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.