like structure, and substantially alter the structure of the highly conserved tertiary structure of filamin

repeat 15. Furthermore, this structure will affect the residues following the beta-sheet and linking repeat 15 to hinge1. Although there is no way to predict what will happen to those linking residues, we believe it will affect the overall conformation of the protein, and likely influence the interaction between filamin A and other molecules.

Figure 3: Detection of alternative splicing and 3D protein model. (A) Diagram of four FLNA transcripts in fibroma cells: transcript 1 and 2, which carry the 48bp deletion at the end of exon 31, as well as the normal transcripts 3 and 4. (B) RT-PCR result from Agilent 2100 Bioanalyzer. Lane 1 is the product of the fibroblasts of 1III:6, which has a predominant longer isoform. Lane 2-4 and 8 are four control human fibroblasts. Lanes 5-7 show RT-PCR products that were obtained from fibroma cells of 1III:6, the normal bands from two FLNA isoforms, and two extra shorter bands, which are faint in lane 6 (left fifth finger)and lane 7 (fifth toe of the left foot) , whereas lane 5 (right fifth finger)shows four dark bands. (C) Sanger sequencing results of c.5858T>C and c.5217G>A in fibroblast and fibroma cells of 1III:6. (D) The 3D model of FLNA domain 15. The deleted 16 amino acids are marked in gray. Beta-strands are marked in red. Green represents a turn. Yellow indicates a 3/10 helix. Random coils are colored in cyan.



The precise mechanism of TOD remains unclear. However, like other X-linked diseases, X chromosome inactivation (XCI) might be a key component of how the disease develops. The developmental role of *FLNA* is borne out by the presence of the skeletal and skin malformations at birth. Multiple fibroma on digits start occurring in the first years. Fibromas spontaneously stop by the age of five. Skewed XCI is known to vary in different tissues and to correlate with age under the pressure of secondary selection[40]. Several mechanisms may contribute to the skewing, including stochastic effects, a selective growth advantage of cell that carries either the mutated or the normal allele (secondary cell selection) and genetic processes yielding preferential inactivation of specific alleles. Primarily the X inactivation choice is random, but during cell proliferation either in all cells or in a tissue specific manner, cells that carry an activate mutated allele may have a significant disadvantage, are gradually lost or selected against and are thus less represented in the adult female[41]. Disorders caused by defects in *FLNA* gene often show skewed XCI pattern[26], suggesting that cells need normal filamin to survive. Several studies in TOD families, showed patients had skewed X-inactivation, while unaffected individuals had random inactivation[1, 6]. We examined the X-inactivation pattern in family 1 (1I:2, 1II:4 and 1III:6) and 3 (3I:2, 3II:3, 3II:4 and 3II:5, Figure 1A) by AR assay. Apart from the uninformative 1III:6, all the patients, 1II:4, 3I:2, 3II:4 and 3II:5 showed extremely skewed X-inactivation (0/100%), while the normal family member 1I:2 showed random X-inactivation (30/70%) together with 3II:3 (50/50%). Since there was no mutant allele detectable in the RNA of normal fibroblast, we deduced 1III:6 also had 100% skewed XCI with the preferential inactivation of the mutant allele. Baroncini tested the XCI of 2II:4 and 2III:5, and both showed 100% skewing[6]. 2II:4 was interpreted by the authors as unaffected. However, we assume 2II:4 is a carrier of TOD, as she only has mild manifestations (multiple frenula in the mouth). She probably has skewed X-inactivation at a very early stage. Overall as well as local X inactivation patterns may influence the severity of the phenotype of carrier females and are also associated with selective female survival in male lethal X-linked dominant disorders.

Taken together, these data suggest that TOD unique variant c.5217G>A (p.Val1724_Thr1739del) in the *FLNA* gene. The variant is not found in other databases, has not been seen in other patients with pathogenic *FLNA* variants, segregates with the disease, and locates in Xq28 where the potential mutated gene causing this disorder was mapped previously. The mutation was found in six unrelated families. It will affect splicing, and causes a deletion of 16 amino acids in protein level. The missing region in the filamin A protein is hypothesized to affect or prevent the interaction of filamin A with other proteins.

Acknowledgements

We would like to thank the patients and their family members for their willingness to join the project, CSC scholarship for supporting Yu Sun's studies in the Netherlands, Filip Kluin for sending and Hans Morreau for isolating DNA from paraffin embedded tissue, Tobias Messemaker for helping us with western blot, the Leiden Genome Technology Center (LGTC), and Laboratory for Diagnostic Genome Analysis (LDGA) for help with sequencing, DNA extraction, X-inactivation detection. X-exome capture was implemented in collaboration with ServiceXS (Leiden, www.servicexs.com). The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreements 223026 (NMD-chip), 223143 (TechGene) and 200754 (Gen2Phen).

Web Resource

Accession numbers and URLs for data presented herein are as follows:
FLNA gene variant database, http://www.lovd.nl/FLNA
SureSelect manual, http://www.genomics.agilent.com/files/Manual/G3360-90020_SureSelect_Indexing_1.0.pdf
UCSC Genome Browser, http://genome.ucsc.edu/
Online Mendelian Inheritance in Man(OMIM), http://www.ncbi.nlm.nih.gov/entrez/Omim/
RefSeq, http://www.ncbi.nlm.nih.gov/Refseq/, for human FLNA[accession number NM_001110556.1], for human chromosome X [accesssion number NC_000023.9]
dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/
1000 genome project, http://www.1000genomes.org/page.php
Bowtie, http://bowtie-bio.sourceforge.net/index.shtml
Human Splicing Finder, http://www.umd.be/HSF
YASARA, http://www.yasara.org/

1. Bacino, C.A., Stockton, D.W., Sierra, R.A., Heilstedt, H.A., Lewandowski, R., and Van den Veyver, I.B. (2000). Terminal osseous dysplasia and pigmentary defects: clinical characterization of a novel male lethal X-linked syndrome. Am J Med Genet 94, 102-112.

2. Zhang, W., Amir, R., Stockton, D.W., Van Den Veyver, I.B., Bacino, C.A., and Zoghbi, H.Y. (2000). Terminal osseous dysplasia with pigmentary defects maps to human chromosome Xq27.3-xqter. Am J Hum Genet 66, 1461-1464.

3. Horii, E., Sugiura, Y., and Nakamura, R. (1998). A syndrome of digital fibromas, facial pigmentary dysplasia, and metacarpal and metatarsal disorganization. Am J Med Genet 80, 1-5.

4. Drut, R., Pedemonte, L., and Rositto, A. (2005). Noninclusion-body infantile digital fibromatosis: a lesion heralding terminal osseous dysplasia and pigmentary defects syndrome. Int J Surg Pathol 13, 181-184.

5. Breuning, M.H., Oranje, A.P., Langemeijer, R.A., Hovius, S.E., Diepstraten, A.F., den Hollander, J.C., Baumgartner, N., Dwek, J.R., Sommer, A., and Toriello, H. (2000). Recurrent digital fibroma, focal dermal hypoplasia, and limb malformations. Am J Med Genet 94, 91-101.

6. Baroncini, A., Castelluccio, P., Morleo, M., Soli, F., and Franco, B. (2007). Terminal osseous dysplasia with pigmentary defects: clinical description of a new family. Am J Med Genet A 143, 51-57.

7. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25.

8. Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., and Beroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37, e67.

9. Robertson, S.P. (2005). Filamin A: phenotypic diversity. Curr Opin Genet Dev 15, 301-307.

10. Agarwal, N., Kutlar, F., Mojica-Henshaw, M.P., Ou, C.N., Gaikwad, A., Reading, N.S., Bailey, L., Kutlar, A., and Prchal, J.T. (2007). Missense mutation of the last nucleotide of exon 1 (G->C) of beta globin gene not only leads to undetectable mutant peptide and transcript but also interferes with the expression of wild allele. Haematologica 92, 1715-1716.

11. Yamada, K., Fukao, T., Zhang, G., Sakurai, S., Ruiter, J.P., Wanders, R.J., and Kondo, N. (2007). Single-base substitution at the last nucleotide of exon 6 (c.671G>A), resulting in the skipping of exon 6, and exons 6 and 7 in human succinyl-CoA:3-ketoacid CoA transferase (SCOT) gene. Mol Genet Metab 90, 291-297.

12. Kuivaniemi, H., Tromp, G., Bergfeld, W.F., Kay, M., and Helm, T.N. (1995). Ehlers-Danlos syndrome type IV: a single base substitution of the last nucleotide of exon 34 in COL3A1 leads to exon skipping. J Invest Dermatol 105, 352-356.

13. Maestrini, E., Patrosso, C., Mancini, M., Rivella, S., Rocchi, M., Repetto, M., Villa, A., Frattini, A., Zoppe, M., Vezzoni, P., et al. (1993). Mapping of two genes encoding isoforms of the actin binding protein ABP-280, a dystrophin like protein, to Xq28 and to chromosome 7. Hum Mol Genet 2, 761-766.

14. Holbrook, J.A., Neu-Yilik, G., Hentze, M.W., and Kulozik, A.E. (2004). Nonsense-mediated decay approaches the clinic. Nat Genet 36, 801-808.

15. Kim, C.E., Gallagher, P.M., Guttormsen, A.B., Refsum, H., Ueland, P.M., Ose, L., Folling, I., Whitehead, A.S., Tsai, M.Y., and Kruger, W.D. (1997). Functional modeling of vitamin responsiveness in yeast: a common pyridoxine-responsive cystathionine beta-synthase mutation in homocystinuria. Hum Mol Genet 6, 2213-2221.

16. Kubota, T., Nonoyama, S., Tonoki, H., Masuno, M., Imaizumi, K., Kojima, M., Wakui, K., Shimadzu, M., and Fukushima, Y. (1999). A new assay for the analysis of X-chromosome inactivation based on methylation-specific PCR. Hum Genet 104, 49-55.

17. Fokkema, I.F., den Dunnen, J.T., and Taschner, P.E. (2005). LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26, 63-68.

18. Kyndt, F., Gueffet, J.P., Probst, V., Jaafar, P., Legendre, A., Le Bouffant, F., Toquet, C., Roy, E., McGregor, L., Lynch, S.A., et al. (2007). Mutations in the gene encoding filamin A as a cause for familial cardiac valvular dystrophy. Circulation 115, 40-49.

19. Gargiulo, A., Auricchio, R., Barone, M.V., Cotugno, G., Reardon, W., Milla, P.J., Ballabio, A., Ciccodicola, A., and Auricchio, A. (2007). Filamin A is mutated in X-linked chronic idiopathic intestinal pseudo-obstruction with central nervous system involvement. Am J Hum Genet 80, 751-758.

20. Fox, J.W., Lamperti, E.D., Eksioglu, Y.Z., Hong, S.E., Feng, Y., Graham, D.A., Scheffer, I.E., Dobyns, W.B., Hirsch, B.A., Radtke, R.A., et al. (1998). Mutations in filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. Neuron 21, 1315-1325.

21. Sheen, V.L., Dixon, P.H., Fox, J.W., Hong, S.E., Kinton, L., Sisodiya, S.M., Duncan, J.S., Dubeau, F., Scheffer, I.E., Schachter, S.C., et al. (2001). Mutations in the X-linked filamin 1 gene cause periventricular nodular heterotopia in males as well as in females. Hum Mol Genet 10, 1775-1783.

22. Moro, F., Carrozzo, R., Veggiotti, P., Tortorella, G., Toniolo, D., Volzone, A., and Guerrini, R. (2002). Familial periventricular heterotopia: missense and distal truncating mutations of the FLN1 gene. Neurology 58, 916-921.

23. Zenker, M., Rauch, A., Winterpacht, A., Tagariello, A., Kraus, C., Rupprecht, T., Sticht, H., and Reis, A. (2004). A dual phenotype of periventricular nodular heterotopia and frontometaphyseal dysplasia in one patient caused by a single FLNA mutation leading to two functionally different aberrant transcripts. Am J Hum Genet 74, 731-737.

24. Sheen, V.L., Jansen, A., Chen, M.H., Parrini, E., Morgan, T., Ravenscroft, R., Ganesh, V., Underwood, T., Wiley, J., Leventer, R., et al. (2005). Filamin A mutations cause periventricular heterotopia with Ehlers-Danlos syndrome. Neurology 64, 254-262.

25. Robertson, S.P. (2007). Otopalatodigital syndrome spectrum disorders: otopalatodigital syndrome types 1 and 2, frontometaphyseal dysplasia and Melnick-Needles syndrome. Eur J Hum Genet 15, 3-9.

26. Robertson, S.P., Twigg, S.R., Sutherland-Smith, A.J., Biancalana, V., Gorlin, R.J., Horn, D., Kenwrick, S.J., Kim, C.A., Morava, E., Newbury-Ecob, R., et al. (2003). Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. Nat Genet 33, 487-491.

27. Robertson, S.P., Thompson, S., Morgan, T., Holder-Espinasse, M., Martinot-Duquenoy, V., Wilkie, A.O., and Manouvrier-Hanu, S. (2006). Postzygotic mutation and germline mosaicism in the otopalatodigital syndrome spectrum disorders. Eur J Hum Genet 14, 549-554.

28. Hidalgo-Bravo, A., Pompa-Mera, E.N., Kofman-Alfaro, S., Gonzalez-Bonilla, C.R., and Zenteno, J.C. (2005). A novel filamin A D203Y mutation in a female patient with otopalatodigital type 1 syndrome and extremely skewed X chromosome inactivation. Am J Med Genet A 136, 190-193.

29. Marino-Enriquez, A., Lapunzina, P., Robertson, S.P., and Rodriguez, J.I. (2007). Otopalatodigital syndrome type 2 in two siblings with a novel filamin A 629G>T mutation: clinical, pathological, and molecular findings. Am J Med Genet A 143A, 1120-1125.

30. Zenker, M., Nahrlich, L., Sticht, H., Reis, A., and Horn, D. (2006). Genotype-epigenotype-phenotype correlations in females with frontometaphyseal dysplasia. Am J Med Genet A 140, 1069-1073.

31. Giuliano, F., Collignon, P., Paquis-Flucklinger, V., Bardot, J., and Philip, N. (2005). A new three-generational family with frontometaphyseal dysplasia, male-to-female transmission, and a previously reported FLNA mutation. Am J Med Genet A 132A, 222.

32. Robertson, S.P. (2004). Molecular pathology of filamin A: diverse phenotypes, many functions. Clin Dysmorphol 13, 123-131.

33. Zhou, A.X., Hartwig, J.H., and Akyurek, L.M. (2010). Filamins in cell signaling, transcription and organ development. Trends Cell Biol 20, 113-123.

34. Nakamura, F., Osborn, T.M., Hartemink, C.A., Hartwig, J.H., and Stossel, T.P. (2007). Structural basis of filamin A functions. J Cell Biol 179, 1011-1025.

35. Gardel, M.L., Nakamura, F., Hartwig, J.H., Crocker, J.C., Stossel, T.P., and Weitz, D.A. (2006). Prestressed F-actin networks cross-linked by hinged filamins replicate mechanical properties of cells. Proc Natl Acad Sci U S A 103, 1762-1767.

36. Arron, J.R., Pewzner-Jung, Y., Walsh, M.C., Kobayashi, T., and Choi, Y. (2002). Regulation of the subcellular localization of tumor necrosis factor receptor-associated factor (TRAF)2 by TRAF1 reveals mechanisms of TRAF2 signaling. J Exp Med 196, 923-934.

37. Awata, H., Huang, C., Handlogten, M.E., and Miller, R.T. (2001). Interaction of the calcium-sensing receptor and filamin, a potential scaffolding protein. J Biol Chem 276, 34871-34879.

38. Nikki, M., Merilainen, J., and Lehto, V.P. (2002). FAP52 regulates actin organization via binding to filamin. J Biol Chem 277, 11432-11440.

39. Krieger, E., Koraimann, G., and Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins 47, 393-402.

40. Sharp, A., Robinson, D., and Jacobs, P. (2000). Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Hum Genet 107, 343-349.

41. Orstavik, K.H. (2009). X chromosome inactivation in clinical practice. Hum Genet 126, 363-373.

Supplementary information

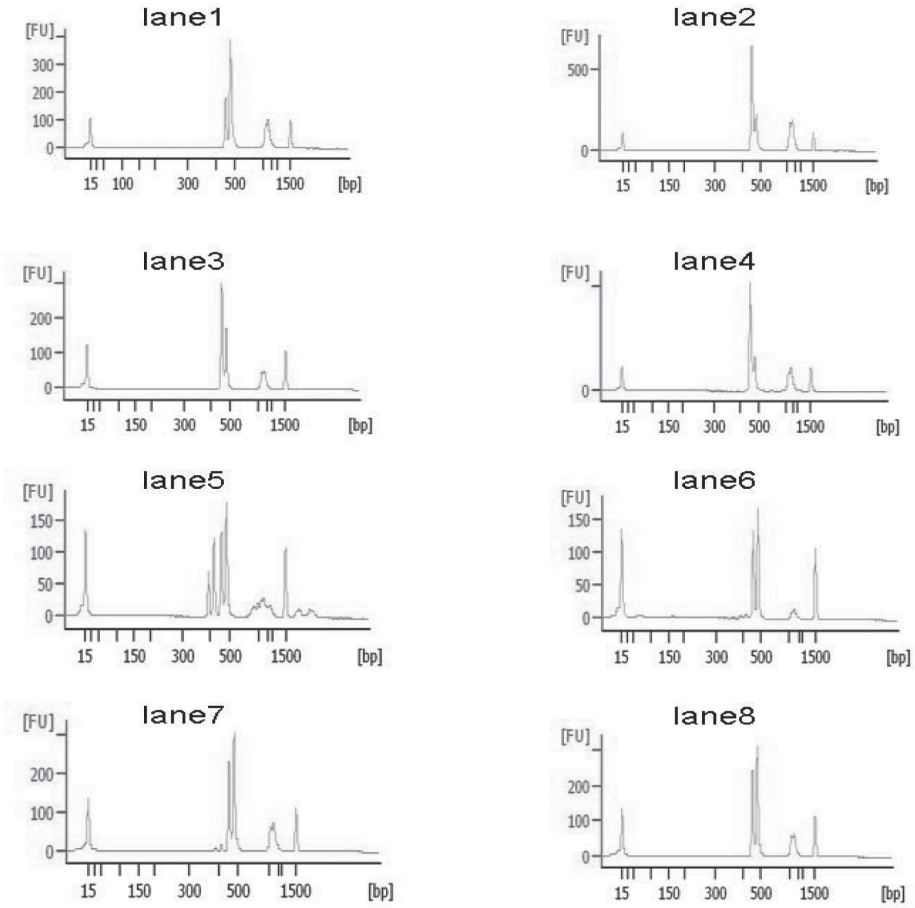
Table S1: The overview of the data generated by GAII.

	1II:4	2III:5
Run	Paired-end	Paired-end
Total reads	36,010,190	28,960,586
Read F	18,005,095	14,480,293
Read R	18,005,095	14,480,293
Aligned reads	33,054,043	18.948,705
Aligned in pair	30,018,244	11,012,526
Read length	51	51

Table S2: F LNA primer list.

location		Primer sequences (5'-3')	Size (bp)
Exon 31-32	DNA (blood, buccal cells)	F:GTCATCTGTGTGCGCTTTGG	222
		R:AGCTGCTGAGACCGTAGAGG	
Exon 31	DNA (paraffin embedded tissue)	F:GGGCAAATACGTCATCTGTGT	104
		R:agacacccctgctgacctac	
Exon 29-32	RNA	F:CCTGGGCGTAGGTGTACTGT	416 (short isoform)
		R:CATCAAGTACGGTGGTGACG	440 (long isoform)
Exon 35-37	DNA, RNA	F:ACATACGCATGGAGTCGTCA	577 (DNA)
		R:TCAACTGTGGCCATGTCACT	294 (RNA)

Figure S3: The 2100 bioanalyzer traces of RT-PCR on c.5217G>A from lane 1 to 8. The peak around 15 bp is the lower ladder and the signal round 1500 bp is the upper ladder.




7

Exome sequencing identifies a branch point variant in Aarskog-Scott syndrome

Emmelien Aten, * Y. Sun, * R. Almomani, G.W.E Santen, T. Messemaker, S.M. Maas, M.H Breuning, J.T. den Dunnen

**These authors contributed equally to this work*

Human Mutation. 2012. In Press.

Abstract

Aarskog-Scott syndrome (ASS) is a rare disorder with characteristic facial, skeletal, and genital abnormalities. Mutations in the *FGD1* gene (Xp11.21) are responsible for ASS. However, mutation detection rates are low. Here, we report a family with ASS where conventional Sanger sequencing failed to detect a pathogenic change in *FGD1*. To identify the causative gene we performed whole-exome sequencing in two patients. An initial analysis did not reveal a likely candidate gene. After relaxing our filtering criteria, accepting larger intronic segments, we unexpectedly identified a branch point (BP) variant in *FGD1*. Analysis of patient-derived RNA showed complete skipping of exon 13, leading to premature translation termination. The BP variant detected is one of very few reported so far proven to affect splicing. Our results show that, besides digging deeper to reveal non-obvious variants, isolation and analysis of RNA provides a valuable but under-appreciated tool to resolve cases with unknown genetic defects.

Key words
FGD1 protein, Exome sequencing, Aarskog syndrome, RNA splice sites, Branch point mutations

Introduction

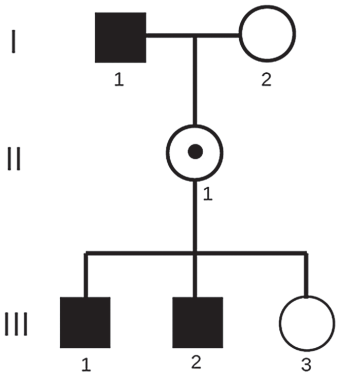
The large scale sequencing of thousands of exomes and genomes from disease cohorts using next generation sequencing (NGS) is expected to uncover causal gene variants in many thus far unresolved cases. Currently, high-throughput clinical sequencing of the entire genome is still too expensive but selective enrichment of genomic regions of interest, such as the exome, provides a cost-effective and scalable approach. The normal strategy is to compare variants with clear functional consequences (nonsense, frame shift, splice site, conserved missense) from several patients/families with identical phenotypes and searching for a shared defective gene or from parent-child trios searching for *de novo* variants in the child. The approach has revealed the cause of a range of rare Mendelian disorders. However, exome sequencing comes with several disadvantages. It is labor intensive (costly), it misses variants outside the targeted (protein coding exons) regions and it generates sequence data with a large variability in coverage across the regions sequenced, missing 5-15% of the targeted area. Furthermore, to cope with the sheer number of variants encountered, choices have to be made regarding the variants that deserve follow-up work (truncating/missense variants, intronic variants, splice site variants, etc). When the pathogenic consequence of a variant is unclear, the effect has to be studied in detail at other levels (e.g. RNA, protein, in vitro functional assay), in other tissues or in additional patients/families making the final proof of causality a tremendous effort. In many cases this may not be possible at all, e.g. since the function of the gene identified is not known or because no other patients are available.
Besides the first nucleotides, intronic regions are usually not studied. This has two main reasons: first introns contain many variants yielding too many candidates to follow-up. Second, RNA is often not available making it impossible to prove predicted effects on RNA splicing.

Aarskog-Scott syndrome (ASS or faciogenital dysplasia, OMIM #305400) is a clinically and genetically heterogeneous disorder characterized by proportionate short stature, short limbs (usually rhizomelic), broad hands and feet, genital hypoplasia (shawl scrotum) and facial dysmorphisms. Mental retardation has been described but is not a common finding (Orrico et al., 2005; Kaname et al., 2006; Shalev et al., 2006). Both autosomal dominant and recessive inheritance patterns have been demonstrated (Teebi et al., 1988; van de Vooren et al., 1983) but currently only one gene has been directly linked to ASS, *FGD1* at Xp11.21. Both substitutions and deletions have been described in *FGD1* in X-linked cases (Shalev et al., 2006; Bedoyan et al., 2009).

The *FGD1* gene (faciogenital dysplasia 1) is ubiquitously expressed and composed of 18 exons encoding a 961 amino acid guanine nucleotide exchange factor (GEF), binding

specifically to the Rho protein Cdc42. The Cdc42 pathway is involved in regulating cell growth and differentiation and plays a role in skeletal development (Pasteris et al., 1997; Hou et al., 2003). The protein consists of several important segments such as a proline-rich SH3-binding motif, a RhoGEF homology domain (DH), two pleckstrin-homology domains (PH) and a FYVE-zinc finger domain (ZF). In ASS, variants have been reported throughout the gene, both missense and protein truncating but there is no clear correlation between the nature and location of variants and disease severity (Orrico et al., 2007). Most variants are private, with the exception of three (c.529dupC, c.1966C>T and several affecting amino acid Arg443; www.LOVD.nl/FGD1) (Orrico et al., 2010; Orrico et al., 2004). *FGD1* variants are found in only ~ 20% of Aarskog-Scott syndrome patients (Orrico et al., 2004). This can be attributed to inaccurate clinical diagnosis, the existence of overlapping syndromes (notably Noonan, LEOPARD, Teebi hypertelorism and Robinow syndrome) or genetic heterogeneity (Bottani et al., 2007). In diagnostics, *FGD1* analysis of the entire coding region is performed using DHPLC and/or Sanger sequencing and deletion/duplication analysis using MLPA.
Two siblings (III-1 and III-2) from a family of Dutch origin were referred with a clinical diagnosis of ASS. The phenotype is summarized in Figure 1 and Table 1 and is illustrated in Figure S1. Conventional diagnostic Sanger sequencing revealed no changes in the *FGD1* gene (Greenwood Genetic Center, USA).

Figure 1: Pedigree of the Dutch Aarskog family.



SNP-array analysis of II-2, III-1 and III-2 excluded the presence of large causative deletions/duplications in the genome, including the region containing *FGD1*. To resolve the genetic basis of ASS in this family, we performed exome sequencing of the two siblings (III-1 and III-2). Illumina GAIIX generated 51 nt, paired end reads. A coverage plot can be found in

Table 1: Phenotype overview of the Dutch ASS family.
na = not applicable

	III-1	III-2	II-1	I-1
Craniofacial				
Widow's peak	-	+	+	+
Cow's lick	+	+	+	+
Hypertelorisme	+	+	+	+
Ptosis	-	-	-	-
Downward slant palpebral fissures	-	-	-	-
Short nose	+	+	+	-
Wide philtrum	+	+	-	+
Cleft lip and palate	+	-	-	-
Underdeveloped maxillae	-	-	-	-
Crease below lower lip	-	-	-	-
Abnormal auricles	+	+	-	-
Skeletal				
Short stature	+	+	+	+
Short broad hands	-	-	-	-
Clinodactyly fifth finger	-	-	-	-
Midl interdigital webbing	-	-	-	-
Joint laxity	+	+	+	-
Contractures of fingers	+	-	-	+
Broad feet	+	-	-	-
Genital				
Shawl scrotum	+	+	na	-
cryptorchidism	+	+	na	-

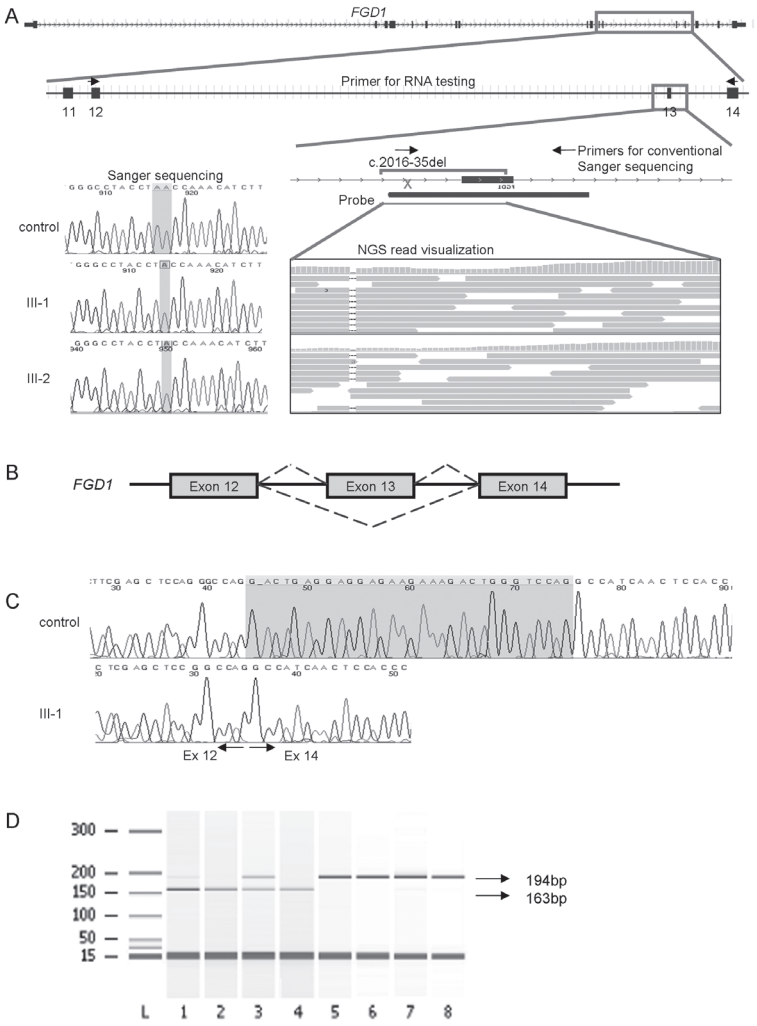
Figure S2. Standard filtering parameters were set for variants with a minimal depth of 8 and initially selecting nonsense, frame shift, splice-site and conserved missense variants not present in variant databases (dbSNP, 1000Genomes, in-house exome variants). Prioritization of variants with functional consequences in targeted exonic regions (Table S1) listed a frameshift insertion in *OXGR1* and a missense variant in *C9orf72*. Additional Sanger Sequencing showed that the *OXGR1* variant did not cosegregate with the disorder in the family. A (GGGGCC)n expansion between exon 1a and 1b in *C9orf72* has recently been reported to cause frontotemporal dementia and/or amyotrophic lateral sclerosis (FTLD/ALS). However, the phenotype of FTLD/ALS is quite different from ASS. Therefore, both variants were excluded as candidate genes for AAS. Criteria for variant analysis were then relieved to include intronic variants up to 50 bp from the splice sites, yielding three candidate variants (Table S2). Unexpectedly, a single nucleotide deletion (c.2016-35del) in the *FGD1* gene emerged that potentially affects the branch point of the splice acceptor site of exon 13. The variant was not present in dbSNP131 or the 1000 Genomes project (May 2011) and has not been reported in the *FGD1* gene variant database (www.LOVD.nl/FGD1), which we established in the course of this study. Using Sanger sequencing we confirmed the variant in the two siblings. In addition we detected perfect co-segregation with ASS in the family (Figure2A).

In Silico analysis predicted the variant to disrupt normal splicing, skipping exon 13, disrupting the reading frame and to lead to a premature stop codon (Desmet et al., 2009) (Figure 2B). This would result in a truncated protein of 677 amino acids.
The effect on splicing was analyzed using RNA derived from cultured lymphocytes. RT-PCR analysis of individuals I-1 (not shown), III-1 (Figure 2C) and III-2 (not shown) showed near complete skipping of exon 13. Semi-quantification of skipped products (Agilent 2100 Bioanalyzer) showed complete skipping in I-1 and III-2, 94% in III-1 and 56% in carrier female II-1 (Figure 2D, Figure S3). We did not observe nonsense-mediated mRNA decay (NMD). RT-qPCR using primers covering exon-exon junctions confirmed these results. Compared to three controls, all affected and carrier individuals had a significantly reduced *FGD1* expression (data not shown).

To explore the general application of exome sequencing for intronic variants, we investigated our datasets to give an estimate of the chance to find an intronic variant by exome sequencing. Although exome sequencing is not designed to detect intronic variants, one can detect them. In our experiment, coverage was sufficient to call variants for ~ 80% of the intronic sequences up to position -35 bp (Figure S4A).We projected the number of variants detected when sequences up to 100 bp intronic would have been targeted by capture (read depth >=8). If probe extension is combined with filtering on conservation scores and information from variant databases (dbSNP, 1000 genomes project, inhouse exome variants), this would increase the number of candidate variants only marginally, making the effort sensible. Figure S4B illustrates this by showing the additional number of candidate variants remaining after standard filtering upon inclusion of an increasing part of the intron (various distances to the splice site) in our dataset.

Exome sequencing has solved several Mendelian diseases successfully. The obvious focus is to look for nonsense, frameshift, splice-site and missense variants and analyze the pathogenic consequence of candidate variants. If the variant cannot be classified, the effect has to be studied in detail at other levels (e.g. RNA, protein, in vitro functional assays). Since RNA is often not available for analysis and since only short intronic segments are analyzed, the true fraction of variants disrupting splicing remains largely unknown but is estimated between 15-50% (Ward and Cooper, 2010). Every intron contains core sequence elements that are essential for splicing: a 5' splice donor site (SD), 3' splice acceptor site (SA) and a branch point site (BP). In experimentally verified distant BPs the region between the BP and 3' SD is devoid of AG dinucleotides, the so called 'AG exclusion zone' (AGEZ) (Gooding et al., 2006) . Since BPs reside deeper into the intron, BP variants proven to affect splicing are quite rare. Observed consequences of variants in BP sites include exon skipping, intron retention and partial intron retention. Several variants have been published that result in genetic diseases. Substitutions resulting in disease have so far only been described in

Figure 2: Genomic analysis of the mutation. A) Analysis result of GAII-X sequencing, displayed in IGV browser. III-1 and III-2 show a single nucleotide deletion, c.2016-35del in intron 12 of *FGD1*. On the left side, DNA Sanger sequencing results confirming the variant in the Dutch Aarskog family (III-1 and III-2 versus a WT control). B) A diagram of the normal and abnormal splicing in exon 12, 13 and 14. C) Sanger sequencing of the RT-PCR shows a skip of exon 13 in individuals III-1 versus a WT control, confirming the branch point variant. The skip of exon 13 leads to a 31 bp shorter mRNA. D) The 2100 Bioanalyzer results show semi-quantification of skipped products. III-1(lane 1) shows a 94% skip, containing two products, a large skipped (163 bp) and a small WT band (194 bp). III-2(lane 2) and I-1(lane4) show a 100% skip. II-1 (mother, lane 3) has a 50-50 distribution of WT and skipped products. Controls(lane 5~8) only carry the WT band.



the two most conserved BPs positions, the BP adenosine itself and the -2 position (relative to the BP)(Kralovicova et al., 2006). To our knowledge, a deletion of the BP has not been described. Useful approaches for follow-up studies of BP variants have been communicated (reviewed by Pagani and Baralle (Pagani and Baralle, 2004)). *In silico prediction* of BP variants, combined with segregation studies and RNA analysis is very useful, especially for clear disease-associated genes. As an alternative, or in conjunction with RNA analysis, *in vitro* splicing reporter assays using minigene reporter constructs and lariat PCR have been used to prove pathogenicity of variants affecting splicing (Gao et al., 2008; Kishore et al., 2008; Kralovicova et al., 2006).

The deletion of the adenosine at position -35 of *FGD1* intron 12 is the first report of an insertion or deletion affecting the BP. Our data make it likely that the A represents the signal for lariat structure formation; 1) its location matches the usual branch point distance from the 3' SA, 2) its sequence "cTaAc" matches the human branch point consensus sequence yUnAy (Gao et al., 2008) , 3) it is conserved (Phastcons score 1), 4) the region between BP and SA is devoid of a AG dinucleotide (an AGEZ , Figure S5) (Gooding et al., 2006) and, most convincingly, 5) we showed abnormal *FGD1* mRNA splicing. Interestingly, since we see a nearly 100% skip of exon 13 in males, the flanking 'A' residue seems not to act as a BP site.

As mentioned, diagnostic *FGD1* Sanger sequencing was performed in the two siblings but failed to identify a causative variant. The 5' primer used to amplify exon 13 covers the deleted A (Figure 2A), but it obviously did not prevent amplification in males and therefore the variant went undetected. Based on the guidelines for diagnostic molecular testing(American college of medical genetics, 2006), there is no clear consensus on the regions that must be included in a standard variant screen. Although the direct splice sites (+1 to +6 and -10 bp to -1) are usually included, the branch point is not. An important reason for this is that BP is difficult to define (weak consensus sequence). Furthermore, RNA is not usually isolated making it difficult to study the effect of any variant detected. It is interesting to note that when blood-derived RNA of the patients would have been analyzed, exon 13 skipping had been detected, triggering an analysis of introns 12 and 13 and the detection of the -35delA variant. The fact that the mutation detection rate for *FGD1* variants in clinically diagnosed ASS patients is less than 20%, may be partly due to the fact that variants affecting splicing are not routinely investigated. We tested four other clinically diagnosed ASS patients with unknown genetic background (3 Dutch, 1 Belgian) for c.2016-35del, but no abnormality was found. Since RNA was not available we could not perform a RNA analysis.

As we demonstrate here, although exome sequencing is not designed to detect intronic variants, they can be partially detected. The intron/exons borders are targeted by the whole-exome capture kits but it does not include the branch point. Even if it is captured, those

variants would be filtered out by the standard strategy to prioritize variants (nonsense, splice site, frameshift). Moreover, sufficient coverage for the branch points is required to reliably call variants while sequence coverage drops significantly near these sites because they only flank the targeted regions. The BP variant in *FGD1* was detected and subjected to detailed analysis mainly because it had a high coverage and because the gene was known to be involved in Aarskog syndrome. Depending on the current capture design, sequences can go up to 200 bp into the introns. To ensure 'standard' capture of the majority of branch points, the exome-kit probe design should be extended with sequences deeper into the intron region taking AGEZ regions into account. Sequencing deeper into the intron will uncover many more variants most of them probably not affecting splicing at all. However, validation of these variants will not be too difficult if they have sufficient coverage.

In Summary, our findings show that traditional clinical gene sequencing is not perfect and, due to practical and cost considerations, may miss causative variants. Application of exome sequencing in a diagnostic setting should give a significant increase in the detection of pathogenic variants by partially covering the intron-exon flanking regions. Ultimately, only full genome sequencing will guarantee that all possible regions are taken into account. Our data indicate that variant analysis in exome sequencing should include analysis of intronic variants, with special attention to branch point sites. In addition, RNA should never be neglected as a valuable source to confirm variants that may affect splicing or to detect such variants directly. In fact, sequencing RNA isolated from the same blood sample as currently used to extract the DNA for exome/genome sequencing may provide an attractive, cost-effective approach when no obvious pathogenic variants are identified.

Acknowledgements

The authors have no conflict of interest to declare.

We are grateful to the family, clinicians, and researchers who contributed to this study. Professor R.C.M. Hennekam is acknowledged for establishing the clinical diagnosis of Aarskog-Scott syndrome in this family. We thank Dr. E.K.Bijlsma, Dr N.S. den Hollander, Dr. A. van Haeringen and Dr. R.McGowan for contributing patient samples not directly used in this study and the technicians in the laboratory of diagnostic genome analysis for their help in lymphocyte culturing. Dr. P.A.C. 't Hoen is acknowledged for his advice and critical reading of the manuscript and J. Laros for the bioinformatic analysis. Y. Sun is supported by China Scholarship Council. This work was supported by the EC's 7th Framework Programme under grant agreement n°s 223026 (NMD-chip), 223143 (TechGene) and 200754 (GEN2PHEN).

1. American college of medical genetics. 2006. Practice guidelines for Sanger Sequencing Analysis and Interpretation .

2. Bedoyan JK, Friez MJ, DuPont B, Ahmad A. 2009. First case of deletion of the faciogenital dysplasia 1 (FGD1) gene in a patient with Aarskog-Scott syndrome. Eur J Med Genet 52:262-264.

3. Bottani A, Orrico A, Galli L, Karam O, Haenggeli CA, Ferey S, Conrad B. 2007. Unilateral focal polymicrogyria in a patient with classical Aarskog-Scott syndrome due to a novel missense mutation in an evolutionary conserved RhoGEF domain of the faciogenital dysplasia gene FGD1. Am J Med Genet A 143A:2334-2338.

4. Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37:e67.

5. Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. Nucleic Acids Res 36:2257-2267.

6. Gooding C, Clark F, Wollerton MC, Grellscheid SN, Groom H, Smith CW. 2006. A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. Genome Biol 7:R1.

7. Hou P, Estrada L, Kinley AW, Parsons JT, Vojtek AB, Gorski JL. 2003. Fgd1, the Cdc42 GEF responsible for Faciogenital Dysplasia, directly interacts with cortactin and mAbp1 to modulate cell shape. Hum Mol Genet 12:1981-1993.

8. Kaname T, Yanagi K, Okamoto N, Naritomi K. 2006. Neurobehavioral disorders in patients with Aarskog-Scott syndrome affected by novel FGD1 mutations. Am J Med Genet A 140:1331-1332.

9. Kishore S, Khanna A, Stamm S. 2008. Rapid generation of splicing reporters with pSpliceExpress. Gene 427:104-110.

10. Kralovicova J, Lei H, Vorechovsky I. 2006. Phenotypic consequences of branch point substitutions. Hum Mutat 27:803-813.

11. Orrico A, Galli L, Buoni S, Hayek G, Luchetti A, Lorenzini S, Zappella M, Pomponi MG, Sorrentino V. 2005. Attention-deficit/hyperactivity disorder (ADHD) and variable clinical expression of Aarskog-Scott syndrome due to a novel FGD1 gene mutation (R408Q). Am J Med Genet A 135:99-102.

12. Orrico A, Galli L, Cavaliere ML, Garavelli L, Fryns JP, Crushell E, Rinaldi MM, Medeira A, Sorrentino V. 2004. Phenotypic and molecular characterisation of the Aarskog-Scott syndrome: a survey of the clinical variability in light of FGD1 mutation analysis in 46 patients. Eur J Hum Genet 12:16-23.

13. Orrico A, Galli L, Faivre L, Clayton-Smith J, Azzarello-Burri SM, Hertz JM, Jacquemont S, Taurisano R, Arroyo C, I, Tarantino E, Devriendt K, Melis D, Thelle T, Meinhardt U, Sorrentino V. 2010. Aarskog-Scott syndrome: clinical update and report of nine novel mutations of the FGD1 gene. Am J Med Genet A 152A:313-318.

14. Orrico A, Galli L, Obregon MG, de Castro Perez MF, Falciani M, Sorrentino V. 2007. Unusually severe expression of craniofacial features in Aarskog-Scott syndrome due to a novel truncating mutation of the FGD1 gene. Am J Med Genet A 143:58-63.

15. Pagani F, Baralle FE. 2004. Genomic variants in exons and introns: identifying the splicing spoilers. Nat Rev Genet 5:389-396.

16. Pasteris NG, Buckler J, Cadle AB, Gorski JL. 1997. Genomic organization of the faciogenital dysplasia (FGD1; Aarskog syndrome) gene. Genomics 43:390-394.

17. Shalev SA, Chervinski E, Weiner E, Mazor G, Friez MJ, Schwartz CE. 2006. Clinical variation of Aarskog syndrome in a large family with 2189delA in the FGD1 gene. Am J Med Genet A 140:162-165.

18. Teebi AS, Naguib KK, Al-Awadi S, Al-Saleh QA. 1988. New autosomal recessive faciodigitogenital syndrome. J Med Genet 25:400-406.

19. van de Vooren MJ, Niermeijer MF, Hoogeboom AJ. 1983. The Aarskog syndrome in a large family, suggestive for autosomal dominant inheritance. Clin Genet 24:439-445.

20. Ward AJ, Cooper TA. 2010. The pathobiology of splicing. J Pathol 220:152-163.

Supplementary Material

Patient DNA

Blood samples of patients and relatives were collected and used for segregation analysis. Consent to genetic testing was obtained from adult probands or parents in the case of minors.

DNA and RNA Analysis

DNA was extracted from whole blood following standard protocols. Leucocyte RNA with or without cycloheximide (CHX) treatment was isolated from blood using Nucleospin RNA II kit (MACHEREY-NAGEL) following the official protocol. cDNA was synthesized from 500 ng of total RNA by RevertAid RNaseH-M-MuLV reverse transcriptase in a total volume of 20 µl according to the protocol provided by the supplier (MBI-Fermentas). Target regions were amplified by RT-PCR (primer sequences are available upon request). The products were evaluated with the Bioanalyzer 2100 DNA chip 1000 (Agilent), according to the manufacturer's instructions.

SNP arrays

SNP arrays (1M-Duov3-0 Illumina Inc., San Diego, CA) were performed and a total of 750 ng DNA was processed according to the manufacturer's instructions. SNP copy number (log R ratio) and B-allele frequency were assessed for II-2, III-1 and III-2 to exclude large copy number variations (CNVs) as a possible genetic cause.

Mutation detection by exome sequencing

To resolve the genetic basis of ASS in this family, we performed exome sequencing of the two siblings (III-1 and III-2) using the SureSelect human all exon kit v2 (Agilent), following the manufacturer's instructions. In brief, 5 ug genomic DNA was fragmented to 300-800 bp using a Covaris S-series and Illumina pair-end adapters were ligated to the fragments. 500 ng adapter ligated DNA was hybridized with 500 ng SureSelect probe mix. Sequencing (one lane per patient) was performed on the Illumina GAIIX. All reads were aligned with BWA-0.5.8 (Li and Durbin, 2010) to the human reference genome hg19. Variant calling, including both single nucleotide substitutions (SNVs) and small insertions, deletions and indels, were done by Samtools-0.1.12 (Li et al., 2009) followed by annotation using SeattleSeq Annot.131.

Sanger sequencing

PCR was performed by using Phire Hot Start II DNA polymerase (Finnzyme) following the official protocol. Primers used in PCR reactions are available upon request. PCR products were first purified by QIAquick PCR purification kit (QIAGEN), then mixed with 25 pmol of the forward or reverse primers respectively and sequenced by the Applied Biosystems 96-capillary 3730XL system.

In Silico Analysis

Prediction of a splice effect was tested using the Human Splicing Finder (HSF) tool which gives a position Weight matrices-derived scores (HSF) for potential splice sites (Desmet et al., 2009), and also integrates the results from the MaxEnt program.

Accession Numbers

GenBank reference sequences: *FGD1*: NM_004463.2
UNIPROT: FYVE, RhoGEF and PH domain-containing protein 1: P98174

Table S1: Standard filtering for exonic regions.

	III-1	III-2
Unique genomic variants	74019	43209
Exonic + 5' and 3'-SS	25642	21491
In sureselect 38M probeset	20080	20187
Depth >= 8	18378	18990
Not coding synonymous	8394	8556
Not validated DbSNP/1K	570	672
Not present In house database	53	169
Conserved variants[a]	30	62
Variants present in two sibs	2	

[a] phylop >3 for SNVs, Phastcons >0.8 for indels

Table S2: Reported variants after standard analysis for intronic positions up to -50 bp.

Sample	Variant	Gene	Function	Depth	dbSNP	OMIM
III-1	NM_018325.2:c.607G>C	C9orf72	missense	9	none	No
III-2	NM_018325.2:c.607G>C	C9orf72	missense	9	none	No
III-1	NM_004463.2:c.2016-35delA	FGD1	intron	11	none	Yes
III-2	NM_004463.2:c.2016-35delA	FGD1	intron	11	none	Yes
III-1	NM_080818.3:c.512_513insA	OXGR1	frameshift	43	none	No
III-2	NM_080818.3:c.512_513insA	OXGR1	frameshift	67	none	No

Figure S1: Phenotype of the two Aarskog siblings (III-1, III-2) and their maternal grandfather (I-1). Facial features are illustrated at age 8 years and 13 years (III-1) and at age 6 years and 11 years (III-2). Craniofacial features, skeletal and genital features are described in Table 1.

III-1

III-2

I-1

Figure S2: Coverage plots of GAIIX sequencing for III-1 and III-2.

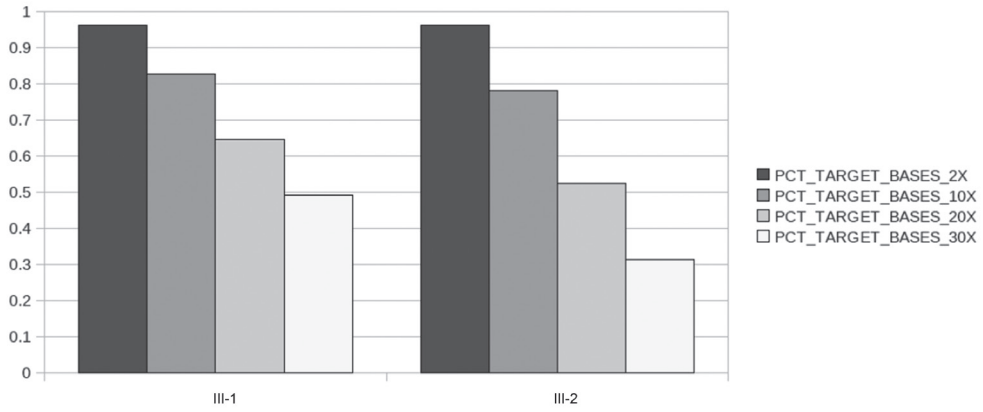


Figure S3: The 2100 Bioanalyzer Traces of RT-PCR show semi-quantification of skipped products. The peaks represented are the lower ladder (15 bp) and upper ladder (1500 bp), the skipped FGD1 product (163 bp) and the WT product.

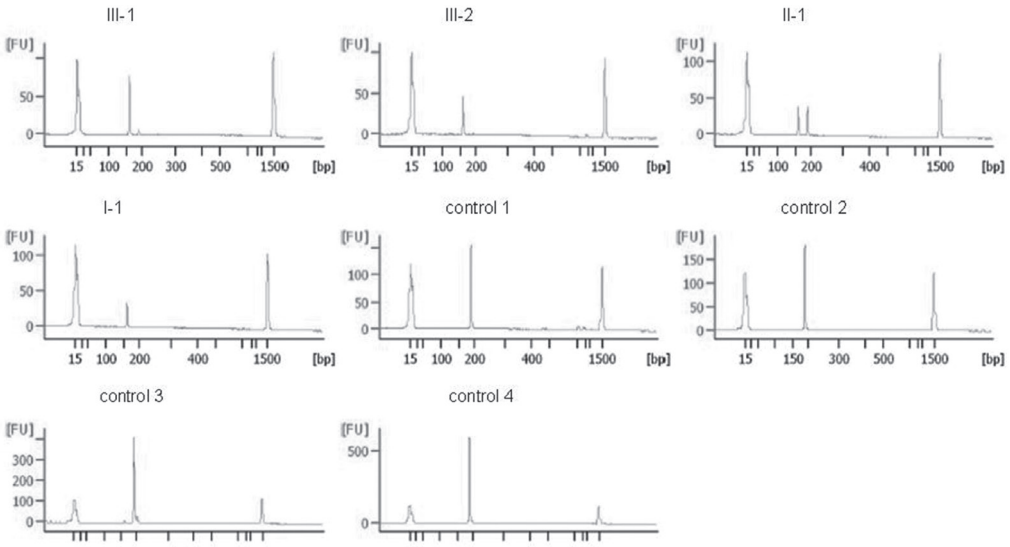


Figure S4: Application of exome sequencing for detection of intronic variants.
A) Intronic coverage plot for a range of distances to the splice site in our dataset. Percentage of the introns covered (minimal depth 8) decreases with an increasing distance from splicesite. B) Calculation of the number of filtered variants upon inclusion of intronic variants by a range of distances to the splice site (depth >8).

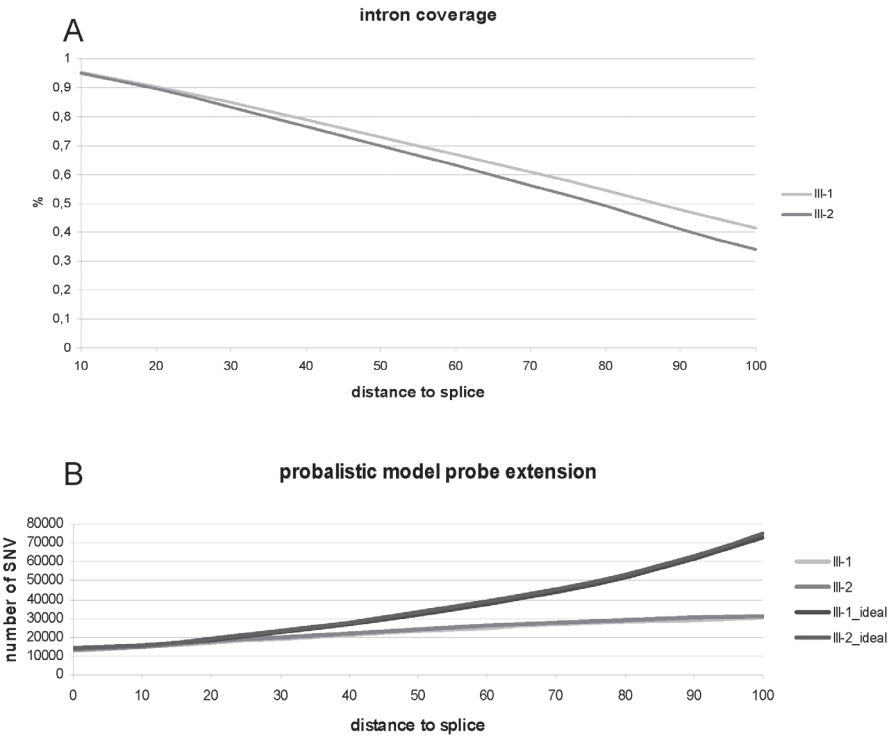
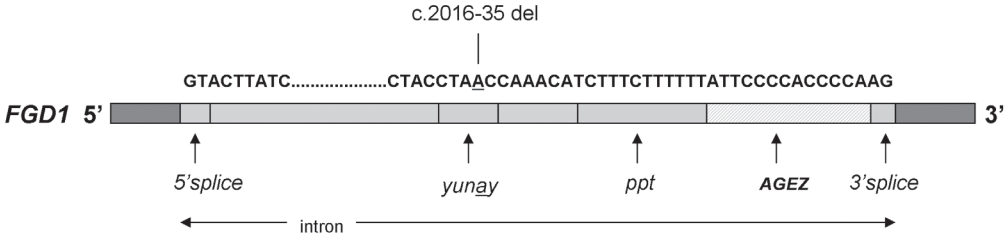


Figure S5: Overview of splice elements of FGD1 exon 12-13. 5'splice site (GU dinucleotide), branch point sequence YUNAY with adenine for lariat formation, AG exlusion zone (AGEZ), polypyrimidine tract (ppt) and 3'splice site (AG dinucleotide).




8

Mutations in Swi/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome

Gijs W.E. Santen, Emmelien Aten, Yu Sun, Rowida Almomani, Christian Gilissen, Maartje Nielsen, Sarina G. Kant, Irina N. Snoeck, Els A.J. Peeters, Yvonne Hilhorst-Hofstee, Marja W. Wessels, Nicolette S. den Hollander, Claudia A.L. Ruivenkamp, Gert-Jan B. van Ommen, Martijn H. Breuning, Johan T. den Dunnen, Arie van Haeringen, Marjolein Kriek

Nature Genetics. 2012;44(4):379-80

Abstract

We identified *de novo* truncating mutations in ARID1B in three individuals with Coffin-Siris syndrome (CSS) by exome sequencing. Array-based copy-number variation (CNV) analysis in 2,000 individuals with intellectual disability revealed deletions encompassing ARID1B in three subjects with phenotypes partially overlapping that of CSS. Together with published data, these results indicate that haploinsufficiency of the ARID1B gene, which encodes an epigenetic modifier of chromatin structure, is an important cause of CSS and is potentially a common cause of intellectual disability and speech impairment.

Main text

Coffin-Siris syndrome (OMIM %135900, CSS)[1] is characterized by developmental delay, severe speech impairment, coarse facial features, hypertrichosis, hypoplastic or absent fifth finger- or toe nails[2] and agenesis of the corpus callosum (Supplementary Table 1). Few published patients fulfill the complete spectrum of the CSS phenotype and it is debated whether all CSS patients have the same syndrome. CSS is generally assumed to display autosomal recessive inheritance, although autosomal dominant inheritance has not been formally excluded[3, 4].

To identify the genetic cause of CSS we performed whole exome sequencing in one patient-parent trio and two sporadic patients with a clinical CSS diagnosis all diagnosed in one hospital by the same clinical geneticist (Fig. 1, Supplementary note, Supplementary Fig. 1, and Supplementary Tables 1 and 2, Supplementary methods). Exome sequencing data are available upon request. Using the GATK sequence analysis pipeline[5, 6] we identified 12,722- 14,642 exonic/splice site variants per individual. Filtering steps using variant databases (dbSNP132 and the 1000 genomes project database) and a selection for coding regions, revealed variants in 34 genes that were shared by all three patients. Filtering for recessive inheritance (discarding all genes with only one heterozygous variant) showed that no gene in agreement with a recessive inheritance model was shared between the patients. Accepting dominant inheritance, we queried heterozygous and *de novo* variants and identified *ARID1B* as the only possible affected gene (Supplementary Table 3). All variants truncate the *ARID1B* reading frame (two nonsense variants -c.5329A>T,p.Lys1777X and c.3223C>T,p.Arg1075X), one frameshift -c.4619_4628del,p.Gln1541ArgfsX35-, Table 1). The mutations were validated using Sanger sequencing and shown to occur *de novo* in all three patients (Supplementary Fig. 2). Since thus far an autosomal recessive inheritance could not be ruled out, the parents of a CSS patient received a recurrence risk of 10%[7]. The identification of *de novo* mutations in *ARID1B* in CSS patients, allowed us to reduce this risk to 1-2%[8].

When we queried our in-house database for patients with potential copy number variants (CNVs) including *ARID1B* we identified three patients with a deletion in 2000 subjects screened for intellectual disability (Fig.1, Supplementary Fig. 1 and 2, Supplementary Table 1). This cohort consists of patients with intellectual disability and/or congenital malformations (syndromic and non-syndromic) referred for array-based CNV analysis. In comparison, we found six patients with the relatively frequent 22q11.2 duplication in this cohort. Patient four has a *de novo* 2.72 Mb deletion of band 6q25 encompassing *ARID1B*, *C6orf35*, *ZDHHC14*, *SNX9* and *SYNJ2*.

Figure 1: Facial features of all patients. Top, left to right: Patient 1 at 4.5 years, patient 2 at 2,5 years, patient 3 at 3 years. Bottom, left to right: Patient 4 at 3.5 years , patient 5 at 45 years, patient 6 at 3,5 years. All patients share coarse facial features, thick eyebrows and broad nasal tips. For further details see Supplementary Table 1. The parents or legal guardians of all affected individuals gave consent for publication of the clinical photographs.

Patient five has a *de novo* 0.73 Mb deletion encompassing *ARID1B, C6orf35* and *ZDHHC14*. Patient six has a *de novo* deletion of 0.88 Mb encompassing the same genes as patient five. Analysis of the phenotypes of these patients shows that, similar to our CSS patients, they have moderate to severe intellectual disability and severe speech delay. These patients also share facial similarities with the CSS patients (Figure 1) but lack the typical CSS abnormalities, such as hypoplastic or absent finger- or toenails. The CSS diagnosis was considered only in patient four because of her hypoplastic fingernails (Supplementary Table 1). Halgren *et al.*[9], describe eight patients with haploinsufficiency of *ARID1B* (in one subject the gene was disrupted by a reciprocal translocation). An additional disruption of *ARID1B* as well as a *de novo* intragenic deletion have been reported in patients with either corpus callosum agenesis or autism[9, 10]. Although the sizes of the published deletions range

Table 1: Overview of the detected variants and CNVs affecting ARID1B.

Patient (DECIPHER ID)	Chr	Position	Size of deletion	cDNA position	Protein Position	Exon	phyloP score
1	6	g.157527604A>T	-	c.5329A>T	p.Lys1777X	20	2.9
2	6	g.157502190C>T	-	c.3223C>T	p.Arg1075X	12	1.9
3	6	g.157522346_157522356del	-	c.4619_4628del	p.Gln1541ArgfsX35	18	3.0
4 (248472)	6	155,797,565 - 158,517,307	2.72 Mb	-	-		-
5 (250455)	6	157,079,676 - 157,806,675	0.73 Mb	-	-		-
6 (257917)	6	157,144,644 - 158,028,969	0.88 Mb	-	-		-

Variants are mapped to the hg19 reference genome. Deletion size represents the maximum deletion. All variants occurred *de novo*. Chr.,chromosome. Subjects 4 and 5 were published previously [9].

from <1 Mb to > 14 Mb, the phenotypes of the patients largely overlap with those of our CSS patients, with the exception of the typical fifth finger CSS abnormalities. Disruption of *ARID1B* therefore seems to be the main driver of the observed phenotype and CSS should be considered in all patients with intellectual disability and speech impairment, particularly in combination with agenesis of the corpus callosum. Based on these findings we conclude that haploinsufficiency of *ARID1B* is likely to be an important cause of CSS.

The public database DECIPHER (http://decipher.sanger.ac.uk) contains 12 patients with haploinsufficiency of *ARID1B* (including the three described here and those described by Halgren *et al.*). Since all deletions and mutations published thus far have been *de novo*, disease penetrance is expected to be high. No truncating or splice-site mutations are reported in the 1000 genomes project, nor in the ~ 5400 exomes on the Exome Variant Server (http://evs.gs.washington.edu/EVS/). However, the Database of Genomic Variants harbors two deletions including exon 1 and exon 2-20 of the *ARID1B* respectively. This could signify reduced penetrance, but an alternative explanation would be that these deletions represent a technical artifact. Both deletions were not validated using alternative methods.

Together with *ARID1A*, *ARID1B* encodes for ARID (AT rich interactive domain), a subunit of the BAF complex. BAF (Brahma associated factor) is one of the two main components of the Switch/sucrose nonfermentable (SWI/SNF)-like chromatin remodeling complex. They act as epigenetic modifiers by altering the structure of chromatin to facilitate access of

transcription factors to DNA. ARID1A and ARID1B proteins have antagonistic functions and they are both important for the regulation of the cell cycle. Although *ARID1B* is predominantly expressed in differentiated cell-types, it has also been suggested to be involved in early development of the brain[11]. *ARID1A* is more abundantly expressed in embryonic tissue and somatic mutations have recently been related to gastric cancer[12]. Recently, Gilissen *et al*[13] remarked that histone modifying proteins seem to have a dual role in developmental disorders and malignancies. One could hypothesize that activating mutations in *ARID1A* might give a clinical phenotype similar to CSS.
In conclusion, CSS can be added to a growing list of syndromes characterized by congenital malformations, and intellectual disability that are caused by mutations in epigenetic genes that encode modifiers of chromatin structure[14, 15].

Accession code. Data for *ARID1B* is deposited in RefSeq under the accession code NM_020732.3.

Acknowledgements

We thank the patients and their parents for participation in this study. We would like to acknowledge Alexander Hoischen for his useful suggestions regarding the interpretation of exome sequencing data. We received funding from the EU FP7 framework program agreements 223026 (NMD-chip) and 223143 (TechGene).

Author contributions

G.W.E.S. analyzed the data and wrote the manuscript. E.A., G.J.v.O, M.H.B., J.T.d.D, A.v.H and M.K. conceived and designed the experiments. E.A., Y.S. and R.a.M. performed the experiments. C.G. contributed analysis tools. M.N., S.K., I.S., E.A.J.P, M.W.W., N.S.d.H, Y.H-H and A.v.H clinically characterized the patients. C.A.L.R. analysed SNP-array data. A.v.H. selected the patients for sequencing. A.v.H. and M.K. jointly supervised the research. All authors contributed to the final manuscript.

Competing financial interests

The authors declare no competing financial interests.

1. Coffin,G.S. & Siris,E. *Am. J. Dis. Child.* **119**, 433-439 (1970).

2. Fleck,B.J., Pandya,A., Vanner,L., Kerkering,K., & Bodurtha,J. *Am. J. Med. Genet.* **99**, 1-7 (2001).

3. Carey,J.C. & Hall,B.D. *Am. J. Dis. Child.* **132**, 667-671 (1978).

4. Haspeslagh,M., Fryns,J.P., & van den Berghe,H. *Clin. Genet.* **26**, 374-378 (1984).

5. DePristo,M.A. *et al. Nat. Genet.* **43**, 491-498 (2011).

6. McKenna,A. *et al. Genome Res.* **20**, 1297-1303 (2010).

7. Levy,P. & Baraitser,M. *J. Med. Genet.* **28**, 338-341 (1991).

8. Harper,P.S. *Practical Genetic Counseling*(Edward Arnold (Publishers) Ltd.,2004).

9. Halgren,C. *et al. Clin. Genet.* **DOI:10.1111/j.1399-0004.2011.01755.x**, (2011).

10. Backx,L., Seuntjens,E., Devriendt,K., Vermeesch,J., & Van,E.H. *Cytogenet. Genome Res.* **132**, 135-143 (2011).

11. Flores-Alcantar,A., Gonzalez-Sandoval,A., Escalante-Alcalde,D., & Lomeli,H. *Cell Tissue Res.* **345**, 137-148 (2011).

12. Wang,K. *et al. Nat. Genet.* **43**, 1219-1223 (2011).

13. Gilissen,C., Hoischen,A., Brunner,H.G., & Veltman,J.A. *Genome Biol.* **12**, 228 (2011).

14. Hargreaves,D.C. & Crabtree,G.R. *Cell Res.* **21**, 396-420 (2011).

15. Bokhoven,v.H. *Annu. Rev. Genet.* **45**, 81-104 (2011).

Supplementary information

The parents of patients 1-3 gave informed consent to participate in this study. The parents or legal guardians of all patients gave consent for publication of the clinical photographs. All DNA samples were isolated from peripheral blood leucocytes according to standard techniques.

Samples for next generation sequencing were prepared according to the manufacturer's instructions. Target enrichment was performed using Agilent's 50 Mb Sureselect v2 exome capture kit. Paired-end reads with a length of 100 nucleotides were sequenced on Illumina's Hiseq 2000. Sequencing statistics are summarized in Supplementary Table 2. BWA[1] was used to align the short reads to the reference genome (hg19), followed by GATK[2] base quality score recalibration, INDEL realignment and duplicate removal. SNP and INDEL discovery and genotyping for each sample were performed separately using standard hard filtering parameters for INDELS and variant quality score recalibration[3] for SNP calls as indicated on the GATK website (http://www.broadinstitute.org/gsa/wiki/index.php/Best_Practice_Variant_Detection_with_the_GATK_v3). Because of the low coverage of our data two modifications were made: (1) the minimum quality for variant calling was set at 20 (default 30) and (2) the minimum number of reads supporting an indel was set at 3 (default 6). Variants were annotated using Seattleseq (http://gvs.gs.washington.edu/) and in-house developed software.

→

Supplementary Table 1: Patient characteristics. Present +, Absent -, not tested/observed: ?. Patients 1-3 fulfill the diagnostic criteria for CSS since they have either fifth ray abnormalities or hypoplastic phalanges in their toes. Patients with deletions of *ARID1B* were not noticed to have these abnormalities, the main reason why they were not diagnosed with CSS. In retrospect a subtle degree of fifth finger/nail hypoplasia is discernable in patient 5.

Patient number	Truncating mutation			Deletion		
	1	2	3	4	5	6
Gender	female	female	female	female	male	female
Age at presentation	4.5 years	6 years	3 years	2 years	46 years	3.5 years
Mutation/CNV *ARID1B*	mutation	mutation	mutation	deletion	deletion	deletion
Inheritance of mutation	*de novo*	*de novo*	*de novo*	*de novo*	*de novo*	*de novo*
Birth (gestational age)	42	40	39	39+4	At term	41+1
Birth weight (g)	3370	3720	normal	2770	3800	2722
OFC at birth (cm)	?	36.9	?	?	?	35
Intellectual disability (mild, moderate, severe)	severe	severe	moderate	severe	severe	moderate
Speech delay (mild, moderate, severe)	severe	severe	severe	no speech	no speech	moderate
Coarse facial features	+	+	+	+	+	-
Thick eyebrows	+	+	+	+	+	+
Low frontal hairline	+	+	+	-	-	+
Hypertrichosis	+	+	+	-	-	-
Joint laxity	+	+	-	-	+	?
Brachydactyly	+	+	-	-	-	-
Brachydactyly fifth finger	+	+	-	-	-	-
Hypoplastic nails	-	-	-	+	-	-
Hypoplastic nail fifth finger	-	+	+	-	-	-
X-ray hand/foot performed	+	+	+	-	-	-
Missing/hypoplastic phalanx of toes	-	+	+	?	?	?
MRI performed	+	+	+	-	-	+
Agenesis of the corpus callosum	+	partial	+	?	?	+
Height (SDS) at last visit	-2,5	-1	0	-1,5	-3,5	-2.6
Weight (SDS) at last visit	0	0	+1	-3,0	0	0.3
OFC (SDS) at last visit	0	+1	+1	0	-0,75	-0.2
Other remarkable features	colpocephaly; autism; strabismus	fetal finger pads	Colpocephaly; strabismus	atrial septum defect; sparse hair; no nail growth; unilateral single palmar crease; nephrocalcinosis at 4 years	anal atresia; double ureters.	pectus excavatum; abnormal palmar creases; striped toe nails

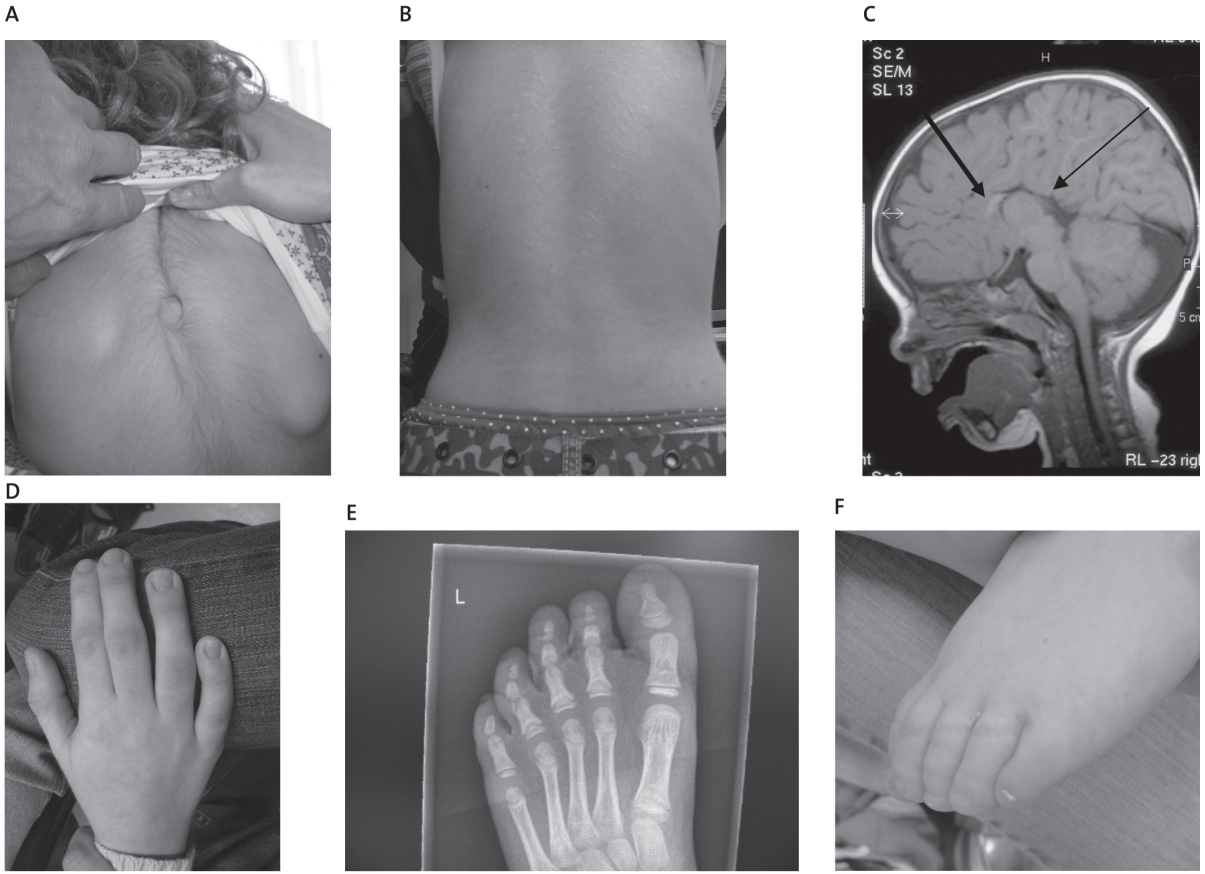
Supplementary Table 2: Sequencing statistics (patients 1, 2 and the parents of patient 1 were sequenced twice).

Patient	1	2	3	Father patient 1	Mother patient 1
Total number of sequenced reads	67,491,568	119,409,964	36,045,692	108,724,122	74,764,116
Aligned reads	64,567,654	114,400,816	35,140,431	104,347,767	71,874,327
Properly paired reads	63,859,108	112,053,996	34,476,740	103,155,322	71,167,452
Mean target coverage	19.5x	34.6x	14.8x	27.7x	21x
% Target bases covered >10x	56	61	57	61	57

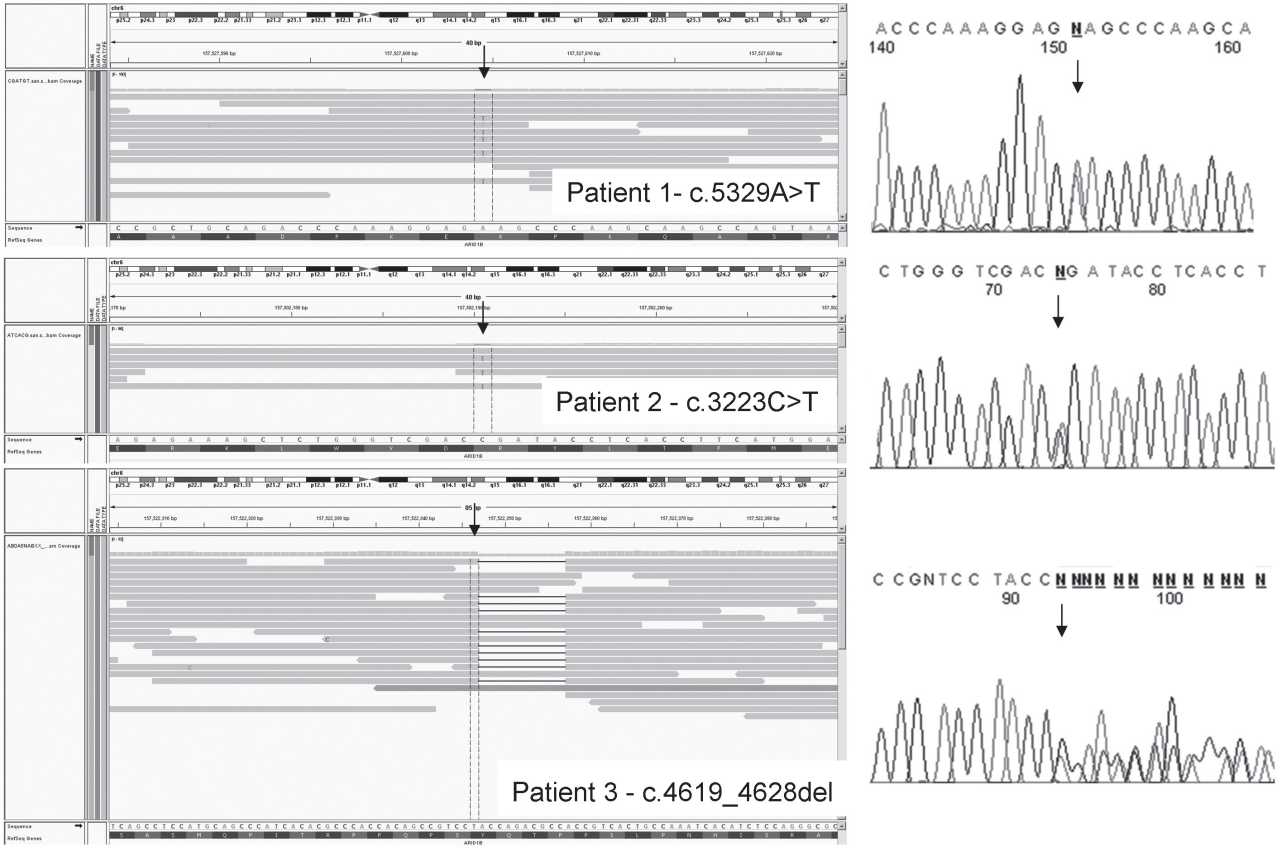
Supplementary Table 3. Filtering steps in 3 CSS patients. Filtering for recessive inheritance did not result in any genes shared by the patients. The number of *de novo* variants in patient one appears high; for six of these variants the coverage in either parent was low (<8). Two variants appear to be inherited (but a slightly different insertion was called in the parents). One 24 bp deletion is called in *SPRR3* but is present in only 3/14 reads. The only compelling variant remaining is the nonsense variant in *ARID1B*.

Filter	Patient 1	Patient 2	Patient 3	# genes with mutation in all patients
Number of variants	147,120	275,502	71,594	
Not in DBSNP/1000G	6,244	12,292	3,089	1,185
Only exonic / splice-site	213	256	313	34
De novo	10			
Heterozygous		223	285	1

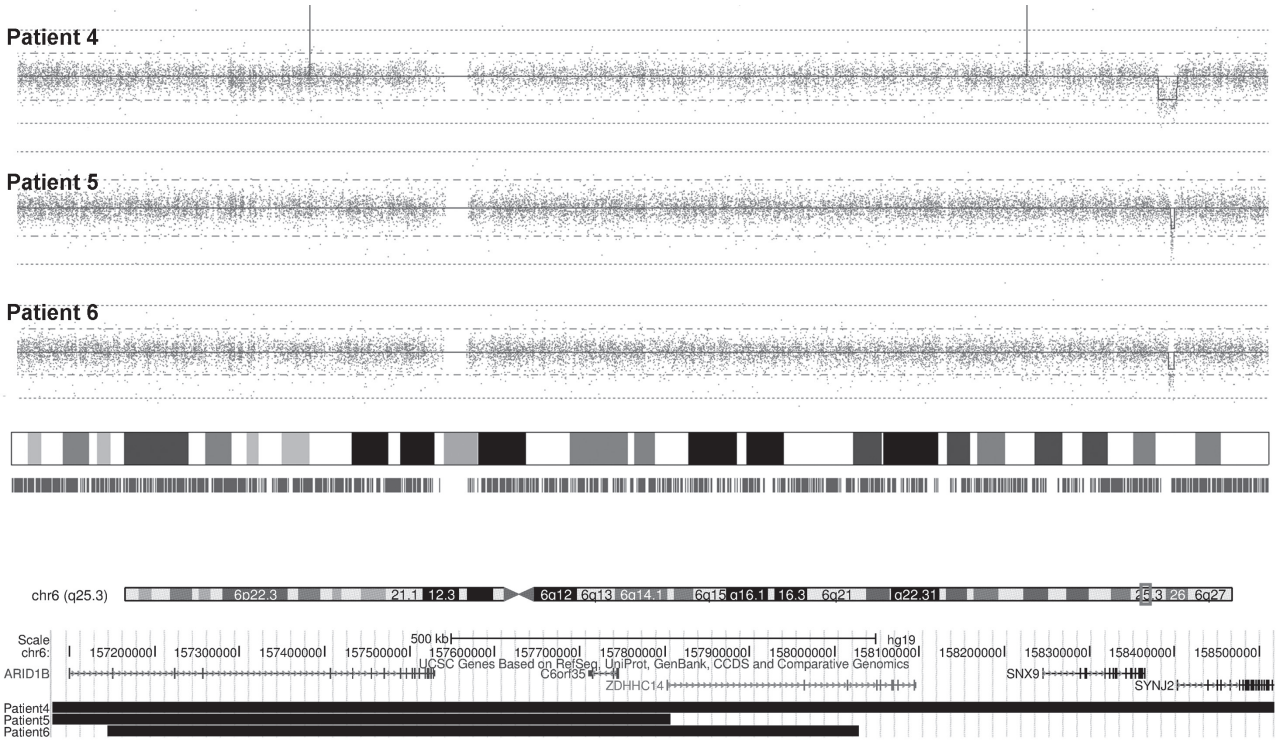
Supplementary Figure 1: Additional features in the CSS patients. Hypertrichosis in patients 1 (A) and 2 (B). Agenesis of the corpus callosum in patient 1 (thin arrow, thick arrow points to a small part of the corpus callosum that is present) (C). Brachydactyly of the fifth finger in patient 2 (D), missing terminal phalanx of the fifth toe in patient 3 (E), hypoplastic nail in patient 4 (F.



Supplementary Figure 2: On the left the raw data around the pathogenic variant in the Integrative Genomics Viewer[3] f or each of the 3 patients. Arrows indicate the location of the variant. On the right the Sanger sequencing validation results.



Supplementary Figure 3: Top: Plot of the genomic deletions detected in patients 4 to 6. Red dots represent the raw data, the blue line represents the calculated CNV score. All deletions are *de novo* (data of parents not shown). No other possibly pathogenic CNVs were found in these patients. Bottom: graph of the size and location of each of the deletion.



1. Li,H. & Durbin,R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **25**, 1754-1760 (2009).

2. McKenna,A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297-1303 (2010).

3. Robinson,J.T. et al. Integrative genomics viewer. Nat. Biotechnol. **29**, 24-26 (2011).


9

General discussion

Discoveries made in the 20th century had a major impact in genetics. In less than 60 years, the field has moved from discovering the structure of DNA to potentially identifying all the variants in an individual's genome. Current (diagnostic) research focuses on identifying the genetic basis of disorders, both at the fine (sequence variants) and large scale (structural variants). While Sanger sequencing has been the golden standard for detection of sequence variants for many years, replacement by whole exome sequencing (WES) is already taking place, soon to be followed by whole genome sequencing (WGS). WES now merely focuses on detection of sequence variation but, although limited, allows detection of structural variation as well. For detection of structural variation, conventional karyotyping has been replaced by microarray platforms, and the resolution of the technique has increased from the 5-10 Mb level towards the kilobase level. With NGS, even smaller structural variation will be detected. For balanced rearrangements and complex rearrangements, although WGS can identify breakpoints, conventional karyotyping and FISH are still valuable techniques.

Several techniques described in this thesis have made their way into clinical diagnostic laboratories. Chapters in this thesis show that each technique has specific applications, and which technical approach to use depends on the resolution and throughput of the technique in combination with the strategy for the identification of the disease gene. While chapter 2 and chapter 4 describe more methodological work on techniques to detect sequence variation and copy number variation, chapters 3 and 5 describe the application of these techniques to the identification of pathogenic variants. Chapter 6 and 7 reflect the rapid introduction of NGS and the successful application of exome sequencing in clinical genetics.

Technological innovations are ongoing. For some of the research published in this thesis we would now already apply other techniques. For instance, the development of a 1400-plex bead assay to screen for CNVs in patients with intellectual disability (chapter 2) is now obsolete, as microarrays made their way into diagnostic laboratories. However, this targeted approach can still be considered useful to circumvent dilemmas like the occurrence of unsolicited findings connected to genome-wide screening. Detailed delineation of microdeletion/microduplication syndromes (chapter 3) is time consuming, and exome sequencing will probably aid in establishing genotype-phenotype correlations within these patients. Exome sequencing of mentally retarded patients with unknown aetiology may reveal pathogenic variants in genes previously described in microdeletion/microduplication intervals. This has been the case for Coffin-Siris syndrome (chapter 8) where pathogenic variants in *ARID1B* were identified in patients using exome sequencing. This gene was previously described in patients carrying a larger deletion affecting *ARID1B,* indicating haploinsuffiency as the causative effect. Additionally, in some of the cases, exome sequencing will allow detection of pathogenic variants on the other allele in genes within

microdeletion/microduplication intervals, confirming a recessive mode of inheritance.

Gene identification in Mendelian disorders

In general, knowledge of advantages and disadvantages of different laboratory techniques will aid in setting up experiments, and allows a critical assessment of subsequent findings. The identification of genes responsible for Mendelian disorders enables molecular diagnosis of patients, as well as testing gene carriers and prenatal testing. Uncovering genetic defects underlying monogenic inherited diseases (this thesis) requires a broad knowledge of genetics. Strategies to identify pathogenic variants include assumptions on inheritance patterns (dominant, *de novo* dominant, recessive, X-linked), careful selection of cases based on phenotype, collection of families and choosing the best experimental approach.

Several strategies for gene identification can be chosen, such as a straightforward candidate gene approach or positional cloning. Previous strategies such as linkage analysis combined with segregation studies and homozygosity mapping have proven successful. Candidate genes within identified genomic loci can then be studied using Sanger sequencing. The combination of classic strategies with novel techniques has proved to be valuable in KFSD research (chapter 4). Data from previously performed linkage analysis were combined with SNP array data to reduce the critical interval of the locus. HRMA was then used as a screening method to identify the pathogenic variant in the candidate genes. There is no doubt that, in this case, X-exome sequencing would have performed equally in identifying the pathogenic variant and would have been faster. This strategy was applied in TOD (Chapter 6), where a splice-site variant was identified in the linkage interval [1]. Remarkably, the variant was missed using other techniques. The effect of the variant on gene splicing could only be shown on patient-derived tumour tissue that had been stored for 15 years. This story illustrates several problems related to sequence variation. Not only can a variant remain undetected due to technical failure, it is also quite possible that a variant is indeed detected, but not recognized as being pathogenic.

The underlying gene defect is still unknown for many rare monogenic diseases but the development of WES has allowed us to resolve many more cases. In particular, *de novo* variants for sporadic cases of previously unknown genetic disorders were identified [2-6].

Genome-wide techniques such as SNP arrays have already changed our approach to genetic disorders and this will continue as the use of WES/WGS will eventually become a standard technique in clinical diagnostic laboratories. The identification of causative variants in patients moved from clinical recognition of a syndrome and collecting patients with overlapping clinical features (phenotype-driven research), towards collecting patients with similar genomic aberrations and, secondly, delineating the phenotypic overlap (genotype-

driven research). The research on Coffin-Siris syndrome (chapter 8) reflects the success of accurate phenotyping of several sporadic patients and searching for a shared defective gene under assumption of different inheritance patterns. The fact that this syndrome is now proven to be *de novo* instead of, as assumed, following autosomal recessive inheritance has a major impact on recurrence risk for parents. The identification allowed us to reduce this risk from 25% to 1-2%. The success of identifying pathogenic variants in a single gene, in multiple unrelated patients with a similar phenotype, is highly dependent on the involvement of a clinical geneticist. Therefore, selection of patients for WES should be performed in close collaboration between molecular and clinical geneticists.

The discovery of causal variants and candidate genes responsible for Mendelian disorders and complex disorders will also help in understanding their biological function.
In addition to WES, experiments can be performed to provide more insight into disease mechanisms. For instance, KFSD (chapter 5), TOD (chapter 6), and AAS (chapter 7) all show X-linked inheritance. In principle, females have random X-inactivation. Thus, for some X-linked disorders, carrier females may show variable and usually milder symptoms of the disease, based on the pattern of X-inactivation. The X-inactivation studies performed for KFSD and TOD showed imbalances in allelic expression, perfectly matched with skewed levels of X-inactivation. In addition, for KFSD, functional *in vitro* assays were developed to prove that the identified variant leads to loss of proteolytic function of the MBTPS2 protein. For AAS, the putative branch point variant was proven to have an effect on splicing by performing RNA analysis. Together, these experiments were critical for understanding the disease mechanism and explaining the severity of phenotype in these disorders. For Coffin-Siris syndrome, functional assays are being developed to provide more insight into epigenetic modifications in the SWI/SNF complex.

Advantages and disadvantages in variant detection with WES

Currently, high-throughput sequencing of the entire genome is still too expensive to be applied in a clinical setting, so a targeted approach such as exome sequencing is the most practical approach towards finding causative variants for genetic disorders. However, NGS is becoming cheaper at an accelerating rate and whole genome sequencing will soon become affordable.
Exome sequencing comes with several disadvantages. It is labour intensive, and includes many steps, with risks of sample swaps and pipetting errors. There is significant variability in capturing efficiency, leading to variability in coverage across the genome. Low input yield and preferential amplification of shorter sequences in the PCR step leads to overrepresentation of one allele, and some reads will be exact copies of each other (PCR

duplicates). This will result in problems regarding variant detection in samples. Removal of these duplicates is possible during data analysis, but insufficient coverage due to capture inefficiency cannot be disregarded. An adequate depth is especially critical for identifying heterozygote variants in *de novo* dominant disorders, or recessive disorders caused by compound heterozygosity. Standards for generation of high-quality exome sequence data are rapidly emerging [7, 8] but there are no clear guidelines yet on the number of reads needed for identification of heterozygote SNV calls, and thresholds are determined by the user when performing data analysis.

Another challenge of NGS is read length. Current NGS machines produce shorter read lengths than Sanger sequencing, which makes the assembly of repeat-rich sequences more difficult [9]. Together with mapping difficulties, this leads to missed variants or an excess of variant calls [10]. Paired-end sequencing (where sequences are retrieved from both ends of the same molecule) helps in reducing mapping errors although indels and structural variants remain challenging. To solve this, other computational methods (read-pair, read-depth, split-read) have been developed [11, 12]. False positives SNV calls most often arise from incorrect mapping and sequencing errors. Sequencing errors can be removed by comparing the test sample to previously sequenced samples, stored in an in-house-database (INHDB). Although detection of structural variants with WES is possible [13], conventional techniques such as karyotyping, FISH, array-CGH and SNP-arrays will still provide useful information on structural variation until WES or WGS becomes a routine diagnostic approach.

The human reference genome plays an important role in several steps of WES. It is used to design chemically synthesized nucleic acid molecules and to design probes to capture regions of interest in the pool of nucleic acids (DNA or RNA). This means that targeted assays, such as WES, do not cover unknown or yet-to-be-identified exons, regulatory sequences or evolutionary conserved coding regions [10]. Moreover, if causal variants lie within exons that are not targeted, they will not be identified. This shortcoming of incomplete capture has occurred in the first attempts by several groups, including our own, to identify the genetic cause for Kabuki syndrome [2].

WES is not targeted towards intronic sequences, promoter or enhancer elements. In routine diagnostics, apart from the direct splice sites, causative variants outside coding regions are not studied. Therefore the true fraction of variants disrupting splicing remains largely unknown. However, the research that was performed for Aarskog-Scott syndrome (Chapter 7) shows that WES can be used to identify intronic variants up to -200 bp from the splice site with sufficient coverage, although it is not specifically designed to do so. A branch point variant was identified that was missed with Sanger sequencing. Detection of branch site variants is possible with WES due to probe overhang. Future probe design for targeted WES approaches may take advantage of this finding. On the other hand, if there is

a continuous demand to increase the coverage or size of targeted regions, whole genome sequencing is probably the best option in the long run. The disadvantage of extending into the non-coding sequences will be the large number of variants detected, and the need to verify aberrant splicing using RNA analysis which is not always possible if the gene is not expressed in available cells.

Whole genome sequencing (WGS) using a paired-end approach should eventually allow the identification of all types of currently known genomic variation in a single experiment, and will gradually replace WES. Improving NGS techniques and lowering the costs, together with the development of affordable analysis software to cope with the millions of variants called per genome, is essential to pave the way for diagnostic application of WGS. Development of software that can combine data from several sources containing information on reported variants has a high priority.

Interpretation of variants using NGS

The more we learn about the human genome, the more we are confronted with the need to separate 'the wheat from the chaff' i.e. distinguishing pathogenic variants from neutral variants. In NGS experiments, individual sequences are aligned to a reference genome. The choice for a specific reference genome (for instance Caucasian or Yoruba) will significantly influence variant detection in the sample of interest. It is important to realize that the human reference genome does not represent any one person's genome, but merely comprises of a mix of sequences from different chromosomes as well as individuals. This means that the human reference genome also contains variations from several donors, including somatic variants. Moreover, the haploid reference sequence has a bias toward a European population background (International Human Genome Consortium, 2004).

In general, variant positions are compared between the sample of interest and the reference, followed by a number of filtering steps to separate benign variants from potential pathogenic variants. The main strategy employed to identify causative variants in WES is to focus on variants with clear functional consequences (nonsense variants, splice-site variants, frameshift, conserved missense and indels). Depending on the enrichment kit, the sequencing platform and the pipeline used for mapping, alignment and variant calling, approximately 20000-50000 variants are identified per exome. After filtering noncoding and synonymous variants, the number of variants remaining is ~5000 [14]. For Mendelian disorders, most of the common and less likely causal variants can be efficiently removed using data from control databases and prediction tools, and between 150-500 non-synonymous and splice site variants remain to be prioritized as potentially pathogenic [15]. Downstream analysis is based on certain assumptions of inheritance (e.g. dominant, recessive, *de novo*) Further filtering strategies are needed to find the causative variant. The

success rate also depends on prior genetic mapping approaches (linkage, homozygosity mapping, CNV analysis) and availability of other patients and/or family members. This means that exome sequencing is highly suitable for cases where linkage intervals had been established, but the causal variant could not be found.

For some Mendelian disorders where the classic linkage study design was unsuitable and a candidate gene approach was not possible or unsuccessful (in very small families, in a few unrelated cases from different families, or even in sporadic cases), exome sequencing also offers opportunities. As there is no candidate genomic locus known, choices have to be made regarding the variants that deserve follow-up work, and proof of pathogenicity has to be obtained. This will require even closer collaboration between clinical and molecular geneticists. Segregation studies are often used to establish co-inheritance with the putative causal variant. For non-coding genomic variants, genomic information alone is not enough to prove pathogenicity. Sequencing RNA isolated from the same blood samples used to extract DNA may provide an attractive approach. Functional studies on RNA and protein (animal models, *in vitro* assays) will become even more important.

Interpreting new findings and translating these to practical healthcare remains a challenge. In addition to an increase in sequence variants, there is currently little experience in interpreting small structural variants, and the interpretation of somatic changes on a large scale. A number of studies have successfully applied WES to study effects of somatic mutations and/or epigenetic regulation and mosaicism, especially in cancer development [16-18].

With the discovery of new SNVs and CNVs in the DNA diagnostic laboratories, the need for practical guidelines on interpretation of genomic variation increased. Both molecular geneticist, cytogeneticists and clinical geneticists are learning to deal with the benefits and drawbacks of applying new techniques in daily clinical practice. Communicating the consequences of unclassified variants is often ambiguous, and efforts should be made to obtain reasonable certainty on the clinical consequences of the variant. Database information (e.g. LOVD, OMIM, Decipher, DGV [19-23]) outcomes, or even suggestions for further studies (e.g. functional studies on RNA or protein, segregation studies) and *in silico* predictions on functional consequences of the variant are useful information to include in the analysis report. Communication of unclassified variants requires special attention because the information given is often not completely understood or misinterpreted by patients and their family [24].

Unanticipated chance findings that are medically relevant (incidental findings) will provide further challenges for counselling. In general, it is very difficult to discuss the disclosure of particular findings to the counselee beforehand, as the spectrum of possibilities is immense. However, as exome sequencing is making its way in the diagnostic field, improved guidelines

for proper informed consent are being developed. One way to minimize these challenges is to analyse only known genes or a disease-specific gene set. This, however, means missing an opportunity to find novel disease loci or make a different diagnosis, especially in clinically and genetically heterogeneous disorders.

An additional formidable challenge is correlating low frequency sequence variants and CNVs to multifactorial diseases such as intellectual disability, autism, schizophrenia, and many other disorders. For complex disorders, Genome-Wide Association Studies (GWAS) have aimed to associate common variants to common diseases by studying large case-control cohorts. However, most of these associated SNVs have a very small effect, and confer a small risk to disease [25]. Identification of all variants in an individual with a common disease, for example autism, may reveal more rare functional variants that have a strong impact on disease, sometimes within a protein network [26]. On the other hand, many of these risk variants are not recurrent but *de novo*. Thus, both *de novo* and extremely rare inherited SNVs and CNVs contribute together to the overall genetic risk. This means that disorders with extreme genetic heterogeneity require the analysis of very large numbers of clinically well-defined patients and extensive sequence data for validation, because the study of individual families will not answer which combinations may or may not be pathogenic [26, 27].
It is currently impossible to predict the outcome of a pregnancy of a foetus carrying a CNV/SNV of this type. Since microarray analysis is increasingly used also in the prenatal setting, such findings create difficult problems in genetic counselling.
In addition, by extending family studies of these apparent strong risk factors for congenital malformation and intellectual disability we will identify young healthy carriers who will want to know the risk of disability in their offspring, and will ask for accurate prenatal diagnosis. The knowledge gained from studying Mendelian disorders and complex diseases will eventually complement each other and aid in genetic counselling.

Databases

Studies on healthy and disease cohorts have identified many variants, which are stored in several databases to enhance clinical interpretation of unclassified variants (SNVs, CNVs). Accumulation of unclassified variants led to the need of dedicated locus specific databases (LSDBs) to enhance clinical interpretation. For all disorders studied in this thesis (KFSD-*MBTPS2*, TOD-*FLNA*, Aarskog-Scott syndrome-*FGD1*, Coffin Siris syndrome-*ARID1B*), LSDBs have been set-up and are publicly available. Variant information can be stored and easily queried by users worldwide, and are subject to editorial screening by curators. In theory, storage in databases available on the internet is effective. In practice, however, a large number of different databases have emerged, leading to the lack of centralised

data storage. Data on genomic variation becomes scattered, depending on where the user decides to upload their experimental data. This could be solved by reaching definite (international) agreements on sharing of variants by laboratories, and honouring the existing commitments. Larger and more general data repositories (including NCBI and UCSC) have requested of locus-specific variant databases (LSDBs) to share their data [28]. The HGMD represents an attempt to collate known gene lesions responsible for human inherited disease [29]. HGMD also includes some variant data from those LSDBs that are in the public domain. Researchers and diagnosticians, genetic counsellors and physicians can use this information of practical diagnostic importance. The need for large, trustworthy databases will only increase when the use of next generation sequencing for diagnostic purposes gains in popularity.

Conclusion

This thesis outlines the rapid development of molecular techniques for detecting genomic variation. It describes advantages and disadvantages of specific techniques, and shows their application in the search for pathogenic variants in patients with genetic disorders. Next generation sequencing brings genetics into a new era of identifying sequence and structural variants in the context of disease. Both the techniques itself and the associated analysis tools will continue to improve in the future, eventually resulting in an approach that can detect all genomic variation in one assay. Every technical advance in genetic analysis has revealed a new level of variation within our genome i.e. initially chromosomal variation, then copy number changes of decreasing size, down to nucleotide level. As further methodological developments can be anticipated, it is reasonable to assume that there are additional levels of genomic complexity still to be revealed. Through identification of new disease genes, the NGS techniques will increase our knowledge of the molecular pathogenesis of genetic disorders. Interpreting these data and translating findings to improve healthcare will be a great challenge. The availability of new techniques for genomic analysis, and the associated benefits for patients, should be balanced against each patient's right not to know what their genome contains. Collaboration between research and diagnostics is pivotal for determining the most appropriate protocols for returning information to counselees. With NGS, the focus in clinical genetics the coming years will shift from identification to interpretation of variants, requiring more follow-up work. Further technical advances are necessary to enable investigation of other genetic mechanisms, including somatic mutations, epigenetic dysregulation and polygenic disruption, each of which is likely to account for a significant proportion of disease.

1. Sun Y, Almomani R, Aten E, Celli J, van der HJ, Venselaar H, Robertson SP, Baroncini A, Franco B, Basel-Vanagaite L, Horii E, Drut R, Ariyurek Y, den Dunnen JT, and Breuning MH (2010) Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. Am J Hum Genet 87 (1):146-153

2. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, and Shendure J (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42 (9):790-793

3. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, and Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42 (1):30-35

4. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van LB, Steehouwer M, van RJ, Kant SG, Roepman R, Knoers NV, Veltman JA, and Brunner HG (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet 87 (3):418-423

5. Hoischen A, van Bon BW, Gilissen C, Arts P, van LB, Steehouwer M, de VP, de RR, Wieskamp N, Mortier G, Devriendt K, Amorim MZ, Revencu N, Kidd A, Barbosa M, Turner A, Smith J, Oley C, Henderson A, Hayes IM, Thompson EM, Brunner HG, de Vries BB, and Veltman JA (2010) De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet 42 (6):483-485

6. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, and Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461 (7261):272-276

7. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, and Akey JM (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337 (6090):64-69

8. Do R, Kathiresan S, and Abecasis GR (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet 21 (R1):R1-R9

9. Bolouri A (2009) personal genomics and personalized medicine.

10. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, and Jabado N (2011) What can exome sequencing do for you? J Med Genet 48 (9):580-589

11. Albers CA, Lunter G, Macarthur DG, McVean G, Ouwehand WH, and Durbin R (2011) Dindel: accurate indel calls from short-read data. Genome Res 21 (6):961-973

12. Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25 (21):2865-2871

13. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, and Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318 (5849):420-426

14. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, and Bamshad MJ (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42 (1):30-35

15. Gilissen C, Hoischen A, Brunner HG, and Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20 (5):490-497

16. Yan XJ, Xu J, Gu ZH, Pan CM, Lu G, Shen Y, Shi JY, Zhu YM, Tang L, Zhang XW, Liang WX, Mi JQ, Song HD, Li KQ, Chen Z, and Chen SJ (2011) Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet 43 (4):309-315

17. Grossmann V, Tiacci E, Holmes AB, Kohlmann A, Martelli MP, Kern W, Spanhol-Rosseto A et al (2011) Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. Blood 118 (23):6153-6163

18. Vissers LE, Fano V, Martinelli D, Campos-Xavier B, Barbuti D, Cho TJ, Dursun A, Kim OH, Lee SH, Timpani G, Nishimura G, Unger S, Sass JO, Veltman JA, Brunner HG, Bonafe L, Dionisi-Vici C, and Superti-Furga A (2011) Whole-exome sequencing detects somatic mutations of IDH1 in metaphyseal chondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA). Am J Med Genet A 155A (11):2609-2616

19. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van VS, Moreau Y, Pettett RM, and Carter NP (2009) DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am J Hum Genet 84 (4):524-533

20. Swaminathan GJ, Bragin E, Chatzimichali EA, Corpas M, Bevan AP, Wright CF, Carter NP, Hurles ME, and Firth HV (2012) DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. Hum Mol Genet 21 (R1):R37-R44

21. Database of Genomic Variants. Http://projects.tcag.ca/variation/

22. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, and den Dunnen JT (2011) LOVD v.2.0: the next generation in gene variant databases. Hum Mutat 32 (5):557-563

23. Online Mendelian Inheritance in Man. www.omim.org

24. Vos J, Gomez-Garcia E, Oosterwijk JC, Menko FH, Stoel RD, van Asperen CJ, Jansen AM, Stiggelbout AM, and Tibben A (2010) Opening the psychological black box in genetic counseling. The psychological impact of DNA testing is predicted by the counselees' perception, the medical impact by the pathogenic or uninformative BRCA1/2-result. Psychooncology
25. Cirulli ET and Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11 (6):415-425

26. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, and Eichler EE (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485 (7397):246-250

27. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, Moreno-De-Luca D, Chu SH et al (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70 (5):863-885

28. den Dunnen JT, Sijmons RH, Andersen PS, Vihinen M, Beckmann JS, Rossetti S, Talbot CC, Jr., Hardison RC, Povey S, and Cotton RG (2009) Sharing data between LSDBs and central repositories. Hum Mutat 30 (4):493-495

29. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, and Cooper DN (2009) The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Hum Genomics 4 (2):69-72


T C C G A G T T C C C T G G A
T C C G A G T T C C C T G G A
G T T C C T T C C G A G T T C
A A T G A G A A T C C G C C G
G T G A G A G C C C C G T C T
G T A C C T A C T G A G T T C
G A G T T A T G G T T C C T T
T C C G A G T T G T A A T T T
A A A T T T G A A A T C T G G
T G C C T A C T G A G T T C C
G T G C C T A C T G A G T T C
G A G C C C C G T C T G G T A
G T T C C T T C C G A G T T C
G G T T C C T T C C G A G T T
T T C C T T C C G A C T T C C

Summary

Variation in structure and composition of the DNA are found throughout our genome. All types of variation are collectively called 'genomic variation'. Identification and analysis of genomic variation is important to distinguish neutral variants ('non pathogenic') from variants involved in disease ('pathogenic'). Identification of new disease genes will increase our knowledge of the molecular pathogenesis of genetic disorders. Every technical advance in genetic analysis has revealed new levels of variation, ranging from single nucleotide differences to full chromosome changes. Chromosome changes larger than 5-10 Mb in size can be detected by conventional karyotyping. Other techniques, notably Fluorescent in situ hybridisation (FISH), Array comparative genome hybridisation (Array-CGH), and SNP-arrays but also amplification methods such as Multiplex ligation-dependent probe amplification (MLPA) allow the detection of aberrations below the resolution of conventional karyotyping. These submicrosopic changes are called copy number variations (CNVs). High resolution melting curve analysis (HRMA) and Sanger sequencing are used to study alterations at the single nucleotide level (sequence variation). The application of new technologies, whole exome sequencing (WES) and whole genome sequencing (WGS) aid in bridging the technical divide between quantitative (CNV) and qualitative DNA differences (sequence variants). As new DNA methods are applied, increasing numbers of variants with unclear significance to disease (UVs) are identified and choices have to be made regarding the variants that deserve follow-up work. When the pathogenic consequence of a variant is unclear, the effect has to be studied in detail at other levels (functional studies, RNA studies, in silico analysis tools, and databases). This research described in this thesis outlines the rapid development and application of molecular techniques for detecting (pathogenic) genomic variation in the context of genetic disorders.

Chapter 1 provides a general introduction to human genetics and the different types of genomic variation, in relation to health and disease. An overview of cytogenetic and molecular techniques was given, and the advantages and disadvantages of several techniques were discussed. Options for interpretation and classifications of variants are summarized.

In **Chapter 2**, methods to detect CNVs in the human genome are reviewed. A 1400-plex targeted CNV assay was designed to study 320 patients with intellectual disability. In 9% of the analysed cases, a pathogenic CNV was found. Furthermore, HRMA was used as a new method to confirm the presence of CNVs detected using SNP-arrays.

In **Chapter 3**, Array-CGH and SNP-array were used to delineate a *de novo* chromosome 19p deletion in a patient with intellectual disability, Split Hand Foot Malformation (SHFM) and Tetralogy of Fallot. In addition to this case study, MLPA was used to study the presence of copy number changes in genes within the deletion interval in 21 other SHFM patients. Based on its function, EPS15L1 was postulated as a candidate gene for Split-Hand-Foot-Malformation (SHFM).

In the following chapters **(4 and 5)** several applications of HRMA are discussed and show the succes of HRMA as a presequence screening technique. HRMA was used to identify MBTPS2 as the causative gene for Keratosis Follicularis Spinulosa Decalvans (KFSD). In three families (Dutch, American and English), a missense mutation was identified. Functional studies have demonstrated this mutation to be pathogenic. Studying X-inactivation (XCI), differences in allelic expression could be correlated to the clinical phenotype of carrier females. Mutations in MBTPS2 have also been described in Ichthyosis Follicularis Atrichia Photophobia (IFAP) syndrome, a syndrome that shows clinical overlap with KFSD.

Chapter 6 describes three families with Terminal Osseous Dysplasia (TOD) where X-exome sequencing was applied to identify the causative gene, FLNA, in the linkage interval on Xq27.3-q28. The missense variant was predicted to affect splicing. Initial RNA analysis in patient-derived fibroblasts did not detect expression of the mutant allele. Additional RNA analysis in 15-year-old fibroma tissue showed alternative splicing confirming activation of a cryptic splice site, and results in an in-frame deletion. To study the pathogenetic mechanism of the FLNA mutation in TOD, a 3D protein model was built indicating that the deleted region affects or prevents important protein-protein interactions. XCI patterns in patients were established and suggest early skewing with preferential inactivation of the mutant allele is key to disease development.

Chapter 7 illustrates the use of WES to identify the causative gene in a family with Aarskog-Scott syndrome. A branch point variant (-35 bp intronic) in FGD1 was identified that was missed with conventional Sanger sequencing. RNA analysis confirmed an effect on splicing. This work shows, although WES is not specifically designed to do so, deeper intronic variants can be detected.

In **Chapter 8** mutations in the SWI/SNF chromatin remodelling complex ARID1B were identified as the genetic cause of Coffin-Siris syndrome (CSS). One case-parent trio and two sporadic patients were studied according to a recessive inheritance model using WES. After initial analysis did not reveal causative variants, a dominant inheritance model identified De novo truncating mutations in all patients. By excluding an autosomal recessive inheritance pattern for CSS, the recurrence risk for parents with a child diagnosed with CSS could be reduced to 1-2%. Mutations in other genes affecting the SWI/SNF complex have been published in CSS patients (Tsurusaki et al, Nat Genet. 2012; 44(4):376-378), in other syndromes (Clapier et al, Annu Rev Biochem. 2009;78:273-304) and are associated with tumorigenesis (Wilson et al, Nat Rev Cancer. 2011;11(7):481-92 and Wang et al, Nat Genet. 2011;43(12):1219-23), indicating that mutations in chromatin remodelling factors contribute significantly to human diseases.

Finally, in **Chapter 9** the evolution of techniques, strategies for gene identification and variant interpretation are discussed.


T C C G A G T T C C C T G G A
T C C G A G T T C C C T G G A
G T T C C T T C C G A G T T C
A A T G A G A A T C C G C C G
G T G A G A G C C C C G T C T
G T A C C T A C T G A G T T C
G A G T T A T G G T T C C T T
T C C G A G T T G T A A T T T
A A A T T T G A A A T C T G G
T G C C T A C T G A G T T C C
G T G C C T A C T G A G T T C
G A G C C C C G T C T G G T A
G T T C C T T C C G A G T T C
G G T T C C T T C C G A G T T
T T C C T T C C G A C T T C C

Nederlandse samenvatting

Ons genoom toont variatie in structuur en samenstelling van het DNA. Alle soorten variatie tezamen worden 'genomische variatie' genoemd. De identificatie en analyse van genomische variatie is belangrijk om neutrale varianten ('niet-pathogeen') te onderscheiden van varianten die een rol spelen bij ziekten ('pathogeen'). Identificatie van nieuwe 'ziekte-genen' zal onze kennis van de moleculaire pathogenese van genetische aandoeningen vergroten. Iedere technische vooruitgang in de genetische analyse openbaarde nieuwe niveaus van variatie, uiteenlopend van enkelvoudige nucleotide verschillen tot volledige veranderingen in het chromosoom. Veranderingen in het chromosoom die groter zijn dan 5-10 Mb zijn op te sporen met de gebruikelijke karyotypering. Andere technieken maken het mogelijk om kleinere afwijkingen te ontdekken, die zich onder de resolutie van de conventionele karyotypering bevinden. Dit betreft met name Fluorescent in situ hybridisation (FISH), Array comparative genome hybridisation (Array-CGH) en SNP-arrays, maar ook andere methoden zoals Multiplex ligation-dependent Probe Amplification (MLPA). Deze submicroscopische veranderingen heten CNV's: copy number variations. High resolution melting curve analysis (HRMA) en Sanger sequencing zijn ontwikkeld om nucleotide veranderingen (sequence variatie) te detecteren. Toepassing van nieuwe technologieën als Whole-exome sequencing (WES) en whole genome sequencing (WGS) helpen om de technische leemte te overbruggen tussen kwantitatieve (CNV) en kwalitatieve (sequentie-varianten) DNA-verschillen.
De toepassing van nieuwe DNA-methoden resulteert in toenemende mate in de ontdekking van varianten waarvan de betekenis voor ziekten onduidelijk is ('unclassified variants') en dit vraagt om het maken van keuzes in de varianten die vervolgstudie verdienen. Als er onduidelijkheid bestaat over de pathogene consequentie van een variant moet het effect in detail worden bestudeerd op andere niveaus (functionele studies, RNA-studies, in silico analyses en databases).
Dit promotieonderzoek beschrijft de ontwikkeling en de toepassing van moleculaire technieken voor het ontdekken van (pathogene) genomische variatie in de context van genetische aandoeningen.

Hoofdstuk 1 voorziet in een algemene introductie in humane genetica en de verschillende typen van genomische variatie in relatie tot ziekte en gezondheid. Het geeft een overzicht van cytogenetische en moleculaire technieken en bespreekt de voordelen en de nadelen van verschillende technieken. Opties voor interpretatie en classificatie van varianten zijn toegevoegd.

Hoofdstuk 2 bespreekt methoden om CNV's in het menselijk genoom te vinden. Voor de bestudering van 320 patiënten met verstandelijke beperking werd een specifieke 1400-

plex array ontwikkeld. In 9% van deze populatie werd een pathogene CNV gevonden. Bovendien werd HRMA gebruikt als nieuwe methode om CNVs te bevestigen, die eerder met arrays gevonden waren.

Hoofdstuk 3 rapporteert over het gebruik van Array-CGH en SNP-array voor de beschrijving van een *de novo* chromosoom 19p-deletie bij een patiënt met verstandelijke beperking, Split Hand Foot Malformation (SHFM) en Tetralogie van Fallot (hartafwijking). Als aanvulling op deze casus werden 21 andere SHFM-patiënten bestudeerd met MLPA om de aanwezigheid van CNVs te onderzoeken in genen binnen het deletie-interval. EPS15L1 werd op grond van zijn functie aangemerkt als kandidaat-gen voor Split Hand Foot Malformation (SHFM).

De hoofdstukken 4 en 5 bespreken verscheidene toepassingen van HRMA en demonstreren het succes van HRMA als pre-sequencing-techniek. Met behulp van HRMA werd gevonden dat mutaties in het MBTPS2-gen verantwoordelijk zijn voor het ontstaan van Keratosis Follicularis Spinulosa Decalvans (KFSD). In drie families (Nederlands, Amerikaans en Engels) werd een missense mutatie geïdentificeerd. Functionele studies toonden aan dat dit een pathogene mutatie is. Bij bestudering van de X-inactivatie konden verschillen in allelische expressie gekoppeld worden aan het klinische phenotype van vrouwelijke dragers. Mutaties in MBTPS2 zijn ook beschreven in het Ichthyosis Follicularis Atrichia Photophobia (IFAP) syndroom, dat klinisch overlap toont met KFSD.

Hoofdstuk 6 beschrijft drie families met Terminal Osseous Dysplasia (TOD), waarbij X-exoom sequencing werd toegepast om de genetisch oorzaak voor deze aandoening op te sporen. In voorgaande studies werd een locatie gevonden op het X-chromosoom (Xq27.3-q28) waarin het veroorzakende gen zou kunnen liggen. Met behulp van WES werd een missense-variant gevonden in het FLNA-gen. De voorspelling was dat deze variant splicing zou veroorzaken. Aanvankelijk toonde RNA-analyse in huidcellen van de patiënt geen expressie van het mutante allel. Echter, aanvullende RNA-analyse in 15 jaar oud fibroomweefsel bevestigde de activering van een cryptische splice-site, resulterend in een deletie, die het leesraam niet verstoort. Om het pathogenetische mechanisme van de FLNA-mutatie in TOD te bestuderen, werd een 3D-eiwit model gebouwd. Hieruit bleek dat de deletie leidt tot het verlies van aminozuren die belangrijk zijn voor de eiwitstructuur en dat dit mogelijk belangrijke eiwit-eiwit-interacties beïnvloedt. In de patiënten en gezonde familieleden werd naar hun X-inactivatie patroon gekeken. Hierbij werd vastgesteld dat in patiënten het X-chromosoom met het mutante allel volledig geïnactiveerd was, terwijl gezonde familieleden een willekeurige inactivatie van een van de twee X-chromosomen toonden. Dit suggereert dat er al vroeg in de ontwikkeling selectie plaatsvindt tegen

cellen met expressie van een mutant FLNA-eiwit en dat dit een belangrijke rol speelt bij de ontwikkeling van de ziekte.

Hoofdstuk 7 belicht het gebruik van WES bij het opsporen van de genetische oorzaak in een familie met het Aarskog-Scott-syndroom. Hierbij werd een branchpoint-mutatie in het FGD1-gen gevonden (-35 bp intronisch), die met Sanger sequencing in een diagnostisch laboratorium niet werd gevonden. RNA-analyse bevestigde een effect op splicing. Dit werk laat zien dat WES geschikt is om diepere intronische varianten te detecteren, hoewel de techniek daarvoor niet specifiek ontworpen is.

Hoofdstuk 8 beschrijft de ontdekking van mutaties in het ARID1B-gen, een SWI/SNF chromatin-remodelling complex, als de genetische oorzaak van het Coffin-Siris-syndroom (CSS). Eén casus van een trio (ouders+kind) en twee sporadische patiënten werden met WES getest, uitgaande van een autsomaal recessief overervingsmodel. Nadat initiële analyse geen relevante varianten opleverde, werd een dominant overervingsmodel toegepast. Hierbij werden *de novo* truncerende mutaties gevonden in alle patiënten. Door aan te tonen dat CSS geen autosomaal recessief overervende aandoening is, kon het herhalingsrisico voor ouders op een volgend kind met CSS worden teruggebracht tot 1-2%. Ook in andere CSS-patiënten, in patiënten met andere syndromen en voor bepaalde vormen van kanker zijn mutaties beschreven in genen die betrokken zijn bij SWI/SNF chromatin remodelling. Dit wijst erop dat mutaties in chromatine-remodelling factoren een significante bijdrage leveren aan menselijke ziekten.

In **hoofdstuk 9** wordt de ontwikkeling van technieken bediscussieerd met de daarmee gepaard gaande strategieën om ziektegenen op te sporen en de betekenis van varianten te interpreteren.


T C C G A G T T C C C T G G A
T C C G A G T T C C C T G G A
G T T C C T T C C G A G T T C
A A T G A G A A T C C G C C G
G T G A G A G C C C C G T C T
G T A C C T A C T G A G T T C
G A G T T A T G G T T C C T T
T C C G A G T T G T A A T T T
A A A T T T G A A A T C T G G
T G C C T A C T G A G T T C C
G T G C C T A C T G A G T T C
G A G C C C C G T C T G G T A
G T T C C T T C C G A G T T C
G G T T C C T T C C G A G T T
T T C C T T C C G A C T T C C

List of publications

Aten E, Sun Y, Almomani R, Santen GWE, Messemaker T, Maas SM, Breuning MH, den Dunnen JT.
Exome sequencing identifies a branch point variant in Aarskog-Scott syndrome.
Human Mutation 2012 (in press)

Aten E, Brasz LC, Bornholdt D, Hooijkaas IB, Porteous ME, Sybert VP, Vermeer MH, Vossen RH, van der Wielen MJ, Bakker E, Breuning MH, Grzeschik KH, Oosterwijk JC, den Dunnen JT.
Keratosis Follicularis Spinulosa Decalvans is caused by mutations in MBTPS2.
Human Mutation. 2010;31(10):1125-33.

Aten E, den Hollander N, Ruivenkamp C, Knijnenburg J, van Bokhoven H, den Dunnen J, Breuning M.
Split hand-foot malformation, tetralogy of Fallot, mental retardation and a 1 Mb 19p deletion-evidence for further heterogeneity?
Am J Med Genet A. 2009;149A(5):975-81.

Aten E, White SJ, Kalf ME, Vossen RH, Thygesen HH, Ruivenkamp CA, Kriek M, Breuning MH, den Dunnen JT. *Methods to detect CNVs in the human genome.*
Cytogenet Genome Res. 2008;123(1-4):313-21.

Sun Y, Almomani R, Breedveld G, Santen GWE, Aten E, Lefeber DJ, Hoff JI, Brusse E, Verheijen FW, Verdijk RM, Kriek M, Oostra B, Breuning MH, Losekoot M, denDunnen JT, van de Warrenburg BP, Maat-Kievit AJA.
Autosomal recessive spinocerebellar ataxia 7 (SCAR7) is caused by variants in TPP1, the gene involved in late-infantile neuronal ceroid lipofuscinosis (CLN2).
Submitted

Almomani R, Yu Sun Y, Aten E, Hilhorst-Hofstee Y, Peeters-Scholte CMPCD, van Haeringen A, Hendriks YMC, den Dunnen JT, Breuning MH, Kriek M, Santen GWE.
GPSM2 and Chudley-McCullough Syndrome: a Dutch founder variant brought to North America.
Am J Med Genet A 2012 (*in press*)

Nielsen M, Vermont CL, Aten E, Ruivenkamp CA, van Herrewegen F, Santen GW, Breuning MH. *Deletion of the 3q26 region including the EVI1 and MDS1 genes in a neonate with congenital thrombocytopenia and subsequent aplastic anaemia.*
J Med Genet. 2012;49(9):598-600.

Santen GW, Aten E, Sun Y, Almomani R, Gilissen C, Nielsen M, Kant SG, Snoeck IN, Peeters EA, Hilhorst-Hofstee Y, Wessels MW, den Hollander NS, Ruivenkamp CA, van Ommen GJ, Breuning MH, den Dunnen JT, van Haeringen A, Kriek M.
Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome.
Nature Genetics. 2012;44(4):379-80.

Sun Y, Almomani R, Aten E, Celli J, van der Heijden J, Venselaar H, Robertson SP, Baroncini A, Franco B, Basel-Vanagaite L, Horii E, Drut R, Ariyurek Y, den Dunnen JT, Breuning MH.
Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene.
Am J Hum Genet. 2010;87(1):146-53.

Vossen RH, Aten E, Roos A, den Dunnen JT.
High-resolution melting analysis (HRMA): more than just sequence variant screening.
Human Mutat. 2009;30(6):860-6.

Herbert MA, Beveridge CJ, McCormick D, Aten E, Jones N, Snyder LA, Saunders NJ.
Genetic islands of Streptococcus agalactiae strains NEM316 and 2603VR and their presence in other Group B streptococcal strains.
BMC Microbiol. 2005;5:31.


T C C G A G T T C C C T G G A
T C C G A G T T C C C T G G A
G T T C C T T C C G A G T T C
A A T G A G A A T C C G C C G
G T G A G A G C C C C G T C T
G T A C C T A C T G A G T T C
G A G T T A T G G T T C C T T
T C C G A G T T G T A A T T T
A A A T T T G A A A T C T G G
T G C C T A C T G A G T T C C
G T G C C T A C T G A G T T C
G A G C C C C G T C T G G T A
G T T C C T T C C G A G T T C
G G T T C C T T C C G A G T T
T T C C T T C C G A C T T C C

Dankwoord

409 maanden nadat ik het eerste levenslicht zag is het zover: het is voltooid, verricht, uitgevoerd, afgerond, beëindigd, afgemaakt, vervuld, tot stand gebracht, bewerkstelligt, 'gecheft'. Mission accomplished.

Velen hebben hierbij een rol gespeeld:

Maanden 331-409

Martijn, als hoofd van de afdeling Klinische Genetica, en als mijn promotor, heb je mij de mogelijkheid geboden om onderzoek te doen en mij als wetenschapper te ontwikkelen. Daarnaast heb ik de kans gekregen om ook mijn klinische ervaring uit te breiden als klinisch geneticus in opleiding. Bedankt voor de ondersteuning, je input en flexibiliteit in dit traject. Johan, als mijn tweede promotor, heb je mij een plek geboden in het lab en mij kennis laten maken met veel nieuwe technieken. Het oppakken van het onderzoek naar KFSD was jouw idee en heeft mij, naast veel plezier, een mooie publicatie opgeleverd. Bedankt voor al je tips en ideeën.

Bert Bakker, Michiel van der Wielen en Jan Oosterwijk, ook jullie wil ik graag bedanken voor al het (reeds) verrichte onderzoek dat nodig was om het gen voor KFSD te vinden.

Yu and Rowida, thanks for the pleasant collaboration. As part of the Breuning research-group, I think we've stuck together and share some really nice papers because of this.

All colleagues from Lab-J, you have all contributed tot his thesis in one way or the other. There has always been support for presentations, help in the lab, useful work discussions and social events. All (former) fellow-PhD students thanks for your support. Especially Maaike, Dwi, Pietro, Herman and Eddy. Thank you for your friendship.

I'd like to thank my Lab-J room mates: Matt, Ivo and Michel for the beers, for trying to introduce met to "pearle" and linux (but failed to convince me) and many other things. Annemieke (and later Isabella), thanks for 'adopting' me. Naast het feit dat ik veel van jullie heb geleerd, was het erg gezellig. Huntington-groupers (Willeke, Menno, Tassos, Barry & Melvin (master of pixels), bedankt voor het jarenlang delen van het lab en het bijbrengen van inventieve western blot technieken. Peter-Bram, de reis naar Kentucky na het ASHG was een uitdaging, maar hebben we precies gered en gaf ons mooi de tijd om (bij) te praten. Dank voor alles!

LGTC-ers, bedankt voor jullie hulp bij diverse zaken. Yavuz en Rolf ,fijn dat ik altijd bij jullie terecht kon met technische vragen (en niet-technische) en acute tekorten aan platen en reagentia.

Margot, Stefan en Marjolein. Jullie hebben me als "de nieuwe Stefan" meteen op weg geholpen met de experimenten zoals MLPA en de MR-goldengate array en hebben de basis gelegd voor de eerste publicaties. Stefan, special thanks to you for your useful comments, correcting my english and your great sense of humour. Marjolein, super dat jij mijn paranimf wil zijn. Het is heel waardevol om zo'n positieve, gezellige collega te hebben die de ins- and outs, up's and down's van promoveren kent. Daarnaast ben je natuurlijk gesequenced en dan tel je pas echt in de wereld van genetica. Ik hoop dat we nog lang kunnen samenwerken!

Alle collega's van de kliniek, ook jullie input en hulp is zeer belangrijk geweest voor de totstandkoming van dit proefschrift. Jullie enthousiasme voor de patiëntenzorg is inspirerend en ik hoop dat we samen in de toekomst klinische kennis en research kunnen blijven combineren. Lieve mede-AIOS; de meesten van jullie weten hoe het is om je opleiding te combineren met het voltooien van een proefschrift. Dank voor jullie steun. Dietje, het was fijn om een kamergenoot te hebben die in hetzelfde schuitje zat! Marije, pita omonia says it all. Gijs, dank voor al je (exome) inzet, het heeft een prachtig Nature Genetics artikel opgeleverd. Ik hoop dat we ook in de toekomst gezamenlijke doelen kunnen realiseren.

Staf (i.o.) en analisten van het LDGA, bedankt voor de samenwerking. Ook jullie hebben op allerlei manieren bijgedragen aan de totstandkoming van een aantal artikelen.

Mijn studenten Ingeborg, Tobias en Lisa wil ik bedanken voor hun inzet. Al hun werk heeft bijgedragen aan mijn publicaties.

Maanden 276-409

Lieve Kirsten (Chica), heel erg leuk dat jij vandaag mijn paranimf bent! Dank voor je humor, interesse, gesprekken (medisch en niet-medisch) en overall enthousiasme.

Maanden 181-409

DFN (Anne-Marleen, Martine, Annemarie, Marieke, Marsha, Gabrielle en Janneke), ontzettend bedankt voor alle gezellige etentjes, jullie verhalen en interesse in "mijn avonturen"!

Maanden 0-409

Pap, mam, Jan Joost (& Marjolein +2), Niels (& Annemieke). Dank voor jullie belangstelling, steun, motiverende woorden en hulp in al deze 409 maanden. Mam, extra dank voor je inzet en input voor het perfectioneren van mijn proefschrift! Josje, bedankt voor je 409 maanden vriendschap (+2 dagen (eigen)wijsheid). Marinde, 313 maanden belevenissen, en ja.. "dat boekje" is nu eindelijk af. F.F.!

Maanden 288-409

Thea en familie van Meggelen, jullie hebben altijd naar de voortgang van mijn proefschrift geïnformeerd, dank!

Lieve Bassy, jij hebt alle jaren onderzoek van dichtbij meegemaakt, met bijbehorende successen (live vanuit Washington) en dieptepunten. Jouw zorgzaamheid, geduld, luisterend oor èn kritische blik zijn onmisbaar!–x-


T C C G A G T T C C C T G G A
T C C G A G T T C C C T G G A
G T T C C T T C C G A G T T C
A A T G A G A A T C C G C C G
G T G A G A G C C C C G T C T
G T A C C T A C T G A G T T C
G A G T T A T G G T T C C T T
T C C G A G T T G T A A T T T
A A A T T T G A A A T C T G G
T G C C T A C T G A G T T C C
G T G C C T A C T G A G T T C
G A G C C C C G T C T G G T A
G T T C C T T C C G A G T T C
G G T T C C T T C C G A G T T
T T C C T T C C G A C T T C C

Curriculum Vitae

Emmelien Aten werd geboren op 31 December 1978 in Amersfoort en groeide op in Soest. Ze behaalde haar VWO diploma in 1997 aan het Herman Jordan Lyceum in Zeist. In 1997 begon zij met een studie Biomedische Wetenschappen. Na het behalen van haar propedeuse Biomedische Wetenschappen ging zij tevens Geneeskunde studeren. Tijdens de studie deed zij binnen het Leids Universitair Medisch Centrum (LUMC) onderzoek op de afdelingen Reumatologie en ImmunoHematologie&Bloedbank. Haar afstudeerstage verrichtte zij op de afdeling Neonatologie van het John Radcliffe Hospital te Oxford (GB). In 2004 behaalde zij haar doctoraalexamens Biomedische Wetenschappen en Geneeskunde. Na haar semi-arts-stage op de afdeling Klinische Genetica (LUMC), legde zij in 2006 haar artsexamen af.

Vanaf juli 2006 was zij werkzaam als arts-onderzoeker op de afdeling Humane en Klinische Genetica, onder begeleiding van Prof. Dr. M.H. Breuning en prof.Dr. J.T. den Dunnen. Haar promotieonderzoek betrof het ontwikkelen en toepassen van nieuwe technieken voor opsporing van genetische afwijkingen bij patiënten met een verstandelijke beperking en/of aangeboren afwijkingen. Tijdens haar promotieonderzoek was zij bestuurslid van de Vereniging voor (arts) onderzoekers (VAO) van het LUMC. Sinds 1 Januari 2010 is zij in het LUMC werkzaam als Klinisch Geneticus in opleiding.


T C C G A G T T C C C T G G A
T C C G A G T T C C C T G G A
G T T C C T T C C G A G T T C
A A T G A G A A T C C G C C G
G T G A G A G C C C C G T C T
G T A C C T A C T G A G T T C
G A G T T A T G G T T C C T T
T C C G A G T T G T A A T T T
A A A T T T G A A A T C T G G
T G C C T A C T G A G T T C C
G T G C C T A C T G A G T T C
G A G C C C C G T C T G G T A
G T T C C T T C C G A G T T C
G G T T C C T T C C G A G T T
T T C C T T C C G A C T T C C

List of abbreviations

List of abbreviations

AAS	Aarskog-Scott Syndrome
AD	Autosomal Dominant
AR	Autosomal Recessive
BAC	Bacterial Artificial Chromosome
Bp	Base pair
CGH	Comparative Genome Hybridisation
CNV	Copy Number Variant
CSS	Coffin Siris Syndrome
DECIPHER	DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
ECARUCA	European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations
FISH	Flourescent in situ Hybridisation
FoSTes	Fork Stalling and Template switching
GWAS	Genome Wide Association Studies
HGVD	Human Genome Variation Database
HGVS	Human Genome Variation Society
HRMA	High Resolution Melting analysis
IBD	Identity by Descent
ID	Intellectual Disability
INHDB	In House Database
Kbp	Kilo base pairs (one thousand base pairs)
LSDB	Locus Specific Database
KFSD	Keratosis Follicularis Spinulosa Decalvans
LOVD	Leiden Open Variation Database
Mbp	Mega base pairs (one million base pairs)
MCA	Multiple Congenital Anomalies
MLPA	Multiplex Ligation-dependent Probe Amplification
MR	Mental Retardation
NAHR	Nonallelic Homologous Recombination
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NHEJ	NonHomologous End Joining

OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
PGD	Preimplantation Genetic Diagnosis
RNA	Ribonucleic acid
SHFM	Split Hand Foot Malformation
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variation
TOD	Terminal Osseous Dysplasia
UPD	Uniparental Disomy
UV	Unclassified Variant
WES	Whole Exome Sequencing
WGS	Whole Genome Seqeuncing
XL	X-linked

Addendum: color figures

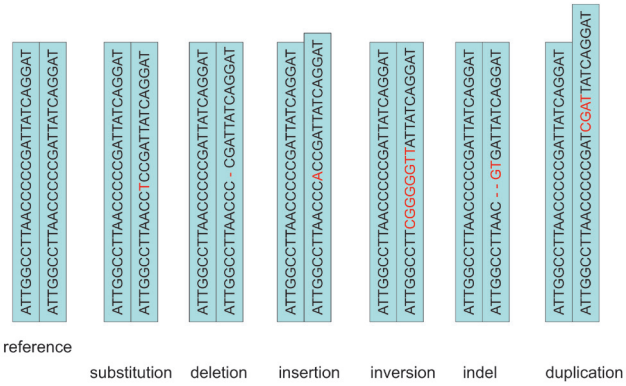
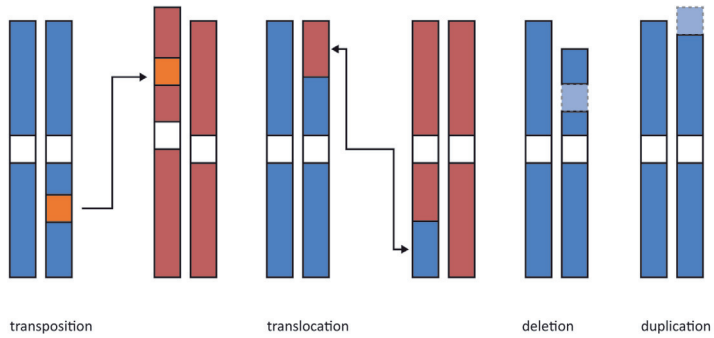


Figure 1a: Types of variation in the human genome (sequence view). These can be single base changes (substitution, deletions, and insertions) or involve larger segments of DNA (deletions, insertions, inversions, indels, and duplications). Adapted from Frazer et al [17]

Figure 1b: Types of variation in the human genome (chromosome view). Examples of structural variation. 1) transposition-transfer of segment of DNA to a new position. 2) translocation-balanced event where two DNA segments are interchanged. 3) Copy number variation: deletion and duplication-loss or gain of DNA segments.



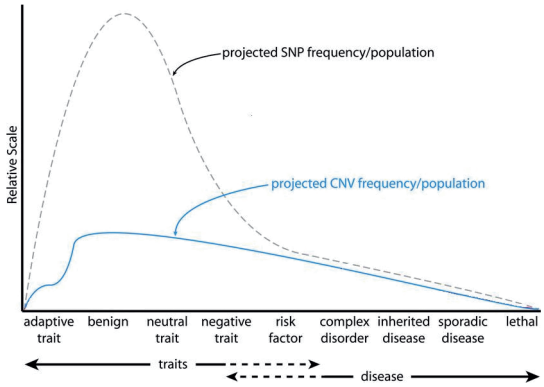


Figure 2: Conceptual curves of SNV and CNV characteristics. Projected frequencies in the population for SNVs (dashed grey) and CNVs (blue) show the relation between genomic variation and a spectrum of phenotypic impact. Adapted from Buchanan et al [28].

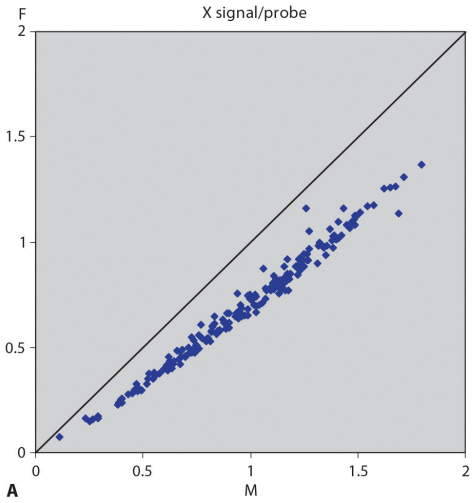


Figure 1: 1400-plex CNV bead assay. **A)** Strength of signal for probes on the X-chromosome obtained in males (M, X-axis) versus females (F, Y-axis). Signals are clearly different yet the difference in F:M signal obtained is not 1:0.5 as expected but 1:~0.7

Figure 2: Fast-MLPA. Detection of a trisomy 21 (top left), Turner (45, X0; top right) and 49, XXXXY case (bottom) using fast-MLPA.

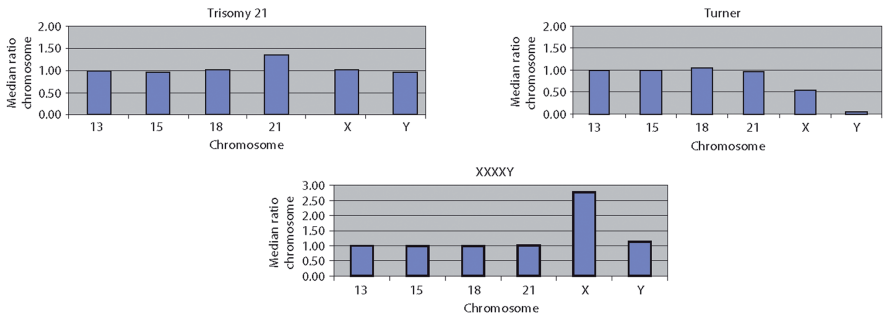


Figure 3: hrMCA CNV. High-resolution Melting Curve Analysis of two DNA samples being homozygous for the opposite SNP-alleles (rs213950:G>A) mixed in different proportions (from 8:0 to 0:8). The difference plots of all mixes can be easily discriminated, giving a sensitivity of at least 12.5%, suggesting that octaploid alleles could be typed.

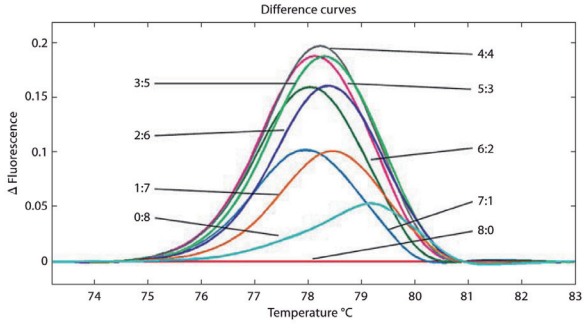


Figure 1: Clinical features of the index patient. (A) Phenotype of the patient (here six years old) with subtle facial features including a wide mouth, fair hair and light skin. Central ray deficiencies as detected at birth. Left foot (B) with a cleft, hypoplasia of the second digit and a nubbin representing the third digit. Right foot (C) with a proximally placed first digit and presence of digit V. Digits II-IV are absent. Right hand (D) with a central cleft leading to absence of the third digit. Digit IV and V show cutaneous syndactyly.

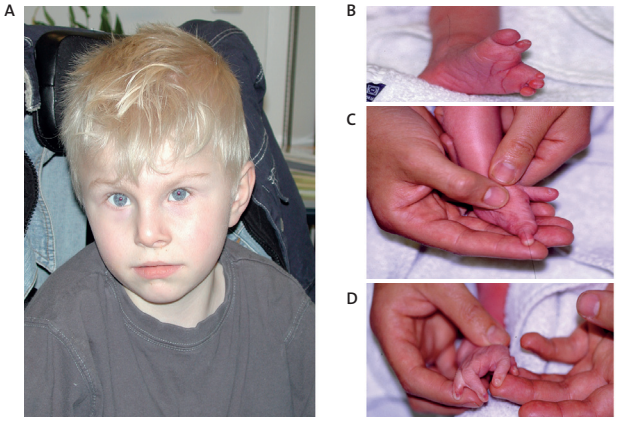


Figure 2: (A) Array-CGH showed one deleted BAC clone (RP11-413M18) on 19p13.11 at 16.8 Mb with a maximal deleted size of 1.44 Mb. **(B)** SNP array refined the deletion from rs10411936 (19: 16,409,375) - rs4808641 (19:17,408,292) with a maximal deleted size of 0.99 Mb

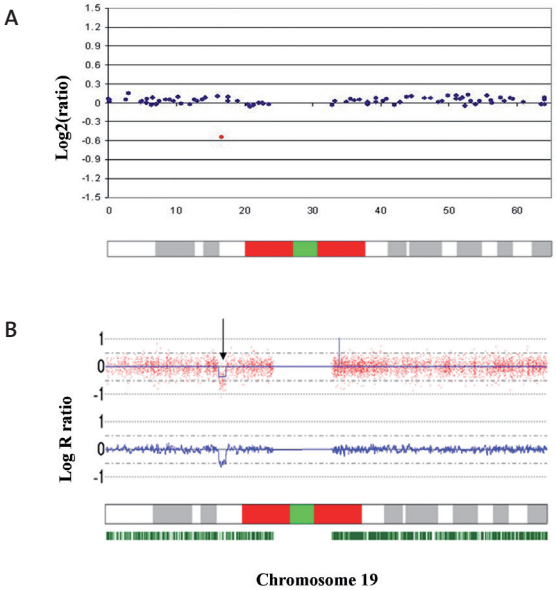
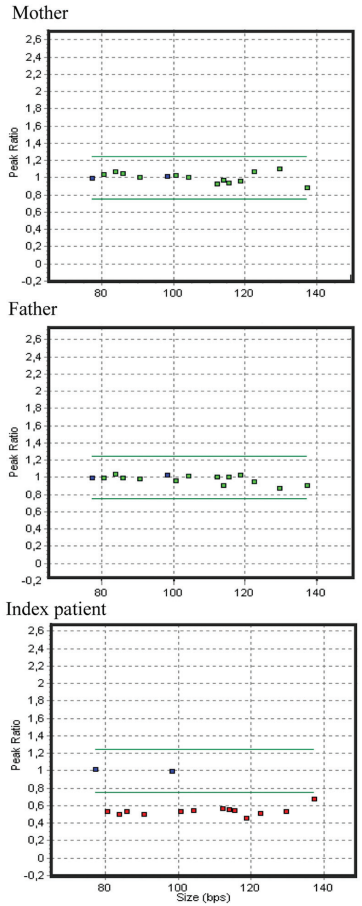
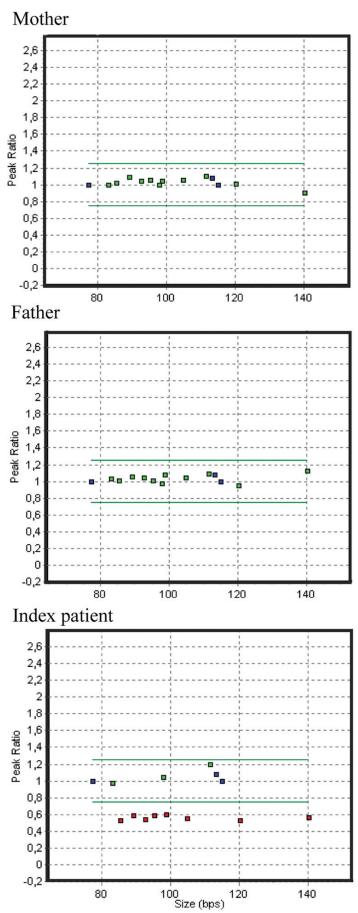


Figure 3: MLPA results for the index patient and his parents. **(A)** Peak ratio plots for the two probe sets, confirming the deletion to be *De Novo*. Control probes are represented by blue squares. **(B)** Peak ratio results for the mother (1), the father (2) and the index patient (3). Peak ratios in the index patient are normal for probes located in KLF2, AP1M1 and SLC27A1, coinciding with the proximal and distal deletion breakpoints.

A



B



		REF -	1	2	3
1	*control1	77.2	0.996	0.989	1.016
2	ANKLE1_1	80.4	1.014	0.970	0.520
3	ANKLE1_2	83.8	1.066	1.037	0.501
4	CPAMD8	86.0	1.044	0.990	0.526
5	EPS15L1	90.6	1.000	0.979	0.500
6	*control2	98.5	1.018	1.026	0.995
7	MY09B	100.7	1.030	0.962	0.534
8	MED26	104.2	1.000	1.018	0.541
9	USHBP1	112.1	0.923	0.999	0.570
10	ABHD8	114.1	0.969	0.904	0.559
11	CALR3	115.6	0.917	0.979	0.538
12	BST2	118.9	0.952	1.012	0.456
13	TMEM38A	122.6	1.071	0.943	0.509
14	C19ORF44	129.8	1.088	0.866	0.529
15	SLC35E1	137.6	0.881	0.898	0.671

		REF -	1	2	3
1	*control1	77.4	1.000	1.000	1.000
2	KLF2	83.4	1.016	1.051	0.991
3	FZRL3	85.5	1.097	1.083	0.566
4	ANKLE1	89.5	1.102	1.069	0.590
5	SIN3B	92.9	1.088	1.089	0.567
6	CHERP	95.3	1.086	1.041	0.605
7	SLC27A1	97.9	1.016	0.988	1.053
8	OCEL1	99.0	1.053	1.089	0.601
9	NR2F6	105.0	1.070	1.059	0.563
10	AP1M1	111.6	1.106	1.094	1.196
11	*control2	113.2	1.076	1.072	1.072
12	*control3	115.2	1.000	1.000	1.000
13	MRPL34	120.3	1.024	0.970	0.534
14	C19ORF42	140.3	0.898	1.113	0.560

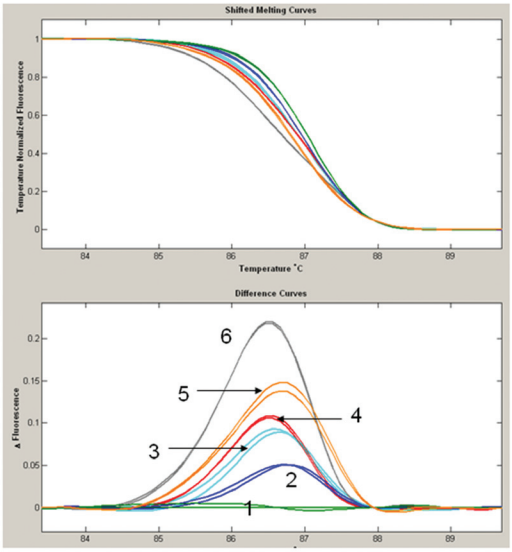
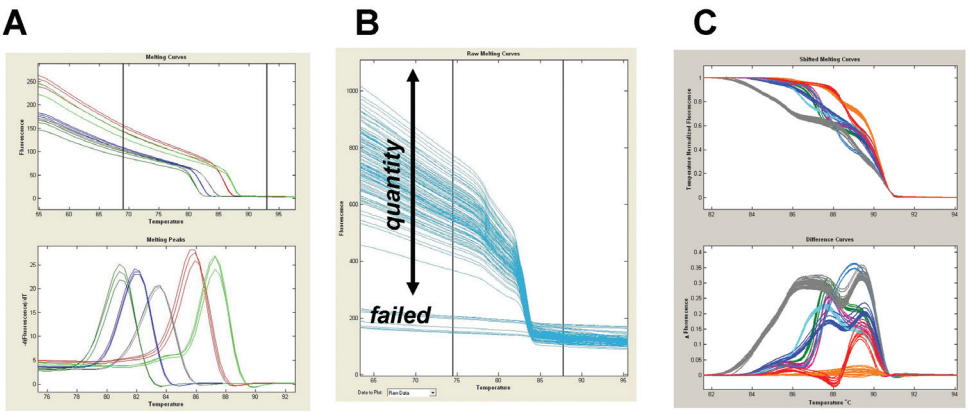


Figure 4: CA repeat analysis using HRMA. Six different DNA samples were analyzed in duplicate, all heterozygotes. Top panel: normalized temperature shifted melting curves. Bottom panel: derivative plot. Allele lengths were 1=18/21, 2=17/21, 3=14/17, 4=14/18, 5=15/21, 6=14/21.

Figure 6: HRMA as alternative for gel electrophoresis. A: Analysis of a series of five different PCR fragment; because the fragments have clearly different melt profiles all five fragments can be analyzed in one analysis. When the melt profiles partly overlap, analysis can be done per fragment. B: Analysis of a PCR performed on 96 different samples. Some PCRs failed (only background) fluorescence, yield of the others can be estimated from the level of fluorescence. Purity, including absence of primer dimers, can be checked by analysis of the HRMA difference plots (not shown). C: HRMA after insert PCR of 384 phage display clones after second round selection. Several clear groups of melt profiles are identified, an indication that the clones contain identical inserts. Note that to identify all groups present, clones recognized after a first analysis need to be removed and software grouping must be repeated. This procedure has to be repeated until no further groups are recognized.



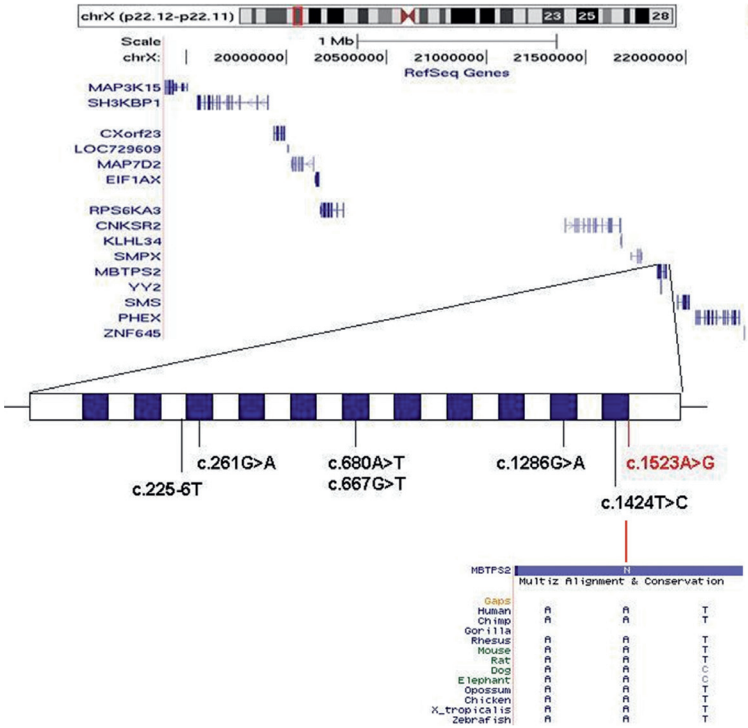


Figure 2: The KFSD locus. The KFSD linkage locus was redefined at Xp22.12-Xp22.11. High Resolution Melting curve Analysis (HRMA) identified *MBTPS2* as a candidate gene in the Dutch KFSD family. The detected variant in *MBTPS2* is indicated in red/gray. Previously identified mutations in IFAP syndrome are indicated in black.

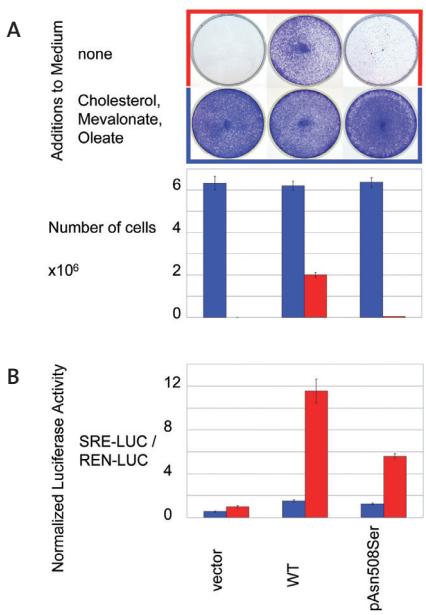


Figure 4: Complementation assay and luciferase reporter assay. Functional studies of the c.1523A>G (p.Asn508Ser) variant using an in vitro assay testing sterol responsiveness. A. Complementation assay: Growth of stably transfected CHO-K1-M19 cells (lacking hamster MBTPS2) was measured in cholesterol rich medium (blue) and cholesterol deficient medium (red). The proportion of cells capable of growth is documented as framed photographs of the cultures and, graphically in bars by counts of growing stably transfected cells. Colours indicate absence (red) or presence (blue) of sterols. B. Luciferase reporter assay: luciferase activity functions as an indirect measure of the ability of MBTPS2 mutants to restore sterol-regulated transcriptional activity in transfected CHO-K1-M19 cells. Cells transfected with the c.1523A>G (p.Asn508Ser) variant are less able to restore sterol-regulated transcription compared to wild type when transferred from a cholesterol rich (blue bars) to a cholesterol deficient medium (red bars).

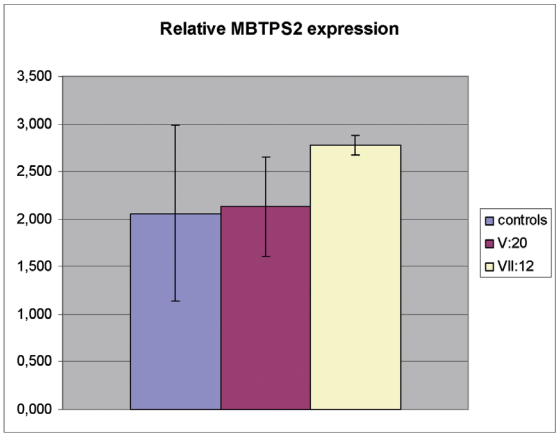


Figure S1: *MBTPS2* levels in KFSD males compared to control males.

Figure 3: *MBTPS2* mutation analysis in the Dutch KFSD family. a) Sanger sequencing of exon 11 in the *MBTPS2* gene identified a c.1523A>G (p.Asn508Ser) in all affected males of the Dutch KFSD family, not present in WT controls. The variant co-segregates with the disease in this family; affected male (VI-20) c.1523A>G , carrier female (VI-27) c.1523AG and unaffected male (VII-6) c.1523G. b) RNA analysis in fibroblasts in the Dutch KFSD family. cDNA sequencing in fibroblasts confirmed the presence of the c.1523A>G variant in an affected male. Carrier females (VI-19, VII-11, VI-27, and VII-10) showed variability in the expression of the mutated and the wild type allele.

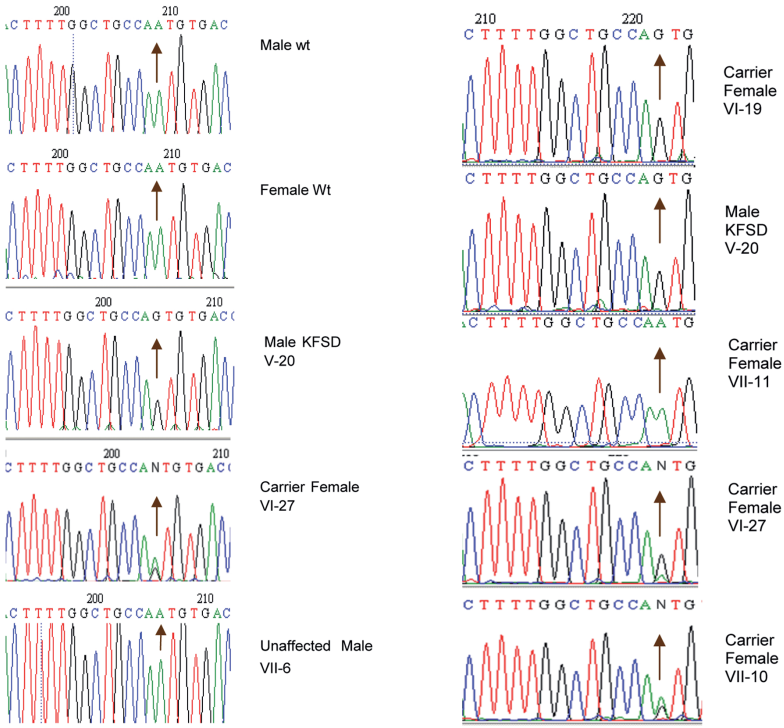
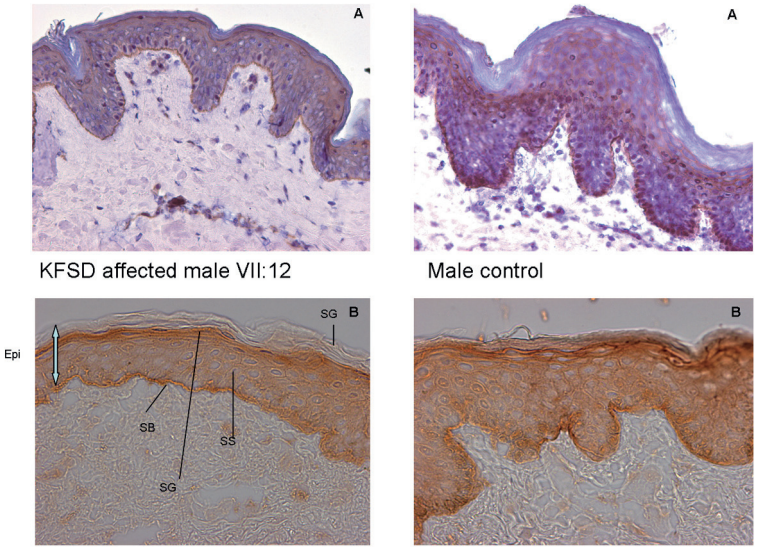


Figure S2: Skin biopsies stained with αM*BTPS2* (brown) with **(A)** and without Hematoxylin **(B)**. 40x and 20x magnification.
Epi=epidermis, SC=stratum corneum, SG=stratum granulosum, SS=stratum spinulosum, SB=stratum basale.



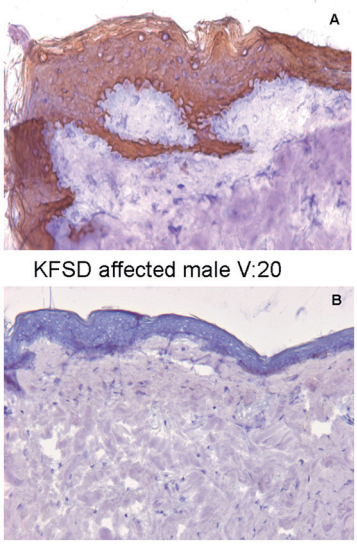
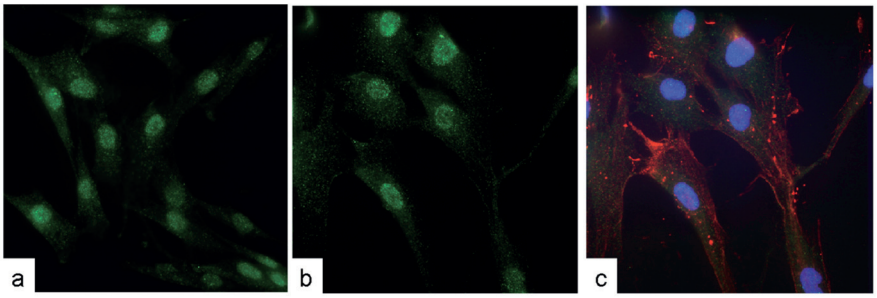


Figure S3: Skin biopsy stained with αKeratin-10 **(A)** and only counterstaining with Hematoxylin **(B)**. 20x and 10x magnification. Keratin-10 is expressed in the epidermis, except the stratum basale.

Figure S4: Cellular localisation of MBTPS2 in fibroblasts. There is no difference between a control male (a) and a male KFSD patient VII-12 (b). *MBTPS2* (green fluorescence) is found both in the cytoplasm and nucleus. Nuclei are stained in blue with DAPI. Red is a control staining of β-actin.



Chapter 6

Figure 1: The pedigrees and the phenotype of family 3. (A) The pedigrees investigated in this study, in family 3 X inactivation patterns show the silencing of the X chromosome which carries the mutant allele. (B) shows the hands of 3I:2. (C) Multiple frenula of 3II:4. (D) The right hands of 3II:4. She has clinodactyly and digital fibroma. (E) shows the right upper accessory frenulum of 3II:5. (F) The right hand of 3II:5.

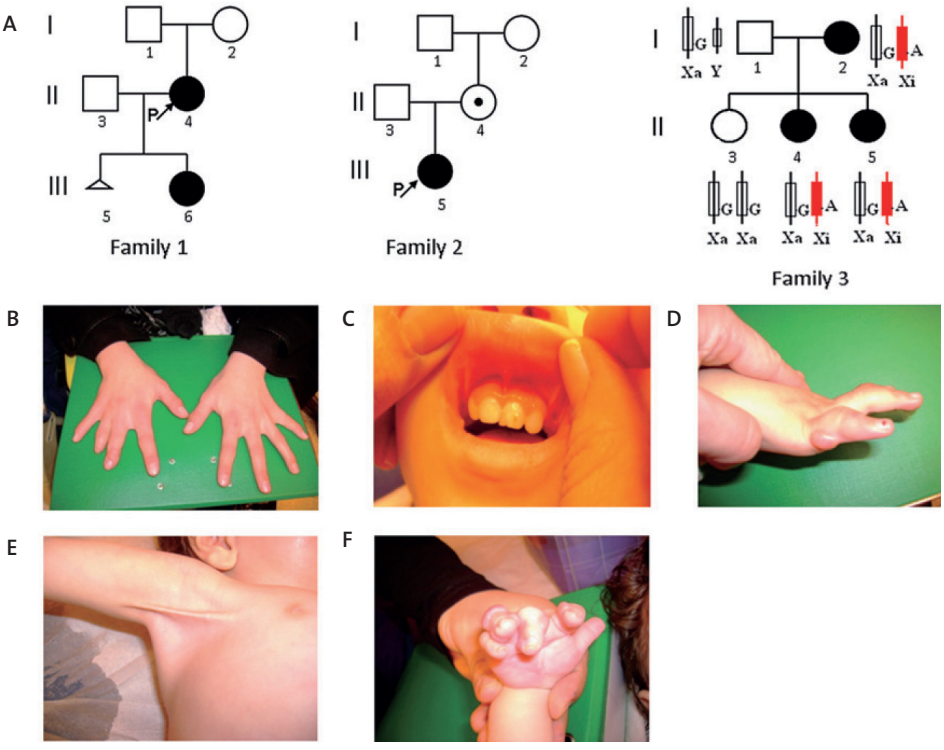
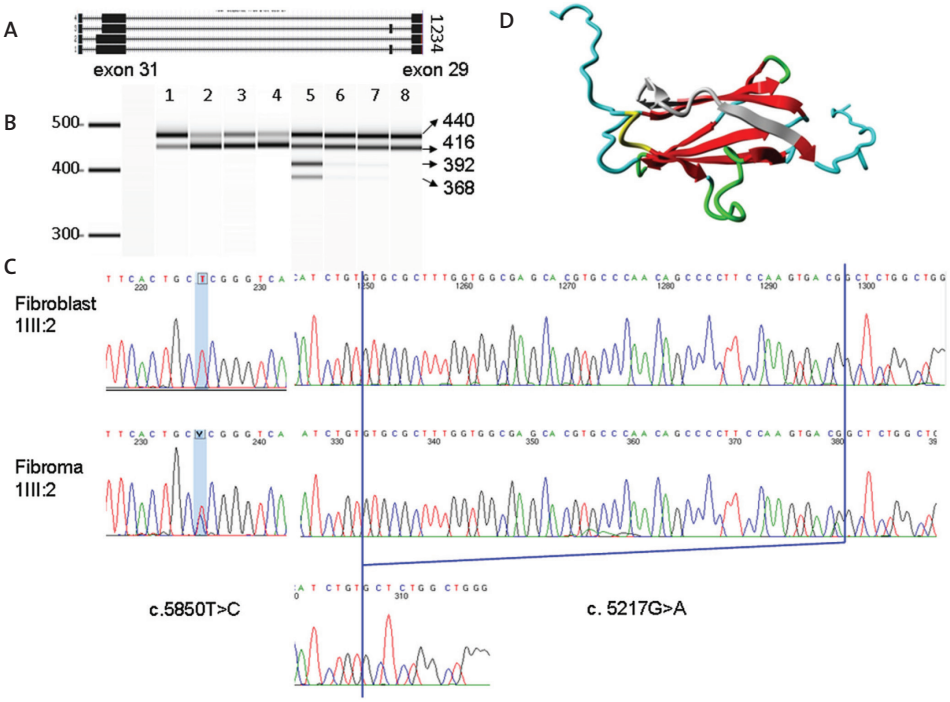


Figure 3: Detection of alternative splicing and 3D protein model. (A) Diagram of four FLNA transcripts in fibroma cells: transcript 1 and 2, which carry the 48bp deletion at the end of exon 31, as well as the normal transcripts 3 and 4. (B) RT-PCR result from Agilent 2100 Bioanalyzer. Lane 1 is the product of the fibroblasts of 1III:6, which has a predominant longer isoform. Lane 2-4 and 8 are four control human fibroblasts. Lanes 5-7 show RT-PCR products that were obtained from fibroma cells of 1III:6, the normal bands from two FLNA isoforms, and two extra shorter bands, which are faint in lane 6 (left fifth finger)and lane 7 (fifth toe of the left foot) , whereas lane 5 (right fifth finger)shows four dark bands. (C) Sanger sequencing results of c.5858T>C and c.5217G>A in fibroblast and fibroma cells of 1III:6. (D) The 3D model of FLNA domain 15. The deleted 16 amino acids are marked in gray. Beta-strands are marked in red. Green represents a turn. Yellow indicates a 3/10 helix. Random coils are colored in cyan.



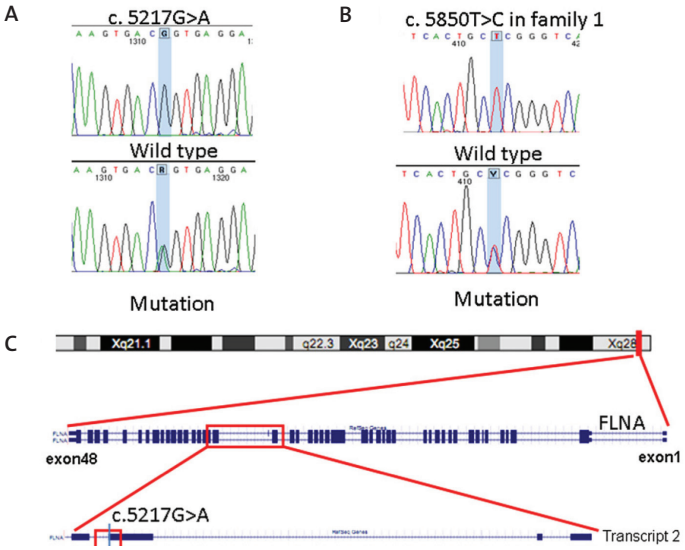
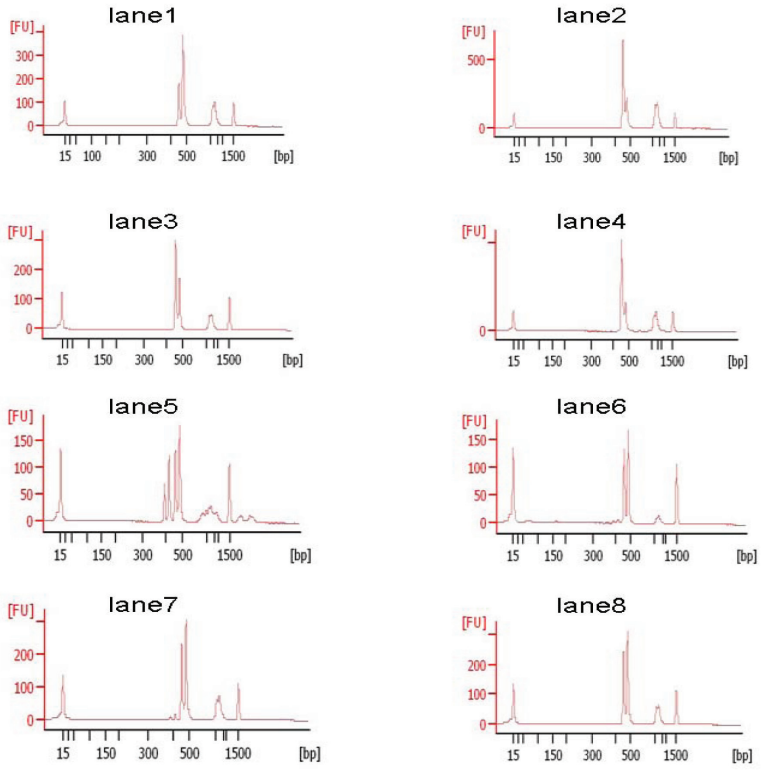


Figure 2: Genomic Structure and mutation analysis of *FLNA*. (A) c.5217G>A was confirmed by Sanger sequencing in all the patients. The unaffected family members and controls carry the homozygous normal allele. (B) shows the sequence of c.5850T>C in family 1. (C) FLNA is located in Xq28, the target region of linkage analysis. C.5217G>A alters the last nucleotide of exon 31 of FLNA.

Figure S3: The 2100 bioanalyzer traces of RT-PCR on c.5217G>A from lane 1 to 8. The peak around 15bp is the lower ladder and the signal round 1500bp is the upper ladder.



Chapter 7

Figure S2: Coverage plots of GAIIX sequencing for III-1 and III-2.

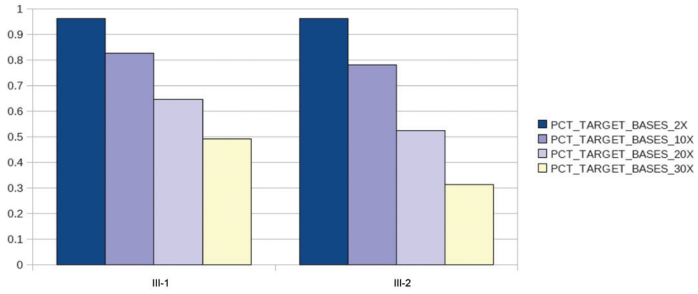


Figure 2: Genomic analysis of the mutation. A) Analysis result of GAII-X sequencing, displayed in IGV browser. III-1 and III-2 show a single nucleotide deletion, c.2016-35del in intron 12 of *FGD1*. On the left side, DNA Sanger sequencing results confirming the variant in the Dutch Aarskog family (III-1 and III-2 versus a WT control). B) A diagram of the normal and abnormal splicing in exon 12, 13 and 14. C) Sanger sequencing of the RT-PCR shows a skip of exon 13 in individuals III-1 versus a WT control, confirming the branch point variant. The skip of exon 13 leads to a 31 bp shorter mRNA. D) The 2100 Bioanalyzer results show semi-quantification of skipped products. III-1(lane 1) shows a 94% skip, containing two products, a large skipped (163 bp) and a small WT band (194 bp). III-2(lane 2) and I-1(lane4) show a 100% skip. II-1 (mother, lane 3) has a 50-50 distribution of WT and skipped products. Controls (lane 5~8) only carry the WT band.

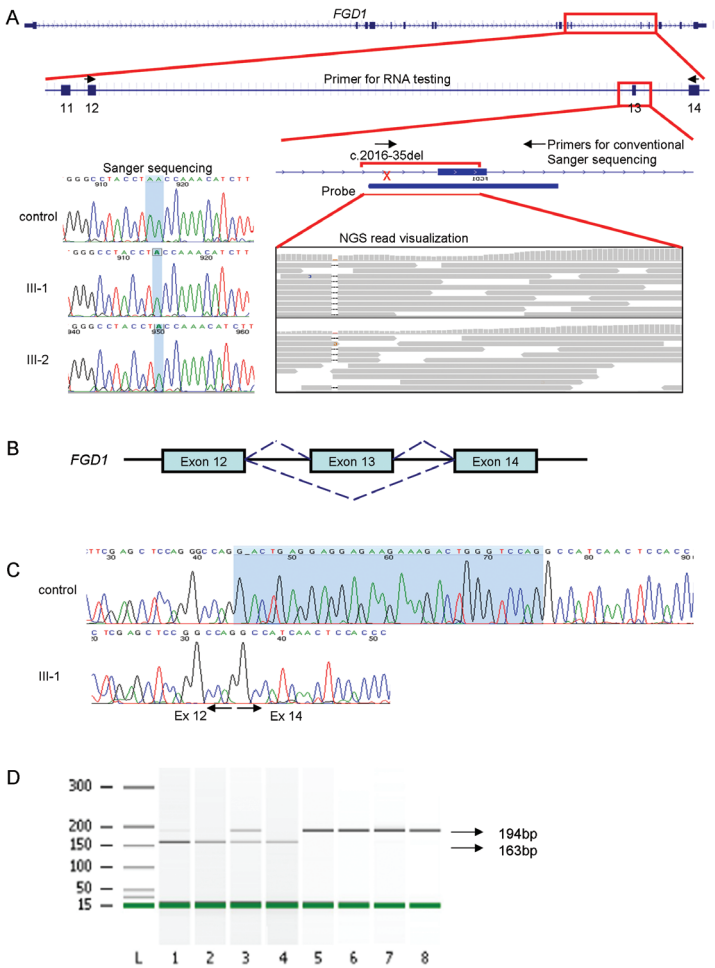


Figure S1: Phenotype of the two Aarskog siblings (III-1, III-2) and their maternal grandfather (I-1). Facial features are illustrated at age 8 years and 13 years (III-1) and at age 6 years and 11 years (III-2). Craniofacial features, skeletal and genital features are described in Table 1.

Figure S3: The 2100 Bioanalyzer Traces of RT-PCR show semi-quantification of skipped products. The peaks represented are the lower ladder (15 bp) and upper ladder (1500 bp), the skipped FGD1 product (163 bp) and the WT product.

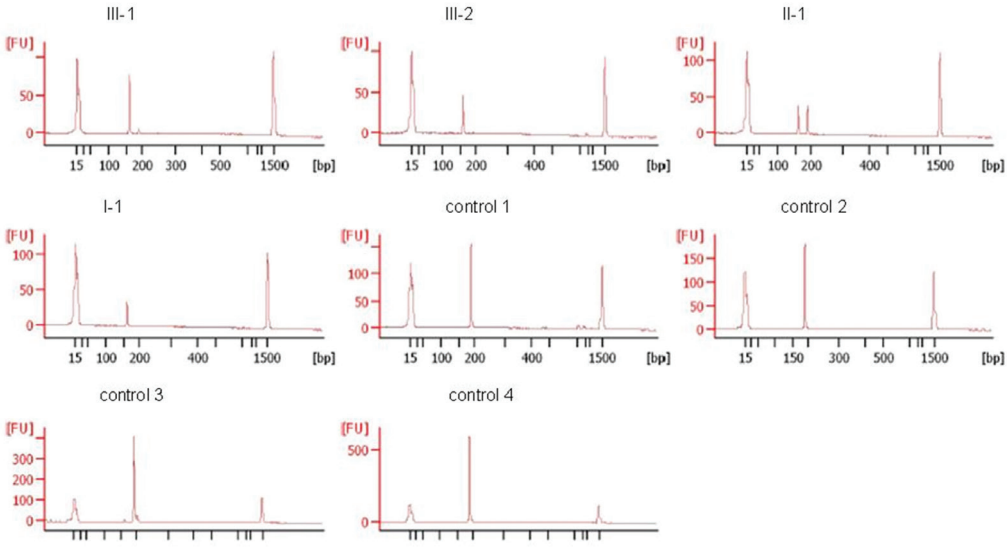


Figure S4: Application of exome sequencing for detection of intronic variants.
A) Intronic coverage plot for a range of distances to the splice site in our dataset. Percentage of the introns covered (minimal depth 8) decreases with an increasing distance from splicesite. B) Calculation of the number of filtered variants upon inclusion of intronic variants by a range of distances to the splice site (depth >8).

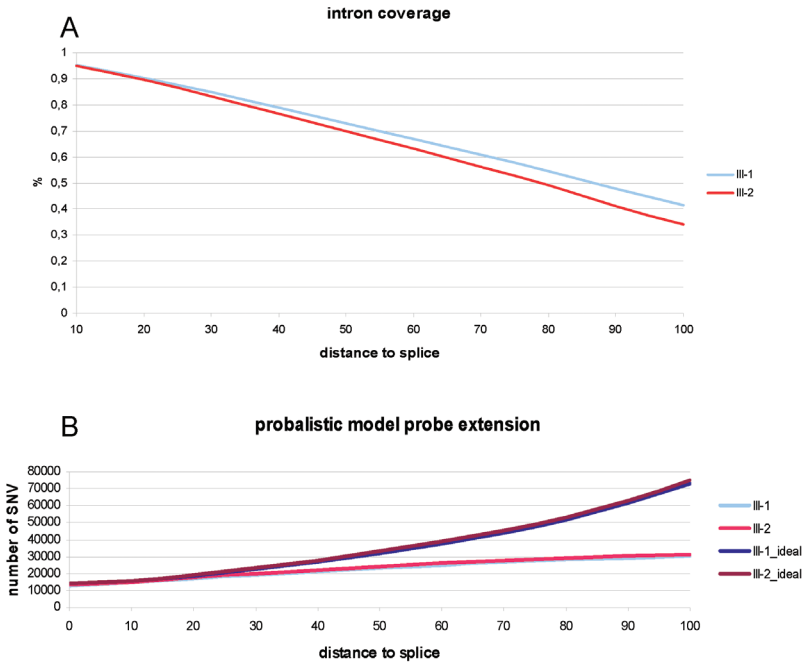


Figure S5: Overview of splice elements of FGD1 exon 12-13. 5'splice site (GU dinucleotide), branch point sequence YUNAY with adenine for lariat formation, AG exlusion zone (AGEZ), polypyrimidine tract (ppt) and 3'splice site (AG dinucleotide).

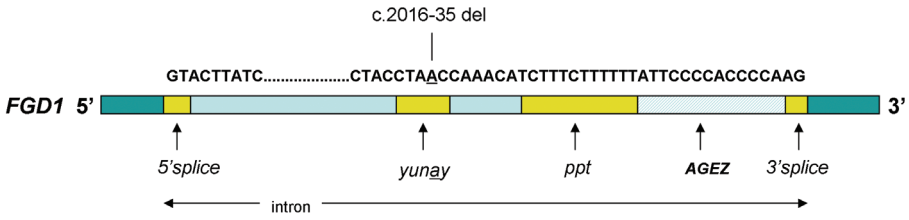
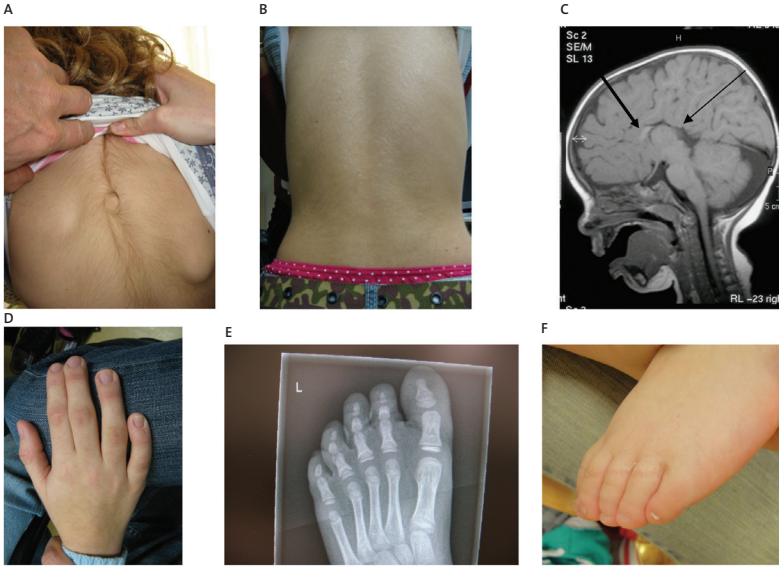
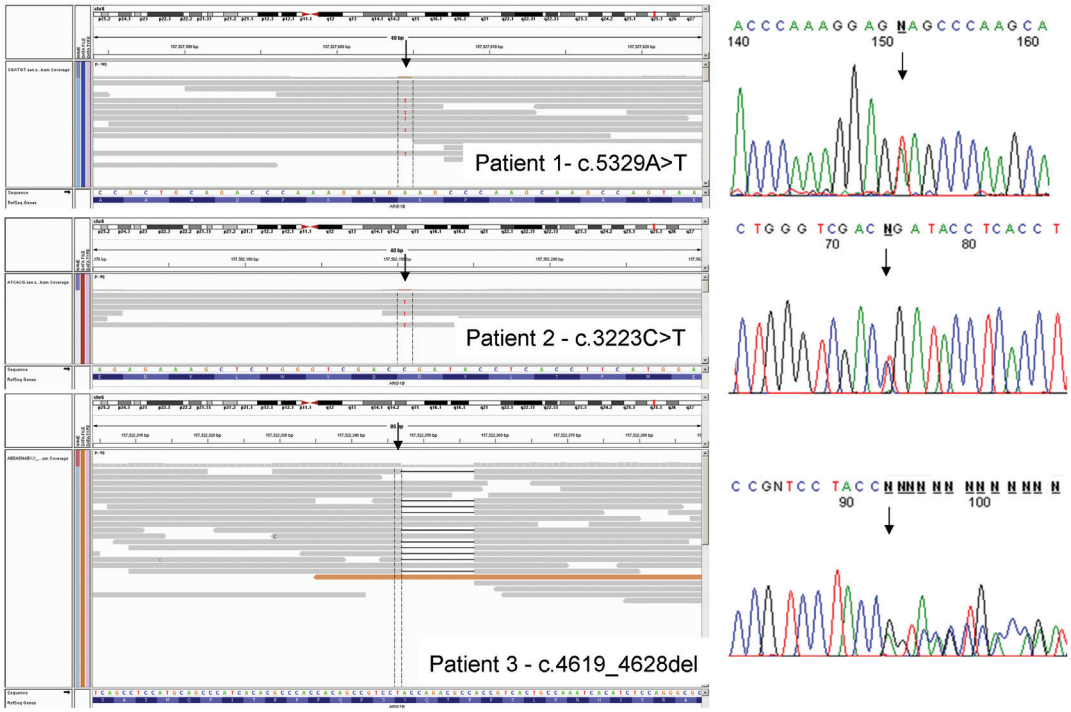


Figure 1: Facial features of all patients. Top, left to right: Patient 1 at 4.5 years, patient 2 at 2,5 years, patient 3 at 3 years. Bottom, left to right: Patient 4 at 3.5 years , patient 5 at 45 years, patient 6 at 3,5 years. All patients share coarse facial features, thick eyebrows and broad nasal tips. For further details see Supplementary Table 1. The parents or legal guardians of all affected individuals gave consent for publication of the clinical photographs.

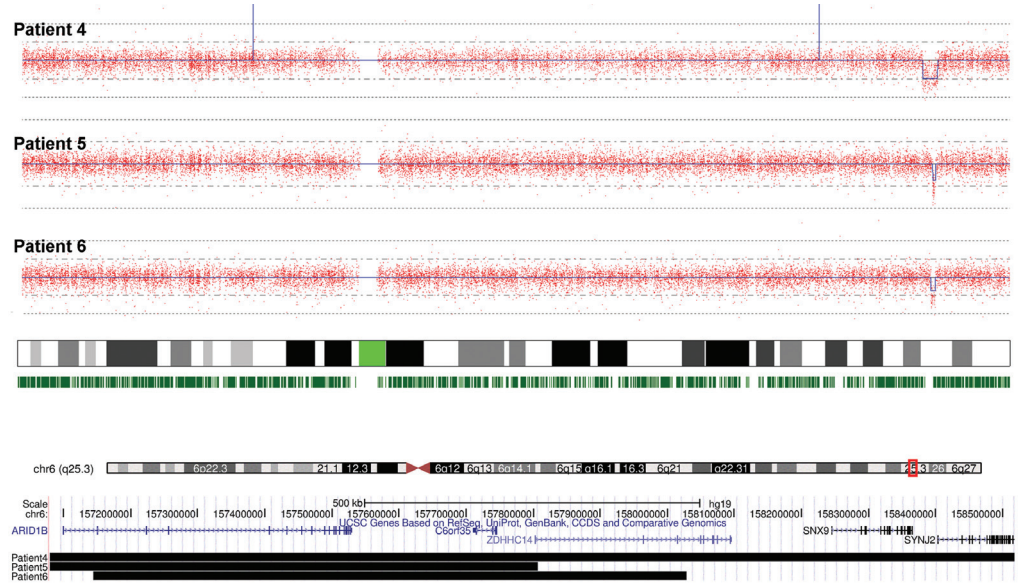
Supplementary Figure 1: Additional features in the CSS patients. Hypertrichosis in patients 1 (A) and 2 (B). Agenesis of the corpus callosum in patient 1 (thin arrow, thick arrow points to a small part of the corpus callosum that is present) (C). Brachydactyly of the fifth finger in patient 2 (D), missing terminal phalanx of the fifth toe in patient 3 (E), hypoplastic nail in patient 4 (F.



Supplementary Figure 2: On the left the raw data around the pathogenic variant in the Integrative Genomics Viewer[3] for each of the 3 patients. Arrows indicate the location of the variant. On the right the Sanger sequencing validation results.



Supplementary Figure 3: Top: Plot of the genomic deletions detected in patients 4 to 6. Red dots represent the raw data, the blue line represents the calculated CNV score. All deletions are *de novo* (data of parents not shown). No other possibly pathogenic CNVs were found in these patients. Bottom: graph of the size and location of each of the deletion.




GAGTTCCCTGGAACGGC
GAGTTCCCTGGAACGGG
CCTTCCGAGTTCCCTGC
GAGAATCCGCCGAACAC
AGAGCCCCGTCTGGTAC
CCTACTGAGTTCCCTGC
TTATGGTTCCTTCCGAC
GAGTTGTAATTTGTAGA
TTTGAAATCTGGCTCCT
CTACTGAGTTCCCTGGA
CCTACTGAGTTCCCTGC
CCCCGTCTGGTAGGACA
CCTTCCGAGTTCCCTGC
TCCTTCCGAGTTCCCTC
CTTCCGACTTCCCTGGA