The use of new technology to improve genetic testing

Rowida Almomani

The use of new technology to improve genetic testing

Rowida Almomani

Printed by Wöhrmann Print Service B.V., Zutphen

The use of new technology to improve genetic testing

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van de Rector Magnificus Prof.Mr. C.J.J.M. Stolker,

volgens besluit van het College voor Promoties

te verdedigen op woensdag 19 juni 2013

klokke 15:00 uur

door

Rowida Almomani

geboren te Al-Ruseifa, Jordanië

in 1979

Promotiecommissie

Promotores: Prof.dr. M.H. Breuning

Prof.dr. E. Bakker

Co-promotor: dr. Ieke B. Ginjaar

Overige leden: Prof.dr. S. van der Maarel (Leiden University Medical Center)

Prof.dr. J. Verschuuren (Leiden University Medical Center)

Prof.dr. R. Sinke (University Medical Centre Groningen)

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

Table of Contents

Chapter 1

1. General introduction

The identification of the molecular genetic basis is crucial for the definitive diagnosis of individuals with congenital malformations and inherited diseases and for the risk evaluation in relatives. Currently, molecular genetic diagnosis depends on the recognition of distinctive clinical features (syndrome), on linking the syndrome to a known underlying defect, and on the availability of a laboratory that offers the diagnostic test for the particular gene(s) of interest (1-2).

In the majority of cases, one sequences the gene(s) of interest to look for genetic variation that can explain the clinical phenotype. Various techniques are available for targeted sequencing, for example, conventional Sanger sequencing (3) and Next Generation Sequencing (NGS) (4-8). Sanger sequencing is the gold standard for diagnostic analysis of single candidate genes. Most diagnostic DNA laboratories worldwide use this method for the identification of disease causing mutations. Although DNA testing by Sanger sequencing is useful for most Mendelian diseases, its use is still hampered by limited throughput and high cost. However, one can enhance the speed and reduce the cost by using any one of several pre-sequencing screening methods for DNA fragment analysis (9-12). One can then sequence only those fragments that contain the variants. We have described the various pre-sequencing methods in **Chapter 2**. Among these, High Resolution Melting Curve Analysis (HR-MCA) offers a cost efficient, fast and convenient method for assessing the presence of variants in a diagnostic setting (13).

However, when there are too many samples and /or too many possible candidate genes to be tested in patients with genetically heterogeneous disorders, this combined technology (HR-MCA followed by Sanger sequencing) becomes time consuming, labour intensive and inefficient. For these diseases, such as, cardiomyopathies (14), retinitis pigmentosa (15), deafness (16), Noonan syndrome (17), one can use NGS to search for gene mutations. NGS circumvents the bottleneck

by interrogating all candidate genes or genomic regions of interest simultaneously (4-7). It is useful for sequencing DNA on a massive-scale, even an entire human genome (8, 18-21). The problem is that Whole Genome Sequencing (WGS) is very time consuming for many applications and is likely to remain prohibitively expensive for some time. However, a number of methods are now available to select targeted regions for sequencing in a more cost and time efficient manner (22). Such methods, described collectively as genome capture or genome enrichment technologies, include PCR- based methods (long range PCR or multiplexed short PCR) (23, 24), capture by hybridization (on–array and in–solution) (25-29), and capture by circularization (30).

The aim of the work reported in this thesis was to optimize, test and apply different new molecular techniques, which include HR-MCA, targeted sequencing and exome enrichment followed by NGS. The purpose of this was to facilitate the detection of disease causing mutations in several disorders with suspected Mendelian inheritance, to speed up the identification of disease genes and to provide a systematic tool for classifying previously intractable genetic diseases.

We review a number of the above-mentioned new techniques below. We also present several technical approaches for detecting candidate pathogenic variants in different Mendelian disorders.

1.1 High Resolution – Melting Curve Analysis (HR-MCA)

The introduction of the Sanger sequencing method over 30 years ago marked a milestone in the history of genetic analysis. It quickly became indispensable for basic biological research and in various applied fields, such as biotechnology, diagnosis and forensic biology. The key principle of this methodology is to utilize dideoxynucleotide triphosphates as DNA chain terminators and to use specifically labelled nucleotides to read a DNA template during DNA synthesis (3).

A series of technical modifications and innovations have slowly improved the accuracy and efficiency of the Sanger method and have finally led to the development of automated Sanger sequencing (31-34). The automated Sanger method became the method of choice for DNA sequencing and has dominated the field for almost two decades. During this time, the capacity of

the Sanger method has become so enhanced that it can read up-to 1000 base pairs per sequencing reaction (34).

Sanger sequencing of single candidate genes for a given disease is useful for diagnosis if minimal locus heterogeneity and distinctive clinical symptoms exist. However, the high cost of sequencing and the rarity of many conditions still hampers the application of this method to many Mendelian disorders. The number of clinical situations where results of testing for DNA mutations has implications for the management and treatment of patients is increasing. The already overburdened laboratories with serious financial constraints have to deal with the growing demand for rapid turnaround time. Therefore, one needs cost-effective techniques that are simple to perform, and which show high sensitivity and specificity. HR-MCA has been attracting more and more attention among analytical nucleic acid techniques in recent years. It is a simple, powerful and robust post-PCR analysis method for scanning of variants and for genotyping (13).

HR-MCA is based on a melting (dissociation) curve and does not require sample processing after PCR. This technology was made possible by the recent advances in real-time PCR instrumentation (that allows for highly controlled temperature transitions and data acquisition) and by the availability of improved double-stranded DNA (dsDNA)–binding saturation dyes and hardware and software to monitor and analyze the melting. These advances allow accurate detection of sequence variations based on melting analysis. The advantage here is that one does not need to use labelled probes and not all amplicons need to be sequenced (35-37). The fluorescent binding dyes, known as intercalating dyes, do not inhibit PCR. They have a high fluorescence when bound to dsDNA and low fluorescence in the unbound state (13). HR-MCA is done after PCR amplification of the region of interest in the presence of a dsDNA dye. The amplicons are warmed up and the fluorescence data is collected at 55–95°C at a temperature transition rate of 0.1 °C/s and 200 data points/°C (Idaho Technology Inc., Salt Lake City, UT). Specific PCR amplification of the intended targets is critical, requires careful design of primers, the correct length of the PCR product and optimal number of cycles.

At the beginning of the HR-MCA, there is a high level of fluorescence in the sample, but as the sample is heated up and the dsDNA melts into single strands, the dye is released resulting in a

change in fluorescence. The fluorescence reduces as the number of double stranded DNA fragments decrease (35, 36, 38). A camera in the machine monitors this process and the machine plots the data in a graph known as a melting curve that represents the level of fluorescence and the temperature (Idaho Technology Inc., Salt Lake City, UT). The fluorescence signal is plotted against the temperature and the fluorescence data that is generated is assessed, based on the shape of the melting curve or on the melting temperature (Tm) (38, 39). The highest rate of decrease of the fluorescence signal occurs at the melting temperature (Tm) of the DNA amplicon. The Tm is the temperature at which 50% of the DNA sample is double stranded and the other 50% is single stranded. In general, the Tm is higher when DNA fragments are long and/or have a high GC content (40). The fluorescence data collected during the HR-MCA ranges from pre-melt (initial fluorescence) to post-melt (final fluorescence) signals. The raw data is first normalized by selecting the pre-melt and post-melt regions for each primer set, onto a 0 to 100% scale (Figure 1A) before plotting the HR MCA curves. The temperature is then shifted in order to reduce the well-to-well variations. The amplicons with heterozygous variants can then be separated from the wild type and this is visible in the distinct shapes of the melting curves (Figure 1B) (38-40).

To visualize normalized data, difference plots can be generated (Figure 1 C) by subtracting the curves of the samples that are analysed from a reference curve. The machine automatically clusters samples with similar melting curves/genotypes into groups. So the distinct shape of the melting curve, the derivative plot, and/or the difference plot can be used for amplicon analysis (Idaho Technology Inc., Salt Lake City, UT). Amplicons that amplify poorly will have low fluorescence and should not be analyzed further.

The melting temperature of an amplicon at which the double DNA strands separate is predictable and depends on the sequence of the DNA bases. This allows one to compare different samples, which should give the same shaped melting curve, for the same amplicon. However, if one of the selected samples has variant in the amplified DNA sequence, it will alter the melting curve profile. In diploid organisms, which have two alleles, there are three possibilities for a given variant: both alleles contain the variant (homozygote variant), either allele has the variant (heterozygote), neither allele contains a variant (homozygous wild type).

One can distinguish between homozygous samples by a shift in the Tm, and between heterozygous samples by changes in the shape of the melting curve (13, 41, 42). However, not all homozygotes can be readily distinguished; homozygous or hemizygous variants can easily be missed due to subtle differences between some variants. The standard solution to overcome this problem is to mix patient DNA with wild-type DNA in order to generate heteroduplexes.

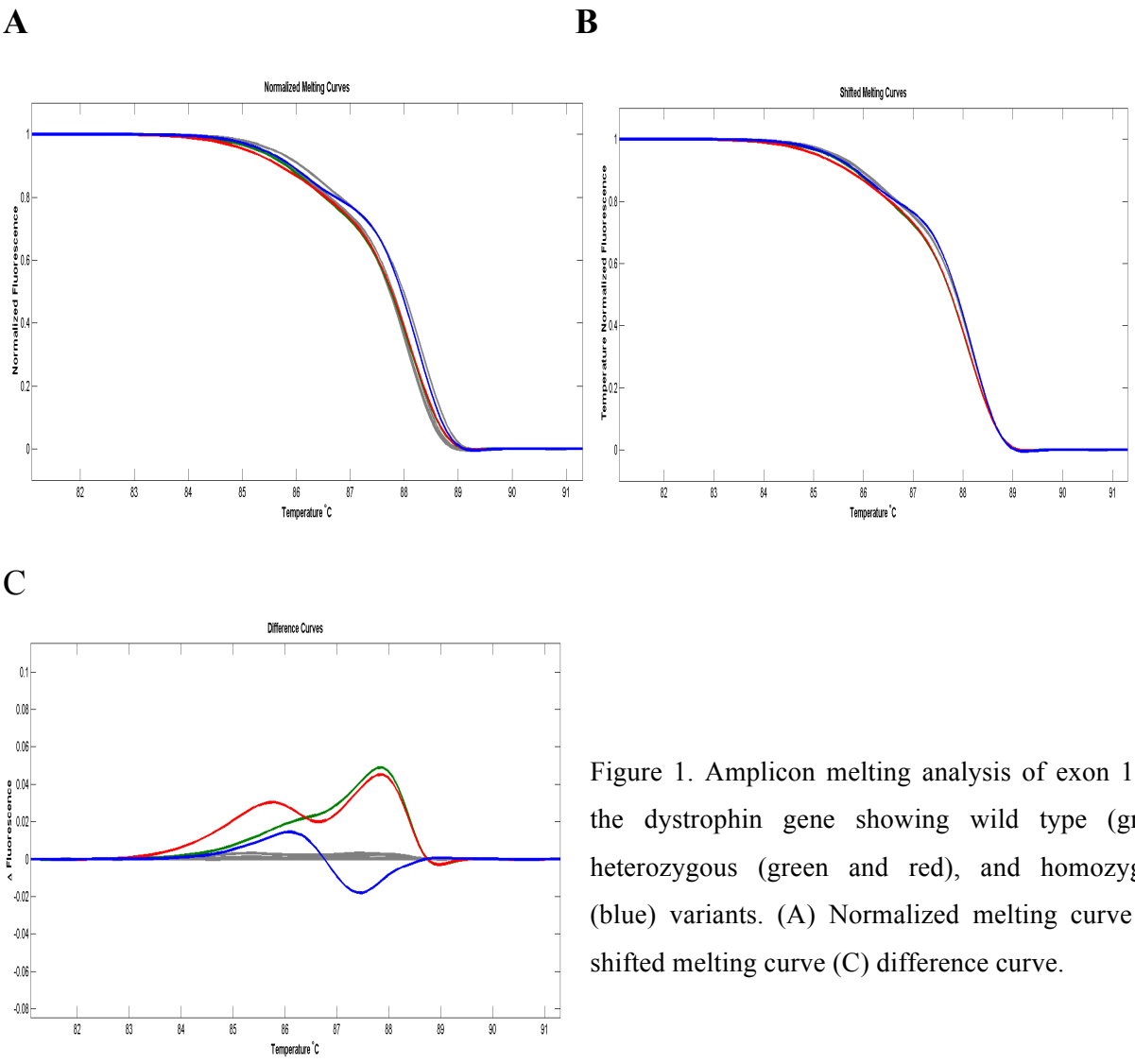


Figure 1. Amplicon melting analysis of exon 17 of the dystrophin gene showing wild type (gray), heterozygous (green and red), and homozygous (blue) variants. (A) Normalized melting curve (B) shifted melting curve (C) difference curve.

The melting profile of a PCR product depends on its GC content, its length and its sequence (39). Short PCR products normally melt in a single transition while longer PCR products often melt in multiple transitions corresponding to melting domains of different stability.

The HR-MCA method is simple and flexible, has minimal requirement for optimization and has superb specificity and sensitivity. For this reason it is used by a wide range of disciplines for a variety of applications such as DNA methylation analysis (43-45), genetic mapping (46), species identification (47), HLA compatibility typing (48), genotyping and mutation detection; the last of these is discussed in detail in the following section.

1. 1.1 Genotyping and mutation detection

It is important to determine differences in the genetic make-up (genotyping) for the study of genes and gene variants associated with disease. HR-MCA was first used as a genotyping technique (13). However, in most cases, the shape of an HR-MCA curve in itself is not sufficient to type a specific variant (40); moreover, the type of base change, the presence of a homozygous and/or a non-pathogenic variant (common variants) may complicate the interpretation of the melting profiles. Either adding an unlabeled probe (49, 50) or sequencing the fragment can solve the problem. Unlabeled probes are around 30 base pairs in length and are blocked at their 3´-end to prevent extension. An excess amount of the strand complementary to the probe is produced by asymmetric PCR (1:5 or 1:10 ratio). Unlabeled probes can be designed to match either the variant or wild type sequence. Data can be viewed by using the derivative plot: if two melting regions are visible, the allele that is complementary to the probe will show a single peak at the highest temperature, whereas other alleles will produce a peak at lower temperatures. Typically, heterozygotes will display two peaks representing the two alleles (39, 51).

Hundreds of variants in many genes that are associated with genetic diseases, with autosomal recessive, autosomal dominant or X-linked inheritance, have been examined; more than 60 different genes have been analyzed using HR-MCA (35, 39, 51-61).

In **Chapter 2** of this thesis, we show the successful application of HR-MCA as a pre-sequencing screening method. We have optimized and validated the HR-MCA method and used it in combination with dsDNA dye LCGreen Plus to scan all coding exons and the exon/intron junctions and to genotype frequently found variants in the largest known human gene to date, the DMD gene. Mutations in the DMD gene can cause Duchenne and Becker muscular dystrophy (DMD). We found that amplicons up to 600 base pairs and more can be used for HR-MCA but the technique is more sensitive when shorter fragments, that result in melting profiles with no

more than two melting domains are tested. In addition, we found that HR-MCA is capable of distinguishing amplicons that differ by a single base pair. Therefore, we can use it to detect single nucleotide substitutions and small deletions and duplications. HR-MCA is a highly reliable and a quick method for mutation scanning and genotyping. It is now ready for routine diagnostic use on patients with Duchenne or Becker muscular dystrophy and on female carriers.

1.2 Next generation sequencing (NGS)

The ability to read the sequence of bases that comprise polynucleotide molecules such as DNA has had an enormous impact on biological and medical research. Sanger sequencing is the gold standard sequencing technology since 1977 (3). It has led to a number of monumental accomplishments, including the completion of the human genome sequence (62). However, the limited throughput of Sanger sequencing makes it expensive, laborious and time consuming, and therefore unsuitable for large-scale sequencing projects. The advent of Next-Generation Sequencing (NGS) technologies in 2005 has changed the paradigm of DNA sequencing and has opened fascinating new opportunities in biomedicine (4-6). NGS technologies have made it possible to process hundreds of thousands to millions of DNA templates in parallel. This results in a high throughput (gigabase) scale and low cost per base (6, 7). The cost of DNA sequencing keeps reducing due to rapid innovations in sequencing technology. The inexpensive production of huge amounts of sequence data is the main advantage of NGS over the Sanger sequencing method.

NGS uses a number of different technologies that have appeared since 2005. In several NGS methods, fragmented genomic DNA ligated to universal adaptors amplifies into PCR colonies. Each colony has many copies of the same fragment, and some NGS methods can sequence all of them in parallel, whereas other NGS methods read single DNA sequences. Older NGS technologies read sequences from one end while newer platforms allow for paired-end sequences (4-8). Once the NGS produces a sequence, the sequenced data is mapped to a reference genome, such as the human reference genome, which provides the basis for all subsequent steps of data analysis. Several NGS platforms are now available on the market and among them, the Roche/ 454, Illumina (Genome Analyzer/ Hiseq), and the Life Technologies SOLiD System are the

commercially dominant ones. Table 1 shows a summary and a comparison between these three different platforms.

Table 1. A summary and a comparison between three NGS platforms.

Companies	**Roche GS FLX**	**Illumina-Solexa**	**Life Technologies**
Company homepage	http://www.454.com	www.illumina.com	http://www.appliedbiosystems.com
Platforms	GS FLX Titanium, GS 20	HiSeq 2000, Genome Analyzer II (GA II), Solexa platform	ABI SOLiD, SOLiD 4
Sample requirements	1 µg for shotgun library, 5 µg for paired end	<1 µg for single or paired-end	<2 µg for shotgun library, 5–20 µg for paired end,
duraton of library prep/feature generation (days)	3–4	2	2–4
Method of feature generation	Bead-based/emulsion PCR	Isothermal bridge PCR amplification on flow cell surface	Bead-based/emulsion PCR
Chemistry	Pyrosequencing (sequencing-by-synthesis with pyrophosphate)	Reversible Dye Terminators	Sequencing by ligation
Reads per run	1 million	up to 3 billion	1.2 to 1.4 billion
Raw accuracy	99.99%	98%	99.99%
Read length	700 bases	50 to 250 bp bases	50+35 or 50+50 bp
Sequencing run time	10-24 hours	1 to 10 days (based on the sequencer)	6 days
References	63-67	68-70	66, 67, 71

In a relatively short time, NGS technologies have revolutionized the research on the human genome. They have been applied to genomic sequences, the trancriptome (RNA seq) and chromatin immunoprecipitation in combination with DNA microarray (ChIP- chip) or sequencing (ChIP-seq) (72-98).

In the following sections, we address the important features and applications of NGS for whole-genome and targeted re-sequencing.

1.2.1 Whole-genome and targeted re-sequencing

The most common use of NGS platforms has been re-sequencing (99). The introduction of these technologies has made it possible for some laboratories to sequence an entire human genome.

Several completely sequenced human genomes have been published so far, for example: human genome sequences of the well known James D. Watson and of an African, a Chinese and two Korean individuals (18, 19, 100- 102). Sequencing whole human genomes will lead to a deeper understanding of the full spectrum of genetic variation. It will also throw some light on the role of genetic variation in phenotypic variation and disease susceptibility. However, the cost and capacity required for whole genome sequencing (WGS) is still significant. Routine sequencing of large numbers of whole human genomes is not yet feasible. Nevertheless, we expect that it will become routine in the near future. For the time being, for time and cost effectiveness, we have to select and enrich genomic regions of interest before sequencing. Moreover, the resulting data from target-enrichment methods is significantly less cumbersome to analyze. Thus, for some projects, the sequencing of large numbers of samples after targeted enrichment provides more answers to biological questions than the sequencing of the whole genomes of fewer individuals.

Several methods to target specific areas in the genome prior to NGS have been developed (86). In general, there are three categories: PCR based methods (103- 105), capture-by-circularization (106-108) and capture-by-hybridization (25, 27, 28, 109).

1.2.1.1 PCR

For over 20 years, PCR has been the most widely used pre-sequencing technique for sample preparation, as it is compatible with Sanger sequencing (3). PCR is also potentially well suited for NGS platforms, but to make full use of the high throughput, a large number of amplicons or samples must be pooled and sequenced together. However, multiplex PCR is difficult to perform since simultaneous use of many primer pairs can lead to differential amplification, formation of primer dimmers and to high rates of mispriming events (110, 111). Some amplicons may even fail to amplify. In addition, the length of amplicons that can be generated by long range PCR is limited (111, 112).

In our experience, working with very long PCR fragments tends to be laborious, time consuming and expensive. This is because each individual PCR has to be optimized, with a maximum length of 11 kilobases, in order to make amplification as efficient as possible. Theoretically, you can design primers for long range PCR for all desired targets. However, in practice, not all reactions will yield a PCR product. This can also be a problem when amplifying fragments with a low

DNA integrity. Furthermore, the presence of SNPs in the primer annealing regions may lead to amplification of only one allele. These problems can be overcome by redesigning and optimizing the primer, and using a combination of short and long range PCR. After PCR amplification, the concentration of uniplex PCR products should be normalized to avoid sequencing one dominant PCR product from one sample or amplicon above all others. Only then, can one pool the PCR products before sequencing. The RainStorm method circumvents many of the problems of the standard PCR–based approach. RainDance Technologies have developed this method (http://www.raindancetechnologies.com), which involves the use of emulsion PCR (70).

1.2.1.2 Capture-by-circularization

Molecular Inversion Probe (MIP) belongs to the category of molecular techniques that capture sequences by circularization. MIPs were initially developed for multiplex target detection and SNP genotyping (113, 114). In this technique, single-stranded oligonucleotide molecule (probe) consists of two target complementary arms separated by a linker region (one or two sequence tags and two amplification primers common to all MIPs). This assay is performed in three steps: hybridization of probes to the target sequences, circularization of bounded MIPs by ligase and amplification using common primers. The main advantages of capture-by-circularization with MIPs are reproducibility and high specificity with high levels of multiplexing (at least 300,000 independent targets). It can be applied directly to genomic DNA, which needs low amounts of starting material. Moreover, MIP amplification products can be directly sequenced by NGS (22, 107, 108).

1.2.1.3 Capture-by-hybridization (Hybridize capture)

Capture-by-hybridization can selectively target any specific area in the genome, such as genes of interest, linkage regions, whole chromosomes (X-exome) and all exons (exome sequencing) (22, 115). This approach relies on the hybridization of fragmented genomic DNA libraries to a complex mixture of capture probes. The capture probes may potentially be in solution (28) or fixed to a solid matrix such as a microbead or a glass surface (on array) (25, 26). This method has clear advantages over other methods that are based on the extension or ligation by an enzyme. These include the possibility of greater degree of multiplexing and potentially higher tolerance for polymorphisms that overlap with the present capture probes (22).

In the on-array capture methodology, total genomic DNA is fragmented and applied to the probes; non-bound fragments are removed by washing after hybridization. The hybridized DNA fragments are then eluted and enriched for sequencing (25) (Figure 2). Roche/NimbleGen was the first to adapt this technique. They provide different types of microarrays with different capture size varying from 5 Mb to 34 Mb on a single array, with different numbers of probes with lengths ranging from 60 to 90 basepairs. Agilent also offers commercial kits implementing this technology.

In our hands, on-array target enrichment of large targeted regions showed several advantages over PCR methods: it is quicker, cheaper and less laborious. However, there are also disadvantages: working with large numbers of samples is not feasible, because arrays that are hybridized at the same time must also be eluted together and a single person cannot perform more than 20 arrays per day. Working with arrays requires expensive equipment such as a hybridization station. In addition, irrespective of whether the on-array capture experiment is for 100 kb or an entire exome, a large amount of input DNA is needed (10–20 μg) to start a library preparation.

In **Chapter 3**, we describe custom high-density microarrays (NimbleGen) to enrich exons and intron-exon junctions of 112 distinct genes potentially involved in mental retardation and congenital malformation, which were sequenced on the Illumina analyzer (Solexa). Our results show that this methodology offers a versatile tool for successfully selecting sequences of interest from the total human genome. In addition, we have discussed a number of advantages and disadvantages characteristics of this methodology. To overcome many of the disadvantages, Agilent, NimbleGen, and other companies have developed an in-solution-based target enrichment. In-solution capture is similar to the on-array capture method, but the probes in this technology are biotinylated (DNA or RNA bait) and not attached to a solid support.

Many different methods have been described for targeted hybrid capture but only a few have been extended to capture the whole human exome (exome sequencing) (http://www.illumina.com/products/truseq_exome_enrichment_kit.ilmn, http://www.genomics.agilent.com/,http://www.nimblegen.com/exomev3launcheq/)

1.2.1.3.1 Exome sequencing (ES)

Elucidation of the genetic basis of rare and common human diseases is the main goal of genetics and molecular biology. To date, causal variants for about 3,000 different Mendelian disorders have been identified (Online Mendelian Inheritance in Man OMIM, http://www.ncbi.nlm.nih.gov/omim) (116, 117).

In the past two decades, linkage studies followed by positional cloning and homozygosity mapping have been very successful in identifying causal mutations for Mendelian disorders (118, 119) due to a perfect segregation pattern of the causal variant with the disorder. However, these traditional methods are not suitable for studying all Mendelian disorders. Several factors limit the power of these methods such as, availability of patients for investigation, small number of cases or families (extremely rare Mendelian disorders), sporadic cases where variants have arisen de novo during meiosis, reduced penetrance and locus heterogeneity (117).

Exome sequencing encompasses 1% of the genome and includes 180,000 exons from more than 20,000 genes. It has become the main tool for studying the genetic causes of Mendelian disorders and of sporadic cases in which traditional methods have failed (120-123). Even when the traditional methods are expected to succeed, ES provides a tool for accelerating the discovery of disease genes (124). **In Chapter 5**, we have shown that we were able to rapidly identify a missense mutation and a splice site mutation in *TPP1*, which causes the autosomal recessive spinocerebellar ataxia type 7 (SCAR7), by ES of only one patient.

ES starts with random shearing of genomic DNA and flanking the library fragments by adaptors. Next, the library is enriched for sequences corresponding to exons either by in-solution or on–array hybridization capture. After that, the fragments are hybridized to the probes in the presence of blocking oligonucleotides. The recovered hybridized fragments are then enriched by PCR amplification, and this is followed by NGS (Agilent sureselect/ Roche/Nimblegen) (Figure 2). For sample indexing, barcodes can be introduced during the initial library construction or during post-PCR capture amplification (28). After NGS, data mapping and analysis of candidate variants are performed. At the Human and Clinical Genetics Center in Leiden, we have built a data analysis pipeline, which automatically maps the data to the reference genome, retrieves

sequence variants, checks their presence in known databases (e.g. dbSNP, 1000 genome) and predicts the potential consequences at the level of protein translation.

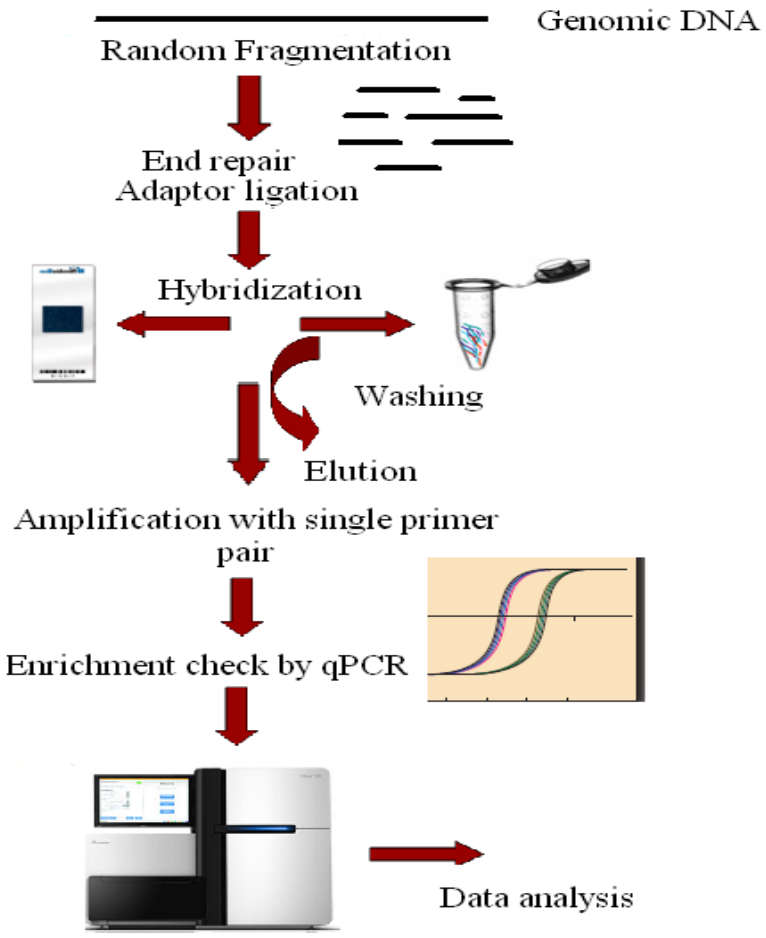


Figure 2. On- array and in-solution hybrid capture protocol.

At least three vendors (Agilent, Illumina and Nimblegen) offer in-solution whole exome capture kits. There are technical differences between them: for instance, Agilent uses RNA probes while Illumina and Nimblegen use DNA probes. These kits also differ in the definition of the exome (fraction of the genome targeted). We found the performance of exome kits from Agilent and Nimblegen (34 Mb) to be largely equivalent (see Table 2); each is scalable in 96-well plates by using a thermal cycler with no need for special equipment.

Table 2. The performance of whole exome kits from Agilent and Nimblegen (34 Mb).

	Agilent Sure Select	NimbleGen
# bases covered by baits (target base)	33,149,893	31,389,337
# bases covered by reads	32,878,642	31,233,008
% target base covered by reads	99.18%	99.50%
Average (base coverage) for the target	30.32x	29.97x
# consecutive-bait regions (CBR)	150,742	164,191
% CBR not covered by reads	1.18%	0.61%
% CBR max(base_coverage) < 5x	1.60%	1.09%
% CBR avg(base_coverage) >= 10x	86.09%	85.63%
% CBR avg(base_coverage) >= 20x	60.53%	60.74%
% CBR min(base_coverage) >= 20x	19.29%	20.96%

Despite the fundamental limitation of the current ES technology, which does not cover non-coding regions, it is a powerful strategy for discovering genes that underlie Mendelian disorders. This is for two reasons: First, large fractions of rare protein-altering variants, which are predicted to be deleterious, are located in exons (22); second, the price of ES is lower than that of sequencing an entire human genome.

Many proof-of-concept studies using ES to identify new disease genes for Mendelian disorders (125, 126) have been carried out in the past two years. There are an increasing number of successful studies that have found pathogenic variants of different diseases. More than 100 genes have been identified by ES in several Mendelian disorders with dominant, recessive and X-linked inheritance (Table 3), and this number is expected to increase.

Table 3. Overview of Mendelian disease genes identified by NGS, based on Rabbani *et al. (*2012).

Disorder	Inheritance	Gene identified	Reference
Miller syndrome	AR	DHODH	127
Autoimmune lymphoproliferative syndrome	AR	FADD	128
Nonsyndromic hearing loss	AR	GPSM2	124
Combined hypolipidemia	AR	ANGPTL3	129
Perrault syndrome	AR	HSD17B4	130
Complex I deficiency	AR	ACAD9	131
Hyperphosphatasia mental retardation syndrome	AR	PIGV	132
Sensenbrenner syndrome	AR	WDR35	133
Cerebral cortical malformations	AR	WDR62	134
3MC syndrome	AR	MASP1	135
Kabuki syndrome	AD	MLL2	121
Schinzel–Giedion syndrome	AD	SETBP1	136
Spinocerebellar ataxia	AD	TGM6	137
Terminal osseous dysplasia	XLD	FLNA	138
Nonsyndromic mental retardation	AR	TECR	139
Retinitis pigmentosa	AR	DHDDS	140
Osteogenesisimperfecta	AR	SERPINF1	141
Skeletal dysplasia	AR	POP1	142
Combined malonic and methylmalonicaciduria	AR	ACSF3	143
Amelogenesis	AR	FAM20A	144
Chondrodysplasia and abnormal joint development	AR	IMPAD1	145
Progeroid syndrome	AR	BANF1	146
Infantile mitochondrial cardiomyopathy	AR	AARS2	147
Heterotaxy	AR	SHROOM3	148
Mosaic variegated aneuploidy syndrome	AR	CEP57	149
Hypomyelinatingleukoencephalopathy	AR	POLR3A, POLR3B	150
Spastic ataxia-neuropathy syndrome	AR	AFG3L2	151
Dilated cardiomyopathy	AR	GATAD1	152
Gonadal dysgenesis	AR	PSMC3IP	153
Autosomal recessive progressive external ophthalmoplegia	AR	RRM2B	154
Knobloch syndrome	AR	ADAMTS18	155
Spinocerebellar ataxia with psychomotor retardation	AR	SYT14	156
Adams–Oliver syndrome	AR	DOCK6	157
Steroid-resistant nephrotic syndrome	AR	MYO1E, NEIL1	158
Complex bilateral occipital cortical gyration abnormalities	AR	LAMC3	159
Intellectual disability	AR	AP4S1, AP4B1, AP4E1	160

Hypertrophic cardiomyopathy	AR	MRPL3	161
Retinitis pigmentosa	AR	MAK	162
3M syndrome	AR	CCDC8	163
Seckel syndrome	AR	CEP152	164
ADK deficiency	AR	ADK	165
Nephronophthisis-like nephropathy	AR	WDR19	166
Psuedo–Sjögren–Larsson syndrome	AR	ELOVL4	167
Idiopathic infantile hypercalcemia	AR	CYP24A1	168
EMARDD	AR	MEGF10	169
Gray platelet syndrome	AR	NBEAL2	170
Immunodeficiency, centromeric instability and facial anomalies	AR	ZBTB24	171
Leber congenital amaurosis	AR	KCNJ13	172
Hereditary spastic paraparesis	AR	KIF1A	173
Ohdo syndrome	AD	KAT6B	174
Paroxysmal kinesigenicdyskinesias	AD	PRRT2	175
Hajdu–Cheney syndrome	AD	NOTCH2	176
Bohring–Opitz syndrome	AD	ASXL1	177
Hereditary diffuse leukoencephalopathy with spheroids	AD	CSF1R	178
Spondyloepimetaphyseal dysplasia	AD	KIF22	179
Adult neuronal ceroid-lipofuscinosis	AD	DNAJC5	180
KBG syndrome	AD	ANKRD11	181
Dendritic cell, monocyte, B and NK lymphoid deficiency	AD	GATA-2	182
Late-onset Parkinson's disease	AD	VPS35	183
Sensory neuropathy with dementia and hearing loss	AD	DNMT1	184
Dilated cardiomyopathy	AD	BAG3	185
High myopia	AD	ZNF644	186
Autosomal dominant retinitis pigmentosa	AD	RPE65	187
Charcot–Marie–Tooth disease	AD	DYNC1H1	188
Hereditary hypotrichosis simplex	AD	RPL21	189
Geleophysic and acromicric dysplasia	AD	FBN1	190
Myhre syndrome	AD	SMAD4	191
Leucoencephalopathy	XLR	MCT8	192
Split hand and foot malformation	AR	DLX5	193
Global eye developmental defects	AR	ATOH7	194
Primary hypertrophic osteoarthropathy	AR	SLCO2A1	195
Bartsocas–Papas Syndrome	AR	RIPK4	196
Familial aplastic anemia	AR	MPL	197
Peeling skin syndrome	AR	CHST8	198
Sengers syndrome	AR	AGK	199
Hypertension	AR/AD	KLHL3, CUL3	200
Weaver syndrome	AD	EZH2	201
Genitopatellar syndrome	AD	KAT6B	202

Hypothyroidism	AD	THRA	203
Floating–Harbor syndrome	AD	SRCAP	204
Hereditary spastic paraplegia type 12	AD	RTN2	205
Microcephaly associated with lymphedema	AD	KIF11	206
Congenital disorders of glycosylation (CDG)	AR	DDOST	207
Congenital mirror movements	AD	RAD51	208
Mandibulofacialdysostosis with microcephaly	AD	EFTUD2	209
Limb-girdle muscular dystrophy	AD	DNAJB6	210
Congenital stationary night blindness	AR	GPR179	211
Autosomal recessive primary microcephaly	AR	CEP135	212
Aplastic anemia and myelodysplasia	AD	SRP72	213
Acrodysostosis	AD	PDE4D	214
Olmsted syndrome	AD	TRPV3	215
Familial diarrhea	AR	GUCY2C	216
Nager syndrome	AD	SF3B4	217
Infantile cerebellar retinal degeneration	AR	ACO2	218
Coffin–Siris syndrome	AD	ARID1B	219
Joubert syndrome	AR	C5ORF42	220
Cerebroretinal microcephaly with calcifications and cysts	AR	CTC1	221
Kohlschutter–Tonz Syndrome	AR	ROGDI	222
UV-sensitive syndrome	AR	UVSSA	223
Pulmonary arterial hypertension	AD	CAV1	224

In **Chapters 4-7,** we have shown several examples of successful application of ES for detecting pathogenic mutations in various diseases. Moreover, ES has been used for detecting somatic mutations in tumours (225). The strength of ES in both research and as a diagnostic tool is becoming increasingly evident. It is used to find pathogenic mutations and to confirm the clinical diagnosis. ES can also be used as a genetic screening method to determine the carrier status of an individual, with respect to mutations that cause a particular autosomal recessive disorder (125, 126, 226). The diagnostic application of ES has been demonstrated by several examples (126, 225). However, a number of technical factors are still challenging for which ES requires further optimization and standardization. It is likely that ES, at least for the time being, will coexist with other NGS-based strategies, namely the targeted NGS and WGS in molecular diagnostics. Although, the total cost of WGS is much higher for the present, it is expected to become much more affordable soon (227). Obviously, one approach does not fit all different diagnostic applications and one needs to select the best approach based on available resources. For example,

a targeted NGS or ES approach is suitable for detecting mutations in disorders with genetic heterogeneity. Similarly, the diagnosis of X-linked disorders would require an NGS approach targeting the genes located on the X chromosome. We have used X-exome sequencing in two affected individuals from two unrelated families to detect the pathogenic mutation (c.5217G>A) in *FLNA* causing the X-linked Terminal Osseous Dysplasia (TOD) (**Chapter 4**). For other diseases such as mental retardation and congenital malformations, which are often due to copy number variations (CNV) or small mutations, the first step is to apply ES. Failure of ES to identify the causative genetic defect would suggest a possible extragenic location of the pathogenic mutation. WGS could then be used to detect deep intronic mutations or variants in remote regulatory elements. The advantage of WGS is that it does not require 30× or more coverage; so, sequencing of paired-end or mate-pair libraries with sufficient coverage across the genome is enough to identify CNVs, small mutations and chromosomal rearrangements. Recently, a cryptic fusion oncogene in acute promyelocytic leukemia was identified by WGS (228).

Definitive genetic diagnosis for a particular Mendelian disease cannot be established based on only one newly identified variant; for that, additional cases are required. However, it may be difficult to find additional cases in rare disorders to validate the potential candidate mutation. In that case, one has to perform functional studies of the putative pathological variant to confirm the pathogenicity. This method should be considered when the mutated gene has a key role in a well-defined molecular pathology of a certain disease. Indeed, the discovery of pathogenic variants and candidate genes responsible for different Mendelian disorders will help in understanding the function of the gene and the related biological pathways underlying health and disease. For example, we discovered that pathogenic variants in the *SMCHD1* (Structural Maintenance of Chromosomes Flexible Hinge Domain containing ***1***) gene responsible for Facioscapulohumeral Dystrophy type 2 (FSHD2) act as epigenetic modifiers of the D4Z4. This has provided new insights into the possible role of *SMCHD1* mutations in modifying the epigenetic repression of other genomic regions (**Chapter 7**).

In summary, the widespread availability of NGS technologies and the ever evolving field of gene sequencing is changing the approach to detecting genes, which cause Mendelian diseases as well as those responsible for complex traits. In particular, ES is expediting the detection of genes for

Mendelian diseases and the detection of mutations because of a rapid and straightforward laboratory workflow. Thousands of variants are identified per individual genome by NGS but only one or two of these variants may explain the Mendelian disease. Therefore, the interpretation of variants has become the new challenge. The number of variants that are identified in ES varies depending on the exome kit used, the NGS platform and the algorithms used for data mapping, and variant calling; this number can be huge. The following are, therefore, crucial for the identification of disease genes: prioritization of the variants, use of suitable bioinformatics tools and automated variants annotation algorithms, and the characterization of the functional impact of variants. It is expected that in due course standards and guidelines for ES or WGS will be established. The new technologies are therefore likely to become the most commonly used tools for the detection of genes for Mendelian diseases as well as for other diseases in the coming years.

References:

1. Peter J. Bowler. (1989) The Mendelian Revolution: The Emergence of Hereditarian Concepts in Modern Science and Society. Baltimore: Johns Hopkins University Press.
2. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods.
3. Sanger F, Nicklenl S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463-7.
4. Bonetta L. (2006) Genome sequencing in the fast lane. Nat Methods 3:141-147.
5. von Bubnoff A. (2008) Next-generation sequencing: the race is on. Cell 132: 721-723.
6. Schuster SC. (2008) Next-generation sequencing transforms today's biology. Nat Methods 5: 16-18.
7. Shendure J, Ji H. (2008) Next-generation DNA sequencing. Nat Biotechnol. 26:1135-45.
8. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.
9. Hofstra RM, Mulder IM, Vossen R, de Koning-Gans PA, Kraak M, Ginjaar IB, van der Hout AH, Bakker E, Buys CH, van Ommen GJ, van Essen AJ, den Dunnen JT. (2004) DGGE-based whole-gene mutation scanning of the dystrophin gene in Duchenne and Becker muscular dystrophy patients. Hum Mutat. 23: 57-66.
10. Bennett RR, den Dunnen J, O'Brien KF, Darras BT, Kunkel LM. (2001) Detection of mutations in the dystrophin gene via automated DHPLC screening and direct sequencing. BMC Genet. 2: 17.
11. Tuffery S, Moine P, Demaille J, Claustres M. (1993) Base substitutions in the human dystrophin gene: detection by using the single-strand conformation polymorphism (SSCP) technique. Hum Mutat. 2: 368-374.
12. Ashton EJ, Yau SC, Deans ZC, Abbs SJ. (2008) Simultaneous mutation scanning for gross deletions, duplications and point mutations in the DMD gene. Eur J Hum Genet. 16:53-61.
13. Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ. (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. *Clin. Chem.* 49: 853–60
14. Paul M, Zumhagen S, Stallmeyer B, Koopmann M, Spieker T, Schulze-Bahr E. (2009) Genes causing inherited forms of cardiomyopathies. A current compendium Herz. 34 :98-109.
15. Hartong DT, Berson EL, Dryja TP. (2006) Retinitis pigmentosa. Lancet 368:1795-809.
16. Hilgert N, Smith RJ, Van Camp G. (2009) Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics? Mutat Res. 681:189-96.

17. Noonan JA. (1994) Noonan syndrome. An update and review for the primary pediatrician. Clin Pediatr (Phila) 33:548-55.
18. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60–65.
19. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing.Nature 452: 872–876.
20. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53–59.
21. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, et al. (2010) Complete Khoisan and Bantu genomes from southern Africa. Nature 463: 943–947.
22. Turner EH, Ng SB, Nickerson DA, Shendure J. (2009) Methods for genomic partitioning. Annu Rev Genomics Hum Genet. 10: 263–284.
23. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase Science 239:487-91.
24. Edwards MC, Gibbs RA. (1994) Multiplex PCR: advantages, development, and applications. PCR Methods Appl. 3: S65–S75.
25. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4: 903-5.
26. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4: 907-9
27. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet. 39: 1522-7.
28. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 27: 182-9.
29. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig, CJ, Dahl F. et al. (2007) Multiplex amplification of large sets of human exons. Nat. Methods 4: 931–936

30. Li JB, Gao Y, Aach J, Zhang K, Kryukov GV, Xie B, Ahlford A, Yoon JK, Rosenbaum, AM, Zaranek AW, et al. (2009) Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. Genome Res. 19: 1606–1615.
31. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE. (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679.
32. Ansorge W, Sproat BS, Stegemann J, Schwager. (1986) A non-radioactive automated method for DNA sequence determination. J Biochem Biophys Methods 13:315–323.
33. Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. (1987) Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucleic Acids Res. 15:4593–4602.
34. Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, Erfle H, Caskey CT, Ansorge W. (1990) Automated DNA sequencing of the human HPRT locus. Genomics 6:593–608.
35. Krypuy M, Ahmed AA, Etemadmoghadam D, Hyland SJ, DeFazio A, Fox SB, Brenton JD, Bowtell DD, Dobrovic A. (2007) High resolution melting for mutation scanning of TP53 exons 5-8. BMC Cancer 7: 168.
36. Reed GH, Kent JO, Wittwer CT. (2007) High-resolution DNA melting analysis for simple and efficient molecular diagnostics. Pharmacogenomics 6: 597–608.
37. Pornprasert S, Phusua A, Suanta S, Saetung R, Sanguansermsri T. (2008) Detection of alpha-thalassemia-1 Southeast Asian type using real-time gap-PCR with SYBR Green1 and high resolution melting analysis. Eur. J. Haematol. 6510–4.
38. Herrmann MG, Durtschi JD, Bromley LK, Wittwer CT, Voelkerding KV. (2006) Amplicon DNA melting analysis for mutation scanning and genotyping: cross-platform comparison of instruments and dyes. Clin Chem. 52: 494-503.
39. Montgomery J, Wittwer CT, Palais R, Zhou L. (2007) Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis. Nat Protoc. 2: 59-66.
40. 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. 6:876-83.
41. Gundry CN, Vandersteen JG, Reed GH, Pryor RJ, Chen J, Wittwer CT. (2003) Amplicon melting analysis with labeled primers: a closed-tube method for differentiating homozygotes and heterozygotes. Clin Chem. 3:396-406.
42. Reed GH, Wittwer CT. (2004) Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin Chem. 10:1748-54.

43. Tucker KL. (2001) Methylated cytosine and the brain: a new base for neuroscience. *Neuron* 3: 649–652.
44. Fraga MF, Esteller M. (2002) DNA methylation: a profile of methods and applications. *BioTechniques* 3: 632–49.
45. Wojdacz TK, Dobrovic A. (2007) Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation. *Nucleic Acids Res.* 6: e41.
46. Lehmensiek A, Sutherland MW, McNamara RB. (2008) The use of high resolution melting (HRM) to map single nucleotide polymorphism markers linked to a covered smut resistance gene in barley. Theor Appl Genet. 5:721-8.
47. Cheng JC, Huang CL, Lin CC, Chen CC, Chang YC, Chang SS, Tseng CP. (2006) Rapid detection and identification of clinically important bacteria by high-resolution melting analysis after broad-range ribosomal RNA real-time PCR. Clin Chem. 11:1997-2004.
48. Zhou L, Vandersteen J, Wang L, Fuller T, Taylor M, Palais B, Wittwer CT. (2004) High-resolution DNA melting curve analysis to establish HLA genotypic identity. Tissue Antigens. 2:156-64.
49. 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 Breastcancer (kConFab), Sangrajrang S, Lesueur F, Tavtigian SV (2009) Description andvalidation of high-throughput simultaneous genotyping and mutation scanningby high-resolution melting curve analysis. Hum Mutat. 6:884-90.
50. Van Der Stoep N, van Paridon CDM, Janssens T, Krenkova P, Stambergova A, MacekM, Matthijs G, Bakker E(2009) Diagnostic guidelines for high resolution meltingcurve (HRM) analysis: an inter-laboratory validation of BRCA1 mutation scanning using the 96-well LightScannerTM. Hum Mutat. 6:899-909.
51. Zhou L, Myers AN, Vandersteen JG, Wang L, Wittwer CT. (2004) Closed-tube genotyping with unlabeled oligonucleotide probes and a saturating DNA dye. Clin Chem. 8: 1328–1335.
52. Erali M, Voelkerding KV, Wittwer CT. (2008) High resolution melting applications for clinical laboratory medicine. Exp Mol Pathol. 1:50-8.
53. Erali M, Palais R, Wittwer C. (2008) SNP genotyping by unlabeled probe melting analysis. Methods Mol Biol. 429:199-206. Erali M, Palais R, Wittwer C. (2008) SNP genotyping by unlabeled probe melting analysis. Methods Mol Biol. 429:199-206.
54. Wittwer CT. (2010) Making DNA melting useful. Clin Chem. 9:1500-1.

55. Chou LS, Lyon E, Wittwer CT. (2005) A comparison of high-resolution melting analysis with denaturing high-performance liquid chromatography for mutation scanning: cystic fibrosis transmembrane conductance regulator gene as a model. Am J Clin Pathol. 124: 330-338.
56. Willmore C, Holden JA, Zhou L, Tripp S, Wittwer CT, Layfield LJ. (2004) Detection of c-kit-activating mutations in gastrointestinal stromal tumors by high-resolution amplicon melting analysis. Am J Clin Pathol. 2:206-16.
57. Dobrowolski SF, McKinney JT, Amat di San Filippo C, Giak Sim K, Wilcken B, Longo N. (2005) Validation of dye-binding/high-resolution thermal denaturation for the identification of mutations in the SLC22A5 gene. Hum Mutat. 3:306-13.
58. Takano E, Mitchell G, Fox S, Dobrovic A. (2008) Rapid detection of carriers with *BRCA1* and *BRCA2* mutations using high resolution melting analysis. BMC Cancer 1: 59.
59. Palles C, Johnson N, Coupland B *et al.* (2008) Identification of genetic variants that influence circulating IGF1 levels: a targeted search strategy. Hum Mol Genet. 10: 1457–1464.
60. Dobrowolski SF, Ellingson C, Coyne T, Grey J, Martin R, Naylor EW, Koch R, Levy HL. (2007) Mutations in the phenylalanine hydroxylase gene identified in 95 patients with phenylketonuria using novel systems of mutation scanning and specific genotyping based upon thermal melt profiles.Mol Genet Metab. 3:218-27.
61. Olsen RK, Dobrowolski SF, Kjeldsen M, Hougaard D, Simonsen H, Gregersen N, Andresen BS. (2010) High-resolution melting analysis, a simple and effective method for reliable mutation scanning and frequency studies in the ACADVL gene. J Inherit Metab Dis. 3:247-60.
62. Yamey G. (2000) Scientists unveil first draft of human genome. BMJ. 321:7.
63. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.
64. Nyrén P, Lundin A. (1985) Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. Anal Biochem. 2:504-9.
65. Dressman D, Yan H, Traverso G, Kinzler KW,Vogelstein B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci USA. 100: 8817–8822.
66. Zheng Z, Advani A, Melefors O, Glavas S, Nordström H, Ye W, Engstrand L, Andersson AF. (2010) Titration-free massively parallel pyrosequencing using trace amounts of starting material. Nucleic Acids Res. 13:e137.
67. Mardis ER. (2008) Next-generation DNA sequencing methods. Annu Rev Genom Hum. 9: 387–402.

68. Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E. (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms.Nucleic Acids Res. 28, E87.
69. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 34, e22.
70. Ansorge WJ. (2009) Next-generation DNA sequencing techniques N Biotechnol. 4:195-203.
71. Housby JN, and Southern EM. (1998) Fidelity of DNA ligation: anovel experimental approach based on the polymerisation oflibraries of oligonucleotides. Nucleic Acids Res. 26: 4259–4266.
72. Bhaijee F, Pepper DJ, Pitman KT, Bell D. (2011) New developments in the molecular pathogenesis of head and neck tumors: a review of tumor-specific fusion oncogenes in mucoepidermoid carcinoma,adenoid cystic carcinoma, and NUT midline carcinoma. Ann Diagn Pathol. 15:69–77.
73. Fouse SD, Nagarajan RP, Costello JF. (2010) Genome-scale DNA methylation analysis. Epigenomics 2:105–117.
74. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao YJ, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4, 651–657.
75. Mikkelsen TS, Ku MC, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, BrockmanW, Kim TK, Koche RP, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553–560.
76. Aparicio O, Geisberg JV, Struhl K. (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol17-17 17.
77. Dodge JE, Ramsahoye BH, Wo ZG, Okano M, Li E. (2002) De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. Gene 1: 41–48.
78. Jacquier A. (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat RevGenet. 10:833–844.
79. Wang Z, Gerstein M, Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10:57–63.
80. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5: 613–619.
81. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628.

82. Tang FC, Barbacioru C, Nordman E, Li B, Xu NL, Bashkirov V I, Lao KQ, Surani MA. (2010) RNA-Seq analysis to capture the transcriptome landscape of a single cell. Nat Protoc. 5:516–535.
83. Sugarbaker DJ, Richards WG, Gordon GJ, Dong L, De Rienzo A, Maulik G, Glickman, JN, Chirieac LR, Hartman ML,Taillon BE, et al. (2008) Transcriptome sequencing of malignant pleural mesothelioma tumors. Proc Natl Acad Sci USA. 105:3521–3526.
84. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A,Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960.
85. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. (2009) Transcriptome Sequencing to Detect Gene Fusions in Cancer. Nature 7234: 97–101.
86. Eisen JA. (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol. 3:e82.
87. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC , Rohwer F. (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7: 57.
88. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. (2007) The human microbiome project. Nature 449: 804–810.
89. Qin JJ, Li RQ, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, et al, and the MetaHIT Consortium. (2010) A human gut microbialgene catalogue established by metagenomic sequencing. Nature 464: 59–65.
90. Goldberg SMD, Johnson J, Busam D, Feldblyum T, Ferriera S,Friedman R, Halpern A, Khouri H, Kravitz SA, Lauro FM, et al. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci USA. 103: 11240–11245.
91. Durfee T, Nelson R, Baldwin S, Plunkett G, Burland V, Mau B, Petrosino JF, Qin X, Muzny DM, Ayele M, et al. (2008) The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. J Bacteriol. 190:2597–2606.
92. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2, e1326.

93. Diguistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, *et al* (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol. 9:R94.
94. D'Argenio V, Petrillo M, Cantiello P, Naso B, Cozzuto L, Notomista E, Paolella G, Di Donato A, Salvatore F. (2011) De novo sequencing and assembly of the whole genome of Novosphingobium sp. strain PP1Y. J Bacteriol. 16:4296.
95. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q *et al.* (2010) The sequence and de novo assembly of the giant panda genome. Nature 7279:311-7.
96. Imelfort M, Edwards D. (2009) De novo sequencing of plant genomes using second-generation technologies. Brief Bioinform. 6:609-18.
97. Li Y, Zheng H, Luo R, Wu H, Zhu H, Li R, Cao H, Wu B, Huang S, Shao H, Ma H, Zhang F, Feng S, Zhang W, Du H, Tian G, Li J, Zhang X, Li S, Bolund L, Kristiansen K, de Smith AJ, Blakemore AI, Coin LJ, Yang H, Wang J. (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. Nat Biotechnol. 8:723-30.
98. Huang S, Du Y, Wang X, et al. (2009) Recent developments of the cucumber genome initiative—an international effort to unlock the genetic potential of an orphan crop using novel genomic technology. Plant and Animal Genomes XVII, San Diego, USA.
99. NowrousianM. (2010) Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. Eukaryot Cell. 91300-10.
100. Pushkarev D, Neff NF, Quake SR. (2009) Single-molecule sequencing of an individual human genome. Nat Biotechnol. 27:847–852.
101. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, et al. (2009) A highly annotated whole genome sequence of a Korean individual. Nature 460:1011–1015.
102. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, et al. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 19:1622–1629.
103. Fredriksson S, Baner J, Dahl F, Chu A, Ji H, et al. (2007) Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. Nucleic Acids Res. 35:e47.
104. Meuzelaar LS, Lancaster O, Pasche JP, Kopal G, Brookes AJ. (2007) MegaPlex PCR: a strategy for multiplex amplification. Nat. Methods 4:835–37.
105. Varley KE, Mitra RD. (2008) Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Genome Res. 18:1844–50.

106. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-9.
107. Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M. (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res. 33:e71.
108. Dahl F, Stenberg J, Fredriksson S,Welch K, Zhang M, et al. (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc. Natl. Acad. Sci. USA 104:9387–92.
109. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M (2005) Direct genomic selection. Nat. Methods 2:63–69.
110. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111-8.
111. Barnes WM (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc. Natl. Acad. Sci. USA. 91:2216–2220.
112. Out AA, van Minderhout IJ, Goeman JJ, Ariyurek Y, Ossowski S, Schneeberger K, Weigel D, van Galen M, Taschner PE, Tops CM, Breuning MH, van Ommen GJ, den Dunnen JT, Devilee P, Hes FJ (2009) Deep sequencing to reveal new variants in pooled DNA samples.Hum Mutat. 30:1703-12.
113. Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary BP, Landegren U (1994) Padlock probes: circularizing oligonucleotides for localized DNA detection. Science 265:2085–2088.
114. Landegren U, Schallmeiner E, Nilsson M, Fredriksson S, Banér J, Gullberg M, Jarvius J, Gustafsdottir S, Dahl F, Söderberg O, Ericsson O, Stenberg J (2004) Molecular tools for a molecular medicine: analyzing genes, transcripts and proteins using padlock and proximity probes. J. Mol. Recognit. 17:194–197.
115. Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet. 19: 45-51.
116. McKusick V A. (2007) Mendelian Inheritance in Man and its online version, OMIM. Am. J. Hum. Genet. 80:588–604.
117. Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet. 7:277-282.

118. Lander ES, Botstein D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567-1570.
119. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073-1080.
120. Biesecker L G (2010) Exome sequencing makes medical genomics a reality. Nature Genet. 42:13–14.
121. 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, Shendure J (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 42:790-793.
122. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, van Lier B, Arts P, Wieskamp N, del Rosario M, van Bon BW, Hoischen A, de Vries BB, Brunner HG, Veltman JA (2010) A de novo paradigm for mental retardation. Nat Genet. 42:1109-1112.
123. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 43:585-589.
124. Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M (2010) Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet. 87:90-4.
125. Gilissen C, Hoischen A, Brunner HG, Veltman JA (2011) Unlocking Mendelian disease using exome sequencing. Genome Biol. 12:228.
126. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. (2012) Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet. 57:621-32.
127. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 42:30-35.
128. Bolze A, Byun M, McDonald D, Morgan NV, Abhyankar A, Premkumar L, Puel A, Bacon CM, Rieux-Laucat F, Pang K, Britland A, Abel L, Cant A, Maher ER, Riedl SJ, Hambleton S, Casanova JL. (2010) Whole-exome-sequencing-based discovery of human FADD deficiency. Am J Hum Genet. 87:873-881.

129. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, Sougnez C, Garimella KV, Fisher S, Abreu J, Barry AJ, Fennell T, Banks E, Ambrogio L, Cibulskis K, Kernytsky A, Gonzalez E, Rudzicz N, Engert JC, DePristo MA, Daly MJ, Cohen JC, Hobbs HH, Altshuler D, Schonfeld G, Gabriel SB, Yue P, Kathiresan S. (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med. 363:2220-2227.
130. Pierce SB, Walsh T, Chisholm KM, Lee MK, Thornton AM, Fiumara A, Opitz JM, Levy-Lahad E, Klevit RE, King MC. (2010) Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. Am J Hum Genet. 87:282-288.
131. Haack TB, Danhauser K, Haberberger B, Hoser J, Strecker V, Boehm D, Uziel G, Lamantea E, Invernizzi F, Poulton J, Rolinski B, Iuso A, Biskup S, Schmidt T, Mewes HW, Wittig I, Meitinger T, Zeviani M, Prokisch H. (2010) Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. Nat Genet. 42:1131-1134.
132. Krawitz PM, Schweiger MR, Rödelsperger C, Marcelis C, Kölsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Köhler S, Jäger M, Grünhagen J, de Condor BJ, Doelken S, Brunner HG, Meinecke P, Passarge E, Thompson MD, Cole DE, Horn D, Roscioli T, Mundlos S, Robinson PN. (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet. 42:827-829.
133. Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG, Roepman R, Knoers NV, Veltman JA, Brunner HG. (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet. 87:418-423.
134. Bilgüvar K, Oztürk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoğlu D, Tüysüz B, Cağlayan AO, Gökben S, Kaymakçalan H, Barak T, Bakircioğlu M, Yasuno K, Ho W, Sanders S, Zhu Y, Yilmaz S, Dinçer A, Johnson MH, Bronen RA, Koçer N, Per H, Mane S, Pamir MN, Yalçinkaya C, Kumandaş S, Topçu M, Ozmen M, Sestan N, Lifton RP, State MW, Günel M. (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 467:207-210.
135. Sirmaci A, Walsh T, Akay H, Spiliopoulos M, Sakalar YB, Hasanefendioğlu-Bayrak A, Duman D, Farooq A, King MC, Tekin M. (2010) MASP1 mutations in patients with facial, umbilical, coccygeal, and auditory findings of Carnevale, Malpuech, OSA, and Michels syndromes. Am J Hum Genet. 87:679-86.

136. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, de Vries P, de Reuver R, 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, Veltman JA. (2010) De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 42:483-485.
137. 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, Tang BS. (2010) TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain.133:3510-3518.
138. 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. (2010) Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. Am J Hum Genet. 87:146-153.
139. Çalışkan M, Chong JX, Uricchio L, Anderson R, Chen P, Sougnez C, Garimella K, Gabriel SB, dePristo MA, Shakir K, Matern D, Das S, Waggoner D, Nicolae DL, Ober C. (2011) Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. Hum Mol Genet. 20:1285-1289.
140. Züchner S, Dallman J, Wen R, Beecham G, Naj A, Farooq A, Kohli MA, Whitehead PL, Hulme W, Konidari I, Edwards YJ, Cai G, Peter I, Seo D, Buxbaum JD, Haines JL, Blanton S, Young J, Alfonso E, Vance JM, Lam BL, Peričak-Vance MA. (2011) Whole-exome sequencing links a variant in DHDDS to retinitis pigmentosa. Am J Hum Genet. 88:201-206.
141. Becker J, Semler O, Gilissen C, Li Y, Bolz HJ, Giunta C, Bergmann C, Rohrbach M, Koerber F, Zimmermann K, de Vries P, Wirth B, Schoenau E, Wollnik B, Veltman JA, Hoischen A, Netzer C. (2011) Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. Am J Hum Genet. 88:362-371.
142. Glazov EA, Zankl A, Donskoi M, Kenna TJ, Thomas GP, Clark GR, Duncan EL, Brown MA. (2011) Whole-exome re-sequencing in a family quartet identifies POP1 mutations as the cause of a novel skeletal dysplasia. PLoS Genet. 7:e1002027.
143. Sloan JL, Johnston JJ, Manoli I, Chandler RJ, Krause C, Carrillo-Carrasco N, Chandrasekaran SD, Sysol JR, O'Brien K, Hauser NS, Sapp JC, Dorward HM, Huizing M; NIH Intramural Sequencing Center Group, Barshop BA, Berry SA, James PM, Champaigne NL, de Lonlay P, Valayannopoulos V, Geschwind MD, Gavrilov DK, Nyhan WL, Biesecker LG, Venditti CP. (2011) Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria. Nat Genet. 43:883-6.

144. O'Sullivan J, Bitu CC, Daly SB, Urquhart JE, Barron MJ, Bhaskar SS, Martelli-Júnior H, dos Santos Neto PE, Mansilla MA, Murray JC, Coletta RD, Black GC, Dixon MJ. (2011) Whole-Exome sequencing identifies FAM20A mutations as a cause of amelogenesis imperfecta and gingival hyperplasia syndrome. Am J Hum Genet. 88:616-20.
145. Vissers LE, Lausch E, Unger S, Campos-Xavier AB, Gilissen C, Rossi A, Del Rosario M, Venselaar H, Knoll U, Nampoothiri S, Nair M, Spranger J, Brunner HG, Bonafé L, Veltman JA, Zabel B, Superti-Furga A. (2011) Chondrodysplasia and abnormal joint development associated with mutations in IMPAD1, encoding the Golgi-resident nucleotide phosphatase, gPAPP. Am J Hum Genet. 88:608-615.
146. Puente XS, Quesada V, Osorio FG, Cabanillas R, Cadiñanos J, Fraile JM, Ordóñez GR, Puente DA, Gutiérrez-Fernández A, Fanjul-Fernández M, Lévy N, Freije JM, López-Otín C. (2011) Exome sequencing and functional analysis identifies BANF1 mutation as the cause of a hereditary progeroid syndrome. Am J Hum Genet. 88:650-656.
147. Götz A, Tyynismaa H, Euro L, Ellonen P, Hyötyläinen T, Ojala T, Hämäläinen RH, Tommiska J, Raivio T, Oresic M, Karikoski R, Tammela O, Simola KO, Paetau A, Tyni T, Suomalainen A. (2011) Exome sequencing identifies mitochondrial alanyl-tRNA synthetase mutations in infantile mitochondrial cardiomyopathy. Am J Hum Genet. 88:635-42.
148. Tariq M, Belmont JW, Lalani S, Smolarek T, Ware SM. (2011) SHROOM3 is a novel candidate for heterotaxy identified by whole exome sequencing. Genome Biol.12:R91.
149. Snape K, Hanks S, Ruark E, Barros-Núñez P, Elliott A, Murray A, Lane AH, Shannon N, Callier P, Chitayat D, Clayton-Smith J, Fitzpatrick DR, Gisselsson D, Jacquemont S, Asakura-Hay K, Micale MA, Tolmie J, Turnpenny PD, Wright M, Douglas J, Rahman N. (2011) Mutations in CEP57 cause mosaic variegated aneuploidy syndrome. Nat Genet. 43:527-9.
150. Saitsu H, Osaka H, Sasaki M, Takanashi J, Hamada K, Yamashita A, Shibayama H, Shiina M, Kondo Y, Nishiyama K, Tsurusaki Y, Miyake N, Doi H, Ogata K, Inoue K, Matsumoto N. (2011) Mutations in POLR3A and POLR3B encoding RNA Polymerase III subunits cause an autosomal-recessive hypomyelinating leukoencephalopathy. Am J Hum Genet. 89:644-651.
151. Pierson TM, Adams D, Bonn F, Martinelli P, Cherukuri PF, Teer JK, Hansen NF, Cruz P, Mullikin For The Nisc Comparative Sequencing Program JC, Blakesley RW, Golas G, Kwan J, Sandler A, Fuentes Fajardo K, Markello T, Tifft C, Blackstone C, Rugarli EI, Langer T, Gahl WA, Toro C. (2011) Whole-exome sequencing identifies homozygous AFG3L2 mutations in a spastic ataxia-neuropathy syndrome linked to mitochondrial m-AAA proteases. PLoS Genet. 7:e1002325.

152. Theis JL, Sharpe KM, Matsumoto ME, Chai HS, Nair AA, Theis JD, de Andrade M, Wieben ED, Michels VV, Olson TM. (2011) Homozygosity mapping and exome sequencing reveal GATAD1 mutation in autosomal recessive dilated cardiomyopathy. Circ Cardiovasc Genet. 4:585-94.
153. Zangen D, Kaufman Y, Zeligson S, Perlberg S, Fridman H, Kanaan M, Abdulhadi-Atwan M, Abu Libdeh A, Gussow A, Kisslov I, Carmel L, Renbaum P, Levy-Lahad E. (2011) XX ovarian dysgenesis is caused by a PSMC3IP/HOP2 mutation that abolishes coactivation of estrogen-driven transcription. Am J Hum Genet. 89:572-9.
154. Takata A, Kato M, Nakamura M, Yoshikawa T, Kanba S, Sano A, Kato T. (2011) Exome sequencing identifies a novel missense variant in RRM2B associated with autosomal recessive progressive external ophthalmoplegia. Genome Biol. 12:R92.
155. Aldahmesh MA, Khan AO, Mohamed JY, Alkuraya H, Ahmed H, Bobis S, Al-Mesfer S, Alkuraya FS. (2011) Identification of ADAMTS18 as a gene mutated in Knobloch syndrome. J Med Genet. 48:597-601.
156. Doi H, Yoshida K, Yasuda T, Fukuda M, Fukuda Y, Morita H, Ikeda S, Kato R, Tsurusaki Y, Miyake N, Saitsu H, Sakai H, Miyatake S, Shiina M, Nukina N, Koyano S, Tsuji S, Kuroiwa Y, Matsumoto N. (2011) Exome sequencing reveals a homozygous SYT14 mutation in adult-onset, autosomal-recessive spinocerebellar ataxia with psychomotor retardation. Am J Hum Genet. 89:320-7.
157. Shaheen R, Faqeih E, Sunker A, Morsy H, Al-Sheddi T, Shamseldin HE, Adly N, Hashem M, Alkuraya FS. (2011) Recessive mutations in DOCK6, encoding the guanidine nucleotide exchange factor DOCK6, lead to abnormal actin cytoskeleton organization and Adams-Oliver syndrome. Am J Hum Genet. 89:328-33.
158. Sanna-Cherchi S, Burgess KE, Nees SN, Caridi G, Weng PL, Dagnino M, Bodria M, Carrea A, Allegretta MA, Kim HR, Perry BJ, Gigante M, Clark LN, Kisselev S, Cusi D, Gesualdo L, Allegri L, Scolari F, D'Agati V, Shapiro LS, Pecoraro C, Palomero T, Ghiggeri GM, Gharavi AG. (2011) Exome sequencing identified MYO1E and NEIL1 as candidate genes for human autosomal recessive steroid-resistant nephrotic syndrome. Kidney Int. 80:389-96.
159. Barak T, Kwan KY, Louvi A, Demirbilek V, Saygı S, Tüysüz B, Choi M, Boyacı H, Doerschner K, Zhu Y, Kaymakçalan H, Yılmaz S, Bakırcıoğlu M, Cağlayan AO, Oztürk AK, Yasuno K, Brunken WJ, Atalar E, Yalçınkaya C, Dinçer A, Bronen RA, Mane S, Ozçelik T, Lifton RP, Sestan N, Bilgüvar K, Günel M. (2011) Recessive LAMC3 mutations cause malformations of occipital cortical development. Nat Genet. 43:590-4.

160. Abou Jamra R, Philippe O, Raas-Rothschild A, Eck SH, Graf E, Buchert R, Borck G, Ekici A, Brockschmidt FF, Nöthen MM, Munnich A, Strom TM, Reis A, Colleaux L. (2011) Adaptor protein complex 4 deficiency causes severe autosomal-recessive intellectual disability, progressive spastic paraplegia, shy character, and short stature. Am J Hum Genet. 88:788-795.
161. Galmiche L, Serre V, Beinat M, Assouline Z, Lebre AS, Chretien D, Nietschke P, Benes V, Boddaert N, Sidi D, Brunelle F, Rio M, Munnich A, Rötig A. (2011) Exome sequencing identifies MRPL3 mutation in mitochondrial cardiomyopathy. Hum Mutat. 32:1225-31.
162. Ozgül RK, Siemiatkowska AM, Yücel D, Myers CA, Collin RW, Zonneveld MN, Beryozkin A, Banin E, Hoyng CB, van den Born LI; European Retinal Disease Consortium, Bose R, Shen W, Sharon D, Cremers FP, Klevering BJ, den Hollander AI, Corbo JC. (2011) Exome sequencing and cis-regulatory mapping identify mutations in MAK, a gene encoding a regulator of ciliary length, as a cause of retinitis pigmentosa. Am J Hum Genet. 89:253-64.
163. Hanson D, Murray PG, O'Sullivan J, Urquhart J, Daly S, Bhaskar SS, Biesecker LG, Skae M, Smith C, Cole T, Kirk J, Chandler K, Kingston H, Donnai D, Clayton PE, Black GC. (2011) Exome sequencing identifies CCDC8 mutations in 3-M syndrome, suggesting that CCDC8 contributes in a pathway with CUL7 and OBSL1 to control human growth. Am J Hum Genet. 89:148-53.
164. Kalay E, Yigit G, Aslan Y, Brown KE, Pohl E, Bicknell LS, Kayserili H, Li Y, Tüysüz B, Nürnberg G, Kiess W, Koegl M, Baessmann I, Buruk K, Toraman B, Kayipmaz S, Kul S, Ikbal M, Turner DJ, Taylor MS, Aerts J, Scott C, Milstein K, Dollfus H, Wieczorek D, Brunner HG, Hurles M, Jackson AP, Rauch A, Nürnberg P, Karagüzel A, Wollnik B. (2011) CEP152 is a genome maintenance protein disrupted in Seckel syndrome. Nat Genet. 43:23-26.
165. Bjursell MK, Blom HJ, Cayuela JA, Engvall ML, Lesko N, Balasubramaniam S, Brandberg G, Halldin M, Falkenberg M, Jakobs C, Smith D, Struys E, von Döbeln U, Gustafsson CM, Lundeberg J, Wedell A. (2011) Adenosine kinase deficiency disrupts the methionine cycle and causes hypermethioninemia, encephalopathy, and abnormal liver function. Am J Hum Genet. 89:507-515.
166. Bredrup C, Saunier S, Oud MM, Fiskerstrand T, Hoischen A, Brackman D, Leh SM, Midtbø M, Filhol E, Bole-Feysot C, Nitschké P, Gilissen C, Haugen OH, Sanders JS, Stolte-Dijkstra I, Mans DA, Steenbergen EJ, Hamel BC, Matignon M, Pfundt R, Jeanpierre C, Boman H, Rødahl E, Veltman JA, Knappskog PM, Knoers NV, Roepman R, Arts HH. (2011) Ciliopathies with skeletal anomalies and renal insufficiency due to mutations in the IFT-A gene WDR19. Am J Hum Genet. 89:634-643.

167. Aldahmesh MA, Mohamed JY, Alkuraya HS, Verma IC, Puri RD, Alaiya AA, Rizzo WB, Alkuraya FS. (21011) Recessive mutations in ELOVL4 cause ichthyosis, intellectual disability, and spastic quadriplegia. Am J Hum Genet. 89:745-50.
168. Dauber A, Nguyen TT, Sochett E, Cole DE, Horst R, Abrams SA, Carpenter TO, Hirschhorn JN. (2011) Genetic defect in CYP24A1, the vitamin D 24-hydroxylase gene, in a patient with severe infantile hypercalcemia. J Clin Endocrinol Metab. 97:E268-74.
169. Logan CV, Lucke B, Pottinger C, Abdelhamed ZA, Parry DA, Szymanska K, Diggle CP, van Riesen A, Morgan JE, Markham G, Ellis I, Manzur AY, Markham AF, Shires M, Helliwell T, Scoto M, Hübner C, Bonthron DT, Taylor GR, Sheridan E, Muntoni F, Carr IM, Schuelke M, Johnson CA. (2011) Mutations in MEGF10, a regulator of satellite cell myogenesis, cause early onset myopathy, areflexia, respiratory distress and dysphagia (EMARDD). Nat Genet. 43:1189-92.
170. Albers CA, Cvejic A, Favier R, Bouwmans EE, Alessi MC, Bertone P, Jordan G, Kettleborough RN, Kiddle G, Kostadima M, Read RJ, Sipos B, Sivapalaratnam S, Smethurst PA, Stephens J, Voss K, Nurden A, Rendon A, Nurden P, Ouwehand WH. (2011) Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. Nat Genet. 43:735-7.
171. de Greef JC, Wang J, Balog J, den Dunnen JT, Frants RR, Straasheijm KR, Aytekin C, van der Burg M, Duprez L, Ferster A, Gennery AR, Gimelli G, Reisli I, Schuetz C, Schulz A, Smeets DF, Sznajer Y, Wijmenga C, van Eggermond MC, van Ostaijen-Ten Dam MM, Lankester AC, van Tol MJ, van den Elsen PJ, Weemaes CM, van der Maarel SM. (2011) Mutations in ZBTB24 are associated with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2. Am J Hum Genet. 88:796-804.
172. Sergouniotis PI, Davidson AE, Mackay DS, Li Z, Yang X, Plagnol V, Moore AT, Webster AR. (2011) Recessive mutations in KCNJ13, encoding an inwardly rectifying potassium channel subunit, cause leber congenital amaurosis. Am J Hum Genet. 89:183-90.
173. Erlich Y, Edvardson S, Hodges E, Zenvirt S, Thekkat P, Shaag A, Dor T, Hannon GJ, Elpeleg O. (2011) Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res. 21:658-64
174. Clayton-Smith J, O'Sullivan J, Daly S, Bhaskar S, Day R, Anderson B, Voss AK, Thomas T, Biesecker LG, Smith P, Fryer A, Chandler KE, Kerr B, Tassabehji M, Lynch SA, Krajewska-Walasek M, McKee S, Smith J, Sweeney E, Mansour S, Mohammed S, Donnai D, Black G. (2011) Whole-exome-sequencing identifies mutations in histone acetyltransferase gene KAT6B in individuals with the Say-Barber-Biesecker variant of Ohdo syndrome. Am J Hum Genet. 89:675-681.

175. Chen WJ, Lin Y, Xiong ZQ, Wei W, Ni W, Tan GH, Guo SL, He J, Chen YF, Zhang QJ, Li HF, Lin Y, Murong SX, Xu J, Wang N, Wu ZY. (2011) Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia. Nat Genet. 43:1252-5.
176. Simpson MA, Irving MD, Asilmaz E, Gray MJ, Dafou D, Elmslie FV, Mansour S, Holder SE, Brain CE, Burton BK, Kim KH, Pauli RM, Aftimos S, Stewart H, Kim CA, Holder-Espinasse M, Robertson SP, Drake WM, Trembath RC. (2011) Mutations in NOTCH2 cause Hajdu-Cheney syndrome, a disorder of severe and progressive bone loss. Nat Genet. 43:303-5.
177. Hoischen A, van Bon BW, Rodríguez-Santiago B, Gilissen C, Vissers LE, de Vries P, Janssen I, van Lier B, Hastings R, Smithson SF, Newbury-Ecob R, Kjaergaard S, Goodship J, McGowan R, Bartholdi D, Rauch A, Peippo M, Cobben JM, Wieczorek D, Gillessen-Kaesbach G, Veltman JA, Brunner HG, de Vries BB. (2011) De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat Genet. 43:729-31.
178. Rademakers R, Baker M, Nicholson AM, Rutherford NJ, Finch N, Soto-Ortolaza A, Lash J, Wider C, Wojtas A, DeJesus-Hernandez M, Adamson J, Kouri N, Sundal C, Shuster EA, Aasly J, MacKenzie J, Roeber S, Kretzschmar HA, Boeve BF, Knopman DS, Petersen RC, Cairns NJ, Ghetti B, Spina S, Garbern J, Tselis AC, Uitti R, Das P, Van Gerpen JA, Meschia JF, Levy S, Broderick DF, Graff-Radford N, Ross OA, Miller BB, Swerdlow RH, Dickson DW, Wszolek ZK. (2011) Mutations in the colony stimulating factor 1 receptor (CSF1R) gene cause hereditary diffuse leukoencephalopathy with spheroids. Nat Genet. 44:200-205.
179. Min BJ, Kim N, Chung T, Kim OH, Nishimura G, Chung CY, Song HR, Kim HW, Lee HR, Kim J, Kang TH, Seo ME, Yang SD, Kim DH, Lee SB, Kim JI, Seo JS, Choi JY, Kang D, Kim D, Park WY, Cho TJ. (2011) Whole-exome sequencing identifies mutations of KIF22 in spondyloepimetaphyseal dysplasia with joint laxity, leptodactylic type. Am J Hum Genet. 89:760-766.
180. Nosková L, Stránecký V, Hartmannová H, Přistoupilová A, Barešová V, Ivánek R, Hůlková H, Jahnová H, van der Zee J, Staropoli JF, Sims KB, Tyynelä J, Van Broeckhoven C, Nijssen PC, Mole SE, Elleder M, Kmoch S. (2011) Mutations in DNAJC5, encoding cysteine-string protein alpha, cause autosomal-dominant adult-onset neuronal ceroid lipofuscinosis. Am J Hum Genet. 89:241-52.
181. Sirmaci A, Spiliopoulos M, Brancati F, Powell E, Duman D, Abrams A, Bademci G, Agolini E, Guo S, Konuk B, Kavaz A, Blanton S, Digilio MC, Dallapiccola B, Young J, Zuchner S, Tekin M. (2011) Mutations in ANKRD11 cause KBG syndrome, characterized by intellectual disability, skeletal malformations, and macrodontia. Am J Hum Genet. 89:289-94.

182. Dickinson RE, Griffin H, Bigley V, Reynard LN, Hussain R, Haniffa M, Lakey JH, Rahman T, Wang XN, McGovern N, Pagan S, Cookson S, McDonald D, Chua I, Wallis J, Cant A, Wright M, Keavney B, Chinnery PF, Loughlin J, Hambleton S, Santibanez-Koref M, Collin M. (2011) Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency. Blood 118:2656-8.
183. Vilariño-Güell C, Wider C, Ross OA, Dachsel JC, Kachergus JM, Lincoln SJ, Soto-Ortolaza AI, Cobb SA, Wilhoite GJ, Bacon JA, Behrouz B, Melrose HL, Hentati E, Puschmann A, Evans DM, Conibear E, Wasserman WW, Aasly JO, Burkhard PR, Djaldetti R, Ghika J, Hentati F, Krygowska-Wajs A, Lynch T, Melamed E, Rajput A, Rajput AH, Solida A, Wu RM, Uitti RJ, Wszolek ZK, Vingerhoets F, Farrer MJ. (2011) VPS35 mutations in Parkinson disease. Am J Hum Genet. 89:162-7.
184. Klein CJ, Botuyan MV, Wu Y, Ward CJ, Nicholson GA, Hammans S, Hojo K, Yamanishi H, Karpf AR, Wallace DC, Simon M, Lander C, Boardman LA, Cunningham JM, Smith GE, Litchy WJ, Boes B, Atkinson EJ, Middha S, B Dyck PJ, Parisi JE, Mer G, Smith DI, Dyck PJ. (2011) Mutations in DNMT1 cause hereditary sensory neuropathy with dementia and hearing loss. Nat Genet. 43:595-600.
185. Norton N, Li D, Rieder MJ, Siegfried JD, Rampersaud E, Züchner S, Mangos S, Gonzalez-Quintana J, Wang L, McGee S, Reiser J, Martin E, Nickerson DA, Hershberger RE. (2011) Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy. Am J Hum Genet. 88:273-282.
186. Shi Y, Li Y, Zhang D, Zhang H, Li Y, Lu F, Liu X, He F, Gong B, Cai L, Li R, Liao S, Ma S, Lin H, Cheng J, Zheng H, Shan Y, Chen B, Hu J, Jin X, Zhao P, Chen Y, Zhang Y, Lin Y, Li X, Fan Y, Yang H, Wang J, Yang Z. (2011) Exome sequencing identifies ZNF644 mutations in high myopia. PLoS Genet. 7:e1002084.
187. Bowne SJ, Humphries MM, Sullivan LS, Kenna PF, Tam LC, Kiang AS, Campbell M, Weinstock GM, Koboldt DC, Ding L, Fulton RS, Sodergren EJ, Allman D, Millington-Ward S, Palfi A, McKee A, Blanton SH, Slifer S, Konidari I, Farrar GJ, Daiger SP, Humphries P. (2011) A dominant mutation in RPE65 identified by whole-exome sequencing causes retinitis pigmentosa with choroidal involvement. Eur J Hum Genet. 19:1074-81.
188. Weedon MN, Hastings R, Caswell R, Xie W, Paszkiewicz K, Antoniadi T, Williams M, King C, Greenhalgh L, Newbury-Ecob R, Ellard S. (2011) Exome sequencing identifies a DYNC1H1 mutation in a large pedigree with dominant axonal Charcot-Marie-Tooth disease. Am J Hum Genet. 89:308-12.

189. Zhou C, Zang D, Jin Y, Wu H, Liu Z, Du J, Zhang J. (2011) Mutation in ribosomal protein L21 underlies hereditary hypotrichosis simplex. Hum Mutat. 32:710-714.
190. Le Goff C, Mahaut C, Wang LW, Allali S, Abhyankar A, Jensen S, Zylberberg L, Collod-Beroud G, Bonnet D, Alanay Y, Brady AF, Cordier MP, Devriendt K, Genevieve D, Kiper PÖ, Kitoh H, Krakow D, Lynch SA, Le Merrer M, Mégarbane A, Mortier G, Odent S, Polak M, Rohrbach M, Sillence D, Stolte-Dijkstra I, Superti-Furga A, Rimoin DL, Topouchian V, Unger S, Zabel B, Bole-Feysot C, Nitschke P, Handford P, Casanova JL, Boileau C, Apte SS, Munnich A, Cormier-Daire V. (2011) Mutations in the TGFβ binding-protein-like domain 5 of FBN1 are responsible for acromicric and geleophysic dysplasias. Am J Hum Genet. 89:7-14.
191. Le Goff C, Mahaut C, Abhyankar A, Le Goff W, Serre V, Afenjar A, Destrée A, di Rocco M, Héron D, Jacquemont S, Marlin S, Simon M, Tolmie J, Verloes A, Casanova JL, Munnich A, Cormier-Daire V. (2011) Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre syndrome. Nat Genet. 44:85-88.
192. Tsurusaki Y, Osaka H, Hamanoue H, Shimbo H, Tsuji M, Doi H, Saitsu H, Matsumoto N, Miyake N. (2011) Rapid detection of a mutation causing X-linked leucoencephalopathy by exome sequencing. J Med Genet. 48:606-9.
193. Shamseldin HE, Faden MA, Alashram W, Alkuraya FS. (2012) Identification of a novel DLX5 mutation in a family with autosomal recessive split hand and foot malformation. J Med Genet. 49:16-20.
194. Khan K, Logan CV, McKibbin M, Sheridan E, Elçioglu NH, Yenice O, Parry DA, Fernandez-Fuentes N, Abdelhamed ZI, Al-Maskari A, Poulter JA, Mohamed MD, Carr IM, Morgan JE, Jafri H, Raashid Y, Taylor GR, Johnson CA, Inglehearn CF, Toomes C, Ali M. (2012) Next generation sequencing identifies mutations in Atonal homolog 7 (ATOH7) in families with global eye developmental defects. Hum Mol Genet. 21:776-83.
195. Zhang Z, Xia W, He J, Zhang Z, Ke Y, Yue H, Wang C, Zhang H, Gu J, Hu W, Fu W, Hu Y, Li M, Liu Y. (2012) Exome sequencing identifies SLCO2A1 mutations as a cause of primary hypertrophic osteoarthropathy. Am J Hum Genet. 90:125-32.
196. Mitchell K, O'Sullivan J, Missero C, Blair E, Richardson R, Anderson B, Antonini D, Murray JC, Shanske AL, Schutte BC, Romano RA, Sinha S, Bhaskar SS, Black GC, Dixon J, Dixon MJ. ((2012) Exome sequence identifies RIPK4 as the Bartsocas-Papas syndrome locus. Am J Hum Genet. 90:69-75.
197. Walne AJ, Dokal A, Plagnol V, Beswick R, Kirwan M, de la Fuente J, Vulliamy T, Dokal I. (2012) Exome sequencing identifies MPL as a causative gene in familial aplastic anemia. Haematologica 97:524-8.

198. Cabral RM, Kurban M, Wajid M, Shimomura Y, Petukhova L, Christiano AM. (2012) Whole-exome sequencing in a single proband reveals a mutation in the CHST8 gene in autosomal recessive peeling skin syndrome. Genomics 99:202-8.
199. Mayr JA, Haack TB, Graf E, Zimmermann FA, Wieland T, Haberberger B, Superti-Furga A, Kirschner J, Steinmann B, Baumgartner MR, Moroni I, Lamantea E, Zeviani M, Rodenburg RJ, Smeitink J, Strom TM, Meitinger T, Sperl W, Prokisch H. (2012) Lack of the mitochondrial protein acylglycerol kinase causes Sengers syndrome. Am J Hum Genet. 90:314-20.
200. Boyden LM, Choi M, Choate KA, Nelson-Williams CJ, Farhi A, Toka HR, Tikhonova IR, Bjornson R, Mane SM, Colussi G, Lebel M, Gordon RD, Semmekrot BA, Poujol A, Välimäki MJ, De Ferrari ME, Sanjad SA, Gutkin M, Karet FE, Tucci JR, Stockigt JR, Keppler-Noreuil KM, Porter CC, Anand SK, Whiteford ML, Davis ID, Dewar SB, Bettinelli A, Fadrowski JJ, Belsha CW, Hunley TE, Nelson RD, Trachtman H, Cole TR, Pinsk M, Bockenhauer D, Shenoy M, Vaidyanathan P, Foreman JW, Rasoulpour M, Thameem F, Al-Shahrouri HZ, Radhakrishnan J, Gharavi AG, Goilav B, Lifton RP. (2012) Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. Nature. 482:98-102.
201. Gibson WT, Hood RL, Zhan SH, Bulman DE, Fejes AP, Moore R, Mungall AJ, Eydoux P, Babul-Hirji R, An J, Marra MA; FORGE Canada Consortium, Chitayat D, Boycott KM, Weaver DD, Jones SJ. (2012) Mutations in EZH2 cause Weaver syndrome. Am J Hum Genet. 90:110-8.
202. Simpson MA, Deshpande C, Dafou D, Vissers LE, Woollard WJ, Holder SE, Gillessen-Kaesbach G, Derks R, White SM, Cohen-Snuijf R, Kant SG, Hoefsloot LH, Reardon W, Brunner HG, Bongers EM, Trembath RC. (2012) De novo mutations of the gene encoding the histone acetyltransferase KAT6B cause Genitopatellar syndrome. Am J Hum Genet. 90:290-294.
203. Bochukova E, Schoenmakers N, Agostini M, Schoenmakers E, Rajanayagam O, Keogh JM, Henning E, Reinemund J, Gevers E, Sarri M, Downes K, Offiah A, Albanese A, Halsall D, Schwabe JW, Bain M, Lindley K, Muntoni F, Vargha-Khadem F, Dattani M, Farooqi IS, Gurnell M, Chatterjee K. (2012) A mutation in the thyroid hormone receptor alpha gene. N Engl J Med. 366:243-9.
204. Hood RL, Lines MA, Nikkel SM, Schwartzentruber J, Beaulieu C, Nowaczyk MJ, Allanson J, Kim CA, Wieczorek D, Moilanen JS, Lacombe D, Gillessen-Kaesbach G, Whiteford ML, Quaio CR, Gomy I, Bertola DR, Albrecht B, Platzer K, McGillivray G, Zou R, McLeod DR, Chudley AE, Chodirker BN, Marcadier J; FORGE Canada Consortium, Majewski J, Bulman DE, White SM, Boycott KM. (2012) Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. Am J Hum Genet. 90:308-13.

205. Montenegro G, Rebelo AP, Connell J, Allison R, Babalini C, D'Aloia M, Montieri P, Schüle R, Ishiura H, Price J, Strickland A, Gonzalez MA, Baumbach-Reardon L, Deconinck T, Huang J, Bernardi G, Vance JM, Rogers MT, Tsuji S, De Jonghe P, Pericak-Vance MA, Schöls L, Orlacchio A, Reid E, Züchner S. (2012) Mutations in the ER-shaping protein reticulon 2 cause the axon-degenerative disorder hereditary spastic paraplegia type 12. J Clin Invest. 122:538-44.
206. Ostergaard P, Simpson MA, Mendola A, Vasudevan P, Connell FC, van Impel A, Moore AT, Loeys BL, Ghalamkarpour A, Onoufriadis A, Martinez-Corral I, Devery S, Leroy JG, van Laer L, Singer A, Bialer MG, McEntagart M, Quarrell O, Brice G, Trembath RC, Schulte-Merker S, Makinen T, Vikkula M, Mortimer PS, Mansour S, Jeffery S. (2012) Mutations in KIF11 cause autosomal-dominant microcephaly variably associated with congenital lymphedema and chorioretinopathy. Am J Hum Genet. 90:356-362.
207. Jones MA, Ng BG, Bhide S, Chin E, Rhodenizer D, He P, Losfeld ME, He M, Raymond K, Berry G, Freeze HH, Hegde MR. (2012) DDOST mutations identified by whole-exome sequencing are implicated in congenital disorders of glycosylation. Am J Hum Genet. 90:363-8.
208. Depienne C, Bouteiller D, Méneret A, Billot S, Groppa S, Klebe S, Charbonnier-Beaupel F, Corvol JC, Saraiva JP, Brueggemann N, Bhatia K, Cincotta M, Brochard V, Flamand-Roze C, Carpentier W, Meunier S, Marie Y, Gaussen M, Stevanin G, Wehrle R, Vidailhet M, Klein C, Dusart I, Brice A, Roze E. (2012) RAD51 haploinsufficiency causes congenital mirror movements in humans. Am J Hum Genet. 90:301-307.
209. Lines MA, Huang L, Schwartzentruber J, Douglas SL, Lynch DC, Beaulieu C, Guion-Almeida ML, Zechi-Ceide RM, Gener B, Gillessen-Kaesbach G, Nava C, Baujat G, Horn D, Kini U, Caliebe A, Alanay Y, Utine GE, Lev D, Kohlhase J, Grix AW, Lohmann DR, Hehr U, Böhm D; FORGE Canada Consortium, Majewski J, Bulman DE, Wieczorek D, Boycott KM. (2012) Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. Am J Hum Genet. 90:369-77.
210. Harms MB, Sommerville RB, Allred P, Bell S, Ma D, Cooper P, Lopate G, Pestronk A, Weihl CC, Baloh RH. (2012) Exome sequencing reveals DNAJB6 mutations in dominantly-inherited myopathy. Ann Neurol. 71:407-16.
211. Audo I, Bujakowska K, Orhan E, Poloschek CM, Defoort-Dhellemmes S, Drumare I, Kohl S, Luu TD, Lecompte O, Zrenner E, Lancelot ME, Antonio A, Germain A, Michiels C, Audier C, Letexier M, Saraiva JP, Leroy BP, Munier FL, Mohand-Saïd S, Lorenz B, Friedburg C, Preising M, Kellner U, Renner AB, Moskova-Doumanova V, Berger W, Wissinger B, Hamel CP, Schorderet DF, De Baere E, Sharon D, Banin E, Jacobson SG, Bonneau D, Zanlonghi X, Le Meur G, Casteels I, Koenekoop R, Long VW, Meire F, Prescott K, de Ravel T, Simmons I,

Nguyen H, Dollfus H, Poch O, Léveillard T, Nguyen-Ba-Charvet K, Sahel JA, Bhattacharya SS, Zeitz C. (2012) Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness. Am J Hum Genet. 90:321-30.

212. Hussain MS, Baig SM, Neumann S, Nürnberg G, Farooq M, Ahmad I, Alef T, Hennies HC, Technau M, Altmüller J, Frommolt P, Thiele H, Noegel AA, Nürnberg P. (2012) truncating mutation of CEP135 causes primary microcephaly and disturbed centrosomal function. Am J Hum Genet. 90:871-8.
213. Kirwan M, Walne AJ, Plagnol V, Velangi M, Ho A, Hossain U, Vulliamy T, Dokal I. (2012) Exome sequencing identifies autosomal-dominant SRP72 mutations associated with familial aplasia and myelodysplasia. Am J Hum Genet. 90:888-892.
214. Lee H, Graham JM Jr, Rimoin DL, Lachman RS, Krejci P, Tompson SW, Nelson SF, Krakow D, Cohn DH. (2012) Exome sequencing identifies PDE4D mutations in acrodysostosis. Am J Hum Genet. 90:746-751.
215. Lin Z, Chen Q, Lee M, Cao X, Zhang J, Ma D, Chen L, Hu X, Wang H, Wang X, Zhang P, Liu X, Guan L, Tang Y, Yang H, Tu P, Bu D, Zhu X, Wang K, Li R, Yang Y. (2012) Exome sequencing reveals mutations in TRPV3 as a cause of Olmsted syndrome. Am J Hum Genet. 90:558-564.
216. Fiskerstrand T, Arshad N, Haukanes BI, Tronstad RR, Pham KD, Johansson S, Håvik B, Tønder SL, Levy SE, Brackman D, Boman H, Biswas KH, Apold J, Hovdenak N, Visweswariah SS, Knappskog PM. (2012) Familial diarrhea syndrome caused by an activating GUCY2C mutation. N Engl J Med. 366:1586-1595.
217. Bernier FP, Caluseriu O, Ng S, Schwartzentruber J, Buckingham KJ, Innes AM, Jabs EW, Innis JW, Schuette JL, Gorski JL, Byers PH, Andelfinger G, Siu V, Lauzon J, Fernandez BA, McMillin M, Scott RH, Racher H; FORGE Canada Consortium, Majewski J, Nickerson DA, Shendure J, Bamshad MJ, Parboosingh JS. (2012) Haploinsufficiency of SF3B4, a component of the pre-mRNA spliceosomal complex, causes Nager syndrome. Am J Hum Genet. 90:925-33.
218. Spiegel R, Pines O, Ta-Shma A, Burak E, Shaag A, Halvardson J, Edvardson S, Mahajna M, Zenvirt S, Saada A, Shalev S, Feuk L, Elpeleg O. (2012) Infantile cerebellar-retinal degeneration associated with a mutation in mitochondrial aconitase, ACO2. Am J Hum Genet. 90:518-23.
219. 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. (2012) Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nat Genet. 44:379-380.

220. Srour M, Schwartzentruber J, Hamdan FF, Ospina LH, Patry L, Labuda D, Massicotte C, Dobrzeniecka S, Capo-Chichi JM, Papillon-Cavanagh S, Samuels ME, Boycott KM, Shevell MI, Laframboise R, Désilets V; FORGE Canada Consortium, Maranda B, Rouleau GA, Majewski J, Michaud JL. (2012) Mutations in C5ORF42 cause Joubert syndrome in the French Canadian population. Am J Hum Genet. 90:693-700.
221. Polvi A, Linnankivi T, Kivelä T, Herva R, Keating JP, Mäkitie O, Pareyson D, Vainionpää L, Lahtinen J, Hovatta I, Pihko H, Lehesjoki AE. (2012) Mutations in CTC1, encoding the CTS telomere maintenance complex component 1, cause cerebroretinal microangiopathy with calcifications and cysts. Am J Hum Genet. 90:540-549.
222. Schossig A, Wolf NI, Fischer C, Fischer M, Stocker G, Pabinger S, Dander A, Steiner B, Tönz O, Kotzot D, Haberlandt E, Amberger A, Burwinkel B, Wimmer K, Fauth C, Grond-Ginsbach C, Koch MJ, Deichmann A, von Kalle C, Bartram CR, Kohlschütter A, Trajanoski Z, Zschocke J. (2012) Mutations in ROGDI Cause Kohlschütter-Tönz Syndrome. Am J Hum Genet. 90:701-707.
223. Nakazawa Y, Sasaki K, Mitsutake N, Matsuse M, Shimada M, Nardo T, Takahashi Y, Ohyama K, Ito K, Mishima H, Nomura M, Kinoshita A, Ono S, Takenaka K, Masuyama R, Kudo T, Slor H, Utani A, Tateishi S, Yamashita S, Stefanini M, Lehmann AR, Yoshiura K, Ogi T. (2012) Mutations in UVSSA cause UV-sensitive syndrome and impair RNA polymerase IIo processing in transcription-coupled nucleotide-excision repair. Nat Genet. 44:586-92.
224. Austin ED, Ma L, LeDuc C, Berman Rosenzweig E, Borczuk A, Phillips JA 3rd, Palomero T, Sumazin P, Kim HR, Talati MH, West J, Loyd JE, Chung WK. (2012) Whole exome sequencing to identify a novel gene (caveolin-1) associated with human pulmonary arterial hypertension. Circ Cardiovasc Genet. 5:336-43.
225. Le Gallo M, O'Hara AJ, Rudd ML, Urick ME, Hansen NF, O'Neil NJ, Price JC, Zhang S, England BM, Godwin AK, Sgroi DC; NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program, Hieter P, Mullikin JC, Merino MJ, Bell DW. (2012) Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nat Genet. 10.1038/ng.2455.
226. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL, Basehore MJ, Broeckel U, Tomita-Mitchell A, Arca MJ, Casper JT, Margolis DA, Bick DP, Hessner MJ, Routes JM, Verbsky JW, Jacob HJ, Dimmock DP. (2012) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 13:255-62.
227. Koboldt DC, Ding L, Mardis ER, Wilson RK. (2010) Challenges of sequencing human genomes.

Brief Bioinform. 11(5), 484–498.

228. Welch JS, Westervelt P, Ding L, Larson DE, Klco JM, Kulkarni S, Wallis J, Chen K, Payton JE, Fulton RS, Veizer J, Schmidt H, Vickery TL, Heath S, Watson MA, Tomasson MH, Link DC, Graubert TA, DiPersio JF, Mardis ER, Ley TJ, Wilson RK. (2011) Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA. 305:1577-84.

Chapter **2**

Rapid and cost effective detection of small mutations in the DMD gene by high resolution melting curve analysis

Rowida Almomani, Nienke van der Stoep, Egbert Bakker, Johan T. den Dunnen, Martijn H. Breuning, Ieke B. Ginjaar

Neuromuscul Disord. 2009; 19:383-90.

Abstract

Duchenne/Becker muscular dystrophy (DMD/BMD) is caused by large deletions or duplications in two-thirds of the cases. The remaining one-third DMD patients have small mutations in the DMD gene. Screening for such small mutations is a daunting and costly task. High resolution melting curve analysis (HR-MCA) followed by sequencing for amplicons with altered melting profiles can be used to scan DNA for small alterations. We first validated the technique as screening procedure for the DMD gene and then screened a group of unrelated 22 DMD/BMD patients and 11 females. We managed to identify all previously found mutations by means of HR-MCA, which provided its validation. Furthermore, 17 different pathogenic mutations were found in the screening group, of which 10 were novel. Our results provide validation of HR-MCA as a powerful and inexpensive pre-sequencing scanning method. This technology is now ready for routine diagnostic use on DMD/BMD patients and female carriers.

Introduction

Duchenne muscular dystrophy (DMD) is a fatal neuromuscular disorder, characterized by rapidly progressive muscle weakness and wasting. DMD is one of the most common types of muscular dystrophy, with an incidence of one in 3500 newborn boys [1]. The onset of symptoms is generally before the age of 5. Affected individuals are confined to a wheelchair before the age of 12 and usually die in the course of the second or third decade, due to respiratory or heart failure [2].

Becker muscular dystrophy (BMD) shows a milder phenotype and is less common, with an incidence of 1:20,000 newborn males. BMD is characterized by delayed onset of muscle weakness and clinical symptoms. Many BMD patients remain ambulant later in life and have a longer life span than DMD patients [2].

DMD and BMD are allelic X-linked recessive diseases, caused by mutations in one of the largest human genes known to date, the DMD gene, which is distributed over about 2.4 million base pairs [3]. The vast majority of affected individuals are boys. However, a few affected females have been reported, in whom the disease was associated with a translocation with a breakpoint within Xp21 locus [4] or due to skewed X-inactivation, in which the majority of muscle cells used the mutated DMD gene, while the normal gene is inactivated through non-random X-inactivation [5].

The DMD gene has 79 exons [6], coding for a 14 kb mRNA transcript. The 427 kDa cytoskeletal dystrophin protein is localized to the cytoplasmic face of the sarcolemma [7]. Dystrophin protein is an important component of the dystrophin–glycoprotein complex that stabilizes the membrane of striated muscle. The absence of dystrophin leads to sarcolemmal fragility, muscle weakness, and eventually muscle degeneration [8].

The extremely large size of the dystrophin gene makes it vulnerable to structural changes. Many pathogenic mutations have been reported among DMD patients; 60% of these mutations are intragenic deletions ranging from one to several exons, and 5–10% are duplications [3]. The remaining one-third of sequence changes are mutations at the nucleotide level [3, 9].

There is a hypothesis known as the reading frame rule. It predicts that deletions or duplications,

which shift the reading frame of dystrophin messenger RNA, produce premature, truncated, nonfunctional protein and cause the severe DMD phenotype. On the other hand, BMD is caused by inframe deletions/duplications, which allow the generation of partially functional, internally deleted or duplicated protein. The reading frame hypothesis holds true for over 92% of all DMD and BMD patients [10].

The great majority of deletions and duplications cluster in a minor and a major hot spot within the DMD gene. The first one spans exons 2–20, while the second, major one spans exons 45–53 [3] and [11]. These mutations can be detected by a variety of methods including Southern blotting [11], multiplex PCR [12] and [13], multiplex amplifiable probe hybridization (MAPH) [14], and recently multiplex ligation-dependent probe amplification (MLPA). The last allows fast and reliable detection of deletions and duplications throughout the DMD gene [15].

A number of scanning methodologies have been developed to enhance small pathogenic mutation detection in patients without detectable large deletions and duplications. These methods include denaturing gradient gel electrophoresis (DGGE) [16], denaturing high performance liquid chromatography (dHPLC) [17], single strand conformation polymorphism analysis (SSCP) [18], fluorescent multiplex conformation sensitive capillary electrophoresis (FM-CSCE) [19], direct sequencing [20], and the protein truncation test (PTT) [21], each with its particular advantages and disadvantages.

The first aim of our study was to evaluate the HR-MCA as a mutation scanning method in the DMD gene and to minimize the cost of mutation scanning. The second was to implement an effective and convenient diagnostic strategy in BMD/DMD patients and carriers to detect small mutations.

Materials and methods

Patients

HR-MCA was performed on a group of 22 patients (12 DMD and 10 BMD) and a group of 11 females: five obligate carriers, five possible carriers (mothers and sisters of isolated DMD patients) and one young symptomatic female in whom cytogenetic analysis had excluded a translocation with a breakpoint in Xp21. All 12 patients suspected of suffering from DMD

exhibited severe phenotypes and elevated serum CK levels. DMD phenotype was confirmed by absence of dystrophin using immunohistochemical analysis in eight cases, in one case only reduced expression of dystrophin was observed and in three no muscle biopsy had been taken. The diagnosis of BMD was based on clinical criteria and elevated serum CK levels. No muscle biopsy had been taken from the majority of BMD-like patients (7 out of 10). Reduced dystrophin expression on Western blot was detected in one case and a weak patchy dystrophin pattern on muscle sections was observed in the other two. Genomic DNA was isolated from peripheral blood by standard procedures [22]. Large deletions and duplications in the DMD gene had been previously excluded by MLPA in all cases.

Validation

In order to determine the efficiency of HR-MCA for mutation scanning, we tested 40 heterozygous and 34 hemizygous variants in 45 different amplicons. These samples were selected from previous studies. In order to enhance heteroduplex formation for hemizygous variants, each hemizygous variant was tested in three ways: without mixing with wild type DNA, mixing with other male genomic wild type DNA before PCR amplification and post-PCR mixing with other male wild type PCR product.

Primers

Sequencing primers with M13 tails were designed previously by using primer 3 software (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3.cgi). All 79 exons and adjacent intron/exon junctions were amplified and optimized for high resolution melting curve analysis. To maximize the sensitivity of the technique, exons 3, 23, 48, 53, 61, 67, 68, 76, and 79, which had three melting domains, were split into multiple amplicons. In addition, new primers were designed for exon 19, with smaller fragment size to avoid having three melting domains and exon 65, which failed to give a PCR product.

All new primers were designed using either primer 3 or light scanner primer design software (Idaho Technology). To predict the number of melting domains, these primers were tested using the melting program (version 1.0; INGENY International BV, Goes, The Netherlands). In 10 amplicons, 19, 3-GC, 61A, 61B, 68A, 68B, 79B2, 79C2, 79D1 and 79D2, a short GC stretch was

added to avoid three melting domains. The total number of amplicons was 96 (Supplementary Table 1).

Probes

We designed unlabeled probes (incorporating a 3′ phosphate in order to prevent polymerase extension) that perfectly match the five most frequent variants in five different exons of the DMD gene. The probe sequences, annealing temperatures and primer ratios for genotyping are shown in Table 1.

Table 1: Sequences and PCR conditions for five different probes.

Amplicon	Probe sequences	Annealing temperature (C°)	Primer ratios	No. of cycles	Variant
17	CTGAAGTCTTTCGAGCAATGTCTGACC	61	1 to 5	55	c.1993 -37T>G
37	AAACTTGATGGCAAACCACGGTGAC	61	1 to 10	55	c.5234G>A
48b	AGAAGGACCATTTGACGTTAAGGTAGG	61	1 to 5	55	c.7096C>A
54	GCATTCATAAAAGGTATGAATTATATTAT	61	1 to 5	55	c.8027+11C>T
66	CAGATGTAAGTCGTGTATACTAATGCTG	61	1 to 5	55	c.9649 +15T>C

PCR

PCR was performed in 96-well, non-transparent plates (ABgene) in 10 µl total volume with: 1 × PCR buffer (Roche), 2 mM $MgCl_2$, 2 mM dNTPs, 3 pmol of each primer, 1 × LCGreen Plus (Idaho Technology), 0.5 U of fast start Taq DNA polymerase (Roche) and 20 ng of DNA template. All PCR wells were covered with 15 µl of mineral oil (Sigma), and centrifuged at 2500 RPM for 1 min before PCR.

PCR was carried out in a gradient cycler (Bio-RAD). The thermo-cycling protocol was as follows: 10 min at 95 °C, 40 cycles of 20 s at 95 °C, 30 s at the annealing temperature, 40 s at 72 °C, and 5 min at 72 °C. In order to promote heteroduplex formation, samples were

denaturated by heating to 95 °C for 1 min and cooling down to 15 °C in the thermo-cycler before HR-MCA.

Asymmetric PCR

Asymmetric PCR was performed whenever unlabeled probes were used. Primer asymmetry ratios of 1:5 to 1:10 produced sufficient single stranded product for probe annealing (see Table 1). PCR reactions were performed as described above with some minor modifications as follows: 1 pmol of the forward primer, 5 or 10 pmol of the reverse primer (see Table 1), and 5 pmol of each unlabeled probe. The thermo-cycling protocol was done as described above but with 55 cycles and an annealing temperature of 61 °C.

Post PCR mixing

Samples from hemizygous males were mixed post-PCR. After successful PCR amplification, which was tested by a light scanner (Idaho Technology), post-PCR mixing was performed between amplicons of two different non-related male patients. As males have only one X-chromosome, mixing is necessary to ensure that heteroduplex formation can occur. Post-PCR mixing was done as follows:

Ten microliters of PCR product from each patient was covered with 15 μl of mineral oil (Sigma), centrifuged at 2500 RPM for 1 min, heated to 95 °C for 5 min and cooled down to 15 °C before HR-MCA.

Melting analysis

After PCR, the plates were imaged in a 96-well Light Scanner (Idaho Technology). The fluorescence data were collected from 65 to 98 °C for the amplicon scanning, and from 55 to 98 °C for the unlabeled probe genotyping at a temperature transition rate of 0.1 °C/s. Melting curves were analyzed by using the commercial light scanner software on the high sensitivity setting as previously described [23], [24] and [25]. After exponential background subtraction, fluorescence data were normalized between 0% and 100%. Slight temperature errors or buffer differences between wells or runs were corrected by temperature shifting in regions of low fluorescence and high temperature (2–5% normalized fluorescence). This facilitated clustering of

curves for heterozygous samples. Difference plots of normalized and temperature overlaid curves were obtained by subtracting the fluorescence values of each curve from the mean reference values, which were defined as the most popular genotype (wild type).

Sequencing

Since HR-MCA is a non-destructive method, all amplicons that produced abnormal melting profiles were sequenced from the original PCR reactions for female carrier samples. All altered male patient amplicons were confirmed by sequencing of independent PCR products. PCR was performed in MicroAmp reaction tubes (Applied Biosystems) in 25 µl total volumes containing: 10 × commercial PCR buffer or 5 × STR buffer which contains (0.5 M $(NH4)_2SO4$, 0.5 M Tris–HCL, pH 8.8, 1 M $MgCl_2$, 10 mM EDTA, 14 M β-mercapto-ethanol, and ultra-pure water), 1.5–3 mM $MgCl_2$ (Supplementary Table 2), 2.5 mM dNTPs, 2.5 pmol each primer, 1 U of Taq DNA polymerase (Promega) and 200 ng of DNA template. PCR was carried out in a Biometra T-Professional (Westburg). The thermo-cycling protocol was as follows: 5 min at 95 °C, 35 cycles of 20 s at 95 °C, 30 s at 55 °C, 30 s at 72 °C, and 5 min at 72 °C. After amplification, the PCR products were purified by the AMPure PCR purifications system using solid-phase paramagnetic bead technology (Agencourt). Sequencing was performed in both sense and antisense direction using uniform BigDye (Terminator v3.1 sequencing reactions, Applied Biosystems) with PAGE purified M13F (−21M13) or M13R (M13REV) sequencing primer. Sequencing reactions were then purified using a column filtration procedure (DTR V3 96-wells plates, Edge Biosystems) and final analysis was done using the ABI 3730 [26]. After electrophoresis, data processing was automated using SeqScape 2.1.1 software (ABI). Base calls with quality values below QV = 25 were checked manually. The primer sequences (with M13 tail) that were used for amplification of DMD amplicons are shown in (Supplementary Table 3).

Results

Data from the HR-MCA

-Validation

DNA samples from patients or female carriers with known sequence changes, were used to optimize parameters for HR-MCA mutation scanning. Scanning for variants relies on differences

in the melting curve profile, which are most apparent in difference plots. The most common genotype, wild type, was selected as a reference to form the baseline, while the variant samples showed clearly distinctive melting curves.

We could not detect the different known sequence variants in exons 3, 23, 48, 53 and 79B. All of these fragments had three melting domains that could mask the presence of the variants. Therefore, new primer sets were designed for all fragments with three melting domains, to reduce the number of melting domains per fragment. Our data show that this approach enhanced the resolution of HR-MCA.

The initial testing correctly identified all 40 heterozygous and 24/34 hemizygous variants. The remaining 10 hemizygous variants were detected only after post-PCR mixing with wild type male DNA, an example is shown in Fig. 1A and B. Panel A shows exon 16 with one aberrant melting profile for a heterozygous variant (c.1961T > C) from a female sample and no aberrant melt profile for the male sample. Whereas panel B shows the result for the same exon after post-PCR mixing for the males samples, with two aberrant melt profiles, which represent the heterozygous (c.1961T > C) and the hemizygous (c.1869C > T) variants.

In several exons there was clustering of different sequence variations, which were readily distinguishable from each other and from the wild type, showing different melting curves.

Variants could also be distinguished in homozygous and heterozygous form. In exon 53, the abnormal curve produced by the same sequence variant (c.7728T > C) in homozygous form differs from that caused by heteroduplex formation in the heterozygous form (Fig. 1C).

In order to reduce the burden of sequencing, five unlabeled probes were designed to identify frequently found (see Table 2) variants (c.1993-37T > G, c.5234G > A, c.7096C > A, c.8027 + 11C > T, c.9649 + 15T > C) in amplicons 17, 37, 48b, 54, and 66, respectively. We successfully detected both heterozygous and homozygous–hemizygous variants. All three possible genotypes within the tested samples set were recognized by a single unlabeled probe. A perfectly matched probe-target hybrid had a higher Tm than the mismatched ones. Heterozygous amplicons, on the other hand, showed two peaks with two different temperatures representing

both genotypes, this is exemplified by sequence variant (c.5234G > A) in exon 37 (Fig. 2A and B).

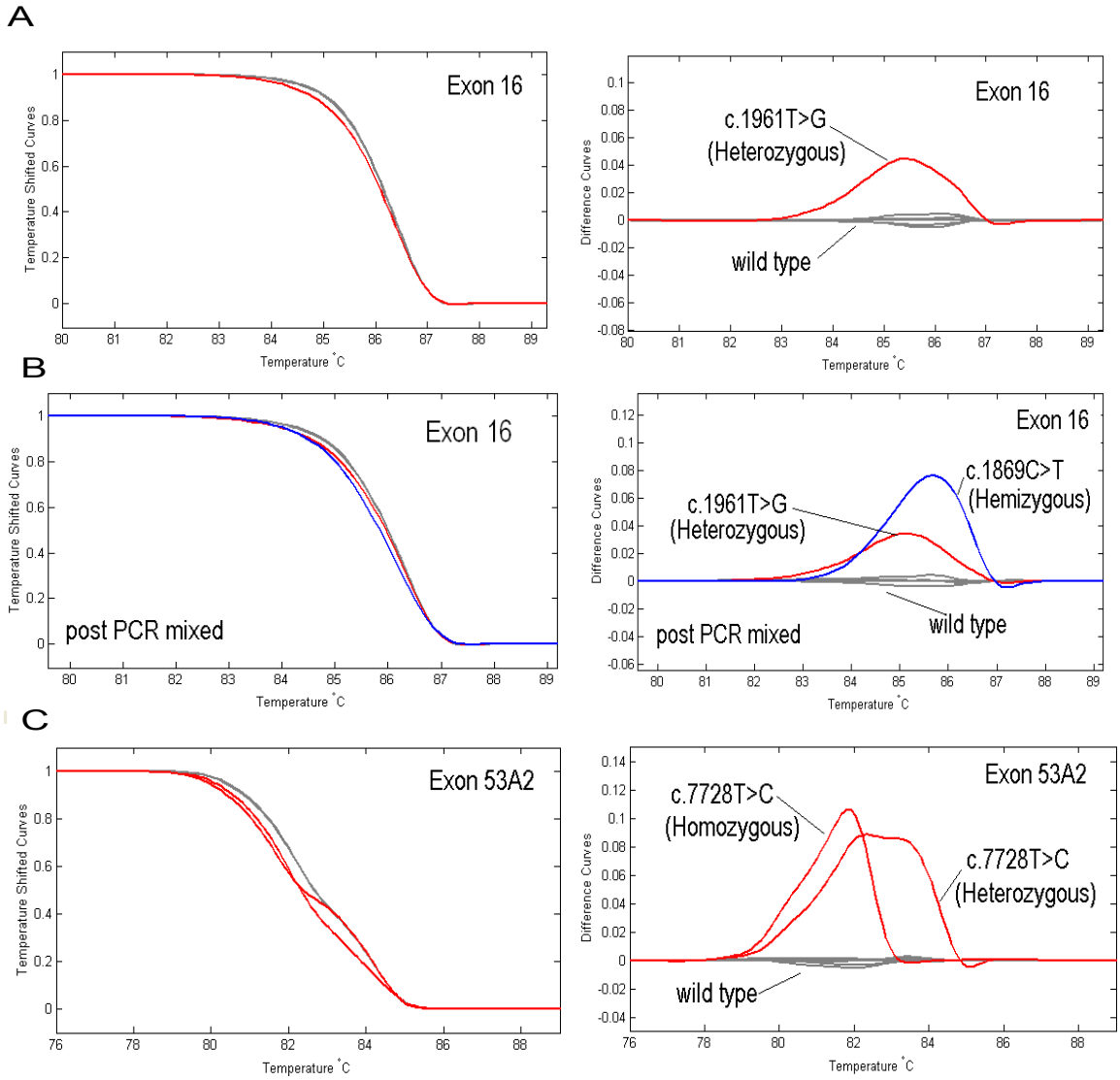


Fig. 1. Temperature shifted (left) and subtractive difference plots (right) of wild type and variants. (A and B) Exon 16, on (A) only the heterozygous variant from female sample is detected, while in (B) the hemizygous variant is only detected after post-PCR mixing. (C) The different melting curve profiles for the same variant in heterozygous and homozygous state.

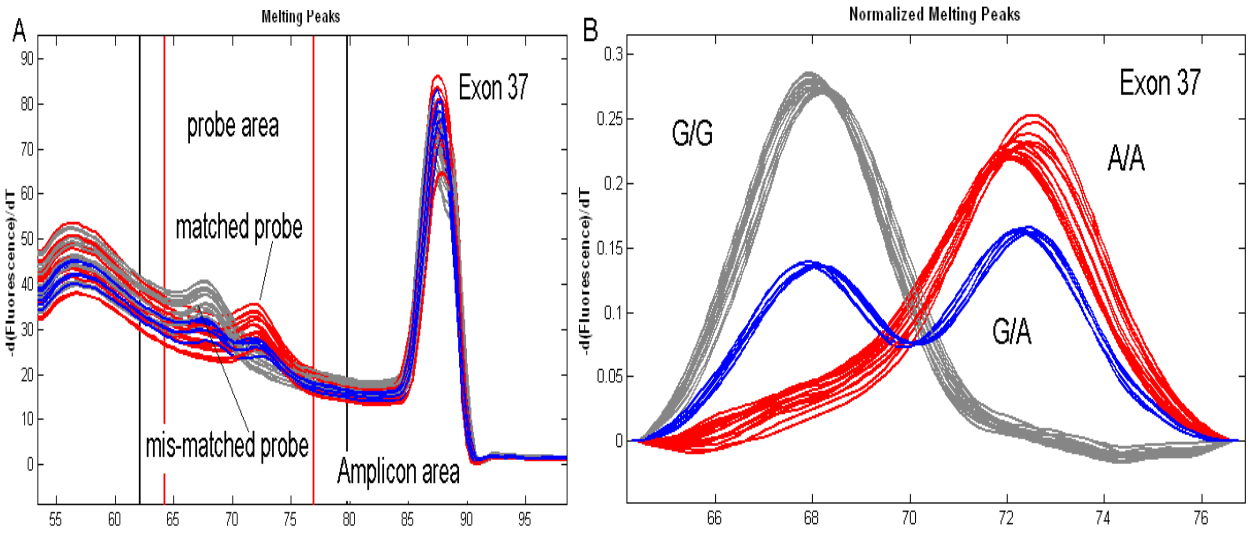


Fig. 2. (A and B) Common DMD gene variants in exon 37 detected by genotyping with unlabeled probes. (A) Both the amplicon and probe area of exon 37. (B) The enlargement of the probe area.

-Patient screening group

After validation of the scanning method, we tested the group of 22 (BMD, DMD) patients and 11 females. Amplicons with abnormal melting curves were sequenced to determine the changed variant. Five different heterozygous pathogenic mutations were detected directly within the female group. Furthermore, seven out of 12 hemizygous pathogenic mutations were found in male patients. Five out of 12 hemizygous pathogenic mutations (c.187-2A > G, c.3097_3098del, c.3603 + 2T > A, c.5771_5772del, c.6611dup) were detected only after post-PCR mixing, because they showed an altered fluorescence curve compared to the wild type profile.

The only deletion/insertion mutation (c.597_614delinsCTAGTTTC), in exon 7 in a DMD male patient, was detected directly without post-PCR mixing (Fig. 3A). However, the abnormal curve produced by the same hemizygous mutation became clearer after post-PCR mixing (Fig. 3B).

The results of genotyping our patients and carriers show that there is a great advantage of having oligonucleotide probes corresponding to the frequently occurring variants, because it reduces the number of sequencing reactions (Table 2).

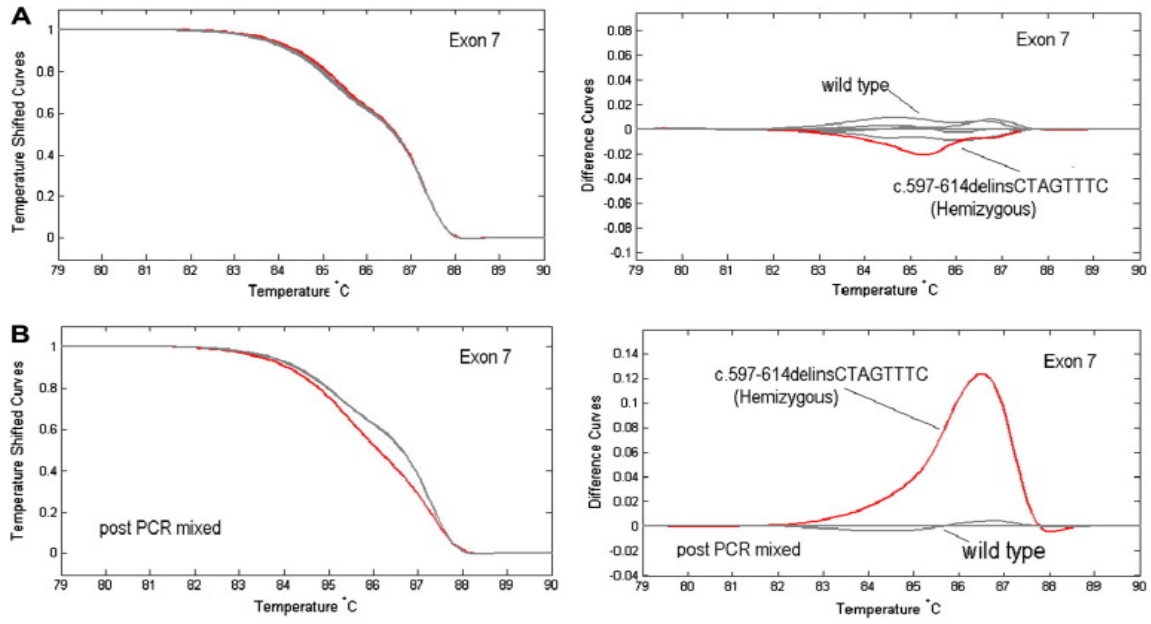


Fig. 3. Temperature shifted (left) and subtractive difference plots (right) of wild type and a mutation in exon 7. (A) Exon 7 from a male patient without post-PCR mixing; the deletion/insertion mutation can be detected. (B) The post-PCR mixing of the same exon, that produces a different and clearer melting curve.

Data from sequencing analysis

In total, 17 different pathogenic mutations were detected in 33 cases (12 DMD, 10 BMD and 11 female carriers) of which 10 were novel (Table 3). Most mutations were identified in obligate carriers (4/5) and DMD patients (10/12). Mutations were identified in all eight DMD patients with absence of dystrophin in the muscle tissue. A mutation was also found in two patients from whom no muscle tissue was available. No mutation was found in the other two, one of whom had reduced dystrophin expression. Seven of the 10 mutations in the DMD patients were novel (see Table 3).

Table 2. variants detected by HR-MCA and sequencing in BMD/DMD patients and carriers.

Exon	DNA change	Protein	Frequency/Remarks
3	c.94-9dupT		4⁄33
3	c.186+35A>T		1⁄33
6	c.530+19C>T		once +patho
9	c.832-18_832-17delinsGA		once +patho
14	c.1635A>G		3⁄33
14	c.1704+51T>C		3⁄33
17	c.2168 +13T>C		1⁄33
17	**c.1993-37T>G**		**23⁄33**
21	c.2645G>A	p.Gly882Asp	12⁄33
23	c.3021G>A		1⁄33
25	c.3406A>T	p.Thr1136Ser	1⁄33
27	c.3734C>T	p.Thr1245Ile	once +patho
31	c.4234-13A>G		2⁄33
33	c.4519-34T>A		once +patho
34	c.4675 -53G>T		2⁄33
37	**c.5234G>A**	**p.Arg1745His**	**19⁄33**
43	c.6290+27T>A		2⁄33
44	c.6291-115G>A		5⁄33
45	c.6463C>T	p.Arg2155Trp3/33	3⁄33
48	c.6913-114A>T		Once
48b	**c.7096C>A**	**p.Gln2366Lys**	**27⁄33**
49	c.7200+53C>G		11⁄33
53	c.7728T>C		8⁄33
54	c.8027+13T>G		once +patho
54	**c.8027+11C>T**		**14⁄33**
59	c.8762A>G	p.His2921Arg	once +patho
59	c.8810A>G	p.Gln2937Arg	3⁄33
64	c.9361+138T>C		7⁄33
66	**c.9649+15T>C**		**28/33**
75	c.10789C>T		2⁄33
75	c.10797+42C>G		1⁄33
79	c.*477_*484del		2⁄33
79	c.*491_*492dupCA		6⁄33
79	c.*1051_*1052ins		1⁄33
79	c.*1447A>G		1⁄33

- + patho, when a variant is found in combination with a pathogenic mutation.
- All variants in bold were detected by HR-MCA/ probes.

Table 3. Pathogenic mutations found in BMD/DMD patients and carriers.

Test Sample #	Phenotype	Sex	Exon	DNA change	Protein	New	Dystrophin (IHC)
21	BMD	male	4	c.187-2A>G		No (27)	Reduced
17	DMD	male	7	c.597_614delinsCTAGTTTC	p.Phe200X	Yes	Absent
8	DMD	male	15	c.1721G>A	p.Trp574X	Yes	ND
24	DMD	male	22	c.2929C>T	p.Gln977X	Yes	ND
9	DMD	male	23	c.3097_3098del	p.Ser1033LeufsX5	Yes	Absent
16	DMD	male	23	c.3151C>T	p.Arg1051X	No (27)	Absent
23	obligate carrier	female	26	c.3516G>A	p.Trp1172X	Yes	ND
26	DMD	male	26	c.3603+2T>A	p.Lys1201_Arg1202insX25	No (27)	Absent
5	BMD	male	34	c.4845+1G>A		Yes	ND
12	obligate carrier	female	40	c.5697del	p.Lys1899AsnfsX2	No (27)	ND
31	DMD	male	41	c.5771_5772del	p.Glu1924GlyfsX7	No (27)	Absent
10	DMD	male	44	c.6291-1G>T		Yes	Absent
30	DMD	male	45	c.6611dup	p.Arg2205GlufsX18	Yes	Absent
11	obligate carrier	female	51	c.7538dup	p.Lys2514GlufsX34	Yes	ND
18	symptomatic carrier	female	58	c.8641del	p.Leu2881X	Yes	Mosaic
1	DMD	male	67	c.9807+1G>C		No (27)	Absent
34	obligate carrier	female	70	c.10141C>T	p.Arg3381X	No (27)	ND

ND, not done.
IHC, immunohistochemistry.

In three cases splice-site mutations were found: two of these (c.3603 + 2T > A; c.9807 + 1G > C) have been described before in DMD patients [27]. One novel mutation (c.6291-1G > T) in the splice-site of exon 44 was predicted to skip exon 44 thereby shifting the reading frame of the DMD gene. Two mutations were found among 10 patients suspected of having BMD based on clinical symptoms. In a 33-year-old BMD patient from a large BMD family with six patients, a novel splice-site mutation of exon 34 was found (c.4845 + 1G > A) predicting an "in-frame" skip of exon 34 in this family. The second mutation was found in a 10-year-old sporadic BMD patient with reduced dystrophin levels on a Western blot. A splice-site mutation of exon 4 (c.187-2A > G) was identified, which is likely to skip (in-frame) exon 4. The same mutation has been reported before in a BMD patient [27]. No mutation was detected in the remaining eight BMD cases. So, it is possible that these patients are suffering from other types of muscular dystrophy such as LGMD. Most of the BMD-like patients are sporadic except for one family in which recent haplotyping showed that X-linked inheritance is unlikely. A novel heterozygous frameshift mutation was detected in a young symptomatic female (c.8641delC; p.Leu2881ArgfsX13), who appeared to be a DMD carrier, and in whom previous cytogenetic analysis had excluded a translocation in band Xp21. A mutation was identified in four out of five obligate carriers of this study: two frameshift mutations, one of which is novel (c.7538dupA; p.Lys2514GlufsX34), and two nonsense mutations, one of which is novel (c.3516G > A; p.Trp1172X). However, no mutation was identified in any of the five possible DMD carriers. In addition to these mutations we identified 30 different variants. All of these have been reported before [27] and are shown in Table 2.

Discussion

HR-MCA in combination with dsDNA dye LCGreen Plus was used to scan the DMD gene and to genotype frequent variants. LCGreen Plus dye does not inhibit Taq polymerase and can be used at a concentration that will saturate newly synthesized double stranded DNA during PCR. Saturating all the available double stranded sites is a critical characteristic that eliminates the potential for a dye molecule to redistribute during the melting process of the PCR product. Another advantage is that because the dye is added to the PCR before amplification, no further processing or labeling of primers is required.

For HR-MCA one needs to pay careful attention to the design of the primers as mutation detection is easier when there are only one or two melting domains. We found that breaking up exons with three melting domains into multiple fragments allowed the detection of all variants that were tested. This is exemplified by exon 23 in which the two variants (c.2994T > A) and (c.3059C > G) were not observed initially before breaking up the exons. We also manipulated the melting by adding a GC stretch (7–11 bp) to 10 primers in order to avoid three melting domains and to maximize the sensitivity of the technique.

After optimizing the various parameters, all 40 heterozygous and 24 out of 34 of the hemizygous variants that were located anywhere between two primers could be detected. Although the majority of the X-linked hemizygous variants were detected directly, 10 of the hemizygous variants were missed, indicating that post-PCR-mixing from two non-related patients is needed to ensure heteroduplex formation and mutation detection. To avoid the risk that a variant would be missed using this approach because two patients may carry the same variant in the DMD gene, post-PCR mixing between patient sample and a non-affected male control sample would remove this risk. Post-PCR mixing is preferred because pre-PCR mixing requires an accurate quantification of DNA [28], and non amplification of one of the fragments could lead to false negative results.

All amplicons with abnormal melting profiles were sequenced, as there is no distinction between polymorphisms and pathogenic mutations. To avoid part of the sequencing, five most frequent variants throughout the DMD gene were genotyped by HR-MCA. All three possible genotypes within the tested sample set were recognized by a single unlabeled probe. In addition unexpected sequence variants under the probe could be detected. Use of unlabeled probes conveniently eliminates the need for expensive fluorescent labeled probes [29].

There have been numerous methods employed to detect small mutations in the DMD gene, such as DGGE [16], dHPLC [17], SSCP [18], FM-CSCE [19], PTT [21] and sequencing [20]. All of these technologies require post-PCR processing and separation on a gel or another matrix, which makes these techniques laborious and time consuming, as compared to HR-MCA, which is fast and has minimal post-PCR processing requirements. We conclude from our data that HR-MCA is at least as sensitive as DGGE, dHPLC and FM-CSCE. However, a comparative study has

recently shown that HR-MCA has higher sensitivity and specificity than dHPLC [30]. Moreover, all fragments analyzed by dHPLC need to be run under different denaturing conditions to maximize the mutation detection.

HR-MCA has an advantage over the DGGE method. DGGE requires considerable effort to design and optimize, making it more labor-intensive than HR-MCA for routine diagnostic use.

HR-MCA is a mutation scanning technique that requires accurate PCR amplification with normal unlabeled primers, whereas the FM-CSCE method needs the fluorophore labeling of one primer for each pair of primers. However, the major advantage of the FM-CSCE method is that nearly all mutation types can be detected simultaneously [19], whereas HR-MCA is suitable for the detection of only small mutations. For a complete mutation scanning strategy, HR-MCA should be combined with other methods such as MLPA [15].

The HR-MCA method makes mutation detection cost effective as it significantly reduces the amount of sequencing that needs to be performed. Furthermore, HR-MCA is a non-destructive and high throughput method for mutation scanning and genotyping, that can analyze 96 or 384 samples per run, and is thus exquisitely suitable for the screening of large multi-exonic genes, like the DMD gene.

As compared to RNA based methods, such as PTT, our HR-MCA technique is less laborious and less time consuming. PTT requires isolation of RNA, preferably from muscle tissue, which is not always available from affected patients. Although isolation of dystrophin mRNA is also possible from lymphocytes, the yield is very low [16]. Furthermore, only truncating mutations can be detected by PTT, whereas HR-MCA is able to detect all sequence changes, missense mutations, silent mutations, single nucleotide polymorphism (SNP's) and variations of unknown significance.

After sequencing of amplicons with abnormal melting profiles, about 83% of small mutations could be identified in our population suspected of suffering from DMD. It is very likely that a higher percentage of mutations would have been found if DMD had been confirmed in all cases by dystrophin analysis of muscle tissue. In two of the DMD-like cases in which no mutations were found, it seems plausible that other types of muscular dystrophies were involved. The fact

that no mutations were found in the majority of BMD-like patients (8/10), suggests that there is a large clinical overlap between BMD and other types of muscular dystrophy such as LGMD. It is, therefore, recommended that immunobiochemical analysis of muscle tissue be performed in patients suspected of having BMD before screening for small mutations. HR-MCA has been shown to be a quick and sensitive technique for further screening for small mutations in cases where dystrophin is absent or reduced in muscle tissue. An explanation for cases where no small mutation is found may be that the mutation is located either deep in an intron or in a regulatory region. Pathogenic mutations were found in four of the five obligate carrier females (80%). Determination of the carrier status is important for prenatal diagnosis, genetic counseling and prevention of the disease. It is possible that in the only family without a mutation, DMD is caused by a mutation deep in one of the introns activating cryptic exons or by a mutation in the promoter area of the DMD gene.

The majority of mutations that were identified were novel (60%), and were scattered throughout the gene. There were six different nonsense mutations which resulted in a truncated, nonfunctional protein, six different frame shift mutations and five changes that are expected to affect the splicing. All these 17 pathogenic mutations that were detected are unique to each family.

In conclusion, HR-MCA was found to be a highly reliable and quick method for mutation scanning and genotyping, requiring only direct analysis of the PCR reaction with a simple instrument. This technique offers many advantages over other techniques, and is a welcome addition to the screening strategy of laboratories involved in the diagnostic service for Duchenne and Becker muscular dystrophy.

Acknowledgments

We would like to thank M.J.R. van der Wielen, C.D.M. van Paridon, D. van Heusden, and A.L.J. Kneppers for their expert technical assistance, and Dr K. Madan for critical reading of the manuscript.

References

1. van Essen AJ, Busch HF, te Meerman GJ, ten Kate LP. Birth and population prevalence of Duchenne muscular dystrophy in The Netherlands. Hum Genet. 1992; 88: 258-266.
2. Emery AE. The muscular dystrophies. Lancet 2002; 359: 687-695.
3. Den Dunnen JT, Grootscholten PM, Bakker E et al. Topography of the Duchenne muscular dystrophy (DMD) gene: FIGE and cDNA analysis of 194 cases reveals 115 deletions and 13 duplications. Am J Hum Genet 1989; 45: 835-847.
4. Jacobs PA, Hunt PA, Mayer M, Bart RD. Duchenne muscular dystrophy (DMD) in a female with an X/autosome translocation: further evidence that the DMD locus is at Xp21. Am J Hum Genet 1981; 33: 513-518.
5. Yoshioka M, Yorifuji T, Mituyoshi I. Skewed X inactivation in manifesting carriers of Duchenne muscular dystrophy. Clin Genet 1998; 53: 102-107.
6. Koenig M, Monaco AP, Kunkel LM. The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein. Cell 1988; 53:219-228.
7. Zubrzycka-Gaarn EE, Bulman DE, Karpati G et al. The Duchenne muscular dystrophy gene product is localized in sarcolemma of human skeletal muscle. Nature 1988; 333: 466-469.
8. Ervasti JM, Sonnemann KJ. Biology of the striated muscle dystrophin-glycoprotein complex. Int Rev Cytol 2008; 265: 191-225.
9. Roberts RG, Bobrow M, Bentley DR. Point mutations in the dystrophin gene. Proc Natl Acad Sci U S A 1992; 89: 2331-2335.
10. Monaco AP, Bertelson CJ, Liechti-Gallati S, Moser H, Kunkel LM. An explanation for the phenotypic differences between patients bearing partial deletions of the DMD locus. Genomics1988; 2: 90-95.
11. Koenig M, Hoffman EP, Bertelson CJ et al. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell 1987; 50: 509-517.
12. Chamberlain JS, Gibbs RA, Ranier JE, Nguyen PN, Caskey CT. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification**.** Nucleic Acids Res 1988 ; 16:11141-11156.
13. Beggs AH, Koenig M, Boyce FM, Kunkel LM, Detection of 98% of DMD/BMD gene deletions by polymerase chain reaction. Hum Genet 1990; 86: 45-48.
14. White S, Kalf M, Liu Q, Villerius M et al. Comprehensive detection of genomic duplications and deletions in the DMD gene, by use of multiplex amplifiable probe hybridization. Am J Hum Genet 2002; 71: 365-374.
15. Lalic T, Vossen RH, Coffa J et al. Deletion and duplication screening in the DMD gene using MLPA. Eur J Hum Genet 2005; 13: 1231-1234.
16. Hofstra RM, Mulder IM, Vossen R et al. DGGE-based whole-gene mutation scanning of the dystrophin gene in Duchenne and Becker muscular dystrophy patients. Hum Mutat 2004; 23: 57-66.
17. Bennett RR, den Dunnen J, O'Brien KF, Darras BT, Kunkel LM. Detection of mutations in the dystrophin gene via automated DHPLC screening and direct sequencing. BMC Genet 2001; 2: 17.
18. Tuffery S, Moine P, Demaille J, Claustres M. Base substitutions in the human dystrophin gene: detection by using the single-strand conformation polymorphism (SSCP) technique. Hum Mutat 1993; 2: 368-374.
19. Ashton EJ, Yau SC, Deans ZC, Abbs SJ. Simultaneous mutation scanning for gross deletions, duplications and point mutations in the DMD gene. Eur J Hum Genet 2008;16:53-61.

20. Flanigan KM, von Niederhausern A, Dunn DM, Alder J, Mendell JR, Weiss RB. Rapid direct sequence analysis of the dystrophin gene. Am J Hum Genet 2003; 72:931-939.
21. Roest PA, Roberts RG, van der Tuijn AC, Heikoop JC, van Ommen GJ, den Dunnen JT. Protein truncation test (PTT) to rapidly screen the DMD gene for translation terminating mutations Neuromuscul Disord 1993; 3: 391-394.
22. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 1988; 16: 1215.
23. Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ. High-resolution genotyping by amplicon melting analysis using LCGreen. Clin Chem 2003; 49: 853-860.
24. Herrmann MG, Durtschi JD, Bromley LK, Wittwer CT, Voelkerding KV. Amplicon DNA melting analysis for mutation scanning and genotyping: cross-platform comparison of instruments and dyes. Clin Chem 2006; 52: 494-503.
25. Montgomery J, Wittwer CT, Palais R, Zhou L., Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis. Nat Protoc 2007; 2: 59-66.
26. Trip J, Drost G, Verbove DJ et al. In tandem analysis of CLCN1 and SCN4A greatly enhances mutation detection in families with non-dystrophic myotonia. Eur J Hum Genet 2008.
27. Aartsma-Rus A, Van Deutekom JC, Fokkema IF, Van Ommen GJ, Den Dunnen JT. Entries in the Leiden Duchenne muscular dystrophy mutation database: an overview of mutation types and paradoxical cases that confirm the reading-frame rule. Muscle Nerve 2006; 34: 135-144.
28. Palais RA, Liew MA, Wittwer CT. Quantitative heteroduplex analysis for single nucleotide polymorphism genotyping. Anal Biochem 2005; 346: 167-175.
29. Gundry CN, Vandersteen JG, Reed GH, Pryor RJ, Chen J, Wittwer CT. Amplicon melting analysis with labeled primers: a closed-tube method for differentiating homozygotes and heterozygotes. Clin Chem 2003; 49: 396-406.
30. Chou LS, Lyon E, Wittwer CT. A comparison of high-resolution melting analysis with denaturing high-performance liquid chromatography for mutation scanning: cystic fibrosis transmembrane conductance regulator gene as a model. Am J Clin Pathol 2005; 124: 330-338.

Supplementary data:

Supplementary Table 1 Sequences of primers, annealing temperatures and fragment sizes for HR-MCA.

Amplicon Number	Primer sequences	Fragment size	Annealing temp.
DMDEX01-F	TGTAAAACGACGGCCAGTGCAGGTCCTGGAATTTGA	405	61Cº
DMDEX01-R	CAGGAAACAGCTATGACCCAAACTAAACGTTATGCCACA		
DMDEX02-F	TGTAAAACGACGGCCAGTCACTAACACATCATAATGG	269	61Cº
DMDEX02-R	CAGGAAACAGCTATGACCGATACACAGGTACATAGTC		
DMDEX03 A-F	TGTAAAACGACGGCCAGTTCATCCGTCATCTTCGGCAGATTAA	176	61Cº
DMDEX03 A-R	CAGGAAACAGCTATGACCCAGGCGGTAGAGTAtgccaaatgaaaatca		
DMDEX03B-F	TCTTCAGTGACCTACAGGATGG	184	64Cº
DMDEX03B-R	CGCCCGCCGtgctgtttcaatcagtacctagtca		
DMDEX04-F	TGTAAAACGACGGCCAGTTTGTCGGTCTCTCTGCTGGTCAGTG	233	60Cº
DMDEX04-R	CAGGAAACAGCTATGACCCCAAAGCCCTCACTCAAAC		
DMDEX05-F	TGTAAAACGACGGCCAGTCAACTAGGCATTTGGTCTC	261	61Cº
DMDEX05-R	CAGGAAACAGCTATGACCTTGTTTCACACGTCAAGGG		
DMDEX06-F	TGTAAAACGACGGCCAGTTGGTTCTTGCTCAAGGAATG	335	61Cº
DMDEX06-R	CAGGAAACAGCTATGACCTGGGGAAAAATATGTCATCAG		
DMDEX07-F	TGTAAAACGACGGCCAGTCTATGGGCATTGGTTGTC	296	60Cº
DMDEX07-R	CAGGAAACAGCTATGACCAAAAGCAGTGGTAGTCCAG		
DMDEX08-F	TGTAAAACGACGGCCAGTTCGTCTTCCTTTAACTTTG	343	61Cº
DMDEX08-R	CAGGAAACAGCTATGACCTCTTGAATAGTAGCTGTCC		
DMDEX09-F	TGTAAAACGACGGCCAGTTCTATCCACTCCCCCAAACC	318	61Cº
DMDEX09-R	CAGGAAACAGCTATGACCAACAAACCAGCTCTTCAC		
DMDEX10-F	CGACGTTGTAAAACGACGGCCAGTGGAACAATCTGCAAAGAC	350	61Cº
DMDEX10-R	CAGGAAACAGCTATGACCAAAGGATGACTTGCCATTATAAC		
DMDEX11-F	TGTAAAACGACGGCCAGTCAAATAAAACTCAAAACCACACC	337	61Cº
DMDEX11-R	CAGGAAACAGCTATGACCCTTCCAAAACTTGTTAGTCTTC		
DMDEX12-F	TGTAAAACGACGGCCAGTCTTTCAAAGAGGTCATAATAGG	305	61Cº
DMDEX12-R	CAGGAAACAGCTATGACCCATCTGTGTTACTGTGTATAGG		
DMDEX13-F	TGTAAAACGACGGCCAGTGCAAATCATTTCAACACAC	387	60Cº
DMDEX13-R	CAGGAAACAGCTATGACCTCTTTAAATCACAGCACTTC		
DMDEX14-15-F	TGTAAAACGACGGCCAGTTGGCAAATTATTCATGCCATT	548	63Cº
DMDEX14-15-R	CAGGAAACAGCTATGACCTGATCCAAGCAAAAATAAACATT		
DMDEX16-F	TGTAAAACGACGGCCAGTATGCAACCCAGGCTTATTC	286	61Cº
DMDEX16-R	CAGGAAACAGCTATGACCCTGTAGCATGATAATTGGTATCAC		
DMDEX17-F	TGTAAAACGACGGCCAGTTTTTCCTTTGCCACTCCAAG	362	61Cº
DMDEX17-R	CAGGAAACAGCTATGACCCACCACCAACAAAACTGCTG		
DMDEX18-F	CGACGTTGTAAAACGACGGCCAGTTGTCAGGCAGGAGTCTCAgat	339	63Cº
DMDEX18-R	CAGGAAACAGCTATGACCCGGAGTTTACAAGCAGCACA		

DMDEX19-F	TGTAAAACGACGGCCAGT gattcacgtgataagctgacaga	286	63Cº
DMDEX19-R	CAGGAAACAGCTATGACCCGCCCGCCGCGCttcagctgataaatatgaacc tatgt		
DMDEX20-F	TGTAAAACGACGGCCAGTTGGCTTTCAGATCATTTCTTTC	393	61Cº
DMDEX20-R	CAGGAAACAGCTATGACCAAATACCTATTGATTATGCTCC		
DMDEX21-F	TGTAAAACGACGGCCAGTGCAAAATGTAATGTATGCAAAG	355	63Cº
DMDEX21-R	CAGGAAACAGCTATGACCATGTTAGTACCTTCTGGATTTC		
DMDEX22-F	TGTAAAACGACGGCCAGTAGGAAAACATGGCAAAGTGTG	370	63Cº
DMDEX22-R	CAGGAAACAGCTATGACCTGCTCAATGGGCAAACTACC		
DMDEX23 A-F	TGTAAAACGACGGCCAGTACTCATCAATTATTAttcatcaattagggt	126	61Cº
DMDEX23 A-R	CAGGAAACAGCTATGACCCATCTCTTTCACAGTGGTGC		
DMDEX23 B-F	TGTAAAACGACGGCCAGTAGCAACAAAGTGGCCTATAC	135	61Cº
DMDEX23 B-F	CAGGAAACAGCTATGACCGCTGGGAGGAGAGCTTC		
DMDEX23 C-F	TGTAAAACGACGGCCAGTTTGAAGAAATTGAGGGACGC	175	61Cº
DMDEX23C-R	CAGGAAACAGCTATGACCCTTTACAGTTTACAGTGTATcgttagg		
DMDEX24-F	TGTAAAACGACGGCCAGTTTGGGCCTGTGTTTAGACATA	327	63Cº
DMDEX24-R	CAGGAAACAGCTATGACCAAATCCACCCCAGCTGTAAAA		
DMDEX25-F	TGTAAAACGACGGCCAGTTGTGGCAGTAATTTTTTTCAG	296	61Cº
DMDEX25-R	CAGGAAACAGCTATGACCAGGAAATCTTAGTTAAGTACG		
DMDEX26-F	TGTAAAACGACGGCCAGTTGAGTGTATCTGATCCCCATGA	438	61Cº
DMDEX26-R	CAGGAAACAGCTATGACCTGTTGCATTTCTTTCTTTTTC		
DMDEX27-F	TGTAAAACGACGGCCAGTTGGGATGTTGTGAGAAAGAA	365	63Cº
DMDEX27-R	CAGGAAACAGCTATGACCTGACCATGTATTGACATATCATTGA		
DMDEX28-F	TGTAAAACGACGGCCAGTGAAGTTTTAATAATGAAATGGCaaaa	311	61Cº
DMDEX28-R	CAGGAAACAGCTATGACCTGACCTCTTTTAATACTGCATAT		
DMDEX29-F	TGTAAAACGACGGCCAGTCCAATGTATTTAGAAAAAAAAGGAG	279	63Cº
DMDEX29-R	CAGGAAACAGCTATGACCGCAAATTAGATTAAAGAGAtttttCAC		
DMDEX30-F	TGTAAAACGACGGCCAGTTACAGAAAAGCTATCAAGAG	297	61Cº
DMDEX30-R	CAGGAAACAGCTATGACCAAAAACAAAAGAATGGAAGC		
DMDEX31-F	TGTAAAACGACGGCCAGTATGGTAGAGGTGGTTGAGGA	296	61Cº
DMDEX31-R	CAGGAAACAGCTATGACCTATAATGCCCAACGAAAACA		
DMDEX32-F	TGTAAAACGACGGCCAGTCAGTTATTGTTTGAAAGGCAAA	322	61Cº
DMDEX32-R	CAGGAAACAGCTATGACCCTTCTTAATGAGGAAAGTCAAGG		
DMDEX33-F	CGACGTTGTAAAACGACGGCCAGTTGGAATAGCAATTAAGGG	393	60Cº
DMDEX33-R	CAGGAAACAGCTATGACCGCTAAGACTCTAATCATAC		
DMDEX34-F	TGTAAAACGACGGCCAGTCAGAAATATAAAAGTTCCaaataagtg	374	61Cº
DMDEX34-R	CAGGAAACAGCTATGACCCATGTTAATACTTCCTTACAAAATC		
DMDEX35-F	TGTAAAACGACGGCCAGTCCGTTTCATAAGCATTAAATC	307	61Cº
DMDEX35-R	CAGGAAACAGCTATGACCAGCTTCTAGCCTTTTCTC		
DMDEX36-F	CGACGTTGTAAAACGACGGCCAGTTGTCTAACCAATAATGCcatg	257	64Cº
DMDEX36-R	CAGGAAACAGCTATGACCCTGGTGTACAATTTGGACA		
DMDEX37-F	CGACGTTGTAAAACGACGGCCAGTCTTTCTCACTCTTCTCGctcac	377	61Cº
DMDEX37-R	CAGGAAACAGCTATGACCTTCGCAAGAGACCATTTAGCAC		

DMDEX38-F	TGTAAAACGACGGCCAGTTTTAGCAACAGGAGGTTGAA	267	64Cº
DMDEX38-R	CAGGAAACAGCTATGACCTTCTTTCCAAATATTTATTTCCACT		
DMDEX39-F	TGTAAAACGACGGCCAGTCTCTGTTAACAATGTACAGCTTTTT	365	64Cº
DMDEX39-R	CAGGAAACAGCTATGACCAAAAACCACAGGCAAGGTAT		
DMDEX40-F	TGTAAAACGACGGCCAGTTACAAAAAGATGAGGGAC	387	61Cº
DMDEX40-R	CAGGAAACAGCTATGACCAATAGAAACAAGAACATCAAC		
DMDEX41-F	TGTAAAACGACGGCCAGTGTTAGCTAACTGCCCTGGGccctgtattg	311	61Cº
DMDEX41-R	CAGGAAACAGCTATGACCTAGAGTAGTAGTTGCaaacacatacgtgg		
DMDEX42-F	TGTAAAACGACGGCCAGTATGGAGGAGGTTTCACTGTT	408	61Cº
DMDEX42-R	CAGGAAACAGCTATGACCCCATGTGAAAGTCAAAATGC		
DMDEX43-F	TGTAAAACGACGGCCAGTTTTCTATAGACAGCTAATTCATTTTT	287	63Cº
DMDEX43-R	CAGGAAACAGCTATGACCACAGTTCCCTGAAAACAAATC		
DMDEX44-F	TGTAAAACGACGGCCAGTGTTACTTGAAACTAAACTCTGCaaatg	444	61Cº
DMDEX44-R	CAGGAAACAGCTATGACCACAACAACAGTCAAAAGTAAtttccatc		
DMDEX45-F	TGTAAAACGACGGCCAGTTTCTTTGCCAGTACAACTGC	357	61Cº
DMDEX45-R	CAGGAAACAGCTATGACCTCTGCTAAAATGTTTTCATTCC		
DMDEX46-F	TGTAAAACGACGGCCAGTCCAGTTTGCATTAACAAATAGtttgag	409	64Cº
DMDEX46-R	CAGGAAACAGCTATGACCAGGGTTAAGAAGAAATAAAgttgtgag		
DMDEX47-F	TGTAAAACGACGGCCAGTTGATAGACTAATCAATAGaagcaaagac	399	61Cº
DMDEX47-R	CAGGAAACAGCTATGACCAACAAAACAAAACAACAATccacatacc		
DMDEX48 A-F	TGTAAAACGACGGCCAGTTTTGGCTTATGCCTTGAGAAT	175	61Cº
DMDEX48 A-R	CAGGAAACAGCTATGACCATAACCACAGCAGCAGATG		
DMDEX48 B-F	TGTAAAACGACGGCCAGTGCTTGAAGACCTTGAAGAGC	185	61Cº
DMDEX48 B-R	CAGGAAACAGCTATGACCAAATGAGAAAATTCAGTGATATTGCC		
DMDEX49-F	TGTAAAACGACGGCCAGTGTGCCCTTATGTACCAGGCAGAAATTG	475	61Cº
DMDEX49-R	CAGGAAACAGCTATGACCGCAATGACTCGTTAATAGCCTTAAGATC		
DMDEX50-F	TGTAAAACGACGGCCAGTCACCAAATGGATTAAGATGTTCATGAAT	307	64Cº
DMDEX50-R	CAGGAAACAGCTATGACCTCTCTCTCACCCAGTCATCACTTCATAG		
DMDEX51-F	TGTAAAACGACGGCCAGTGAAATTGGCTCTTTAGCTTGTGTTTC	424	64Cº
DMDEX51-R	CAGGAAACAGCTATGACCGGAGAGTAAAGTGATTGGTGGAAAATC		
DMDEX52-F	TGTAAAACGACGGCCAGTGTGTTTTGGCTGGTCTCACA	298	63Cº
DMDEX52-R	CAGGAAACAGCTATGACCCATGCATCTTGCTTTGTGTGT		
DMDEX53 A-F	TGTAAAACGACGGCCAGTAAGAATCCTGTTGTTCATCATCCTAGC	252	64Cº
DMDEX53 A-R	CAGGAAACAGCTATGACC CCAGCCATTGTGTTGAATCCTTTAAC		
DMDEX53 B-F	TGTAAAACGACGGCCAGT AGTACAAGAACACCTTCAGAACCG	278	64Cº
DMDEX53 B-R	CAGGAAACAGCTATGACCactttacattaaacatcattaaattacaatctatgg		
DMDEX54-F	CGACGTTGTAAAACGACGGCCAGTGTATTCTGACCTGAGGATTC	378	61Cº
DMDEX54-R	CAGGAAACAGCTATGACCCATGGTCCATCCAGTTTC		
DMDEX55-F	TGTAAAACGACGGCCAGTAATTTAGTTCCTCCATCTTTCTCT	445	61Cº
DMDEX55-R	CAGGAAACAGCTATGACCAAATACATCAGGCTGTATAAAAGC		
DMDEX56-F	TGTAAAACGACGGCCAGTATTCTGCACATATTCTTCTTCCTGC	353	63Cº

DMDEX56-R	CAGGAAACAGCTATGACCGGATGATTTACGTAGACATGTGAG		
DMDEX57-F	TGTAAAACGACGGCCAGTCAATGGAATTGTTAGAATCATCA	320	63Cº
DMDEX57-R	CAGGAAACAGCTATGACCCACTGGATTACTATGTGCTTAACAT		
DMDEX58-F	TGTAAAACGACGGCCAGTTTTTGAGAAGAATGCCACAAGCC	315	63Cº
DMDEX58-R	CAGGAAACAGCTATGACCAAAATATGAGAGCTATCCAGACCC		
DMDEX59-F	TGTAAAACGACGGCCAGTAAAGAATGTGGCCTAAAACC	433	64Cº
DMDEX59-R	CAGGAAACAGCTATGACCTTGTGGGAAGATAACACTGC		
DMDEX60-F	TGTAAAACGACGGCCAGTTAAATATTCTCATCTTCCAATTTGC	267	63Cº
DMDEX60-R	CAGGAAACAGCTATGACCTTACTGTAACAAAGGACAACAATG		
DMDEX61A-F	TGTAAAACGACGGCCAGTCGCCCGCCGCttgctttagtgttctcagtcttgg	169	63Cº
DMDEX61A-R	CAGGAAACAGCTATGACCAAAGTCCCTGTGGGCTTCAT		
DMDEX61B-F	TGTAAAACGACGGCCAGTCGTCGAGGACCGAGTCAG	210	63Cº
DMDEX61B-R	CAGGAAACAGCTATGACCCGCCCGCCGCcaggatgatttatgcttctactgc		
DMDEX62-F	TGTAAAACGACGGCCAGTTAATGTTGTCTTTCCTGTTTGCG	221	63Cº
DMDEX62-R	CAGGAAACAGCTATGACCATACAGGTTAGTCACAATAAATGC		
DMDEX63-F	TGTAAAACGACGGCCAGTTACTCATTGTAAATGCTAAAGTC	229	63Cº
DMDEX63-R	CAGGAAACAGCTATGACCTAGCAAGTAACTTTCACACTGC		
DMDEX64-F	TGTAAAACGACGGCCAGTTTCTGATGGAATAACAAATGCT	322	61Cº
DMDEX64-R	CAGGAAACAGCTATGACCCATTCTAGGCAAACTCTAGGC		
DMDEX65-F	TGTAAAACGACGGCCAGTagtgtggttcacgtttggt	386	64Cº
DMDEX65-R	CAGGAAACAGCTATGACCtgtacgctaagcctcctgtg		
DMDEX66-F	TGTAAAACGACGGCCAGTGTCAGTAATTGTTTTCTGCTTTG	246	61Cº
DMDEX66-R	CAGGAAACAGCTATGACCATAAGAACAGTCTGTCATTTCCC		
DMDEX67 A-F	TGTAAAACGACGGCCAGTTCAGGTTCTGCTGGCATC	172	60Cº
DMDEX67 A-R	CAGGAAACAGCTATGACCTGCAACTTCACCCAACTGTC		
DMDEX67 B-F	TGTAAAACGACGGCCAGTGCCTCCTTCTGCATGATT	187	61Cº
DMDEX67 B-R	CAGGAAACAGCTATGACCAGAAAACGAAGCTCTGTGG		
DMDEX68 A-F	TGTAAAACGACGGCCAGTCGCCCGCCcagcctagctttgcaaccat	249	61Cº
DMDEX68 A-R	CAGGAAACAGCTATGACCACTGGGGTTCCAGTCTCATC		
DMDEX68 B-F	TGTAAAACGACGGCCAGTAGCGGCCCTCTTCCTAGACT	236	61Cº
DMDEX68 B-R	CAGGAAACAGCTATGACCCGCCCGC taacagcaactggcacagga		
DMDEX69-F	TGTAAAACGACGGCCAGTGAACGTGGTAGAAGGTTTATTAAA	267	61Cº
DMDEX69-R	CAGGAAACAGCTATGACCCTAACTCTCACGTCAGGCTG		
DMDEX70-F	TGTAAAACGACGGCCAGTTGGTCATTAGTTTTGAAATCATC	273	63Cº
DMDEX70-R	CAGGAAACAGCTATGACCCATCAAACAAGAGTGTGTTCTG		
DMDEX71-F	TGTAAAACGACGGCCAGTGGCTGAGTTTGCGTGTGTCT	174	61Cº
DMDEX71-R	CAGGAAACAGCTATGACCGAGCGAATGTGTTGGTGGTA		
DMDEX72-F	TGTAAAACGACGGCCAGTAAGCATTCTAGGCCATGTGT	261	61Cº
DMDEX72-R	CAGGAAACAGCTATGACCGGTTAGCTTTCCTTGGTTAGTT		
DMDEX73-F	TGTAAAACGACGGCCAGTACGTCACATAAGTTTTAATGAGC	238	63Cº
DMDEX73-R	CAGGAAACAGCTATGACCATGCTAATTCCTATATCCTGTGC		
DMDEX74-F	TGTAAAACGACGGCCAGTATAAGGGGGGGAAAAAAC	290	63Cº
DMDEX74-R	CAGGAAACAGCTATGACCTGCAAGTGTATGCACTCTG		

DMDEX75-F	TGTAAAACGACGGCCAGTTCTTTTTTACTTTTTTGATGC	380	60Cº
DMDEX75-R	CAGGAAACAGCTATGACCAGTGCTCTCTGAGGTTTAG		
DMDEX76 A-F	TGTAAAACGACGGCCAGTacaatctttgggagggcttc	231	63Cº
DMDEX76 A-R	CAGGAAACAGCTATGACCCTGACTGCTGTCGGACCTCT		
DMDEX76 B-F	TGTAAAACGACGGCCAGTCACAACGGTGTCCTCTCCTT	216	63Cº
DMDEX76 B-R	CAGGAAACAGCTATGACCttcagtggctccctgatacc		
DMDEX77-F	TGTAAAACGACGGCCAGTTAATCATGGCCCTTTAATATCTG	306	63Cº
DMDEX77-R	CAGGAAACAGCTATGACCGATACTGCGTGTTGGCTTCC		
DMDEX78-F	TGTAAAACGACGGCCAGTTTCTGATATCTCTGCCTCTTCC	267	61Cº
DMDEX78-R	CAGGAAACAGCTATGACCCATGAGCTGCAAGTGGAGAGG		
DMDEX79 A-F	TGTAAAACGACGGCCAGTAGAGTGATGCTATCTATCTGCAC	385	61Cº
DMDEX79 A-R	CAGGAAACAGCTATGACCTGCATAGACGTGTAAAACCTGCC		
DMDEX79B 1 – F	TGTAAAACGACGGCCAGTTTGTGAAGGGTAGTGGTATTATACTG	323	60Cº
DMDEX79B 1 – R	CAGGAAACAGCTATGACCTGCCTCAAAGTTTTGTGTGTG		
DMDEX79B 2 – F	TGTAAAACGACGGCCAGTCGCCCGCCGAACGCATTTTGGGTTGTTT	284	60Cº
DMDEX79B 2 – R	CAGGAAACAGCTATGACCTCAAATGAGCAGTGTGTAGTAGTCA		
DMDEX79C1-F	TGTAAAACGACGGCCAGTCTTCCTCTACCACCACACCAA	242	60Cº
DMDEX79 C1-R	CAGGAAACAGCTATGACCAAGCAGGTAAGCCTGGATGA		
DMDEX79 C2-F	TGTAAAACGACGGCCAGTTGTTCATGTCACATCCTAATAGAAA	309	60Cº
DMDEX79 C2-R	CAGGAAACAGCTATGACCCGCCCGCCGTAGCAGCAGGAAGCTGAATG		
DMDEX79D 1 – F	TGTAAAACGACGGCCAGTCGCCCGCCGCGAGTAATCGGTTGGTTGGttga	265	60Cº
DMDEX79D 1 -R	CAGGAAACAGCTATGACC TCCTTCACTTAAAGAGTGGCCTA		
DMDEX79D 2 – F	TGTAAAACGACGGCCAGTGCTGGAGGGCTATGGATTC	280	60Cº
DMDEX79D 2 – R	CAGGAAACAGCTATGACCCGCCCGCCGTCACAAATGTGATGGGGCTA		
DMDEX79E –F	TGTAAAACGACGGCCAGTAATAAACTTTGGGAAAAGGTG	536	64Cº
DMDEX79E –R	CAGGAAACAGCTATGACCGAAGCCGTGTTTGATGTTAAT		
DMDEX79F-F	TGTAAAACGACGGCCAGTGAGAGTGGGCTGACATCAA	532	61Cº
DMDEX79F-R	CAGGAAACAGCTATGACCTCACTCCAGAGCTAATGTGTCT		
DMDEX79G-F	TGTAAAACGACGGCCAGTAGTAAGTTTCATTCTAAAATCAGAGG	531	61Cº
DMDEX79G-R	CAGGAAACAGCTATGACCGTGTTTTCACTGTCTTTCTGGA		

Supplementary Table 2 Exon (fragment) number, PCR buffers and $MgCl_2$ concentrations for sequencing.

Fragment #	**Buffer system**	**$MgCl_2$**
5,22,48,53	10x AT(Applied Biosystems)	1.5mM
4,6,7,8,9,10,12,13, 16,17,18,19,20,21, 25,26,29,30 ,32, 35, 36, 37,38,39,40,41,42,44, 45,46,47,49,50,51,54, 55,58,59,60,61,62,64,66,68,69,72,73, 75,76,77,79A,79G	10x ST (Promega)	1.5mM
1,2,23,24,31,33,34,43, 56,57,71,74	10x ST (Promega)	3 mM
3,11,27,28,63,65,67,70,78,79B,79C,79D,79F	5x STR (LDGA)	1.5 mM
14/15,52	5x STR (LDGA)	3 mM

LDGA: Laboratory of Diagnostic Genome Analysis

Supplementary Table 3 Primer sequences (with M13 tail) that were used for the amplification of DMD amplicons for sequencing.

Amplicon number	Primer sequences
DMDEX01-F4	TGTAAAACGACGGCCAGTGCAGGTCCTGGAATTTGA
DMDEX01-R4	CAGGAAACAGCTATGACCCAAACTAAACGTTATGCCACA
DMDEX02-F5	TGTAAAACGACGGCCAGTCACTAACACATCATAATGG
DMDEX02-R2	CAGGAAACAGCTATGACCGATACACAGGTACATAGTC
DMDEX03-F5	TGTAAAACGACGGCCAGTTCATCCGTCATCTTCGGCAGATTAA
DMDEX03-R4	CAGGAAACAGCTATGACCCAGGCGGTAGAGTATGCCAAATGAAAATCA
DMDEX04-F3	TGTAAAACGACGGCCAGTTTGTCGGTCTCTCTGCTGGTCAGTG
DMDEX04-R2	CAGGAAACAGCTATGACCCCAAAGCCCTCACTCAAAC
DMDEX05-F3	TGTAAAACGACGGCCAGTCAACTAGGCATTTGGTCTC
DMDEX05-R3	CAGGAAACAGCTATGACCTTGTTTCACACGTCAAGGG
DMDEX06-F6	TGTAAAACGACGGCCAGTTGGTTCTTGCTCAAGGAATG
DMDEX06-R6	CAGGAAACAGCTATGACCTGGGGAAAAATATGTCATCAG
DMDEX07-F3	TGTAAAACGACGGCCAGTCTATGGGCATTGGTTGTC
DMDEX07-R3	CAGGAAACAGCTATGACCAAAAGCAGTGGTAGTCCAG
DMDEX08-F5	TGTAAAACGACGGCCAGTTCGTCTTCCTTTAACTTTG
DMDEX08-R5	CAGGAAACAGCTATGACCTCTTGAATAGTAGCTGTCC
DMDEX09-F4	TGTAAAACGACGGCCAGTTCTATCCACTCCCCCAAACC
DMDEX09-R4	CAGGAAACAGCTATGACCAACAAACCAGCTCTTCAC
DMDEX10-F1	CGACGTTGTAAAACGACGGCCAGTGGAACAATCTGCAAAGAC
DMDEX10-R1	CAGGAAACAGCTATGACCAAAGGATGACTTGCCATTATAAC
DMDEX11-F5	TGTAAAACGACGGCCAGTCAAATAAAACTCAAAACCACACC
DMDEX11-R3	CAGGAAACAGCTATGACCCTTCCAAAACTTGTTAGTCTTC
DMDEX12-F2	TGTAAAACGACGGCCAGTCTTTCAAAGAGGTCATAATAGG

DMDEX12-R2	CAGGAAACAGCTATGACCCATCTGTGTTACTGTGTATAGG
DMDEX13-F3	TGTAAAACGACGGCCAGTGCAAATCATTTCAACACAC
DMDEX13-R3	CAGGAAACAGCTATGACCTCTTTAAATCACAGCACTTC
DMDEX14-F2	TGTAAAACGACGGCCAGTTGGCAAATTATTCATGCCATT
DMDEX14-R2	CAGGAAACAGCTATGACCTGATCCAAGCAAAAATAAACATT
DMDEX16-F3	TGTAAAACGACGGCCAGTATGCAACCCAGGCTTATTC
DMDEX16-R3	CAGGAAACAGCTATGACCCTGTAGCATGATAATTGGTATCAC
DMDEX17-F6	TGTAAAACGACGGCCAGTTTTTCCTTTGCCACTCCAAG
DMDEX17-R4	CAGGAAACAGCTATGACCCACCACCAACAAAACTGCTG
DMDEX18-F1	CGACGTTGTAAAACGACGGCCAGTTGTCAGGCAGGAGTCTCAGAT
DMDEX18-R1	CAGGAAACAGCTATGACCCGGAGTTTACAAGCAGCACA
DMDEX19-F4	TGTAAAACGACGGCCAGTGATGGCAAAAGTGTTGAGAAAAAGTC
DMDEX19-R4	CAGGAAACAGCTATGACCTTCTACCACATCCCATTTTCTTCCA
DMDEX20-F2	TGTAAAACGACGGCCAGTTGGCTTTCAGATCATTTCTTTC
DMDEX20-R2	CAGGAAACAGCTATGACCAAATACCTATTGATTATGCTCC
DMDEX21-F3	TGTAAAACGACGGCCAGTGCAAAATGTAATGTATGCAAAG
DMDEX21-R3	CAGGAAACAGCTATGACCATGTTAGTACCTTCTGGATTTC
DMDEX22-F2	TGTAAAACGACGGCCAGTAGGAAAACATGGCAAAGTGTG
DMDEX22-R2	CAGGAAACAGCTATGACCTGCTCAATGGGCAAACTACC
DMDEX23-F3	TGTAAAACGACGGCCAGTTCATCTACTTTGTTTACATGTTTGAA
DMDEX23-R3	CAGGAAACAGCTATGACCACAGTGTATCGTTAGGGAAAAA
DMDEX24-F4	TGTAAAACGACGGCCAGTTTGGGCCTGTGTTTAGACATA
DMDEX24-R3	CAGGAAACAGCTATGACCAAATCCACCCCAGCTGTAAAA
DMDEX25-F2	TGTAAAACGACGGCCAGTTGTGGCAGTAATTTTTTTCAG
DMDEX25-R2	CAGGAAACAGCTATGACCAGGAAATCTTAGTTAAGTACG
DMDEX26-F3	TGTAAAACGACGGCCAGTTGAGTGTATCTGATCCCCATGA
DMDEX26-R1	CAGGAAACAGCTATGACCTGTTGCATTTCTTTCTTTTTC
DMDEX27-F2	TGTAAAACGACGGCCAGTTGGGATGTTGTGAGAAAGAA
DMDEX27-R4	CAGGAAACAGCTATGACCTGACCATGTATTGACATATCATTGA
DMDEX28-F2	TGTAAAACGACGGCCAGTGAAGTTTTAATAATGAAATGGCAAAA
DMDEX28-R3	CAGGAAACAGCTATGACCTGACCTCTTTTAATACTGCATAT
DMDEX29-F4	TGTAAAACGACGGCCAGTCCAATGTATTTAGAAAAAAAAGGAG
DMDEX29-R5	CAGGAAACAGCTATGACCGCAAATTAGATTAAAGAGATTTTTCAC
DMDEX30-F4	TGTAAAACGACGGCCAGTTACAGAAAAGCTATCAAGAG
DMDEX30-R3	CAGGAAACAGCTATGACCAAAAACAAAAGAATGGAAGC
DMDEX31-F2	TGTAAAACGACGGCCAGTATGGTAGAGGTGGTTGAGGA
DMDEX31-R2	CAGGAAACAGCTATGACCTATAATGCCCAACGAAAACA
DMDEX32-F2	TGTAAAACGACGGCCAGTCAGTTATTGTTTGAAAGGCAAA
DMDEX32-R2	CAGGAAACAGCTATGACCCTTCTTAATGAGGAAAGTCAAGG
DMDEX33-F1	CGACGTTGTAAAACGACGGCCAGTTGGAATAGCAATTAAGGG
DMDEX33-R1	CAGGAAACAGCTATGACCGCTAAGACTCTAATCATAC
DMDEX34-F3	TGTAAAACGACGGCCAGTCAGAAATATAAAAGTTCCAAATAAGTG
DMDEX34-R3	CAGGAAACAGCTATGACCCATGTTAATACTTCCTTACAAAATC

DMDEX35-F2	TGTAAAACGACGGCCAGTCCGTTTCATAAGCATTAAATC
DMDEX35-R3	CAGGAAACAGCTATGACCAGCTTCTAGCCTTTTCTC
DMDEX36-F1	CGACGTTGTAAAACGACGGCCAGTTGTCTAACCAATAATGCCATG
DMDEX36-R1	CAGGAAACAGCTATGACCCTGGTGTACAATTTGGACA
DMDEX37-F1	CGACGTTGTAAAACGACGGCCAGTCTTTCTCACTCTTCTCGCTCAC
DMDEX37-R1	CAGGAAACAGCTATGACCTTCGCAAGAGACCATTTAGCAC
DMDEX38-F3	TGTAAAACGACGGCCAGTTTTAGCAACAGGAGGTTGAA
DMDEX38-R3	CAGGAAACAGCTATGACCTTCTTTCCAAATATTTATTTCCACT
DMDEX39-F3	TGTAAAACGACGGCCAGTCTCTGTTAACAATGTACAGCTTTTT
DMDEX39-R3	CAGGAAACAGCTATGACCAAAAACCACAGGCAAGGTAT
DMDEX40-F2	TGTAAAACGACGGCCAGTTACAAAAAGATGAGGGAC
DMDEX40-R2	CAGGAAACAGCTATGACCAATAGAAACAAGAACATCAAC
DMDEX41-F2	TGTAAAACGACGGCCAGTGTTAGCTAACTGCCCTGGGCCCTGTATTG
DMDEX41-R2	CAGGAAACAGCTATGACCTAGAGTAGTAGTTGCAAACACATACGTGG
DMDEX42-F3	TGTAAAACGACGGCCAGTATGGAGGAGGTTTCACTGTT
DMDEX42-R3	CAGGAAACAGCTATGACCCCATGTGAAAGTCAAAATGC
DMDEX43-F6	TGTAAAACGACGGCCAGTTTTCTATAGACAGCTAATTCATTTTT
DMDEX43-R6	CAGGAAACAGCTATGACCACAGTTCCCTGAAAACAAATC
DMDEX44-F5	TGTAAAACGACGGCCAGTGTTACTTGAAACTAAACTCTGCAAATG
DMDEX44-R2	CAGGAAACAGCTATGACCACAACAACAGTCAAAAGTAATTTCCATC
DMDEX45-F5	TGTAAAACGACGGCCAGTTTCTTTGCCAGTACAACTGC
DMDEX45-R4	CAGGAAACAGCTATGACCTCTGCTAAAATGTTTTCATTCC
DMDEX46-F4	TGTAAAACGACGGCCAGTCCAGTTTGCATTAACAAATAGTTTGAG
DMDEX46-R4	CAGGAAACAGCTATGACCAGGGTTAAGAAGAAATAAAGTTGTGAG
DMDEX47-F4	TGTAAAACGACGGCCAGTTGATAGACTAATCAATAGAAGCAAAGAC
DMDEX47-R4	CAGGAAACAGCTATGACCAACAAAACAAAACAACAATCCACATACC
DMDEX48-F4	TGTAAAACGACGGCCAGTTTGAATACATTGGTTAAATCCCAACATG
DMDEX48-R4	CAGGAAACAGCTATGACCCCTGAATAAAGTCTTCCTTACCACAC
DMDEX49-F4	TGTAAAACGACGGCCAGTGTGCCCTTATGTACCAGGCAGAAATTG
DMDEX49-R4	CAGGAAACAGCTATGACCGCAATGACTCGTTAATAGCCTTAAGATC
DMDEX50-F3	TGTAAAACGACGGCCAGTCACCAAATGGATTAAGATGTTCATGAAT
DMDEX50-R2	CAGGAAACAGCTATGACCTCTCTCTCACCCAGTCATCACTTCATAG
DMDEX51-F3	TGTAAAACGACGGCCAGTGAAATTGGCTCTTTAGCTTGTGTTTC
DMDEX51-R3	CAGGAAACAGCTATGACCGGAGAGTAAAGTGATTGGTGGAAAATC
DMDEX52-F4	TGTAAAACGACGGCCAGTGTGTTTTGGCTGGTCTCACA
DMDEX52-R4	CAGGAAACAGCTATGACCCATGCATCTTGCTTTGTGTGT
DMDEX53-F3	TGTAAAACGACGGCCAGTTCCTCCAGACTAGCATTTAC
DMDEX53-R3	CAGGAAACAGCTATGACCTTAGCCTGGGTGACAGTG
DMDEX54-F2	CGACGTTGTAAAACGACGGCCAGTGTATTCTGACCTGAGGATTC
DMDEX54-R2	CAGGAAACAGCTATGACCCATGGTCCATCCAGTTTC
DMDEX55-F3	TGTAAAACGACGGCCAGTAATTTAGTTCCTCCATCTTTCTCT
DMDEX55-R7	CAGGAAACAGCTATGACCAAATACATCAGGCTGTATAAAAGC
DMDEX56-F2	TGTAAAACGACGGCCAGTATTCTGCACATATTCTTCTTCCTGC

DMDEX56-R2	CAGGAAACAGCTATGACCGGATGATTTACGTAGACATGTGAG
DMDEX57-F2	TGTAAAACGACGGCCAGTCAATGGAATTGTTAGAATCATCA
DMDEX57-R2	CAGGAAACAGCTATGACCCACTGGATTACTATGTGCTTAACAT
DMDEX58-F4	TGTAAAACGACGGCCAGTTTTTGAGAAGAATGCCACAAGCC
DMDEX58-R4	CAGGAAACAGCTATGACCAAAATATGAGAGCTATCCAGACCC
DMDEX59-F6	TGTAAAACGACGGCCAGTAAAGAATGTGGCCTAAAACC
DMDEX59-R6	CAGGAAACAGCTATGACCTTGTGGGAAGATAACACTGC
DMDEX60-F3	TGTAAAACGACGGCCAGTTAAATATTCTCATCTTCCAATTTGC
DMDEX60-R3	CAGGAAACAGCTATGACCTTACTGTAACAAAGGACAACAATG
DMDEX61-F3	TGTAAAACGACGGCCAGTCATTGTTTTAATTGTTCCTCATT
DMDEX61-R3	CAGGAAACAGCTATGACCTTCAACTCTTAATTCTTTTGTTTTT
DMDEX62-F4	TGTAAAACGACGGCCAGTTAATGTTGTCTTTCCTGTTTGCG
DMDEX62-R4	CAGGAAACAGCTATGACCATACAGGTTAGTCACAATAAATGC
DMDEX63-F3	TGTAAAACGACGGCCAGTTACTCATTGTAAATGCTAAAGTC
DMDEX63-R3	CAGGAAACAGCTATGACCTAGCAAGTAACTTTCACACTGC
DMDEX64-F2	TGTAAAACGACGGCCAGTTTCTGATGGAATAACAAATGCT
DMDEX64-R2	CAGGAAACAGCTATGACCCATTCTAGGCAAACTCTAGGC
DMDEX65-F4	TGTAAAACGACGGCCAGTGGTTTTACTCTTTGAGTCATTTGT
DMDEX65-R4	CAGGAAACAGCTATGACCTACGCTAAGCCTCCTGTGAC
DMDEX66-F5	TGTAAAACGACGGCCAGTGTCAGTAATTGTTTTCTGCTTTG
DMDEX66-R3	CAGGAAACAGCTATGACCATAAGAACAGTCTGTCATTTCCC
DMDEX67-F3	TGTAAAACGACGGCCAGTGAAGTAACCCCACTCTGTGGAA
DMDEX67-R2	CAGGAAACAGCTATGACCAAACGAAGCTCTGTGGGTTT
DMDEX68-F1	TGTAAAACGACGGCCAGTTAATCGAACTGATATACACCTCC
DMDEX68-R1	CAGGAAACAGCTATGACCACTAACAGCAACTGGCACAGG
DMDEX69-F3	TGTAAAACGACGGCCAGTGAACGTGGTAGAAGGTTTATTAAA
DMDEX69-R3	CAGGAAACAGCTATGACCCTAACTCTCACGTCAGGCTG
DMDEX70-F3	TGTAAAACGACGGCCAGTTGGTCATTAGTTTTGAAATCATC
DMDEX70-R3	CAGGAAACAGCTATGACCCATCAAACAAGAGTGTGTTCTG
DMDEX71-F5	TGTAAAACGACGGCCAGTGGCTGAGTTTGCGTGTGTCT
DMDEX71-R3	CAGGAAACAGCTATGACCGAGCGAATGTGTTGGTGGTA
DMDEX72-F3	TGTAAAACGACGGCCAGTAAGCATTCTAGGCCATGTGT
DMDEX72-R3	CAGGAAACAGCTATGACCGGTTAGCTTTCCTTGGTTAGTT
DMDEX73-F2	TGTAAAACGACGGCCAGTACGTCACATAAGTTTTAATGAGC
DMDEX73-R2	CAGGAAACAGCTATGACCATGCTAATTCCTATATCCTGTGC
DMDEX74-F1	TGTAAAACGACGGCCAGTATAAGGGGGGGAAAAAAC
DMDEX74-R1	CAGGAAACAGCTATGACCTGCAAGTGTATGCACTCTG
DMDEX75-F1	TGTAAAACGACGGCCAGTTCTTTTTTACTTTTTTGATGC
DMDEX75-R1	CAGGAAACAGCTATGACCAGTGCTCTCTGAGGTTTAG
DMDEX76-F4	TGTAAAACGACGGCCAGTGGGTCAAAATTTATGAGTCCTG
DMDEX76-R3	CAGGAAACAGCTATGACCTTCATGTCCCTGTAATACGACT
DMDEX77-F1	TGTAAAACGACGGCCAGTTAATCATGGCCCTTTAATATCTG
DMDEX77-R1	CAGGAAACAGCTATGACCGATACTGCGTGTTGGCTTCC

DMDEX78-F2	TGTAAAACGACGGCCAGTTTCTGATATCTCTGCCTCTTCC
DMDEX78-R3	CAGGAAACAGCTATGACCCATGAGCTGCAAGTGGAGAGG
DMDEX79-F2	TGTAAAACGACGGCCAGTAGAGTGATGCTATCTATCTGCAC
DMDEX79-R2	CAGGAAACAGCTATGACCTGCATAGACGTGTAAAACCTGCC
DMDEX79-F3	TGTAAAACGACGGCCAGTATTTTTGTGAAGGGTAGTGGT
DMDEX79-R3	CAGGAAACAGCTATGACCGAAAAAGTCAGTCTATAGAAATTCG
DMDEX79-F4	TGTAAAACGACGGCCAGTCCACCACACCAAATGACTAC
DMDEX79-R4	CAGGAAACAGCTATGACCATCTAAATCGTGGCATTGCT
DMDEX79-F5	TGTAAAACGACGGCCAGTAGTAATCGGTTGGTTGGTTG
DMDEX79-R5	CAGGAAACAGCTATGACCAACACAGTTCATGGGCTTCT
DMDEX79-F6	TGTAAAACGACGGCCAGTAATAAACTTTGGGAAAAGGTG
DMDEX79-R6	CAGGAAACAGCTATGACCGAAGCCGTGTTTGATGTTAAT
DMDEX79-F7	TGTAAAACGACGGCCAGTGAGAGTGGGCTGACATCAA
DMDEX79-R7	CAGGAAACAGCTATGACCTCACTCCAGAGCTAATGTGTCT
DMDEX79-F8	TGTAAAACGACGGCCAGTAGTAAGTTTCATTCTAAAATCAGAGG
DMDEX79-R8	CAGGAAACAGCTATGACCGTGTTTTCACTGTCTTTCTGGA

Chapter **3**

Experiences with array-based sequence capture; toward clinical applications

Rowida Almomani, Jaap van der Heijden, Yavuz Ariyurek, Yuching Lai, Egbert Bakker, Michiel van Galen, Martijn H Breuning and Johan T den Dunnen

Eur J Hum Genet. 2011; 19: 50–55.

Abstract

Although sequencing of a human genome gradually becomes an option, zooming in on the region of interest remains attractive and cost saving. We performed array-based sequence capture using 385K Roche NimbleGen, Inc. arrays to zoom in on the protein-coding and immediate intron-flanking sequences of 112 genes, potentially involved in mental retardation and congenital malformation. Captured material was sequenced using Illumina technology. A data analysis pipeline was built that detects sequence variants, positions them in relation to the gene, checks for presence in databases (eg, db single-nucleotide polymorphism (SNP)) and predicts the potential consequences at the level of RNA splicing and protein translation. In the samples analyzed, all known variants were reliably detected, including pathogenic variants from control cases and SNPs derived from array experiments. Although overall coverage varied considerably, it was reproducible per region and facilitated the detection of large deletions and duplications (copy number variations), including a partial deletion in the *B3GALTL* gene from a patient sample. For ultimate diagnostic application, overall results need to be improved. Future arrays should contain probes from both DNA strands, and to obtain a more even coverage, one could add fewer probes from densely and more probes from sparsely covered regions.

Introduction

For many years, the amplification of target sequences by PCR, followed by Sanger sequencing, has been the gold standard for screening of variants in terms of both read length and accuracy of sequencing.[1] However, when it comes to conditions with highly heterogeneous etiology, a large number of different genes need to be screened for mutations. In such cases, gathering information becomes laborious, expensive and time-consuming. There are many examples of diseases that can be caused by mutations in many different genes, including mental retardation (MR),[2] Charcot–Marie–Tooth disease,[3] cardiomyopathy,[4] retinitis pigmentosa,[5] autism,[6] hearing loss[7] and congenital disorders of glycosylation.[8] Extensive resequencing of many disease-associated genes is required to explore, at the sequence and structural level, the genomic variation that might be involved in causing such diseases.

Several next-generation sequencing (NGS) platforms are now available and they have allowed the sequencing and analysis of large numbers of genes in one experiment,[9, 10, 11] and are able to

generate a massive amount of sequence data and have considerably reduced the cost of DNA sequencing.[12] However, although NGS platforms have enormously increased throughput and have permitted whole-genome sequencing, high cost still prevents routine whole human genome resequencing projects. Therefore, zooming in on the region of interest is an attractive option. In addition, it circumvents the problem of identifying variants in genes for which the analyses were not intended (with associated ethical problems).

Microarray-based genomic selection combined with massively parallel high-throughput sequencing is the method of choice to analyze large numbers of genes in a more comprehensive and cost-effective manner.[13, 14, 15] We have used custom high-density microarrays (Roche NimbleGen, Inc., Madison, WI, USA) for the enrichment of 112 distinct genes potentially involved in MR and congenital malformation, followed by sequencing on the Illumina Genome Analyzer I platform (Illumina, San Diego, CA, USA).

The first aim of our study was to apply and validate the array-based enrichment method as an efficient and convenient strategy to capture any desired portion of the human genome. The second aim was to accelerate the detection of sequence and copy number variations (CNV) in the selected candidate genes with lower costs, especially for the genes that are potentially involved in MR.

Materials and methods

Sample selection and validation

Six DNA samples were used in this study, including two controls containing known pathogenic variants. Sample S-2 contains a known *MECP2* (OMIM 300005) pathogenic point mutation (c.538C>T); the second sample, patient S-6, carries a large deletion spanning exons 8–15 in one allele and a splice site mutation (c.660+1G) at the other allele of the *B3GALTL* (OMIM 610308) gene.

The other four DNA samples were from patients with MR with an unknown cause. Single-nucleotide polymorphism (SNP) array data were available for two samples: S-7 with 250K Nsp Affymetrix and S-5 with 317K Illumina data. We used these data to validate the sequences

obtained after capture-array and Illumina sequencing. Causative large deletions and duplications had been previously excluded by SNP array testing in S-3, S-5, S-7 and S-8.

Exon array design

Microarrays with 385K probe capacity (Roche NimbleGen, Inc.) were used to capture all exons, the splice site and the immediately adjacent intron sequence of 112 human genes. On the basis of searches in OMIM and literature, we selected 112 human genes known to cause MR, either as part of a known syndrome or in isolation (Supplementary Table 1). Primary sequence data from all exons were extracted from NCBI's genome (Build 36). Microarrays were designed by Roche NimbleGen, Inc. with long oligonucleotide probes (54–99 nucleotides) that span each target region, overlapped and shifted on an average of seven bases.[13] The oligonucleotides were designed to achieve isothermal hybridization across the arrays capturing one strand only. All highly repetitive regions were excluded from the probe selection in order to avoid nonspecific capturing of genomic regions. Using all criteria listed, for 2% of the target sequences, no capture probe could be designed (note that, theoretically, these sequences can be covered partly through capture from directly flanking unique sequences). Four of the arrays were reused at least twice.

Genomic DNA library preparation and target capture

The methods used for target capture, enrichments and elution followed previously described protocols with slight modifications (Roche NimbleGen, Inc.).[16] Genomic DNA (20–10 μg) was fragmented using a nebulizer or Bioruptor according to instructions from the manufacturer to yield fragments from 250–1000 bp (nebulization) or 250–600 bp (Bioruptor). Adapter oligonucleotides from Illumina (single reads) were ligated to the ends. After the ligation was completed, successful adapter ligation was confirmed by PCR. The DNA-adapter ligated fragments were then hybridized to the sequence capture microarray for 65 h. After hybridization and washing, the DNA fragments bound to the array were eluted, using 300 μl of the elution buffer (Qiagen, Valencia, CA, USA) on each array. A gasket (Agilent) was applied and placed on the thermal elution device (homemade) for 20 min at 95°C. We repeated this process once by adding 200 μl of elution buffer (Qiagen). DNA from each eluted sample was enriched by 18-cycle PCR using a high-fidelity polymerase and a single primer pair corresponding to the Illumina adapters ligated earlier.

Check enrichments by qPCR

To verify successful hybridization capture, we performed qPCR (quantitative PCR) on DNA samples (S-2, S-3, S-5, S-7, S-6 and S-8) before and after array enrichment. The primers amplified five loci from *MBL2*, *DMD* and *BRCA1* (100 bp) as negative controls (no capture probes on the array) and four loci from *MECP2*, *CREBBP* and *NSD1* genes as positive controls (capture probes on the array) (Supplementary Table 2). All primers for qPCR were designed using Primer 3 (http://frodo.wi.mit.edu/).

The qPCR assays were performed in triplicate in the Lightcycler using 384-well plates (Roche NimbleGen, Inc.) in 10 μl total volume: 5 μl of 2 × SYBR Green master Rox (Roche NimbleGen, Inc.), 0.25 μl of each primer (10 pmol/μl), 2 μl of DNA template and 2.5 μl of ultrapure water. The thermo-cycling protocol was carried out as follows: 10 min at 95°C, 45 cycles of 10 s at 95°C, 30 s at 60°C, 20 s at 72°C and 5 min at 72°C, followed by melting curve analysis in order to determine the specific and nonspecific amplified products and other artifacts that might interfere with CP values. To calculate the relative fold enrichment of the targeted regions, we compared amplification of the positive *versus* negative controls. The relative fold enrichment, R, was calculated using the values of ΔCP (ie, the difference between average CP of non-captured and average CP of captured samples) according to $R=E^N$, where E is the efficiency of the qPCR assay for a particular amplicon and N=ΔCP (crossing point).

DNA Sequencing

The eluted enriched DNA fragments were sequenced using the Illumina GAI platform at the Leiden Genome Technology Center (LGTC). Single-end sequencing of 36 or 50 nucleotides was performed following the instructions of the manufacturer.

Reads mapping and data analysis

Sequence read mapping was carried out by ELAND and ELAND-extended programs, which were a part of the Illumina GAI data analysis package. Only reads of high-quality scores were mapped to the human reference genome (NCBI, BUILD 36.2), allowing up to two mismatches. We created different Perl scripts to extract and process data from the ELAND files. Coverage was calculated at the target level (gene–exons), the nucleotide level and at the per probe region.

SNP calling was performed by searching for nucleotides discordant with the reference genome with a base call quality score of 30 (99.9% base call accuracy), a read depth of 8 or greater and the variant allele larger than 30% of the total coverage. Thereafter, all variants were checked for their presence in known databases, for example, dbSNP. Perl scripts were designed to predict the potential consequences at the level of RNA splicing and protein translation on the basis of Ensemble v.51. Furthermore, we designed a Perl script to facilitate detection of small deletions/insertions (up to three nucleotides). All Perl scripts are available on request.

Sanger sequencing

A total of 21 variants detected by Illumina GAI analyzer were selected and confirmed by Sanger sequencing using the standard Sanger sequencing protocol at the Leiden Genome Technology Center (LGTC). The primer sequences (with M13 tail) used are shown in Supplementary Table 3.

Results

The methodology used starts with fragmentation of the genomic DNA. Linker and primer addition can then be performed either before or after array-capture target enrichment. To facilitate limited amplification of the expected low-yield array elution, we decided to perform full Illumina sample preparation before array capture. Initially, experiments were conducted using 20 μg genomic DNA, later we reduced this to 10 μg. We used qPCR, comparing targeted (four positive controls) and non-targeted regions (five negative controls), to check successful array enrichment and to estimate the fold enrichment obtained (see Supplementary Tables 4 and 5 for examples). As enrichment varies significantly from locus to locus, we tested multiple loci to obtain an accurate estimate. Samples in which qPCR did not indicate clear enrichment (>100 ×) were discarded. The ultimate enrichments achieved varied from experiment to experiment with a tendency to increase over time, indicating that lab experience is an important aspect of the array capture technology. As the fold enrichments determined by qPCR correlate positively with

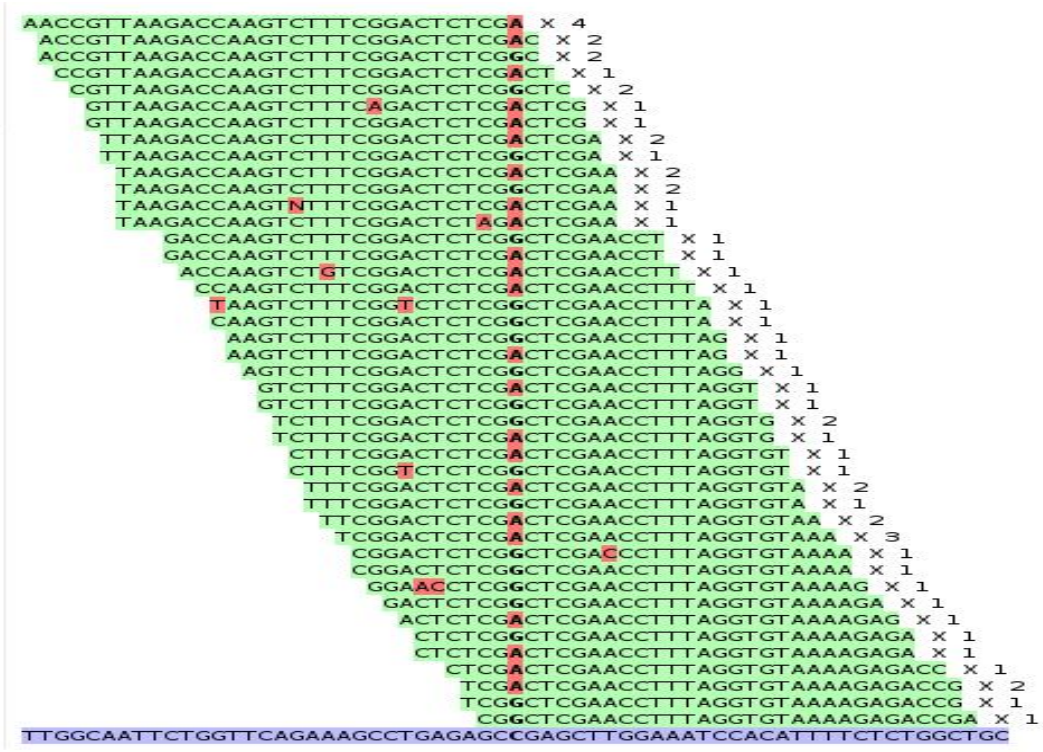


Figure 1 Detection of sequence variants. A total of 32 nucleotide NGS reads (top, sequence mismatches in red) aligned with the genomic reference sequence (bottom). The center of the alignment shows a variant present in the heterozygous state. ' × *n*' behind the read indicates how many identical reads were obtained.

the average sequence depth obtained, we conclude that qPCR provides an effective and cost-saving check for successful enrichment (examples are listed in Supplementary Tables 4 and 5).

Sequence data

The custom arrays used contained 112 different human genes that are known to be or potentially involved in MR and congenital malformation. Samples were run on one channel of the Illumina GAI. For sequence analysis, we used only those QC-filtered reads that map back uniquely to the reference sequence (M0) or with one or two mismatches (M1, M2) (Figure 1). Using these settings, 85–92% of the targeted nucleotides were covered by at least eight reads (Table 1) and 94–98% by at least one read (note that for 2% of the targeted sequences, no probe could be designed, see M&M). Effectively, this means that for 78% of the targeted sequences on the array, coverage was sufficient (>20 ×) to detect any variants that were present.

Table 1 Sequence summary results of the different array-capture experiments performed
Abbreviations: F, female; M, male; MM# reads, number of reads with # mismatches to the reference sequence; QC, quality control.

Sample ID, sex	Total reads × 10^3	Reads passing QC filter × 10^3	Total number of reads mapped × 10^3	MM0 reads × 10^3	MM1 reads × 10^3	MM2 reads × 10^3	Coverage per nucleotide	% of Nucleotides were covered ≥8 times	% of nucleotides were covered 0 times	Read length	Array reused
S-2, F	6.744	4.804	2.428	1.359	691	378	138	87.11	6.22	50	No
S-3, M	7.305	5.354	2.176	1.225	618	333	100	90.71	4.49	50	No
S-5, M	10.43	7.237	5.576	4.935	499	142	120	92.42	2.09	32	No
S-7, M	15.771	6.112	4.719	3.885	638	196	100	91.13	2.7	32	No
S-6, M	12.154	6.575	6.575	5.914	486	174	99	99.24	7.08	32	Yes, 2nd time
S-8, F	11.077	3.531	3.531	2.301	736	485	44	85.38	4.43	49	Yes, 3rd time

Two of the samples had been previously analyzed using SNP arrays. The region selected using the capture array included 67 different SNPs that had been present on the SNP arrays. We observed a perfect agreement (100%) between array-based SNP calls and those obtained using NGS (67/67 variants) (Supplementary Table 6).

To determine our ability to detect pathogenic mutations, we included one sample from a female patient (S-2) harboring a dominant pathogenic point mutation in the *MECP2* gene, (c.538C>T) on the X chromosome. Our results clearly detected the change in the heterozygous state (Supplementary Table 7). Similarly, we detected a homozygous change in the *B3GALTL* gene in a Peter's Plus patient (c.660+1G>A, Supplementary Table 7, see below).

We next selected 21 variants detected in samples S-2, S-3, S-5, S-7 and S-8 and checked these by traditional Sanger sequencing. We were able to confirm 21 of the 21 variants, including their status being homozygous or heterozygous (Supplementary Table 7). The analysis of the variants found in all 112 genes of the patients did not reveal a clear cause of their MR Supplementary Table 8 and 9.

CNV

Changes that cannot be easily detected using the sequence itself include deletions and duplications (CNVs). However, such variants can be expected to yield quantitative changes in coverage. To determine whether overall coverage can be used to detect quantitative changes, we first analyzed the 39 genes located on the X chromosome. Indeed, when coverage was normalized using autosomal genes (Figure 2a), samples from females showed a clearly higher X-chromosome coverage compared with male samples (Figure 2b). Furthermore, as expected, the gene on the Y chromosome (*NLGN4Y*) gave no coverage in the female sample (Figure 2b). To determine the sensitivity of our method for detecting smaller CNVs, we carefully analyzed a sample from a compound heterozygous patient (S-6) carrying a partial deletion (exons 8–15) and a splice site mutation (c.660+1G>A, intron 8) in the *B3GALTL* gene. The splice site mutation was evident as no wild-type sequence was present. The presence of a deletion emerged as, compared with other samples, we observed a significantly lower average coverage for the *B3GALTL* gene (53 × *versus* 155 ×, 150 ×, 140 ×) (Figure 2c). In addition, although the splice site mutation in exon 8 was detected in the 'homozygous' state (similar to all nine variants downstream), we observed variants in the first exons (1–7) also in heterozygous state (Supplementary Table 10). These data show that not only have we obtained an excellent specificity of the capture process but we have also been able to distinguish between male and female samples.

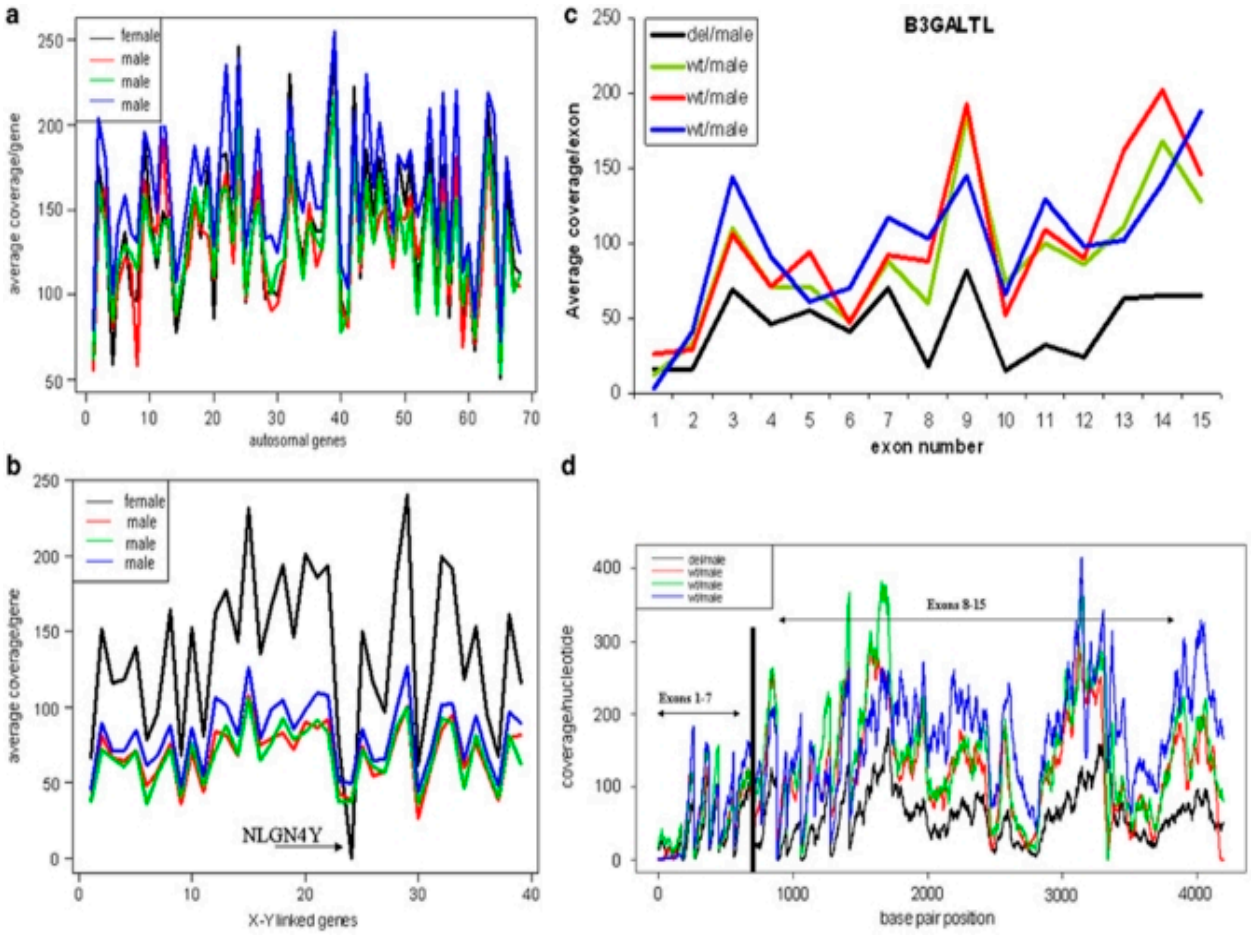


Figure 2 Average coverage obtained for different genes in four different samples. (**a**) Shows average coverage of 69 autosomal genes from four different samples. (**b**) Shows average coverage of 39 genes located on X and one gene (*NLGN4Y*) located on the Y chromosome; a female sample exhibited an absence of hybridization on the captured array, with no coverage in the regions corresponding to the *NLGN4Y*. The female sample shows a higher average coverage per gene for all genes located on X-chromosome compared with male samples. (**c**) Lower average coverage of *B3GALTL* gene in a male patient sample with a known large deletion compared with three wild-type male samples. (**d**) Coverage per nucleotide/position for the whole *B3GALTL* gene: the patient sample shows lower coverage for the second half (exons 8–15) compared with wild type samples. del=deletion, wt=wild type.

Discussion

Array-based genomic selection offers several advantages for large-scale targeted DNA isolation over other approaches such as PCR-based methods (long-range PCR or multiplexed short PCR),[17, 18, 19] selector technology[20, 21] and BACs technology.[22] PCR-based methods become

laborious, time-consuming and costly if hundreds to thousands of regions (exons) need to be amplified, especially if all the sequences are required. Furthermore, when PCRs are multiplexed, it becomes difficult to check successful amplification per fragment, the chance of obtaining artifacts increases and equimolar loading before sequencing becomes very difficult. New approaches for massive individual PCR have been introduced recently[23] but experiences with these are still limiting. Selector technology[20, 21] seems attractive but it largely depends on proper in-house probe design, and experience thus far is very limited. Successful genomic selection using BACs has been demonstrated but has several limitations. As a BAC is the unit of selection, multiple BACs are required to isolate discontinuous regions of interest.

In this study, we have tested array-based sequence capture to determine the sequence of 112 genes potentially involved in MR. We show that array-based sequence capture technology is an efficient, quick and reliable method for the parallel sequencing of a range of genes of interest. Known variants (array-based calls) for 67 SNPs matched perfectly with those obtained using NGS Supplementary Table 6. Two positive controls with known pathogenic changes in the *MECP2* gene (sample S-2) and *B3GALTL* gene (sample S-6) were readily detected. In addition, 21/21 selected variants found in the five samples analyzed could be confirmed using Sanger sequencing (Supplementary Table 7). Sequence coverage of the nucleotide of interest is critical for reliably detecting sequence changes. If coverage is too low, both false positives (caused by sequence errors) and false negatives (if only one allele from a heterozygous sample is observed) will occur.

The coverage we obtained differs significantly not only between targeted genomic regions (genes) but also between different samples (Supplementary Table 1, Figure 2a). As the overall methodology is rather complex, particularly the collection of the hybridized array-enriched DNA sequences, the difference between samples is most probably influenced by technical factors such as variations in hybridization, washing conditions and potential reuse of the capture array. Furthermore, coverage is influenced by array design, including probe sequence (melting temperature, GC content), probe density and spacing (Supplementary Table 1). Our data show that AT-rich regions (>55%), regions with an overall low probe density (<3) and small exons (on average 90 bp) yield a low coverage, which also varies significantly between experiments. For a second-generation capture array, the results obtained could be used to change the probe density,

that is, decreased in well-covered and increased in low-covered regions. Our data show that longer reads (50 bp) improve accuracy and selectivity of read mapping to the reference genome, which influenced the SNP calling by having less false positives and slightly better coverage.

As CNVs (deletions/duplications) are a significant cause in the etiology of MR,[24] we tested the feasibility of detecting large CNVs using array capture and NGS. Our results indicate that, if coverage is sufficiently high, array capture can also be used to detect such quantitative changes. Our array contained one gene from the Y chromosome that gave no coverage in females (Figure 2b), whereas the 39 X-linked genes when compared with the 69 autosomal genes yielded overall 50% lower coverage in male samples (Figure 2b). Another example derives from a sample containing a partial *B3GALTL* gene deletion on one allele (exons 8–15) and a splice site mutation on the other allele (c.660+1G>A). Although coverage over the entire gene seems reduced (experimental variation/coincidence), coverage for the second half of the gene clearly drops below that of normal (Figure 2d). An algorithm for detecting local deviations from the average coverage is currently under development.

Regarding probe design (performed by Roche NimbleGen, Inc.), it should be noted that all array probes are from one strand (coding DNA strand) and thus DNA molecules from only the non-coding strand are captured. This has several consequences. First, the sequence obtained is from one strand only, whereas for diagnostic applications, quality assurance requires that sequences be obtained in forward and reverse orientation. Sequencing this one strand in both directions is partly fooling oneself. Second, we observed that the sequences obtained relative to the array probes extend in a 5′ but not in a 3′ direction. The most probable cause for the latter is steric hindrance during array hybridization, preventing non-hybridizing tails at the surface side of the array. When capture probes are attached with their 3′ ends, this has consequences for probe design at the edges of the targeted regions; on the 5′ side, coverage will be significantly better than on the 3′ side. Both effects could be overcome simply by reversing the probe sequence of every other nucleotide on the array. Theoretically, this would also mean that the overall yield of enriched DNA would double, as both strands from the sample will be captured.

To save costs, we have reused the arrays up to three times by hybridizing different samples. The danger of this approach is of course contamination, if hybridized DNA from a previous

experiment is not eluted completely. Indeed, in some experiments, we observed low-level contamination, for example, through heterozygous calls from X-chromosome sequences in male samples. It should be noted, however that cross-contamination can be easily controlled when samples containing differently tagged linkers are used in subsequent experiments.

Using the current design, low coverage was obtained mainly at the edges of the regions targeted, especially the 3′ side (see above), that is, direct gene flanking or intronic regions. Although coverage varied widely, 78% of all regions targeted and present on the array were covered effectively by the sequence obtained. Note that there is a clear correlation between fragment size of the genomic DNA used and the coverage, the larger the fragment size used the lower the target coverage achieved, as more flanking DNA is captured. Especially for array-based capture, because of the steric hindrance described, this effect will be significant near the array-attached end of a probe-targeted region. Assuming that second-generation capture arrays will be more effective (ie, complete and with even coverage) and sequence power will improve further, it should soon be possible to sequence-tag, mix and simultaneously analyze different samples in one experiment, giving a significant cost reduction.

Recently in-solution capture was presented as an alternative to array-based capture.[25] Besides advantages of simplicity, a reduced workload and a potential for automation, when attempted, in-solution capture will not show the effect of steric hindrance we observed. However, capturing both strands would be complicated by the fact that capture probes will hybridize with each other. Initial experiences in our lab with in-solution capture were successful and for future projects we will change to this approach.

Overall, we conclude that array-based sequence capture followed by NGS offers a versatile tool for successfully selecting sequences of interest from a total human genome. The approach will be especially helpful in speeding up the identification of the pathogenic mutation(s) in diseases in which the genomic region to be scanned is large. Our results indicate that the methodology can still be improved, in particular, with respect to probe design, obtaining a more even coverage of the targeted regions. On the basis of initial experiences and publications, we expect that array capture will be quickly replaced by in-solution capture. Ultimately, the cost of this approach is

determined by the minimal coverage, which in turn determines the sensitivity required for the detection of potential sequence variants.

Acknowledgments

We thank the Leiden Genome Technology Center (LGTC), in particular Sophie Greve-Onderwater, Matthew Hestand and Rolf Vossen, for their expert technical assistance; Antoinette Gijsbers for sharing the SNP data; and Kamlesh Madan for critical reading of the paper. 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) and 223143 (the TechGene).

References

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 1977;74:5463–5467.
2. Chelly J, Khelfaoui M, Francis F, Chérif B, Bienvenu T. Genetics and pathophysiology of mental retardation. Eur J Hum Genet. 2006;14:701–713.
3. Szigeti K, Lupski JR. Charcot-Marie-Tooth disease. Eur J Hum Genet. 2009;17:703–710.
4. Paul M, Zumhagen S, Stallmeyer B, Koopmann M, Spieker T, Schulze-Bahr E. Genes causing inherited forms of cardiomyopathies. A current compendium. Herz. 2009;34:98–109.
5. Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. Lancet. 2006;368:1795–1809.
6. Muhle R, Trentacoste SV, Rapin I. The genetics of autism. Pediatrics. 2004;113:472–486.
7. Hilgert N, Smith RJ, Van Camp G. Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics. Mutat Res. 2009;681:189–196.
8. Freeze H. Genetic defects in the human glycome. Nat Rev Genet. 2006;7:537–551.
9. Bonetta L. Genome sequencing in the fast lane. Nat Methods. 2006;3:141–147.
10. von Bubnoff A. Next-generation sequencing: the race is on. Cell. 2008;132:721–723.
11. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods. 2008;5:16–18.
12. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–1145.
13. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods. 2007;4:903–905.
14. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods. 2007;4:907–909.
15. Hodges E, Xuan Z, Balija V, et al. Genome-wide *in situ* exon capture for selective resequencing. Nat Genet. 2007;39:1522–1527.
16. Roche NibleGen NimbleGen services user's guides: sequence capture service . http://www.nimblegen.com/products/lit/SeqCap_UsersGuide_Service_v3p0.pdf .
17. Edwards MC, Gibbs RA. Multiplex PCR: advantages, development, and applications. PCR Methods Appl. 1994;3:S65–S75
18. Markoulatos P, Siafakas N, Moncany M. Multiplex polymerase chain reaction: a practical approach. J Clin Lab Anal. 2002;16:47–51.
19. Cutler DJ, Zwick ME, Carrasquillo MM, et al. High-throughput variation detection and genotyping using microarrays. Genome Res. 2001;11:1913–1925.
20. Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res. 2005;33:71.
21. Dahl F, Stenberg J, Fredriksson S, et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci USA. 2007;104:9387–9392.
22. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. Direct genomic selection. Nat Methods. 2005;2:63–69.

23. Tewhey R, Warner JB, Nakano M, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol. 2009;27:1025–1031.

24. Shaw-Smith C, Redon R, Rickman L, et al. Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. J Med Genet. 2004;41:241–248.

25. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27:182–189.

Supplementary data:

Supplementary Table 1 Average coverage for all selected genes for six different samples. S-2, S-8, are female samples while the rest are male samples. All genes with bold have low GC or /and low probe density. Columns in bold represent data from re-used arrays.

Gene Name	Chromosome location	**S_6**	S_2	S_3	S_5	S_7	**S_8**
ACSL4	X	**37.58**	66.24	37.68	45.04	55.24	**32.2**
AFF2	X	**81.6**	152.41	72.51	89.61	92.34	**57.94**
AGTR2	X	**66.93**	115.49	66.91	71.58	96.6	**48.06**
ALG1	16	**55.84**	61.43	62.68	79.63	46.89	**21.58**
ALG12	22	**147.22**	176.4	164.69	203.65	85.58	**41.45**
ALG2	9	**163.81**	147.3	139.8	180.21	186.49	**54.07**
ALG6	1	**80.79**	59.33	84.49	99.93	129.98	**37.87**
ALG8	11	**105.88**	113.45	118.26	140.59	136.9	**48.35**
ALG9	11	**121.78**	137.43	129.53	158.8	143.43	**52.15**
AMMECR1	X	**64.75**	118.66	60.42	70.6	83.18	**49.71**
ARHGEF6	X	**71.53**	139.8	71.27	84.35	82.22	**56.31**
ARX	X	**47.8**	78.39	35.73	61.96	20.88	**23.93**
ASPM	1	**117.55**	100.77	124.26	136.38	180.53	**52.79**
B3GALTL	13	**58.33**	96.4	114.61	130.66	155.11	**40.4**
B4GALT1	9	**167.08**	193.19	159.43	195.85	144.86	**52.82**
BBS1	11	**142.78**	153.75	138.72	183.65	107.54	**41.16**
BBS10	12	**136.27**	116.27	121.91	148.89	186.85	**49.54**
BBS12	4	**191.59**	149.38	141.62	216.64	235.1	**71.96**
BBS2	16	**135.96**	139.07	143.86	167.26	155.66	**52.06**
BBS7	4	**87.54**	78.35	88.24	107.63	132.32	**40.22**
BRWD3	X	**57.78**	95.96	56.43	68.09	72.16	**47.02**
CA2	8	**111**	102.69	108.39	133.41	126	**42.57**
CC2D1A	19	**117.69**	140.17	130.99	158.96	66.25	**29.97**

CDK5RAP2	9	**153.43**	162.19	163.23	187.89	146.64	**58.83**
CDKL5	X	**75.99**	165.03	73.1	88.4	70.16	**54.41**
CENPJ	13	**138.59**	133.45	141.94	164.93	176.59	**54.91**
COG7	16	**135.39**	176.24	163.72	186.49	111.66	**51.73**
CRBN	3	**108.94**	86.16	109.99	128.55	152.04	**38.71**
CREBBP	16	**153.08**	180.96	159.75	182.66	109.2	**48.82**
CUL4B	X	**36.41**	68.41	40.24	47.46	49.6	**34.64**
DHCR7	11	**171.56**	183.15	163.11	234.75	106.68	**50.25**
DLG3	X	**69.9**	153.26	75.57	86.07	57.36	**47.41**
DNMT3B	20	**118.41**	129.94	130.4	156.44	95.96	**36.4**
DPAGT1	11	**193.16**	245.7	207.64	239.34	156.49	**63.91**
DPM1	20	**97.19**	95.5	99.55	127.71	125.07	**45.49**
DYRK1A	21	**138.62**	140.82	139.63	158.83	171.06	**55.53**
EP300	22	**172.71**	171.73	155.09	197	138.73	**49.63**
ERCC8	5	**103.87**	97.76	117.7	132.83	146.93	**48**
FGFR3	4	**91.04**	104.19	100.77	134.5	53.29	**25.78**
FKTN	9	**94.39**	99.05	116.64	125.17	128.33	**43.87**
FMR1	X	**43.64**	80.63	48.33	54.48	68.83	**42.42**
FOXP2	7	**116.96**	118.59	120.86	150.1	134.51	**46.95**
FTSJ1	X	**84.3**	162.65	71.18	105.87	67.19	**50.11**
GDI1	X	**81.34**	177.75	88.1	102.05	52.92	**47.42**
GLI3	7	**170.51**	229.96	193.16	214.24	119.51	**57.33**
GRIA3	X	**68.43**	143.32	68.85	81.17	76.74	**56.3**
GRIK2	6	**141.74**	140.26	138.97	164.42	157.43	**59.8**
HRAS	11	**110.45**	119.66	109.32	148.37	50.77	**26.55**
HSD17B10	X	**108.3**	232.65	105.43	126.58	76.49	**58.98**
IL1RAPL1	X	**75.14**	135.02	64.74	79.09	81.97	**56.15**
JAG1	20	**153.84**	149.33	143.06	177.83	134.17	**48.77**

JARID1C	X	**78.79**	166.5	75.9	98.34	55.94	**40.82**
KRAS	12	**37.5**	27.35	38.21	55.47	74.39	**14.28**
L1CAM	X	**83.47**	194.58	92.76	104.82	46.23	**44.07**
MAOA	X	**72.42**	146.33	78.53	85.28	89.76	**57.2**
MCPH1	8	**116.95**	137.58	132.3	150.93	134.11	**50.26**
MECP2	X	**90.3**	202.17	83.87	98.16	70.25	**55.29**
MED12	X	**87.01**	185.32	90.93	109.51	70.96	**53.62**
MGAT2	14	**128.29**	137.63	126.84	151.21	129.3	**43.37**
MPDU1	17	**191.91**	201.65	178.4	210.59	143.1	**54.87**
MPI	15	**208.85**	238.66	216.49	254.72	147.26	**62.87**
MYCN	2	**94.44**	97.38	78.66	117.71	65.2	**25.33**
NF1	17	**81.22**	82.78	89.34	103.68	103.15	**35.33**
NLGN3	X	**92.27**	193.94	83.89	107.85	60.81	**48.03**
NLGN4X	X	**43.01**	76.28	37.59	50.9	54.92	**28.06**
NLGN4Y	Y	**38.17**	0	37.83	49.38	59.67	**0**
NSD1	5	**184.55**	222.02	187.13	208.31	179.62	**72.56**
NUFIP1	13	**7.36**	7.19	11.08	18.4	21.56	**3.08**
OPHN1	X	**74.32**	150.2	72.16	85.18	77.89	**54.7**
PAFAH1B1	17	**113.96**	109.42	118.84	136.76	160.18	**42.43**
PAFAH1B3	19	**162.95**	185.67	166.94	229.95	92.43	**43.41**
PAK3	X	**54.5**	115.72	60.31	64.4	64.17	**46.98**
PHF6	X	**57.49**	96.99	57.08	65.4	87.77	**47.64**
PHF8	X	**83.55**	188.65	87.66	98.96	67.6	**57.96**
PMM2	16	**130.66**	146.18	130.15	162.48	104.29	**40.29**
POMT1	9	**146.51**	180.39	171.95	201.25	108.67	**47.81**
PQBP1	X	**100.71**	240.55	99.96	127.36	56.04	**52.63**
PRPS1	X	**26.32**	61.85	35.36	44.24	38.85	**23.34**
PRSS12	4	**152.22**	154.55	139.87	171.21	143.31	**52.15**

PTPN11	12	**10.42**	9.97	13.07	20.86	18.3	**3.7**
RAB3GAP2	1	**116**	116.02	120.94	136.63	151.15	**47.11**
RAF1	3	**147.08**	180.38	158.42	182.34	130.38	**54.88**
RAI1	17	**144.02**	152.31	124.16	173.36	70.74	**31.93**
REST	4	**158.76**	178.87	150.5	185	161.32	**55.18**
RNF135	17	**121.02**	101.12	88.77	145.5	117.24	**43.3**
RPS6KA3	X	**60.25**	117.99	60.92	70.45	83.05	**47.97**
SATB2	2	**139.83**	150.23	142.16	164.9	156.18	**54.62**
SCN8A	12	**170.77**	187.94	171.23	209.05	162.32	**63.16**
SHANK3	22	**92.14**	96.03	87.73	128.76	40.77	**23.37**
SHROOM4	X	**84.92**	199.93	92.8	102.24	73.7	**59.54**
SIL1	5	**167.93**	176.09	153.93	219	117.45	**52.09**
SLC16A2	X	**95.29**	191.86	89.71	102.77	75.19	**56.68**
SLC35A1	6	**94.52**	86.78	105.94	118.68	130.38	**42.01**
SLC35C1	11	**181.41**	178.83	157.21	220.06	118.75	**46.87**
SLC6A8	X	**6.11**	11.01	6.31	14.48	6.55	**2.61**
SNRPN	15	**69.64**	85.84	93.84	113.43	89.68	**31.19**
SOS1	2	**113.1**	114.37	125.16	130.21	145.75	**48.56**
SOX3	X	**59.69**	118.15	46.04	64.87	29.19	**31.19**
SUZ12	17	**71.62**	67.57	74.81	86.36	123.04	**34.85**
TCF4	18	**107.87**	120.1	130.09	145.84	113.73	**43.27**
TSC1	9	**191.94**	211.41	191.59	218.62	185.63	**65.81**
TSC2	16	**139.23**	161.14	155.1	205.55	78.3	**40.38**
TSPAN7	X	**76.43**	153.85	81.13	94.13	82.21	**52.58**
UBE2A	X	**58.21**	96.51	60.21	68.49	87.91	**44.96**
UBE3A	15	**57.82**	51	53.43	72.9	89.32	**23.11**
UPF3B	X	**38.43**	67.6	39.57	48.28	52.12	**31.34**
WHSC1	4	**149.14**	168.67	150.08	181.19	130.69	**51.42**

WHSC2	4	**107.28**	117.25	102.2	142.55	57.45	**28.88**
ZFHX1B	2	**105.2**	112.69	110.64	124.93	125.11	**39.93**
ZNF41	X	**80.14**	161.3	80.93	96.93	91.05	**55.53**
ZNF674	X	**3.54**	11.65	9.41	11.36	8.58	**3.15**
ZNF81	X	**80.94**	115.7	62.7	89.15	98.13	**59.43**

Supplementary Table 2 Primer sequences used for quantitative PCR (qPCR) to determine successful target enrichment. Genes in bold have capture probes on the array (positive controls), the others are negative controls.

Gene name	Target	Forward primer	Reverse primer
MECP2	Exon 1	CACCAGTTCCTGCTTTGATGT	CCCTAACATCCCAGCTACCAT
CREBBP	Exon 4	CACAAGTCCATTTGGACAGC	GTTGACCATGCTCTGTTTGC
CREBBP	Exon 5	CAGTGGGAATTGTACCCACAC	GAGCATGAAGCAGTAGAACCAG
NSD1	Exon 7	GTGAAGAGGAAAGCCTTCTAGC	AGAACTGGAGGCTCTTCTTTGG
MBL2	Exon 1	CCTGTTTCCATCACTCCCTCT	CACTGCAGGGCAGGTCTTTT
MBL2	Exon 4	AAGTGAAGGCCTTGTGTGTCA	AAGGCTTCCTCCTTGATGAGAT
BRCA1	Exon 10	CCCTTTGAGAGTGGAAGTGACA	CTGGGCTCCATTTAGACCTGA
DMD	Exon 20	TGCCAGTTGCTAAGTGAGAGAC	GCAGTAGTTGTCATCTGCTCCA
DMD	Exon 51	GGAAACTGCCATCTCCAAAC	CCAGTCGGTAAGTTCTGTCCA

Supplementary Table 3 Primer sequences (with M13 tail) used for amplification of targets for Sanger sequencing.

Gene name	Target	Forward primer	Reverse primer
DHCR7	exon4	TGTAAAACGAcggccagctcccacacagagcctcttagg	CAGGAAACAGCTATgacccccagacaaatggaaggactac
MED12	exon 6	TGTAAAACGACGGCcagcccagttgtggttctcttcatc	CAGGAAACAGCTATGACCaggggaccctgctctaacattt
ALG6	exon 6	TGTAAAACgacggccagctgggttcattgttgtaggtactg	CAGGAAACAGCTATGACCcttttccccaaacaaacacc
SCN8A	exon 17	TGTAAAACGACggccaggtctgtcacgtgaagtccattg	CAGGAAACAGCTATGACCctgtttctaggctgggaccttac

B4GALT1	exon 2	TGTAAAACGACggccagagtcctgtggcctgctaacttct	CAGGAAACAGCTATGACCgtctgtgaaatcactccccttc
B3GALTL	exon 5	TGTAAAACGACGGCCAGagtagtcaattcatacttatc Ttttcgg	CAGGAAACAGCTATGACCtgaggaaaaccacaccctct
B3GALTL	exon 6	TGTAAAACGACGGCCAGTtgccattctgtgtacccttc	CAGGAAACAGCTATGACCtggctcattataagctctgtcc
B3GALTL	exon 9	TGTAAAACGACGGCCAGTtgtgttctgctttcccttga	CAGGAAACAGCTATGACCagaatcagcagaagcccaaa
B3GALTL	exon 10	TGTAAAACGACGGCCagtatgccggaaatatgtttggt	CAGGAAACAGCTATGACCttgaaatggtgcaatgagga
B3GALTL	exon 13	TGTAAAAcgacgggcagtcagagtgggatgtaagaacca	CAGGAAACAGCTATGACCtcccagtgccagagacctac
B3GALTL	exon 15	TGTAAAACGACGGCCAGTggtagtgaagtaaagcag tccactt	CAGGAAACAGCTATGACCaagtcaggaagcacccaatg
NSD1	exon 23	TGTAAAACGACGGCCAGAACCTCCTGCTGACA CCAAC	CAGGAAACAGCTATGACCTGGGAACTGAGGTTTT CTCC
NSD1	exon 17	TGTAAAACGACGGCCAGTagcattggtcgatttttgtg	CAGGAAACAGCTATGACCgcccggctatttctgatctt
NSD1	exon 5	TGTAAACGACGGCCAGTAACCTCGTAAGCGCA TGAAC	CAGGAAACAGCTATGACCgggaaaagggcttctggtaa
TSC1	exon 23	TGTAAAACGACGGCCAGCCTAACCCCCTCTCA TTTACCT	CAGGAAACAGCTATGACCGGGACAAAACCAGACT TACCTG
RAB3GAP2	exon 35	TGTAAAACGACGGCCAGTTACAGAGTAGCAGC ACTGGAAAG	CAGGAAACAGCTATGACCCCAAGTTTCTTTGACT AGCCTCCT
EP300	exon 31	TGTAAAACGACGGCCAGGACTCAGCACCGATA ACTCAGACT	CAGGAAACAGCTATGACCCGGCTACTGCACAGTT CTTATG
CENPJ	exon 16	TGTAAAACGACggccaggtacaaacttgctccaccctct	CAGGAAACAGCTATGACCcaggtgtcacactgagtggttt

Supplementary Table 4 qPCR fold-enrichment values for four targets from different samples.

Sample /sex	Amplicon Name	PCR efficiency	Delta-CP	qPCR fold enrichment	% Average coverage per target	Coverage per nucleotide in all targets	Coverage for all exons
S-2 (female)	MECP2 amplicon 1	1.98	9.26	558	203	128	119
S-2 (female)	CREBBP amplicon 4	1.8	9.9	336	182		
S-2 (female)	CREBBP amplicon 5	1.79	10.39	428	182		
S-2 (female)	NSD1 amplicon 7	1.84	9.51	328	222		
S-8 (female)	MECP2 amplicon 1	1.98	8.37	304	56	44	41

S-8 (female)	CREBBP amplicon 4	1.8	8.56	153	49		
S-8 (female)	CREBBP amplicon 5	1.79	8.64	153	49		
S-8 (female)	NSD1 amplicon 7	1.84	7.95	127	73		
S-5 (male)	MECP2 amplicon 1	1.98	10.23	1000	99	120	111
S-5 (male)	CREBBP amplicon 4	1.8	11.5	862	184		
S-5 (male)	CREBBP amplicon 5	1.79	11.35	741	184		
S-5 (male)	NSD1 amplicon 7	1.84	10.59	637	208		
S-6 (male)	MECP2 amplicon 1	1.98	9.82	819	91	100	92
S-6 (male)	CREBBP amplicon 4	1.8	11.2	723	154		
S-6 (male)	CREBBP amplicon 5	1.79	10.5	452	154		
S-6 (male)	NSD1 amplicon 7	1.84	9.87	411	184		

Supplementary Table 5 CP values for non-targeted regions (negative controls) before and after array enrichment.

Sample ID	CP value (negative targets) before capture	CP value (negative targets) after capture	Delta-CP value
	MBL2 amplicon 1	**MBL2 amplicon 1**	
S-2	26	39	-13
S-8	26	38	-12
S-5	28	31	-3
S-6	26	31	-5
	MBL2 amplicon 4	**MBL2 amplicon 4**	
S-2	25	36	-11
S-8	26	27	-1
S-5	26	29	-3
S-6	25	37	-12
	BRCA1 amplicon 10	**BRCA1 amplicon 10**	
S-2	26	36	-10

S-8	26	40	-14
S-5	26	40	-14
S-6	27	40	-13
	DMD amplicon 10	**DMD amplicon 10**	
S-2	25	40	-15
S-8	26	39	-13
S-5	27	32	-5
S-6	26	31	-5
	DMD amplicon 20	**DMD amplicon 20**	
S-2	25	40	-15
S-8	26	28	-2
S-5	27	32	-5
S-6	28	40	-12

Supplementary Table 6 Single nucleotide polymorphism (SNPs) detected by Illumina sequencing and confirmed by SNP array data from two patients.

Sample id	Chromosome position	Gene name	SNP ID	Location	Ref. Sequence	Genotype array	Genotype Illumina	Sequence depth	Wild type	Mutant
S-7	chr22_48683439	ALG12	rs1321	exon	A	AG	AG	93	45	48
S-7	chrX_147856958	AFF2	rs16994895	Intron	T	GG	GG	95	0	95
S-7	chrX_147887801	AFF2	rs6641482	exon	A	GG	GG	66	0	66
S-7	chr4_123883877	BBS12	rs13135766	exon	G	GC	GC	156	84	72
S-7	chr13_30749059	B3GALTL	rs1409373	Intron	A	GG	GG	77	0	77
S-7	chr13_30803641	B3GALTL	rs912603	exon	G	GA	GA	153	73	80
S-7	chr3_3167927	CRBN	rs1620675	Intron	T	GG	GG	66	0	66
S-7	chr9_122348097	CDK5RAP2	rs10739564	Intron	T	CC	CC	56	0	56

S-7	chr9_122370634	CDK5RAP2	rs4837782	Intron	T	CC	CC	145	0	145
S-7	chr9_122209539	CDK5RAP2	rs2282168	Intron	C	GG	GG	59	0	59
S-7	chr22_39867180	EP300	rs6002267	Intron	G	TT	TT	206	0	206
S-7	chr22_39844101	EP300	rs4822005	Intron	G	AA	AA	16	0	16
S-7	chrX_53244068	JARID1C	rs2182285	Intron	A	GG	GG	11	0	11
S-7	chr12_25253819	KRAS	rs712	exon	A	AC	AC	27	15	12
S-7	chrX_28717690	IL1RAPL1	rs12690144	Intron	T	CC	CC	20	0	20
S-7	chrX_28717669	IL1RAPL1	rs6526807	Intron	A	GG	GG	35	0	35
S-7	chr8_6290433	MCPH1	rs2584	exon	G	GA	GA	166	72	94
S-7	chrX_70269142	MED12	rs10521349	Intron	T	CC	CC	49	0	49
S-7	chr13_44422083	NUFIP1	rs1175384	Intron	A	CC	CC	151	77	74
S-7	chr9_133371932	POMT1	rs10448341	Intron	G	AA	AA	19	0	19
S-7	chr3_12601516	RAF1	rs3729931	Intron	G	GA	GA	112	50	62
S-7	chr5_138484893	SIL1	rs11750382	Intron	G	GA	GA	100	49	51
S-7	chr5_138385110	SIL1	rs3749665	Intron	A	AG	AG	37	23	14
S-7	chr5_138414758	SIL1	rs3828600	Intron	C	CA	CA	153	74	79
S-7	chr2_199953490	SATB2	rs1348813	Intron	C	CG	CG	154	72	82
S-7	chr12_50449589	SCN8A	rs303809	Intron	G	CC	CC	108	0	108
S-7	chr12_50449515	SCN8A	rs303810	Intron	A	GG	GG	174	0	174
S-7	chr12_50469752	SCN8A	rs303816	Intron	C	TT	TT	20	0	20
S-5	chr1_195337065	ASPM	rs3762271	Exon	A	AC	AC	143	60	83
S-5	chr2_144878372	ZEB2	rs13009259	Intron	G	AA	AA	25	12	13
S-5	chr2_199845853	SATB2	rs2881208	Intron	T	CC	CC	30	0	30
S-5	chr4_57492171	REST	rs3796529	Intron	G	AG	AG	148	80	68
S-5	chr7_42054747	GLI3	rs846266	Exon	A	GG	GG	222	0	221
S-5	chr9_122211576	CDK5RAP2	rs2297454	Intron	A	AG	AG	42	24	18
S-5	chr9_133375257	POMT1	rs2296949	Exon	A	GG	GG	311	0	311
S-5	chr9_122211576	CDK5RAP2	rs2297454	Intron	T	TC	TC	42	18	24

S-5	chr9_134760121	TSC1	rs2809243	Exon	G	AA	AA	153	0	153
S-5	chr11_45789511	SLC35C1	rs1139266	Exon	G	AA	AA	133	0	133
S-5	chr11_66010661	DPP3	rs11550299	Exon	C	AC	AC	114	53	61
S-5	chr11_66028718	DPP3	rs1671063	Exon	A	GG	GG	109	0	109
S-5	chr11_66028813	DPP3	rs2305535	Exon	G	AG	AG	227	113	114
S-5	chr11_66038671	BBS1	rs2298806	Exon	G	AG	AG	210	106	104
S-5	chr11_66053939	BBS1	rs3816492	Exon	C	CT	CT	194	98	96
S-5	chr11_111229343	ALG9	rs10502151	Exon	G	AG	AG	115	57	58
S-5	chr12_25251108	KRAS	rs13096	Exon	A	AG	AG	53	27	26
S-5	chr12_50450056	SCN8A	rs303808	Intron	G	AG	AG	198	108	94
S-5	chr12_50470538	SCN8A	rs303815	Exon	A	AG	AG	115	47	68
S-5	chr13_30789746	B3GALTL	rs1041073	Exon	G	AA	AA	135	0	135
S-5	chr16_8849319	PMM2	rs2075827	Exon	A	CC	CC	210	0	210
S-5	chr17_7431901	MPDU1	rs4227	Exon	C	AA	AA	229	0	229
S-5	chr17_17637480	RAI1	rs11649804	Exon	C	CA	CA	237	130	107
S-5	chr18_51282486	TCF4	rs3760600	Intron	C	AC	AC	50	28	22
S-5	chr19_13890269	CC2D1A	rs2305776	Intron	A	AC	AC	21	11	10
S-5	chr20_10566574	JAG1	rs8708	Exon	A	GG	GG	201	0	201
S-5	chr20_10568275	JAG1	rs1051421	Exon	C	CT	CT	164	83	81
S-5	chr20_10581313	JAG1	rs6040055	Intron	A	AG	AG	206	98	108
S-5	chr20_30860197	DNMT3B	rs2424932	Exon	A	GG	GG	94	0	94
S-5	chr20_48986311	DPM1	rs2294902	Intron	A	GG	GG	63	0	63
S-5	chr22_49480384	SHANK3	rs13055562	Intron	G	AG	AG	106	45	61
S-5	chrX_5820574	NLGN4X	rs3810686	Exon	G	AA	AA	62	0	62
S-5	chrX_47212055	ZNF41	rs5905607	Exon	T	GG	GG	114	0	114
S-5	chrX_53980349	PHF8	rs7892782	Exon	T	CC	CC	45	0	45
S-5	chrX_54036020	PHF8	rs7061449	Intron	C	TT	TT	140	0	140
S-5	chrX_69590536	DLG3	rs2274309	Intron	T	CC	CC	36	0	36

S-5	chrX_69640819	DLG3	rs1044422	Exon	G	AA	AA	125	0	125
S-5	chrX_118853103	UPF3B	rs2239963	Intron	A	CC	CC	74	0	74
S-5	chrX_152945374	MECP2	rs2734647	Exon	T	CC	CC	105	0	105

Supplementary Table 7 Different variants detected by Illumina sequencing and confirmed by Sanger Sequencing

sample ID	chr. position	gene name	location	Ref. sequence	observed genotype	Change	sequence depth	wild type	variant	Sanger sequencing
S-2	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	17	0	17	AA
S-2	30789746	B3GALTL	Exon	G	AA	c.1108G>A	21	0	21	AA
S-2	30801834	B3GALTL	UTR	G	TT	c.*29G>T	94	0	94	TT
S-2	30719240	B3GALTL	Intron	C	CT	c.347+4C>T	19	5	14	CT
S-2	30801841	B3GALTL	UTR	A	GA	*36A>G	38	50	88	GA
S-2	30789743	B3GALTL	Exon	G	GA	c.1105G>A	11	3	8	GA
S-2	176571910	NSD1	Intron	C	CG	c.3796+108C>G	41	18	23	CG
S-2	176570797	NSD1	Exon	C	CG	c.2791C>G	66	28	38	CG
S-2	176653878	NSD1	Exon	G	GC	c.6903G>C	71	33	38	GC
S-2	152949971	MECP2	Exon	C	CT	c.538C>T	34	12	22	CT
S-3	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	23	0	23	AA
S-3	30719256	B3GALTL	Intron	C	GG	c.347+20C>G	15	0	15	GG
S-3	30789746	B3GALTL	Exon	G	AA	c.1108G>A	32	0	32	AA
S-3	30801834	B3GALTL	UTR	G	TT	c.*29G>T	93	0	93	TT
S-5	30719256	B3GALTL	Intron	C	CG	c.347+20C>G	113	33	80	CG
S-5	30719992	B3GALTL	Exon	T	CC	c.348T>C	13	0	13	CC
S-5	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	93	0	93	AA
S-5	30789746	B3GALTL	Exon	G	AA	c.1108G>A	135	0	135	AA
S-5	30801834	B3GALTL	UTR	G	TT	c.*29G>T	261	0	261	TT
s-5	30801841	B3GALTL	UTR	A	AG	c.*36A>G	156	33	123	GA
s-5	30719240	B3GALTL	Intron	C	CT	c.347+4C>T	107	31	76	CT
S-5	176571910	NSD1	Intron	C	GG	c.3796+108C>G	99	0	99	GG
S-6	30741415	B3GALTL	Intron	G	AA	c.660+1G>A	27	0	27	AA
S-7	30719256	B3GALTL	Intron	C	CG	c.347+20C>G	120	56	64	CG
S-7	30801841	B3GALTL	UTR	A	GA	c.*36A>G	160	83	77	GA
S-7	30719992	B3GALTL	Exon	T	CC	c.348T>C	17	0	17	CC
S-7	30746823	B3GALTL	Intron	A	AG	c.780+58A>G	218	94	124	AG
S-7	30748989	B3GALTL	Intron	G	GA	c.850+81G>A	168	84	84	AG
S-7	30789746	B3GALTL	Exon	G	GA	c.1108G>A	66	23	43	GA

S-7	30801834	B3GALTL	UTR	G	TG	c.*29G>T	131	74	57	TG
S-7	33125238	B4GALT1	Exon	C	CT	c.597C>T	67	31	36	CT
S-7	63644620	ALG6	Exon	T	CT	c.391T>C	146	85	61	CT
S-7	70832913	DHCR7	Intron	G	AG	c.99-4G>A	92	51	41	AG
S-7	70257894	MED12	Intron	A	CC	c.736-8A>C	45	0	45	CC
S-7	50449090	SCN8A	Exon	C	CT	c.3076C>T	125	57	68	CT
S-7	134761154	TSC1	UTR	T	del T	c.*289delT	52		52	c.*289delT
S-7	218389523	RAB3GAP2	UTR		insAAC	c.*866+827insAAC	44	0		c.*866+827insAAC
S-7	39904953	EP300	UTR	CAA	del CAA	*47_*49del	38	0	38	*47_*49del
S-7	24356236	CENPJ	Intron	CAA	del CAA	c.3704_15delCAA	44	0	44	c.3704_15delCAA
S-7	176653878	NSD1	Exon	G	GC	c.6903G>C	129	57	72	GC
S-7	176571910	NSD1	Intron	C	GG	c.3796+108C>G	109	0	109	GG
S-8	30748989	B3GALTL	Intron	G	AA	c.850+81G>A	9	0	9	AA
S-8	30801834	B3GALTL	UTR	G	TT	c.*29G>T	43	0	43	TT
S-8	30719240	B3GALTL	Intron	C	TT	c.347+4C>T	30	0	30	TT

Supplementary Table 8 Number of variants detected in six different samples in UTR and introns.

Patient ID	Variants in introns	Variants in untranslated region
S-2	188	77
S-3	171	64
S-5	279	107
S-7	245	76
S-6	363	103
S-8	136	81

Supplementary Table 9 All variants detected in exons in six different samples (S-2, S-3, S-5, S-6, S-7, S-8).

Chromosome position	Gene name	Variant type	Variant	SNP ID
chrX_147842900	AFF2	Silent	c.1488G>A	rs12011040
chr22_48687480	ALG12	Silent	c.885A>G	rs8135963
chr1_195337438	ASPM	Silent	c.7566A>G	rs1412640
chr1_195358160	ASPM	Silent	c.3579T>A	rs4915337
chr1_195360653	ASPM	Silent	c.3138G>A	rs6676084
chr1_195379156	ASPM	Silent	c.849C>T	rs6677082
chr1_195337330	ASPM	Silent	c.7674C>T	rs41308365
chr1_195339043	ASPM	Silent	c.5961A>G	rs41310925
chr1_195340555	ASPM	Silent	c.4449A>G	rs2878749

chr1_195375569	ASPM	Silent	c.1977T>C	rs17550662
chr1_195337399	ASPM	Silent	c.7605G>A	rs10922162
chr13_30801679	B3GALTL	Silent	c.1371A>G	
chr13_30719992	B3GALTL	Silent	c.348T>C	rs4943266
chr9_33125238	B4GALT1	Silent	c.597C>T	rs1065765
chr4_123883877	BBS12	Silent	c.1380G>C	rs13135766
chr4_123883907	BBS12	Silent	c.1410C>T	rs13135445
chr4_123884369	BBS12	Silent	c.1872A>G	rs13102440
chr4_123883559	BBS12	Silent	c.1062G>C	rs34296401
chr4_123883697	BBS12	Silent	c.1200G>A	rs309371
chr4_123883706	BBS12	Silent	c.1209G>A	rs17006092
chr4_123883895	BBS12	Silent	c.1398C>T	rs2292493
chr8_86576655	CA2	Silent	c.562T>C	rs703
chr19_13891689	CC2D1A	Silent	c.1281T>C	rs10410239
chr9_122202874	CDK5RAP2	Silent	c.5418C>T	rs3739822
chr9_122222023	CDK5RAP2	Silent	c.4041G>A	rs6478475
chr9_122260650	CDK5RAP2	Silent	c.2274T>C	rs2501727
chr9_122202874	CDK5RAP2	Silent	c.5418C>T	rs3739822
chr13_24364955	CENPJ	Silent	c.3042A>G	rs3742165
chr16_3717837	CREBBP	Silent	c.7212A>G	rs55916120
chr11_70824339	DHCR7	Silent	c.1158T>C	rs760241
chr11_70830109	DHCR7	Silent	c.438T>C	rs949177
chr11_70832801	DHCR7	Silent	c.207T>C	rs1790334
chr11_70832819	DHCR7	Silent	c.189G>A	rs1044482
chr11_70832777	DHCR7	Silent	c.231C>T	rs4316537
chr11_70824225	DHCR7	Silent	c.1272C>T	rs909217
chr20_30850008	DNMT3B	Silent	c.1572T>C	rs6058891
chr20_30850110	DNMT3B	Silent	c.1674T>C	rs2424922
chr11_118484236	DPAGT1	silent	c.16A>G	rs6589717
chr22_39880985	EP300	silent	c.3183T>A	rs20552
chr22_39903214	EP300	silent	c.5553T>C	
chr5_60236422	ERCC8	silent	c.435T>C	rs4647100
chr4_1777692	FGFR3	silent	c.1959G>A	rs7688609
chr4_1773502	FGFR3	silent	c.882T>C	rs2234909
chr7_41971125	GLI3	silent	c.4071C>T	rs34089404
chr7_41972361	GLI3	silent	c.2835G>C	rs61758978
chr7_42046290	GLI3	silent	c.900C>T	rs35961850
chr7_42054757	GLI3	silent	c.537C>T	rs3898405
chrX_122364958	GRIA3	silent	c.1200T>C	rs502434
chr6_102610010	GRIK2	silent	c.2424G>A	rs2227283
chr11_524242	HRAS	silent	c.81T>C	rs12628
chr20_10568275	JAG1	silent	c.3528C>T	rs1051421
chr20_10568386	JAG1	silent	c.3417T>C	rs1051419

chr20_10581237	JAG1	silent	c.765C>T	rs1131695
chr20_10573804	JAG1	silent	c.2214A>C	rs1801140
chr20_10585057	JAG1	silent	c.744A>G	rs10485741
chr20_10601469	JAG1	silent	c.267G>A	rs1051415
chr20_10587222	JAG1	silent	c.588C>T	rs1801138
chr12_25259729	KRAS	silent	c.483G>A	rs4362222
chr12_25254044	KRAS	silent	c.519T>C	rs1137282
chrX_43475980	MAOA	silent	c.891G>T	rs6323
chrX_43488335	MAOA	silent	c.1410T>C	rs1137070
chr8_6290433	MCPH1	silent	c.1782G>A	rs2584
chr8_6466586	MCPH1	silent	c.2418C>A	rs2912016
chr8_6259807	MCPH1	silent	c.228G>T	rs2305022
chr8_6466394	MCPH1	silent	c.2226C>T	rs2912010
chrX_70266672	MED12	silent	c.3930A>C	rs5030619
chrX_70277813	MED12	silent	c.6276G>A	
chr15_72976983	MPI	silent	c.1131A>G	rs1130741
chr17_26577611	NF1	silent	c.2034G>A	rs2285892
chr17_26532901	NF1	silent	c.702G>A	rs1801052
chr17_26507234	NF1	silent	c.168C>T	rs17881168
chr5_176569488	NSD1	silent	c.675C>T	rs1363405
chr5_176569755	NSD1	silent	c.1749G>A	rs3733874
chr5_176653804	NSD1	silent	c.6829T>C	rs28580074
chr5_176653878	NSD1	silent	c.6903G>C	rs11740250
chr9_133377309	POMT1	silent	c.1113C>T	rs3739494
chr9_133385395	POMT1	silent	c.1758G>A	rs34954751
chrX_48644671	PQBP1	silent	c.510G>A	
chr4_119422614	PRSS12	silent	c.2553A>C	
chr4_119456796	PRSS12	silent	c.1281G>A	rs2292597
chr1_218397295	RAB3GAP2	silent	c.3495G>A	rs11547779
chr17_17638979	RAI1	silent	c.1992G>A	rs8067439
chr17_17637824	RAI1	silent	c.837G>A	rs11078398
chr4_57471795	REST	silent	c.234G>T	rs61748752
chr4_57492946	REST	silent	c.3165G>A	rs2227901
chr17_26322577	RNF135	silent	c.360G>T	rs7224960
chrX_20114382	RPS6KA3	silent	c.798C>A	rs12009120
chr12_50367232	SCN8A	silent	c.576C>T	rs4761829
chr12_50470538	SCN8A	silent	c.4509T>C	rs303815
chr12_50487009	SCN8A	silent	c.5472C>A	rs60637
chrX_50367414	SHROOM4	silent	c.3468A>G	rs3747282
chr5_138484714	SIL1	silent	c.153A>G	rs3088052
chr18_51046529	TCF4	silent	c.1941A>G	rs8766
chr9_134762538	TSC1	silent	c.2829C>T	rs4962081
chr9_134772042	TSC1	silent	c.1335A>G	rs7862221

chr16_2078585	TSC2	silent	c.5397G>C	rs1051771
chr16_2076341	TSC2	silent	c.4809C>T	
chr16_2078270	TSC2	silent	c.5202T>C	rs1748
chr16_2074493	TSC2	silent	c.4269G>A	rs45438898
chrX_147856140	AFF2	missense	c.3040G>A	
chr22_48683892	ALG12	missense	c.1177A>G	rs3922872
chr1_63654140	ALG6	missense	c.911C>T	rs4630153
chr1_63644620	ALG6	missense	c.391T>C	rs35383149
chr11_77501439	ALG8	missense	c.803G>A	rs61995925
chr11_111229343	ALG9	missense	c.352G>A	rs10502151
chr1_195337524	ASPM	missense	c.7480T>C	rs964201
chr1_195327709	ASPM	missense	c.9395T>G	rs36004306
chr1_195337065	ASPM	missense	c.7939C>A	rs3762271
chr1_195337320	ASPM	missense	c.7684A>G	rs41310927
chr1_195339155	ASPM	missense	c.5849C>T	
chr13_30789743	B3GALTL	missense	c.1105G>A	rs34638481
chr13_30789746	B3GALTL	missense	c.1108G>A	rs1041073
chr11_66038671	BBS1	missense	c.378G>A	rs2298806
chr12_75264280	BBS10	missense	c.1616C>T	rs35676114
chr4_123883654	BBS12	missense	c.1157G>A	rs309370
chr4_123883896	BBS12	missense	c.1399G>A	rs13135778
chr16_55106002	BBS2	missense	c.209G>A	rs4784677
chr16_55102676	BBS2	missense	c.367A>G	rs11373
chrX_79830225	BRWD3	missense	c.3863A>G	rs3122407
chr19_13899791	CC2D1A	missense	c.2402C>T	rs2305777
chr19_13891753	CC2D1A	missense	c.1345G>A	
chr9_122210554	CDK5RAP2	missense	c.4618G>C	rs4837768
chr9_122330857	CDK5RAP2	missense	c.865G>C	rs4836822
chr9_122245733	CDK5RAP2	missense	c.3134G>C	rs3780679
chr9_122245802	CDK5RAP2	missense	c.3065G>A	rs34523498
chr13_24377541	CENPJ	missense	c.2635T>G	rs17402892
chr13_24384911	CENPJ	missense	c.253C>A	rs9511510
chrX_69582011	DLG3	missense		c.235G>A
chr11_118472968	DPAGT1	missense	c.1177A>G	rs643788
chr11_66010661	DPP3	missense	c.435G>T	rs11550299
chr11_66028813	DPP3	missense	c.2033G>A	rs2305535
chr22_39877954	EP300	missense	c.2989A>G	rs20551
chr9_107406555	FKTN	missense	c.608G>A	rs34787999
chr9_107437316	FKTN	missense	c.1336A>G	rs41313301
chrX_106733095	FRMPD3	missense	c.5269C>G	
chrX_106733134	FRMPD3	missense	c.5308C>G	
chr7_42054747	GLI3	missense	c.547A>G	rs846266
chr7_41972203	GLI3	missense	c.2993C>T	rs929387

chr20_10570501	JAG1	missense	c.2612C>G	rs35761929
chr8_6283958	MCPH1	missense	c.513G>T	rs2442513
chr8_6289591	MCPH1	missense	c.940G>C	rs930557
chr8_6289826	MCPH1	missense	c.1175A>G	rs2515569
chr8_6466450	MCPH1	missense	c.2282C>T	rs1057090
chr8_6487952	MCPH1	missense	c.2482C>T	rs1057091
chr8_6325714	MCPH1	missense	c.2045C>A	rs12674488
chr17_7431534	MPDU1	missense	c.685G>A	rs10852891
chr2_15999836	MYCN	missense	c.199T>G	
chr17_26725232	NF1	missense	c.8453C>A	
chr5_176569846	NSD1	missense	c.1840G>T	rs3733875
chr5_176570182	NSD1	missense	c.2176T>C	rs28932178
chr5_176570797	NSD1	missense	c.2791C>G	
chr13_44461464	NUFIP1	missense	c.108C>G	rs1140993
chrX_67569473	OPHN1	missense	c.115G>A	rs41303733
chr9_133375257	POMT1	missense	c.752A>G	rs2296949
chr9_133376602	POMT1	missense	c.979G>A	rs4740164
chr4_119422669	PRSS12	missense	c.2498G>A	rs17594503
chr4_119422617	PRSS12	missense	c.2550T>G	
chr4_119493160	PRSS12	missense	c.164G>C	rs13119545
chr1_218391338	RAB3GAP2	missense	c.4060A>G	rs59190330
chr1_218397828	RAB3GAP2	missense	c.3275G>C	rs2289189
chr17_17637256	RAI1	missense	c.269G>C	rs3803763
chr17_17637480	RAI1	missense	c.493C>A	rs11649804
chr17_17639523	RAI1	missense	c.2536T>G	
chr17_17647830	RAI1	missense	c.5601T>C	rs3818717
chr17_17641713	RAI1	missense	c.4726C>T	
chr4_57492171	REST	missense	c.2390C>T	rs3796529
chr4_57491857	REST	missense	c.2076G>T	rs2227902
chr17_26322430	RNF135	missense	c.213C>G	rs7225888
chr17_26322539	RNF135	missense	c.322T>C	rs7211440
chr12_50401628	SCN8A	missense	c.1667T>G	
chr12_50401630	SCN8A	missense	c.1669T>C	
chr12_50449090	SCN8A	missense	c.3076C>T	
chrX_73558294	SLC16A2	missense	c.319T>C	rs6647476
chr15_22770605	SNRPN	missense	c.694T>C	rs705
chr9_134761574	TSC1	missense	c.3364G>A	
chr9_134776725	TSC1	missense	c.965T>C	rs1073123
chrX_152949971	MECP2	nonsense	c.538C>T	

Supplementary Table 10 All variants detected in B3GALTL gene in sample S-6. Variants at the first exons (1-7) can be heterozygous or homozygous while all variants at the rest of the gene (8-15) have only a homozygous genotype.

Chromosome position	location	Ref. sequence	observed genotype Illumina	Wt	Variant	Mutation
chr13_30694969	intron 2	G	CC	0	41	c.121-120G>C
chr13_30719240	intron 5	C	CT	59	42	c.347+4C>T
chr13_30719469	intron 5	A	GG	0	9	c.347+233A>G
chr13_30719992	exon 6	T>	CC	0	11	c.348T>C
chr13_30733375	intron 7	G	GA	33	32	c.596+156G>A
chr13_30741415	intron 8	G	AA	0	27	c.660+1G>A
chr13_30746535	intron 8	G	AA	0	37	c.661-111G>A
chr13_30741664	intron 8	A	GG	0	14	c.660+250A>G
chr13_30746823	intron 9	A	GG	0	91	c.780+58A>G
chr13_30748714	intron 9	G	CC	0	10	c.781-125G>C
chr13_30749059	intron 10	A	GG	0	43	c.850+151A>G
chr13_30758769	intron 11	G	CC	0	9	c.965-88G>C
chr13_30757039	intron 11	T	AA	0	23	c.964+141T>A
chr13_30759059	intron 12	G	AA	0	37	c.1064+103G>A
chr13_30789561	intron 12	T	CC	0	13	c.1065-142T>C

Chapter **4**

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

***Both authors contributed equally to this work**

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

Abstract

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 six unrelated cases (three families and three sporadic cases) and was not found in 400 control X chromosomes, 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, because of nonrandom X chromosome inactivation, the mutant allele was not expressed in patient fibroblasts. RNA expression of the mutant allele was detected only 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.

Main text

Terminal osseous dysplasia (MIM 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 had been discovered.

In the present study, we examined terminal osseous dysplasia (TOD) in three families and three sporadic case individuals (patients 1, 2, and 3 described by Horii,[3] Drut,[4] and Breuning[5]). The Dutch family (Figure 1A, family 1) and Italian family (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 and is nonconsanguineous 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 has had multiple

minor surgeries to remove small skin lesions from her hands and legs. On clinical examination at the age of 25 yrs, her head circumference was 54 cm (25%–50%), her height was 170 cm (75%–90%), and her arm span was 171 cm. Her right hand showed brachydactyly of digit III-V, a short fingernail on digitus IV, and lateral deviation of the fifth digit. On her left hand, there was lateral deviation of the fourth digit, with a small lesion on the lateral aspect of the distal phalanx, and clinodactyly of the fifth digit (Figure 1B). Her right foot showed a short and highly implanted fourth digit. There was bilateral widening of the distal portion of the second–fifth digits. 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 mo of age to remove small skin lesions from her hands, feet, and gingiva. On clinical examination at the age of 3 yrs, she had a head circumference of 48 cm (25%–50%), a height of 85 cm (< 3%), and a weight of 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 second and fifth digits and bilateral clinodactyly of the fifth digit (Figure 1D). Her feet showed a lesion in her third toes and thickening of the nail of the fifth toes bilaterally. A skeletal X-ray survey revealed bilateral lytic lesions in the proximal humerus and the proximal femur, as well as 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 third digit, and bilateral overriding of the fourth toe. Echocardiogram at birth showed persistent foramen ovale. On clinical examination at the age of 6 mo, her head circumference was 42 cm (25%–50%), her height was 58.8 cm (< 3%), and her weight was 5.1 kg (< 3%). She has mild hypertelorism, three brownish pigmented spots of different sizes (3 mm to 1.5 cm) in her right temporal groove, mild retrognathia, a right upper accessory frenulum, 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 second–fifth digits, clinodactyly and lateral deviation of her second and third digits, and a narrow fifth digit with an absent distal crease. Her left hand showed skin lesions on her second-fourth digits. Her second digit was narrow and laterally deviated. There was camptodactyly of the third-fifth digits,

brachydactyly and clinodactyly of the fifth digit, and absence of a distal crease. In her feet, she had bilateral plantar pits. The right foot has distal broadening of the second-fifth toe and brachydactyly of the second and third toe accompanied by syndactyly. There was overriding of the third and fourth toe. On her left foot, the second-fifth toes were distally broad. She had a overlapping of the second and fourth toes over her third toe, brachydactyly of the third toe that was proximally 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.

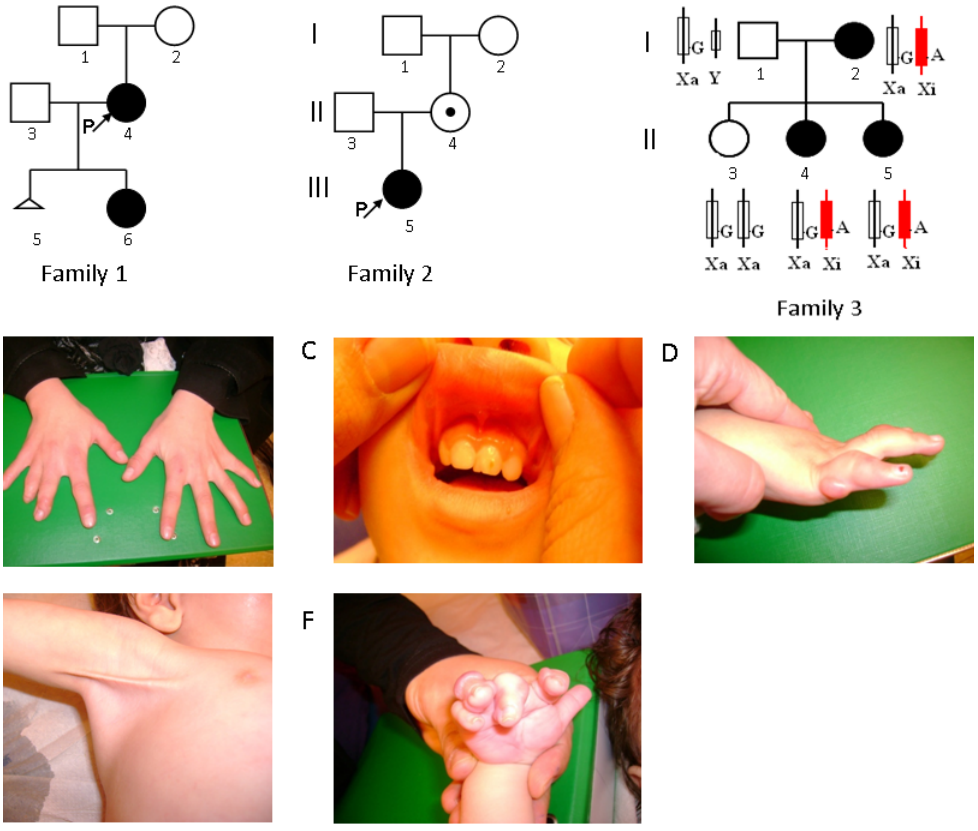


Figure 1 The Pedigrees and the Phenotype of Family 3. (A) The pedigrees investigated in this study. In family 3, XCI patterns show the silencing of the X chromosome that carries the mutant allele. (B) The hands of 3I:2. (C) Multiple frenula of 3II:4. (D) The right hand of 3II:4. She has clinodactyly and digital fibroma. (E) The right axillary pterygium of 3II:5. (F) The right hand of 3II:5.

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	Argentinian	Dutch
Age at 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 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	+							−		

mo: month

DNA of patients and family members were extracted from peripheral blood (families 1, 2, and 3), buccal cells (patient 1), or paraffin-embedded tissue (patients 2 and 3). Two probands (1II:4 and 2III:5) of the Dutch and the Italian families were tested with 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,

500 ng of DNA was fragmented (Bioruptor, Diagenode) according to manufacturer's instructions to yield fragments from 200 to 300 bp. Paired-end adaptor oligonucleotides from Illumina were added to both ends. The DNA-adaptor-ligated fragments were then hybridized to 250 ng of SureSelect X chromosome oligo capture library (SureSelect, Agilent) for 14 hr. After hybridization, washing, and elution, the elute was amplified to create sufficient DNA template for downstream applications. The eluted-enriched DNA fragments were sequenced with 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] (Table S1, available online). 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 was selected for further study for the following reasons: (1) c.5217G>A, the last nucleotide of exon 31, was predicted to affect splicing by Human Splicing Finder.[8] The score of the splicing donor site dropped from 91.2 to 80.63, indicating that the wild-type site may not function as usual. (2) 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 and 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, in whom affected cases (2II:4 and 2III:5) carry exactly the same variant, c.5217G>A, together with another exonic variant, c.5814C>T. Unfortunately, we did not have access to material from both parents and therefore could not determine whether the mutations occurred de novo. 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 in unaffected family members (1I:2, 3I:1, and 3II:3).

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 the HGVS numbers were generated using longest isoforms if multiple transcripts exist.

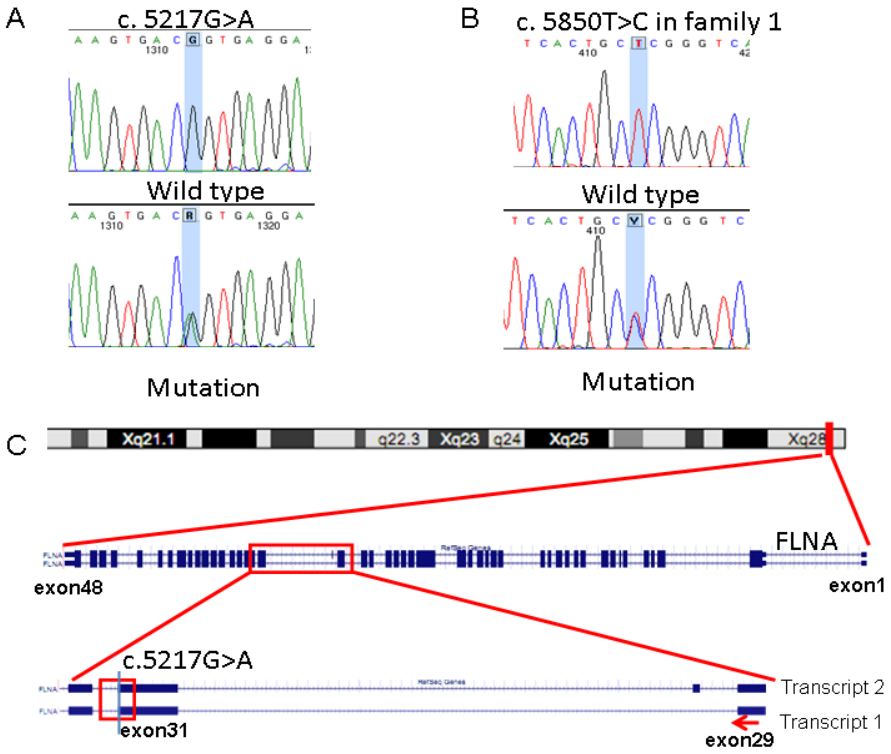


Figure 2 Genomic Structure and Mutation Analysis of *FLNA*. (A) c.5217G>A was confirmed by Sanger sequencing in all of the patients. The unaffected family members and controls carry the homozygous normal allele. (B) The sequence of c.5850T>C in family 1. (C) *FLNA* is located in Xq28, the target region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

Variant c.5217G>A affects the last nucleotide of exon 31 of the *FLNA* gene (Figure 2C). At the 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 anesthesia with informed consent. Cells were cultured in standard medium for human fibroblasts (Dulbecco's modified Eagle's medium with 10% FBS, 1% penicillin/streptomycin, 1% glucose, 1% glutamax) with 5% CO2 in 37°C. RNA was extracted with the RNeasy Mini Kit (QIAGEN). 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 with the use of the primers listed in Table S2. The products were evaluated with the Bioanalyzer 2100 DNA chip 1000 (Agilent), according to the manufacturer's instructions. RNA from patient fibroblasts showed only normal transcripts, both

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, lanes 2–4 and 8) and higher expression of transcript 2 in patient fibroblasts (Figure 3B, lane 1). Both bands were isolated from the agarose gel by the Qiaquick Gel Extraction Kit (QIAGEN) and analyzed by Sanger sequencing. Interestingly, we detected no expression of the mutant allele. This could be due to nonsense-mediated decay[14] and/or skewed X chromosome inactivation (XCI). To test the first possibility, the fibroblasts were treated with cycloheximide[15] for 4.5 hr followed by RNA analysis using the same procedures as those for RNA from untreated cells. The mutant allele was still absent in RNA from cycloheximide treated cells. XCI was analyzed with the Androgen Receptor (AR) assay.[16] The assay showed random XCI in 1I:2 versus 100% XCI 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 yr, patient 1III:6 had fibroma tissue from the fifth digits of both hands and the fifth toe of the left foot surgically removed and stored in liquid nitrogen. We cultured these cells and analyzed RNA. In the fibroma cells, we observed two sets of two bands (Figure 3B, lanes 5–7), indicating altered splicing. One set had the same length as that observed in normal fibroblasts (Figure 3A, transcripts 1 and 2), and the other set was shorter (Figure 3A, transcripts 3 and 4, faint from RNA of a tumor in left fifth finger and toe; Figure 3B, lanes 6 and 7). Note that the fibroma always contains a mixture of tumor and normal stroma cells. 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 the literature thus far (83 in total; see *FLNA* mutation database), including the variants described here. The c.5217G>A variant detected in TOD patients has not been described before; it is listed neither in dbSNP nor in the pilot study 1 of the 1000 Genomes Project. Finally, over 400 chromosomes have been sequenced and the mutant allele was not found (data not shown).

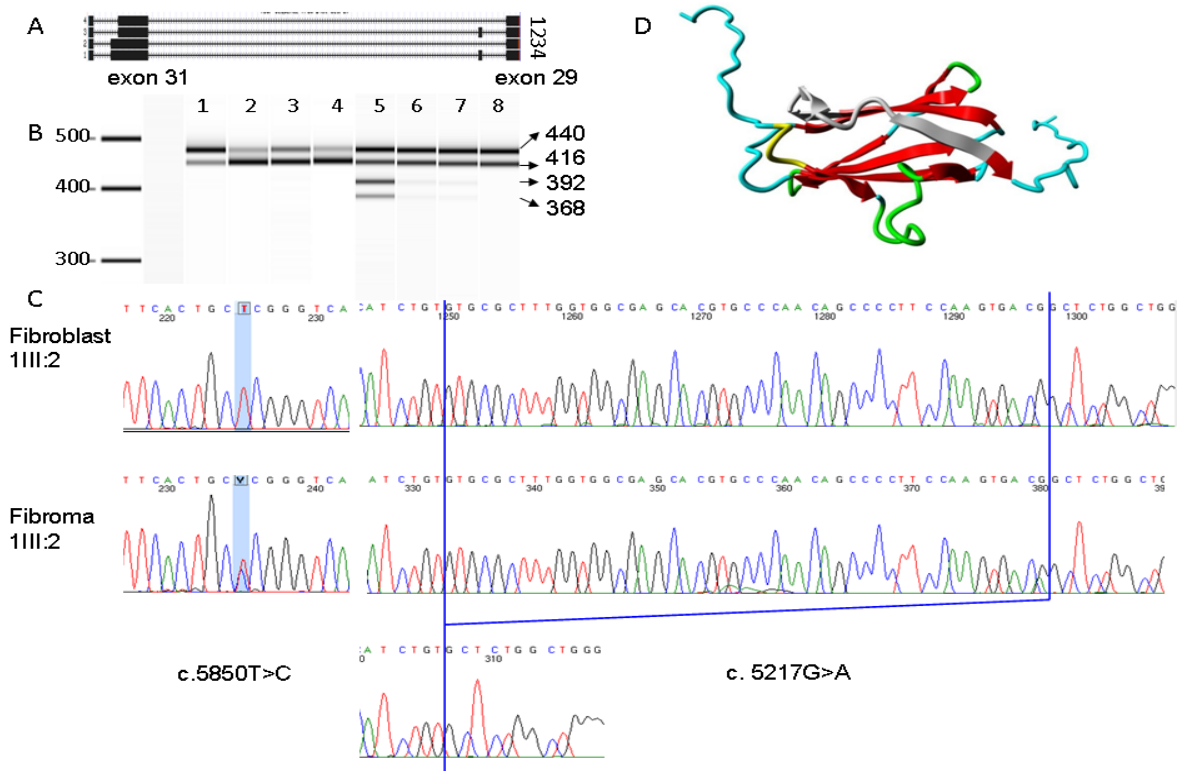


Figure 3 Detection of Alternative Splicing and 3D Protein Model. (A) Diagram of four *FLNA* transcripts in fibroma cells: transcripts 1 and 2, which carry the 48 bp deletion at the end of exon 31, as well as the normal transcripts 3 and 4. (B) RT-PCR result from Agilent 2100 Bioanalyzer. Lane 1 is the product of the fibroblasts of 1III:6, which has a predominant longer isoform. Lanes 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.

Mutations in *FLNA* have been reported to cause a wide range of developmental malformations in the brain, bones, limbs, heart,[18] and other organs[19] in human,[9] including periventricular heterotopia (PVNH [MIM 300049])[20–24] and otopalatodigital (OPD) spectrum disorders,[25] which include otopalatodigital syndrome type 1 (MIM 311300)26–28 and type 2 (MIM 304120),[26,29] frontometaphyseal dysplasia (MIM 305620),[26,30,31] and Melnick-Needles syndrome (OMIM 309350).[26,27] Although each of the OPD spectrum disorders are characterized by specific clinical symptoms, there clearly is a clinical overlap with TOD, including a generalized bone dysplasia that includes craniofacial anomalies and anomalies in 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 terminus and connects with 24 immunoglobin-like filamin repeats, divided by two hinges, one between repeats 15 and 16 and one between repeats 23 and 24. To check the stability of filamin A in patient cells, protein was extracted from both fibroblast and fibroma cells. Immunoblot was performed with the use of mouse human filamin A monoclonal antibody, MAB1680, from Millipore. No difference in molecular weight or quantity was observed. The difference of 18 amino acids was likely too small to be distinguished by immunoblot. 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, and found in all ten affected patients from six different unrelated families. In addition, the mutation introduced abnormal splicing in fibroma cells. At the protein level, c.5217G>A encodes the second-to-last amino acid of repeat 15, which is immediately adjacent to hinge 1. Recent studies demonstrated repeats 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, such as TRAF1, TRAF2,[36] CaR extracellular Ca2+ receptor,[37] and FAP52.[38] Because no crystal structure has yet been described for this region, the crystal structure of repeat 15 in filamin B(PDB file 2 dmb), which shows the highest identity (58%) with this region of interest, was used as a template for building a 3D model (Figure 3D). The model was built with the use of the WHAT IF and YASARA twinset.[39] Repeat 15 consists of two 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 (gray part of Figure 3D). These residues are likely to form some kind of beta strand-like structure and to 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 hinge 1. 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.

The precise mechanism of TOD remains unclear. However, like other X-linked diseases, 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 begins to occur in the first years, and 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 the cell that carries either the mutated or the normal allele (secondary cell selection), and genetic processes yielding preferential inactivation of specific alleles. Primarily the XCI choice is random, but during cell proliferation, either in all cells or in a tissue specific manner, cells that carry an active 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 the *FLNA* gene often show a skewed XCI pattern,[26] suggesting that cells need normal filamin to survive. Several studies in TOD families showed that patients had skewed XCI, while unaffected individuals had random inactivation.[1,6] We examined the XCI pattern in family 1 (1I:2, 1II:4, and 1III:6) and family 3 (3I:2, 3II:3, 3II:4, and 3II:5; Figure 1A) by AR assay. Apart from the uninformative patient 1III:6, all of the other patients—1II:4, 3I:2, 3II:4, and 3II:5—showed extremely skewed XCI (0/100%), whereas the normal family member 1I:2 showed random XCI (30/70%), as did 3II:3 (50/50%). Because there was no mutant allele detectable in the RNA of normal fibroblast, we deduced that 1III:6 also had 100% skewed XCI with the preferential inactivation of the mutant allele. We 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 that 2II:4 is a carrier of TOD, given that she has only mild manifestations (multiple frenula in the mouth). She probably has skewed XCI at a very early stage. Local XCI 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 is caused by a 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 is located 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 it causes a deletion of 16 amino acids at the protein level. The missing region in the filamin A protein is hypothesized to affect or prevent the interaction of filamin A with other proteins.

Acknowledgments

We would like to thank the patients and their family members for their willingness to join the project, the China Scholarship Council (CSC) scholarship for supporting Yu Sun's studies in The Netherlands, Filip Kluin for sending paraffin-embedded tissue and Hans Morreau for isolating DNA from the tissue, Tobias Messemaker for helping us with immunoblotting, and the Leiden Genome Technology Center (LGTC) and the Laboratory for Diagnostic Genome Analysis (LDGA) for help with sequencing, DNA extraction, and XCI detection. X-exome capture was implemented in collaboration with ServiceXS (Leiden, http://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). B.F. was funded by the Italian Telethon Foundation

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/

References:

1. Bacino CA, Stockton DW, Sierra RA, Heilstedt HA, Lewandowski R, Van den Veyver IB. Terminal osseous dysplasia and pigmentary defects: clinical characterization of a novel male lethal X-linked syndrome. Am J Med Genet. 2000;94:102–112.
2. Zhang W, Amir R, Stockton DW, Van Den Veyver IB, Bacino CA, Zoghbi HY. Terminal osseous dysplasia with pigmentary defects maps to human chromosome Xq27.3-xqter. Am J Hum Genet. 2000;66:1461–1464.
3. Horii E, Sugiura Y, Nakamura R. A syndrome of digital fibromas, facial pigmentary dysplasia, and metacarpal and metatarsal disorganization. Am J Med Genet. 1998;80:1–5.
4. Drut R, Pedemonte L, Rositto A. Noninclusion-body infantile digital fibromatosis: a lesion heralding terminal osseous dysplasia and pigmentary defects syndrome. Int J Surg Pathol. 2005;13:181–184.
5. Breuning MH, Oranje AP, Langemeijer RA, Hovius SE, Diepstraten AF, den Hollander JC, Baumgartner N, Dwek JR, Sommer A, Toriello H. Recurrent digital fibroma, focal dermal hypoplasia, and limb malformations. Am J Med Genet. 2000;94:91–101.
6. Baroncini A, Castelluccio P, Morleo M, Soli F, Franco B. Terminal osseous dysplasia with pigmentary defects: clinical description of a new family. Am J Med Genet Part A. 2007;143:51–57.
7. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10:R25.
8. Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37:e67.
9. Robertson SP. Filamin A: phenotypic diversity. Curr Opin Genet Dev. 2005;15:301–307.
10. Agarwal N, Kutlar F, Mojica-Henshaw MP, Ou CN, Gaikwad A, Reading NS, Bailey L, Kutlar A, Prchal JT. 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. 2007;92:1715–1716.
11. Yamada K, Fukao T, Zhang G, Sakurai S, Ruiter JP, Wanders RJ, Kondo N. 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. 2007;90:291–297.
12. Kuivaniemi H, Tromp G, Bergfeld WF, Kay M, Helm TN. 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. 1995;105:352–356.
13. Maestrini E, Patrosso C, Mancini M, Rivella S, Rocchi M, Repetto M, Villa A, Frattini A, Zoppè M, Vezzoni P. 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. 1993;2:761–766.
14. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. Nonsense-mediated decay approaches the clinic. Nat Genet. 2004;36:801–808.
15. Kim CE, Gallagher PM, Guttormsen AB, Refsum H, Ueland PM, Ose L, Folling I, Whitehead AS, Tsai MY, Kruger WD. Functional modeling of vitamin responsiveness in yeast: a common pyridoxine-responsive cystathionine beta-synthase mutation in homocystinuria. Hum Mol Genet. 1997;6:2213–2221.
16. Kubota T, Nonoyama S, Tonoki H, Masuno M, Imaizumi K, Kojima M, Wakui K, Shimadzu M, Fukushima Y. A new assay for the analysis of X-chromosome inactivation based on methylation-specific PCR. Hum Genet. 1999;104:49–55.
17. Fokkema IF, den Dunnen JT, Taschner PE. LOVD: easy creation of a locus-specific sequence variation database using an “LSDB-in-a-box” approach. Hum Mutat. 2005;26:63–68.
18. Kyndt F, Gueffet JP, Probst V, Jaafar P, Legendre A, Le Bouffant F, Toquet C, Roy E, McGregor L, Lynch SA. Mutations in the gene encoding filamin A as a cause for familial cardiac valvular dystrophy. Circulation. 2007;115:40–49.

19. Gargiulo A, Auricchio R, Barone MV, Cotugno G, Reardon W, Milla PJ, Ballabio A, Ciccodicola A, Auricchio A. Filamin A is mutated in X-linked chronic idiopathic intestinal pseudo-obstruction with central nervous system involvement. Am J Hum Genet. 2007;80:751–758.
20. Fox JW, Lamperti ED, Ekşioğlu YZ, Hong SE, Feng Y, Graham DA, Scheffer IE, Dobyns WB, Hirsch BA, Radtke RA. Mutations in filamin 1 prevent migration of cerebral cortical neurons in human periventricular heterotopia. Neuron. 1998;21:1315–1325.
21. Sheen VL, Dixon PH, Fox JW, Hong SE, Kinton L, Sisodiya SM, Duncan JS, Dubeau F, Scheffer IE, Schachter SC. Mutations in the X-linked filamin 1 gene cause periventricular nodular heterotopia in males as well as in females. Hum. Mol. Genet. 2001;10:1775–1783.
22. Moro F, Carrozzo R, Veggiotti P, Tortorella G, Toniolo D, Volzone A, Guerrini R. Familial periventricular heterotopia: missense and distal truncating mutations of the FLN1 gene. Neurology. 2002;58:916–921.
23. Zenker M, Rauch A, Winterpacht A, Tagariello A, Kraus C, Rupprecht T, Sticht H, Reis A. 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. 2004;74:731–737.
24. Sheen VL, Jansen A, Chen MH, Parrini E, Morgan T, Ravenscroft R, Ganesh V, Underwood T, Wiley J, Leventer R. Filamin A mutations cause periventricular heterotopia with Ehlers-Danlos syndrome. Neurology. 2005;64:254–262.
25. Robertson SP. Otopalatodigital syndrome spectrum disorders: otopalatodigital syndrome types 1 and 2, frontometaphyseal dysplasia and Melnick-Needles syndrome. Eur J Hum Genet. 2007;15:3–9.
26. Robertson SP, Twigg SR, Sutherland-Smith AJ, Biancalana V, Gorlin RJ, Horn D, Kenwrick SJ, Kim CA, Morava E, Newbury-Ecob R, OPD-spectrum Disorders Clinical Collaborative Group Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. Nat Genet. 2003;33:487–491.
27. Robertson SP, Thompson S, Morgan T, Holder-Espinasse M, Martinot-Duquenoy V, Wilkie AO, Manouvrier-Hanu S. Postzygotic mutation and germline mosaicism in the otopalatodigital syndrome spectrum disorders. Eur J Hum Genet. 2006;14:549–554.
28. Hidalgo-Bravo A, Pompa-Mera EN, Kofman-Alfaro S, Gonzalez-Bonilla CR, Zenteno JC. 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. 2005;136:190–193.
29. Mariño-Enríquez A, Lapunzina P, Robertson SP, Rodríguez JI. 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. 2007;143A:1120–1125.
30. Zenker M, Nährlich L, Sticht H, Reis A, Horn D. Genotype-epigenotype-phenotype correlations in females with frontometaphyseal dysplasia. Am J Med Genet Part A. 2006;140:1069–1073.
31. Giuliano F, Collignon P, Paquis-Flucklinger V, Bardot J, Philip N. A new three-generational family with frontometaphyseal dysplasia, male-to-female transmission, and a previously reported FLNA mutation. Am J Med Genet Part A. 2005;132A:222.
32. Robertson SP. Molecular pathology of filamin A: diverse phenotypes, many functions. Clin Dysmorphol. 2004;13:123–131.
33. Zhou AX, Hartwig JH, Akyürek LM. Filamins in cell signaling, transcription and organ development. Trends Cell Biol. 2010;20:113–123.
34. Nakamura F, Osborn TM, Hartemink CA, Hartwig JH, Stossel TP. Structural basis of filamin A functions. J Cell Biol. 2007;179:1011–1025.
35. Gardel ML, Nakamura F, Hartwig JH, Crocker JC, Stossel TP, Weitz DA. Prestressed F-actin networks cross-linked by hinged filamins replicate mechanical properties of cells. Proc Natl Acad Sci USA. 2006;103:1762–1767.
36. Arron JR, Pewzner-Jung Y, Walsh MC, Kobayashi T, Choi Y. Regulation of the subcellular localization of tumor necrosis factor receptor-associated factor (TRAF)2 by TRAF1 reveals

mechanisms of TRAF2 signaling. J Exp Med. 2002;196:923–934.
37. Awata H, Huang C, Handlogten ME, Miller RT. Interaction of the calcium-sensing receptor and filamin, a potential scaffolding protein. J Biol Chem. 2001;276:34871–34879.
38. Nikki M, Meriläinen J, Lehto VP. FAP52 regulates actin organization via binding to filamin. J Biol Chem. 2002;277:11432–11440.
39. Krieger E, Koraimann G, Vriend G. Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. Proteins. 2002;47:393–402.
40. Sharp A, Robinson D, Jacobs P. Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Hum Genet. 2000;107:343–349.
41. Orstavik KH. X chromosome inactivation in clinical practice. Hum Genet. 2009;126:363–373

Supplementary data

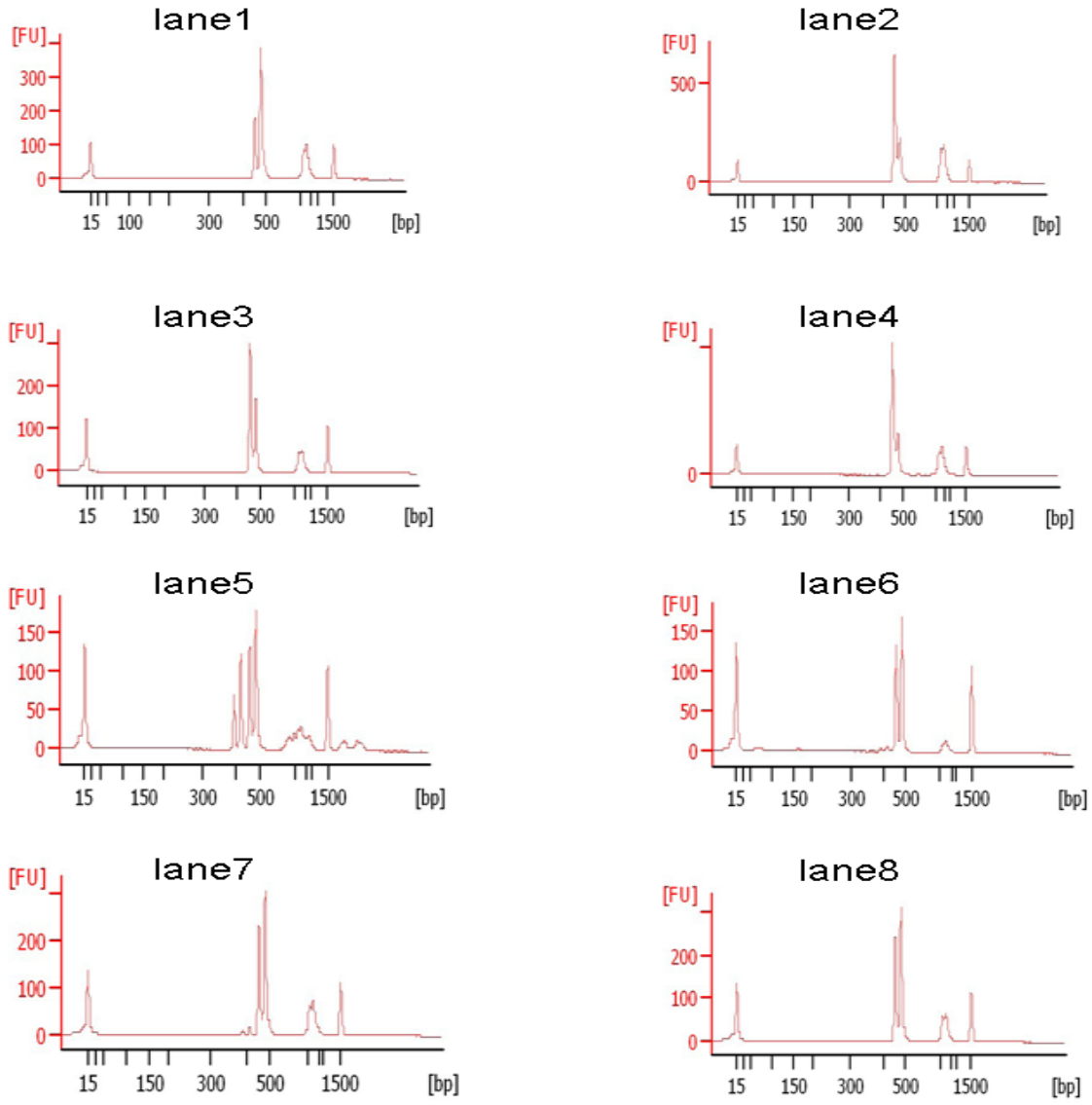
Supplementary Table S1 Overview of the data generated by GAII

	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

Supplementary Table S2 FLNA primer list

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

Supplementary Figure S3 2100 bioanalyzer traces of RT-PCR on c.5217G>A from lane1 to 8. The peak around 15bp is the lower ladder and the signal round 1500bp is the upper ladder.



Chapter 5

Autosomal Recessive Spinocerebellar Ataxia 7 (SCAR7) is Caused by Variants in *TPP1*, The Gene Involved in Classic Late-Infantile Neuronal Ceroid Lipofuscinosis 2 Disease (CLN2 Disease)

Yu Sun*, Rowida Almomani*, Guido Breedveld, Gijs W.E. Santen, Emmelien Aten, Dirk J. Lefeber, Jorrit I. Hoff, Esther Brusse, Frans W. Verheijen, Rob M. Verdijk, Marjolein Kriek, Ben Oostra, Martijn H. Breuning, Monique Losekoot, Johan T. den Dunnen, Bart P. van de Warrenburg, and Anneke J.A. Maat-Kievit

***The authors contributed equally to the work**

Hum Mutat. 2013; 34:706-13.

Abstract

Spinocerebellar ataxias are phenotypically, neuropathologically and genetically heterogeneous. The locus of autosomal recessive spinocerebellar ataxia type 7 (SCAR7) was previously linked to chromosome band 11p15. We have identified *TPP1* as the causative gene for SCAR7 by exome sequencing. A missense and a splice site variant in *TPP1*, cosegregating with the disease, were found in a previously described SCAR7 family and also in another patient with a SCAR7 phenotype. *TPP1*, encoding the tripeptidyl peptidase 1 enzyme, is known as the causative gene for neuronal ceroid lipofuscinosis disease 2 (CLN2). CLN2 is characterized by epilepsy, loss of vision, ataxia and a rapidly progressive course, leading to early death. SCAR7 patients showed ataxia and low activity of tripeptidyl peptidase 1, but no ophthalmologic abnormalities or epilepsy. Also, the slowly progressive evolution of the disease until old age and absence of ultra structural curvilinear profiles is different from the known CLN2 phenotypes.

Our findings now expand the phenotypes related to *TPP1*-variants to SCAR7. In spite of the limited sample size and measurements a putative genotype-phenotype correlation may be drawn: we hypothesize that loss of function variants abolishing TPP1 enzyme activity lead to CLN2, while variants that diminish TPP1 enzyme activity lead to SCAR7.

Introduction

Spinocerebellar ataxias are phenotypically, neuropathologically and genetically heterogeneous, with over 50 genes and loci associated with genetic forms of spinocerebellar ataxias (Matilla-Duenas, 2012; Vermeer, et al., 2011). The inheritance of the disease can be either autosomal dominant, autosomal recessive, X-linked or mitochondrial. Due to genetic heterogeneity of the the hereditary ataxias, it is time and money consuming to check all known genes by Sanger sequencing (Sailer and Houlden, 2012). Recently developed genomic techniques, such as exome sequencing that targets only the coding portion of the genome, offer an alternative strategy to rapidly sequence all genes in a comprehensive manner and its utility has been demonstrated in more diagnostic settings (Sailer, et al., 2012).

Breedveld *et al.* previously reported a unique Dutch family with a childhood onset, slowly progressive autosomal recessive spinocerebellar ataxia, referred to as SCAR7 (OMIM 609207) and distinguished from other recessive ataxia types (Suppl. Table S1) by locus, onset and/or clinical findings. A genome-wide linkage study mapped the causative gene on a 5.9 cM region on chromosome band 11p15, which contains more than 200 genes. No obvious candidate gene could be assigned, as genes for ataxia mostly have different functions and features (Breedveld, et al., 2004).

Here we report the results of exome sequencing in the Dutch family revealing disease-causing variants in the *TPP1* gene (OMIM 607998), encoding the lysosomal enzyme tripeptidyl peptidase 1.

Homozygous or compound heterozygous variants in *TPP1* usually lead to neuronal ceroid lipofuscinosis 2 disease (CLN2; OMIM 204500) (Williams and Mole, 2012), a neurodegenerative disorder generally characterized by onset at 2-4 years of age with seizures, ataxia and a progressive cognitive and motor dysfunction, and visual impairment later in the course of the disease, followed by death at the end of the first decade or beginning of the second (Santavuori, 1988; Williams, et al., 1999). Our findings expand the phenotypes of *TPP1* mutations (Kousi, et al., 2012) to SCAR7.

Materials & Methods

Patients

The clinical data of the original Dutch sib ship (pedigree family A, Figure 1A) have been reported previously (Breedveld, et al., 2004). In summary, patients of family A, suffer from a childhood-onset spinocerebellar ataxia with pyramidal signs and posterior column involvement and a postural tremor, without other (non-) neurological features. Neuroimaging shows atrophy of cerebellum, vermis, pons, and medulla oblongata.

Patient B-II.1 (Figure 1B), a 51 year-old woman reported an onset of symptoms with diplopia at age 18. Two years later, subtle gait changes occurred. At the age of 28 years, she was diagnosed with cerebellar atrophy. Symptoms have been very slowly progressive since then; she still walks unsupported, although with occasional falls. She volunteered some loss of dexterity, mild speech and swallowing difficulties, and urinary urgency. Family history was negative and there was no consanguinity known in the parents. On examination, we observed normal cognitive functions; square-wave jerks, jerky pursuit, and hypermetric saccades; cerebellar dysarthria; mild proximal leg muscle weakness; no extrapyramidal features; very mild gait and appendicular ataxia; clear hyperreflexia with ankle jerk clonus; equivocal plantar responses; and normal sensory examination. MRI showed diffuse cerebellar atrophy (Figure 2). Full ophthalmologic evaluation was completely normal. Negative or normal outcomes were obtained for molecular genetic testing of various SCA genes (1, 2, 3, 6, 7, 12, 13, 14, and 17), APTX, SETX, FXN, SACS, SPG7, and ANO10, as well as measurements of creatine kinase level, alpha-fetoprotein, vitamins, and acanthocytes, and lysosomal enzymes. However, increased activity of plasma chitotriosidase as a marker for lysosomal disorders (280 nmol/h/ml, reference <160) and decreased TPP1 activity was noted. The phenotype, TPP1 enzyme activity and *TPP1* mutations of these SCAR7 and other Dutch CLN2 patients (C-O) and relatives are described in Table 1.

All patients in this study provided informed consent for DNA studies, and for diagnostic procedures.

Exome sequencing

Genomic DNA from patients and relatives from family A was extracted from peripheral blood using the salt precipitation method (PUREGENE, QIAGEN). Exome sequencing was performed on patient A.III-2 using the SureSelect 50Mb exome capture kit (Agilent) following the manufacturer's protocol. The captured fragments were subsequently sequenced by Illumina HiSeq as previously described (Santen, et al., 2012), paired end mode. Read length is 100bp.

The raw Fastq files were aligned by bwa-0.5.9 (Li and Durbin, 2009). SAM/BAM files were manipulated by Samtools-0.1.10 (Li, et al., 2009) and Picard-1.57. Variations were called by GATK (McKenna, et al., 2010). The output vcf file was annotated by uploading to SeattleSeq 134 (http://snp.gs.washington.edu/SeattleSeqAnnotation134/). The responsible gene for autosomal recessive ataxia was mapped in a 5.9 cM linkage interval (4.5 Mb), so only variants in this region were considered as candidates.

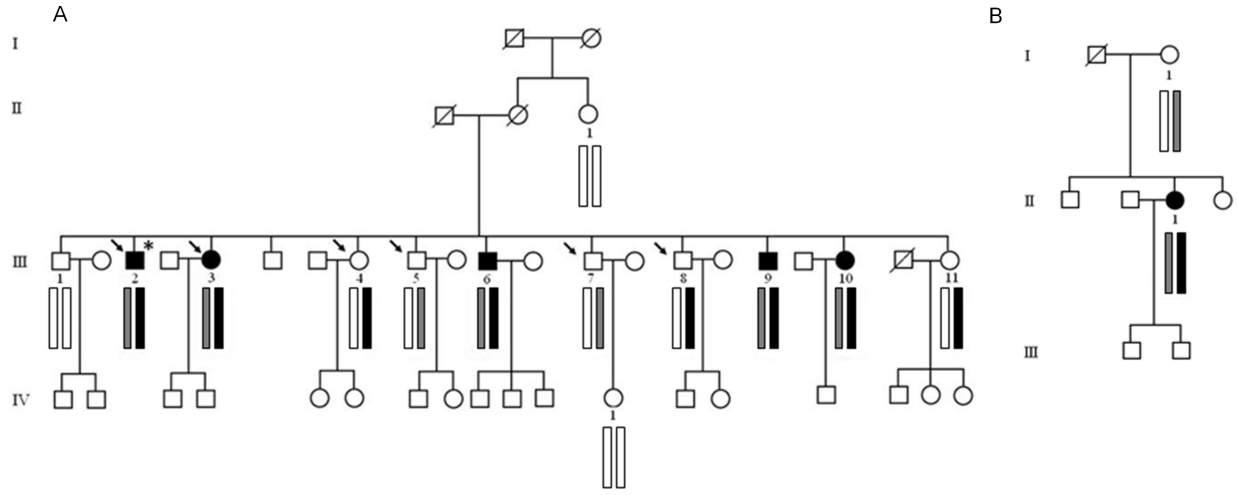


Figure 1. (A) Pedigree of family A and (B) family B with autosomal recessive spinocerebellar ataxia (SCAR7). Patients are marked with black symbols, unaffected relatives with open symbols. *TPP1* genotypes are shown below individuals (open bars indicate normal allele, black bars indicate alleles with c.509-1G>C, grey bars represent alleles with c.1397T>G). Genotype analysis shows co-segregation of variants with disease.

* =exome sequencing performed; → =RNA analysis performed.

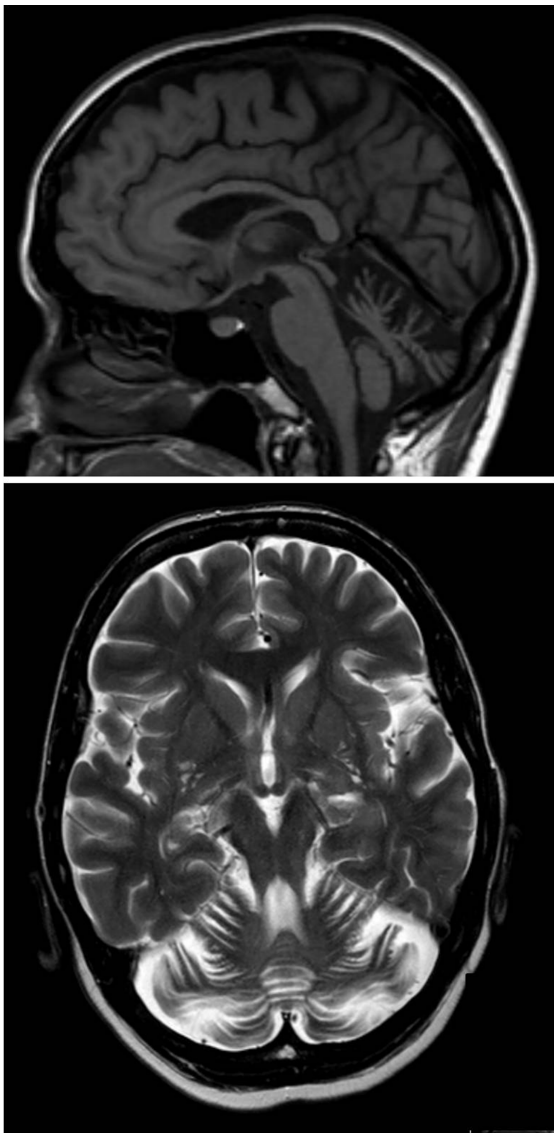


Figure 2. MRI of the brain of patient B. Sagittal T1-weighted (top) and transversal T2-weighted (bottom) show diffuse cerebellar atrophy.

Sanger sequencing

Sanger sequencing was applied to confirm the finding of exome sequencing and cosegregation of the variations in family A. PCR was performed by using Phire Hot Start II DNA polymerase (Finnzyme) following the official protocol. Primers used in PCR reactions are shown in Supp. Table S2. PCR products were first purified by QIAquick PCR purification kit (QIAGEN), subsequently mixed with 25pmol of either forward or reverse primers and sequenced.

For family B, the coding region and flanking intron sequences of the *TPP1* gene were examined by Sanger sequencing in a diagnostic setting, using standard procedures (The protocol and the primer sequences are available upon request).

RNA analysis

Leucocyte RNA from affected compound heterozygous patients (A.III-2, A.III-3), unaffected c.509-1G>C carriers (A.III-4, A.III-8), and unaffected c.1397T>G carriers (A.III-5, A.III-7) was isolated from blood using RNABEE following the official protocol. cDNA synthesis and RT-PCR was performed as previously described (Sun, et al., 2010). The primer sequences are listed in Supp. Table S2. The RT-PCR products were examined by 2% agarose gel, and followed by Sanger sequencing.

TPP1 enzyme activity assay

Enzyme activity of tripeptidyl peptidase 1 was assayed in leucocytes and fibroblasts of SCAR7 patients via the determination of fluorescent 7-amino-4 methylcoumarin, released from the substrate Ala-Ala-Phe 7-amido-4-methylcoumarin by incubation in cell homogenates as described previously (Van Diggelen, et al., 2001). Also TPP1 activity in leucoytes of SCAR7 carriers was measured. The TPP1 enzyme activity of B-II.1 was tested in Radboud University Nijmegen Medical Centre (normal range: 37-209 nmol/h/mg protein), while family A and other CLN2 samples (C-O) are analyzed in Erasmus Medical Center, with the normal range, 125-340 nmol/h/mg protein.

EM study

Fibroblasts from one of the patients of family A-III.2 and patient B-II.1 were fixed in glutaric-aldehyde, postfixated with osmium tetroxide and embedded in Epon (Hexion Specialty Chemicals, Inc, Danbury, Connecticut) and examined by electron microscopy.

Results

Exome sequencing reveals candidate variants

Exome sequencing was performed to target protein coding sequences in the human genome for potential disease-causing variants. An overview of the data obtained is listed in Supp. Table S3. Only variants inside the 5.9 cM linkage interval (from D11S4088 to D11S1331, genomic location: chr11.hg19:g.2,754,951-7,292,210) were analyzed. If the variant allele frequency in the NHLBI ESP exomes (http://evs.gs.washington.edu/EVS/) was larger than 5%, the variant was removed from the candidate list. We then selected stop-gain, stop-loss, missense, splice site, frameshift and in-frame coding indel variants and concordance with autosomal recessive inheritance (i.e. homozygous or compound heterozygous variants in one gene). Variants in three genes, *C11orf40*, *TPP1* and *DCHS1* (Table 2), fulfilled these criteria. However, *C11orf40* and *DCHS1,* located on the same allele, showed no co-segregation with the disease, they were excluded.. Thus, the only candidate gene left was *TPP1*, encoding the lysosomal serine protease with tripeptidyl-peptidase 1 activity. Two *TPP1* variants, a splice site variant, c.509-1G>C and a missense variant, c.1397T>G, p.(Val466Gly) (Figure 3A), co-segegrated with the disease (Figure

1A). Later, identical variants in the *TPP1* gene were found in a sporadic patient B, in whom *TPP1* variant analysis was requested because of the low TPP1 enzyme activity, obtained through the diagnostic work-up. Analysis of the *TPP1* closest homologs in other species (Figure 3B) showed that the Val466 residue is highly conserved during evolution, suggesting a functional role for this amino acid, and a deleterious effect for the predicted Val466Gly change (SIFT: deleterious, Polyphen: possibly damaging). Since both patients carried identical disease-causing variants we considered the possibility of the presence of founder alleles. Genealogical studies showed no close relation between the two families. Sanger sequencing of six closely linked variants (five of which are low frequent) covering the *OR56A3*, *TPP1* and *DCHS1* (Supp. Table S4) shows the presence of two haplotypes in patients of family A, one allele T-***C***-C-A-T, includes c.1397T>G and the other C-A-***G***-G-C, includes c.509-1G>C. They differ with the genotypes of patient B.II-1 at the two outer variants g.5968589C>T and g.6662466C>T, giving a maximal length of shared haplotype of 700 kb. The two families are not closely related, since the length is small. However, this does not exclude the variants derive from a common ancestor in the Dutch population. The results of molecular *TPP1* testing of patients from family A, patient B-II.1 and Dutch CLN2 patients (C-O) are summarized in Table 1.

Table 1. TPP1 enzyme activity, *TPP1* mutations and phenotype of patients from family A, B and other Dutch patients (C-O)

Patient	TPP1 activity leucocytes (nmol/h/m	TPP1 activity fibroblasts	TPP1 mutations	Predicted protein	Curvilinear profiles	Phenotype	Onset	Death yr/ (current age)	Affected relatives
A-III.2	16	7.2	c.509-1G>C/ c.1397T>G	splice defect/ p.Val466Gly	no	SCAR7	childhood	no/ (55)	4 sibs SCAR7
A-III.3	8	nd[1]	c.509-1G>C/ c.1397T>G	splice defect/ p.Val466Gly		SCAR7	childhood	no/ (59)	4 sibs, SCAR7
A-III.6	18	nd	c.509-1G>C/ c.1397T>G	splice defect/ p.Val466Gly		SCAR7	childhood	no/ (66)	4 sibs, SCAR7
A-III.9	Nd	nd	c.509-1G>C/ c.1397T>G	splice defect/ p.Val466Gly		SCAR7	childhood	68	4 sibs, SCAR7
A-III.10	32	nd	c.509-1G>C/ c.1397T>G	splice defect/ p.Val466Gly		SCAR7	childhood	no/ (73)	4 sibs, SCAR7
B-II.1	4.0-13.0[2]	nd	c.509-1G>C/ c.1397T>G	splice defect/ p.Val466Gly	no	SCAR7	18yr	no (51)	No
C	2.16	0	c.509 -1G>C/ c.509 -1G>C	splice defect/splice defect		CLN2	3 yr	no (13)	1 sib, CLN2
D	3.19	nd	nd	nd		CLN2	4.5 yr	23	No
E1	21.6	0.7-1.1	c.509 -1G>C/ c.509 -1G>C	splice defect/splice defect		CLN2	3.5 yr	11	1 sib, CLN2
E2	3.73	nd	c.509 -1G>C/ c.509 -1G>C	splice defect/splice defect		CLN2	3 yr	no (14)	1 sib, CLN2
F	3.88-9.04	nd	nd	nd		CLN2	4 yr	8	No
G	4.8- 5.14	0.4	c.509-1G>C/ c.622C>T	splice defect/p.Arg208X		CLN2	1.5 yr	no (11)	No
H	5.18	nd	nd	nd		CLN2	4 yr	no (6)	No
I	10.1-10.8	nd	c.622C>T/ c.1266G>C	p.Arg208X/p.Gln422His		CLN2	4 yr	no (9)	No
L	15.4	nd	nd	nd		CLN2	4 yr	12	No
M	24.6-32.3	0.3	c.509-1G>C/ c.622C>T	splice defect/p.Arg208X	yes	CLN2	3 yr	8	No
N	25.6-33.1	nd	c.509-1G>C/ c.622C>T	splice defect/p.Arg208X		CLN2	1.5 yr	no (7)	No
O1	27.7	nd	c.225A>G/ c.622C>T	splice defect/p.Arg208X	yes	CLN2	3 yr	10	1 sib, CLN2
O2	23.5	1.5	c.225A>G/ c.622C>T	splice defect/p.Arg208X	yes	CLN2	3 yr	no (17)	1 sib, CLN2

[1]nd = not done. [2] this sample was tested in Nijmegen; other samples in Rotterdam.

Table 2. The candidate variant list

Gene	Chromosome	Position	Reference base	Sample genotype	HGVS nomenclature	Function GVS
TPP1	11	6636430	A	A/C	NM_000391.3:c.1397T>G	Missense
TPP1	11	6638385	C	C/G	NM_000391.3:c.509-1G>C	splice-3
DCHS1	11	6645264	G	A/G	NM_003737.2:c.7643C>T	Missense
DCHS1	11	6662466	C	C/T	NM_003737.2:c.379G>A	Missense
C11orf40	11	4594558	-	-/G	NM_144663.1:c.286_287insC	Frameshift
C11orf40	11	4598956	C	C/T	NM_144663.1:c.95G>A	Nonsense

RNA test

The consequences of the splice site variant c.509-1G>C were studied by RT-PCR analysis of leucocyte RNA from patients in family A. Besides the expected normal fragment of 550 bp, a longer band (697 bp) was observed in the heterozygous carriers (Figure 3C). The bands were isolated from the agarose gel, and Sanger sequencing revealed that intron 5 was retained in the longer RNA fragment (r.[508_509ins508+1_509-1;509-1g>c]), indicating inactivation of the splice site by the c.509-1G>C change. The insertion of intron 5 caused a premature termination of translation 29 amino acids (p.Val170Glyfs*29).downstream of the splice site. Only one band was found for the RT-PCR around the c.1397T>G variant, indicating it had no effect on RNA processing (Figure 3D). The variant allele is therefore "active" and likely generating a p.Val466Gly missense variant at protein level.

TPP1 enzyme activity

TPP1 enzyme activity in leucocytes and fibroblasts of SCAR7 patients of family A, patient B-II.1 and several Dutch CLN2 patients (C-O) is described in Table 1. In all affected individuals, deficient activity of tripeptidyl peptidase 1 was found. There is however considerable overlap in enzyme activity in leucocytes from CLN2 and SCAR7 patients. For the affected individuals in family A, residual activity in leucocytes varied from 8 to 32 nmol/h/mg protein. The mean residual activity is 15% of the lowest control in family A and 10% in patient B-II.1 (4 nmol/h/mg protein). For CLN2 patients (C-O) the mean residual activity in leucocytes was 9% of the lowest control (2.16-23.5 nmol/h/mg protein), almost comparable with the SCAR7 patients. In fibroblasts however, this difference is more substantial with a residual enzyme activity of 0.4% of the lowest control in CLN2 patients and of 5% in patient A-III.2 (Table 1) from family A. The mean TPP1 activity in leucocytes of carriers with the splice site variant, c.509-1G>C was 132 nmol/h/mg protein and 139 nmol/h/mg protein in carriers with the missense variant, c.1397T>G, both as expected within the normal range (data not shown, but available upon request).

Electron microscopy

Electron microscopy of a skin biopsy tissue of one of the patients of family A (A-III.2) did not show the typical curvilinear profiles seen in patients with a typical CLN2 phenotype but some

granular osmiophilic deposits (GROD) and fingerprint profiles were noticed (Supp. Figure S1). Electron microscopy of a skin biopsy tissue of patient B-II.1 was normal; in particular there were no changes compatible with neuronal ceroid lipofuscinosis.

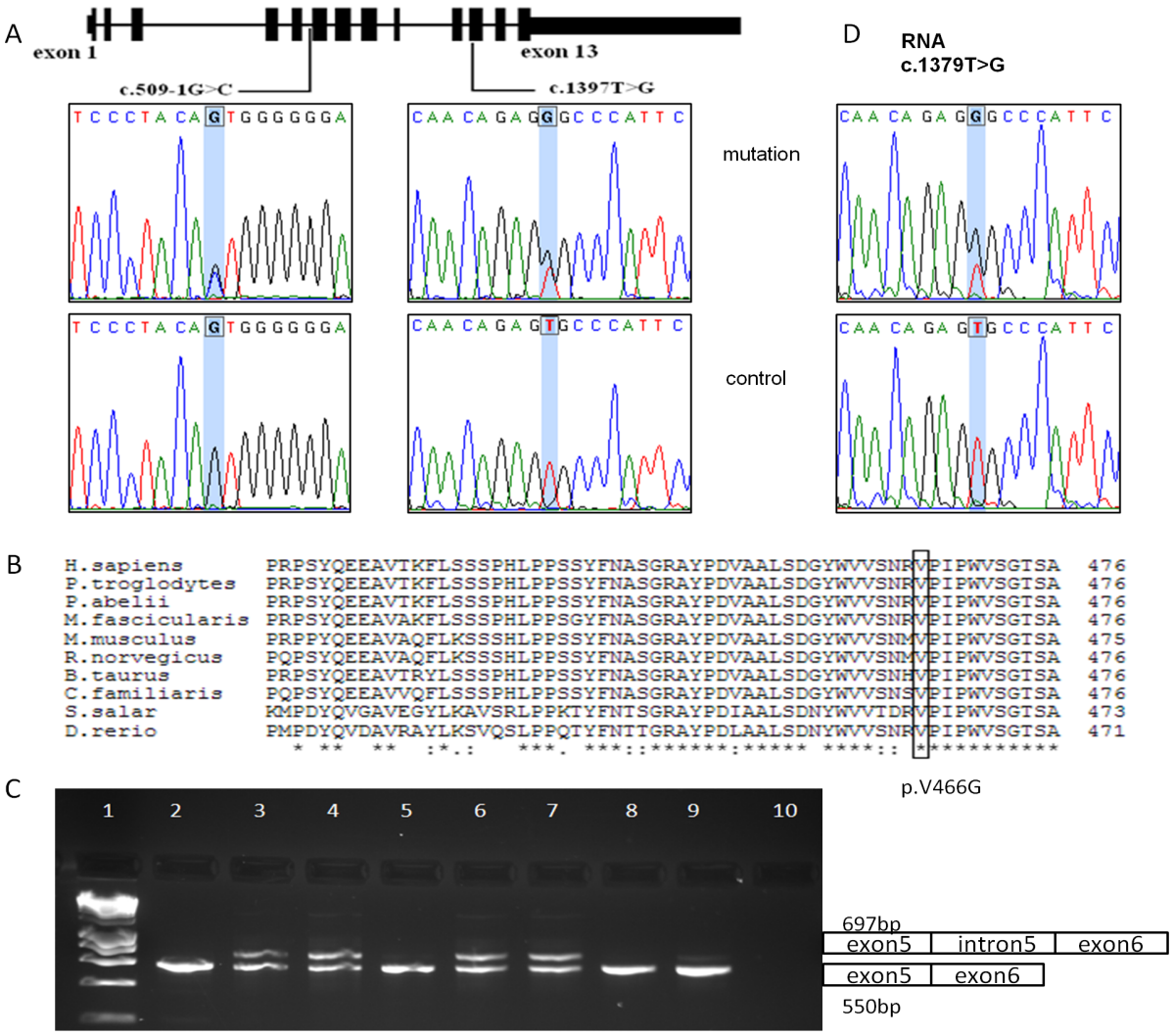


Figure 3. *TPP1* gene structure, variant analysis by Sanger sequencing and RT-PCR results. (A) *TPP1* gene structure with c.509-1G>C variant in exon 5 and c.1397T>G variant in exon 11 (top). Sanger sequencing confirmation of the two variants in patient sample compared to control (bottom). (B) Alignment of *TPP1* protein homologs. The amino acid Valine (V) at position 466, replaced by the missense variant in family A and patient B is highly conserved in species and highlighted in a rectangle (C) RT-PCR products from affected patients III-3, III-2 with compound heterozygous variants (c.509-1G>C, c.1397T>G) (lanes 3 and 6), un-affected III-4, III-8 (c.509-1G>C carriers) (lanes 7 and 4), unaffected III-5, III-7 (c.1397T>G carriers) (lanes 2 and 5), controls (lanes 8 and 9), and a blank (lane 10) analyzed in 2% agarose gel. Lane 1 size standard (Smartladder, Eurogentec). A longer band is observed in all heterozygous carriers of c.509-1G>C. (D) The Sanger sequencing result of c.1397T>G from RT-PCR, the variant is a true missense change.

Discussion

Exome sequencing in family A and Sanger sequencing, as part of a diagnostic workup in patient B-II.1, showed these compound heterozygous variants in *TPP1* as the cause of SCAR7. Defects in *TPP1* have previously been linked to CLN2. In the majority of cases, the age of onset of CLN2 is late infantile, between 2 to 4 years. It can also be infantile with onset before the age of 1 year (Ju, et al., 2002; Simonati, et al., 2000) or even juvenile, with disease onset between 6 and 10 years and a more protracted course (Bessa, et al., 2008; Elleder, et al., 2008; Hartikainen, et al., 1999; Kohan, et al., 2009; Sleat, et al., 1999; Wisniewski, et al., 1999). Developmental studies of TPP1 distribution in human brain and visceral organs, showed that the enzyme is not expressed in the developing neurons of the human fetus (Kida, et al., 2001; Kurachi, et al., 2001; Oka, et al., 1998). It appears in the neurons of the central nervous system at the age of 5 months and expression increases gradually to reach stable levels at the age of 3 years. This finding may explain why CLN2 and SCAR7 do not start in the early beginning of life.

The lipopigment pattern seen most often in CLN2 consists of curvilinear profiles, detectable by electron microscopy, in various cell types. There is a relationship between *TPP1* mutations, TPP1 activity, and curvilinear profiles (Mole, et al., 2005; Sleat, et al., 1999). In SCAR7 patients A.III-2 and B.II-1, however, the skin biopsy showed no curvilinear profiles. In CLN2 with a later onset and more protracted course, curvilinear profiles are not the only ultrastructural features found, also fingerprint profiles and GROD may appear (Wisniewski, et al., 1999). The skin biopsy of A.III-3 showed some GROD and fingerprint profiles, but no ultra structural features were found in patient B.II-1. It is suggested that there is a spectrum of ultra structural features in diseases caused by mutations in *TPP1,* ranging from curvilinear profiles in classic CLN2, mixed ultra structural features consisting of curvilinear- and fingerprint profiles and GROD in CLN2 with a late onset and protracted course, to only some GROD and fingerprint profiles or even absence of ultra structural features in SCAR7. Ultra structural findings show a correlation with the severity and course of the phenotype due to TPP1 deficiency, as was also shown before in mice (Sleat, et al., 2008).

The *TPP1* gene is composed of 13 exons. It encodes a member of the sedolisin family of serine proteases, tripeptidyl peptidase 1, mainly expressed in the lysosome and melanosome. The protease cleaves the N-terminal tripeptides from substrates, and it has a weak endopeptidase

activity. It is synthesized as a catalytically-inactive enzyme which is activated and auto-proteolyzed upon acidification. The TPP1 protein starts with a 19 amino acid signal peptide (Lin, et al., 2001; Sleat, et al., 1997) and a pro peptide of 176 amino acids, which will be removed in the mature form. The last part of the protein consists of the 368 amino acid tripeptidyl-peptidase 1 chain. The majority of the mutations in *TPP1* are located in the tripeptidyl-peptidase 1 domain, while only three mutations are localized in the propeptide domain, and none in signal peptide section. However, in the general population (data derived from the 1000 Genomes Project and GoNL project, including 500 unrelated Dutch individuals), the number of variations in propeptide domain and tripeptidyl-peptidase 1 chain is comparable (Supp. Table S5). This suggests that the propeptide section is more tolerant to variation, possibly due to the fact that this part of the protein is removed from the mature form and therefore may not have a significant effect on the function of the protein. The low variation in the signal peptide indicates the significance of that part of the protein. Without a recognizable signal peptide, the protein will not reach its destination nor will it be cleaved, and its function will probably be lost.

To facilitate genotype – phenotype studies, we examined the *TPP1* database, summarizing all variants published in the literature (http://www.ucl.ac.uk/ncl/cln2.shtml, date August 14, 2012). The mutations reported so far in relation to *TPP1*, including missense, nonsense, insertion, deletion, splice site, were scattered throughout the whole gene. Several mutations are recurrent, such as the stop codon p.(Arg208*), a splice site variant c.509-1G>C, which is present in patients of family A and B, and a Newfoundland founder variant p.(Gly284Val) (Ju, et al., 2002). The missense mutation found in patients of family A and patient B.II-1, c.1397T>G, Val466Gly, was not reported before, but showed conservation in evolution and is located in the peptidase region of the protein.

Genotype – phenotype relations in the neuronal ceroid lipofuscinoses have been reviewed and tested in Chinese hamster cells (Kousi, et al., 2012; Mole, et al., 2005; Walus, et al., 2010). A loss of TPP1 function will cause the CLN2-late infantile, which means the TPP1 enzyme activity will be extremely reduced or absent. By examining the variant spectrum of *TPP1*, some variants will evidently truncate the protein, while some missense and in-frame insertion or deletion variants are observed to impair enzyme function of the protein. We have summarized the *TPP1* genotype with at least one missense variant reported in the literature in Supp. Table S6, and link

the experimental data (Guhaniyogi, et al., 2009; Lin and Lobel, 2001; Pal, et al., 2009; Walus, et al., 2010) and prediction tools (Desmet, et al., 2009) to elucidate the potential effect of those mutations. The majority of variants clearly inactivate the gene by creating an early stop of translation by introducing either a nonsense, or a frameshift variant. Some mutations, like c.1266G>C and c.380G>A, are located near the splice sites, and may therefore alter splicing and disrupt the reading frame (Chao, et al., 2010; Sun, et al., 2010). Other variants are predicted to generate a new splice site by Human Splicing Finder (Desmet, et al., 2009) (http://www.umd.be/HSF/, date July 06, 2012). For those mutations, it is worthwhile to study RNA to verify the predicted truncating effect. Another category of variant are those located within the active site of the protein (c.827A>T (Kohan, et al., 2009), c.1424C>T (Sleat, et al., 1999)), so even a minor change might affect the function of the protein significantly.

In patients of family A and patient B-II.1, RNA analysis of the splice site variant c.509-1G>C, showed that retention of intron 5 in the reading frame generates a premature stop codon, leading to haplo-insufficiency through nonsense mediated decay. For the amino acid changes from Valine to Glycine, unlike other CLN2 "missense" mutations, prediction tools do not show that it will produce a severe splicing alteration (HSF), indicating the protein product translated from this allele might still work actively.

Although there is considerable overlap in mean enzyme activity in leucocytes from CLN2 (9%) and SCAR7 patients (10-15%), in fibroblasts it differs about a factor 10 (0.4% in CLN2 patients and 5% in SCAR7 patient A.III-2), although the number of patients studied is very small (Table 1). The low activity in blood and especially in fibroblasts give an indication of the overall TPP1 enzyme activity in the central nervous system. Sleat et al showed in CLN2 mutant mice that low TPP1 levels attenuated disease. Compound heterozygosity for a null allele and a presumed hypomorphic p.Arg447His missense variant resulted in a later onset and a protracted disease with survival into the third or fourth decade of life. Mice homozygous for this hypomorphic mutation, showed locomotor deficits at a later age, with a slower disease progression, compared to homozygous null allele mutated mice and compound heterozygote mice and also showed a greatly extended life span, approaching that of normal mice. The brains of these mice showed approximately 3% of normal TPP1 activity compared to homozygous null allele mutated mice expressing 0.2% of normal levels (Sleat, et al., 2008). A semantic data mining approach

comparing model organism and clinical phenotype data (Chen, et al., 2012) identified this homozygous TPP1 hypomorphic mouse as the 25th best match for SCAR7 out of the 25,141 mouse models annotated at MGI (Damian Smedley, personal communication). The mouse model shows the clinical features through the presence of ataxia, Purkinje cell degeneration, neurodegeneration, tremor, and audiogenic seizures, which highly resemble the SCAR7 phenotype.

We infer the following genotype – phenotype correlation: loss of function variants abolishing TPP1 enzyme activity lead to CLN2, while variants that diminish TPP1 enzyme activity lead to SCAR7 (Table 3). Therapeutic approaches, causing a small increase in TPP1 enzyme activity in brain, might change the course of the disease and extend the lifespan of CLN2 patients by pushing them towards a more SCAR7-like phenotype, but higher levels will be required to cure the disease. Further investigations are needed to confirm this hypothesis.

Table 3. The phenotypes and genotypes of patients with *TPP1* mutations

	CLN2, late infantile	CLN2, juvenile	SCAR7
General	very severely affected	less severely affected	mild phenotype and protracted course
Age of onset	2-4 years	10-Jun	childhood or teenage
Age of death	5-15 years	> 12-40 years	> 60 years
Clinical findings	seizures, dementia, visual loss, ataxia and cerebral atrophy	seizures, dementia, visual loss, ataxia and/or cerebral atrophy, protracted course	Cerebellar ataxia, pyramidal signs, deep sensory loss, cerebellar atrophy
TPP1 enzyme activity	extremely low or none	residual or very low	Residual
Ultrastructural features (EM)	curvilinear bodies	curvilinear bodies/GROD/fingerprint profiles	some GROD/fingerprint profiles/none
Alleles	null/null	null/partial affected	null/minor modification

To conclude, SCAR7 is caused by compound heterozygous variants in *TPP1*. The genetic background of cerebellar ataxias are even more heterogeneous than the neuronal ceroid lipofuscinoses with a still growing number of subtypes and we here add *TPP1* to the list of genes implicated in the autosomal recessive ataxias. The phenotype associated with *TPP1* variants is expanded now by an autosomal recessive form of slowly progressive cerebellar ataxia. Diagnostic work-up for unexplained spinocerebellar ataxias should thus include analysis of TPP1 enzyme activity, particularly if the family history or the age of onset is suggestive of an

autosomal recessive disorder. Other features that could suggest *TPP1* mutations, i.e. CNL2 features such as visual regression, epilepsy or curvilinear profiles in a skin biopsy, can be absent. This finding again illustrates the sometimes unexpected clinical spectrum of variants in known genes. We will encounter this phenomenon with increasing frequency using new techniques such as whole exome sequencing.

Acknowledgments

We would like to thank the patients for their kind participation and Leiden Genome Technology Center (LGTC), Laboratory for Diagnostic Genome Analysis (LDGA) and Department of Clinical Genetics, Erasmus Medical Center Rotterdam for technical support. Yu Sun was supported by China Scholarship Council.

References

Bessa C, Teixeira CA, Dias A, Alves M, Rocha S, Lacerda L, Loureiro L, Guimaraes A, Ribeiro MG. 2008. CLN2/TPP1 deficiency: the novel mutation IVS7-10A>G causes intron retention and is associated with a mild disease phenotype. Mol Genet Metab 93:66-73.

Breedveld GJ, van Wetten B, te Raa GD, Brusse E, van Swieten JC, Oostra BA, Maat-Kievit JA. 2004. A new locus for a childhood onset, slowly progressive autosomal recessive spinocerebellar ataxia maps to chromosome 11p15. J Med Genet 41:858-66.

Chao SC, Chen JS, Tsai CH, Lin JM, Lin YJ, Sun HS. 2010. Novel exon nucleotide substitution at the splice junction causes a neonatal Marfan syndrome. Clin Genet 77:453-63.

Chen CK, Mungall CJ, Gkoutos GV, Doelken SC, Kohler S, Ruef BJ, Smith C, Westerfield M, Robinson PN, Lewis SE and others. 2012. MouseFinder: Candidate disease genes from mouse phenotype data. Hum Mutat 33:858-66.

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.

Elleder M, Dvorakova L, Stolnaja L, Vlaskova H, Hulkova H, Druga R, Poupetova H, Kostalova E, Mikulastik J. 2008. Atypical CLN2 with later onset and prolonged course: a neuropathologic study showing different sensitivity of neuronal subpopulations to TPP1 deficiency. Acta Neuropathol 116:119-24.

Guhaniyogi J, Sohar I, Das K, Stock AM, Lobel P. 2009. Crystal structure and autoactivation pathway of the precursor form of human tripeptidyl-peptidase 1, the enzyme deficient in late infantile ceroid lipofuscinosis. J Biol Chem 284:3985-97.

Hartikainen JM, Ju W, Wisniewski KE, Moroziewicz DN, Kaczmarski AL, McLendon L, Zhong D, Suarez CT, Brown WT, Zhong N. 1999. Late infantile neuronal ceroid lipofuscinosis is due to splicing mutations in the CLN2 gene. Mol Genet Metab 67:162-8.

Ju W, Zhong R, Moore S, Moroziewicz D, Currie JR, Parfrey P, Brown WT, Zhong N. 2002. Identification of novel CLN2 mutations shows Canadian specific NCL2 alleles. J Med Genet 39:822-5.

Kida E, Golabek AA, Walus M, Wujek P, Kaczmarski W, Wisniewski KE. 2001. Distribution of tripeptidyl peptidase I in human tissues under normal and pathological conditions. J Neuropathol Exp Neurol 60:280-92.

Kohan R, Cismondi IA, Kremer RD, Muller VJ, Guelbert N, Anzolini VT, Fietz MJ, Ramirez AM, Halac IN. 2009. An integrated strategy for the diagnosis of neuronal ceroid lipofuscinosis types 1 (CLN1) and 2 (CLN2) in eleven Latin American patients. Clin Genet 76:372-82.

Kousi M, Lehesjoki AE, Mole SE. 2012. Update of the mutation spectrum and clinical correlations of over 360 mutations in eight genes that underlie the neuronal ceroid lipofuscinoses. Hum Mutat 33:42-63.

Kurachi Y, Oka A, Itoh M, Mizuguchi M, Hayashi M, Takashima S. 2001. Distribution and development of CLN2 protein, the late-infantile neuronal ceroid lipofuscinosis gene product. Acta Neuropathol 102:20-6.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9.

Lin L, Lobel P. 2001. Expression and analysis of CLN2 variants in CHO cells: Q100R represents a polymorphism, and G389E and R447H represent loss-of-function mutations. Hum Mutat 18:165.

Lin L, Sohar I, Lackland H, Lobel P. 2001. The human CLN2 protein/tripeptidyl-peptidase I is a serine protease that autoactivates at acidic pH. J Biol Chem 276:2249-55.

Matilla-Duenas A. 2012. The Ever Expanding Spinocerebellar Ataxias. Editorial. Cerebellum

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and others. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-303.

Mole SE, Williams RE, Goebel HH. 2005. Correlations between genotype, ultrastructural morphology and clinical phenotype in the neuronal ceroid lipofuscinoses. Neurogenetics 6:107-26.

Oka A, Kurachi Y, Mizuguchi M, Hayashi M, Takashima S. 1998. The expression of late infantile neuronal ceroid lipofuscinosis (CLN2) gene product in human brains. Neurosci Lett 257:113-5.

Pal A, Kraetzner R, Gruene T, Grapp M, Schreiber K, Gronborg M, Urlaub H, Becker S, Asif AR, Gartner J and others. 2009. Structure of tripeptidyl-peptidase I provides insight into the molecular basis of late infantile neuronal ceroid lipofuscinosis. J Biol Chem 284:3976-84.

Sailer A, Houlden H. 2012. Recent advances in the genetics of cerebellar ataxias. Curr Neurol Neurosci Rep 12:227-36.

Sailer A, Scholz SW, Gibbs JR, Tucci A, Johnson JO, Wood NW, Plagnol V, Hummerich H, Ding J, Hernandez D and others. 2012. Exome sequencing in an SCA14 family demonstrates its utility in diagnosing heterogeneous diseases. Neurology 79:127-31.

Santavuori P. 1988. Neuronal ceroid-lipofuscinoses in childhood. Brain Dev 10:80-3.

Santen GW, Aten E, Sun Y, Almomani R, Gilissen C, Nielsen M, Kant SG, Snoeck IN, Peeters EA, Hilhorst-Hofstee Y and others. 2012. Mutations in SWI/SNF chromatin remodeling complex gene ARID1B cause Coffin-Siris syndrome. Nat Genet 44:379-80.

Simonati A, Santorum E, Tessa A, Polo A, Simonetti F, Bernardina BD, Santorelli FM, Rizzuto N. 2000. A CLN2 gene nonsense mutation is associated with severe caudate atrophy and dystonia in LINCL. Neuropediatrics 31:199-201.

Sleat DE, Donnelly RJ, Lackland H, Liu CG, Sohar I, Pullarkat RK, Lobel P. 1997. Association of mutations in a lysosomal protein with classical late-infantile neuronal ceroid lipofuscinosis. Science 277:1802-5.

Sleat DE, El-Banna M, Sohar I, Kim KH, Dobrenis K, Walkley SU, Lobel P. 2008. Residual levels of tripeptidyl-peptidase I activity dramatically ameliorate disease in late-infantile neuronal ceroid lipofuscinosis. Mol Genet Metab 94:222-33.

Sleat DE, Gin RM, Sohar I, Wisniewski K, Sklower-Brooks S, Pullarkat RK, Palmer DN, Lerner TJ, Boustany RM, Uldall P and others. 1999. Mutational analysis of the defective protease in classic late-infantile neuronal ceroid lipofuscinosis, a neurodegenerative lysosomal storage disorder. Am J Hum Genet 64:1511-23.

Sun Y, Almomani R, Aten E, Celli J, van der Heijden J, Venselaar H, Robertson SP, Baroncini A, Franco B, Basel-Vanagaite L and others. 2010. Terminal osseous dysplasia is caused by a single recurrent mutation in the FLNA gene. Am J Hum Genet 87:146-53.

Van Diggelen OP, Keulemans JL, Kleijer WJ, Thobois S, Tilikete C, Voznyi YV. 2001. Pre- and postnatal enzyme analysis for infantile, late infantile and adult neuronal ceroid lipofuscinosis (CLN1 and CLN2). Eur J Paediatr Neurol 5 Suppl A:189-92.

Vermeer S, van de Warrenburg BP, Willemsen MA, Cluitmans M, Scheffer H, Kremer BP, Knoers NV. 2011. Autosomal recessive cerebellar ataxias: the current state of affairs. J Med Genet 48:651-9.

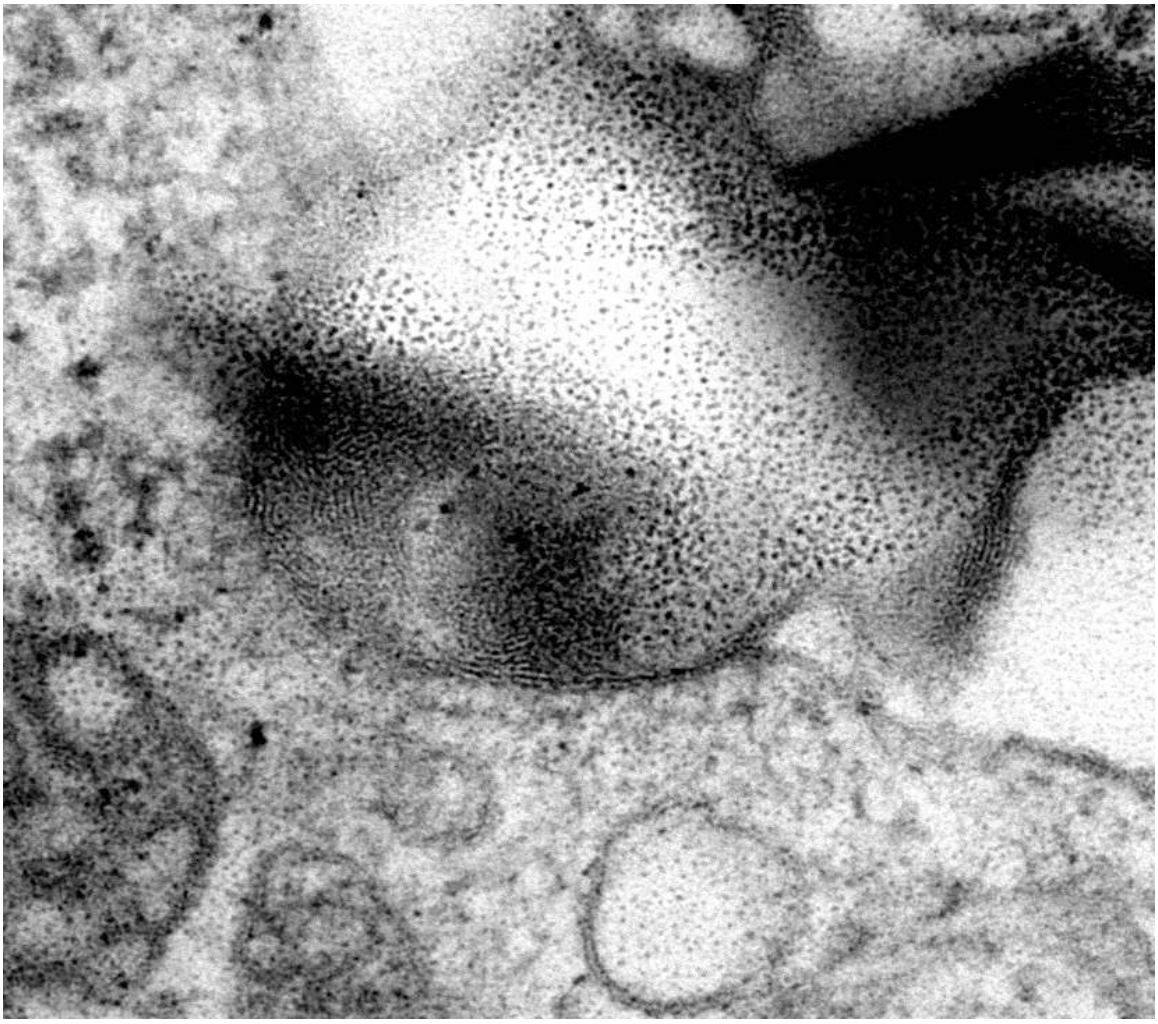
Walus M, Kida E, Golabek AA. 2010. Functional consequences and rescue potential of pathogenic missense mutations in tripeptidyl peptidase I. Hum Mutat 31:710-21.

Williams RE, Boyd S, Lake BD. 1999. Ultrastructural and electrophysiological correlation of the genotypes of NCL. Mol Genet Metab 66:398-400.

Williams RE, Mole SE. 2012. New nomenclature and classification scheme for the neuronal ceroid lipofuscinoses. Neurology 79:183-91.

Wisniewski KE, Kaczmarski A, Kida E, Connell F, Kaczmarski W, Michalewski MP, Moroziewicz DN, Zhong N. 1999. Reevaluation of neuronal ceroid lipofuscinoses: atypical juvenile onset may be the result of CLN2 mutations. Mol Genet Metab 66:248-52.

Supplementary data:



Supplementary Figure 1. Electron microscopy of a skin biopsy tissue of patient A.III-2 which shows granular osmiophilic deposits (GROD) and fingerprint profiles. The magnification of EM image was 20,000 times.

Supplementary. Table S1. SCAR types

SCAR	Locus	Gene	Clinical findings other than spinocerebellar ataxia
1	9q34.13	*SETX*	onset 10-25 yr, progressive, ocular apraxia, axonal neuropathy, tremor, pyramidal signs, elevated AFP
2	9q34-qter	*?*	congenital, ID, small head, cataracts, pyramidal signs, intention tremor, short stature
3	6p23-p21	*?*	early-onset, hearing impairment, optic atrophy
4	1p36	*?*	onset 3rd decade, progressive, pyramidal signs, myoclonic jerks, fasciculations, impaired joint position sense
5	15q25.3	*ZNF592*	Congenital, severe psychomotor retardation, short stature, pyramidal signs, microcephaly, optic atrophy, speech defect, abnormal osmiophilic pattern of skin vessels (CAMOS)
6	20q11-q13	*?*	onset in infancy, nonprogressive; delayed motor and speech development, no ID, hypotonia, pes planus
7	11p15	*TPP1*	childhood-onset, slowly progressive
8	6q25.1-q25.2	*SYNE1*	late-onset, slow progression
9	1q42.13	*ADCK3*	childhood onset, progressive, cerebellar atrophy, seizures, developmental delay, hyperlactatemia
10	3p22.1	*ANO10*	onset teenage-young adulthood, hyperreflexia, nystagmus, atrophy lower limbs with fasciculations tortuosity of the conjunctival vessels, ID, pes cavus
11	1q32.2	SYT14	onset 6th decade, progressive, ID
12	16q21-q23	?	onset early-childhood, generalized seizures, delayed psychomotor development, ID
13	6q24.3	GRM1	onset infancy, slowly progressive, ID with poor or absent speech, hyperreflexia, eye movement abnormalities

Supplementary Table S2. Primer list

Primer	Sequence (5'-3')	DNA/RNA
DCHS2_ex21_F	GTCAGCTGCAGCCACTGTTA	DNA
DCHS2_ex21_R	TGTGGCTGTGACTGAAGACC	DNA
DCHS2_ex2_F	GGTGCCAAAAGCTGTATGCT	DNA
DCHS2_ex2_R	TGCAGATTGATGAGGAGCAG	DNA
TPP1_ex11_F	AGGGGTTCTAGGTGCAAGGT	DNA
TPP1_ex11_R	CCAGGAACCTTTCCTCATCA	DNA
TPP1_ex5-6_F	TGTTATTGCTGGTGCCAGAG	DNA
TPP1_ex5-6_R	CAGGGATGCTCAGAGGTAGC	DNA
NOP56_ex11_F	AAGGAGTCCTCAGAGCACCA	DNA
NOP56_ex11_R	CCACTGTGAAACACGACCAC	DNA
TPP1_RNA_ex5_F2	GTCTCACCTTTGCCCTGAGA	RNA
TPP1_RNA_ex6_R2	AGGAACTGGGCACAGGCTT	RNA
TPP1_RNA_ex11_F	CTGATGGCTACTGGGTGGTC	RNA
TPP1_RNA_ex12-13_R	AGCCACGGGTTACATCAAAG	RNA

Supplementary Table S3. The overview of the exome sequencing

Patient	*A III-2*
total reads	77479996
Read 1	38739998
read length 1 (nt)	100
Read 2	38739998
read length 2 (nt)	100
aligned reads	75802233
PCT_aligned_reads	97.83%
properly paired	74211004
PCT_properly_paired	95.78%
PERCENT_DUPLICATION	34.65%
MEAN_BAIT_COVERAGE	48.930747
PCT_TARGET_BASES_10X	86.41%
PCT_SELECTED_BASES	82.89%
FOLD_ENRICHMENT	38.026774
ZERO_CVG_TARGETS_PCT	4.39%
FOLD_80_BASE_PENALTY	2.825506

Supplementary Table S4. The genotype of six variants around the *TPP1* mutations, and the inferred haplotypes in family A

Gene	OR56A3	TPP1	TPP1	TPP1	DCHS1	*DCHS1*
Variation in Transcript	c.13C>T	c.1542A>T	c.1397T>G	c.509-1G>C	c.7643C>T	c.379G>A
Genomic Location	g.5968589C>T	g.6636106T>A	g.6636430A>C	g.6638385C>G	g.6645264G>A	g.6662466C>T
A.III-2	C/T	T/A	A/C	C/G	G/A	C/T
A.III-3	C/T		A/C	C/G	G/A	C/T
A.III-6			A/C	C/G		
A.III-10			A/C	C/G		
A.III-9			A/C	C/G		
A.III-1	C/C		A/A	C/C	G/G	C/C
A.III-8	C/C		A/A	C/G	G/G	C/C
A.III-7	C/T		A/C	C/C	G/A	C/T
A.III-5	C/T		A/C	C/C	G/A	C/T
A.IV-1	C/C		A/A	C/C	G/G	C/C
A.III-11	C/C		A/A	C/G	G/G	C/C
A.II-1			A/A	C/C		
A.III-4			A/A	C/G		
B.II-1	C/C	T/A	A/C	C/G	G/A	C/C
B.I-1	C/C	T/A	A/C	C/C	G/A	C/C
Haplotype allele 1 in patients of family A	T		C	C	A	T
haplotype allele 2 in patients of family A	C		A	G	G	C

Supplementary Table S5. The variants found in normal populations

	1000 Genomes		GoNL	
	Nonsynonymous	Missense	Nonsynonymous	Missense
Signal peptide	1	0	0	1
Propeptide	4	6	0	3
Protease	4	6	5	2

Supplementary Table S6. Reported genotypes in *TPP1* with at least one missense, inframe change or intronic variant and information from literatures and prediction tools. When the phenotype is not available, the background is marked grey. Here CLN2=CLN2, late infantile, JNCL=CLN2, juvenile, INCL= CLN2, infantile. SD = splice donor site, SA = splice acceptor site, BS = splice branch site, HSF = predicted by Human Splicing Finder.

Missense Allele 1	protein domain	Allele 2	phenotype	Splicing effect (Experiment and prediction)	Functional Study by Walus 2010			Structure study by Pal 2008	Reference
					TPP1 activity[1]	Lysosomal transport	Half-lives of proenzymes[2]		
c.225A>G, p.(Gln75Gln)	Propeptide	c.509-1G>C	CLN2	Splice defect according to the article (prediction)					Sleat 1999
c.225A>G, p.(Gln75Gln)	Propeptide	c.1678_1679delCT, p.(Leu560Thrfs*47)	NA	Splice defect according to the article (prediction)					Kousi 2012
c.184T>A, p.(Ser62Thr)	Propeptide	?	NA	HSF- new BS					Kousi 2012
c.229G>A, p.(Gly77Arg)	Propeptide	c.509-1G>C	CLN2	Last nucleotide of exon 3	1	PA	4.5		Sleat 1999
c.229G>A, p.(Gly77Arg)	Propeptide	c.640C>T, p.(Gln214*)	CLN2	Last nucleotide of exon 3	1	PA	4.5		Kousi 2012
c.229G>A, p.(Gly77Arg)	Propeptide	c.790C>T, p.(Gln264*)	CLN2	Last nucleotide of exon 3	1	PA	4.5		Kousi 2012

c.229G>A, p.(Gly77Arg)	Propeptide	c.1062del, p.(Leu355Serfs*72)	CLN2	Last nucleotide of exon 3	1	PA	4.5		Kousi 2012
c.299A>G, p.(Gln100Arg)	Propeptide	c.1266+5G>A	NA	HSF-enhancer interrupted					Kousi 2012
c.1266+5G>A	tripeptidyl-peptidase 1 chain	c.299A>G, p.(Gln100Arg)	NA	HSF- broken +new SD					Kousi 2012
c.380G>A, p.(Arg127Gln)	Propeptide	c.380G>A, p.(Arg127Gln)	NA	Last nucleotide of exon 4	43.6	N	1.7	NA, surface exposure	Kousi 2012
c.380G>A, p.(Arg127Gln)	Propeptide	c.622C>T, p.(Arg208*)	CLN2	Last nucleotide of exon 4	43.6	N	1.7	NA, surface expo-sure	Steinfeld 2002
c.380G>A, p.(Arg127Gln)	Propeptide	c.509-1G>C	CLN2	Last nucleotide of exon 4	43.6	N	1.7	NA, surface exposure	Zhong 2000
c.380G>A, p.(Arg127Gln)	Propeptide	?	NA	Last nucleotide of exon 4	43.6	N	1.7	NA, surface exposure	Kousi 2012
c.381-17_-4del	Propeptide	c.229G>T, p.(Gly77*)	CLN2	HSF-SA Broken					Chang 2012
c.457T>C, p.(Ser153Pro)	Propeptide	?	NA, JNCL		NA	NA	NA		Cillaud 1999; Mole 2001
c.524G>A, p.(Arg175His)	Propeptide	?	NA	HSF-enhancer interrupted					Kousi 2012
c.605C>T, p.(Pro202Leu)	Tripeptidyl-peptidase 1 chain	?	NA		0	A	6.5, homodimer of proenzyme		Mole 2001
c.616C>T, p.(Arg206Cys)	Tripeptidyl-peptidase 1 chain	?	NA		0.7	A	2.3		Mole 2001
c.616C>T, p.(Arg206Cys)	Tripeptidyl-peptidase 1 chain	c.616C>T, p.(Arg206Cys)	CLN2		0.7	A	2.3		Tessa 2000
c.617G>A, p.(Arg206His)	Tripeptidyl-peptidase 1 chain	?	NA	HSF-SD Enhancer disrupted					Kousi 2012

c.625T>C, p.(Tyr209His)	Tripeptidyl-peptidase 1 chain	c.625T>C, p.(Tyr209His)	NA	HSF- new BS					Kousi 2012
c.646G>A, p.(Val216Met)	Tripeptidyl-peptidase 1 chain	c.1551+1G>A	CLN2	HSF-SD broken Enhancer disrupted	NA	NA	NA		Wang 2011
c.650G>T, p.(Gly217Val)	Tripeptidyl-peptidase 1 chain	c.640C>T, p.(Gln214*)	CLN2	HSF-SD broken	NA	NA	NA		Chang 2012
c.797G>A, p.(Arg266Gln)	Tripeptidyl-peptidase 1 chain	c.1015C>T, p.(Arg339Gln)	NA	HSF-new SA	NA	NA	NA		Kousi 2012
c.827A>T, p.(Asp276Val)	Tripeptidyl-peptidase 1 chain	c.827A>T, p.(Asp276Val)	CLN2		NA	NA	NA	Active site	Kohan 2009
c.827A>T, p.(Asp276Val)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	CLN2		NA	NA	NA	Active site	Kohan 2009
c.829G>A, p.(Val277Met)	Tripeptidyl-peptidase 1 chain	?	CLN2		0	PA	6.8, homodimer of proenzyme	Might affect active site	Ju 2002
c.833A>C, p.(Gln278Pro)	Tripeptidyl-peptidase 1 chain	?	CLN2	HSF- SA broken disrupt an alpha-helix	NA	NA	NA	Might affect active site	Ju 2002
c.843G>T, p.(Met281Ile)	Tripeptidyl-peptidase 1 chain	?	NA	HSF-SD broken					Kousi 2012
c.851G>T, p.(Gly284Val)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	CLN2	HSF-SD broken, enhancer disrupted	0.4	A	3.9, homodimer of proenzyme		Zhong 2000; Ju 2002
c.851G>T, p.(Gly284Val)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	CLN2	HSF- SD broken	0.4	A	3.9, homodimer of proenzyme		Ju 2002
c.851G>T, p.(Gly284Val)	Tripeptidyl-peptidase 1 chain	c.851G>T, p.(Gly284Val)	CLN2	HSF- SD broken	0.4	A	3.9		Ju 2002
c.851G>T, p.(Gly284Val)	Tripeptidyl-peptidase 1 chain	?	CLN2	HSF- SD broken	0.4	A	3.9, homodimer of proenzyme		Ju 2002

c.851G>T, p.(Gly284Val)	Tripeptidyl-peptidase 1 chain	Uncharacterised 1bp deletion in exon 7	CLN2	HSF- SD broken	0.4	A	3.9, homodimer of proenzyme		Ju 2002
c.857A>G, p.(Asn286Ser)	Tripeptidyl-peptidase 1 chain	c.857A>G, p.(Asn286Ser)	CLN2	HSF-new SA	0	PA	3.1, homodimer of proenzyme	Loss N-linkd glycos-ylation, surface exposure	Steinfeld 2002
c.860T>A, p.(Ile287Asn)	Tripeptidyl-peptidase 1 chain	?	NA		0.1	PA	3.8, homodimer of proenzyme		Sleat 1999
c.887-10A>G	Tripeptidyl-peptidase 1 chain	c.196C>T, p.(Gln66*)/c.89+4A>G	JNCL, milder than CLN2	New SA by RT-PCR, 9 nt inserted between c.886_887insAAAATCCAG	NA	NA	NA	NA	Kohan 2009
c.887-10A>G, p.(Gly296delinsGluAsnProGly)	Tripeptidyl-peptidase 1 chain	c.887-10A>G, p.(Gly296delinsGluAsnProGly)	JNCL, much milder than CLN2	New SA by RT-PCR, 9 nt inserted between c.886_887insAAAATCCAG	NA	NA	NA	NA	Bessa 2008
c.887-18A>G, p.(Gly296delinsGlyLysLysLysAsnPoGly)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	CLN2	Splice defect according to the article (prediction)					Sleat 1999
c.984_986del, p.(Asp328del)	Tripeptidyl-peptidase 1 chain	?	NA	splice site, according to the article HSF-enhancer dirupted					Kousi 2012
c.987_989delinsCTC, p.(Glu329_Asp330delinsAspSer)	Tripeptidyl-peptidase 1 chain	?	NA	HSF-SA broken	NA	NA	NA	NA	Kousi 2012
c.1015C>T, p.(Arg339Gln)	Tripeptidyl-peptidase 1 chain	c.797G>A, p.(Arg266Gln)	NA	HSF-new SA					Kousi 2012
c.797G>A, p.(Arg266Gln)	Tripeptidyl-peptidase 1	c.1015C>T, p.(Arg339Gln	NA	HSF-new SA					Kousi 2012

	chain	)							
c.1015C>T, p.(Arg339Gln)	Tripeptidyl-peptidase 1 chain	?	NA	HSF-new SA					Kousi 2012
c.1027G>A, p.(Glu343Lys)	Tripeptidyl-peptidase 1 chain	c.1027G>A, p.(Glu343Lys)	CLN2		0	A	ND		Sleat 1999
c.1057A>C, p.(Thr353Pro)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	CLN2		NA	NA	NA		Steinfeld 2002
c.1064T>C, p.(Leu355Pro)	Tripeptidyl-peptidase 1 chain	?	NA		NA	NA	NA		Hofmann 2002, Kousi 2012
c.1093T>C, p.(Cys365Arg)	Tripeptidyl-peptidase 1 chain	c.1595dupA, p.(Glu534Profs*74)	CLN2	HSF-SD broken	0	PA	7.8	Loss of disulfide-bond, surface exposure	Sleat 1999
c.1094G>A, p.(Cys365Tyr)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	CLN2	HSF-SD broken	NA	NA	NA	Loss of disulfide bond	Sleat 1999
c.1094G>A, p.(Cys365Tyr)	Tripeptidyl-peptidase 1 chain	c.1094G>A, p.(Cys365Tyr)	CLN2	HSF-SD broken	NA	NA	NA	Loss of disulfide bond	Sleat 1999
c.1094G>A, p.(Cys365Tyr)	Tripeptidyl-peptidase 1 chain	c.1361C>A, p(Ala454Glu)	CLN2	HSF-SD broken	NA	NA	NA	Loss of disulfide bond	Sleat 1999
c.1094G>A, p.(Cys365Tyr)	Tripeptidyl-peptidase 1 chain	c.1525C>T, p.(Gln509X)	NA	HSF-SD broken	NA	NA	NA		Kousi 2012
c.1094G>A, p.(Cys365Tyr)	Tripeptidyl-peptidase 1 chain	?	NA	HSF-SD broken	NA	NA	NA		Kousi 2012
c.1146C>G, p.(Ser382Arg)	Tripeptidyl-peptidase 1 chain	c.887-10A>G	NA	first nucleotide of exon 10					Kousi 2012
c.1154T>A, p.(Val385Asp)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	CLN2	HSF-SD broken	NA	NA	NA		Sleat 1999

c.1166G>A, p.(Gly389Glu) + c.299A>G, p.(Gln100Arg)	Tripeptidyl-peptidase 1 chain	c.1166G>A, p.(Gly389Glu) + c.299A>G, p.(Gln100Arg)	CLN2	HSF-SD broken; two homozygous missense mutations	NA	NA	NA		Sleat 1999
c.1266G>C, p.(Gln422His)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	CLN2, NA	Last nucleotide of exon 10	0	PA	5.1		Sleat 1999; Kousi 2012
c.1266G>C, p.(Gln422His)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	CLN2	Last nucleotide of exon 10	0	PA	5.1		Sleat 1999; Steinfeld 2002
c.1266G>C, p.(Gln422His)	Tripeptidyl-peptidase 1 chain	c.1266G>C, p.(Gln422His)	CLN2	Last nucleotide of exon 10	0	PA	5.1		Sleat 1999
c.1269G>C, p.(Glu423Asp)	Tripeptidyl-peptidase 1 chain	?	CLN2	Third nucleotide of exon 11	NA	NA	NA		Sleat 2001
c.1284G>T, p.(Lys428Asn)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	CLN2	HSF-SA broken	NA	NA	NA		Zhong 2000 , Ju 2002
c.1340G>A, p.(Arg447His)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	JNCL		1.8	PA	7.1		Wisniewski 1999; Sleat 1999
c.1340G>A, p.(Arg447His)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	JNCL		1.8	PA	7.1		Wisniewski 1999; Sleat 1999
c.1343C>T, p.(Ala448Val); c.1501G>T, p.(Gly501Cys)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	NA						Kousi 2012
c.1358C>T, p.(Ala453Val)	Tripeptidyl-peptidase 1 chain	c.311T>A, p.(Leu104*)	CLN2	HSF-new SD	NA	NA	NA		Kohan 2009
c.1417G>A, p.(Gly473Arg)	Tripeptidyl-peptidase 1 chain	p.(Ser62Glyfs*25)	CLN2	HSF- new SA, SD broken	NA	NA	NA	Disturb catalytic activity	Lam 2001
c.1417G>A,	Tripeptidyl-peptidase 1	c.622C>T,	CLN2	HSF- new SA, SD	NA	NA	NA	Disturb catalytic	Zhong

p.(Gly473Arg)	chain	p.(Arg208*)		broken				activity	2000
c.1424C>T, p.(Ser475Leu)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	CLN2, NA	Second last nucleotide of exon 11	0	N	1.2	Active site, surface exposure, impair processing	Sleat 1999; Kousi 2012
c.1424C>T, p.(Ser475Leu)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	NA	Second last nucleotide of exon 11	0	N	1.2	Active site, surface exposure	Kousi 2012
c.1424C>T, p.(Ser475Leu)	Tripeptidyl-peptidase 1 chain	c.1424C>T, p.(Ser475Leu)	NA	Second last nucleotide of exon 11	0	N	1.2	Active site, surface exposure	Kousi 2012
c.1424C>T, p.(Ser475Leu)	Tripeptidyl-peptidase 1 chain	?	NA	Second last nucleotide of exon 11	0	N	1.2	Active site, surface exposure	Kousi 2012
c.1439T>G, p.(Val480Gly)	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	JNCL		NA	NA	NA		Elleder 2008
c.1439T>G, p.(Val480Gly)	Tripeptidyl-peptidase 1 chain	?	NA		NA	NA	NA		Elleder 2008
c.1442T>G, p.(Phe481Cys)	Tripeptidyl-peptidase 1 chain	c.851G>T, p.(Gly284Val)	INCL	HSF- new SD	NA	NA	NA		Ju 2002
c.1444G>C, p.(Gly482Arg)	Tripeptidyl-peptidase 1 chain	c.1444G>C, p.(Gly482Arg)	NA	HSF-SD broken	NA	NA	NA		Kousi 2009
c.1551+5_1551+6delinsTA	Tripeptidyl-peptidase 1 chain	c.622C>T, p.(Arg208*)	NA	Splice defect according to the article					Kousi 2012
c.1551+5_1551+6delinsTA	Tripeptidyl-peptidase 1 chain	c.1551+5_1551+6delinsTA	NA	Splice defect according to the article					Kousi 2012
c.1397T>G, p.(Val466Gly)	Tripeptidyl-peptidase 1 chain	c.509-1G>C	SCAR 7						This study

[1] % of wild type TPP1
[2] unit = hr, wild type TPP1 : 1.9 hr

Chapter 6

GPSM2 and Chudley–McCullough Syndrome: A Dutch Founder Variant Brought to North America

Rowida Almomani*, Yu Sun*, Emmelien Aten, Yvonne Hilhorst-Hofstee, Cacha M.P.C.D. Peeters-Scholte, Arie van Haeringen, Yvonne M.C. Hendriks, Johan T. den Dunnen, Martijn H. Breuning, Marjolein Kriek, and Gijs W.E. Santen

***Both authors contributed equaly to this work**

Am J Med Genet Part A , 2013.

Abstract

Chudley–McCullough syndrome (CMS) is characterized by profound sensorineural hearing loss and brain anomalies. Variants in GPSM2 have recently been reported as a cause of CMS by Doherty et al. In this study we have performed exome sequencing of three CMS patients from two unrelated families from the same Dutch village. We identified one homozygous frameshift GPSM2 variants c.1473delG in all patients. We show that this variant arises from a shared, rare haplotype. Since the c.1473delG variant was found in Mennonite settlers, it likely originated in Europe. To support DNA diagnostics, we established an LOVD database for GPSM2 containing all variants thus far described.

Introduction

Chudley–McCullough syndrome (CMS) is characterized by profound congenital sensorineural hearing loss associated with (partial) agenesis of the corpus callosum, colpocephaly (enlargement of the occipital horns), hydrocephaly, and other brain abnormalities such as arachnoid cysts, gray matter hetero- topia, and cortical dysplasia [Ostergaard et al., 2004]. This syndrome was first recognized in a brother and sister by Chudley et al. [1997]. Based on affected sibs of both sexes from phenotypically normal parents the syndrome was assumed to be an autosomal recessive trait. Subsequent reports have supported this assumption, describing parental consanguinity or origin from a small community [Chudley et al., 1997].

Patients with CMS may either be hearing or deaf at birth. However, hearing loss is always profound by the age of 3 years [Hendriks et al., 1999; Lemire and Stoeber, 2000; Welch et al., 2003; Ostergaard et al., 2004; Matteucci et al., 2006]. It is has been suggested that some cases of CMS may not be detected because the hydrocephalus does not progress and is compensated [Welch et al., 2003].

Inactivating mutations in the GPSM2 gene have been linked to both autosomal recessive non-syndromic (DFNB82) and syndromic hearing loss. Doherty et al. [2012] recently linked inactivating mutations in GPSM2 to CMS. They successfully applied exome sequencing in conjunction with homozygosity mapping to identify four deleterious mutations (c.1473delG (c.1471delG in Doherty et al.), c.742delC (c.741delC in Doherty et al.), c.1661C>A and

c.1062þ1G>T) in affected individuals with CMS from eight fam- ilies. In this study, two other families with CMS were investigated; the same c.1473delG variant in three patients from two unrelated Dutch families was identified. Together, the c.742delC and c.1473delG founder mutations seem to be a frequent cause of CMS, as they are observed in homozygous form in 8/10 families reported thus far, and in heterozygous form in an additional family. In the present study we confirm that mutations in GPSM2 gene are responsible for CMS and show that at least a part of the CMS cases are due to a founder effect.

Materials and Methods

Patients

In our clinic we had three CMS patients in two different families (Fig. 1). In the first family, described previously by Hendriks et al. [1999], two affected sisters had a combination of congenital sensorineural hearing loss, partial agenesis of the corpus callosum, arachnoid cyst, and hydrocephalus. They had normal development and no distinctive physical anomalies. Their parents were non- consanguineous but originated from the same Dutch village, were phenotypically normal and both had normal hearing and no brain abnormalities. Hendriks et al. postulated that the two affected sibs may have had a different syndrome than that described by Chudley et al. Welch et al. later commented that the two affected girls most likely had CMS [Hendriks et al., 1999; Welch et al., 2003]. Recently the two sisters were re-examined at the age of 17 and 25 years, respectively, and had normal intelligence.

The second family included a single affected patient who was born after an uneventful pregnancy. A structural ultrasound study at a gestational age of 20 weeks did not show abnormalities. During pelvic examination at labor a hydrocephalus was suspected and a subsequent ultrasound revealed ventriculomegaly and cesarean section was performed. The patient was born at term. Postnatal brain imaging (MRI) revealed colpocephaly, agenesis of the corpus callosum, heterotopia, an interhemispheric cyst at the dorsum of the third ventricle, polymicrogyria of frontal lobes and cerebellar dysgenesis. The patient was also diagnosed with severe sensorineu- ral hearing loss (no response at 90 dB). She had a normal develop- ment at the age of 2 years except for her delay in speech development. After she received a cochlear implant at the age of 2 years and 4 months, there was an improvement in hearing (30 –40 dB hearing

level) and in active vocabulary. The family historywas unremarkable. Physical findings were normal (head circum- ference: +2 SDS).

Metabolic and DNA analysis of several genes related to hearing loss (GJB2, GJB6, SLC26A4 gene) or brain anomalies (GPR56 gene) did not result in an obvious cause for the malformations. Based on the presence of these rather specific clinical findings, the patient was diagnosed as having CMS.

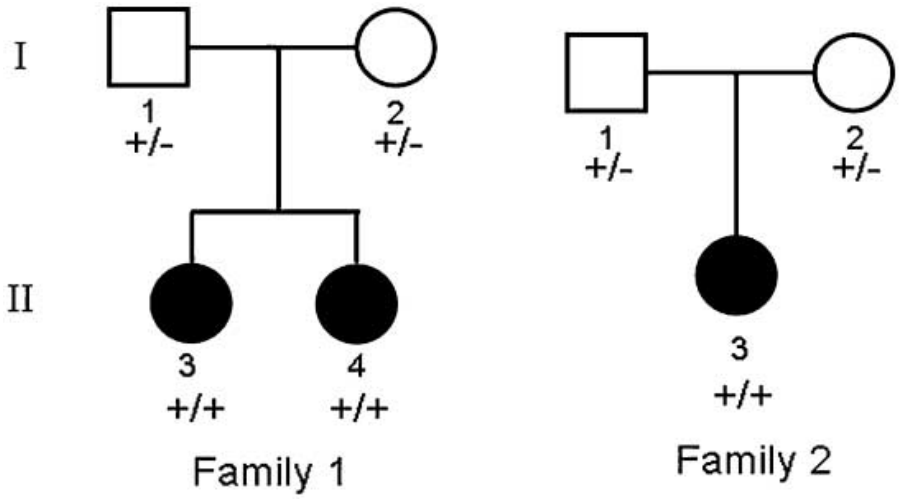


Figure 1. Pedigrees of the two families with Chudley-McCullough syndrome. Black symbols: affected patients. +/+: homozygous for the c.1473delG variant; +/-: heterozygous for the c.1473delG variant.

Exome Sequencing

As the gene for CMS was not yet identified at the time of this investigation, we have applied exome sequencing to resolve the genetic basis of CMS in these two families. We sequenced the exomes of the sibs from family one and the sporadic patient from family two. Whole exome capture was performed using Agilent's 50 Mb Sure- select exome capture kit following instructions provided by the manufacturers (SureSelect, Agilent). In brief, 3 mg of DNA was fragmented (Covaris) to yield fragments of 300–400 bp. Paired- end adaptors with index from Illumina were added to both ends. The DNA-adaptor-ligated fragments were then hybridized to 250 ng of SureSelect whole exome probes capture library (SureSelect, Agilent) for 30h. After capture, a qPCR assay was done to calculate the relative fold-enrichment prior for sequencing. The eluted-enriched DNA fragments were sequenced using the Hiseq 2000 platform (Illumina). BWA [Li and Durbin, 2009] was used to map the data to human genome build 37 (hg19), and

GATK tools [McKenna et al., 2010] were used to perform the data analysis with minor modifications described elsewhere [Santen et al., 2012].

Sanger Sequencing

PCR was done by using Phire Hot Start II DNA polymerase following the official protocol. Primers used in PCR reactions are available upon request. After PCR, fragments were first purified by QIAquick PCR purification kit (Qiagen), then mixed with 10 pmol of the forward or reverse primers and sequenced by the Applied Biosystems 96-capillary (3730XL system).

Results

Using the GATK sequence analysis pipeline, we identified 26,487 possibly shared variants in the two siblings and 22,901 variants in the sporadic patient located in the exons and exon/intron junctions. After filtering to exclude all known variants in databases (dbSNP135, 1000 Genomes Project, and the in-house database) 278 and 181 variants remained, respectively. Furthermore, filtering for recessive inheritance left only a homozygous single base pair deletion in *GPSM2* which causes a frameshift and a premature stop (NM_013296.4:c.1473delG, p.Phe492SerfsX5). This variant was confirmed in all patients and in heterozygous form in their parents by Sanger sequencing.

Since both of our families came from the same small Dutch village and shared the same homozygous variant, we used the exome data to reveal a possible link between the two families. We found two other rare homozygous variants shared by all three patients, both of which within 0.5 Mbp of the *GPSM2* variant (g.109477462G>C; NM_015127.3(CLCC1):c.1336C>G; p.Pro446Ala and g.109909853A>C; NM_002959.5(SORT1):c.440+177T>G). To further define the size of the haplotype, we looked at high quality calls of known SNPs in this genomic region and found that the region of homozygous SNPs shared by all three patients is flanked by rs6672483 and rs333967 (Fig. 2), indicating that the maximal size of the shared haplotype is 2.2 Mb. The shared genotype for the two sisters spans 14.7 Mb (Fig. 2). The sporadic patient has a homozygous stretch of 19 Mb (Fig. 2). None of the three rare variants shared by the three patients was identified in the Genome of the Netherlands (consisting of 250 completely sequenced trios, 500 independent genomes, www.nlgenome.nl). The g.109477462G>C variant

was found in 0.2% of the subjects with European ancestry sequenced in the National Heart, Lung, and Blood Institute Grand Opportunity Exome Sequencing Project (NHLBI GO ESP).

We further investigated whether the distant relation of the three patients was reflected by the mitochondrial DNA (i.e. the maternal line). We found 19 positions with variants on the mitochondrial DNA covered at least 10x in all of the subjects. There was a 100% concordance between the two sisters and only one position in which the sporadic patient differed. We compared the concordance with three other Dutch patients from different projects and found a much lower concordance in those patients (44-55%).

To support DNA diagnostic studies we established a Leiden Open Variation Database (LOVD) for *GPSM2* containing all variants that have been published thus far (www.LOVD.nl/GPSM2).

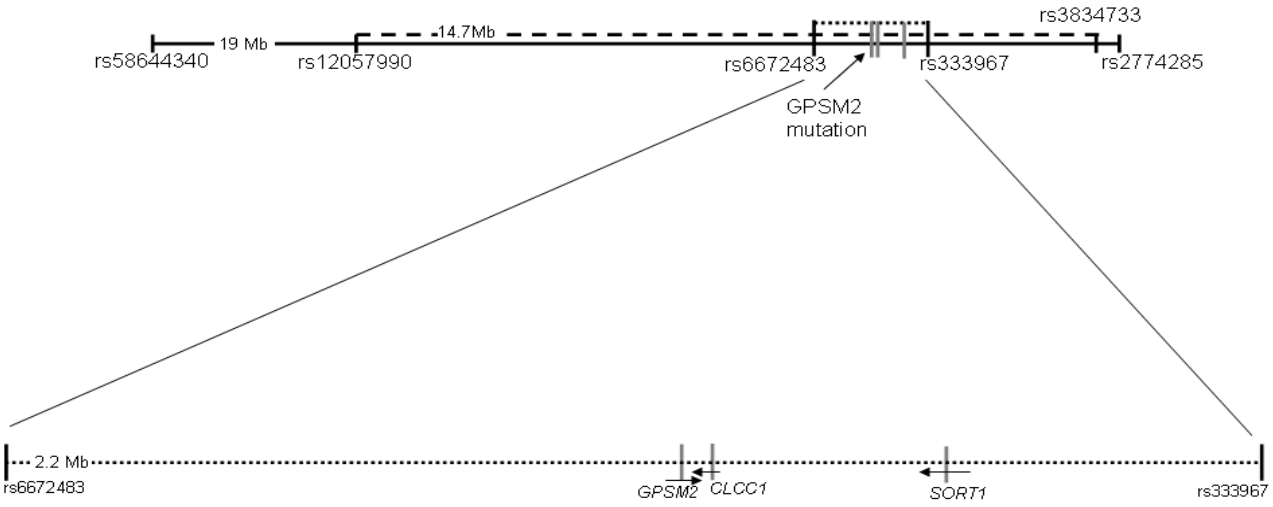


Figure 2. Representation of the haplotype information around the *GPSM2* mutation. The three vertical gray lines represent the rare homozygous variants shared by all three patients in *GPSM2, CLCC1* and *SORT1*, respectively. The dotted line represents the homozygous stretch shared by all three patients (2.2 Mb). The striped line represents the stretch where the two sisters share the same genotype (14.7 Mb). The continuous line represents the homozygous stretch for the sporadic patient (19 Mb).

Discussion

GPSM2 was linked to autosomal recessive non-syndromic hearing loss DFNB82 by Walsh et al, who described the successful application of exome sequencing in conjunction with homozygosity mapping to identify a nonsense variant in *GPSM2* (c.379C>T;p.Arg127*) [Walsh et al., 2010]. Subsequently, Yariz et al. [2012] reported a second truncating variant in *GPSM2*

(c.1684C>T;p.Q562*) by autozygosity mapping followed by candidate gene analysis in a consanguineous Turkish family with nonsyndromic hearing loss. While preparing this manuscript, Doherty et al. [2012] reported the identification of variants in *GPSM2* as a cause of CMS. They identified the c.1473delG variant in patients from Mennonite ancestry (reported as c.1471delG) [Doherty et al., 2012]. The authors hypothesize that it is from European origin, which we can now confirm. The size we computed for the common haplotype was 2.2 Mb, which is smaller but in the same order of magnitude as the 2.9 Mb haplotype observed in the Mennonite families [Doherty et al., 2012]. Since none of the variants was present in the database of the Genome of The Netherlands, and one of the variants was present in a low frequency in the Exome Variant Server (NHLBI GO ESP), we conclude that this haplotype is very rare, and represents a founder haplotype in the village of origin. This is further strengthened by the mitochondrial data, which shows that the two families are related in the maternal line, and by the fact that we have identified another unrelated family with the same homozygous mutation from this village (data not shown).

We have created a variant database (www.LOVD.nl/GPSM2) for *GPSM2*. The value of this database is in enhanced interpretation for diagnostic use, but also facilitates comparison between studies. The c.1473delG variant that we identified was erroneously annotated as c.1471delG in Doherty et al., 2012. The creation of a database which checks for HGVS nomenclature partly resolves such differences between papers.

GPSM2 (the G protein signaling modulator 2) also known as *LGN* and *Pins,* contains 14 exons, and spans 55,073 bp on chromosome 1p13.3. It encodes a 684 amino acid protein. *GPSM2* has 6 transcripts according to the Ensembl database, ranging in size from 571 to 3310 bp. The GPSM2 protein is widely expressed [Blumer et al., 2002]. However, highest expression is seen during embryonic development. Its functional role relates to cell polarity and spindle orientation, for example, in cells of the developing cerebral cortex in mice [discussed by Doherty et al., 2012]. The protein is comprised of seven N-terminal tetratricopeptide (TPR) motifs, a linker domain, and four C-terminal (GoLoco) motifs which are involved in guanine nucleotide exchange [Du et al., 2001; Johnston et al., 2009; Willard et al., 2008]. The 1 bp deletion in exon 13 locates within the C-terminal GoLoco motif and creates a frame shift starting at codon Phe492 and ends in a stop codon four positions downstream, leading to a functional absence of the GPSM2 protein.

Our results confirm that inactivating mutations in *GPSM2* cause Chudley McCollough syndrome. The c.1473delG mutation in *GPSM2* associated with CMS appears to be an ancient founder mutation brought to North America by Mennonite settlers originating from Western Europe. Together, the c.742delC and c.1473delG founder mutations seem to be a frequent cause of CMS. Future work will need to show if an ascertainment bias has inflated the importance of these mutations.

Acknowledgments

We thank the patients and their parents for participation in this study. Yu Sun was supported by China Scholarship Council. We received funding from the EU FP7 framework program agreements 223026 (NMD-chip) and 223143 (TechGene).

Web Resources

GPSM2, Leiden Open Variant Database: http://www.lovd.nl/gpsm2

Genome of the Netherlands: http://www.nlgenome.com

NHLBI GO ESP: http://evs.gs.washington.edu/EVS/

References

Blumer JB, Chandler LJ, Lanier SM. 2002. Expression analysis and subcellular distribution of the two G-protein regulators AGS3 and LGN indicate distinct functionality. Localization of LGN to the midbody during cytokinesis. J Biol Chem 277:15897–15903.

Chudley AE, McCullough C, McCullough DW. 1997. Bilateral sensorineural deafness and hydrocephalus due to foramen of Monro obstruction in sibs: a newly described autosomal recessive disorder. Am J Med Genet 68:350-6.

Doherty D, Chudley AE, Coghlan G, Ishak GE, Innes AM, Lemire EG, Rogers RC, Mhanni AA, Phelps IG, Jones SJ, Zhan SH, Fejes AP, Shahin H, Kanaan M, Akay H, Tekin M; FORGE Canada Consortium, Triggs-Raine B, Zelinski T. 2012. *GPSM2* mutations cause the brain malformations and hearing loss in Chudley-McCullough syndrome. Am J Hum Genet 90:1088-93.

Du Q, Stukenberg PT, Macara IG. 2001. A mammalian partner of inscrutable binds NuMA and regulates mitotic spindle organization. Nat Cell Biol 3:1069–1075.

Johnston CA, Hirono K, Prehoda KE, Doe CQ. 2009. Identification of an Aurora-A/PinsLINKER/Dlg spindle orientation pathway using induced cell polarity in S2 cells. Cell 138: 1150–1163.

Hendriks YMC, Laan LAEM, Vielvoye GJ, Van Haeringen A. 1999. Bilateral sensorineural deafness, partial agenesis of the corpus callosum, and arachnoid cysts in two sisters. Am J Med Genet 86:183–186.

Lemire EG, Stoeber GP. 2000. Chudley-McCullough syndrome: bilateral sensorineural deafness, hydrocephalus, and other structural brain abnormalities. Am J Med Genet 90:127–130.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Matteucci F, Tarantino E, Bianchi MC, Cingolani C, Fattori B, Nacci A, Ursino F. 2006. Sensorineural deafness, hydrocephalus and structural brain abnormalitiesin two sisters: the Chudley-McCullough syndrome. Am J Med Genet 140:1183–1188.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303.

Ostergaard E, Pedersen VF, Skriver EB, Brøndum-Nielsen K. 2004. Brothers with Chudley-McCullough syndrome: sensorineural deafness, agenesis of the corpus callosum, and other structural brain abnormalities. Am J Med Genet 124:74-8.

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. 2012. Mutations in SWI/SNF chromatin remodeling complex gene *ARID1B* cause Coffin-Siris syndrome. Nat Genet 44:379-80

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M. 2010. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet 87:90-4.

Welch KO, Tekin M, Nance WE, Blanton SH, Arnos KS, Pandya A. 2003. Chudley-McCullough syndrome: expanded phenotype and review of the literature. Am J Med Genet 119:71–76.

Willard FS, Zheng Z, Guo J, Digby GJ, Kimple AJ, Conley JM, Johnston CA, Bosch D, Willard MD, Watts VJ, Lambert NA, Ikeda SR, Du Q, Siderovski DP. 2008. A point mutation to Galphai selectively blocks GoLoco motif binding: Direct evidence for Galpha.GoLoco complexes in mitotic spindle dynamics. J Biol Chem 283:36698–36710.

Yariz KO, Walsh T, Akay H, Duman D, Akkaynak AC, King MC, Tekin M. 2012. A truncating mutation in GPSM2 is associated with recessive non-syndromic hearing loss. Clin Genet 81:289-93.

Chapter 7

Digenic inheritance of an *SMCHD1* mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2

Richard J.L.F. Lemmers*, Rabi Tawil*, Lisa M. Petek, Judit Balog, Gregory J. Block, Gijs W.E. Santen, Amanda M. Amell, Patrick J. van der Vliet, Rowida Almomani, Kirsten R. Straasheijm, Yvonne D. Krom, Rinse Klooster, Yu Sun, Johan T. den Dunnen, Quinta Helmer, Colleen M. Donlin-Smith, George W. Padberg, Baziel G.M. van Engelen, Jessica C. de Greef, Annemieke M. Aartsma-Rus, Rune R. Frants, Marianne de Visser, Claude Desnuelle, Sabrina Sacconi, Galina N. Filippova, Bert Bakker, Michael J. Bamshad, Stephen J. Tapscott, Daniel G. Miller, Silvère M. van der Maarel

***Contributed equally to this work**

Nat Genet. 2012; 44:1370-4.

Abstract

Facioscapulohumeral dystrophy (FSHD) is characterized by chromatin relaxation of the D4Z4 macrosatellite array on chromosome 4 and expression of the D4Z4-encoded *DUX4* gene in skeletal muscle. The more common form, autosomal dominant FSHD1, is caused by contraction of the D4Z4 array, whereas the genetic determinants and inheritance of D4Z4 array contraction–independent FSHD2 are unclear. Here, we show that mutations in *SMCHD1* (encoding structural maintenance of chromosomes flexible hinge domain containing 1) on chromosome 18 reduce SMCHD1 protein levels and segregate with genome-wide D4Z4 CpG hypomethylation in human kindreds. FSHD2 occurs in individuals who inherited both the *SMCHD1* mutation and a normal-sized D4Z4 array on a chromosome 4 haplotype permissive for *DUX4* expression. Reducing SMCHD1 levels in skeletal muscle results in D4Z4 contraction–independent DUX4 expression. Our study identifies *SMCHD1* as an epigenetic modifier of the D4Z4 metastable epiallele and as a causal genetic determinant of FSHD2 and possibly other human diseases subject to epigenetic regulation.

Main

FSHD (MIM 158900) is clinically characterized by the initial onset of facial and upper-extremity muscle weakness that is often asymmetric and progresses to involve both upper and lower extremities[1]. FSHD1 and FSHD2 are phenotypically indistinguishable, and both are associated with DNA hypomethylation and decreased repressive heterochromatin at the D4Z4 macrosatellite array, which we collectively refer to as chromatin relaxation[2, 3, 4, 5, 6, 7, 8] (Supplementary Fig. 1). Each D4Z4 unit encodes a copy of the *DUX4* retrogene (encoding double homeobox 4)[9, 10, 11, 12, 13], a transcription factor expressed in the germ line that is epigenetically repressed in somatic tissues. D4Z4 chromatin relaxation in FSHD results in inefficient epigenetic repression of *DUX4* and a variegated pattern of DUX4 protein expression in a subset of skeletal muscle nuclei[14] (Supplementary Fig. 1). Ectopic expression of DUX4 in skeletal muscle activates the expression of stem cell and germline genes[15], and, when overexpressed in somatic cells, DUX4 can ultimately lead to cell death[12, 16, 17, 18, 19, 20]. Chromatin relaxation in FSHD1 is associated with contraction of the array to 1–10 D4Z4 repeat units and has a dominant inheritance pattern linked to the contracted array. In FSHD2, chromatin

relaxation is independent of the size of the D4Z4 array and occurs at both chromosome 4 D4Z4 arrays and at the highly homologous arrays on chromosome 10 (Supplementary Fig. 1)[2, 7, 8, 21, 22].

D4Z4 chromatin relaxation must occur on a specific chromosome 4 haplotype to cause FSHD1 and FSHD2. This haplotype contains a polyadenylation signal to stabilize *DUX4* mRNA in skeletal muscle[13, 23, 24, 25, 26, 27]. Chromosomes 4 and 10 that lack this polyadenylation signal do not produce DUX4 protein; consequently, D4Z4 chromatin relaxation and transcriptional derepression on these nonpermissive haplotypes does not lead to disease. Because chromatin relaxation occurs at D4Z4 repeats on both chromosomes 4 and 10 in FSHD2, we sought to determine whether an inherited defect in a modifier of D4Z4 repeat–mediated epigenetic repression might cause FSHD2 when combined with an FSHD-permissive *DUX4* allele.

To quantify D4Z4 chromatin relaxation, we determined the percentage of CpG methylation on the basis of measurements following cleavage with the methylation-sensitive FseI endonuclease, in an assay that averaged the percentage of D4Z4 methylation on both alleles of chromosomes 4 and 10 in a cohort of 72 controls, 93 individuals with FSHD1 and 53 individuals with FSHD2. In FSHD2-affected individuals, D4Z4 methylation was at least 2 s.d. below the average levels in the general population (44% ± 10% for the general population and 11 ± 5% for individuals with FSHD2; Fig. 1a, Supplementary Fig. 2 and Supplementary Note). Using a stringent methylation threshold of <25%, we discovered that, in some kindreds identified by a proband with FSHD2, D4Z4 hypomethylation segregated in a pattern consistent with autosomal dominant inheritance that was not linked to the chromosome 4 or 10 D4Z4 array haplotype (Fig. 1b). In these kindreds, individuals with FSHD2 inherited both the hypomethylation trait and the FSHD-permissive chromosome 4 haplotype with the *DUX4* polyadenylation signal, suggesting that two independently segregating loci cause and determine the penetrance of FSHD2.

To identify the locus controlling the D4Z4 hypomethylation trait, we performed whole-exome sequencing[28] of 14 individuals in 7 unrelated families with FSHD2: 5 with dominant segregation of the hypomethylation trait and 2 with sporadic hypomethylation and FSHD2. Detailed genetic analysis of the repeat lengths and haplotypes did not provide evidence of non-paternity in these families (Fig. 1b). Families were stratified according to the criterion that the D4Z4 methylation level had to be <25%, not as a result of contracted repeats on chromosomes 4 and 10

(Supplementary Table 1 and Supplementary Note). We identified rare and potentially pathogenic mutations in the *SMCHD1* gene (encoding structural maintenance of chromosomes flexible hinge domain containing 1) in all individuals with D4Z4 hypomethylation, with the exception of members of one family (Rf854; Table 1). These mutations were not present in public (dbSNP132 and the 1000 Genomes Project) or in house databases or in family members with normal D4Z4 methylation levels.

We confirmed the presence of these mutations by Sanger sequencing and included 12 additional unrelated families with FSHD2 from whom DNA or RNA was available. We identified heterozygous out-of-frame deletions, heterozygous splice-site mutations and heterozygous missense mutations in *SMCHD1* in 15 out of 19 families (79%; Fig. 1b, Table 1 and Supplementary Fig. 3). We also confirmed that the splice-site mutations altered normal *SMCHD1* mRNA by excluding exons and introducing the usage of cryptic splice sites (Supplementary Fig. 4a,b).

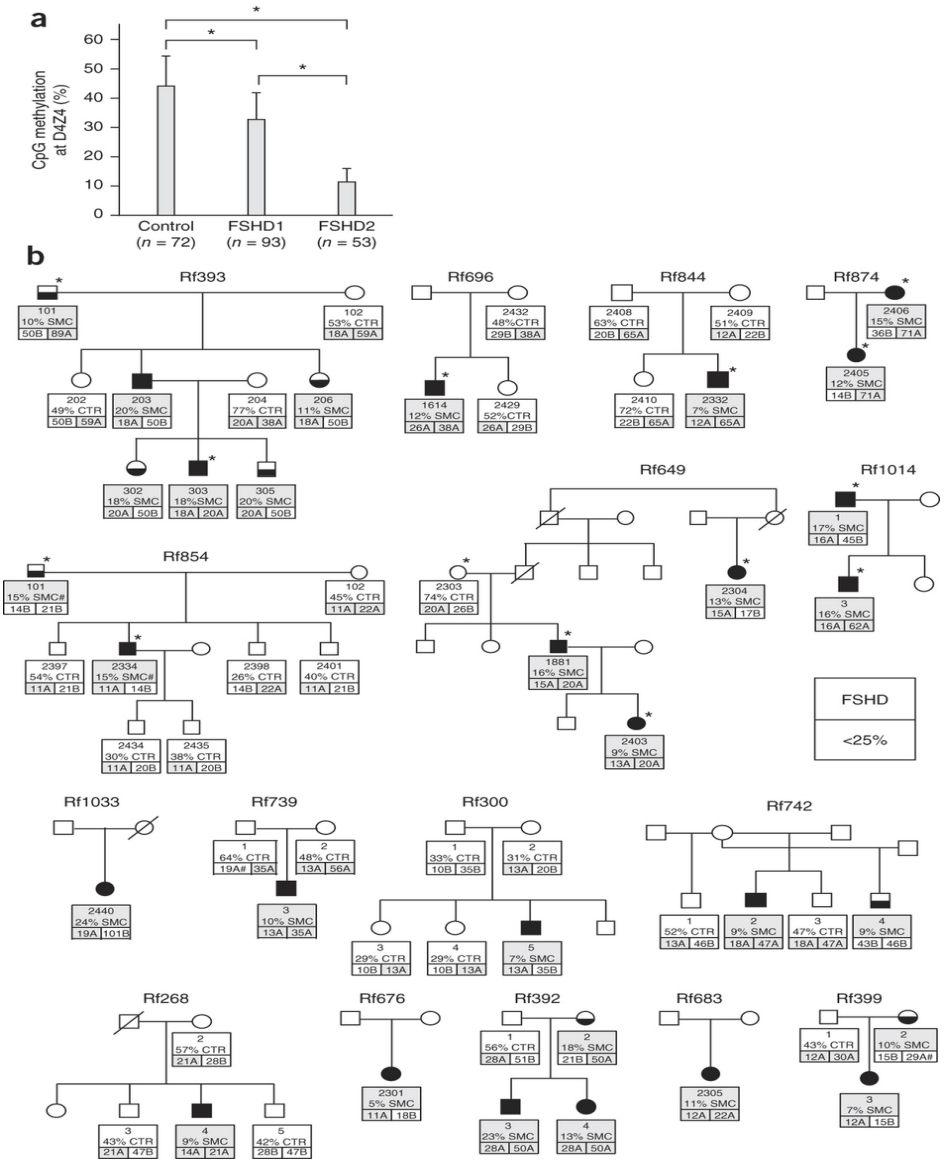


Figure 1 D4Z4 methylation test and families with FSHD2. (**a**) FseI methylation levels in 72 control, 93 FSHD1 and 53 FSHD2 genomic DNA samples. Error bars, s.d. **P* < 0.005. (**b**) Pedigrees of families with FSHD2. For each individual, the subject ID, FseI-detected methylation level (%) and presence (SMC, gray) or absence (CTR, white) of a *SMCHD1* mutation is indicated (upper box). Also indicated in the lower two boxes are the lengths of both D4Z4 arrays on chromosomes 4 in units. Permissive alleles, typically A alleles defined on the basis of a polymorphism distal to the repeat (24), are indicated by gray boxes. B alleles, which are nonpermissive (42), are indicated by white boxes. Some less common subtypes of the A allele are considered to be nonpermissive (41); these are marked with # and colored white (Rf399 and Rf739). Note the independent segregation of D4Z4 hypomethylation and FSHD-permissive alleles. Only in those individuals in whom a permissive allele combines with D4Z4 hypomethylation (<25%) was FSHD diagnosed, whereas individuals with D4Z4 hypomethylation carrying nonpermissive alleles were unaffected by FSHD. Individuals selected for whole-exome sequencing (upper seven pedigrees) are indicated by asterisks. SMC# indicates a coding synonymous SNP identified in Rf854. A key for the symbols is included.

Table 1 SMCHD1 mutations identified

Inheritance/ family	Mutation type	Nr[a]	Position[b]	Chromosome position[c]	Transcript position[d]	Protein position[e]	RNA analysis
Maternal Rf742	Missense	M1	Exon 9	g.2697047A>G	c.1058A>G	p.Tyr353Cys	_
Unknown Rf1033	Deletion	D1	Exon 10	g.2697999_2698003del	c.1302_1306del	p.Tyr434*	WT + mutant transcript[g]
De novo Rf739	Missense	M2	Exon 11	g.2700630G>C	c.1436G>C	p.Arg479Pro	WT + mutant transcript[g]
De novo Rf300	Missense	M3	Exon 12	g.2700743T>C	c.1474T>C	p.Cys492Arg	WT + mutant transcript
Paternal Rf393	Deletion	D2	Exon 12	g.2700875_2700875del	c.1608del	p.Asp537Ilefs*10	WT + mutant transcript[g]
Unknown Rf696	5′ splice site	S1	Intron 12	g.2701019A>G	c.1647+103A>G		WT + skipped exon 12[g] + cryptic splicing of exon12[g]
Maternal Rf399	Missense	M4	Exon 16	g.2707565C>T	c.2068C>T	p.Pro690Ser	WT + mutant Transcript
Unknown Rf268	5′ splice site	S2	Exon 20	g.2722661G>A	c.2603G>A	p.Ser868Asn	_
De novo Rf844	5′ splice site	S3	Intron 25	g.2732488_2732492de	c.3274_3276+2del		WT + skipped exon 25 + cryptic splicing of exon 25
Maternal Rf874	5′ splice site	S3	Intron 25	g.2732488_2732492del	c.3274_3276+2del		_
Paternal	Synonym	CS	Exon	g.2739448T>A[f]	c.3444T>A	p.Pro1148Pro	WT + mutant

Rf854	ous		27				transcript
Paternal Rf649	5′ splice site	S4	Intron 29	g.2743927G>A	c.3801+1G >A		WT + cryptic splicing
Unknown Rf676	5′ splice site	S4	Intron 29	g.2743927G>A	c.3801+1G >A		_
Paternal Rf1014	5′ splice site	S5	Exon 36	g.2762234G>A	c.4566G>A	p.Thr1522Thr	WT + skipped exon 36
Maternal Rf392	5′ splice site	S5	Exon 36	g.2762234G>A	c.4566G>A	p.Thr1522Thr	WT + skipped exon 36 + cryptic splicing of exon 36
Unknown Rf683	Missense	M5	Exon 37	g.2763729T>C	c.4661T>C	p.Phe1554Ser	WT + mutant transcript

–, no RNA available; WT, wild type. [a]The position of each mutation is shown according to mutation number (Nr) in supplementary Figure 3. [b]Exon number is based on Ensembl transcript ENST00000320876. [c]Genomic position is based on hg19. [d]Transcript position is based on NM_015295.2. [c]Protein position is based on NP_056110.2. [f]Present at frequency of 0.0055 in 1000 Genomes Project data. [g]Disrupts ORF.

Because heterozygous *SMCHD1* mutations cosegregated with D4Z4 hypomethylation in families with FSHD2 or occurred *de novo* in individuals with sporadic hypomethylation and FSHD2 (Fig. 1b), we considered *SMCHD1* haploinsufficiency to be a candidate disease mechanism, particularly because many of the mutations were predicted to affect production of the full-length protein. Indeed, fibroblasts from individuals with FSHD2 who had nonsynonymous or splice-site mutations in *SMCHD1* expressed substantially lower amounts of SMCHD1 protein relative to control individuals (Fig. 2a). We found normal levels of SMCHD1 protein in the individual with FSHD2 with hypomethylated D4Z4 in family Rf854 who did not have an *SMCHD1* mutation (Fig. 2a), suggesting that FSHD2 in this family has a genetic cause other than *SMCHD1* haploinsufficiency. Finally, chromatin immunoprecipitation (ChIP) showed the presence of SMCHD1 on the D4Z4 array and detected lower levels of this association in individuals with FSHD2 who had *SMCHD1* mutations (Fig. 2b). Taken together, these results support

haploinsufficiency of *SMCHD1* as a cause of D4Z4 hypomethylation in unrelated kindreds with FSHD2.

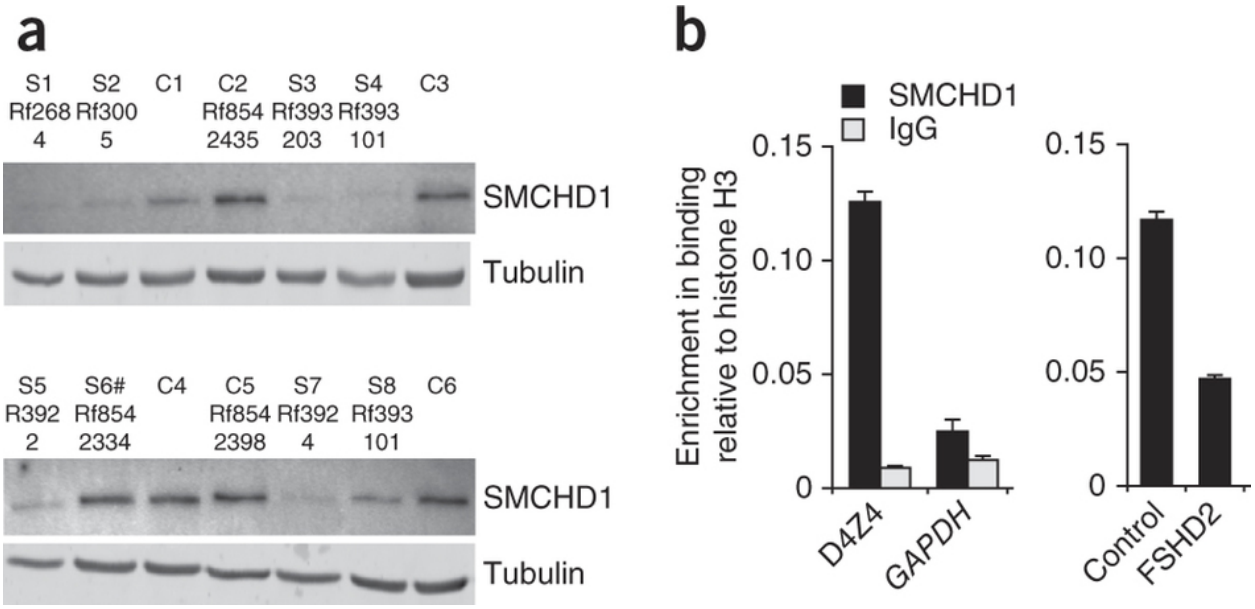


Figure 2 Families with FSHD2 with *SMCHD1* mutations.
a) Protein blot analysis of fibroblast cultures from six controls (C) and eight individuals carrying a *SMCHD1* mutation (S). Sample identifiers refer to the pedigrees in Figure 1b, and S6 denotes the individual with FSHD2 with only a synonymous coding SNP. (b) The results of ChIP analysis showing binding of SMCHD1 to D4Z4 arrays but not to *GAPDH* (left) and weaker binding of SMCHD1 to D4Z4 arrays in individual 2305 with FSHD2 from family Rf683 (right). Error bars, ± 1 s.d. from duplicate experiments. IgG, immunoglubulin G.

FSHD is characterized by low-level variegated expression of DUX4 in skeletal muscle. Therefore, we assessed *DUX4* expression in skeletal muscle cells from control individuals after decreasing *SMCHD1* levels by RNA interference (Fig. 3a,b). We detected no *DUX4* mRNA in primary myotubes from an unaffected individual with a normal-sized and methylated D4Z4 array on the FSHD-permissive *DUX4* polyadenlated haplotype. In contrast, *DUX4* was transcriptionally activated in these myotubes (Fig. 3c) when SMCHD1 transcripts and protein amounts were reduced to <50% of normal. We observed a variegated pattern of DUX4 protein expression in myotubes in all samples with adequate *SMCHD1* knockdown (Fig. 3d); this pattern was similar to that seen in myotubes from individuals with FSHD2. Cells expressing a scrambled or ineffective short hairpin RNA (shRNA) did not express DUX4 (Fig. 3, control and 4059).

To show that the *SMCHD1* splice-site mutations identified in individuals with FSHD2 result in *DUX4* expression, we manipulated *SMCHD1* pre-mRNA splicing in skeletal muscle cells using antisense oligonucleotides directed against exon 29 or 36. These antisense oligonucleotides caused skipping of *SMCHD1* exon 29 or 36 at rates comparable to those detected in some individuals with FSHD2 and resulted in transcription of *DUX4* (Fig. 3e,f). Thus, SMCHD1 activity is necessary for the somatic repression of *DUX4*, and reduction of this activity results in D4Z4 arrays that express DUX4 when an FSHD-permissive *DUX4* haplotype is present, with a pattern of variegated expression similar to that observed in FSHD1 and FSHD2 myotube cultures.

SMCHD1 belongs to the SMC gene superfamily that regulates chromatin repression in many different organisms, mediating the silencing of mating loci in yeast[29], dosage compensation in *Caenorhabditis elegans*[30, 31], position effect variegation in *Drosophila melanogaster*[32] and RNA-directed DNA methylation in *Arabidopsis thaliana*[33]. SMCHD1 was first identified in a mouse mutagenesis screen for modifiers of the variegated expression of a multicopy transgene[34]. Gene targeting confirmed that *Smchd1* was necessary for hypermethylation of a subset of CpG islands associated with X-chromosome inactivation, and continued association of the Smchd1 protein with the inactive X chromosome suggested its continuous requirement in maintaining X-chromosome inactivation[35, 36]. Our observations paint a similar picture of the role of SMCHD1 and the D4Z4 arrays: SMCHD1 is necessary for D4Z4 hypermethylation, SMCHD1 remains associated with the D4Z4 array in skeletal muscle cells, and its continuous expression is required to maintain array silencing. It will be interesting to examine individuals with *SMCHD1* mutations for subclinical abnormalities in X-chromosome inactivation.

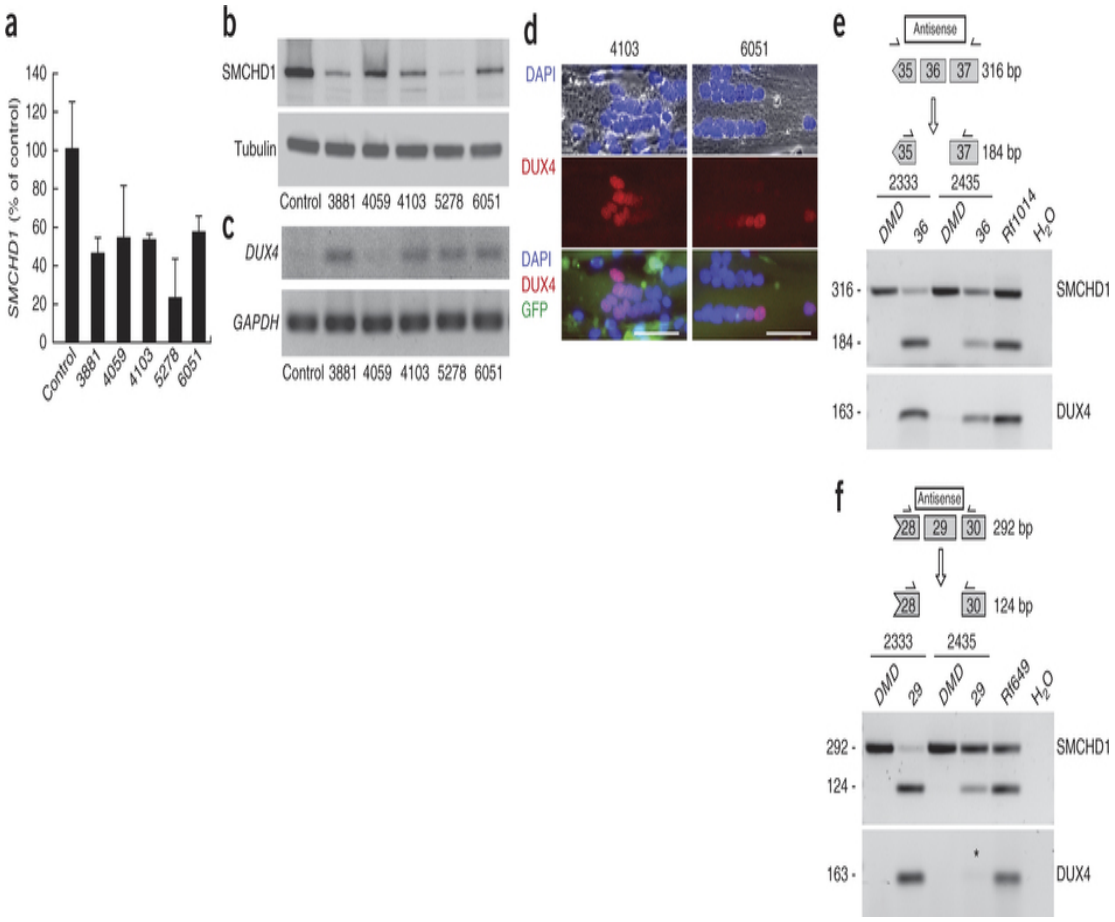


Figure 3 *SMCHD1* haploinsufficiency results in DUX4 expression in normal human myoblasts. **a**,**b**) shRNAs against different regions of *SMCHD1* are effective in reducing the production of SMCHD1 in normal human primary myoblasts at the RNA and protein levels. Numbers below the graph and the gel lanes indicate the regions within the *SMCHD1* transcript that are homologous to the indicated shRNAs. (**a**) *SMCHD1* mRNA levels were measured by quantitative RT-PCR and normalized to *RPP30* transcript levels in a multiplexed reaction. Normalized *SMCHD1* levels are shown as a percentage of the levels found in the same cells treated with a vector expressing a scrambled sequence. Error bars, s.d. of the mean for three separate reactions. (**b**) Protein blot of protein samples from the cultures in **a** normalized to tubulin. (**c**) Semiquantitative RT-PCR analysis of *DUX4* in cells deficient in *SMCHD1*. *GAPDH* was amplified to confirm RNA integrity. (**d**) Examples of DUX4 immunoreactive nuclei observed in myotubes where *SMCHD1* levels were reduced using shRNA 4103 or 6051. Myotubes are shown with nuclei labeled with DAPI (blue) and stained for DUX4 (red). GFP fluorescence produced from the lentivirus vector expressing the shRNAs is also shown. Scale bars, 50 μm. (**e**,**f**) Antisense oligonucleotide–mediated exon skipping of *SMCHD1* exons 36 and 29 in normal human myoblasts 2333 and 2435. Cells were treated with antisense oligonucleotides designed to reproduce this skipping, and primers homologous to flanking exons (shown above each gel) were used to evaluate the proportion of transcripts with skipped exons. *DUX4* expression from the same cells is shown below each panel of *SMCHD1* exon analysis. Results are also shown for myotube RNA from affected individuals in both families with the mutations. (**e**) An 184-bp fragment is produced when exon 36 is skipped. (**f**) An 124-bp fragment is produced when exon 29 is skipped. *, low *DUX4* expression levels consistent with inefficient *SMCHD1* exon skipping. An antisense oligonucleotide targeting exon 50 of the *DMD* gene (encoding dystrophin) was used as a negative control.

The *Smchd1* mutation was originally called the *Momme D1* locus (encoding modifiers of murine metastable epialleles D1)[34]. The term metastable epiallele has been applied to genes that show variable expression because of probabilistic determinants of epigenetic repression[37]. An example of a metastable epiallele in mice is the A^{vy} locus (encoding agouti viable yellow); coat colors of isogenic mice vary on the basis of the epigenetic state of a retrotransposon integrated near the *A* promoter[38]. SmcHD1 is a modifier of metastable epialleles, as *Smchd1* haploinsufficiency results in higher penetrance of agouti expression[34]. In the case of FSHD, lower levels of SMCHD1 resulted in lower D4Z4 CpG methylation and variegated expression of DUX4 in myonuclei. In both FSHD1 and FSHD2, the penetrance is incomplete, and the presentation is often asymmetric. Out of the 26 individuals with hypomethylation at D4Z4 with a *SMCHD1* mutation and carrying a permissive D4Z4 haplotype, 5 (19%) are asymptomatic (Supplementary Table 2). This proportion of clinically unaffected carriers is notably similar to that observed in FSHD1 (ref. 39), although a recent publication corroborates an earlier observation that non-penetrance may be much more frequent[40, 41]. Thus, both features are consistent with FSHD being a metastable epiallele–linked disease. Our demonstration that independently variable modifiers of D4Z4 chromatin relaxation (repeat size in FSHD1 and SMCHD1 activity in FSHD2) modulate the variegated expression of DUX4 suggests that *DUX4* should be regarded as a metastable epiallele causing phenotypic variation in humans.

The disease mechanisms of FSHD1 and FSHD2 converge at the level of D4Z4 chromatin relaxation and the variegated expression of *DUX4* (refs. 14,15). Both FSHD1 and FSHD2 require inheritance of two independent genetic variations: a version of the *DUX4* gene with a polyadenylation signal and a second genetic variant that results in D4Z4 chromatin relaxation. For FSHD1, the genetic variant associated with chromatin relaxation involves contraction of the D4Z4 array and is therefore transmitted as a dominant trait. For FSHD2, mutations in *SMCHD1*, which is on chromosome 18, segregate independently from the FSHD-permissive *DUX4* allele on chromosome 4 and result in a digenic inheritance pattern in affected kindreds. Considering the variable clinical severity and asymmetric disease presentation, as well as the families with FSHD2 without *SMCHD1* mutations, it is likely that other modifier loci will be identified that affect D4Z4 chromatin structure. *SMCHD1* mutations could also modify the penetrance of FSHD1. Finally, many other human diseases show variable penetrance that might be related to

epigenetic control. Our findings establish the possibility that *SMCHD1* mutations may modify the epigenetic repression of other genomic regions and the penetrance of other human diseases.

Study subjects and samples.

Forty-one individuals with FSHD2 were selected on the basis of published clinical and molecular criteria[5, 7, 43, 44] and because they had D4Z4 methylation levels of <25% (Supplementary Table 1). Assessment of the FSHD phenotype was performed by experienced neurologists. Initial testing was performed using pulsed-field gel electrophoresis and hybridization of Southern blots with P13E-11, A and B probes, and SSLP length was determined using an ABI Prism 3100 Genetic Analyzer[41, 45, 46] according to protocols at the Fields Center at the University of Rochester for FSHD Research website (see URLs). Forty of the affected individuals had D4Z4 array sizes of >10 units on both copies of chromosome 4, ruling out diagnosis with FSHD1. One affected individual had two contracted alleles on chromosome 10, possibly explaining the low D4Z4 methylation observed for this subject, and was therefore excluded from further studies. Of the 39 remaining affected families, we had sufficient family information for 13 suggesting dominant inheritance of D4Z4 hypomethylation, whereas the hypomethylation seemed to have occurred *de novo* in 7 affected individuals (Fig. 1b). For exome sequencing, we selected five families with a dominant inheritance pattern and two with *de novo* hypomethylation in the affected individual. In total, 14 individuals from these families were analyzed by exome sequencing. All participants provided written consent, and the institutional review boards (IRBs) of participating institutes approved all studies.

D4Z4 methylation analysis.

Genomic DNA was double digested with EcoRI and BglII overnight at 37 °C, and cleaved DNA was purified using PCR extraction columns (Supplementary Note). Purified DNA digested with EcoRI and BglII was digested with FseI for 4 h, separated by size on 0.8% agarose gels, transferred to a nylon membrane (Hybond XL, Amersham) by Southern blotting and probed using the p13E-11 radiolabeled probe[22]. Probe signals were quantified using a phosphorimager and ImageQuant software. The signal from the 4,061-bp fragment was divided by the total amount of hybridizing fragments at 4,061 bp (methylated) and 3,387 bp (unmethylated) to give

the percentage of methylated FseI sites within the most proximal D4Z4 unit (Supplementary Note).

Exome definition, array design and target masking.

We targeted all protein-coding regions as defined by RefSeq 36.3. Entries were filtered for (i) CDS as the feature type, (ii) transcript name starting with 'NM_' or '-', (iii) reference as the group_label and (iv) not being on an unplaced contig (for example, 17|NT_113931.1). Overlapping coordinates were collapsed for a total of 31,922,798 bases over 186,040 discontiguous regions. A single custom array (Agilent, 1 million features, array–comparative genomic hybridization (aCGH) format) was designed to have probes over these coordinates as previously described, except that the maximum melting temperature (T_m) was raised to 73 °C.

The mappable exome was also determined as previously described, instead using the RefSeq 36.3 exome definition. After masking for 'unmappable' regions, 30,923,460 bases remained as the mappable target.

Targeted capture and massively parallel sequencing.

Genomic DNA was extracted from peripheral blood lymphocytes using standard protocols. DNA (5 μg) from each of the eight individuals was used for construction of a shotgun sequencing library as described previously, using paired-end adaptors for sequencing on an Illumina Genome Analyzer IIx (GAIIx). Each shotgun library was hybridized to an array for target enrichment, which was followed by washing, elution and additional amplification. Enriched libraries were then sequenced on a GAIIx to generate either single-end or paired-end reads.

Read mapping and variant analysis.

Reads were mapped and processed largely as previously described. In brief, reads were quality recalibrated using Eland and then aligned to the reference human genome (hg19) using MAQ. When reads with the same start site and orientation were filtered, paired-end reads were treated as separate single-end reads; this method is overly conservative, and, hence, the actual coverage of the exomes was higher than reported here. Sequence calls were performed using MAQ, and these calls were filtered to coordinates with ≥8× coverage and consensus quality of ≥20.

Insertions and/or deletions (indels) affecting coding sequences were identified as previously described, but we used phaster instead of cross_match and MAQ. Specifically, unmapped reads from MAQ were aligned to the reference sequence using phaster (version 1.100122a) with the parameters -max_ins:21 -max_del:21 -gapextend_ins:-1 -gapextend_del:-1 -match_report_type:1. Reads were then filtered for those with at most two substitutions and one indel. Reads that mapped to the negative strand were reverse complemented and, together with the other filtered reads, were remapped using the same parameters to reduce ambiguity in the called indel positions. These reads were then filtered for (i) having a single indel more than 3 bp from the ends and (ii) having no other substitutions in the read. Putative indels were then called per individual if they were supported by at least two filtered reads that started from different positions. An indel reference was generated as previously described, and all the reads from each individual were mapped back to this reference using phaster with default settings and -match_report_type:1. Indel genotypes were called as previously described.

To determine whether the variants were novel, sequence calls were compared against our previously reported exome data for 1,200 individuals and the 1000 Genomes Project database and dbSNP. Annotations of variants were made on the basis of information in the NCBI and UCSC databases using an in house server (SeattleSeq Annotation). Loss-of-function variants were defined as nonsense mutations (introduction of a premature stop codon) or frameshift indels. For each variant, we also generated constraint scores, as implemented in genomic evolutionary rate profiling (GERP).

Ranking of candidate genes.

Candidate genes were ranked by summation of variant scores calculated by counting the total number of nonsense and nonsynonymous variants across the FSHD2 exomes.

Mutation validation.

Sanger sequencing of PCR amplicons (LGTC, Leiden, Netherlands) from genomic DNA was used to confirm the presence and identity of mutations in *SMCHD1* that were initially detected via exome sequencing and to screen each mutation in the affected and unaffected family members in families with FSHD2.

Cells and culture conditions

Primary human myoblasts were obtained through the Fields Center. Biopsies were obtained after obtaining full consent with an IRB-approved protocol. Consents included the possibility of exome sequencing and sharing of samples with other investigators. Normal human myoblasts were grown on dishes coated with 0.01% calf skin collagen (Sigma-Aldrich) in F10 medium (Invitrogen) supplemented with 20% FBS, 100 U/ml penicillin and 100 μg/ml streptomycin, 4 μg/ml human basic fibroblast growth factor (bFGF) (Invitrogen) and 1 μM dexamethasone (Sigma-Aldrich) in a humidified atmosphere containing 5% CO_2 at 37 °C[13]. Transduction of human myoblasts with retroviral vectors was accomplished by seeding cells at a density of 5×10^4 cells/cm^2 on day −1. On day 0, the medium was changed, and cells were incubated with vector preparations and polybrene (4 μg/ml; Sigma-Aldrich). After 2–4 h, the medium was replaced with fresh medium, and cells were cultured and split at ~75% confluence to prevent differentiation. Human myoblasts transduced with pGIPZ shRNA expression vectors were selected with puromycin (0.5 μg/ml). Differentiation was induced using F10 medium supplemented with 1% horse serum and ITS supplement (insulin 0.1%, 0.000067% sodium selenite, 0.055% transferrin; Invitrogen).

Fibroblasts obtained from individuals with FSHD2 and their family members were cultured in DMEM/F-12 medium supplemented with 20% heat-inactivated FBS, 1% penicillin-streptomycin, 10 mM HEPES and 1 mM sodium pyruvate (all from Invitrogen).

RNA extraction and cDNA synthesis.

Total RNA was extracted using the Qiagen miRNeasy Mini isolation kit with DNase I treatment. The RNA concentration was determined on an ND-1000 spectrophotometer (Thermo-Scientific), and RNA quality was analyzed with an RNA 6000 Nanochip Labchip on an Agilent 2100 Bioanalyzer (Agilent Technologies Netherlands). cDNA was synthesized from 2 μg of total RNA using random hexamer primers (Fermentas) and the RevertAid H Minus M-MuLV First Strand kit (Fermentas Life Sciences) according to the manufacturer's instructions. After completion of cDNA synthesis, 30 μl of water was added to an end volume of 50 μl.

Semiquantitative RNA analysis and sequencing of *SMCHD1* mutations.

Splicing alterations were analyzed by RT-PCR using different primer sets covering the exons surrounding the possible splice-site mutation. Subsequently, PCR fragments generated from control samples and individuals heterozygous for *SMCHD1* mutations were analyzed on 1.5–2% agarose gels. Fragments were separated by size on agarose gels, purified and analyzed by Sanger sequencing (LGTC).

Allelic expression analysis of missense mutations (wild-type versus mutant alleles) was carried out with Sanger sequencing (LGTC) by comparison of the nucleotide peak heights of the wild-type and mutant alleles.

DUX4 mRNA levels were analyzed in duplicate by RT-PCR using the SYBR Green QPCR master mix kit (Stratagene) on a MyiQ (Bio-Rad), running an initial denaturation step at 95 °C for 6 min followed by 40 cycles of 10 s at 95 °C and 30 s at 60 °C (35 cycles for the *DUX4* RT-PCR samples shown in Fig. 3e,f). All PCR products were analyzed on a 2% agarose gel. Expression levels were corrected by those of *GAPDH* and *GUSB*, constitutively expressed standards for cDNA input, and the relative steady-state RNA levels of the genes of interest were calculated by a previously described method[47]. All primers were designed using Primer3 software, and sequences are provided in Supplementary Table 3.

ChIP assays.

Chromatin was prepared from myoblast cell lines fixed with 1% formaldehyde according to a published protocol[48]. Control and FSHD2 myoblasts carried a comparable total number D4Z4 repeat units on permissive and nonpermissive chromosomes. We incubated 60 μg of chromatin with the different antibodies. Every sample was independently studied twice. Antibodies against SMCHD1 (ab31865) and histone H3 (ab1791) were purchased from Abcam. Normal rabbit serum was used to measure unspecific binding of proteins to beads. Immunopurified DNA was quantified with the D4Z4 Q-PCR primer pair[8], and quantitative PCR measurements were performed with the CFX96 Real-Time PCR Detection System using iQ SYBRR Green Supermix. Relative enrichment values were calculated by dividing the ChIP values obtained with the antibodies to SMCHD1 or IgG by the ChIP values obtained with the antibodies to histone H3.

Antisense-mediated exon skipping.

Antisense oligonucleotides for *SMCHD1* exons 29 (29AON5) and 36 (36AON1) were designed on the basis of the guidelines for *DMD* exons (Supplementary Table 3)[49]. All antisense oligonucleotides target exon-intern sequences, consist of 2′-O-methyl RNA with a full-length phosphorothioate backbone and were manufactured by Eurogentec. Human control myoblasts were seeded in 6-well plates or 6-cm dishes at a density of approximately 1×10^4 cells/cm^2 and were cultured for 2 d. Myotubes were obtained by growing myoblasts at 70% confluence for 4 d in differentiation medium (DMEM (with glucose, L-glutamine and pyruvate) supplemented with 2% horse serum). Cells were transfected with 250 nM concentrations of antisense oligonucleotides 4 h after the differentiation medium was added, using 2.5 µl of polyethyleneimine (MBI-Fermentas) per microgram of antisense oligonucleotide according to the manufacturer's instructions. A FAM-labeled antisense oligonucleotide targeting exon 50 of the *DMD* gene was used to confirm the efficiency of transfection and exon skipping. Primers flanking the targeted exons were used to study splicing of the *SMCHD1* and *DMD* genes.

Knockdown of *SMCHD1* mRNA in normal human myoblasts.

SMCHD1 transcripts were targeted for degradation using lentiviral vectors expressing shRNAs from a CMV promoter linked to a puromycin selection cassette controlled by an internal ribosome entry site (IRES). Five different pGIPZ vectors (Open Biosystems) were purchased, and each was tested in normal human myoblasts for the effect on *SMCHD1* transcripts by quantitative PCR, immunofluorescence signal intensity and protein blot analysis.

Antibodies, immunofluorescence and protein blotting.

Immunofluorescence for human DUX4 was performed using a rabbit monoclonal antibody specific to its C terminus (Epitomics, E5-5), as previously described[15]. Immunoreactivity was detected with a mouse Alexa Fluor 594–conjugated secondary antibody to rabbit (Molecular Probes; 1:1,000 dilution).

For protein blotting, fibroblast or myoblast lysates were separated by 7.5% SDS-PAGE and transferred to PVDF membrane. SMCHD1 protein was detected using a commercially available rabbit polyclonal antibody (Sigma-Aldrich, HPA039441; 1:250 dilution), and the reference

protein tubulin was detected with a commercially available mouse monoclonal antibody (Sigma, T6199; 1:2,000 dilution). Bound antibodies were detected with a horseradish peroxidase (HRP)-conjugated donkey secondary antibody to rabbit (Pierce, 31458; 1:5,000 dilution) and an IRDye 800CW–conjugated goat secondary antibody to mouse (Westburg, 926-32210; 1:5,000 dilution), respectively.

Acknowledgments

The authors thank all subjects and family members for their participation. We thank D. Nickerson and J. Shendure for excellent assistance and B. Trask for helpful discussions and critical reading of the manuscript. This work was supported by grants from the US National Institutes of Health (NIH) (National Institute of Neurological Disorders and Stroke (NINDS) P01NS069539, Clinical & Translational Science Award (CTSA) UL1RR024160, National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) R01AR045203 and National Human Genome Research Institute (NHGRI) HG005608 and HG006493), a Netherlands Genomics Initiative (NGI) Horizon Valorization Project Grant (93515504), The University of Washington Center for Mendelian Genomics, the Muscular Dystrophy Association (MDA; 217596), the Fields Center for FSHD Research, the Geraldi Norton and Eklund family foundation, the FSH Society, The Friends of FSH Research, European Union Framework Programme 7 agreements 223026 (NMD-chip), 223143 (TechGene) and 2012-305121 (NEUROMICS) and the Stichting FSHD. Y.S. is supported by the China Scholarship Council.

URLs.

SAMtools, http://samtools.sourceforge.net/; Picard, http://picard.sourceforge.net/; SeattleSeq Annotation, http://snp.gs.washington.edu/SeattleSeqAnnotation131/;1000 Genomes Project, http://www.1000genomes.org/; Alamut, http://www.interactive-biosoftware.com/; Mutalyzer 2.0.beta-21, https://mutalyzer.nl/; NCBI, http://www.ncbi.nlm.nih.gov/; GeneCards, http://www.genecards.org/; Ensembl, http://www.ensembl.org/Homo_sapiens/Info/Index; FSHD genotyping and methylation analysis protocols, http://www.urmc.rochester.edu/fields-center/.

References

1. Statland, J.M. & Tawil, R. Facioscapulohumeral muscular dystrophy: molecular pathological advances and future directions. *Curr. Opin. Neurol.* 2011;24:423–428.
2. Balog, J. *et al*. Correlation analysis of clinical parameters with epigenetic modifications in the *DUX4* promoter in FSHD. *Epigenetics*. 2012;7:579–584.
3. Bodega, B. *et al*. Remodeling of the chromatin structure of the facioscapulohumeral muscular dystrophy (FSHD) locus and upregulation of FSHD-related gene 1 (*FRG1*) expression during human myogenic differentiation. *Bmc Biol.* 2009; 7:41.
4. Cabianca, D.S. *et al*. A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell*. 2012;149:819–831.
5. de Greef, J.C. *et al*. Common epigenetic changes of D4Z4 in contraction-dependent and contraction-independent FSHD. *Hum. Mutat.* 2009; 30:1449–1459.
6. Jiang, G. *et al*. Testing the position-effect variegation hypothesis for facioscapulohumeral muscular dystrophy by analysis of histone modification and gene expression in subtelomeric 4q. *Hum. Mol. Genet.* 2003;12:2909–2921.
7. van Overveld, P.G. *et al*. Hypomethylation of D4Z4 in 4q-linked and non-4q–linked facioscapulohumeral muscular dystrophy. *Nat. Genet.* 2003;35:315–317.
8. Zeng, W. *et al*. Specific loss of histone H3 lysine 9 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). *Plos Genet.* 2009;5:e1000559.
9. Gabriëls, J. *et al*. Nucleotide sequence of the partially deleted D4Z4 locus in a patient with FSHD identifies a putative gene within each 3.3 kb element. *Gene*.1999;236:25–32.
10. Hewitt, J.E. *et al*. Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. *Hum. Mol. Genet.* 1994;3:1287–1295.
11. Lyle, R., Wright, T.J., Clark, L.N. & Hewitt, J.E. The FSHD-associated repeat, D4Z4, is a member of a dispersed family of homeobox-containing repeats, subsets of which are clustered on the short arms of the acrocentric chromosomes. *Genomics* 1995;28:389–397.
12. Snider, L. *et al*. RNA transcripts, miRNA-sized fragments and proteins produced from D4Z4 units: new candidates for the pathophysiology of facioscapulohumeral dystrophy. *Hum. Mol. Genet.* 2009;18:2414–2430.
13. Snider, L. *et al*. Facioscapulohumeral dystrophy: incomplete suppression of a retrotransposed gene. *Plos Genet.* 2010;6: e1001181.
14. van der Maarel, S.M., Tawil, R. & Tapscott, S.J. Facioscapulohumeral muscular dystrophy and DUX4: breaking the silence. *Trends Mol. Med.* 2011;17:252–258.
15. Geng, L.N. *et al*. DUX4 activates germline genes, retroelements, and immune mediators: implications for facioscapulohumeral dystrophy. *Dev. Cell*. 2012;22:38–51.
16. Bosnakovski, D. *et al*. An isogenetic myoblast expression screen identifies DUX4-mediated FSHD-associated molecular pathologies. *Embo J.* 2008;27:2766–2779.
17. Kowaljow, V. *et al*. The *DUX4* gene at the FSHD1A locus encodes a pro-apoptotic protein. *Neuromuscul. Disord.* 2007;17: 611–623.
18. Vanderplanck, C. *et al*. The FSHD atrophic myotube phenotype is caused by DUX4 expression. *Plos One.* 2011;6:e26820.

19. Wallace, L.M. *et al*. *DUX4*, a candidate gene for facioscapulohumeral muscular dystrophy, causes p53-dependent myopathy *in vivo*. *Ann. Neurol.* 2011;69:540–552.

20. Wuebbles, R.D., Long, S.W., Hanel, M.L. & Jones, P.L. Testing the effects of FSHD candidate gene expression in vertebrate muscle development. *Int. J. Clin. Exp. Pathol.* 2010;3:386–400.

21. van Deutekom, J.C. *et al*. FSHD associated DNA rearrangements are due to deletions of integral copies of a 3.2 kb tandemly repeated unit. *Hum. Mol. Genet.* 1993;2:2037–2042.

22. Wijmenga, C. *et al*. Chromosome 4q DNA rearrangements associated with facioscapulohumeral muscular dystrophy. *Nat. Genet.* 1992;2:26–30.

23. Dixit, M. *et al*. *DUX4*, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of *PITX1*. *Proc. Natl. Acad. Sci. Usa*. 2007;104:18157–18162.

24. Lemmers, R.J. *et al*. Facioscapulohumeral muscular dystrophy is uniquely associated with one of the two variants of the 4q subtelomere. *Nat. Genet.* 2002;32:235–236.

25. Lemmers, R.J. *et al*. A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*. 2010;329:1650–1653.

26. Spurlock, G., Jim, H.P. & Upadhyaya, M. Confirmation that the specific SSLP microsatellite allele 4qA161 segregates with fascioscapulohumeral muscular dystrophy (FSHD) in a cohort of multiplex and simplex FSHD families. *Muscle Nerve.* 2010;42:820–821.

27. Thomas, N.S. *et al*. A large patient study confirming that facioscapulohumeral muscular dystrophy (FSHD) disease expression is almost exclusively associated with an FSHD locus located on a 4qA-defined 4qter subtelomere. *J. Med. Genet.* 2007;44:215–218.

28. Bamshad, M.J. *et al*. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* 2011;12:745–755.

29. Bhalla, N., Biggins, S. & Murray, A.W. Mutation of *YCS4*, a budding yeast condensin subunit, affects mitotic and nonmitotic chromosome behavior. *Mol. Biol. Cell*. 2002;13:632–645.

30. Lieb, J.D., Capowski, E.E., Meneely, P. & Meyer, B.J. DPY-26, a link between dosage compensation and meiotic chromosome segregation in the nematode. *Science.*1996;274:1732–1736.

31. Chuang, P.T., Albertson, D.G. & Meyer, B.J. DPY-27:a chromosome condensation protein homolog that regulates *C. elegans* dosage compensation through association with the X chromosome. *Cell*. 1994;9:459–474.

32. Dej, K.J., Ahn, C. & Orr-Weaver, T.L. Mutations in the *Drosophila* condensin subunit dCAP-G: defining the role of condensin for chromosome condensation in mitosis and gene expression in interphase. *Genetics.* 2004;168:895–906.

33. Kanno, T. *et al*. A structural-maintenance-of-chromosomes hinge domain–containing protein is required for RNA-directed DNA methylation. *Nat. Genet.* 2008;40:670–675.

34. Blewitt, M.E. *et al*. An N-ethyl-N-nitrosourea screen for genes involved in variegation in the mouse. *Proc. Natl. Acad. Sci. Usa*. 2005;102:7629–7634.

35. Blewitt, M.E. *et al*. SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. *Nat. Genet.* 2008;40:663–669.

36. Gendrel, A.V. *et al*. Smchd1-dependent and -independent pathways determine developmental dynamics of CpG island methylation on the inactive X chromosome. *Dev. Cell*. 2012;23:265–279.

37. Rakyan, V.K., Blewitt, M.E., Druker, R., Preis, J.I. & Whitelaw, E. Metastable epialleles in mammals. *Trends Genet.* 2002;18:348–351.

38. Duhl, D.M., Vrieling, H., Miller, K.A., Wolff, G.L. & Barsh, G.S. Neomorphic agouti mutations in obese yellow mice. *Nat. Genet.* 1994;8:59–65.

39. van der Maarel, S.M., Frants, R.R. & Padberg, G.W. Facioscapulohumeral muscular dystrophy. *Biochim. Biophys. Acta*. 2007;1772:186–194.

40. Scionti, I. *et al*. Facioscapulohumeral muscular dystrophy: new insights from compound heterozygotes and implication for prenatal genetic counselling. *J. Med. Genet.* 2012;49:171–178.

41. Lemmers, R.J. *et al*. Specific sequence variations within the 4q35 region are associated with facioscapulohumeral muscular dystrophy. *Am. J. Hum. Genet.* 2007;81:884–894.

42. Lemmers, R.J. *et al*. Contractions of D4Z4 on 4qB subtelomeres do not cause facioscapulohumeral muscular dystrophy. *Am. J. Hum. Genet.* 2004;75:1124–1130.

43. van Overveld, P.G. *et al*. Variable hypomethylation of D4Z4 in facioscapulohumeral muscular dystrophy. *Ann. Neurol.* 2005;58:569–576.

44. de Greef, J.C. *et al*. Clinical features of facioscapulohumeral muscular dystrophy 2. *Neurology*. 2010;75:1548–1554.

45. Lemmers, R.J.L. *et al*. Complete allele information in the diagnosis of facioscapulohumeral muscular dystrophy by triple DNA analysis. *Ann. Neurol.* 2001;50:816–819.

46. Lemmers, R.J.L.F. *et al*. Worldwide population analysis of the 4q and 10q subtelomeres identifies only four discrete duplication events in human evolution. *Am. J. Hum. Genet.* 2010;86:364–377.

47. Pfaffl, M.W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 2001;29: e45.

48. Nelson, J.D., Denisenko, O. & Bomsztyk, K. Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nat. Protoc.* 2006;1:179–185.

49. Aartsma-Rus, A. Overview on AON design. *Methods Mol. Biol.* 2012;867:117–129.

Supplementary Note

Genomic DNA isolated from peripheral blood lymphocytes from a large panel of controls, sporadic patients with FSHD and FSHD families were included in this study after obtaining informed consent. The clinical diagnosis of FSHD was based on a standardized clinical form made available through the Fields Center: http://www.urmc.rochester.edu/fields-center/). For all individuals we performed a detailed genotyping, including D4Z4 repeat array length and chromosomal background analysis of chromosomes 4q and 10q.

The observation that in FSHD1 patients D4Z4 hypomethylation is restricted to the disease allele while in FSHD2 patients the repeats on all four chromosomes are affected provides a unique opportunity to develop a more sensitive and specific diagnostic test for FSHD2. Rather than separating the chromosome 4-derived fragments from the chromosome 10-derived fragments by using restriction enzyme BlnI, as done before [1, 2], a collective measurement of D4Z4 methylation on both chromosomes 4 and 10 should yield a more sensitive and specific diagnostic test for FSHD2. From our previous tests involving three methylation-sensitive restriction enzymes, FseI was shown to be the most informative enzyme[1, 2]. Therefore, we redesigned the FseI D4Z4 methylation test so that it interrogates all four alleles simultaneously by omitting BlnI from the digestion (Supplementary Fig. 1). Previously, we showed that the FseI methylation value of the first D4Z4 unit in controls is ~50% on both chromosomes 4q[1,2]. The average FseI methylation level of the first unit in pathogenic chromosomes 4 in FSHD1 patients (n=21) was shown to be 20%[3], while in FSHD2 patients we found for both chromosomes 4 on average a value of 13% (n=32)[4]. While in controls and FSHD1 patients we would expect (near-) normal methylation values (as in FSHD1 the hypomethylation signal from the disease allele would be diluted 3x by the normal methylation levels of the normal chromosome 4 and chromosomes 10), in FSHD2 patients we would expect to see profound hypomethylation. As the activity of restriction enzymes is sensitive to salt or protein impurities in the gDNA we introduced an extra DNA clean-up step preceding digestion with FseI (Supplementary Fig. 1a). This extraction column-based purification step can also be applied to gDNA embedded in agarose plugs and to samples with low gDNA concentrations. Upon digesting with BglII a 4061 bp fragment is released (M in Supplementary Fig. 1c) while digesting with FseI yields a fragment of 3387 bp when the restriction site is unmethylated (U in Supplementary Fig. 1c). The previously used enzyme BlnI

to separate chromosomes 4 (white) from chromosomes 10 (black) is also shown. To validate the modified methylation test, we re-analyzed the same gDNA samples from a previous study[4]. While we obtained nearly identical average methylation levels in all three populations analyzed, the modified methylation test clearly improves discrimination between FSHD1 and FSHD2 by reducing the error bars particularly in FSHD1 patients (Supplementary Fig. 1d). Supplementary Fig. 1b shows a typical example of the D4Z4 methylation analysis on a de novo FSHD2 patient and his unaffected family members. The FSHD2 patient has comparable methylation levels (%) to her unaffected mother who carries a non-permissive alleles (NP) only. The unaffected father has significant lower methylation levels than mother and daughter as quantified by fragment intensities.

Supplementary Table 1 Selection criteria used to prioritize FSHD2 families for whole exome sequencing.

Criteria	Number of Families
FseI Methylation <25%	41
Both chromosome 4q D4Z4 arrays > 10 units	40
Not more than one chromosome 10q D4Z4 array <11 units	39
Inheritance Pattern:	
Dominant inheritance	13
de novo D4Z4 hypomethylation	7
Unknown (not informative)	19

Maximum D4Z4 methylation at FseI site in patients was set at 25%. We excluded families in which one of the individuals with D4Z4 methylation <25% had a D4Z4 repeat array of <10 units on a permissive allele or more than one array of <10 units. Families were further categorized according to the inheritance pattern of D4Z4 hypomethylation.

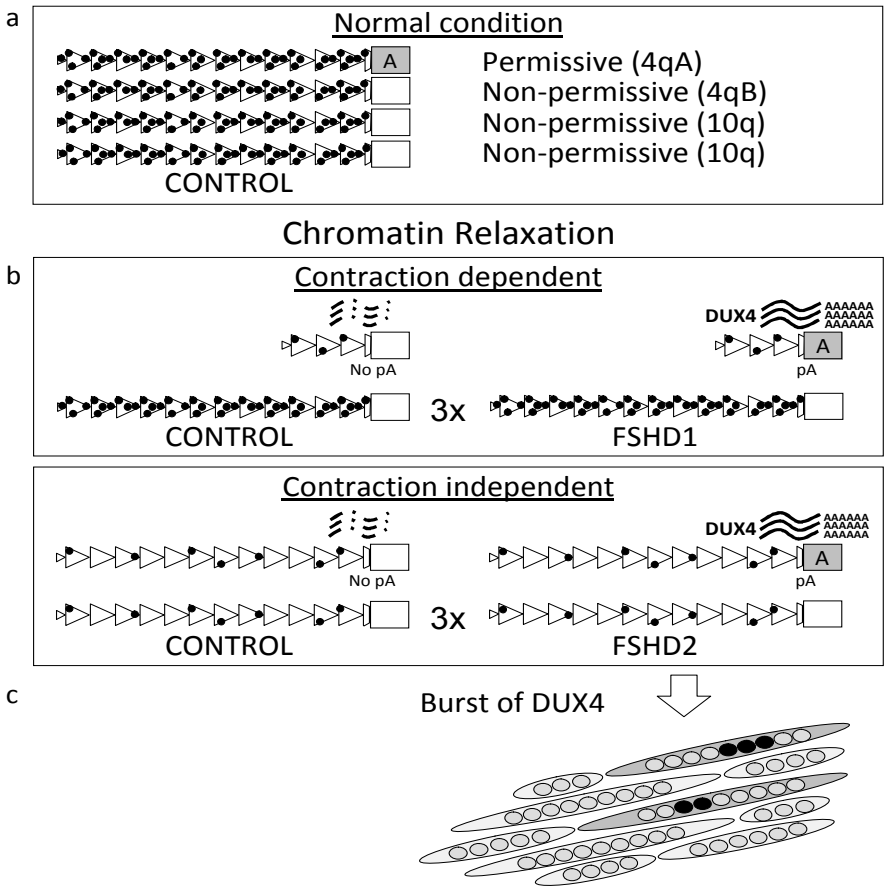
Supplementary Table 2 Information on unaffected SMCHD1 heterozygotes with a permissive D4Z4 haplotype.

Family	Individual	Gender	age	FseI	Units
Rf392	102	F	54	17	50U
Rf393	101	M	75	11	89U
Rf393	206	F	42	11	18U
Rf393	302	F	27	19	20U
Rf393	305	M	34	21	20U

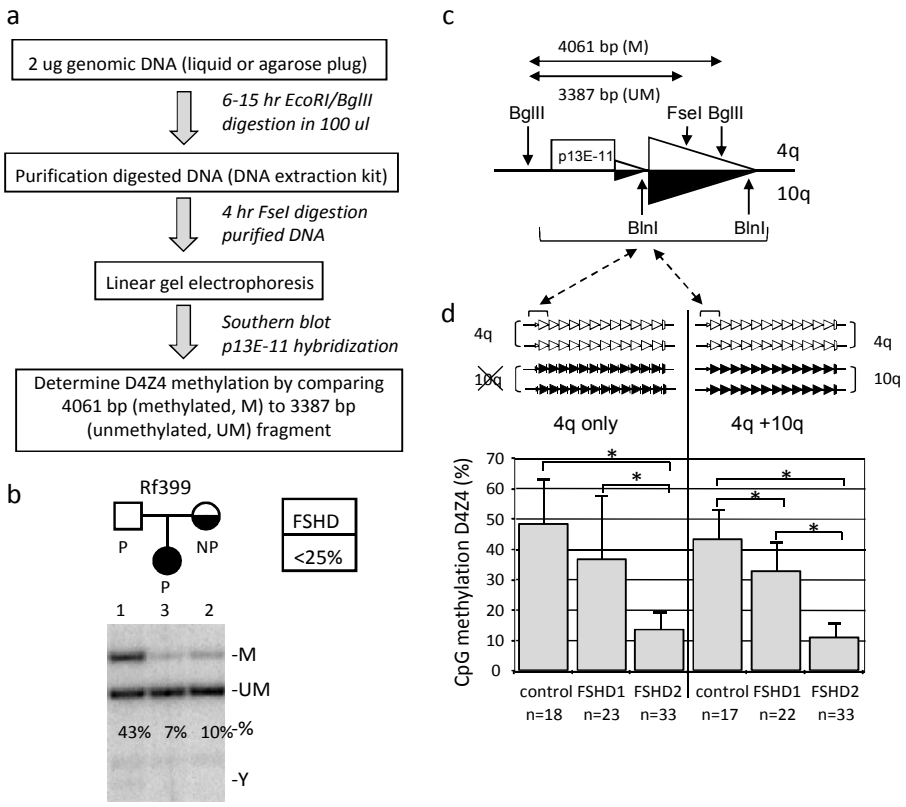
Indicated are family ID, individual ID, gender, age, FseI methylation level and D4Z4 array size in units (U) of smallest permissive allele.

Supplementary Table 3 Sequences primers and antisense oligo nucleotides used in this study

Primers for analysis SMCHD1 splicing at exons 12, 25, 29 and 36		
Name	**Sequence (5' to 3')**	**position SMCHD1**
exon 10F	5'-TGA TCC ATG CTT TCC ATC AA-3'	NM015295_1512F
exon 13R	5'-CCT TCA GCC ACA AAG CAA AT-3'	NM015295_1882R
exon 24F	5'-TCT GGA ACC AGT ATT TTA ACA GGA-3'	NM015295_3151F
exon 26R	5'-TTG CAC ATC AGG AAG CAG AC-3'	NM015295_3518R
exon 28F	5'-CTG GGG TTG GAC TTG ATA GC-3'	NM015295_3779F
exon 30R	5'-GGT GCT GGA TTA TCC CAC TG-3'	NM015295_4070R
exon 35F	5'-TCC AGT TTG GTT TTA TGA TGG A-3'	NM015295_4574F
exon 37R	5'-TTC ACG AAG GGG AAT TCA AG-3'	NM015295_4889R
qPCR primers		
Name	**Sequence (5' to 3')**	**position SMCHD1**
SMCHD1_F (exon 47F)	5'- CGA CAG ATT GTC CAG TTC CTC-3'	NM015295_6125F
SMCHD1_R (exon 48R)	5'- CCA ATG GCC TCT TCT CTC TG-3'	NM015295_6225R
DUX4RT-F2	5'-CCC AGG TAC CAG CAG ACC-3'	
DUX4-pLAMR4	5'-TCC AGG AGA TGT AAC TCT AAT CCA-3'	
hGAPDHFw	5'-AGC ACA TCG CTC AGA CAC-3'	
hGAPDHRev	5'-GCC CAA TAC GAC CAA ATC C-3'	
qPCR GUS fw	5'-CTC ATT TGG AAT TTT GCC GAT T-3'	
qPCR GUS rev	5'-CCG AGT GAA GAT CCC CTT TTT A-3'	
ChIP primers		
Name	**Sequence (5' to 3')**	
DUX4 ChIP F	5'-CCG CGTC CGT CCG TGA AA-3'	
DUX4 ChIP R	5'-TCC GTC GCC GTC CTC GTC-3'	
GAPDH ChIP F	5'-CTG AGC AGT CCG GTG TCA CTA C-3'	
GAPDH ChIP R	5'-GAG GAC TTT GGG AAC GAC TGA G-3'	
Antisense oligo nucleotide		
Name	**Sequence (5' to 3')**	**position**
29AON5	5'-GUC CAG AAA UUA GUU GCA CUC-3'	exon 29 SMCHD1
36AON1	5'-GAU UAG GCA GGA CUU CAA CU-3'	exon 36 SMCHD1
h50AON2	5'-(6FAM)-GGC UGC UUU GCC CUC-3'	exon 50 DMD

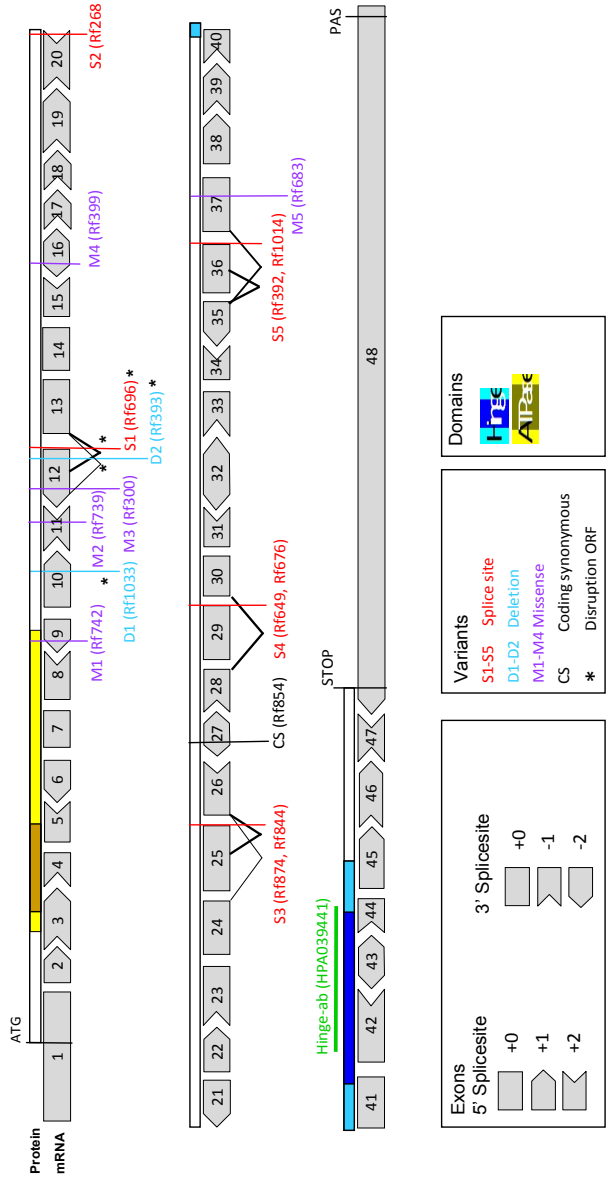


Supplementary Figure 1 Schematic of the FSHD locus. Different combinations of D4Z4 chromatin relaxation are shown with the associated chromosomal context and patient phenotype. The D4Z4 array is shown as a series of white triangles on chromosome 4. The homologous array on chromosome 10 is depicted in grey. The FSHD permissive 4qA, and FSHD non-permissive 10q and 4qB haplotypes are depicted as white and light grey boxes, respectively. (a) In the normal condition, D4Z4 arrays of >10 units are densely CpG methylated (black dots) on all four chromosomes. (b) FSHD1 is associated with D4Z4 array contraction-dependent D4Z4 hypomethylation and DUX4 expression from the deleted chromosome having a FSHD-permissive 4qA haplotype. Permissive 4qA haplotypes have a DUX4 polyadenylation signal (pA) distal to the last unit of the D4Z4 array. This pA signal results in stabilization of DUX4 mRNA. Contraction-dependent chromatin relaxation on non-permissive haplotypes (4qB or 10q) do not cause disease, because they lack this DUX4 pA signal. In FSHD1, D4Z4 hypomethylation is restricted to the contracted array. FSHD2 is caused by D4Z4 array contraction-independent chromatin relaxation of a D4Z4 locus with a permissive haplotype. In this case all four D4Z4 arrays are hypomethylated, and the hypomethylation phenotype can segregate independently of the permissive 4q haplotype within a family. Thus, family members who inherit the hypomethylation phenotype without a permissive haplotype do not develop FSHD2 (CONTROL). Chromosome 10 arrays are not depicted. (c) D4Z4 chromatin relaxation leads to a variegated production of the DUX4 protein in a subset of FSHD1 and FSHD2 myonuclei (black).

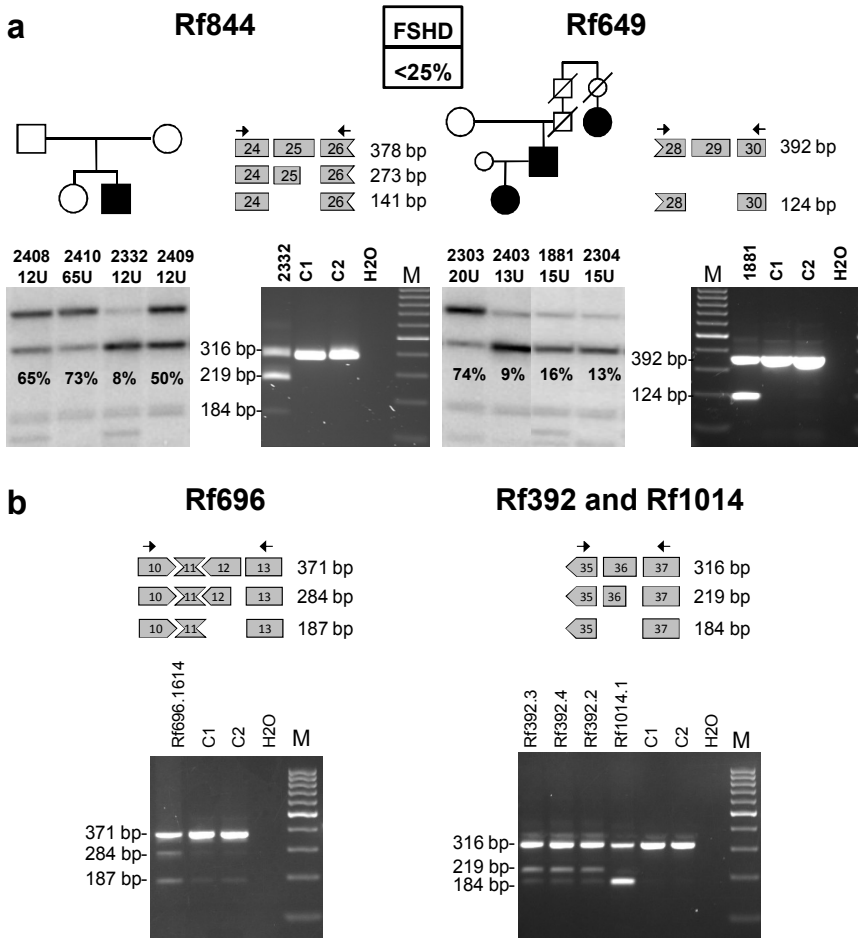


Supplementary Figure 2 Design and results of the D4Z4 methylation test.

(a) Overview of methylation analysis method. (b) Example of methylation analysis in an FSHD2 family. Methylated (M) and unmethylated (UM) D4Z4 fragments are indicated. Below each lane the methylation value is indicated in %. Y indicates cross hybridizing Y fragment. The hypomethylated mother in this family is not affected in the absence of a permissive haplotype. (c) Schematic of methylation test showing the p13E-11 probe region at the proximal end of the D4Z4 repeat array and the expected D4Z4 fragment sizes upon digestion with restriction enzymes EcoRI, BglII and FseI (EcoRI sites are not shown as they are outside the indicated area and the enzyme is only used for additional fragmentation of the gDNA). The position of the chromosome 10q-specific restriction enzyme BlnI (black bottom half) that was previously used for the chromosomes 4q only methylation analysis is indicated. (d) Schematic of FseI methylation analysis for both chromosomes 4 (old method; left panel)4 and chromosomes 4 and 10 (new method; right panel). Bar diagram of average methylation levels in controls (N=17), FSHD1 patients (N=22) and FSHD2 patients (N=33) obtained by the old method (left panel) and same samples by new method (right panel). Error bar represents standard deviation. FSHD2 patients are significantly hypomethylated by this test compared to controls and FSHD1 patients (*: $p<0.005$). Note that FSHD1 patients have methylation levels in between controls (normal methylation at all 4 alleles) and FSHD2 (hypomethylation at all 4 alleles) due to the presence of one hypomethylated allele.



Supplementary Figure 3 Schematic of the human SMCHD1 gene. All exons are indicated with boxes. Information about the SMCHD1 protein domains and Hinge antibody epitope is also given. mutations identified in this study are documented with their (predicted) consequences. The position of the 5 ' and 3 ' splice sites with respect to the coding frame is also indicated. Mutations that result in a frameshift are indicated by an asterisk.



Supplementary Figure 4 Examples of methylation analysis and alternative splicing in SMCHD1 heterozygotes (a) Pedigrees of sporadic (left panel) and familial (right panel) FSHD2 kindreds. Methylation analysis of the FseI site in D4Z4 shows the degree of methylation (left panels). SMCHD1 mRNA analysis in SMCHD1 heterozygotes and controls (C1-2) shows exon skipping or cryptic splice site usage (right panels). (b) RT- PCR analysis of SMCHD1 RNA in controls (C) and individuals heterozygous for SMCHD1 splice site mutations in families Rf696, Rf392 and Rf1014. RT-PCR products were sequence verified. Schematics of alternative splice events are shown on top and primers used to determine splicing are indicated with arrows. The splicing changes in family Rf696 can also be observed at lower frequency in the controls indicating that this variant shifts the balance (compare unspliced product with spliced products).

Chapter 8

General discussion

Molecular confirmation of a clinical diagnosis of an inherited disease or of congenital malformations is of paramount importance for patients and their families. It is the conclusion of the differential diagnostic process, and provides information on the prognosis, in some cases on the therapeutic options, and on the recurrence risk. The cycle of new emerging analytical techniques, the identification of genetic defects and genes, followed by further improvement in molecular diagnosis, is turning with an increasing speed and is contributing to better patient care and management.

Currently, targeted sequencing of gene (s) of interest is the preferred approach for searching for small pathogenic mutations. Several techniques are available for targeted sequencing, for example, conventional Sanger sequencing (1) and the Next Generation Sequencing (NGS) (2-4). Since its development, the Sanger sequencing method has gradually become the gold standard for clinical molecular diagnostics, because of its accuracy in detecting small genetic variants.

Sanger sequencing is often combined with other techniques in order to reduce the cost (5-9). We have implemented High Resolution Melting Curve Analysis (HR-MCA) to screen the entire coding sequence of the *DMD* gene to select fragments for sequencing. This process was quite straightforward, we used a gradient PCR-cycler to quickly determine the most optimal annealing temperature for PCR primers and to determine the number of melting domains for each amplicon. Although large amplicons (more than 600bp) and amplicons with more than three melting domains can be used for HR-MCA, the sensitivity is reduced and the risk of false positives is higher. To solve this problem we divided these amplicons into multiple fragments. HR-MCA requires neither specific skills nor special changes in the laboratory. It is a simple PCR combined with a saturation dye such as LCGreen. A potentially weak point of HR-MCA is that

homozygous or hemizygous variants may not be detected. We have, therefore, used post-PCR sample mixing to generate hetero-duplexes in all male patients with DMD/BMD. We have tested, validated, and adopted this technology for screening the *DMD* gene in patients as well as in female carriers in the Laboratory for Diagnostic Genome Analysis (LDGA) of the Department of Clinical Genetics in Leiden (**chapter 2**).

The diagnosis of monogenic genetic disorders, which depends on the size and complexity of the gene investigated, usually has a reasonable turnaround time with this combined strategy (HR-MCA followed by Sanger sequencing). However, when there are too many samples and/or too many possible candidate genes to be tested, this approach is time consuming, labour intensive and inefficient. Moreover, this method can be difficult or impossible to use in cases where no specific syndrome can be diagnosed, because of atypical or mild clinical features, and one cannot limit the number of candidate genes. Therefore, alternative strategies are needed to reduce time and cost for testing large numbers or even all of the genes.

NGS, which can currently access the primary structure of the entire genome of an individual (10), is likely to become a popular strategy to detect genetic variations that underlie human diseases. This is the ultimate goal but for the time being, because of the complexity of information and high costs, it is necessary to select and enrich particular genomic regions of interest before sequencing (11, 12).

We have tested long range PCR and capture by hybridization (on–array and in–solution). Long range PCR is potentially well suited for NGS platforms, but in practice, working with very long PCR fragments tends to be laborious, time consuming, and expensive. Each individual PCR of a given fragment with specific primers must be first tested and optimized. Also, not all reactions give the desired specific PCR products. Moreover, DNA with impurities or partial degradation does not amplify.

To overcome these problems, we have used the capture by hybridization methods (on–array and in–solution) (12-16). In principle, both the on–array and the in–solution hybridization work in the same way. We first hybridized the fragmented genomic DNA with common adapters to oligonucleotide probes in order to capture the target sequences. We then amplified the captured materials, tested the fold enrichment by quantitative PCR (qPCR) and performed NGS. We

found that qPCR is a crucial step to check successful enrichment as well as to estimate the fold-enrichment obtained for both on–array and in–solution capture methods. It offers a reliable, quick, and cheap check prior to NGS and in our hands all tested samples in which qPCR did not indicate a clear enrichment, results of sequencing were poor, indicating that these samples should not be included for further analysis (**chapter 3**).

Although on–array and in–solution share many similarities, there are several differences that make the in-solution method preferable. For instance, the amount of input DNA required by the in-solution method (around 500ng -1µg) is much less than that required by the on-array methodology, which requires at least 10µg. For this reason the in-solution method is cheaper, is easier to work with and can be used on samples where it is difficult to obtain sufficient amount of DNA. The in-solution method shows also other advantages over the on–array platform. The in–solution methodology is less laborious and less time consuming, does not require special equipment in the laboratory, is highly scalable and can be automated. The on- array capture method, on the other hand, requires lab-experience and expensive equipment such as a hybridization station and an elution apparatus. It is also difficult to automate.

The in-solution capture by hybridization is the ideal method to enrich any desired fragment in the genome. Most Mendelian disorders are caused by exonic or exon/intron junctions variants that alter the amino acid sequence of the affected gene. An exome represents only about 1% a of the human genome (17, 18). However, 85% of disease-related mutations found so far are located in the protein-coding regions (18). In classical strategies for identifying disease-associated mutations, homozygosity mapping or linkage analysis is performed by studying genetically related family members (19, 20). In informative families, candidate regions containing the disease gene may be narrowed down to a specific region. One can then systematically sequence the candidate region. Targeted enrichment, Exome Sequencing (ES) and NGS have brought new ways of addressing monogenic disorders (Mendelian disorders), because of their large capacity and unbiased survey of the sequenced region (21, 22). Previous linkage studies had mapped the potential mutated gene causing the X-linked dominant, male lethal disorder, Terminal Osseous Dysplasia (TOD), to Xq27.3-q28 (23). We used the linkage data to narrow down the candidate region and performed X exome sequencing in two unrelated patients (**chapter 4**). Furthermore, we used the linkage data to filter and select only the heterozygous variants located in the

previously identified TOD linkage interval. With this strategy we were able to identify c.5217G>A as the only heterozygous variant shared by the two patients in one gene, the *FLNA* gene, which causes the disease.

Another example of using the linkage data to narrow down the candidate region is in autosomal recessive spinocerebellar ataxia 7 which is linked to chromosome 11p15 (SCAR7) (24) (**chapter 5**). We investigated the entire coding sequence of this region. By selecting only a single affected individual for ES to obtain sequencing data, we could reduce the number of candidate genes to two (*TPP1* and *DCHS1* genes), for straightforward follow-up by Sanger sequencing. We found that the disease was caused by one splice variant and one missense variant in the *TPP1* gene.

Classical strategies can not be applied in many rare diseases where samples from large families is not available. In addition, a disease locus is not known in many syndromes with congenital malformations and/or intellectual disability. In all these cases an unbiased approach is required, for which ES is the best choice for the moment. The usefulness of ES for identifying causal variants for inherited disorders (recessive and dominant) is well established and many groups have identified the causative variants for a large number of Mendelian disorders (25, 26). Uncovering genetic defects that underlie different human disorders is one of the most obvious applications of ES. Moreover, ES has opened up new avenues towards understanding the mechanisms that underlie specific molecular pathogenesis of genetic disease. For example, we discovered that mutations in the gene *SMCHD1* (Structural Maintenance of Chromosomes flexible Hinge Domain containing 1) act as an epigenetic modifier of the D4Z4 metastable epiallele and thus cause the disease FacioScapuloHumeral Dystrophy type 2 (FSHD2) (**chapter 7**). Epigenetics refers to heritable changes in gene expression that are not caused by changes in DNA sequence and which play a major role in a variety of normal cellular processes. Key players in epigenetic control are DNA methylation and histone modifications (27). Disruption of either of these systems that contribute to epigenetic alterations can cause abnormal activation or silencing of genes and is known to result in various diseases states (27, 28). Thus, ES has provided a better understanding of the pathogenetic mechanism underlying FSHD2 where reducing *SMCHD1* levels in skeletal muscle results in contraction-independent DUX4 expression**.**

However, there are growing pains as we move forward with these new technologies. A key challenge is the interpretation of the enormous number of variants and the ability to identify disease-related alleles among the background of millions of neutral variants, polymorphisms, and sequencing errors, while in many cases we are not even sure whether a single pathogenic variant or a combination of several variants are causing disease. Several different strategies are available for filtering the variants found among the large numbers of sequences, and selecting the possible causal alleles (29). The number of candidate variants that are filtered depends on several factors such as: the mode of inheritance of a trait, the availability of a linkage or homozyosity mapping data, the degree of locus heterogeneity for a given trait, the availability of samples from patients with the same phenotype and the presence of a proper bioinformatics analysis pipeline for exome data.

With each type of disease the most crucial step is to define the character of variants to be prioritized. When looking for a gene causing a rare autosomal recessive disorder, candidate genes must show either homozygous or compound heterozygous variants. With ES one can identify, on average, 30,000-40,000 variants in an individual exome that are different from the reference genomic sequence. It has been reported that, on average, each genome has around 165 homozygous protein truncating or stop loss variants in different genes, involved in several pathways (30) and around 300-400 variants are predicted to alter protein structure (31). Depending on the ethnic background of the sequenced proband, most of these variants (>95%) are known to be polymorphisms in the human population and can be found in databases such as dbSNP (32), the 1000 genomes (31), and in-house exome databases. Based on the assumption that variants with high frequency in the population are not likely to be pathogenic, these are filtered out before any further analysis. Furthermore, variants that are computationally predicted to be benign and non-pathogenic are removed. We have applied this strategy to detect the pathogenic mutation causing Chudley McCullough Syndrome (CMS) (**Chapter 6**). We sequenced affected individuals with the CMS phenotype from two unrelated families. After following the above-mentioned filtering steps and selecting for variants present in one gene, we were able to detect one homozygous frameshift mutation in *GPSM2* as a possible cause for CMS. However, this strategy can miss the pathogenic variants in certain cases. For instance, if the causative variant is located in a poorly covered exon in one or several sequenced individuals, the candidate gene will be falsely removed from the list. Also, in heterogeneous disorders the real

gene may be removed by this strategy, if only a minority of the patients show a mutation, because several different genes are involved (33).

It is known that the interpretation of missense variants is challenging because a change of an amino acid in a long peptide chain in itself is not necessarily meaningful. The change may be entirely harmless or it may obliterate the function of the protein. There are several approaches to obtain evidence for the pathogenicity of missense variants. If, for example, a variant, identified using GERP, PhyloP or PhastCons scores, affects an amino acid position that is evolutionary highly conserved, it is more likely to be pathogenic. We have shown in **chapter 4** that even an apparently neutral variant can alter splicing and in that way become pathogenic. The fact that the different computational algorithms currently in use to asses DNA and protein variants can lead to false positive, or false negative predictions is borne out by the fact that the *FLNA* mutation leading to TOD was overlooked by other authors (34) (**chapter 4**).

Although many studies have shown the successful application of ES for finding causative disease genes (26), it is difficult to know how often this method leads to negative results because results that fail to identify the pathogenic variant are rarely reported. ES is not a panacea for all genetic problems and moreover has limitations similar to other molecular technologies. From our experience we find that not every ES experiment results in the identification of a novel disease gene. We were able to solve nine out of 16 (56%) cases for which we tried to find the disease causing genes with ES. Several technical and/or analytical factors may play a role in the failure of gene discovery: 1) Our knowledge of all truly protein-coding exons in the human genome is still uncertain, so all current capture kits target only exons that have been identified until now but all parts of the genome that we do not recognize as functional are not included. 2) It is possible that some or all exons of the causative gene are not included in the target kit due to failure of the probe design. 3) There may be insufficient coverage of the region that contains the pathogenic mutation. This is because the efficiency of capture probes differs considerably and not all templates are sequenced as effectively. 4) It is possible that the causal variant is well covered but is inaccurately mapped because of miss-mapped reads or errors in the alignment. 5) The causal mutation is located in non-coding sequences (deep intronic) or in distal regulatory elements. 6) Our understanding of the genome and the exons is limited and we are unable to interrogate many variants that may be important for controlling gene transcription or splicing. 7) Current practice

shows clear limitation of exome sequencing for the detection of CNVs, which represent an important cause of Mendelian disorders. 8) It may be difficult to discriminate the causal alleles from the neutral alleles due to genetic heterogeneity of the disorder. If, for example, one gene accounts for only a small fraction of the sequenced cases (depending on the sample size), no single gene will be shared between all cases and at the same time many other genes may have shared neutral variants. 9) Possible non-genetic causes of the disorder can lead to failure of gene discovery.

In conclusion, although ES has several limitations, it is revolutionizing the discovery of Mendelian diseases. Identifying the genetic alteration underlying phenotypic variation is of particular biological and medical interest. The unbiased ES identifies variants in all known genes simultaneously and allows systematic analysis of all coding exons from individual samples and families. This approach is providing significant insights into the genetic causes of Mendelian diseases and the role of rare variants in healthy individuals as well as individuals with genetic diseases. It provides more accurate genotype-phenotype correlations and will improve clinical diagnosis, family counselling and potential future therapeutic intervention. Our studies and many others, show promising results for the development of new technologies for clinical applications. Continuous innovation and improvement of methods and techniques for sequencing, the rapid reduction of cost, the improvement of tools for bioinformatics data analysis, and the improved methods and algorithms for the interpretation of variants will make NGS the preferred approach for clinical diagnosis. However, for the time being, during this early phase, it is a difficult undertaking to confidently pinpoint the causal genetic change. Once large numbers of DNA variants have been collected, and well documented worldwide, and effective pipelines for data analysis are in place, this diagnostic approach will become routine and we can expect that many genetic abnormalities will be resolved. The adaptation of targeted capture and/or ES followed by NGS in clinical diagnostics has begun and it is very likely that ES, and if not whole genome sequencing, will have significant impact in the clinical setting for diagnosis of genetic diseases in the near future.

References

1. Sanger F, Nicklenl S, Coulson AR. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463-7.
2. von Bubnoff A. (2008) Next-generation sequencing: the race is on. Cell 132: 721-723.
3. Schuster SC. (2008) Next-generation sequencing transforms today's biology. Nat Methods 5: 16-18.
4. Shendure J, Ji H. (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135-45.
5. Hofstra RM, Mulder IM, Vossen R, de Koning-Gans PA, Kraak M, Ginjaar IB, van der Hout AH, Bakker E, Buys CH, van Ommen GJ, van Essen AJ, den Dunnen JT. (2004) DGGE-based whole-gene mutation scanning of the dystrophin gene in Duchenne and Becker muscular dystrophy patients. Hum Mutat 23: 57-66.
6. Bennett RR, den Dunnen J, O'Brien KF, Darras BT, Kunkel LM. (2001) Detection of mutations in the dystrophin gene via automated DHPLC screening and direct sequencing. BMC Genet 2: 17.
7. Tuffery S, Moine P, Demaille J, Claustres M. (1993) Base substitutions in the human dystrophin gene: detection by using the single-strand conformation polymorphism (SSCP) technique. Hum Mutat 2: 368-374.
8. Ashton EJ, Yau SC, Deans ZC, Abbs SJ. (2008) Simultaneous mutation scanning for gross deletions, duplications and point mutations in the DMD gene. Eur J Hum Genet 16:53-61.
9. Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ. (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. *Clin. Chem.* 49: 853–60.
10. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60–65.
11. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods.
12. Turner EH, Ng SB, Nickerson DA, Shendure J. (2009) Methods for genomic partitioning. Annu Rev Genomics Hum Genet 10: 263–284.
13. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4: 903-5.
14. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4: 907-9
15. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39: 1522-7.
16. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27: 182-9.
17. Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet . 7:277-282.
18. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. (2011) What can exome sequencing do for you?. J Med Genet. 48:580-9.
19. Lander ES, Botstein D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567-1570.
20. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073-1080.
21. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. (2012) Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet. 57:621-32.

22. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 42:30-35.
23. Zhang W, Amir R, Stockton DW, Van Den Veyver IB, Bacino CA, Zoghbi HY. (2000) Terminal osseous dysplasia with pigmentary defects maps to human chromosome Xq27.3-xqter. Am J Hum Genet. 66:1461–1464.
24. Breedveld GJ, van Wetten B, te Raa GD, Brusse E, van Swieten JC, Oostra BA, Maat-Kievit JA. (2004) A new locus for a childhood onset, slowly progressive autosomal recessive spinocerebellar ataxia maps to chromosome 11p15. J Med Genet. 41:858-66.
25. Gilissen C, Hoischen A, Brunner HG, Veltman JA. (2011) Unlocking Mendelian disease using exome sequencing. Genome Biol. 12:228.
26. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. (2012) Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J Hum Genet. 57:621-32.
27. Khan DH, Jahan S, Davie JR. (2012) Pre-mRNA splicing: role of epigenetics and implications in disease. Adv Biol Regul. 52:377-88.
28. Lu Q, Qiu X, Hu N, Wen H, Su Y, Richardson BC. (2006) Epigenetics, disease, and therapeutic interventions. Ageing Res Rev. 54:449-67.
29. Gilissen C, Hoischen A, Brunner HG, Veltman JA. (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet. 20:490-7.
30. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, Cirulli ET, Fellay J, Dickson SP, Gumbs CE, Heinzen EL, Need AC, Ruzzo EK, Singh A, Campbell CR, Hong LK, Lornsen KA, McKenzie AM, Sobreira NL, Hoover-Fong JE, Milner JD, Ottman R, Haynes BF, Goedert JJ, Goldstein DB. (2010) The characterization of twenty sequenced human genomes. PLoS Genet. 6:9.
31. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR et al. (2010) A map of human genome variation from population-scale sequencing. Nature. 467:1061–73.
32. Sayers EW, Barrett T, Benson DA et al. (2011) Database resources of the national center for biotechnology information. Nucleic Acids Res. 39 (Database issue): D38–D51.
33. Robinson PN, Krawitz P, Mundlos S. (2011) Strategies for exome and genome sequence data analysis in disease-gene discovery projects. Clin Genet. 80:127-32.
34. Brunetti-Pierri N, Lachman R, Lee K, Leal SM, Piccolo P, Van Den Veyver IB, Bacino CA. (2010) Terminal osseous dysplasia with pigmentary defects (TODPD): Follow-up of the first reported family, characterization of the radiological phenotype, and refinement of the linkage region. Am J Med Genet A. 7:1825-31.

Summary

The work presented in this thesis describes the development and application of new techniques for detecting small variations (mutations) in genomic DNA that underlie various disorders. These techniques include High Resolution Melting Curve Analysis (HR-MCA) followed by Sanger sequencing, targeted, X-exome and whole exome capture followed by Next Generation Sequencing (NGS).

Chapter 2 describes the use and implementation of HR-MCA followed by Sanger sequencing as a pre-sequencing routine diagnostic scanning method for all exons and exon-intron junctions of the DMD gene. After validating the technique, we screened a group of 22 unrelated DMD/BMD patients and 11 females in which deletions and duplications of the gene had been excluded. Seventeen different pathogenic mutations were found in the screened group, of which ten were novel. Our results show that HR-MCA is a powerful and inexpensive diagnostic pre-sequencing scanning method to detect small mutations in BMD/DMD patients and carriers.

In chapter 3 we describe the application of array-based sequence capture (385K NimbleGen arrays) to enrich the exons and immediate intron flanking sequences of 112 genes, which are potentially involved in mental retardation and congenital malformation. Captured material was sequenced using Illumina technology and a data analysis pipeline was built. Our data show that: 1) An array-based sequence capture followed by Illumina sequencing, offers a versatile tool for successfully selecting sequences of interest from a total human genome. 2) All known variants were reliably detected. 3) Although overall coverage varied considerably, it was reproducible per region and facilitated the detection of large deletions and duplications (CNVs), including a partial deletion in the B3GALTL gene from a patient sample. 4) There is room for improvement of the methodology for ultimate diagnostic application, in particular with respect to array design that can obtain a more even coverage of the targeted regions.

In chapter 4, we performed X-exome capture followed by Illumina (Genome Analyzer II) sequencing in two probands from Dutch and Italian families with Terminal Osseous Dysplasia (TOD). TOD is an X-linked dominant male-lethal disease, characterized by terminal skeletal dysplasia, pigmentary defects of the skin, and recurrent digital fibroma during infancy. Previous linkage studies have mapped the disease-causing gene to Xq27.3-q28. After analyzing the data,

we identified a silent variant at the last nucleotide of exon 31 of the FLNA gene in both patients. The same variant c.5217G>A was also found in another four unrelated cases but not in 400 control X chromosomes, the 1000 Genomes, or in the database for FLNA gene variants. In families, this variant co-segregated with the disease. Our data show that due to nonrandom X chromosome inactivation, the mutant allele was not expressed in patient fibroblasts. RNA expression of the mutant allele was detected only in cultured fibroma cells obtained from material that had been surgically removed 15 years ago. 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 single unique recurrent mutation in the FLNA gene.

In chapter 5, we have used whole exome sequencing to identify pathogenic mutations causing autosomal recessive Spinocerebellar ataxia type 7 (SCAR7). The locus of SCAR7 has been linked to chromosome band 11p15. We have now identified the causative gene for SCAR7 by exome sequencing in the index family. One missense and one splice site mutation were found in the *TPP1* gene which co-segregated with the disease. The same mutations were found in an unrelated patient with a similar phenotype. Affected individuals showed low activity of tripeptidyl peptidase1, the protein coded by *TPP1*, the gene known to cause the infantile form of Neuronal Ceroid Lipofuscinosis (CLN2). However, the patients that we studied had none of the findings that are characteristic for CLN2: epilepsy, ophthalmic abnormalities, curvilinear bodies in the skin biopsy tissue. Also, the slow progressive evolution of the disease until old age of the patients is clearly different from the relentless progression in infancy known for CLN2.

In chapter 6, we studied the genetic cause of Chudley-McCullough Syndrome (CMS). We sequenced the exomes of three patients with CMS from two unrelated Dutch families from the same village and identified the same homozygous frameshift *GPSM2* variant c.1473delG in all patients. This variant was confirmed by Sanger sequencing in all affected patients and in a heterozygous form in their parents. We have shown that this variant arises from a shared, rare haplotype. Our data confirm the recent finding of Doherty *et al*., who reported *GPSM2* variants as a cause of CMS. The c.1473delG mutation in *GPSM2* associated with CMS appears to be an ancient founder mutation brought to North America by early Mennonite settlers originating from

Western Europe. Furthermore, we have established an LOVD database for *GPSM2* containing all variants thus far described.

In chapter 7, we describe the successful application of whole exome sequencing for finding the genetic cause of Faciocapulohumeral dystrophy type 2 (FSHD2). FSHD is the third most common myopathy, which is characterized by progressive and irreversible weakness of the facial, shoulder and upper arm muscles. In the majority of cases, the FSHD type 1 (FSHD1) is caused by contraction of the D4Z4 repeat array on a specific permissive allele on chromosome 4. This leads to local chromatin relaxation and stable expression of the D4Z4-encoded *DUX4* retrogene in skeletal muscle. In FSHD2, the myopathy results from chromatin relaxation and stable *DUX4* expression but without D4Z4 array contraction. To determine the genetic cause of FSHD2 and to identify the locus controlling the D4Z4 hypomethylation, we performed whole exome sequencing of twelve individuals from seven unrelated FSHD2 families: Five with dominant segregation of the hypomethylation and two with sporadic hypomethylation. Different mutations in *SMCHD1* (Structural Maintenance of Chromosomes flexible Hinge domain containing ***1)*** were identified in all affected individuals except in one family. We used Sanger sequencing to confirm the presence of these mutations and included 12 additional unrelated families with FSHD2 from whom DNA or RNA was available. We identified heterozygous out-of-frame deletions, heterozygous missense, and splice-site mutations in *SMCHD1* in 15/19 (79%) families. Mutations in *SMCHD1* substantially reduce the *SMCHD1* protein levels in skeletal muscle, which leads to contraction-independent *DUX4* expression. Furthermore, we found that mutations in *SMCHD1*, which is on chromosome 18, segregates independently of the FSHD-permissive *DUX4* allele on chromosome 4. This results in a digenic inheritance pattern in affected individuals. FSHD2 occurs exclusively in individuals who inherit both the *SMCHD1* mutation and a normal-sized D4Z4 array on a chromosome 4 haplotype permissive for *DUX4* expression. This showed that *SMCHD1* is an epigenetic modifier of the D4Z4 metastable epiallele and is a key genetic determinant of FSHD2 disorder.

Finally in chapter 8 we have discussed the pros and cons of all the techniques that we have presented in this thesis. These methods have made a significant contribution to accurate molecular diagnosis and to quick identification of disease causing genes.

Nederlandse samenvatting

Dit proefschrift beschrijft de ontwikkeling en toepassing van nieuwe technieken voor het detecteren van kleine variaties (mutaties) in genomisch DNA. Deze technieken zijn:

- Smeltcurve analyse met hoge resolutie (High Resolution Melting Curve Analyse (HR-MCA) als pre-sequencing techniek gevolgd Sanger sequencing
- Gerichte verrijking van exonen op het X chromosoom en alle exonen in het genoom (whole exome) gevolgd door "Next Generation Sequencing" (NGS).

Hoofdstuk 2 beschrijft het toepassen en de implementatie van HR-MCA en Sanger sequencing als een efficiënte diagnostische voorscreeningsmethode voor het opsporen van kleine mutaties in alle exonen en exon-intron overgangen van het DMD (Duchenne Muscular Dystrophy) gen. Na de validatie van de techniek, werd het DNA onderzocht van 22 onafhankelijke DMD/BMD patiënten en 11 vrouwen bij wie eerder deleties en duplicaties in het DMD gen waren uitgesloten. Zeventien verschillende pathogene mutaties werden in deze groep gevonden, waarvan er tien nieuw waren ontstaan. De resultaten tonen aan dat HR-MCA een krachtige en goedkope diagnostische pre-sequencing-techniek is om kleine mutaties te detecteren in het DMD gen van patiënten en dragers.

In hoofdstuk 3 beschrijven we het verrijken van stukjes DNA sequenties met behulp van micro-arrays (385K NimbleGen arrays) om de exonen en de flankerende intron gebieden te kunnen sequencen van 112 genen, die mogelijk betrokken zijn bij mentale retardatie en aangeboren afwijkingen. De sequentie van het geselecteerde materiaal werd bepaald met behulp van Illumina-technologie en een data-analyse pijplijn werd gebouwd. Onze gegevens tonen aan dat:

- ten eerste, het verrijken van DNA sequenties met micro-arrays gevolgd door Illumina sequencing een effectief hulpmiddel is om belangrijke sequenties uit het menselijk genoom te selecteren;
- ten tweede alle bekende varianten met succes werden gedetecteerd;
- ten derde, alhoewel de totale 'coverage' aanzienlijk varieerde, de resultaten per regio reproduceerbaar waren en het de detectie van grote deleties en duplicaties (Copy Number

Variants) vergemakkelijkte, waaronder het opsporen van een gedeeltelijke deletie in het B3GALTL gen in een patiënten monster.

Voor de uiteindelijke diagnostische toepassing, kan de methode nog worden verbeterd, met name met betrekking tot het array ontwerp teneinde een meer gelijkmatige coverage van de te onderzoeken gebieden te verkrijgen.

In hoofdstuk 4, wordt X-exome capture beschreven, gevolgd door Illumina (Genome Analyzer II) sequencing bij twee index patiënten uit een Nederlands en een Italiaans gezin met Terminal Osseous Dysplasia (TOD). TOD is een X-gebonden dominante mannelijk-lethale ziekte, die wordt gekenmerkt door terminale skeletdysplasie, pigment afwijkingen, hypoplasie van de huid en regelmatig terugkerende digitale fibromen tijdens de kinderjaren. Eerdere koppelingsstudies met markers op het X chromosoom plaatsten het ziekte-veroorzakende gen in band Xq27.3-q28. Met behulp van data analyse werd een "stille" variant in het laatste nucleotide in exon 31 van het gen FLNA geïdentificeerd in beide patiënten. Dezelfde variant c.5217G>A werd ook gevonden in vier andere niet verwante patiënten en werd niet gevonden in 400 controle X-chromosomen, de 1000 genomen en de FLNA gen variant database. De variant co-segregeerde met ziekte in deze families. Onze gegevens laten zien dat door non-random X chromosoom inactivatie, het mutante allel niet tot expressie komt in de fibroblasten van een patiënt. RNA expressie van het mutante allel werd alleen gedetecteerd in gekweekte cellen van een chirurgisch verwijderd fibroma van een 3-jarig patiëntje, dat 15 jaar was bewaard. De variant activeert een cryptische splicings site waardoor de laatste 48 nucleotiden van exon 31 worden verwijderd. Op eiwitniveau, resulteert dit in een verlies van 16 aminozuren (p.Val1724_Thr1739del) en naar verwachting de deletie van een sequentie op het oppervlak van filamine repeat 15. Onze gegevens tonen aan dat TOD wordt veroorzaakt door een enkele, unieke, herhaald nieuw ontstane mutatie in het FLNA gen.

Hoofdstuk 5 bespreekt whole exome sequencing om pathogene mutaties te identificeren die het autosomaal recessieve Spino Cerebellaire Ataxie type 7 (SCAR7) veroorzaken. Het SCAR7 locus werd eerder gemapt op chromosoom 11p15. We hebben nu het oorzakelijke gen voor SCAR7 geïdentificeerd middels exome sequencing in de familie van de de index patiënt. Twee verschillende mutaties, een missense en een splice-site mutatie werden gevonden in het TPP1

gen dat co-segregeert met de ziekte. Dezelfde mutaties werden gedetecteerd in een niet verwante andere patiënt met een vergelijkbaar fenotype. De patiënten toonden een lage activiteit van tripeptidyl peptidase-1, het eiwit dat wordt gecodeerd door *TPP1*. Het is al langer bekend dat volledige inactivering van dit gen de infantiele vorm van Neuronale Ceroid Lipofuscinosis (CLN2) veroorzaakt. De patiënten hadden geen epilepsie, geen oog afwijkingen, noch hadden zij curviliniaire lichaampjes bij elektronenmicroscopisch onderzoek van een huidbiopt, die karakteristiek zijn voor CLN2. Het langzaam progressieve beloop van de ziekte tot op oudere leeftijd is duidelijk anders dan de snelle en gestadige achteruitgang in de kinderjaren die kenmerkend is voor CLN2.

Hoofdstuk 6 beschrijft het sequencen van de exomen van drie patiënten met Chudley McCullough Syndroom (CMS) uit twee niet verwante Nederlandse gezinnen uit hetzelfde dorp. We identificeerden bij alle drie de patiënten dezelfde homozygote frameshift variant c.1473delG in het GPSM2 gen. Deze variant werd middels Sanger sequencing homozygoot in de patiënten en in heterozygote vorm in hun ouders aangetoond. We laten zien dat deze variant is ontstaan in een gemeenschappelijk, zeldzaam haplotype. Onze gegevens bevestigen de recente bevinding van Doherty en co-autheurs, die varianten in *GPSM2* als oorzaak voor CMS beschrijven. De c.1473delG mutatie in *GPSM2,* die is geassocieerd met CMS, lijkt afkomstig van een gemeenschappelijke voorouder en is waarschijnlijk door mennonieten (doopsgezinde kolonisten) vanuit West-Europa naar Noord Amerika gebracht. Wij hebben tevens een Leiden Open source Variant Database (LOVD) voor het GPSM2 gen opgezet met alle tot dusver bekende varianten.

In hoofdstuk 7 beschrijven we de succesvolle toepassing van whole exome sequencing bij het zoeken naar de genetische oorzaak van Facioscapulohumerale dystrofie type 2. FSHD is na spinale spieratrofie en de spierziekte van Duchenne de derde meest voorkomende myopathie, die wordt gekenmerkt door geleidelijk toenemende zwakte van de spieren in het gelaat, de schouder en de bovenarm spieren. In de meeste gevallen wordt FSHD veroorzaakt door samentrekking van de D4Z4 repeat op een allel gelegen op chromosoom 4 (FSHD1). Dit leidt tot een lokale verandering van de chromatinestructuur en stabiele expressie van het door D4Z4 gecodeerde DUX4 retrogen in skeletspieren. In andere gevallen komt de myopathie eveneens tot stand door chromatine verandering en stabiele DUX4 expressie, maar zonder de contractie van de D4Z4 repeat (FSHD2). Om de genetische oorzaak van FSHD2 en het locus dat verantwoordelijk is

voor de D4Z4 hypomethylatie te identificeren, werd whole exome sequencing uitgevoerd bij twaalf personen uit zeven onafhankelijke FSHD2 families: vijf met dominant overervende hypomethylatie en twee met sporadische hypomethylatie. Verschillende mutaties in het SMCHD1 gen (structural maintenance of chromosomes flexible hinge domain containing ***1***) werden met uitzondering van één patiënt in alle aangedane individuen geïdentificeerd. We hebben Sanger sequencing gebruikt om de aanwezigheid van deze mutaties te bevestigen. En we hebben 12 aanvullende niet-verwante families met FSHD2 waarvan DNA of RNA beschikbaar was. We identificeerden in 15 van de 19 (79%) families heterozygote frameshift deleties, heterozygote missense en splice-site mutaties in *SMCHD1*.Mutaties in *SMCHD1* verminderen in grote mate het SMCHD1 eiwitgehalte in de skeletspieren met als gevolg algehele hypomethylatie, hetgeen leidt tot DUX4 expressie onafhankelijk van de contractie van de D4Z4 repeat. Daarnaast werd duidelijk dat mutaties in *SMCHD1*, dat is gelegen op chromosoom 18, onafhankelijk van het FSHD-permissieve DUX4 allel op chromosoom 4 wordt doorgegeven in een familie. Dit resulteert in een digeen overervingspatroon bij de aangedane personen. FSHD2 ontstaat uitsluitend bij personen die zowel de SMCHD1 mutatie als een normale grootte van de D4Z4 repeat geërfd hebben op een chromosoom 4 haplotype permissief voor DUX4 expressie. Dus *SMCHD1* is een epigenetische modifier van het D4Z4 metastabiele epi-allel en een belangrijke genetische determinant van FSHD2.

In hoofdstuk 8 worden de voors en tegens van de verschillende hierboven beschreven technieken besproken en komen we tot de conclusie dat de nieuwste technieken ons steeds sneller en beter in staat stellen de oorzakelijke genen voor ziekten op te sporen en moleculaire diagnoses te stellen.

List of publications

Almomani R, van der Stoep N, Bakker E, den Dunnen JT, Breuning MH, Ginjaar IB.Rapid and cost effective detection of small mutations in the DMD gene by high resolution melting curve analysis. ***Neuromuscul Disord***. 2009; 19:383-90.

Almomani R, van der Heijden J, Ariyurek Y, Lai Y, Bakker E, van Galen M, Breuning MH, den Dunnen JT.Experiences with array-based sequence capture; toward clinical applications. ***Eur J Hum Genet***. 2011; 19: 50–55.

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; 87:146-53.***The authors contributed equally to this work**

Yu Sun*, **Rowida Almomani*,** Guido Breedveld, Gijs W.E. Santen, Emmelien Aten, Dirk J. Lefeber, Jorrit I. Hoff, Esther Brusse, Frans W. Verheijen, Rob M. Verdijk, Marjolein Kriek, Ben Oostra, Martijn H. Breuning, Monique Losekoot, Johan T. den Dunnen, Bart P. van de Warrenburg, and Anneke J.A. Maat-Kievit. Autosomal recessive spinocerebellar ataxia 7 (SCAR7) is caused by variants in *TPP1*, the gene involved in classic late-infantile neuronal ceroid lipofuscinosis 2 disease (CLN2 disease). ***Hum mutat.*** 2013;34:706-13. ***The authors contributed equally to the work**

Rowida Almomani*, Yu Sun*, Emmelien Aten, Yvonne Hilhorst-Hofstee, Cacha M.P.C.D. Peeters-Scholte, Arie van Haeringen, Yvonne M.C. Hendriks, Johan T. den Dunnen, Martijn H. Breuning, Marjolein Kriek, and Gijs W.E. Santen. GPSM2 and Chudley–McCullough Syndrome: A Dutch Founder Variant Brought to North America. ***Am J Med Genet Part A.*** 2013. ***The authors contributed equally to this work**

Lemmers RJ, Tawil R, Petek LM, Balog J, Block GJ, Santen GW, Amell AM, van der Vliet PJ, **Almomani R**, Straasheijm KR, Krom YD, Klooster R, Sun Y, den Dunnen JT, Helmer Q, Donlin-Smith CM, Padberg GW, van Engelen BG, de Greef JC, Aartsma-Rus AM, Frants RR,

de Visser M, Desnuelle C, Sacconi S, Filippova GN, Bakker B, Bamshad MJ, Tapscott SJ, Miller DG, van der Maarel SM.Digenic inheritance of an *SMCHD1* mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. ***Nat Genet***. 2012 44:1370-4.

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. ***Nat Genet.*** 2012; 44:379-80.

Aten E*, Sun Y*, **Almomani R**, Santen GW, Messemaker T, Maas SM, Breuning MH, den Dunnen JT.Exome Sequencing Identifies A Branch Point Variant in Aarskog-Scott Syndrome. ***Hum Mutat***. 2013;34:430-4. ***The authors contributed equally to this work**

B. Sikkema-Raddatz, L. F.Johansson, E. N. de Boer, **R. Almomani,** L.G. Boven, M.P. v.d. Berg, K.Y. van Spaendonck-Zwarts, JP van Tintelen, R. Sijmons, J.D.H. Jongbloed , R.J. Sinke. Targeted next generation sequencing ready for clinical diagnostics: Sanger sequencing can be replaced**.** ***Hum Mutat.*** 2013

Karin Y. van Spaendonck-Zwarts, Anna Posafalvi, Denise Hilfiker-Kleiner, Karen Sliwa, Mariel Alders, **Rowida Almomani**, Dirk J. Van Veldhuisen, Irene M. van Langen, Richard J. Sinke, Jolanda van der Velden, Maarten P. van den Berg, J. Peter van Tintelen, Jan D.H. Jongbloed. Targeted next generation sequencing in families with peripartum cardiomyopathy and dilated cardiomyopathy. submitted

Acknowledgements

I would like to thank all the people who have contributed to finish this thesis. It would have never have reached this point without their support, encouragement and help.

First, I would like to thank my promoters Prof. Martijn Breuning and Prof. Bert Bakker who gave me the opportunity to work as a PhD student in the Clinical and Human Genetic Department of the LUMC. Thank you for your kindness, support, patient guidance as well as your academic experience, which have been greatly valuable to me.

I would like to express my thanks to Ieke, for her personal support and great patience at all times, you are just a close family member who cares about me and who is always there.

Many thanks to Kamlesh for critical reading of my thesis and precise sense of the English language contributed to the final draft.

Johan, thanks for your kind assistance and willingness to discuss ideas about different projects.

Marjolein, Emmelien, Yu and Gijs, it was nice to work with all of you, so enjoyable and productive.

Many thanks to my colleagues at LDGA, working at this department has definitely became easier due to the nice and relaxing social environment, all of you were willing to help and answering my questions.

I would like to thank Astrid and Janina for shearing the same office with me. You are one of my best friends. You were nice company.

I want to thank the colleagues in LGTC. Yavuz, you are so smart, you are always willing to help, explain the experiments technical details in so easy way. I still need to learn from you, and still remember your words (no worries, who cares, details, tomorrow is another day). Thanks a lot for your help!

I would like to thank Rolf for introducing the melting curve analysis to me at the beginning of my PhD period. It was nice to work with you all the time at the LGTC.

I would like to thank also my colleagues at my work at the Department of Genetics, University of Groningen, University Medical Centre Groningen (UMCG), the Netherlands. It is indeed a nice place to work in, thanks Richard and Jan for your nice support.

I would like to express my thanks and love to my parents, brothers and my sister; they gave me the courage and support in my life.

In the end, I lack words to express my thanks, feelings, and love to my husband Ali and my two little angles Tala and Yara: you are the source of happiness, peace and love in my life, you were so patient during my PhD period, without your support I would have never finished my work. I was so busy during my PhD work, I know you missed me so much I was seeing that in your eyes, every day I came home and every weekend I was working in the lab. I love you my little family.

Curriculum Vitae

Rowida Almomani was born on 4th of February, 1979 in Al-Ruseifa, Jordan. She passed the general secondary education examination (Tawjihi) in 1997. She then studied Biology at Mutah University in Al-Karak, Jordan, and got her Bachelor degree with the highest average and thus being the top student at the Bachelor program for that year, 2001. She got the Honoree certificate for advanced academic achievements in the Bachelor period. In 2005, Rowida got her Master degree in applied biology from the Jordan University of Science and Technology (JUST) in Irbid, Jordan. Two years later she joined the group of Prof. Martijn H Breuning, at the Clinical and human genetics department of Leiden University Medical Center (LUMC) in Leiden, the Netherlands, to do her PhD. During her PhD work, she has been introduced to different molecular diagnostic technologies including the rapidly developing field of next generation sequencing (NGS). Her research focuses on the application of NGS and the use of new technologies, especially High Resolution Melting Curve Analysis (HR-MCA), targeted and exome sequencing, to be able to identify the causal pathogenic variants in different genetic diseases and to improve genetic testing. In December 2011 until now, she works as a post doc at the department of Genetics of the University Medical Center Groningen (UMCG), the Netherlands. Currently her research subject focuses on finding pathogenic mutations in genes related to Cardiomyopathies and Heart diseases by exome sequencing. The research projects she worked with have provided her with a broad view that makes her able to work independently as well as in a team-work and makes her highly motivated to work in different areas of genetic research.