Chapter 1

Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics

Human Mutation 2013;34(7):1035-42.
DOI: 10.1002/humu.22332
PubMed ID: 23568810

L.F. Johansson[1,2,*], F. van Dijk[1,2,*], E.N. de Boer[1], K.K. van Dijk-Bos[1], L.G. Boven[1], M.P. van den Berg[2], K.Y. van Spaendonck-Zwarts[1], J. Peter van Tintelen[1], R.H. Sijmons[1], J.D. Jongbloed[1], R.J. Sinke[1]

1. University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands
2. University of Groningen, University Medical Center Groningen, Department of Cardiology, Groningen, The Netherlands

Received 2013 Jan 9; Accepted revised manuscript 2013 Apr 2; Published online 2013 Apr 4.

* Contributed equally

Abstract

Mutation detection through exome sequencing allows simultaneous analysis of all coding sequences of genes. However, it cannot yet replace Sanger sequencing (SS) in diagnostics because of incomplete representation and coverage of exons leading tomissing clinically relevant mutations. Targeted next-generation sequencing (NGS), in which a selected fraction of genes is sequenced, may circumvent these shortcomings. We aimed to determine whether the sensitivity and specificity of targeted NGS is equal to those of SS. We constructed a targeted enrichment kit that includes 48 genes associated with hereditary cardiomyopathies. In total, 84 individuals with cardiomyopathies were sequenced using 151 bp paired-end reads on an Illumina MiSeq sequencer. The reproducibility was tested by repeating the entire procedure for five patients. The coverage of $\geq$30 reads per nucleotide, our major quality criterion, was 99% and in total ~21,000 variants were identified. Confirmation with SS was performed for 168 variants (155 substitutions, 13 indels). All were confirmed, including a deletion of 18 bp and an insertion of 6 bp. The reproducibility was nearly 100%. We demonstrate that targeted NGS of a disease-specific subset of genes is equal to the quality of SS and it can therefore be reliably implemented as a stand-alone diagnostic test.

1.1 Introduction

Next-generation sequencing (NGS) techniques have significantly increased the possibilities of genome analysis. If we focus on diagnostic applications,

mutation analysis through exome sequencing (ES) allows for the simultaneous analysis of all coding sequences of genes. One of the first clinical applications of ES was the detection of disease-associated mutations in rare Mendelian diseases, such as Miller syndrome [10], Sensenbrenner syndrome [4], and Schinzel–Giedion syndrome [8]. The advantage of ES is that it does not require a priori knowledge of gene(s) responsible for a disorder using it as a genetic discovery panel. In diagnostics, ES is already used to screen for de novo pathogenic mutations in intellectual disability [2] explained by more than 1,000 different genes. In addition, ES can be used more targeted by analyzing only a panel of genes that may be involved in a particular disease. However, in routine diagnostics, detecting mutations via conventional Sanger sequencing (SS) is still the standard, despite the practical difficulties of keeping up with the ever-increasing numbers of test requests and of disease-associated genes. For instance, hereditary cardiomyopathies can be explained by 40-60 different genes [**?**, 11] and effective analysis of all these genes by SS in a diagnostic setting is not feasible. In practice, it is limited to no more than 10 genes. In contrast, ES would allow the simultaneous analysis of all coding genes through enrichment for these coding regions before sequencing. However, in its current state, ES cannot be used as a reliable substitute for SS in diagnostics. A major shortcoming is incomplete representation and coverage of exons, leading to clinically relevant mutations being missed [4, 13]. Here, amore dedicated targeted enrichment appears to be the method of choice, not only because it allows focusing on the genes relevant for a particular disorder, but also because its highly effective enrichment provides a superior quality of representation and coverage. In addition, focusing on only the genes relevant for a particular disorder minimizes the problems associated with unsolicited findings. Targeted NGS is faster and cheaper than ES, especially for the analysis of certain distinct disease phenotypes. Various enrichment methods have been developed in the last few years, such as solid phase-based microarrays, micro-droplet-based PCR (Rain Dance Technologies, Lexington, MA), amplicon-based or solution phase-based methods such as Sure Select Targeted enrichment and Illumina TruSeq Customenrichment. Different types of platformshave also been developed for high-throughput sequencing. Recently, even bench-top instruments have become available, such as Ion Torrent PGM (Life Technologies Ltd, Paisley, UK), 454 GS Roche Junior (Roche Applied Science, Indianapolis, IN), and the Illumina MiSeq (Illumina, SanDiego,CA) [9]. These are the size of a modern laser printer and offer modest set-up and running costs; they are particularly suited to small projects and allow a fast throughput. The aim of our study was to validate targeted NGS for application in clinical diagnostics and to assess its sensitivity and specificity relative to SS. We therefore developed a SureSelect targeted enrichment kit (Agilent

Technologies, Inc., Santa Clara, CA) for diagnostic testing of patients with hereditary cardiomyopathies. Hereditary cardiomyopathies are highly heterogeneous disorders, and include dilated (DCM), hypertrophic (HCM), and arrhythmogenic right ventricular cardiomyopathies (ARVC), which are leading causes of heart failure and sudden death. Approximately 30%–50% of DCM cases are familial, but with significant genetic and phenotypic heterogeneity [12]. Particularly for DCM, for which more than 50 cardiomyopathy-related genes have been identified, targeted resequencing would be a much better diagnostic platform than SS. The use of a MiSeq bench-top machine would also enable short turn-around times in the laboratory. We compared the outcome of our targeted NGS experiments with results from SS, and discuss our findings in the light of validation, clinical laboratory implementation, and quality assessment in general.

1.2 Material and Methods

1.2.1 Design of the Study

Our study was divided into two parts: a validation phase and an application phase. (1) Validation phase in which:

- sequencing quality of the targeted NGS kit was measured in terms of representation and coverage;
- sequencing reliability was measured in terms of sensitivity compared with SS results for at least six out of 14 different cardiomyopathy-related genes.

(2) Application phase in which:

- novel variants identified by our targeted NGS approach were confirmed by SS to assess the specificity;
- tests for reproducibility were performed. We set the following thresholds for accepting targeted NGS to replace SS in a diagnostic setting:
- Sequencing quality: coverage of at least$\times$30 for each nucleotide, based on a normal binomial contribution, a minimum number of four reads for a call, a 20% allele frequency resulting in a sensitivity of 99.96% for a heterozygote.
- Sequence reliability in validation and application phase: 100% sensitivity for at least 75 variants, including substitutions and indels. The specificity should be at least 98%, that is, amaximum of 2% false-positive variants.

- Reproducibility: 98%, so that a maximum of 2% difference in the variants within one sample was allowed when repeating the entire procedure.

1.2.2 Patients/Samples

For the validation phase, we selected DNA samples of 24 patients diagnosed with dilated or arrhythmogenic cardiomyopathies. These patients had previously been analyzed by SS for up to six out of 14 disease genes (*DES*, *DSC2*, *DSG2*, *LMNA*, *MYBPC3*, *MYH7*, *PKP2*, *PLN*, *RBM20*, *SCN5A*, *TMEM43*, *TNNC1*, *TNNT2*, and *TNNI3*). Here, SS resulted in the identification of a disease-associated mutation in seven out of the 24 patients and a total of 90 variants. Subsequently, for the application phase, we selected a further 60 DNA samples of unrelated cardiomyopathy patients, for whom no causative mutation had been found by routine diagnostic testing by SS. All samples (n = 84) were subjected to targeted NGS (described below). In addition, the entire procedure was repeated for five out of the total of 84 patient samples to test the reproducibility of our method.

1.2.3 Targeted Enrichment Kit Design

The biotinylated cRNA probe solution was manufactured by Agilent Technologies and provided as capture probes. We selected 48 genes known to be involved in isolated forms of cardiomyopathy or in disorders of which cardiomyopathy is a major part of the disease spectrum (mostly neuromuscular disorders) but in which mutations in isolated cardiomyopathy forms have been reported as well. The sequences corresponding to these 48 cardiomyopathy genes (Table 1.1) were uploaded to the Web-based probe design tool eArray (Agilent Technologies, Inc.); in total 1,134 targets with a size of 323,651 bp. The coordinates of the sequence data are based on NCBI build 37 (UCSC hg19). For the probe design, we set the following parameters: 120 bp bait length, per target spaced every 60 bp, centered, two times tiling, and targets to include sequences 40 bp before and after each exon.

Table 1.1: List of genes included in the targeted SureSelect Enrichment Kit

Gene	Chromosome	Basepair position (start-end)[1]	Total number of basepairs covered by baits	Number of exons covered
LMNA	1	156084670-156108971	3,010	12
TNNT2	1	201328298-201346845	2,339	17
PSEN2	1	227058923-227083365	2,701	12
ACTN2	1	236849934-236925959	4,365	21
RYR2	1	237205782-237996012	23,329	105
TTN	2	179391699-179672188	125,455	316
DES	2	220283145-220290507	2,011	8
TMEM43	3	14166654-14183335	2.163	12
SCN5A	3	38595730-38674890	7,117	27
MYL3	3	46899317-46904920[2]	X	7
TNNC1	3	52485251-52488071	966	6
MYOZ2	4	120056899-120107411	1,504	6
SGCD	5	155753727-156186441	2,467	9
DSP	6	7542109-7586986	11,371	24
LAMA4	6	112430565-112575868	9,125	39
PLN	6	118879948-118880328[2]	381	1
TBX20	7	35242002-35293271	1,988	8
PRKAG2	7	151254178-151573745	3,059	16
MYPN	10	69881155-69970283	5,515	19
MYOZ1	10	75391372-75401555	2,021	6
VCL	10	75757926-75878001	5,199	22
LDB3	10	88428388-88492804	4,519	16
ANKDR1	10	92672493-92681072	2,018	9
RMB20	10	112404173-112595790	4,951	15
BAG3	10	121411148-121437369	2,583	4
CSRP3	11	19204110-19223629	1,249	6
MYBPC3	11	47352917-47374293	6,858	33
CRYAB	11	111779310-111782513	931	3
ABCC9	12	21953938-22089668	7,928	39
PKP2	12	32945260-33049705	3,824	14
MYL2	12	111348584-111358444	1,291	7
MYH6	14	23851159-23877526	9,061	39
MYH7	14	23881907-23904910	9,361	41
PSEN1	14	73614463-73686082	2,464	11
ATC1	15	3508225-35087049	1,931	6
TPM1	15	63334989-63363411	2,576	14
TCAP	17	37821573-37822407	669	2
JUP	17	39911956-39928146	3,278	13
DSC2	18	28647949-28682428	2,706	17
DSG2	18	29078175-29126804	8,751	15
CALR3	19	16589835-16606980	1,942	9
TNNI3	19	55663096-55668997	1,340	8
JPH2	20	42743396-42789087	2,032	4
DMD	X	31139907-33357766	19,354	85
GLA	X	100652739-100663041	1,978	7
LAMP2	X	119565097-119603064	2,215	10
EMD	X	153607805-153609597	1,245	6
TAZ	X	153640141-153649402	1,782	11

[1] Basepair position according to NCBI build 37 [2] The original article mistakenly states the start position twice

1.2.4 Sample Preparation

Sample preparation was performed according to the manufacturer's instructions (SureSelect XT Custom 1kb-499kb library, Cat. No. 5190–4806, SureSelect Library prep kit; Agilent Technologies, Inc.). In brief, the quality of each sample was checked on a Nanodrop machine (Thermo Scientific,Waltham, MA) and, before fragmentation by electrophoresis, on a 0.7% agarose gel. Next, 3 µg of each genomic DNA sample was fragmented by Adaptive Focused Acoustics (Covaris S220 one channel, runtime 80 sec, peak power 140.0W, duty factor 10.0%, cycles/burst 200 cycles; Covaris,Woburn, MA), purified according to the QIAquick protocol and eluted in 20 µl (MinElute PCR purification kit, Cat. No. 28006, PCR purification kit, Cat. No. 28106; Qiagen, Hilden, Germany). After end-repair, A-tailing and adapter ligation size se-

lection of the fragments (335– 365 bp) was performed on a LabChip XT DNA Assay (750 chip; Caliper Life Sciences, Hopkinton, MA). After each step, DNA fragments were purified (QIAquick protocol). The resulting DNA fraction was amplified (11 cycles at a concentration of 5 ng/µl) by PCR amplification (Herculase II Fusion Enzyme with dNTP Combo 200 RXN kit, Cat. No. 600677; Agilent Technologies, Inc.) and purified again. The concentration and length of the DNA fragments of each sample were measured with an Experion^{TM} DNA chip (Experion DNA 12K Reagents and Supplies, Cat. No. 700–7165 and Experion DNA chips, Cat. No. 700–7163; Bio-Rad Laboratories Ltd., Hemel Hempstead, Herts, UK).

1.2.5 Capturing/Enrichment

Target enrichment was performed according to the manufacturer's instructions (SureSelect XT Custom 1kb-499kb library Cat. No. 5190–4806, Agilent Target Enrichment kit and Agilent SureSelect MPCapture Library kit; Agilent Technologies, Inc.). Briefly, samples were diluted or concentrated to 500 ng in 3.4 µl milliQ/elution buffer using a Speedvac machine (Savant SpeedVac SPD101B; Thermo Scientific) at a maximum temperature of 40 °C. Capture probes were mixed with RNase block solution and kept on ice. Each genomic DNA fragment library was mixed with SureSelect BlockMix, heated for 5 min at 95 °C, and kept at 65 °C. While maintaining the sample at 65 °C, hybridization buffer was added and the sample was incubated at this temperature for at least 5 min. The capture library mix was added and the sample incubated for 2 min. Then, the hybridization mixture was added to the capture probes, followed by the addition of the DNA fragment library. Solution hybridization was performed for 24 hr at 65 °C. After hybridization, the captured targets were pulled down by biotinylated probe/target hybrids using streptavidin-coated magnetic beads (Dynabeads MyOne Streptavidine T1; LifeTechnologiesLtd.). The magnetic beads were prewashed three times and resuspended in binding buffer. Next, the captured target solution was added to the beads and incubated for 30 min at room temperature. After purification, the captured DNA was eluted from the streptavidin beads and purified again. Finally, fragments were amplified by 14 cycles of PCR using the complete sample as a template. During the amplification step barcoding index tags were ligated to the fragments. The concentration and length of the DNA fragments of each sample were measured with an Experion^{TM} DNA chip (Experion DNA 12K Reagents and Supplies, Cat.No. 700–7165 and Experion DNA chips, Cat.No. 700–7163; Bio-Rad Laboratories Ltd.). The concentration of each sample was adjusted to 10 nmol/l, and 12 samples were pooled. According to the expected number of sequenced basepairs (1

× 109) and the size of the enrichment kit (323,651 bp) running equimolar pools of 12 samples resulted in a theoretical coverage of 257.5 for all targets.

1.2.6 Sequencing

A sample sheet was prepared on the MiSeq sequencer (Illumina) to provide run details. A standard flow-cell was inserted into the flow-cell chamber. The pooled sample was diluted with chilled HT1 buffer to a concentration of 2 nmol/l and an equal amount of 0.2N NaOH to denaturate the sample was added and incubated for five minutes. A PhiX sample at 2 nmol/l was denatured in the same way. Both the sample and the PhiX were diluted to 8 pmol/l and 1% PhiX was added to the sample. Then, 600 µl of the spiked sample with a final concentration of 8 pmol/l was pipetted into the sample well on the MiSeq consumable cartridge before loading in the cooling section of the MiSeq machine. Sequencing was performed on a MiSeq sequencer using 151 bp paired-end reads, including an index run according to the manufacturer's instructions (MiSeq System user guide part #15027617 Rev. C April 2012, MiSeq Reagent kit 300 cycles, Box1 [ref 15026431] and Box2 [ref 15026432]).

1.2.7 Data Analysis and Variant Annotation

Data analysis was performed using the MiSeq reporter program (Illumina) to generate fastq.gz output files. These were unpacked to create fastQ files. In the NextGENe software (v2.2.1; Softgenetics, State College, PA), we performed the following six steps:

1. the fastQ output file was converted into a FASTA file to eliminate reads that were not "paired" and that did notmeet the criteria of the default settings; it was also checked for "Paired Reads Data";

2. duplicate reads were removed;

3. reads from the converted unique FASTA file were aligned to the reference genome (Human_v37.2). The default settings were extra checked for load-paired end, library size range 200–500 bases, and allowing one mismatch or using seeds. After alignment a *.pjt file was created and opened in the NextGENe Viewer;

4. a mutation report was created using the coordinates from the targeted enrichment kit as a *.bed file to enable calling of SNPs and indels in the regions of interest. Data analyses were limited to ±20 bp of exon-flanking intronic sequences;

5. an expression report was created from which the mean, minimal, and maximal coverage per target and targeted nucleotide was calculated. The coverage was defined as the average number of reads representing a given nucleotide in the reconstructed sequence;

6. a mutation report (*.vcf file) was created annotating all variants.

To interpret the data, additional custom-filtering criteria were imposed to minimize false-positive rates. Variants were filtered for those that are novel (not present in dbSNP133, downloadedApril 1, 2011; or 1000 Genomes databases, downloaded May 25, 2011) and were called pathogenic in case of a truncating variant or a missense variant when it was in silico predicted to be pathogenic, described as pathogenic in the literature or showed cosegregation in affected family members.

1.2.8 Validation of Mutations by Sanger Sequencing

Sequencing analysis of a subset of coding exons and flanking intronic sequences in which a novel variation was identified by NGS was carried out using flanking intronic primers (primer sequences are available upon request). The forward primer was designedwith a PT1 tail (5'-TGTAAAACGACGGCCAGT-3') and the reverse primer was designed with a PT2 tail (5'-CAGGAAACAGCTATGACC-3'). PCR was performed in a total volume of 10 µl containing 5 µl AmpliTaq Gold ®Fast PCR Master Mix (Applied Biosystems), 1.5µl of each primer with a concentration of 0.5 pmol/µl(Eurogentec, Serian, Belgium), and 2 µl genomic DNA in a concentration of 40 ng/µl. Samples were PCR amplified according to our standard diagnostic protocols (available upon request). To rule out sample switches during the procedure we performed a concordance check for 12 highly heterogeneous SNP's for which Sanger sequencing of the respective amplicons is performed in parallel.

1.3 Results

1.3.1 Validation Phase

Sequencing quality

The two validation runs, which contained 12 patient samples each, produced totals of 16,414,062 and 15,186,556 reads, respectively, which were aligned and met the Q30 quality criteria meaning that only reads were included in which the error probability for each base has a likelihood of 1/1,000. The

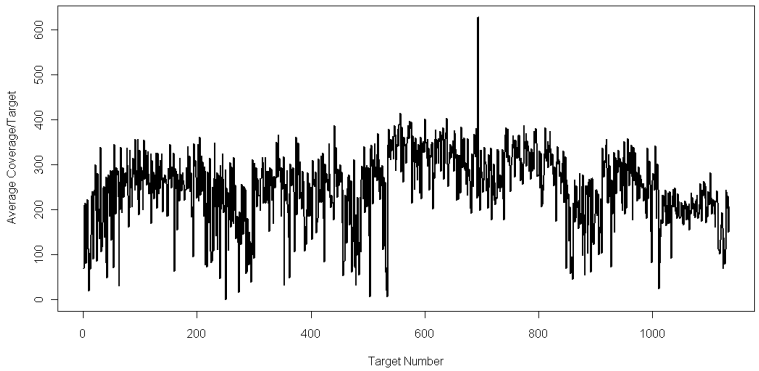

Figure 1.1: Average coverage obtained from 22 different samples of all exon (1,133 exons) and exon/intron junctions ($\pm$20 bp) of 48 genes potentially involved in cardiomyopathy. 99% of the targets show an average coverage of $\geq$30$\times$

pooling was proportional, resulting in a standard deviation between the 12 samples within one run of 0.99% and 0.75%, respectively. The coverage statistics were comparable between both runs (Table 1.2) as well as in subsequent runs (data not shown). The mean coverage per target was 246 and 251 reads, respectively, which is in accordance with our theoretical calculated coverage of 257.5. In 1,084 of the 1,134 targets, the minimal coverage was at least 30 reads in more than 22 out of 24 patients (Fig. 1.1). The validation runs had 99.4% to 99.1% mean coverage >30 of all targets, respectively. For 50 targets, the coverage of at least one basepair position was less than 30 reads in more than two out of the 24 patients. Of these 50 targets, a total of 4,398 bp had a coverage lower than 30 reads. When investigated in more detail, the coverage within such targets varied significantly and in most of these only a few basepairs were covered below 30, resulting in 67 different regionswith a coverage below 30. One example of such a target is shown in Figure 1.2.

Specificity and sensitivity of targeted NGS: confirmation of SS variants

In previous SS analyses, a total of 90 variants in 14 different genes had been identified in the 24 patients used for validation (2 runs). All these variants were also detected with our targeted NGS approach applying the Agilent SureSelect kit (Fig. 1.3) and resulting in no false negatives. This included

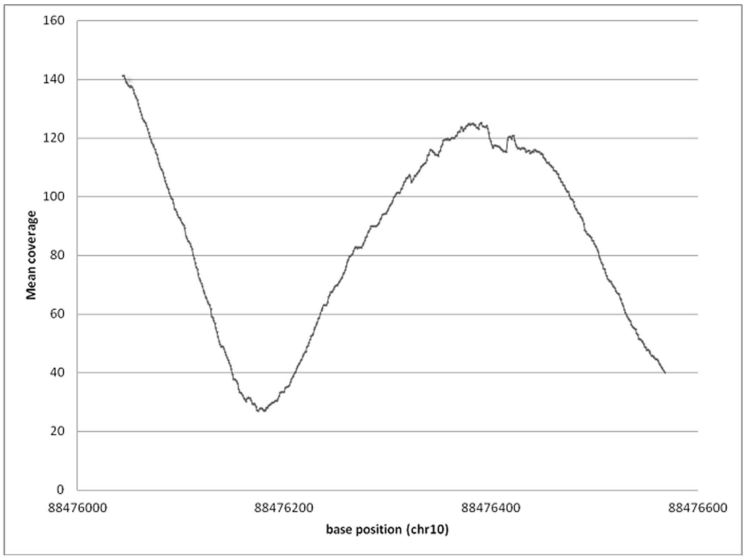

Figure 1.2: Coverage of one target, exon 9 of the *LBD3* gene on chromosome 10 (NCBI build 37, UCSC hg19), representing one region with a coverage $\leq$30 in one patient.

Table 1.2: Overview of the Sequence Performance for the Validation Runs

	Run 1	Run 2	Average of both runs
Cluster density (k/mm^2)	1,289	1,119	1,204
% Cluster PF	89.3	94.6	92.0
Q30	80.3	83.9	82.1
Total reads	17,168,243	15,788,049	16,478,146
Matched reads	16,414,062	15,186,556	15,800,309
% reads in .fasta file aligned	96	96	96
Mean mean coverage targets	246	251	248
Mean min coverage targets	166	179	173
Mean max coverage targets	299	297	298
%Targets Mean $<$30	0.6	0.9	0.7
%Targets Mean $>$30	99.4	99.1	99.3
%Targets Min $<$30	2.6	2.5	2.6
%Targets Min $>$30	97.4	97.5	97.4
%Targets Max $<$30	0.4	0.7	0.5
%Targets Max $>$30	99.6	99.3	99.5

84 substitutions and six indels (four deletions, two insertions). No additional variants were identified in these genes, comprising 55,784 bp. We therefore concluded that for these 24 samples there was full concordance with the SS results.

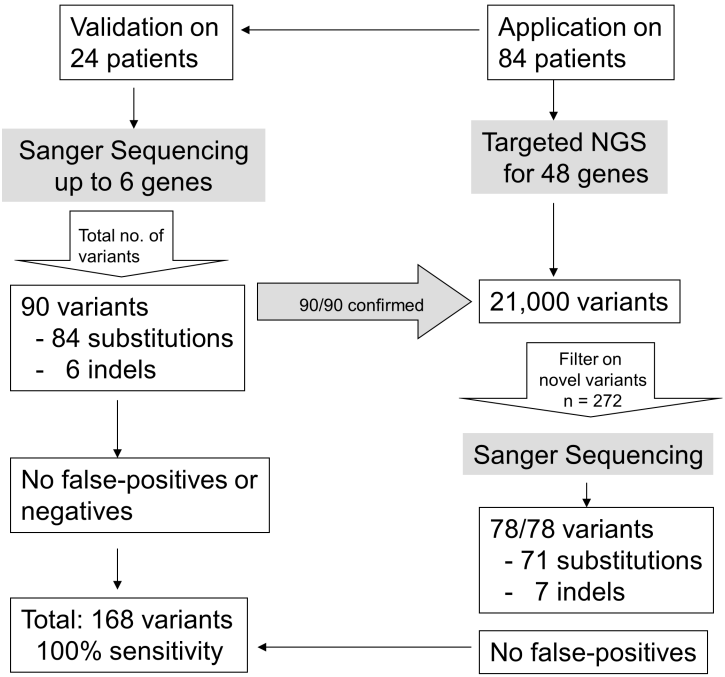

Figure 1.3: Summary of the results of our confirmation analyses.

1.3.2 Application Phase

Sequence specificity of targeted NGS: confirmation of NGS variants

Using targeted NGS of 48 genes, approximately 21,000 variants were identified in 84 unique patients (Fig. 1.3), including the 90 variants that had been previously detected with SS. Of these variants, 272 were novel (245 substitutions, 27 indels). On average, we identified three novel variants per patient. For validationwith SS, 78 out of the 272 novel variants were selected, including detected indels (n = 7). The largest deletion comprised 18 bp and the largest insertion 8 bp. Notably, of the 71 substitutions, one was initially not confirmed by SS. This could be explained by the presence of a SNP in the primer binding site of the forward primer. A subsequent SS experiment, in which an alternative set of primers was used, did confirm the presence of this variant. In summary, a total of 168 variants were confirmed with

SS. Based on these data, we reached a 100% sensitivity (at 95% confidence 97.76%–100%) [15] with the NGS targeted approach.

Diagnostic yield

Applying this targeted NGS strategy, our first results indicate that the diagnostic yield is significantly improved from 15% to about 40%,mostly for DCM. However, this is based on small numbers of samples and the increase in yield may be even higher if this strategy is applied on a regular basis for larger series. Where regular routine diagnostics involves stepwise testing of up to about 10 different genes (which can easily take more than 1 year to complete), using targeted NGS of entire gene-panels on the MiSeq sequencer could theoretically provide reporting times of no more than 2 weeks. At this stage, we are aiming for reporting times of 4–6 weeks, a huge improvement compared with current diagnostic services.

1.3.3 Reproducibility of Targeted NGS

The entire procedure was performed twice for five samples, including sequencing in different runs. On average, 231 variants (198– 268) were detected per sample, and on average, 10 unique variants (8–14) were differently reported between the two analyses of identical samples. In total, 1,007 variants were detected and 51 of these were differently reported in the two separate analyses of the same sample resulting in a nonconcordance rate of 0.00315% according to the number of sequenced bp (five times 323,651 bp of the targeted NGS kit). These differences can be attributed to three underlying causes: (1) in 12 out of 51 cases this was due to coverage differences, which meant variants were either not reported because of a too low coverage or reported when the coverage was just above threshold levels using the default settings; (2) in 24 out of 51 cases this was explained by alignment problems due to poly-T/A stretches, resulting in different annotating of the same variant; and (3) 15 out of 51 were due to differences in heterozygote levels, which meant variants that were present in $<$20% of reads were not reported. Variants that fall within the first two categories are "true variants" that were either missed or reported as the result of analysis software settings or limitations. In contrast, variants in the third category most likely represent recurring technical artefacts, as all were repeatedly reported in a significant number of patients and in different runs, but were nonetheless not reported in the dbSNP and/or 1000 Genomes databases. Considering the artefacts as potential false positives the technical specificity is 0.0009269%. In our future bioinformatic analyses, we will filter for the variants of the third category during our selection for potentially

interesting variants, in addition to other filtering steps.

1.4 Discussion

We present the validation of a targeted resequencing method for cardiomyopathy-associated genes and our results support its implementation in routine diagnostics. In this study, all the 168 variants identified by our NGS-approach were confirmed with SS (Fig. 1.3). The variants included deletions up to 18 bp and insertions up to 8 bp. No false-negative or false-positive results were obtained for variants selected for confirmation. We therefore conclude that, at a coverage of at least 30 times per nucleotide, the performance of our procedure is comparable with SS. ES is likely to become the most commonly used tool for identifying genes in Mendelian diseases in the coming years [5]. This approach has been shown to be successful in cases of rare monogenetic disorders [4, 8, **?**] and of intellectual disability [16]. However, as demonstrated by Gilissen et al. (2012)[5], 2128 (5.7%) of 37,424 disease-causing variant positions from the Human Genome Mutation Database are not covered with the 50Mb SureSelect ES kit (Agilent Technologies, Inc.). From our experience, we know that all the 48 genes we targeted are covered by probes in the 50Mb ES kit. However, the coverage performance varied significantly between exons within a gene and between different genes, and for some regions the coverage was $<$20 times, too low for reliable variant detection. This is exemplified by the *TTN* gene. Recently, Herman et al. (2012)[7] showed that *TTN* truncating mutations are a common cause of DCM, occurring in approximately 25% of familial cases of idiopathic DCM and in 18% of sporadic cases. From our ES data, we have calculated the average coverage per target for the coding regions of *TTN*. We found that 7% of the targets sequenced had an average coverage of $\leq$20 times (around 25 exons) and among those, 12 exons showed an average coverage of $\leq$10 times. It is therefore very likely we would miss clinically relevant variants in these regions of low coverage. In contrast, the targeted region of the *TTN* gene in our designed kit shows a 100% coverage for all exons and the respective nucleotides were all covered $\geq$30 times, with a high reproducibility between different samples. We therefore decided to continue developing our targeted resequencing method to overcome the shortcomings of incomplete representation and coverage of exons in ES experiments. The first prerequisite for high sensitivity of a NGS method is the development of a well-designed enrichment kit. We chose to use the SureSelect kit (Agilent Technologies, Inc.) as the e-Array programme used for kit design offered flexibility in optimizing the respective probe design. The number of tilings of each target can be cho-

sen and extra baits can be added for GC-rich targets to increase coverage. A theoretical 100% representation was reached for all of our targets. Based on our data, the theoretical representation given by e-Array is indicative for the actual coverage. Because the cost of a targeted custom-made enrichment kit is rather high, a good prediction of the coverage is an advantage before ordering such a kit for diagnostic use. The second prerequisite is high coverage of preferably all the targets. Setting the threshold at a coverage $\geq$30, we found only 50 targets out of the 1,134 with less coverage of the nucleotides, mostly in a part of the respective targets. We therefore decided that, parallel to targeted NGS, we will perform SS for targets with a low coverage from those genes of which the clinical relevance is uncontested (e.g., *MYH7*, *TNNI3* or *MYBPC3* for HCM; *LMNA*, *MYH7* or *MYBPC3* for DCM; and *PKP2* for ARVC) to ensure complete coverage of the respective amplicons (see Table 1.3 for general recommendations).

Table 1.3: Resulting Diagnostic Workflow and Implementation Guidelines

Workflow	Recommendations
Enrichtment kit construction	Theoretically 100% horizontal and vertical coverage of all targets
Sample preparation	
Days 1-3	Automated, that is, using a Bravo or Caliper robot (Agilent Technologies, Inc./Caliper Life Sciences, Hopkinton, MA)
Sample Enrichment	Bar-coding samples to a theoretical mean coverage of 250 for all targets resulting in a coverage of at least 30 per nucleotide in p8% of targets
Days 4-6	Avaiding sample-mix-up by spiking unique DNA sequences before the procedure or including a limited SNP analysis for each individual patient
Sequencing on bench-top machine	80% of the reads with Q30
Days 7-8	
Data analysis	
Days 8-10	Minimal coverage of 30 per nucleotide In house (control) variant database for filtering A predifined variant filtering precedure, preferably automated in software programmes like the NGS bench lab from CARTAGENIA (Leuven, Belgium)[1]
Confirmation with Sanger Sequencing	
Days 11-20	Obsolete at a coverage of >30 per nucleotide Coverage of targets structurally below 20: Sanger sequencing in parallel with NGS Incidental coverage below 20: Sanger sequencing depending on the target's clinical relevance Coverage between 20 and 30: visual inspection, Sanger sequencing of novel variants
Total turn-around time	21 days

Valencia et al. (2012)[14] developed a SureSelect enrichment kit for congenital muscular dystrophy for 321 targets (12 genes) and 95% of them had a coverage of at least 20. According to their data, the coverage was below 20 times for two genes due to a high GC content. In contrast, our kit represents amuch better coverage (99% covered more than 30 times). There are several explanations for this difference, for instance the tiling of the baits, differences in the overall GC content, or the number of pooled patients, which make a good comparison difficult. In our approach, 12 samples were

[1] Basepair position according to NCBI build 37

pooled based on the size of the enrichment kit to reach a coverage of at least 30 times per basepair for most of the targets. Because no false-positive or -negative results were detected, this would seem to be a safe threshold. One could even consider whether more than 12 patients could be pooled or the coverage threshold reduced to >20 times instead of >30 times. In Table 1.3, we give some general recommendations on the clinical laboratory implementation and quality assessment of targeted resequencing methods. These recommendations are in line with the general guidelines for assuring the quality of NGS in clinical laboratory practice formulated by the national workgroup of the US Centers for Disease Control and Prevention [3]. A 100% sensitivity (95% confidence: 97.76%–100%) was reached with our approach and a specificity of nearly 100% (0.00315% false positive). Gowrisankar et al. (2010)[6] reported a false-positive rate of 0.011 ±0.002%, close to 100% specificity for 41,475 bp using an Illumina GAII sequencing machine and targeted resequencing of 19 DCM genes. However, four out of the 160 basepair substitutions and three out of 31 indels were missed, including one 18 bp duplication. The basepair substitutions were missed because of insufficient coverage (<30 times), whereas the indels were likely missed due to sequencing of short read lengths (36 bp). In our approach, 151 bp reads were used and we were able to detect an 18 bp deletion, the largest indel detected in our study. In total, 17 indels detected were confirmed with SS, but it is debatable how many and which type of indels should be confirmed by SS for proper validation. Depending on the gene panel to be sequenced, it seems obvious to choose patients with the largest known indels for validation. Gowrisankar et al. (2010)[6] recently reported an 18 bp duplication and Herman et al. (2012)[7] a 13 bp deletion in the titin gene, which seem to be the largest indels associated with cardiomyopathies so far. As indels of that size were detected in our procedure, we are convinced we can retain 100% sensitivity. Moreover, according to our results, we would have missed one variant with SS due to a SNP in the primer sequence. This suggests that resequencing after hybridization-based enrichment of targets may even outperform SS. The importance of longer read lengths was underscored by the results of Voelkerding et al. (2010)[17]. They performed SureSelect enrichment for 12 genes responsible for congenital muscular dystrophy in combination with sequencing on a SOLiD machine. Two out of the 34 identified variants were not confirmed with SS because of sequence read misalignment between two closely related genes. As a probe based method, not only targeted sequences but also highly homologous pseudogenes and other homologous sequences, such as those present in gene families and domain analogs will be captured [1]. Highly homologous sequences coalign to the reference sequence. However, it is uncertain to what extent regions of high-homology may negatively af-

fect the sensitivity and specificity. In general, construction of a unique tiled bait library using differences in the neighboring intron sequences and eventually longer paired end reads can reduce this problem. The reproducibility of our procedure was tested by repeating the procedure for five samples. The 99.99685% concordance of all detected variants demonstrates the high performance of our targeted enrichment and MiSeq resequencing method. Apart from low coverage an alignment problem due to poly A/T stretches resulted in discrepancies. However, these variants will not result in false positives. Discrepancies due to differences in the heterozygote level of 20% might be considered as technical false positives (0.0009269%). However, according to our analyses criteria we would have filtered these variants out. In summary, the differences seen between the separate analyses of the five repeated samples were due to bioinformatic threshold and annotation settings and not due to technical limitations. Variants with an allelic imbalance need careful follow up. This is in line with the first report on a MiSeq-based sequencing method in which drafting genomic sequences of E. coli resulted in an error rate of 0.1 substitutions per 100 bases and a near absence of indel errors[9]. This, together with the almost 100% sensitivity and specificity of our results, raises the question whether a variant still needs to be confirmed with SS, as is often daily practice in clinical diagnostics at the moment. Zhang et al. (2012)[18] felt it was necessary for two reasons: first, to remove incorrect calls due to experimental errors, and second, to confirm a diagnosis. However, as they discussed, confirmation becomes burdensome or impossible when a large number of novel variants need to be confirmed and this would result in long turn-around times. We therefore propose to refrain from confirming results with SS as long as the coverage is >30 times per nucleotide. In addition, targets that are not covered or badly covered can either be excluded from the final report or SS of these targets can be performed in parallel. At a coverage between 30 and 20 times, visual inspection of the regions is recommended (see Table 1.3 for general recommendations).

1.5 Conclusion

Our data convincingly demonstrate that targeted NGS of a disease-specific subset of genes can be reliably implemented as a stand-alone diagnostic test.

1.6 Acknowledgments

We thank Jackie Senior for editorial advice.

Disclosure Statement

The authors declare no conflict of interest.

Bibliography

[1] Emily M. Coonrod, Rebecca L. Margraf, and Karl V. Voelkerding. Translating exome sequencing from research to clinical diagnostics. *Clinical Chemistry and Laboratory Medicine*, 50(7), Jan 2012.

[2] Joep de Ligt, Marjolein H. Willemsen, Bregje W.M. van Bon, Tjitske Kleefstra, Helger G. Yntema, Thessa Kroes, Anneke T. Vulto-van Silfhout, David A. Koolen, Petra de Vries, Christian Gilissen, and et al. Diagnostic exome sequencing in persons with severe intellectual disability. *New England Journal of Medicine*, 367(20):1921–1929, Nov 2012.

[3] Amy S Gargis, Lisa Kalman, Meredith W Berry, David P Bick, David P Dimmock, Tina Hambuch, Fei Lu, Elaine Lyon, Karl V Voelkerding, Barbara A Zehnbauer, and et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology*, 30(11):1033–1036, Nov 2012.

[4] Christian Gilissen, Heleen H. Arts, Alexander Hoischen, Liesbeth Spruijt, Dorus A. Mans, Peer Arts, Bart van Lier, Marloes Steehouwer, Jeroen van Reeuwijk, Sarina G. Kant, and et al. Exome sequencing identifies wdr35 variants involved in sensenbrenner syndrome. *The American Journal of Human Genetics*, 87(3):418–423, Sep 2010.

[5] Christian Gilissen, Alexander Hoischen, Han G Brunner, and Joris A Veltman. Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*, 20(5):490–497, Jan 2012.

[6] Sivakumar Gowrisankar, Jordan P. Lerner-Ellis, Stephanie Cox, Emily T. White, Megan Manion, Kevin LeVan, Jonathan Liu, Lisa M. Farwell, Oleg Iartchouk, Heidi L. Rehm, and et al. Evaluation of second-generation sequencing of 19 dilated cardiomyopathy genes for clinical applications. *The Journal of Molecular Diagnostics*, 12(6):818–827, Nov 2010.

[7] Daniel S. Herman, Lien Lam, Matthew R.G. Taylor, Libin Wang, Polakit Teekakirikul, Danos Christodoulou, Lauren Conner, Steven R. DePalma, Barbara McDonough, Elizabeth Sparks, and et al. Truncations of titin causing

dilated cardiomyopathy. *New England Journal of Medicine*, 366(7):619–628, Feb 2012.

[8] Alexander Hoischen, Bregje W M van Bon, Christian Gilissen, Peer Arts, Bart van Lier, Marloes Steehouwer, Petra de Vries, Rick de Reuver, Nienke Wieskamp, Geert Mortier, and et al. De novo mutations of setbp1 cause schinzel-giedion syndrome. *Nature Genetics*, 42(6):483–485, May 2010.

[9] Nicholas J Loman, Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain, and Mark J Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30(5):434–439, Apr 2012.

[10] Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, and et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1):30–35, Nov 2009.

[11] Nadine Norton, Duanxiang Li, and Ray E. Hershberger. Next-generation sequencing to identify genetic causes of cardiomyopathies. *Current Opinion in Cardiology*, 27(3):214–220, May 2012.

[12] Anna Posafalvi, Johanna C Herkert, Richard J Sinke, Maarten P van den Berg, Jens Mogensen, Jan D H Jongbloed, and J Peter van Tintelen. Clinical utility gene card for: Dilated cardiomyopathy (cmd). *European Journal of Human Genetics*, 21(10), Dec 2012.

[13] Anna-Maija Sulonen, Pekka Ellonen, Henrikki Almusa, Maija Lepistö, Samuli Eldfors, Sari Hannula, Timo Miettinen, Henna Tyynismaa, Perttu Salo, Caroline Heckman, and et al. Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biology*, 12(9):R94, 2011.

[14] C. Alexander Valencia, Devin Rhodenizer, Shruti Bhide, Ephrem Chin, Martin Robert Littlejohn, Lisa Mari Keong, Anne Rutkowski, Carsten Bonnemann, and Madhuri Hegde. Assessment of target enrichment platforms using massively parallel sequencing for the mutation detection for congenital muscular dystrophy. *The Journal of Molecular Diagnostics*, 14(3):233–246, May 2012.

[15] B.L. Van der Waerden. *Mathematische Statistik*. Springer Verlag, Göttingen: Heidelberg, 1957.

[16] Lisenka E L M Vissers, Joep de Ligt, Christian Gilissen, Irene Janssen, Marloes Steehouwer, Petra de Vries, Bart van Lier, Peer Arts, Nienke Wieskamp, Marisol del Rosario, and et al. A de novo paradigm for mental retardation. *Nature Genetics*, 42(12):1109–1112, Nov 2010.

[17] Karl V. Voelkerding, Shale Dames, and Jacob D. Durtschi. Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy. *The Journal of Molecular Diagnostics*, 12(5):539–551, Sep 2010.

[18] Wei Zhang, Hong Cui, and Lee-Jun C. Wong. Application of next generation sequencing to molecular diagnosis of inherited diseases. *Topics in Current Chemistry*, page 19–45, 2012.

List of Tables

List of Figures