Biomolecular Detection and Quantification xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Biomolecular Detection and Quantification

journal homepage: www.elsevier.com/locate/bdq

Research paper

Targeted resequencing and variant validation using pxlence PCR assays

Frauke Coppieters[a,b], Kimberly Verniers[a], Kim De Leeneer[a], Jo Vandesompele[a,b], Steve Lefever[a,b,*]

[a] Center for Medical Genetics, Ghent University, Ghent, Belgium
[b] pxlence, Dendermonde, Belgium

ARTICLE INFO

Article history:
Received 21 June 2015
Received in revised form 14 August 2015
Accepted 6 September 2015
Available online xxx

Keywords:
PCR
Next-generation sequencing
Sanger sequencing
Amplification specificity

ABSTRACT

The advent of next-generation sequencing technologies had a profound impact on molecular diagnostics. PCR is a popular method for target enrichment of disease gene panels. Using our proprietary primer-design pipeline, primerXL, we have created almost one million assays covering over 98% of the human exome. Here we describe the assay specification and both *in silico* and wet-lab validation of a selected set of 2294 assays using both next-generation sequencing and Sanger sequencing. Using a universal PCR protocol without optimization, these assays result in high coverage uniformity and limited non-specific coverage. In addition, data indicates a positive correlation between the predictive *in silico* specificity score and the amount of assay non-specific coverage.

1. Introduction

The Online Mendelian Inheritance in Man (OMIM) database currently contains over 4400 human inherited diseases with a known genetic cause [1]. Over 300 new disease genes are being identified yearly, with novel mutations accumulating at a rate of 10,000 per annum [2]. The establishment of a molecular diagnosis in a family confirms the clinical diagnosis, enables reproductive options, and, more recently, is a prerequisite for gene-specific therapies. Indeed, several ongoing clinical trials for human gene and gene-specific therapy emphasize the advent of personalized genomic medicine [3].

Over the past decade, molecular diagnostic testing has faced an exponential growth due to the replacement of laborious gene-by-gene Sanger sequencing by parallel resequencing of multiple genes with massively parallel or so-called next-generation sequencing (NGS) technologies. Various target enrichment strategies are available, enabling the customer to resequence any region of interest. Molecular diagnostic laboratories often develop customized NGS platforms, offering a specific diagnostic portfolio. In addition to gene centric analyses, both exome and genome sequencing are appealing NGS approaches because of the greatly decreased sequencing cost per base [4]. Thus far, these are not yet routinely used in diagnostics because of data quality and ethical reasons, *i.e.* insufficient coverage for relevant genes and incidental findings, respectively. As most genetic centres are accredited, strict regulations are applicable regarding variant reporting [5,6]. Variants identified through NGS generally require confirmation using either NGS or Sanger sequencing. Based on a survey we did in September 2014, almost 70% of 178 respondents from Europe, USA and Asia indicated they are currently validating their NGS findings using either NGS or Sanger sequencing (Fig. 1A). According to this survey, PCR amplification is the most commonly used target enrichment method, followed by hybridization (Fig. 1B).

So far, the currently available NGS enrichment methods for gene panels are hampered by technical limitations. Capture based enrichment for instance often struggles with GC content or repeat rich regions. On the other hand, major advantages of PCR-based enrichment include high flexibility when using singleplex PCR and cost-effectiveness in case of automation or multiplex PCR [7–9]. Of note, PCR based enrichment is sensitive to allelic dropout and/or lower amplification efficiency caused by single nucleotide polymorphisms (SNP) in primer annealing sites and requires more optimization in case of less efficient PCR assays [10]. We recently developed a primer design pipeline for targeted resequencing PCR assays, called primerXL, tackling these issues [11]. PrimerXL makes use of the third-party software packages primer3 v3.2.2 (primer design), UNAFold v3.8 (secondary structures) and Bowtie v0.12.7 (specificity) and includes optimized settings to maximize target

* Corresponding author at: Center for Medical Genetics, Ghent University, Ghent, Belgium.
E-mail address: steve.lefever@ugent.be (S. Lefever).

http://dx.doi.org/10.1016/j.bdq.2015.09.001

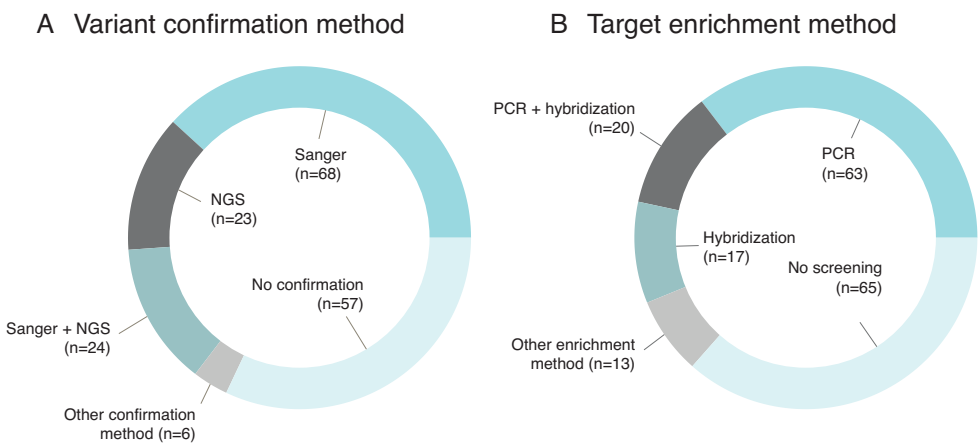

Fig. 1. Distribution of respondents' ($n = 178$) answers on two survey questions, (A) do you validate your NGS findings and if yes, using what method, (B) do you perform screening experiments, if so how do you enrich your targets?

coverage while minimizing off-target amplification [12–14]. Each of the assay designs is put through a series of stringent *in silico* tests to filter out primer pairs harbouring secondary structures or SNPs in their annealing sites, or having non-specific amplification potential (because of sequence homology to off-target regions). Together, this results in robust PCR assays amplifiable under uniform conditions. Using primerXL, two databases containing almost one million pre-designed sets of assays were developed, each covering ~98% of the exonic regions of all human protein coding genes. The main applications of these assays are the development of gene panels and validation of variants located in exonic regions. A selection of the assays has successfully been used in the development of NGS gene panels for congenital blindness, deafness and cancer [7,15,16]. In addition, these assays were used to replace Sanger-based sequencing with NGS for over 200 genes in an ISO15189-accreditated setting by our diagnostics department [8]. The goal of this study was to further determine the applications for these assays and their overall wet-lab success rate.

2. Material and methods

2.1. (Quantitative) PCR

Each reaction was performed in a 10 μL volume: 5 μL Kapa 2G mastermix, 2.5 μL primers (forward and reverse oligos mixed at a 1 μM concentration) and 2.5 μL DNA template (20 ng/μL). For sample 1, an additional 0.5 μL LCGreen Plus (Bioké) was added since this sample was assessed using qPCR. All (q)PCR reactions were run on a Roche LC480 instrument using the following protocol: (1) 180 s at 95 °C, (2) 15 s at 95 °C, (3) 10 s at 60 °C, (4) 15 s at 72 °C, (5) 60 s at 72 °C. Steps 2–4 were repeated 35 times. For sample 1, this was followed by running a melting curve starting at 65 °C up to 95 °C with 0.5 °C temperature increments each 5 s.

2.2. Library prep and sequencing

Following (q)PCR all reactions were pooled (no normalization was performed). Concentration measurement was performed with the Qubit Fluorometer (Life Technologies). A total of 2.5 μg of the pooled PCR product was used as input for the NEBNext DNA Library Prep Master Mix Set for Illumina (New England BioLabs). During each step, 2 μL was retained to assess the quality of the prep by means of a Bioanalyzer analysis. Both samples were sequenced on a single Illumina MiSeq run (2 × 150 cycles).

3. Results

3.1. Assay specifications

Two assay databases were created with different applications in mind. The first catalogue, with ~320,000 assays having amplicon lengths between 350–750 bp (with 65.2% between 350–450), is optimized for high-quality DNA samples. The second catalogue, suited for fragmented DNA (*e.g.* derived from FFPE samples), contains almost 550,000 short assays (amplicon lengths: 125–275 bp). The latter assays are also ideal candidates for multiplex PCR because of their uniform amplicon lengths. Both databases have been generated to cover all exons of all Ensembl canonical transcripts (Ensembl build 63). The exome coverage for the long and short dataset is 97.99% and 98.71%, respectively. Since then, the *in silico* SNP and specificity analysis was reassessed for the longest amplicons using a more recent genome build (Ensembl build 78). In 94.01% of these assays, no SNPs are present in the primer annealing sites. For the remaining assays, 85.74% contain SNP(s) outside the critical 5 bp 3′ region, whereas 92.47% contain only a single SNP. The *in silico* specificity analysis determined the likelihood of non-specific product generation for each assay. This was done by Bowtie-based alignment of an assay to the human genome (hg38), allowing up to 3 mismatches per primer (3 or more mismatches significantly impede the amplification process), and assigning each assay an *in silico* specificity score equal to the minimal number of mismatches across all predicted off-target hits [10,12]. The higher the resulting specificity level, the more specific the assays are predicted to be. A specificity level of 7 means that there are no non-specific hits, whereas a specificity level of 5 for *e.g.* refers to predicted off-target hits with three mismatches in one primer and two mismatches in the other primer. This analysis revealed that the majority (73.1%) of assays attain the most stringent specificity level (*i.e.* level 7) (Fig. 2). All assays are linked to their specificity level and SNP information, which is displayed to the user upon querying the database.

3.2. Wet-lab assay validation

From the 350–750 bp dataset, 2294 assays covering 169 diagnostically relevant diseases were randomly selected. Using these assays, singleplex amplification was performed on two pooled samples containing male and female DNA.

To assess assay performance and end-point equimolarity, quantitative PCR was used for sample 1, while classical PCR was performed for sample 2. Following amplification using a universal protocol (KAPA 2G Robust–spiked with LCGreen Plus for sample

Please cite this article in press as: F. Coppieters, et al., Targeted resequencing and variant validation using pxlence PCR assays, Biomol Detect Quantif (2015), http://dx.doi.org/10.1016/j.bdq.2015.09.001

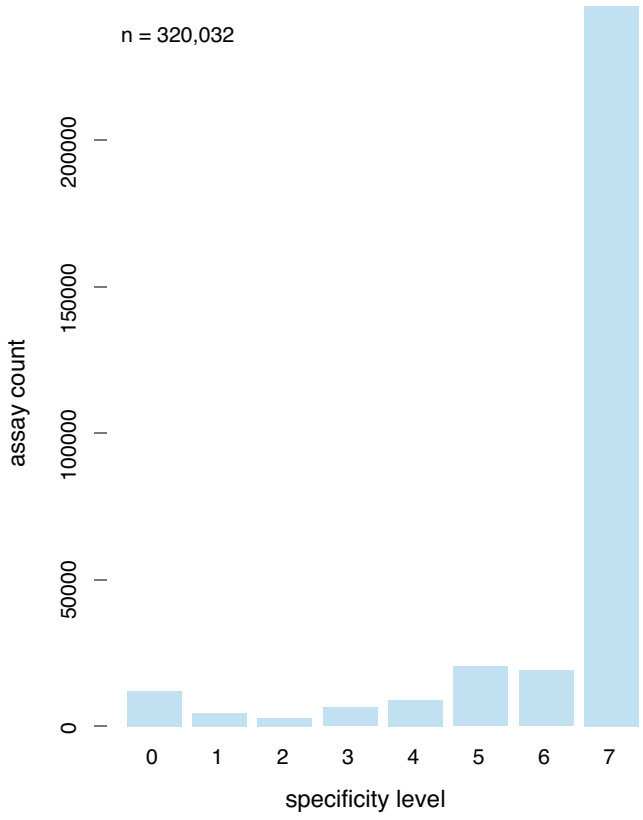

Fig. 2. Distribution of the number of assays in the 350–750 bp assay set in function of the *in silico* specificity score.

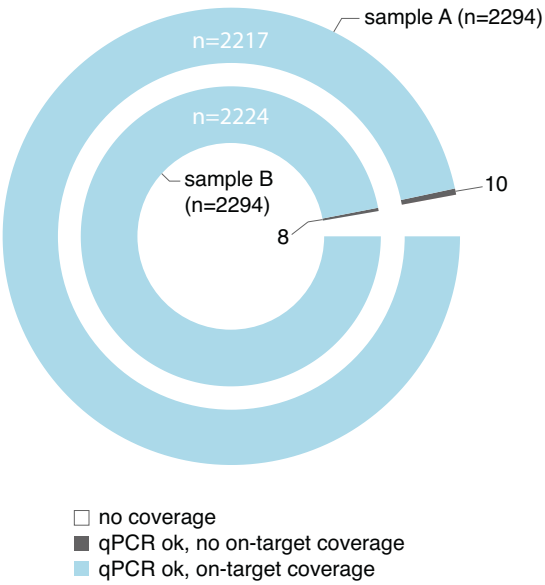

Fig. 3. Overview of the coverage results for the 2294 assays in both samples.

1), products were pooled per sample and prepped for sequencing using the NEBNext DNA Library Prep Master Mix Set for Illumina (New England BioLabs). Both samples were sequenced on a single Illumina MiSeq run (2 × 150 cycles) generating 3,819,421 and 3,747,843 reads, respectively. A custom mapping strategy was applied to determine the specificity level of each assay. To assign reads to their corresponding assay, the first 50 nucleotides of the first read-pair were concatenated with the reverse complement of the first 50 nucleotides of the second read-pair, thus generating a fasta file. This file was used as a reference sequence against which each assay was mapped in paired-mode using Bowtie allowing up to three mismatches and a 100 bp maximum product size. For each sequencing read, the read/assay combination having the least number of mapping mismatches was then considered the correct one. In this way, a total of 3,457,533 (sample A, 90.5%) and 3,319,688 (sample B, 88.6%) reads could be assigned to the various assays. A summary of (q)PCR and sequencing results for both samples can be seen in Fig. 3.

Overall, coverage uniformity per assay was high, with 60% and 58.8% of the assay having a coverage within 2-fold of the mean for sample A (mean = 1551) and B (mean = 1487) respectively (83.3% and 82.6% within 5-fold of the mean) (Fig. 4). No correlation between read-depth and qPCR end-point fluorescence values for sample 1 was observed (data not shown). Of note, the end-point fluorescence values showed highly uniform product equimolarity (95.95% of the assays have an end-point fluorescence value within 2-fold of the mean), obviating the need for time-consuming product normalization prior to sequencing. Coverage specificity per assay was determined by calculating the ratio between the number of associated (on-target) reads overlapping the genomic coordinates of the assay and all reads (both on- and off-target) linked to that assay. Results indicated that less than 12.0% and 12.8% of the assays, for sample A and B respectively, had more than 2% of its associated reads mapping to off-target regions (Fig. 5). As mentioned earlier, each assay is assigned an *in silico* specificity score. For this experiment, the *in silico* assay specificity score was determined by means of a Bowtie alignment with a maximum hit size of 1500 bp. For each score, the percentage of assays with more than 2% off-target was calculated. Fig. 6 shows there is a correlation between the percentage of assays having a higher degree of non-specificity and the *in silico* specificity score, confirming the predictive value of the latter with respect to off-target sequencing coverage.

In addition to NGS, 1900 out of 2294 pxlence assays were also subjected to Sanger sequencing using a universal amplification and sequencing protocol. This was performed in a ISO15189 accredited lab in the context of comparing and replacing Sanger-based diagnostic tests with NGS [8]. In practice, Sanger sequencing was only performed for assays passing LabChip GX (PerkinElmer) assessment. The overall success rate of the pxlence assays on LabChip GX was 95.77%. Subsequent Sanger sequencing of 1900 pxlence assays revealed a Sanger success rate of 88.63%. The most common observations in case of failed Sanger sequencing where (1) completely failed Sanger traces possibly due to the amplicon sequence contents, (2) the presence of pseudogenes and (3) the presence of homopolymeric regions close (downstream) to the primers, causing uninterpretable traces early in the sequence.

4. Discussion

We have generated pre-designed assays covering over 98% of the human exome using an in-house developed primer-design pipeline called primerXL. All assays have thoroughly been tested *in silico* resulting in robust assay performance while minimizing potential aspecific amplification. An extensive singleplex wet-lab qPCR amplification experiment showed that the majority (94.73% have more than 20× on-target coverage) of these off-the-shelf assays work well without any need for optimization, reducing the time required to enrich targets. Although not tested in this study, we anticipate that the short assays could be good candidates for multiplex PCR because of their uniform amplicon lengths and primer properties. Sequencing coverage uniformity is high, with limited sequencing drop-out (3.71% have less than 20× total coverage) while non-specific coverage is kept to a minimum. This, together with the results of our small-scale market study, indicate that these assays, available for both normal quality as well as fragmented

Please cite this article in press as: F. Coppieters, et al., Targeted resequencing and variant validation using pxlence PCR assays, Biomol Detect Quantif (2015), http://dx.doi.org/10.1016/j.bdq.2015.09.001

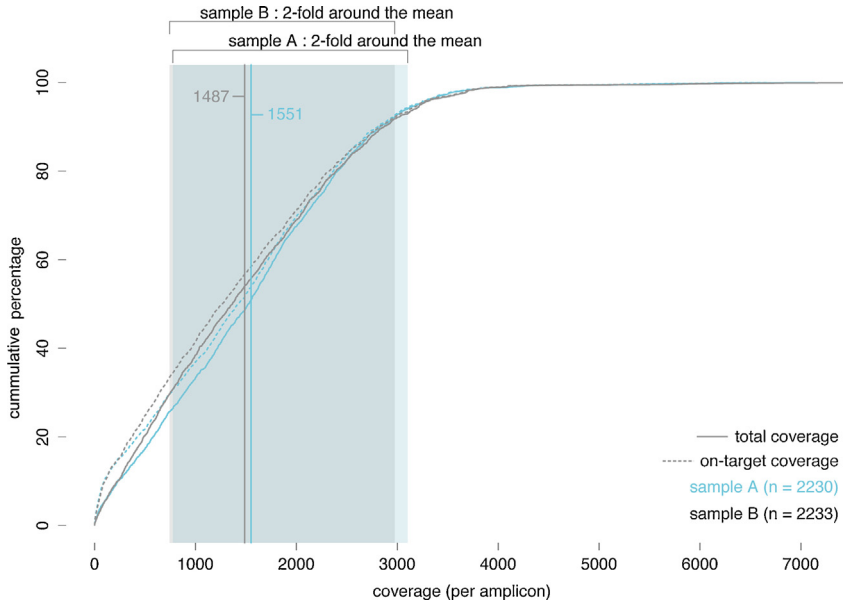

Fig. 4. Cummulative distribution of the coverage per amplicon, both total coverage and on-target coverage. Areas show the 2-fold region around the mean for each sample.

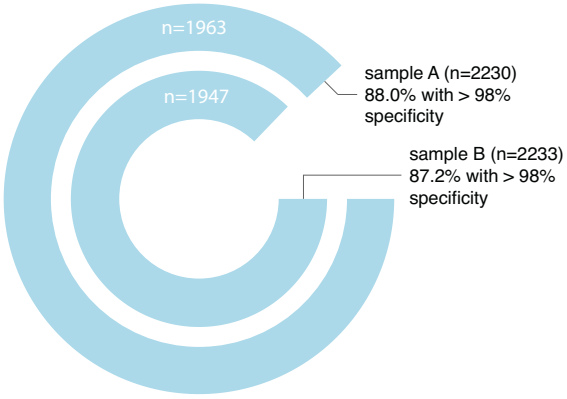

Fig. 5. Percentage of assays with more than 98% of the reads mapping to on-target regions.

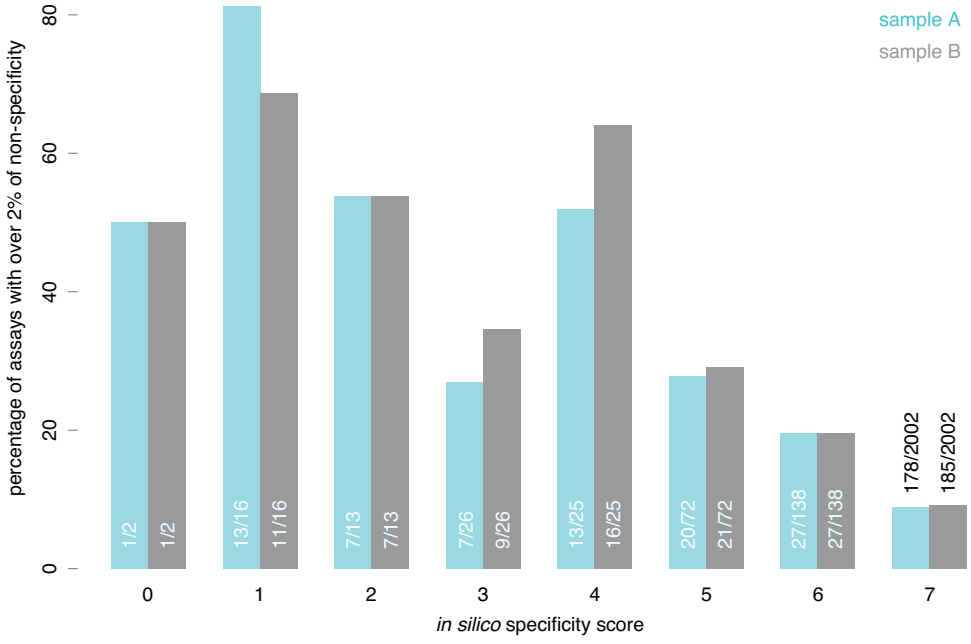

Fig. 6. Percentage of assays within each *in silico* specificity score category showing more than 2% non-specific coverage. Numbers (x/y) in the bars indicate the number of assays for a certain *in silico* specificity score with more than 2% non-specificity (x) across all assays having that *in silico* specificity score (y).

Please cite this article in press as: F. Coppieters, et al., Targeted resequencing and variant validation using pxlence PCR assays, Biomol Detect Quantif (2015), http://dx.doi.org/10.1016/j.bdq.2015.09.001

DNA, could be of particular interest to the research and diagnostic community. As the major aim of this study was to get a global assessment of the quality of the pxlence PCR assays, no efforts were made to obtain 100% sequence coverage for all regions, which is required for diagnostic purposes. However, a subset of these assays are being used in a diagnostic context for 265 different genes [8]. In this set-up, 100% coverage was obtained by redesigning assays using less stringent design settings [8]. In this context, a Ghent University spin-off company called pxlence, whit the goal to commercialize the assays as well as the primer-design pipeline, was recently founded. All assays are available through a webshop reachable *via* www.pxlence.com.

Conflict of interest

SLF, FCP and JVS are co-founders of pxlence.

Acknowledgments

SL and FC are post-doctoral fellows with the Research Foundation—Flanders (FWO).

References

[1] Online Mendelian Inheritance in Man, www.Omim.org.
[2] D.N. Cooper, J.-M. Chen, E.V. Ball, K. Howells, M. Mort, A.D. Phillips, et al., Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics, Hum. Mutat. 31 (2010) 631–655, http://dx.doi.org/10.1002/humu.21260.
[3] ClinicalTrials.gov.
[4] Y. Yu, B.-L. Wu, J. Wu, Y. Shen, Exome and whole-genome sequencing as clinical tests: a transformative practice in molecular diagnostics, Clin. Chem. 58 (2012) 1507–1509, http://dx.doi.org/10.1373/clinchem.2012.193128.
[5] N. Aziz, Q. Zhao, L. Bry, D.K. Driscoll, B. Funke, J.S. Gibson, et al., College of American Pathologists' laboratory standards for next-generation sequencing clinical tests, Arch. Pathol. Lab. Med. 139 (2015) 481–493, http://dx.doi.org/10.5858/arpa.2014-0250-CP.
[6] M.M. Weiss, B. Van der Zwaag, J.D.H. Jongbloed, M.J. Vogel, H.T. Brüggenwirth, R.H. Lekanne Deprez, et al., Best practice guidelines for the use of next-generation sequencing applications in genome diagnostics: a national collaborative study of Dutch genome diagnostic laboratories, Hum. Mutat. 34 (2013) 1313–1321, http://dx.doi.org/10.1002/humu.22368.
[7] B. De Wilde, S. Lefever, W. Dong, J. Dunne, S. Husain, S. Derveaux, et al., Target enrichment using parallel nanoliter quantitative PCR amplification, BMC Genomics 15 (2014) 184, http://dx.doi.org/10.1186/1471-2164-15-184.
[8] K. De Leeneer, J. Hellemans, W. Steyaert, S. Lefever, I. Vereecke, E. Debals, et al., Flexible, scalable, and efficient targeted resequencing on a benchtop sequencer for variant detection in clinical practice, Hum. Mutat. 36 (2015) 379–387, http://dx.doi.org/10.1002/humu.22739.
[9] A. Ruiz, G. Llort, C. Yagüe, N. Baena, M. Viñas, M. Torra, et al., Genetic testing in hereditary breast and ovarian cancer using massive parallel sequencing, Biomed. Res. Int. 2014 (2014) 542541–542548, http://dx.doi.org/10.1155/2014/542541.
[10] S. Lefever, F. Pattyn, J. Hellemans, J. Vandesompele, Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative PCR assays, Clin. Chem. 59 (2013) 1470–1480, http://dx.doi.org/10.1373/clinchem.2013.203653.
[11] S. Lefever, F. Pattyn, B. De Wilde, F. Coppieters, S. De Keulenaer, J. Hellemans, et al. High-throughput PCR assay design for targeted resequencing using primerXL, (2015) in preparation.
[12] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25, http://dx.doi.org/10.1186/gb-2009-10-3-r25.
[13] N.R. Markham, M. Zuker, UNAFold: software for nucleic acid folding and hybridization, Methods Mol. Biol. 453 (2008) 3–31, http://dx.doi.org/10.1007/978-1-60327-429-6-1.
[14] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, Methods Mol. Biol. 132 (2000) 365–386.
[15] F. Coppieters, B. De Wilde, S. Lefever, E. De Meester, N. De Rocker, C. Van Cauwenbergh, et al., Massively parallel sequencing for early molecular diagnosis in Leber congenital amaurosis, Genet. Med. 14 (2012) 576–585, http://dx.doi.org/10.1038/gim.2011.51.
[16] S. De Keulenaer, J. Hellemans, S. Lefever, J.-P. Renard, J. De Schrijver, H. Van de Voorde, et al., Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform, BMC Med. Genomics 5 (2012) 17, http://dx.doi.org/10.1186/1755-8794-5-17.

Please cite this article in press as: F. Coppieters, et al., Targeted resequencing and variant validation using pxlence PCR assays, Biomol Detect Quantif (2015), http://dx.doi.org/10.1016/j.bdq.2015.09.001