METHODS

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

Flexible, Scalable, and Efficient Targeted Resequencing on a Benchtop Sequencer for Variant Detection in Clinical Practice

Kim De Leeneer,[1] Jan Hellemans,[2] Wouter Steyaert,[1] Steve Lefever,[1] Inge Vereecke,[1] Eveline Debals,[1] Brecht Crombez,[1] Machteld Baetens,[1] Mattias Van Heetvelde,[1] Frauke Coppieters,[1] Jo Vandesompele,[1,2] Annelies De Jaegher,[1] Elfride De Baere,[1] Paul Coucke,[1] and Kathleen Claes[1]*

[1] *Center for Medical Genetics Ghent, Ghent University, Ghent, Belgium;* [2] *Biogazelle, Zwijnaarde, Belgium*

Communicated by Mireille Claustres

Received 7 August 2014; accepted revised manuscript 2 December 2014.
Published online 12 December 2014 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22739

ABSTRACT: The release of benchtop next-generation sequencing (NGS) instruments has paved the way to implement the technology in clinical setting. The need for flexible, qualitative, and cost-efficient workflows is high. We used singleplex-PCR for highly efficient target enrichment, allowing us to reach the quality standards set in Sanger sequencing-based diagnostics. For the library preparation, a modified NexteraXT protocol was used, followed by sequencing on a MiSeq instrument. With an innovative pooling strategy, high flexibility, scalability, and cost-efficiency were obtained, independent of the availability of commercial kits. The approach was validated for ~250 genes associated with monogenic disorders. An overall sensitivity (>99%) similar to Sanger sequencing was observed in combination with a positive predictive value of >98%. The distribution of coverage was highly uniform, guaranteeing a minimal number of gaps to be filled with alternative methods. ISO15189-accreditation was obtained for the workflow. A major asset of the singleplex PCR-based enrichment is that new targets can be easily implemented. Diagnostic laboratories have validated assays available ensuring that the proposed workflow can easily be adopted. Although our platform was optimized for constitutional variant detection of monogenic disease genes, it is now also used as a model for somatic mutation detection in acquired diseases.

Hum Mutat 36:379–387, 2015. © 2014 Wiley Periodicals, Inc.

KEY WORDS: targeted resequencing; NGS; uniform target enrichment; clinical implementation; ISO15189 accreditation

Introduction

Since the 1970s, Sanger sequencing [Sanger et al., 1977] has been the gold standard for mutation analysis in molecular diagnostics. However, because of the overall cost [Martinez and Nelson, 2010] and the developments in novel sequencing technologies, it is time for a paradigm shift. Currently, massive parallel sequencing, also called next-generation sequencing (NGS), enables generation of millions of DNA sequences simultaneously in notably reduced turnaround time (TAT) and cost. For clinical use, new sequencing workflows must be at least as specific, sensitive, time- and cost-efficient, scalable, and flexible as the Sanger sequencing approach applied in routine clinical diagnostic laboratories. As for many monogenic conditions, there is no need for whole exome sequencing (WES) or whole genome sequencing (WGS) in routine diagnostics, targeted resequencing of a gene (panel) is preferred and sufficient. Indeed, for genetically heterogeneous disorders, like hereditary blindness, deafness, connective tissue disorders, cardiomyopathies, familial cancer syndromes, customized or commercially available fixed sequence capture-based gene panels are frequently used in a clinical setting [for example, Mook et al., 2013; Sivakumaran et al., 2013; Wang et al., 2013; Arvai et al., 2014; Brownstein et al., 2014; Jin et al., 2014; Pugh et al., 2014].

However, for many monogenic disorders for which only one, a few or new genes need to be investigated, a fixed capturing approach is not an ideal solution. Sequence capture methods do not allow quick and flexible implementation of new targets representing novel disease genes of interest as a redesign of the complete panel is often required. Furthermore, the sensitivity of such methods may drop compared with Sanger sequencing as not all exons may be sufficiently covered. For instance, in a recent article a capture efficiency of 70%–80% was reported [Lupski et al., 2013] especially exons with deviating GC/AT content ratio are known to be less efficiently hybridized and amplified [Clark et al., 2011; Knies et al., 2012]. Another issue is the relatively high proportion of off-target capturing especially if the regions of interest contain long stretches of highly (>90%) homologous sequences. This has a negative effect on the coverage and the variant calling in the target region and may increase the sequencing cost.

An elegant alternative for a capturing approach is PCR-based amplification, with the versatility in design as a major advantage, followed by flexible adjustments of the target regions of interest. Most diagnostic laboratories have already invested in the optimization of singleplex PCR assays for Sanger sequencing of genomic regions of interest. The same amplicons can be used for sequencing on NGS benchtop sequencers.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Kathleen Claes, De Pintelaan 185, 9000 Ghent, Belgium. E-mail: kathleen.claes@ugent.be

Contract grant sponsors: GOA, Ghent University (grant no. BOF10/GOA/019) (KDL, JV, KC); Spearhead financing Ghent University Hospital (KDL); FC and SL are postdoctoral fellows and EDB is a senior clinical investigator of the FWO. (Research Foundation—Flanders). Agency for Innovation by Science and Technology in Flanders (IWT) (MVH)

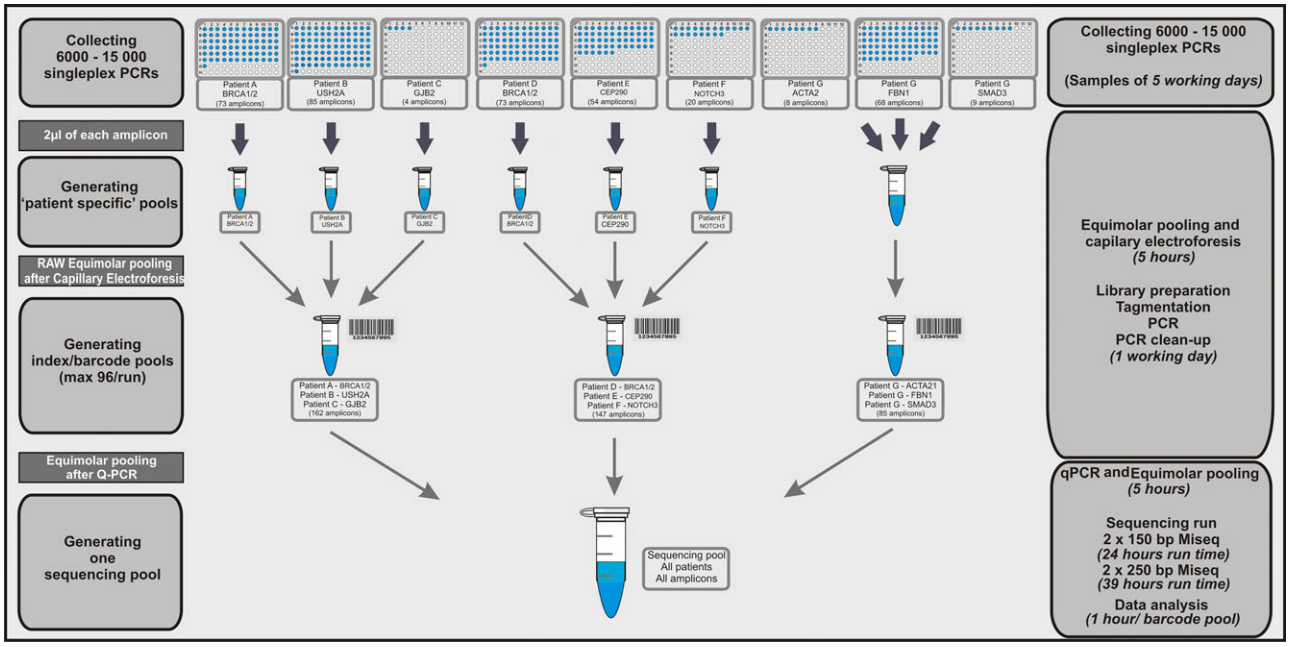

Figure 1. Schematic representation of the different pooling steps. A schematic overview of the different pooling steps during target enrichment and library preparation is shown. In summary, we perform singleplex PCRs for all required amplicons for every patient (for instance, for patient A BRCA1/2 testing is required, we perform 73 singleplex PCRs covering the complete coding region of these genes; for patient B USH2A analysis is requested and we perform 86 singleplex PCRs). In a second step, we take 2 μl of each PCR product and pool these per patient and per amplicon. Next, we pool the PCR-mixes of several patients—several patients can be "attached" to the same barcode/index as long as they are screened for different loci. In a final step, all index pools undergo library preparation and are pooled in a single tube prior to sequencing. Therefore, generally, we pool 6,000–15,000 singleplex PCRs per sequencing experiment (on average 10,000 singleplexes per run).

In order to realize short TATs required for clinical molecular diagnostics, it is important to start sequencing runs on a regular basis. To keep it cost efficient, the capacity of the devices should be optimally used.

We developed a flexible, scalable, and cost-efficient targeted re-sequencing workflow to replace Sanger sequencing in an ISO15189 accredited laboratory with a diagnostic sensitivity and specificity matching the Sanger standard. Easy addition of new target genes was one of the main goals of our approach. To this end, we implemented an efficient PCR-based target enrichment approach and library preparation and thoroughly evaluated sensitivity and specificity, costs, and TATs.

Material and Methods

Selection of Targets

A broad range of different disease genes, associated with monogenic disorders (for instance, genes involved in familial cancer syndromes, connective tissue disorders, hereditary blindness, mental retardation, etc.) were tested in multiple sequencing runs. This workflow was applied for 265 different genes (An overview of the GenBank reference Sequences for the genes studied is given in Supp. Table S1a), with the number of genes continuing to increase. A schematic workflow of the target enrichment steps and library preparation including a time schedule is shown in Figure 1.

Target Enrichment: Generating "Patient Specific" Pools and "Index" Pools

We implemented a high-throughput PCR-based approach for target enrichment. To this end, primers were designed using an in house developed primer design software program (primerXL-www.primerxl.org; Lefever et al., in preparation; similar PCR assays are available through pxlence – www.pxlence.com) and ordered at Integrated DNA Technologies. In total, there are over 4,000 PCRs that cover the complete coding and splice site regions (exon borders ±20 bp) of the 265 genes included in this study.

PCRs were performed in 10 μl total volume. The uniform amplification mixture consists of 5 μl of KAPA2G Robust HotStart ReadyMix (2×; KAPA Biosystems,Wilmington, MA), 50 ng of DNA template, and 2.5 μl of premixed 1 μM of each forward and reverse primer. The temperature cycling protocol consists of initial denaturation step at 95°C for 3 min, followed by 35 cycles of denaturation at 95°C for 15 sec, annealing at 60°C for 10 sec, and an extension at 72°C for 15 sec. Final extension was accomplished at 72°C for 1 min.

In a first pooling step, the regions amplified for 1 DNA sample are pooled together (2 μl/amplicon), generating a "patient specific pool." Patient specific pools are pooled with other patient specific pools when these are analyzed for different regions in the genome. For example, a patient who needs a molecular diagnosis for *BRCA1/2* can easily be pooled with another patient who needs to undergo a screening for a connective tissue disorder. Because of the targeted enrichment, the reads mapping to the region of interest can be traced back to the correct patient. All patient-specific pools are run on a Labchip GX (Caliper Life Sciences, Hopkinton, MA)

instrument to get a rough estimation of the molarity of the pools. For this analysis, the "smear analysis" function in the Labchip GX software is used, the molarity calculated by the program is divided by the number of amplicons present in the patient specific pool. A second pooling step is performed to generate "index pools" (this is the combination of different patient-specific pools which will be attached to the same sequencing index) taking into account capillary electrophoresis based DNA concentration assessment. Hereto, a homemade script and template are used to generate index pools with a similar number of amplicons. The number of amplicons in every pool and the molarity of this pool are exported from the Labchip GX software. Based on this export, the script calculates the average number of amplicons for one index (e.g., 10,000 amplicons/48 indices = 208 amplicons/index). The script takes into account that samples with identical gene names cannot be pooled together and generates the pipetting volumes needed to compose the index pools.

Prior to library preparation, 10 μl of these "index" pools are cleaned by 10 μl of Agencourt Ampure XP beads (Beckmann Coulter, Pasadena, CA) according to the manufacturer's instructions. The purified index pool is measured using Qubit dsDNA HS (high sensitivity) assay kit on the Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA) and diluted to 0.8 ng/μl.

A detailed overview of the different pooling steps is given in Figure 1.

Automation

As we strongly invested in uniform PCR conditions for all amplicons, set-up of the singleplex PCRs can easily be performed by robots. The liquid handling in our set-up is done by a Freedom EVO 150 (Tecan, Männedorf, Switzerland). The pipetting files of the robots are generated with sLIMS (Genohm, Lausanne, Switzerland), to allow a variable set-up every run, since the composition of 1 Miseq experiment depends on the requests received during a certain week, but is never identical to a previous run. Furthermore, each sample or PCR is stored in the online sLIMS system, which enables us to perform detailed sample tracking. In addition, post PCR robots (e.g., EpMotion 7075 LH) are used to pool the amplicons.

Library Preparation

Target enrichment is followed by Nextera XT library preparation (Illumina, San Diego, CA) with several modifications to establish a robust workflow. The input material for the library prep was set on 0.8 ng after Nextera XT tagmentation and PCR for adaptor ligation; the PCR products are cleaned with 50 μl of Ampure XP beads (Beckmann Coulter) and size selection is performed with 83% ethanol.

The beads normalization of the tagmented adapter pools was replaced by qPCR-based normalization using the KAPA qPCR library quantification kit (KAPA Biosystems). Samples are run according to the manufacturer's instructions in triplicate on a LightCycler480 instrument (Roche, Basel, Switzerland). After qPCR quantification of all index pools, the pools are diluted to 2 nM and equimolarly pooled into the final sequencing pool.

Sequencing

The final library pool is denaturated and 12 pM with a 5% PhiX spike-in is loaded onto a flow cell in a MiSeq instrument (Illumina) and subjected to cluster generation and sequencing using a paired-end 2 × 250 bp (v2) cycle protocol according to the manufacturer's instructions.

Data Analysis

The generated Fastq files are mapped with CLC bio Genomics Workbench v6 (Qiagen, Venlo, The Netherlands). An overview of the CLC Bio settings is listed in Supp. Table S2. The variant calls are annotated with VEP (Ensembl-http://www.Ensembl.org). Several in house developed Perl scripts are used to facilitate data analysis and reporting. An example of our output format for one patient screened for the *BRCA1/2* genes is shown in Supp. Worksheets S1–S3. Worksheet S1 provides an overview of all amplicons which need Sanger sequencing (because of too low coverage or to confirm a deleterious variant); Worksheet S2 shows an overview of all detected variants in the regions of interest for this DNA sample; and in Worksheet S3, an overview of the minimal coverage for each individual amplicon is summarized.

All variants mentioned in this manuscript have been submitted to the ClinVar database (http://www.ncbi.nlm.nih.gov/clinvar/).

Results

To evaluate the performance of our approach in a diagnostic setting, a thorough optimization and validation was performed of all pre- and post-sequencing steps. The following crucial parameters were analyzed: flexibility, throughput, coverage, TAT, sensitivity, specificity, and cost efficiency. The workflow should outperform or at least emulate current methods used in molecular diagnostics for several of these parameters to make the implementation worth while.

Flexibility through Target Enrichment and Library Preparation

To retain the same flexibility as with Sanger sequencing, we chose a singleplex PCR-based approach for target enrichment. Primers were designed using the primerXL database, which currently contains designs for all human coding exons; therefore, implementation of additional targets can be done very fast.

Several PCR conditions were tested until we obtained uniform reaction conditions for at least 95% of all designed amplicons. Not surprisingly, amplicons encompassing GC-rich regions are the majority of the failed reactions. Primers were redesigned rather than optimizing the reaction conditions for the less optimal 5%. Intensive validation (evaluation of amplification on control DNA, of primer dimers, of the specificity and robustness, and so on) is performed before the PCRs are implemented into daily routine practice, hence the failure rate downstream is negligible.

Throughput and Coverage

We opted for a threshold of 36-fold coverage based on statistical power calculations previously made by our group [De Leeneer et al., 2011]: 36-fold coverage theoretically guarantees detection of a heterozygous variant with a probability of >99.99% when variants present in at least 15% of the reads are considered as true variants. The rationale for the 15% variant frequency is described below in the section: ''Sensitivity''.

On average 10,000 amplicons are sequenced on a single MiSeq run, guaranteeing a minimal coverage of >36× for at least 97% of the samples. However, the throughput can still be increased. Table 1 shows the results of several experiments. Taking into account the average yield per run (7.2 G) and the average number of reads with

Table 1. Overview of the Sequencing Yield per Run

Run	Total yield (Gb)	Clusters PF (K/mm^2)	Total reads	%>=Q30 (Gb)	# Amplicons sequenced	Average coverage	Amplicons <36× coverage
1	7.2	764	15,846,069	74% (5.3)	9,716	482×	3.00%
2	4.5	459	9,591,477	78% (3.5)	10,258	390×	2.50%
3	7.7	802	16,619,639	74% (5.7)	9,571	884×	1.50%
4	7.3	751	15,809,595	74% (5.4)	9,859	822×	2.30%
5	9.0	961	20,867,166	68% (6.1)	10,302	827×	1.20%
6	7.3	761	15,703,759	75% (5.5)	9,097	901×	1.00%

In comparison, the theoretically predicted yield for a 2 × 250 bp Miseq run is approximately 15 M reads or 7.5 Gb.
PF, passing filter.

a Phred score >30 (5.2 G), theoretically up to 30,000 amplicons can be pooled in one single (2× 250 bp v2 run) run, still resulting in an average coverage of 300×. However, not the average coverage but the minimal coverage determines the number of gaps that need to be completed with alternative techniques. From Table 1, it is clear that the number of amplicons not meeting the 36× coverage threshold is not limited by the yield of our sequencing runs, but by the uniformity of the distribution of the reads. Therefore, a balance needs to be defined between the effort of optimizing uniform distribution and labor intensity of the target enrichment and library preparation versus the decrease of sequencing costs.

Figure 2 shows the detailed distribution of coverage within one run. When the distribution of coverage is compared within one run (= comparison of reads/index), within index pools (= coverage/gene), and within a patient pool (= coverage of amplicons/patient), the uniformity is optimal for the number of reads between the different indices (CV: 0.12), which is not surprising since the composition of the final pool is accurately normalized with qPCR. An intermediate uniformity of coverage is seen between the different genes pooled together for one index (CV range: [0.16–0.33]). With capillary electrophoresis a raw estimation of the molarity for each amplicon is determined. Although the values are not exact, it is sufficient to obtain similar fold coverage ratios for each assay when a patient specific pool of 150 assays is combined with a patient specific pool of only five assays. The largest variation is seen within the amplicons for one patient since we add a fixed volume (2 μl) of each amplicon to the patient specific pool. There is no normalization in terms of intensity or GC content of the PCR product.

Turnaround Time

Previously, the TATs of the analyses included in this project varied from six weeks to six months. A frequent reason for the delays is that for rare Mendelian diseases, one waits until enough samples are collected before starting an analysis, resulting in a long TAT. Our investment in uniform PCR conditions allows automation of PCR set-up for all samples received, no matter which analysis is requested. In our laboratory, the orders received during 5 working days are target of PCR amplification the next week and are sequenced in a MiSeq run 2 weeks after arrival. Our TAT is determined by the frequency of the sequencing runs, currently we are performing MiSeq runs weekly for 6,000–15,000 amplicons. When adding up the time needed from DNA extraction up to data analysis, confirmation of pathogenic mutations with Sanger sequencing and reporting (Fig. 1), each sample can be analyzed within four weeks.

Implementation of quality controls to evaluate the efficiency of all important steps in the whole process, allows us to interfere whenever a step does not fulfill all the requirements and hence to guarantee the TAT. A few examples are listed in the next paragraph:

- A fixed positive control sample (10 amplicons containing a known variant from 10 different genes are pooled together) is analyzed on each run to quickly evaluate if data analysis is done with the use of correct settings.
- The quantification of the input material of our library preparation is crucial, so the known concentration of a commercial human gDNA sample is measured in parallel with our unknown samples. This "standard" gDNA follows the workflow prior to the adapter ligation, as it serves as quality check of the tagmentation reaction as well.

Furthermore, we roughly calculated the minimal number of clusters (400 K/mm^2 with more than 90% CPF) that needs to be sequenced to obtain at least 36× coverage for 95% of the amplicons in an average run. The estimated number of clusters is calculated by the instrument in cycle 25 (approximately 1.5 hr after initiation), which allows us to restart or reload the run immediately when the minimum threshold is not met, instead of after 30 hr (2 × 150) or 40 hr (2 × 250) sequencing time.

Sensitivity

Prior to implementation in diagnostics, a thorough validation of the data analysis pipeline was performed using 387 unique variants. These variants were previously identified in 387 different patients with other mutation detection techniques (262 substitutions, 71 deletions with length <5 bp, 3 deletions affecting 5–10 bp, 28 insertions of <5 nucleotides, 3 insertions of 5–10 nucleotides, 8 deletions and 4 insertions affecting >10 bp, 10 indels). The complete list of the variants evaluated can be found in Supp. Table S1b. For these samples, only the amplicon containing the variant was analyzed using NGS and Sanger sequencing in parallel to rule out sample switches. The presence of these variants was evaluated with different data analysis settings to define the settings with maximal sensitivity. Detection of deletions or insertions is more challenging since NGS analysis software is based on mapping of reads on a reference sequence. Hence, reads lacking one or more nucleotides or containing an insertion will complicate this process. Another important parameter is the minimum allele frequency for a variant to be considered as a true variant. If this parameter is set too low, one will have a very sensitive but less specific test and vice versa. Although only six of the detected variants had an allele frequency of <25%, we finally chose to filter variants with a minimum allele frequency threshold of 15%. These setting were chosen because after duplicate read removal and implementing a quality filter, these six variants (*FBN1* c.1155_1166del, *GPR143* c.222_223insG, *COL2A1*

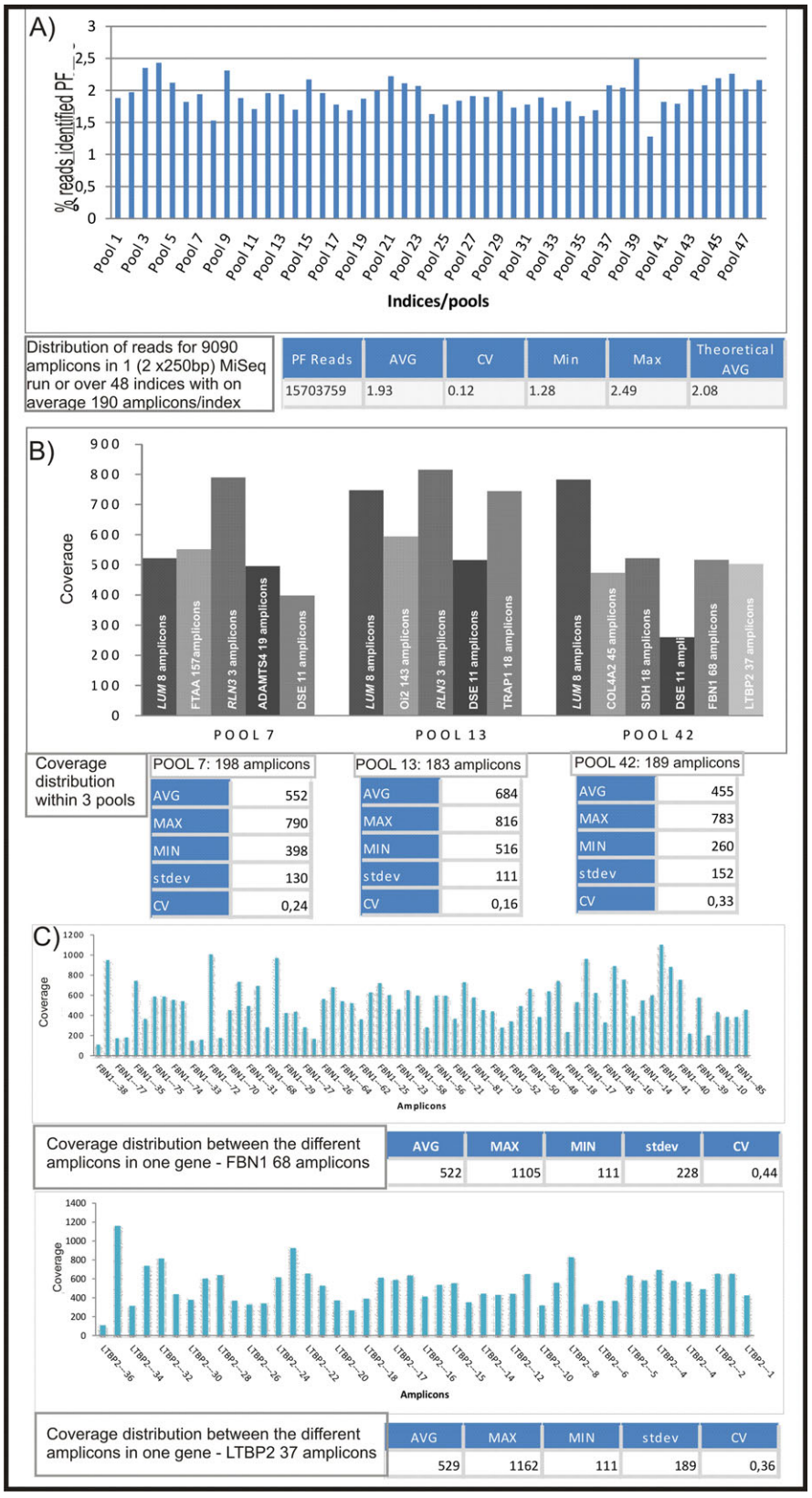

Figure 2. Distribution of coverage within the different pooling steps. **A**: The percentage of reads distributed over all the different indices used in a run. The reads are almost perfectly distributed over the different pools/indices due to an exact measurement of the molarity with the help of qPCR. **B**: The distribution of coverage within three pools is shown. Each bar represents the coverage of a patient specific pool for a certain gene. The variation in coverage is increased compared with panel **A**, since a raw estimation of the molarity by capillary electrophoresis is used to pool. **C**: This panel shows two examples of the coverage of each amplicon within one patient, compared with panels **A** and **B**, the variation in coverage is largest because there is no correction to obtain equimolar pools of the different amplicons.

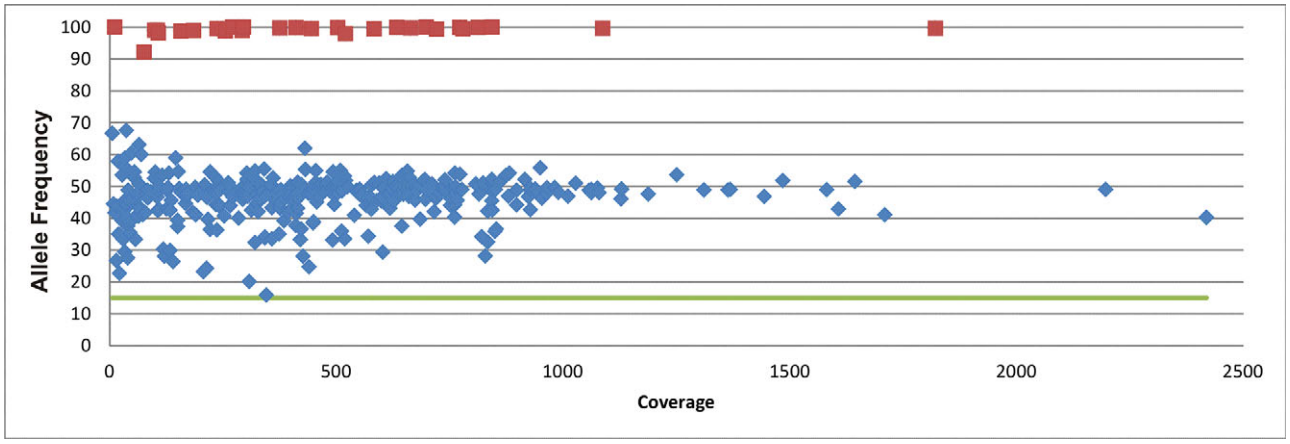

Figure 3. Allele frequency of all positive controls evaluated. The allele frequencies of all positive controls evaluated are shown. The bar depicts the 15% threshold. Variants are called heterozygous in an allele frequency range of 15%–80%. Variants with an allele frequency of 80% and higher are called homozygous.

c.3357_3358insCT, *COL2A1* c.1189_1190insCTCCTGGGT, *BRCA1* c.1009_1010insA, and *BEST1* c.173_174insCA) were detected with an allele frequency between 16% and 25%. We are confident that variants with allele frequency values lower than 15% are sequencing errors or PCR artifacts. The allele frequencies of all variants evaluated are depicted in Figure 3.

Four (1%) of the tested variants could not be detected with our settings. These variants were all deletions, insertions, or combined indels larger than 10 bp in complex repetitive sequence regions with high GC-contents (Fig. 4). All other variants ($n = 16$) spanning more than 10 bp could be detected with the settings applied.

With our settings, the measured sensitivity true positives (TP)/(TP + false negatives [FN]) of this validation set is 99%. By complementing our workflow with (capillary) electrophoresis where the presence of multiple bands of individual PCR products is evaluated for a limited set of genes, we increased the measured sensitivity up to 100%.

Specificity: Positive Predictive Value

To study the specificity, the variants detected with NGS in 61 patients ($n = 298$) in eight different genes (16 patients for *NOTCH3*, 11 patients for *FBN1*, four patients for *GUCY2D*, five patients for *AIPL1*, four patients for *RPE65*, nine patients for *CRB1*, and six patients for *ABCA4* and *USH2A*) were investigated with Sanger sequencing. Seven out of 298 variants detected with a coverage >36× could not be confirmed with Sanger sequencing. Those seven false positives (FP) are representing three different variants (3× RPE65 c.747_748insG, 3× RPE65 c.749T>C, and 1× CRB1 c.3511_3512insT). To calculate specificity, we used the fraction of TP among all positives TP/(TP + FP), also known as positive predictive value (PPV), instead of the standard specificity: true negatives [TN]/(TN + FP). As all nucleotide positions coinciding with the reference sequence will be true negatives, the number of TN will be much larger than the number of FP. In this situation, the standard specificity will always be close to one, as the ratio will be dominated by TN, whereas the PPV will give a more informative value [Mook et al., 2013; Wang et al., 2013]. A PPV of 98% was obtained. In our laboratory, a variant can be named a recurrent FP and be filtered out of the data when it is detected in at least three sequencing runs in more than 75% of the patients analyzed for this region.

In practice, recurrent FP variants are recognized this way and filtered out, which increases the PPV to >99%.

Discussion

An increasing need for genetic testing of Mendelian disorders combined with shorter TATs required a shift in the sequencing methodology in our center. Therefore, we evaluated if genetic testing using NGS could be implemented in an ISO15189-accredited clinical diagnostic setting.

Three main NGS workflows are currently possible for mutation detection in human genetics: targeted sequencing of specific disease genes [e.g., Simpson et al., 2011; Lin et al., 2012; Tang et al., 2012; Beltran et al., 2013; Nikiforova et al., 2013], WES [Yang et al., 2013], and WGS [Gilissen et al., 2014]. Genome wide tests are very useful in patients without a clear clinical diagnosis or for patients who have negative test results for genes known to be associated with the disorder, or for genetically heterogeneous disorders (>100 genes). Consequently, the power of WES and WGS for the discovery of new disease genes or loci is unseen and has found applications in many unsolved cases [Choi et al., 2012; Corton et al., 2013; Nishiguchi et al., 2013; Zheng et al., 2013]. There are some drawbacks however, such as the relatively high cost for sequencing, storing, analysis and interpreting data compared to targeted sequencing, often making it not the ideal strategy in routine diagnostics. Furthermore, WES and WGS increase the chance of incidental results for which innovative approaches and tools need to be developed that enable individuals to exercise their choice about the return of incidental results and which need to take into consideration the predicted increase in workload caused by reporting incidental findings [Ding et al., 2014; Tabor et al., 2014].

For targeted resequencing three different sequence enrichment methodologies are available [Mertes et al., 2011]: PCR-based (Access Array by Fluidigm, Directseq by Raindance Technologies, Ampliseq by Life Technologies, MaSTR [optimized multiplex assays] by Multiplicom), hybridization-based (Sureselect by Agilent and Nimblegen by Roche), and hybridization-extension based (Haloplex by Agilent and TruSight by Illumina). Each of these technologies or methods has its strengths and limitations. Unfortunately at the moment there is no sequencing enrichment strategy available which is perfect for all diseases studied in routine diagnostics. Therefore,

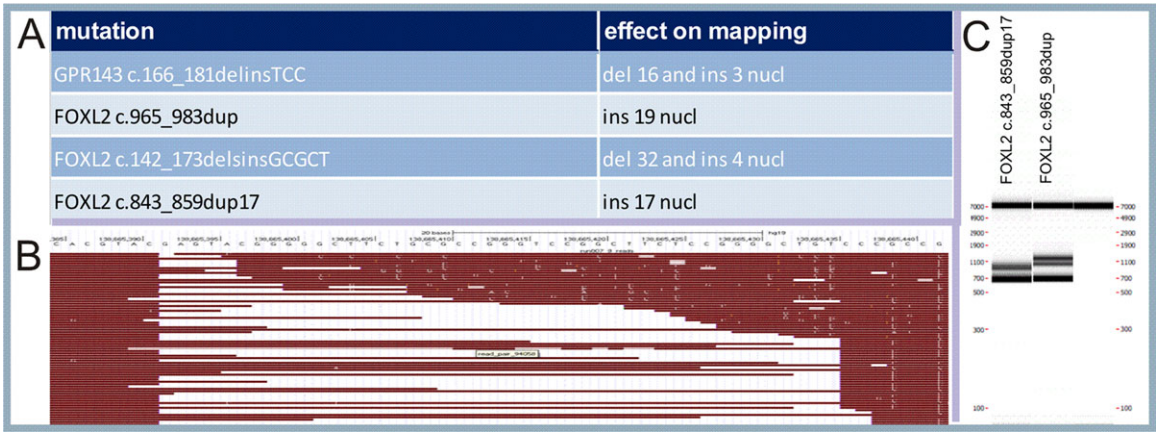

Figure 4. False negatives. **A**: Overview of the four missed variants with NGS. **B**: A detailed mapping of some reads in the neighborhood of a heterozygous *FOXL2* c.142_173delinsGCGCT variant. The variant should be visible at the following coordinates 138665392_138665423, but the mapping fails to align the reads correctly to this area. **C**: The variants *FOXL2* c.965_983dup and c.843_859dup17 are shown and clearly visible on capillary electrophoresis.

laboratories may choose the best suited method and choices may be guided by locus heterogeneity of the disease or other parameters like costs, labor intensity of the library preps, and so on. Our approach is most suitable for small gene panels (e.g., up to 10 genes). PCR-based methods are generally accepted to have the best overall sensitivity and specificity but are limited in target size mainly due to the primer cost [Mamanova et al., 2010]. However, in a diagnostic setting primers can be re-used, making it cost efficient. In addition, costs can be further decreased by performing the PCR reactions in small volumes. Further developments for working with nanoliter volumes are to be expected [De Wilde et al., 2014].

In our center, 265 disease genes associated with Mendelian disorders are currently tested in routine diagnostics. We chose to use singleplex PCRs for target enrichment. We strongly invested in uniform reaction conditions for all PCRs instead of optimizing complicated multiplex protocols. In this way, the highest level of flexibility is retained but upgraded to high throughput and the PCR set-up could be fully automated. We use a combination of a LIMS program and own scripts to generate pipetting files for automated liquid handlers which allowed us to automate such a variable work flow. However, in case of technical issues with the robots, every step can still be performed manually but with a lower throughput. It can be challenging to achieve uniform reaction conditions for thousands of amplicons, as PCR specificity depends on primer design and reaction optimization. Although, by applying the same requirements for each primer pair in primerXL and optimizing PCR conditions for, we found that nonspecific priming in our setting is minimal.

A major challenge of PCR-based target enrichment is the possible presence of variants in primer annealing sites, which may decrease priming efficiency and cause allelic bias and drop-out, the presence of known SNPs in the primers with a minor allele frequency > 0.001 was evaluated and avoided in each primer. But allelic drop out may still be caused by rare variants. As for all PCR-based mutation screening methods, multi-exon deletions or duplications are not detectable, therefore the approach should be complemented with for example MLPA, qPCR or targeted array analysis for the disorders in which these types of variants play an important role.

We strongly invested in an optimal design of the primers allowing uniform reaction conditions, however, diagnostic laboratories may prefer to use the primers previously validated for Sanger sequencing. Especially for the implementation of more complex genes because of the presence of pseudogenes, the use of previously validated primers can be critical. Different target enrichment strategies (e.g., PCR on cDNA, long range PCR) can be used, but as long as it is a PCR based strategy, our unique workflow allows us to bring the amplification products together at the start of the library preparation. The only condition that needs to be taken into account is that fragment lengths of >300 bp are recommended when using the NexteraXT library prep to maximize recovery of smaller fragments out of the AMPureXP cleanup and to ensure even coverage across the length of the DNA fragment.

Throughput of NGS workflows is dependent on the number of amplicons in a single experiment. The uniformity of coverage for the proposed workflow is shown in Figure 2 and found to be sufficient, although out of three major pooling steps, only in the last one an exact correction for equimolarity is performed. This approach was preferred since it allowed us to decrease hands-on time during the library preparation. We are aware about the fact that equimolarity or uniformity in coverage between all amplicons can be improved. However, this may be less cost efficient if the hands-on time for additional quantification and equimolar pooling is taken into account. Now, the cost per amplicon is highly dependent on the number of samples pooled in one run, even with a minimum of 6,000 samples per run, it remains cost efficient to use NGS instead of Sanger sequencing. At the moment, we achieve more than a tenfold reduction of our sequencing cost compared to Sanger sequencing when 10,000 amplicons are sequenced in one MiSeq experiment.

Turnaround times are dependent on the frequency of the MiSeq runs and can be determined by the laboratory in function of the number and type of samples to be analyzed. The proposed workflow allows sequencing complete gene (panels) and single exon analyses within a single run. As for different patients, the same index can be used as long as another target region needs to be sequenced, the combinations are endless and allow absolute flexibility, required for efficient organization of a molecular diagnostic laboratory.

A high sensitivity and specificity was established for our data-analysis pipeline (99% and 98%, respectively). A lower sensitivity was observed for the identification of indels >10 bp, although we believe that it is not the length *per se* which plays a crucial role,

but rather the sequence environment, as we could detect 16/20 larger deletions and insertions. Evaluating these FN into more detail revealed mapping issues because of the similarity of the surrounding repeats rather than an issue with the variant calling. This is in agreement with other studies in which it is believed that medium-sized deletions and insertions should still be detectable with short read sequencing [Lee et al., 2009]. We evaluated all previously detected variants in our pool of 265 genes in our center and the prevalence is very low. By complementing our workflow with capillary electrophoresis, which easily distinguishes DNA fragments starting from 3 bp difference, we feel confident about the sensitivity of our system.

By the application of a detection threshold of 15% allele frequency, some FPs are retained. The bulk of FP was found at the end of the reads, which is a known shortcoming of the sequencing by synthesis chemistry. For Illumina platforms, the average raw error rate is typically 1% or less [Kinde et al., 2011]. The error rate increases toward the end of the reads due to accumulation of asynchrony in the synthesis process [Minoche et al., 2011]. Substitution errors are more frequent than indel errors and certain substitutions (A>C) are more prevalent. Recurrent FPs are artifacts incorporated at stereotypic locations, rather than polymerase errors that occur at random locations and are not reproducible across samples. Recurrent FP artifacts have also been observed by others using the RainDance DeepSeq platform where target enrichment is performed by microdroplet emulsion PCR, prior to sequencing on a MiSeq [Cheng et al., 2014]. At the time of writing, it is not clear if this also observed for other amplicon sequencing platforms, such as the Illumina TruSeq Amplicon assay of IonTorrent Ampliseq system. Because of the recurrence, these artifacts can easily be filtered out of the data.

We present here a flexible, fast, and affordable NGS-based workflow for targeted variant detection on a benchtop sequencer in a clinical setting, with similar sensitivity as the gold standard Sanger sequencing. A major asset of the uniform PCR-based enrichment is that new targets representing novel disease genes of interest can be added and validated in a flexible way. Another advantage is that this workflow can be easily adopted by clinical diagnostic laboratories equipped with a benchtop sequencer. By replacing Sanger sequencing by a home-made flexible singleplex PCR-based NGS workflow, we were able to reduce our sequencing costs per amplicon over ten fold. ISO15189 accreditation was obtained for the complete workflow. Finally, our platform which was optimized for constitutional mutation detection of monogenic disease genes, but is now also used as a model for mutation detection of acquired diseases at the somatic level.

Acknowledgments

We wish to thank the technical support of the CMGG technicians involved in molecular diagnostics.

Disclosure statement: The authors declare no conflict of interest.

References

Arvai K, Horvath P, Balla B, Tokes AM, Tobias B, Takacs I, Nagy Z, Lakatos P, Kosa JP. 2014. Rapid and cost effective screening of breast and ovarian cancer genes using novel sequence capture method in clinical samples. Fam Cancer 13:583–589.

Beltran H, Yelensky R, Frampton GM, Park K, Downing SR, MacDonald TY, Jarosz M, Lipson D, Tagawa ST, Nanus DM, Stephens PJ, Mosquera JM, et al. 2013. Targeted next-generation sequencing of advanced prostate cancer identifies potential therapeutic targets and disease heterogeneity. Eur Urol 63:920–926.

Brownstein Z, Abu-Rayyan A, Karfunkel-Doron D, Sirigu S, Davidov B, Shohat M, Frydman M, Houdusse A, Kanaan M, Avraham KB. 2014. Novel myosin mutations for hereditary hearing loss revealed by targeted genomic capture and massively parallel sequencing. Eur J Hum Genet 22:768–775.

Cheng DT, Cheng J, Mitchell TN, Syed A, Zehir A, Mensah NY, Oultache A, Nafa K, Levine RL, Arcila ME, Berger MF, Hedvat CV. 2014. Detection of mutations in myeloid malignancies through paired-sample analysis of microdroplet-PCR deep sequencing data. J Mol Diagn 16:504–518.

Choi BO, Koo SK, Park MH, Rhee H, Yang SJ, Choi KG, Jung SC, Kim HS, Hyun YS, Nakhro K, Lee HJ, Woo HM, et al. 2012. Exome sequencing is an efficient tool for genetic screening of Charcot-Marie-Tooth disease. Hum Mutat 33:1610–1615.

Clark MJ, Chen R, Lam HY, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M. 2011. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29:908–914.

Corton M, Nishiguchi KM, Avila-Fernandez A, Nikopoulos K, Riveiro-Alvarez R, Tatu SD, Ayuso C, Rivolta C. 2013. Exome sequencing of index patients with retinal dystrophies as a tool for molecular diagnosis. PLoS One 8:e65574.

De Leeneer K, De Schrijver J, Clement L, Baetens M, Lefever S, De Keulenaer S, Van Criekinge W, Deforce D, Van Nieuwerburgh F, Bekaert S, Pattyn F, De Wilde B et al. 2011. Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. PLoS One 6:e25531.

De Wilde B, Lefever S, Dong W, Dunne J, Husain S, Derveaux S, Hellemans J, Vandesompele J. 2014. Target enrichment using parallel nanoliter quantitative PCR amplification. BMC Genomics 15:184.

Ding LE, Burnett L, Chesher D. 2014. The impact of reporting incidental findings from exome and whole-genome sequencing: predicted frequencies based on modeling. Genet Med. (doi: 10.1038/gim.2014.94)

Gilissen C, Hehir-Kwa JY, Thung DT, van de Vorst M, van Bon BW, Willemsen MH, Kwint M, Janssen IM, Hoischen A, Schenck A, Leach R, Klein R, et al. 2014. Genome sequencing identifies major causes of severe intellectual disability. Nature 511:344–347.

Jin X, Qu LH, Meng XH, Xu HW, Yin ZQ. 2014. Detecting genetic variations in hereditary retinal dystrophies with next-generation sequencing technology. Mol Vis 20:553–560.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA 108:9530–9535.

Knies K, Schuster B, Ameziane N, Rooimans M, Bettecken T, de Winter J, Schindler D. 2012. Genotyping of fanconi anemia patients by whole exome sequencing: advantages and challenges. PLoS One 7:e52648.

Lee S, Hormozdiari F, Alkan C, Brudno M. 2009. MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. Nat Methods 6:473–474.

Lin X, Tang W, Ahmad S, Lu J, Colby CC, Zhu J, Yu Q. 2012. Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. Hear Res 288:67–76.

Lupski JR, Gonzaga-Jauregui C, Yang Y, Bainbridge MN, Jhangiani S, Buhay CJ, Kovar CL, Wang M, Hawes AC, Reid JG, Eng C, Muzny DM, et al. 2013. Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. Genome Med 5:57.

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 7:111–118.

Martinez DA, Nelson MA. 2010. The next generation becomes the now generation. PLoS Genet 6:e1000906.

Mertes F, Elsharawy A, Sauer S, van Helvoort JM, van der Zaag PJ, Franke A, Nilsson M, Lehrach H, Brookes AJ. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics 10:374–386.

Minoche AE, Dohm JC, Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol 12:R112.

Mook OR, Haagmans MA, Soucy JF, van de Meerakker JB, Baas F, Jakobs ME, Hofman N, Christiaans I, Lekanne Deprez RH, Mannens MM. 2013. Targeted sequence capture and GS-FLX Titanium sequencing of 23 hypertrophic and dilated cardiomyopathy genes: implementation into diagnostics. J Med Genet 50:614–626.

Nikiforova MN, Wald AI, Roy S, Durso MB, Nikiforov YE. 2013. Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. J Clin Endocrinol Metab 98:E1852–E1860.

Nishiguchi KM, Tearle RG, Liu YP, Oh EC, Miyake N, Benaglio P, Harper S, Koskiniemi-Kuendig H, Venturini G, Sharon D, Koenekoop RK, Nakamura M, et al. 2013. Whole genome sequencing in patients with retinitis pigmentosa reveals pathogenic DNA structural changes and NEK2 as a new disease gene. Proc Natl Acad Sci USA 110:16139–16144.

Pugh TJ, Kelly MA, Gowrisankar S, Hynes E, Seidman MA, Baxter SM, Bowser M, Harrison B, Aaron D, Mahanta LM, Lakdawala NK, McDermott G, et al. 2014. The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. Genet Med 16:601–608.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467.

Simpson DA, Clark GR, Alexander S, Silvestri G, Willoughby CE. 2011. Molecular diagnosis for heterogeneous genetic diseases with targeted high-throughput DNA sequencing applied to retinitis pigmentosa. J Med Genet 48:145–151.

Sivakumaran TA, Husami A, Kissell D, Zhang W, Keddache M, Black AP, Tinkle BT, Greinwald JH, Jr., Zhang K. 2013. Performance evaluation of the next-generation sequencing approach for molecular diagnosis of hereditary hearing loss. Otolaryngol Head Neck Surg 148:1007–1016.

Tabor HK, Auer PL, Jamal SM, Chong JX, Yu JH, Gordon AS, Graubert TA, O'Donnell CJ, Rich SS, Nickerson DA, NHLBI Exome Sequencing Project, Bamshad MJ. 2014. Pathogenic variants for Mendelian and complex traits in exomes of 6,517 European and African Americans: Implications for the return of incidental results. Am J Hum Genet 95:183–193.

Tang W, Qian D, Ahmad S, Mattox D, Todd NW, Han H, Huang S, Li Y, Wang Y, Li H, Lin X. 2012. A low-cost exon capture method suitable for large-scale screening of genetic deafness by the massively-parallel sequencing approach. Genet Test Mol Biomarkers 16:536–542.

Wang X, Wang H, Sun V, Tuan HF, Keser V, Wang K, Ren H, Lopez I, Zaneveld JE, Siddiqui S, Bowles S, Khan A, et al. 2013. Comprehensive molecular diagnosis of 179 Leber congenital amaurosis and juvenile retinitis pigmentosa patients by targeted next generation sequencing. J Med Genet 50:674–688.

Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, et al. 2013. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369:1502–1511.

Zheng Z, Geng J, Yao RE, Li C, Ying D, Shen Y, Ying L, Yu Y, Fu Q. 2013. Molecular defects identified by whole exome sequencing in a child with Fanconi anemia. Gene 530:295–300.