Massive parallel amplicon sequencing of the breast cancer genes BRCA1&2: opportunities, challenges and limitations.

Kim de Leeneer, Jan Hellemans, Joachim de Schrijver, Machteld Baetens, Bruce Poppe, Wim van Criekinge, Anne Depaepe, Paul Coucke, Kathleen B.M. Claes

Human Mutation

Massive parallel amplicon sequencing of the breast cancer genes *BRCA1&2* : opportunities, challenges and limitations.

Journal:	*Human Mutation*
Manuscript ID:	humu-2010-0318.R1
Wiley - Manuscript type:	Methods
Date Submitted by the Author:	29-Oct-2010
Complete List of Authors:	De Leeneer, Kim; Center for Medical Genetics, Ghent University Hospital Hellemans, Jan; Center for Medical Genetics, Ghent University Hospital De Schrijver, Joachim; Laboratory for Bioinformatics and Computational Genomics, Ghent University Baetens, Machteld; Center for Medical Genetics, Ghent University Hospital Poppe, Bruce; Center for Medical Genetics, Ghent University Hospital Van Criekinge, Wim; Laboratory for Bioinformatics and Computational Genomics, Ghent University DePaepe, Anne; Center for Medical Genetics, Ghent University Hospital Coucke, Paul; Center for Medical Genetics, Ghent University Hospital Claes, Kathleen; Ghent University Hospital, Center for Medical Genetics
Key Words:	massive parallel sequencing, BRCA1/2, multiplex, barcoding, amplicon sequencing, breast cancer

Massive parallel amplicon sequencing of the breast cancer genes *BRCA1&2:* opportunities, challenges and limitations.

Kim De Leeneer (1), Jan Hellemans (1), Joachim De Schrijver (2), Machteld Baetens (1), Bruce, Poppe (1), Wim Van Criekinge (2), Anne De Paepe (1), Paul Coucke (1), Kathleen Claes (1)*

1 Center for Medical Genetics, Ghent University Hospital, De Pintelaan 185, B-9000 Gent, Belgium

2 Laboratory for Bioinformatics and Computational Genomics, Ghent University, B- 9000 Gent, Belgium

Corresponding author: Kathleen Claes, Ph.D.
E-mail address: Kathleen.Claes@UGent.be

ABSTRACT

This study describes how the new massive parallel sequencing technology can be implemented in a diagnostic setting by proof of concept studies for the breast cancer susceptibility genes (*BRCA1&2*). Throughput was maximized by increasing uniformity in coverage. This was obtained by a multiplex approach, which outperformed pooling of singleplex PCRs. We evaluated the sensitivity by analysis of 133 distinct sequence variants; 3 (2%) deletions or duplications in homopolymers of ≥7 nucleotides remained undetected, illustrating a limitation of pyrosequencing. Furthermore, other limitations like non random sequencing errors, pseudogene amplification and failure to detect of multi exon deletions are thoroughly described.

Our workflow certainly has the potential for high throughput analysis of large genes in diagnostic settings, which is of great importance to meet the increasing expectations of genetic testing. Implementation of this approach will hopefully lead to a strong reduction in turnaround times, so a wider spectrum of at risk women will be able to benefit from therapeutic interventions for which knowledge of the DNA sequence is essential.

Key Words: massive parallel sequencing, BRCA1/2, multiplex, barcoding, amplicon sequencing

INTRODUCTION

The development of high throughput or massive parallel sequencing (MPS) has opened many new research opportunities. Several platforms are released, of which the three major are the Genome Sequencer from Roche/454 Life Sciences, the Genome Analyzer from Illumina/Solexa and the SOLiD System from Applied Biosystems. These platforms differ in several ways, such as the technology applied, read length and the number of DNA molecules sequenced.

In the case of 454 sequencing, single DNA strands with 5' and 3' adaptor sequences are attached to beads and then clonally amplified by PCR in an oil-water emulsion. The beads are mixed with DNA polymerase and deposited in plates containing more than 1 million wells, with one bead per well. Nucleotides then flow sequentially over the wells and as each nucleotide is added to form complementary DNA strands, pyrophosphate is released and detected in a chemiluminescent flash (pyrosequencing chemistry).

Since the launch of MPS, an increasing number of applications have been published, amongst others *de novo* sequencing (Pearson, et al., 2007; Pol, et al., 2007; Velasco, et al., 2007), whole genome re-sequencing (Albert, et al., 2007; Korbel, et al., 2007), amplicon sequencing (e.g. exon re-sequencing, virus variant detection, DNA methylation) (Dahl, et al., 2007; Korshunova, et al., 2008; Pettersson, et al., 2008; Taylor, et al., 2007; Thomas, et al., 2006) , miRNA and splice variant discovery (Ruby, et al., 2006; Yao, et al., 2007). However, the implementation of this high throughput sequencing in molecular diagnostics remains largely unexplored. Over the last decades, Sanger sequencing (Sanger, et al., 1977) has been the dominant DNA sequencing technology and the "gold standard" for DNA-based mutation

detection. However, due to cost limitations direct sequencing of large genes in diagnostics, is often preceded by a mutation scanning technique, followed by characterization of the variant with Sanger sequencing. Denaturing gradient gel electrophoresis (DGGE) (van der Hout, et al., 2006), denaturing High-performance Liquid chromatography (dHPLC) (Liu, et al., 1998) and High Resolution Melting Curve Analysis (HRMCA) (De Leeneer, et al., 2008; De Leeneer, et al., 2009) are well known examples. For these pre-screening techniques, sensitivities varying from 50-100% and specificities close to 100% were reported (Gerhardus, et al., 2007). Some of these methods are laborious and cannot be fully automated since a sequencing step is required to define the nature of the variant.

We chose the BReast CAncer susceptibility genes *BRCA1* (MIM +113705) and *BRCA2* (MIM +600185) to optimize high throughput amplicon sequencing, because of their size, polymorphic character and lack of mutation hot spots. These genes require high throughput screening as an increasing number of samples need to be tested within shorter turnaround times. With the approval of targeted therapeutic agents like PARP inhibitors with selective toxicity for tumors derived from germline carriers of mutations in *BRCA1&2,* expectations for genetic testing keep increasing (Curtin, 2005). To obtain a cost efficient MPS strategy, an easy set-up, and uniform distribution of coverage are required to maximize throughput. We explored the challenges of optimizing the clinical use of MPS with two different approaches and compared sensitivity and specificity of MPS with Sanger sequencing and prescreening techniques currently used in molecular diagnostics. For our studies we chose the Roche Genome Sequencer FLX (GS FLX) system because of the longer read lengths generated by this instrument compared to the other 2 major platforms.

MATERIAL AND METHODS

DNA samples and sequence variants evaluated

In total 123 DNA samples isolated from blood were evaluated. To evaluate the specificity of our MPS set-up genomic DNA samples from 30 patients previously analysed for the complete coding region of *BRCA1/2* with other mutation detection techniques were used. Eleven of these patients were completely Sanger sequenced. Nineteen were analysed with High resolution melting curve analysis (HRMCA) followed by Sanger sequencing of all aberrant melting curves (De Leeneer, et al., 2008; De Leeneer, et al., 2009). For Sanger sequencing and HRMCA identical primers sets were used as for MPS.

Furthermore, 93 samples with previously characterized (Sanger sequencing) deletion or insertion variants were used as positive control samples to validate the detection capacities of MPS. These samples were only sequenced for the amplicon containing the mutation and other amplicons within the specific multiplex set. An overview of all variants evaluated is shown in Supp. Table 1. In total forty-three 1-3 bp deletions, fourteen 1-3 bp insertions (of which 13 duplications), three combined indels and twenty-two deletions and an insertion larger than 3bp were sequenced. Twenty of these deletions and insertions are located in a homopolymeric region longer than 3bp.

Furthermore, 410 (40 unique) nucleotide substitutions were present in all samples analyzed. The majority (406) were frequent SNPs and 4 were nonsense mutations.

PCR set-up

To cover all coding regions and splice sites, we used primer sets thoroughly validated by HRMCA (De Leeneer, et al., 2008; De Leeneer, et al., 2009) and Sanger sequencing. In run 1&2 we started from equimolar pools of singleplex PCRs, in run 3&4 multiplex PCRs were

generated prior to sequencing. To fuse the amplicon-specific primers with the adaptor-MID (Multiplex Identifiers) barcoded primers, 2 consecutive PCR rounds were used with a universal M13 tail as linker (Hellemans J et al, under review). A schematic representation of the principle and workflow of both approaches is shown in Figure 1.

Run 1 & 2

Per sample 111 singleplex PCRs with amplicon specific primers were performed with identical reaction conditions and primers as published before (De Leeneer, et al., 2008). PCRs were performed on the CFX384 (Bio-Rad) and RFU data (endpoint fluorescence) were used for subsequent normalization to obtain 12 equimolar pools of PCR products per patient.

Run 3 & 4

Sixteen multiplex reactions were optimised to cover the complete coding and splice site regions of *BRCA1* and *BRCA2*, except *BRCA1* exon 2. As this amplicon amplified less efficiently compared to the other PCRs, it was optimized as an individual reaction.

The specifications of each multiplex reaction are given in Supplemental Table 2 & 3. Multiplex PCR was performed in 20µl total volume. For sets 1,2,5,8-15 the amplification mixture included 2X Titanium buffer (Clontech), 3% DMSO (VWR International), 200µM of each dNTP (GE Healthcare), 1x Titanium Taq Polymerase (Clontech) and approximately 100ng of DNA. In 5 reactions (set 3,4,6,7 and 16), 3 mM $MgCl_2$ (Invitrogen), 2X PCR buffer (Invitrogen) and 1.5U Platinum Taq polymerase was used instead of Titanium buffer and polymerase. Two touchdown PCR programs were used (abbreviated as Touch46 and Touch48 in Supp. Table 2). The temperature cycling protocol consists of an initial denaturation step at

94°C for 2 minutes, followed by 12 cycles of denaturation at 94°C for 20 seconds, annealing starting at 60 (58) °C for 20 seconds (decreasing 1°C per cycle) and an extension at 72°C during 1 minute. This initial PCR is followed by 25 additional cycles of denaturation at 94°C for 40 seconds, annealing at 48 (46) °C during 40 seconds and extension at 72°C for 30 seconds. Final extension was accomplished at 72°C for 10 minutes.

Primer concentrations in one multiplex vary between 0.025µM and 0.8µM and were adjusted to obtain equimolar quantities of each amplicon in one reaction (Supp. Table 3).

PCR with MID barcoded primers

MID barcoded primers consist of (Figure 1A): (i) the required sequencing adaptor (A or B), (ii) a 10-nucleotide long MID tag or barcode to identify the patient (MID sequences provided by Roche/454, application note (CRF00104)) (iii) a universal M13-tail (forward primer: cacgacgttgtaaaacgac and reverse primer: caggaaacagctatgacc), identical with the M13 tail used in the first PCR round.

After the initial PCRs (singleplex or multiplex), all samples were diluted 1000 times and 1 µl product was used as a template for a second PCR with MID barcoded primers. Total volume of this reaction was 15µl. The amplification mixture included 1.5 mM $MgCl_2$ (Invitrogen), 1X PCR buffer (Invitrogen) and 1.5U Platinum Taq (Invitrogen), 3% DMSO (VWR International), 200µM of each dNTP (GE healthcare) and 0.2 µM of both forward and reverse primer. Temperature cycling protocol consists out of following steps: 4 minutes at 94°C, 15 cycles of denaturation at 94°C for 30 seconds, annealing at 60°C for 30 seconds, extension at 72°C during 50 seconds, and final extension at 72° for 10 minutes.

PCRs were performed on a CFX384 instrument (Bio-Rad). During optimization FAM labeled MID primers were used to evaluate equimolarity between amplicons within one multiplex reaction and fluorescent peaks were separated on an ABI3730 capillary system.

Sequencing runs and data analysis

PCRs were normalized and equimolarly pooled in relation to the RFU data. This pool was purified on a High Pure PCR Cleanup Micro kit (Roche). Fragment length of this total amplicon pool was evaluated on the Bioanalyzer (Agilent) and compared to the theoretically predicted pattern (Figure 2).

Emulsion PCR and sequencing reactions on the GS-FLX (454- Roche) were performed according to the manufacturer's instructions. The FASTA files were analysed with in house developed variant interpretation pipeline (VIP) software version 1.3 (De Schrijver et al., 2010) and with the commercially available Nextgene software (Softgenetics) version 2.0. Reads were aligned against GenBank reference sequences, NC_000017.10 (41196313-41277467, complement BRCA1) and NC_000013.10 (32889617-32973809;BRCA2) from build 18 of the Human Genome assembly.

RESULTS

We selected crucial characteristics like uniformity of coverage, sensitivity, specificity and throughput, to evaluate the utility of MPS in diagnostics. Furthermore, MPS should outperform current methods used in molecular diagnostics in terms of cost efficiency to make the implementation worthwhile.

The performance evaluation of the selected approach is based on the results of 4 proof of concept (PoC) experiments (Figure 1). In the first two experiments an identical singleplex approach was used for the analysis of 22 different patient samples.

To reduce workload we started in PoC 3&4 from a multiplex set-up. For PoC4 a strong optimization was performed compared to PoC3: primer concentrations within several multiplex sets were adjusted for amplicons poorly covered in PoC3. Furthermore, the composition of a few multiplexes was changed to obtain more uniform coverage.

Evaluation of uniformity in coverage distribution: singleplex vs. multiplex approach

In a cost efficient test, the full capacity of the GS-FLX instrument is used. This can be achieved by pooling different patients and/or disorders in a single lane. To maximize sample size in a single experiment, a uniform distribution of coverage is required. This means that the difference in number of reads between the "less efficient" amplified fragments and the "best performing" fragments should be as small as possible. We evaluated a singleplex (PoC1&2) and a multiplex approach (PoC3&4). A summary on the number of amplicons sequenced, reads mapped and coverage distribution is shown for each PoC study in Table 1. By calculating the "fold difference to mean coverage", we showed that starting from strongly optimized multiplex sets results in a more uniform distribution of coverage compared to the

equimolar pooling of singleplex sets (Figure 3). The "fold difference to mean coverage" can be used as "spread correction factor" to calculate the average coverage one needs to aim for to make sure that also for the less efficiently amplified fragments the minimum required coverage is obtained. The smaller the fold difference to mean coverage, the larger the number of samples that can be pooled in a single experiment, the more cost efficient a test will be. Considering our best optimized multiplex sets (PoC4) we obtained a 3.16-fold difference to mean coverage for 95% of the amplicons. This value can be used to calculate the average coverage we need to aim for to obtain a predefined minimum coverage for at least 95% of the amplicons (see below).

Multiplex optimization

Optimizing the multiplex sets by fragment analysis using FAM labeled MID primers, turned out to be a good strategy: on average, we found a nearly linear correlation between peak height (=relative fluorescence) on capillary electrophoresis within sets and coverage obtained after 454 sequencing as shown for 2 multiplex sets in Figure 4. The better results for the multiplex approach indicate that the second PCR round attaching the MID barcodes, introduces inequimolarities between the amplicons within the pool of singleplex PCRs. These equimolarities are further increased during the emulsion PCR.

Calculation of the number of samples that can be pooled in a single run

Knowing the fold difference to mean coverage, allows calculating the number of patients that can be analyzed in a single standard GS-FLX run (calculation template available in supplemental files of Hellemans et al. (under review):

- With a fold difference to mean coverage of 3.16, the required average coverage to obtain a minimum coverage of 38 = 38*3.16= 120. We opted for a threshold of 38-fold coverage based on statistical analyses made by Hellemans et al. who found that 38-fold coverage is required to detect a particular heterozygous variant with a probability of 99.9% when only variants present in at least 25% of the reads are considered as possible true variants. Reasons for the 25% variant frequency: see "Sensitivity".
- A standard GS-FLX run has 400 000 reads available; based on previous runs we found that the number of reads mapped is 95%= 400 000*95% = 380 000
- 111 amplicons are required to cover the complete coding region and splice sites of *BRCA1&2* and considering a 5% safety margin to correct for possible run errors and differences in MID amplification efficiencies during sequencing: 111*120 + 5% = 13986 reads/patient are required
- therefore, 380 000/13986= 27 patients can be pooled in a standard GS-FLX run, with a maximum of 5% of the amplicons not meeting the 38x coverage threshold.
- With the Titanium chemistry (1,100,000 reads available), the number of patients can be increased to 74.

Figure 5 shows that for PoC4 on average 4 of 111 amplicons (96.4%) did not meet the 38X-treshold. However, based on the capillary electrophoresis panels, higher coverage was expected for these amplicons, indicating that some fragments are less efficiently amplified by the emulsion PCR or had reduced coverage due to experimental variation. Primer concentrations in the multiplex reactions can still be increased for *BRCA2* 11-16, *BRCA2* 11-18, *BRCA2* 11-19 and *BRCA2* 18-1 to improve coverage for these amplicons and may provide a possible solution.

Sensitivity, specificity and filter settings

Sensitivity

In total 503 (133 distinct) sequence variants, previously identified with Sanger sequencing were evaluated in our sample set (30 patients and 93 control samples) with VIP v1.3. (The sensitivity data obtained with NextGene v2.0 are described below.)

All 40 unique substitutions (SNP's, missense, nonsense, splice site mutations) were easily detected. Detection of deletions or insertions is more challenging. MPS analysis software is based on mapping of single stranded reads on a reference sequence; hence, reads lacking one or more nucleotides or containing an insertion will complicate this process. Furthermore, pyrosequencing has its limitations for correct basecalling in homopolymeric regions (Huse, et al., 2007). Therefore, we specifically selected 93 insertion-deletion (indel) mutations, of which 20 deletions and insertions were present in homopolymeric tracts longer than 3 bp, to thoroughly evaluate the limitations of the technology.

Of the 93 indels, 2 remained completely undetectable and 1 additional mutation was filtered out due to low quality scores (<30) and was present in less than 25 % of the reads. Table 2 shows an overview of all undetected variants and their flanking sequences. All undetected variants affect homopolymer stretches of 7 nucleotides.

In total 130/133 of all unique variants were detected with the VIP software resulting in a sensitivity of 98% (100% for substitutions). In general, sensitivity of MPS will be higher, since we introduced a sample bias by selecting variants in complex sequence regions.

Reference bias

A priori, heterozygote variants are assumed to be present in 50% of the reads and homozygous mutant samples in 100%. On average the heterozygous variants were present in 48.4% (95%CI: 47.7-49.1%; range 27-77%) of the reads. Homozygous mutant variants were found with an average variant frequency of 99.0% (range: 93-100%). Based on these data we conclude that the reference bias in mapping is minimal.

Specificity

MPS is more sensitive than Sanger sequencing in terms of random sequencing errors and errors introduced by Taq polymerases, since the technology is based on sequencing of single, clonally amplified molecules. It is challenging to filter out these errors in the data analysis.

The VIP software ((version 1.3)(De Schrijver et al., 2010)) generates a list of variants detected in more than 10% of mapped reads. By analyzing 30 patient samples for the complete *BRCA1/2* coding sequence (3330 amplicons), this program generated a list of 5513 variants, of which only 443 are true variants, leaving 5070 false positives. Based on the analysis of the positive controls, criteria were defined to distinguish false positive from true variants. Reads defining a variant have to fulfill all of them, before being accepted as a possible true variant that requires confirmation with Sanger sequencing.

To filter out as many false positives as possible the following filters were applied in VIP:

1. Min. 38X coverage (cov) is required (Filter 1, cov >38x)
2. The variant needs to be present in at least 25% of the reads (Filter 2, AF > 25%)
3. At least in 1 direction (forward (F) or reverse (R)) a high quality score is required (Filter 3 Quality (Q) >30)

4. Variants in homopolymer stretches longer than 6 basepairs are non reliable calls (Filter 4, Homopolymer (Hp) >6)

An overview of the results is shown in Supplemental Table 3 & Figure 6. Application of these filters allowed obtaining a specificity of ~92 %. Specificity for each PoC separately is shown in Figure 7. Application of filters 1 and 2 resulted in the largest reduction of false positives: 1432 data points were filtered out because of <38X coverage and another 2628 because of being present in less than 25% of the reads, resulting in a specificity of 70% (filter 1&2). Fine tuning occurred by applying filters 3 and 4 decreasing the remaining 1010 false positives to 276. ln total, 104 distinct false positive variants remained. The recurrence of some false positives can be attributed to the Taq polymerase used or pseudogene amplification (Figure 8). 89% (244/276) of the false positives found are indel variations in the close neighborhood of a homopolymer region.

Data analysis with a commercial available software package: Nextgene (Softgenetics)

The performance of our in house developed VIP software package was compared with the commercially available software Nextgene version 2.0 (Softgenetics). This software has intrinsic filters in terms of coverage and frequency of a particular variant being present in the reads. An overall variant score is calculated, which provides an empirical estimation of the likelihood that a given SNP is real and not an artifact of sequencing or alignment. This score is mainly based on the concept of Phred scores where quality scores are logarithmically linked to error probabilities. For example a quality score of 10, gives a chance of 1 out of 10 that the base is incorrectly called. Furthermore, sub scores are integrated in this general score, taking into account the allele frequency, forward and reverse balance and a homopolymeric score, which penalizes indels found in homopolymeric regions. The maximum value for this overall

mutation score is 30. Filters and settings were defined based on our data obtained on the analysis of the positive controls, as we did for VIP.

We obtained an overall specificity of 84% (533 false positives in 30 patients), when all the variants present in 25% of the reads at least 38X coverage were taken into account. This specificity can be highly increased by applying an additional filter with the overall variant score.

Alle 93 indel variations and 40 substitutions were detected (100%), when no filter was set on an overall variant score. To improve specificity, we included an overall variant score > 15 requirement (according to software developers recommendations), this resulted in a specificity of 96%, but also in loss of detection of 3 variants present in homopolymeric regions (shown in Table 2). *BRCA1* c.1010del was also not detected with the VIP software, the remaining 2 are different compared to those missed with VIP, but they are also present in homopolymeric regions >6. Hellemans et al. found that the vast majority of reads for homopolymers up to the length of 6 can be correctly base called. However, for homopolymers of 7 nucleotides or more, the number of correctly called reads decreases. Therefore, detection of mutations in homopolymer tracts of 7 is challenging in in both programs evaluated.

Non random sequencing errors

Currently Sanger sequencing is considered as the gold standard. Since MPS involves the sequencing of single clones, it can be more vulnerable to errors introduced by polymerases. We did not use proofreading polymerase enzymes (except for the emulsion PCR). Since we perform 3 PCRs (amplicon specific, MID and emulsion PCR) in our workflow, non specific errors may occur. Performing our assays with proofreading Taq might lead to a reduction of

some false positives. Only 11% (32) of all false positives were single nucleotide substitutions, most of them can be explained by random sequencing or PCR artefacts and are only seen in a single direction.

For at least one variant we have strong evidence of a non random sequence error caused by the Taq polymerase: *BRCA2* c.9502-44 G>T (exon 26) was found in about 36% (95% CI:30-40% range: 19-61%) of all reads of all patients sequenced for this amplicon with Titanium Taq polymerase. Sanger sequencing of exon 26 and splice sites for all of the patients with Platinum Taq polymerase revealed only the G allele. Sanger sequencing of samples with Titanium Taq polymerase clearly showed the G>T substitution in all samples (Figure 9). These data clearly point at a possible role of the polymerase used for these non-random sequencing errors.

Evaluation specificity of MID barcodes used

No other mutation detection technology allows pooling of several patients in a single lane. MPS made this possible but the specificity of the MID barcodes used to distinguish individual patients is crucial. MID tags are designed in such a way that 2 sequencing errors may occur, without being defined as another MID. We analyzed the data for an experiment containing only MIDs 1-5 and verified whether reads for MIDs 6-60 were generated. For MID10 approximately 900 reads were mapped (2 mismatches allowed), compared to about 14,000 reads for MIDs included in the experiment. Since these reads are randomly scattered over all amplicons, chances on generating false positives are minimal, but not impossible for amplicons with low coverage. Therefore, MID10 should be excluded when 2 mismatches are allowed for mapping. MID10 is not present in the data when allowing only a single mismatch.

Hereby, only a small fraction of the reads (maximum 2000 scattered over all amplicons and patients) are lost and the risk on false positives is strongly reduced. As this is of major importance for reliable analyses in diagnostic settings, this solution is preferred.

Evaluation of multi-exon deletion detection

We evaluated the capacity to detect multi-exon deletions in both approaches. In PoC 2, a patient sample with a heterozygous deletion of *BRCA2* exon 1-4 was included and in PoC 4 a heterozygous deletion of *BRCA1* exon 18-19 was evaluated. By calculating the dosage quotient (DQ) (Goossens, et al., 2009), we obtained values not statistically different for the deleted exons compared to the non-deleted exons. To clearly distinguish a deleted exon, DQ values of 0.5 are expected. Normalizing the read counts by the average coverage of the reference patients for these amplicons, resulted in a DQ of 0.35 (*BRCA2* exon 2), DQ of 1 (*BRCA2* exon 3) and 0.72 (*BRCA2* exon 4). For *BRCA1* exon 18 and 19, we found a dosage quotient of 0.9 for both exons. Therefore, differences in coverage for the deleted exons were in our experiments not significant and we failed to detect this type of mutations.

DISCUSSION

In this study we evaluated whether massive parallel amplicon sequencing on the GS-FLX is suitable for implementation in a diagnostic setting. We used a multiplex bar coded amplicon sequencing approach for *BRCA1* and *BRCA2* as an example. For these genes a thoroughly optimized and efficient mutation detection strategy was already previously optimized in our laboratory (De Leeneer, et al., 2008; De Leeneer, et al., 2009). Therefore, we critically evaluated the progress that could be made using amplicon sequencing on the GS-FLX. Different critical aspects such as uniformity of coverage, sensitivity and specificity and reliability of primers and consumables were considered. Our experiences will be very useful for the optimization of other genes.

The recommended approach for amplicon sequencing of multiple sequences is based on the use of fusion primers which start with an A or B sequence on which the pyrosequencing reaction is initiated, followed by a patient specific barcode (MID) and a target specific sequence at the 3' end. Although this fusion approach is very simple, some impractical issues, in terms of primer management and set-up, will arise when the complexity of the experiment (i.e. number of amplicons and number of patients) increases. Primer costs and workload can be strongly reduced by attaching sequencing adaptors and barcodes to the PCR product by ligation (Meyer, et al., 2007) or nested patch PCR (Varley and Mitra, 2008). In our study, a second PCR was used to attach the MIDs and sequencing adaptors to the amplicon. We preferred this approach because of its simplicity, lower cost and workload compared to some other workflows. Multiplexing PCR products allowed to further reduce workload and consumable cost, and to save patient material. We have chosen to multiplex about 10 amplicons in a single set: optimization of such sets can be obtained with a minimum of extra efforts and results in a reduction of the initial workload by tenfold. Higher degrees of

multiplexing would require significantly more efforts and would result in a relatively small additional increase in efficiency. Furthermore, it would hamper the evaluation of pools by capillary electrophoresis prior to sequencing, which turned out to be a cost effective optimization method. Our results showed a good correlation between peak height (=relative fluorescence) on capillary electrophoresis within sets and coverage obtained after 454 sequencing, indicating a limited influence of the emulsion PCR. We currently have no other explanation than experimental variation for 4% of the PCR fragments that were less efficiently amplified by the emulsion PCR. A correlation between coverage and amplicon length, GC content or other sequence related characteristics could not be found. Although a homopolymeric tract is present in 2 of these amplicons (*BRCA2* 11.18 and 11.19), this cannot explain the lower coverage, since the presence of homopolymer stretches did not influence the amplification efficiency of other homopolymer-rich fragments.

By analysis of 30 patient samples for the complete *BRCA1/2* coding sequence with the VIP software, we obtained a list of 5513 variants present in at least 10% of the reads, of which only 443 are true variants, leaving 5070 false positives (specificity less than 10%). We are the first group defining filters to cope with MPS sequencing "errors" in a diagnostic setting. With this program we succeeded to obtain an overall specificity of 92%.

Filter 2 (AF >25%) is based on the minimal allele frequency found (27%) by analysis of 93 distinct indel variants and should avoid random PCR errors. The majority of false positives were generated by sequencing of homopolymeric regions, a known complication for pyrosequencing, resulting in undercalls and overcalls in homopolymeric stretches (i.e. one nucleotide missing or one nucleotide added compared to the reference sequence). Filter 3 (Q>30) and 4 (Hp>6) were applied to reduce these numbers, since homopolymeric stretches

influenced the Q value assigned to a specific read. These results were compared to those obtained with the commercially available software Nextgene version 2.0. Application of filters 1 and 2 (AF >25% and coverage >38 x), resulted in a specificity of 84%. Using an overall mutation score filter of >15, allowed to increase the specificity to 96% but resulted in loss of detection of some mutations present in homopolymeric regions of >6 nucleotides.

Our results provide evidence that some Taq polymerases introduce non random sequencing errors. For example *BRCA2* c.9502-44 G>T was found in almost all the patients analysed for the relevant amplicon when amplified with Titanium Taq polymerase (ClonTech), but not with Platinum Taq polymerase (Invitrogen). Since these errors will be reproducible in every run in almost all patients, they can be considered as true false positives and can be ignored in the long term.

Furthermore, the specificity of the barcodes used to identify patients was evaluated to confirm that false positives could not have been generated by aligning reads to a wrong MID. Despite the fact that no patients with MID10 were included in one of our experiments, reads for MID10 were detected when allowing 2 mismatches for the MIDs. To avoid possible false positives for amplicons with lower coverage we excluded MID10 from further experiments, since 10 misaligned reads for a given amplicon are sufficient to generate a variant with at a frequency of > 25% if only 38 fold coverage is obtained. In future experiments, data will be mapped allowing only a single mismatch as hereby the fraction of reads lost is minimal and misalignments to an incorrect MID will be avoided.

Homopolymeric stretches turned out to be major sources for false positive and false negative variants. For an efficient workflow a high specificity is required and our study shows that the

number of false positives is strongly reduced by the application of predefined filters. However, it was impossible to define adequate filters in both software programs evaluated without loss of detection of some variants in homopolymeric regions longer than 6 nucleotides. The coding region of *BRCA1/2* contains 7 homopolymer regions with 7 and 3 homopolymer regions with 8 nucleotides, in 11 of our amplicons. Until homopolymer analysis improves, these eleven amplicons are currently analysed with HRMCA in our setting, increasing detection capacity to 100% for all mutations evaluated (De Leeneer, et al., 2008). Sequencing technologies not based on pyrosequencing may outperform 454 sequencing for detection of variants in homopolymer regions.

Breast cancer diagnostics was earlier this year evaluated by deep sequencing on the GAIIx instrument (Illumina) by Morgan, et al. (starting from long-range PCRs) and by Walsh, et al. (DNA capturing by hybridization in solution to custom-designed cRNA oligonucleotide baits). Both reported detection of all variants/mutations evaluated, but the number of variants evaluated was much smaller and insertions/deletions were maximum 19bp in length. The deletion c.5503_5564del62 evaluated in our study may have remained undetected with this sequencing technology with read lengths of 2 × 76-bp paired-end reads (Walsh, et al.) or 51 bp. (Morgan, et al.).

Walsh, et al. did not report any false positives after filtering out variants present in less than 15% of the reads. With an average coverage of 1286 (range 781-1854) a reduction in false positives is indeed expected, however, such high coverage largely increases the cost per sample. We calculated consumable costs and labor time for our MPS approach, and found that our MPS set-up by pooling 74 patients in a single run costs about 345 EUR per sample (consumable cost: 232 EUR), which becomes cost competitive with our HRMCA approach and is much smaller than Sanger sequencing. The current development of commercial user-

friendly software for data analysis, allows MPS to outperform prescreening techniques used in combination with Sanger sequencing. Automation of the workflow will even further decrease the workload.

Turnaround times for genetic testing need to be strongly reduced to meet increasing expectations. Pooling large numbers of patients in a single run will only be useful in a diagnostic setting if different disorders can be pooled in a single run. This requires the development of uniform workflows for different genetic tests.

In some populations, large intragenic founder deletions represent an important fraction of the *BRCA1/2* mutation spectrum. Our MPS set-up allowed to detect point mutations with high sensitivity but turned out to be unreliable for the identification of large exon (or multi-exon) deletions. The application of 3 consecutive PCR rounds prior to sequencing most likely explains why deviations from diploidy remained undetected. Additionally we aimed for an average coverage of 120 (to obtain minimally 38), to allow a cost efficient test by pooling of a large number of samples in a single lane. Studies successfully reporting the detection of large intragenic rearrangements worked with much higher read depths, allowing more reliable quantifications of copy numbers (Goossens, et al., 2009; Walsh, et al.) For a sensitive analysis an additional technique for the detection of copy number variations needs to be included in the mutation detection strategy.

In conclusion, we developed an efficient workflow for high throughput *BRCA1/2* amplicon sequencing. Sensitivity and specificity of MPS amplicon sequencing is high and can be further increased by supplementing MPS assays to overcome issues related to homopolymeric regions. In terms of throughput, diagnostic testing can be highly accelerated and MPS facilitates offering genetic analyses to more at-risk patients. Considering cost efficiency MPS outperforms all other mutation screening techniques, but there is a shift from “wet” lab work

towards data analysis. In our opinion, Sanger sequencing should still be used for confirmation of deleterious variants in diagnostics.

ACKNOWLEDGEMENTS

This project was realized with the funding of an Emmanuel van der Schueren scholarship of the Flemish foundation against cancer to Kim De Leeneer. This research was supported by grant 1.5.150.07 from the Fund for Scientific Research Flanders (FWO) to Kathleen Claes and by GOA grant BOF10/GOA/019 (Ghent University). Bruce Poppe is a senior clinical investigator from FWO. This study was supported by a StepStone grant from the Industrial Research Fund from Ghent University; the Roche 454 GS-FLX instrument is part of the NXTGNT infrastructure, funded by a Hercules grant (middle heavy infrastructure). NXTGNT (initiated and supervised by Sofie Bekaert, Jo Vandesompele, Dieter Deforce, Philippe van Nieuwerburgh, Wim Van Criekinge, Jan Hellemans, Paul Coucke) is a genome analysis platform from Ghent University.

REFERENCES

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ and others. 2007. Direct selection of human genomic loci by microarray hybridization. Nat Methods 4(11):903-5.

Curtin NJ. 2005. PARP inhibitors for cancer therapy. Expert Rev Mol Med 7(4):1-20.

Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, Bicknell D, Bodmer WF, Davis RW, Ji H. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci U S A 104(22):9387-92.

De Leeneer K, Coene I, Poppe B, De Paepe A, Claes K. 2008. Rapid and sensitive detection of BRCA1/2 mutations in a diagnostic setting: comparison of two high-resolution melting platforms. Clin Chem 54(6):982-9.

De Leeneer K, Coene I, Poppe B, De Paepe A, Claes K. 2009. Genotyping of frequent BRCA1/2 SNPs with unlabeled probes: a supplement to HRMCA mutation scanning, allowing the strong reduction of sequencing burden. J Mol Diagn 11(5):415-9.

De Schrijver JM, De Leeneer K, Lefever S, Sabbe N, Pattyn F, Van Nieuwerburgh F, Coucke P, Deforce D, Vandesompele J, Bekaert S and others. Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline. BMC Bioinformatics 11:269.

Gerhardus A, Schleberger H, Schlegelberger B, Gadzicki D. 2007. Diagnostic accuracy of methods for the detection of BRCA1 and BRCA2 mutations: a systematic review. Eur J Hum Genet 15(6):619-27.

Goossens D, Moens LN, Nelis E, Lenaerts AS, Glassee W, Kalbe A, Frey B, Kopal G, De Jonghe P, De Rijk P and others. 2009. Simultaneous mutation and copy number variation (CNV) detection by multiplex PCR-based GS-FLX sequencing. Hum Mutat 30(3):472-6.

Hellemans J, De Leeneer K, De Schrijver J, Clemente L, Baetens M, Lefever S, De Keulenaer S, Claes K, Pattyn F, De Wilde B, Coucke P, Vandesompele J. . Massively parallel sequencing of PCR products, the road to next generation molecular diagnostics. Under review.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8(7):R143.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L and others. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420-6.

Korshunova Y, Maloney RK, Lakey N, Citek RW, Bacher B, Budiman A, Ordway JM, McCombie WR, Leon J, Jeddeloh JA and others. 2008. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. Genome Res 18(1):19-29.

Liu W, Smith DI, Rechtzigel KJ, Thibodeau SN, James CD. 1998. Denaturing high performance liquid chromatography (DHPLC) used in the detection of germline and somatic mutations. Nucleic Acids Res 26(6):1396-400.

Meyer M, Stenzel U, Myles S, Prufer K, Hofreiter M. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 35(15):e97.

Morgan JE, Carr IM, Sheridan E, Chu CE, Hayward B, Camm N, Lindsay HA, Mattocks CJ, Markham AF, Bonthron DT and others. Genetic diagnosis of familial breast cancer using clonal sequencing. Hum Mutat 31(4):484-91.

Pearson BM, Gaskin DJ, Segers RP, Wells JM, Nuijten PJ, van Vliet AH. 2007. The complete genome sequence of Campylobacter jejuni strain 81116 (NCTC11828). J Bacteriol 189(22):8402-3.

Pettersson E, Zajac P, Stahl PL, Jacobsson JA, Fredriksson R, Marcus C, Schioth HB, Lundeberg J, Ahmadian A. 2008. Allelotyping by massively parallel pyrosequencing of SNP-carrying trinucleotide threads. Hum Mutat 29(2):323-9.

Pol A, Heijmans K, Harhangi HR, Tedesco D, Jetten MS, Op den Camp HJ. 2007. Methanotrophy below pH 1 by a new Verrucomicrobia species. Nature 450(7171):874-8.

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell 127(6):1193-207.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463-7.

Taylor KH, Kramer RS, Davis JW, Guo J, Duff DJ, Xu D, Caldwell CW, Shi H. 2007. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. Cancer Res 67(18):8511-8.

Thomas RK, Nickerson E, Simons JF, Janne PA, Tengs T, Yuza Y, Garraway LA, LaFramboise T, Lee JC, Shah K and others. 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nat Med 12(7):852-5.

van der Hout AH, van den Ouweland AM, van der Luijt RB, Gille HJ, Bodmer D, Bruggenwirth H, Mulder IM, van der Vlies P, Elfferich P, Huisman MT and others. 2006. A DGGE system for comprehensive mutation screening of BRCA1 and BRCA2: application in a Dutch cancer clinic setting. Hum Mutat 27(7):654-66.

Varley KE, Mitra RD. 2008. Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. Genome Res 18(11):1844-50.

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J and others. 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2(12):e1326.

Walsh T, Lee MK, Casadei S, Thornton AM, Stray SM, Pennil C, Nord AS, Mandell JB, Swisher EM, King MC. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. Proc Natl Acad Sci U S A 107(28):12629-33.

Yao Y, Guo G, Ni Z, Sunkar R, Du J, Zhu JK, Sun Q. 2007. Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). Genome Biol 8(6):R96.

FIGURE LEGENDS

Figure 1: Schematic overview of the two set-up and different workflows used in the 4 Poc studies.

Panel A: Amplicon specific primers are fused with an universal M13tail. This target specific PCR product is amplified in a second PCR with primers, consisting of an M13tail, MID-tag (patient specific) and at the end an A or B adaptor

Panel B Upper panel: Schematic representation of the approach used in PoC 1 & 2. 111 PCR reactions per patient are performed and equimolarly pooled in 12 pools, followed by a second PCR to attach MID primers and sequencing adaptors. All products are equimolarly pooled prior to emulsion PCR and sequencing.

Lower panel: Approach used in PoC 3 & 4; 16 multiplex reactions were optimized containing the 111 amplicons to be amplified for each patient. By a second PCR round MID barcode and sequencing adaptors were attached. PCR products are equimolarly pooled prior to emulsion PCR and sequencing.

Figure 2: Comparison of fragment length of the total amplicon pool in theory and in practice.

Left panel: Theoretical profile of the amplicon length of the total amplicon pool for one patient when all amplicons are equimolarly pooled together. Amplicon length range is 233bp to 437 bp, with an average amplicon length of 335bp.

Right panel: Amplicon length profile obtained with the Bioanalyzer (Agilent) after column purification of the pool. Comparison of the two patterns is used as a quality control prior to emulsion PCR.

Figure 3: Evolution of distribution of coverage uniformity in the 4 PoC experiments.

In total we performed 4 GS-FLX runs. In our first two experiments (green and orange curve), equimolar pooling of simplex PCR reactions was used. The third (blue) and the fourth (purple) experiment were prepared according to the multiplexing protocol.

Panel A: A distribution plot of the coverage is shown. The experiments with the multiplex approach(PoC 3&4) clearly show a more uniform distribution of coverage compared to the pooled simplex runs (PoC 1&2). 38 fold coverage threshold is depicted by the dotted line. Failure rate in PoC 3 was higher because of poor amplification of some multiplex sets.

Panel B: Fold difference to the mean coverage is shown in function of the fraction of amplicons. The dotted lines depict the factor where 90 or 95% fraction of all amplicons is

considered. Our last experiment (PoC4-purple curve) clearly shows the best result where the smallest difference to mean coverage was obtained. 95% of all amplicons are sequenced with a spread correction factor of 3.16 (Table 1) (X-axis is shown in log10 scale).

Figure 4: Correlation between relative fluorescence and coverage within multiplex sets

The correlation between the relative fluorescence seen on capillary electrophoresis and coverage is shown.

Equimolarity of amplicons within a multiplex set was verified on capillary electrophoresis.

Figure 5: Distribution of coverage for each amplicon in BRCA1/2 (PoC 4)

Coverage for each BRCA1 and BRCA2 amplicon is shown, with the line indicating the threshold of 38 fold coverage. In total 111 amplicons are needed to cover the complete coding region of both genes. In BRCA2, 4 amplicons did on average not reach the 38x coverage threshold; further optimisation for these amplicons is required.

Figure 6: Overview of data generated

Data were analysed with an in house developed software program (VIP). Combined forward and reverse allele frequencies are plotted against coverage. Variants detected in 30 patients and 93 positive control samples are shown. Grey data points are the false positives filtered out when filters are applied. Green data points are true variants (503). The green outlier at almost 80% AF is deletion of 62 nucleotides, probably preferential amplification of the shorter allele has occurred. 80% of the true variants have an allele frequency in the 40-60 range. Red data points are the remaining false positives (276).

Figure 7: Specificity of 4 PoC studies analysed with VIP software

Specificity (%) of each PoC study and in total is plotted, for every filter applied in the VIP software. Filter 1 and 2 clearly result in the largest increase in specificity. The total specificity found is ~92%.

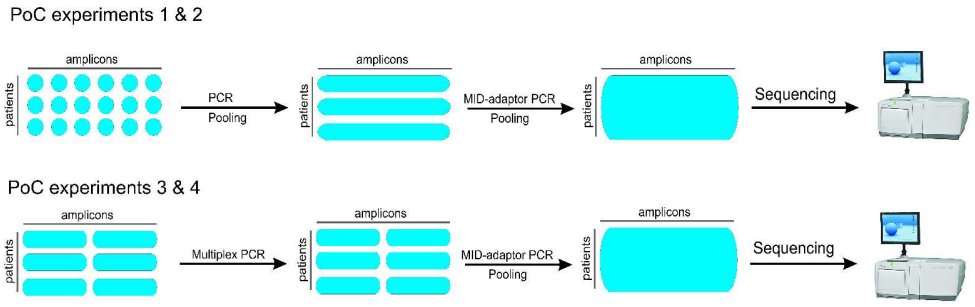
Figure 8: Pseudogene amplification by MPS

A part of the sequence of BRCA1 exon 2 is shown in both panels. In BRCA1 exon 2 for each sample 10-15 variants (in 25-50% of the reads, blue boxes in upper panel) were detected with MSP, but were not observed Sanger sequencing (lower panel, blue boxes represent the

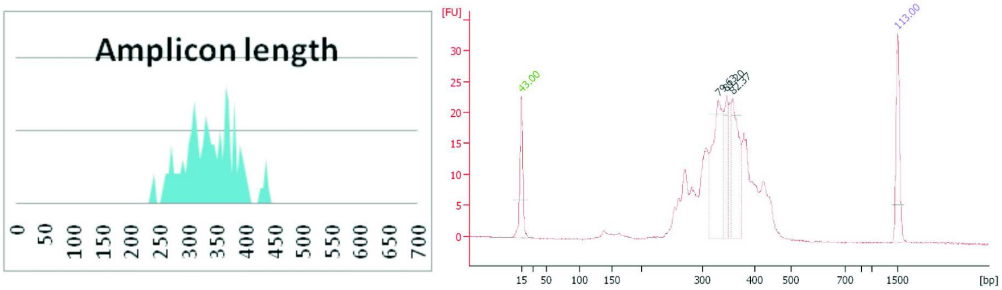
possible psuedogene nucleotides). Blasting of these sequences, showed co-amplification of BRCA1P1 (90% analogy with BRCA1 exon 2) by the reverse primer.

Figure 9: **Sanger sequencing results for *BRCA2* c.9502-44G>T**

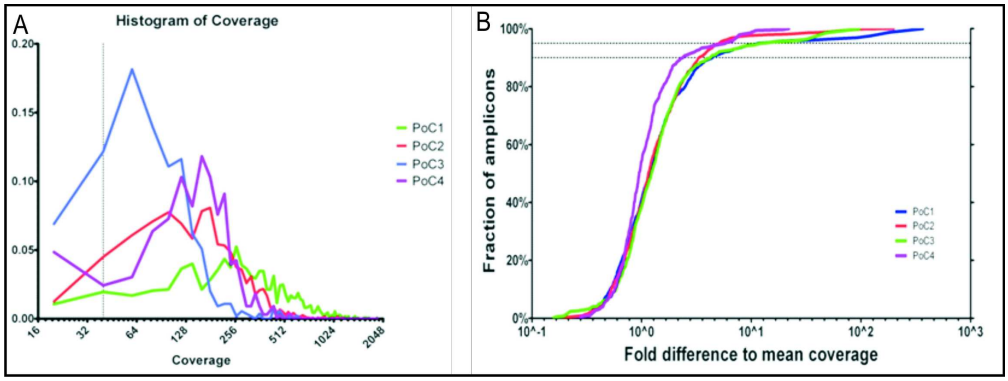
In the upper panel the *BRCA2* a Sanger sequence flanking the G allele at position c.9502-44 is shown for a sample amplified with Platinum Taq polymerase. In the lower panel, the same Sanger sequence for the same sample is shown after amplification with Titanium Taq polymerase. In this sample we see a clear reduction of the G allele and the replacement by a T allele.



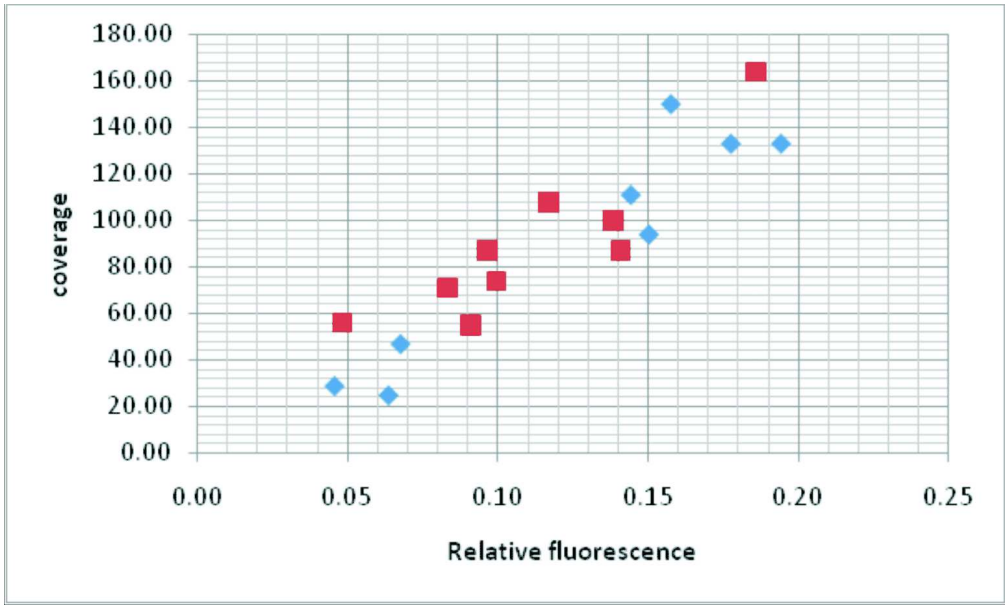
678x215mm (300 x 300 DPI)



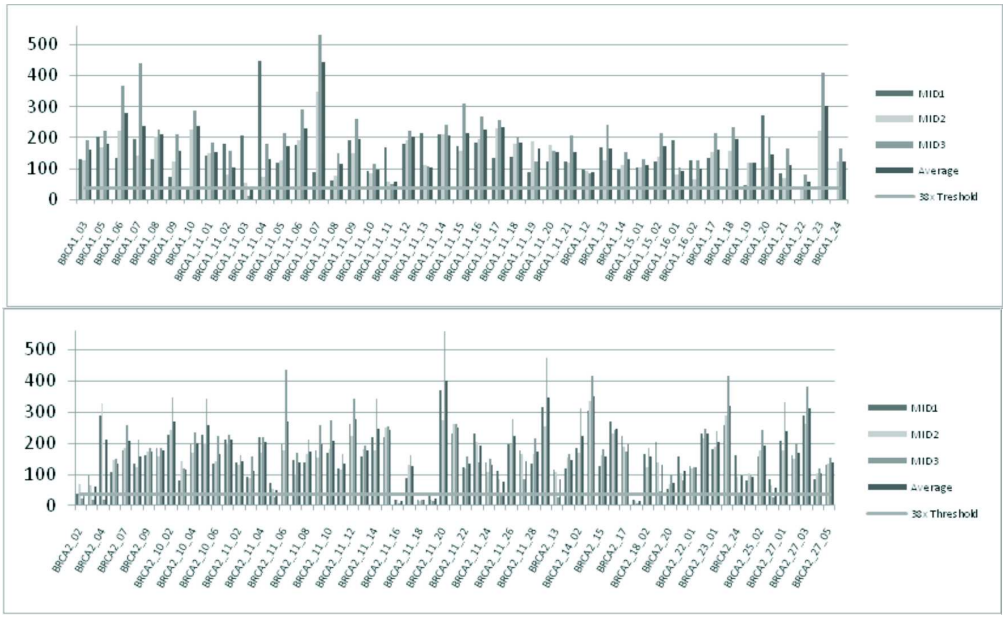
389x122mm (600 x 600 DPI)



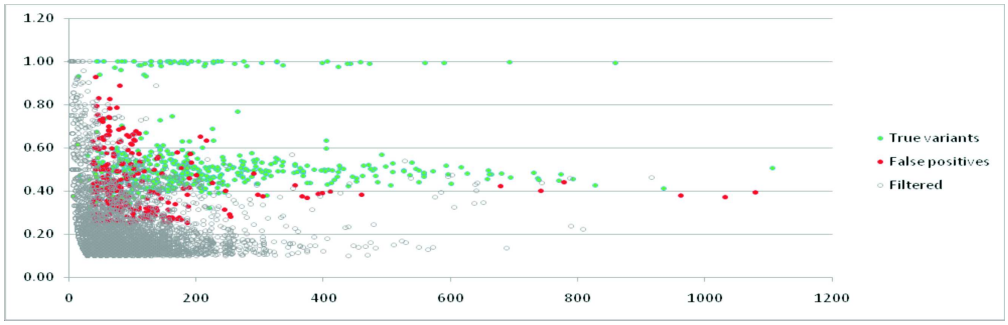
195x72mm (300 x 300 DPI)



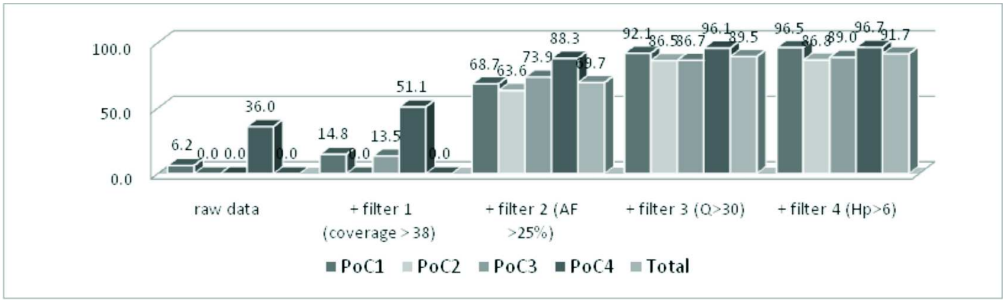
130x78mm (300 x 300 DPI)



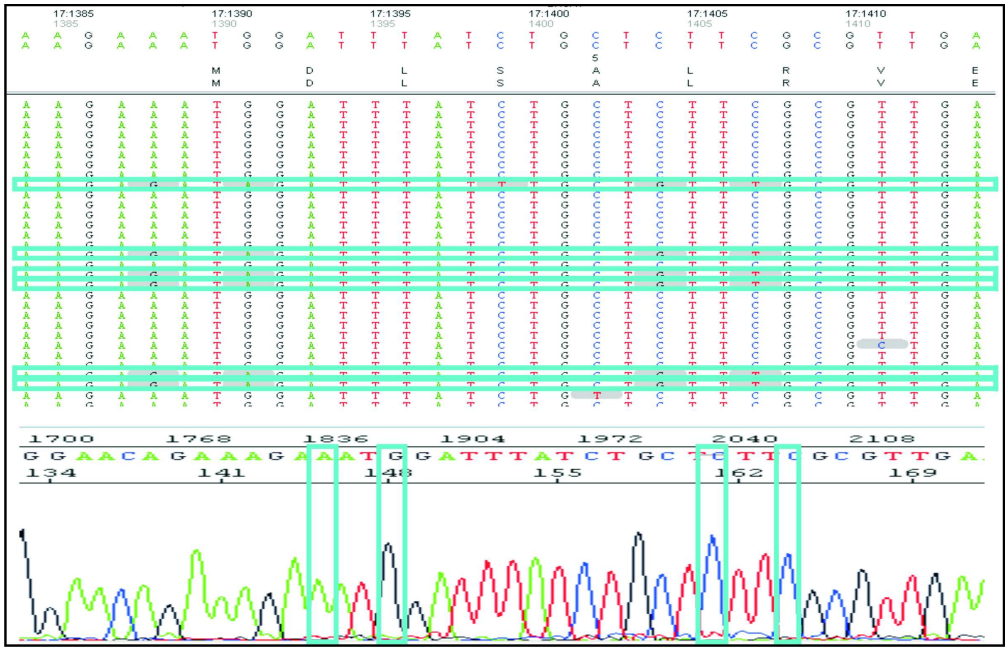
299x182mm (300 x 300 DPI)



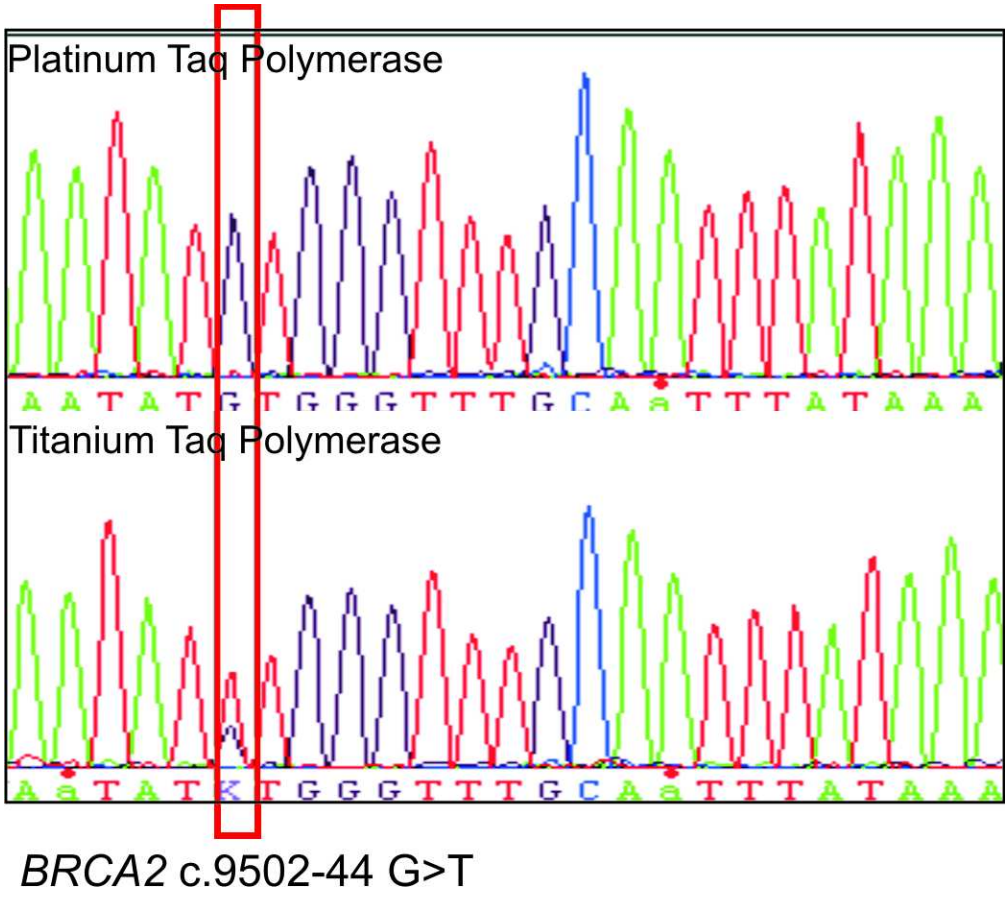
225x71mm (300 x 300 DPI)



286x84mm (300 x 300 DPI)



203x130mm (300 x 300 DPI)



88x77mm (300 x 300 DPI)

Table 1: Overview of characteristics of the 4 PoC experiments

	Singleplex pooling		Multiplex pooling	
	PoC 1	**PoC 2**	**PoC 3**	**PoC 4**
SET UP				
used run reads	100%	100%	~30%*	~25%*
mapped BRCA1/2 reads	515,916	250,884	78,937	55,574
amplicons sequenced	1221	1221	555	333
patients sequenced	11	11	5	3
COVERAGE				
min/average/max (per amplicon)	1/348/1870	1/191/1076	1/144/931	1/168/559
standard deviation	264	140	122	90
Variation coefficient	0.76	0.73	0.84	0.53
Fold difference to mean coverage 90%/95%	4.48/11.86	3.38/5.11	4.2/12.08	2.15/3.16
# amplicons <38 fold coverage (%)	107 (8.76)	84 (6.88)	54 (9.73)	21 (6.31)

* only 30% and 25% of the reads were used in these experiments, since in the same runs pooling different disorders in one experiment was evaluated

Table 2: Examples of variants analysed in homopolymeric regions ≥5

Variant (c.)	flanking sequences	Change of homopolymer length	coverage of nucleotide	Quality	VF (%)
Results with VIP software					
Undetected variants				*F*	
BRCA1 c.1016dup	ACTCCCAGCACAG**AAAAAAA**AGGTAGATCTGAATGCTGATC	7→8	38	n.a	<10%
BRCA2 c.994del	CTAGMAAGACTAGG**AAAAAAA**TTTTCCATGARGCAAACGCT	7→6	219	n.a	<10%
BRCA1 c.1010del	ACTCCCAGCACAG**AAAAAA**AGGTAGATCTGAATGCTGATCC	7→6	39	25	23%
Detected variants in homopolymeric regions					
BRCA1 c.3329dup	ATCCTGAAAT**AAAAAA**AGCAAGAATATGAAGAAGTAGTTC	6→7	99	31	47%
BRCA2 c.5577_5580del	ACATGAAACAA**TTAAAAAA**GTGAAAGACATATTTACAGAC	4→6	105	34	50%
BRCA1 c.2989_2990dup	AAAACTAAATGTAAG**AAAAAA**AATCTGCTAGAGGAAAACTTT	5→7	167	32	40%
Results with Nextgene sofware					
Examples of variants with low mutation score				Mutation score	
BRCA2 c.6351dup	GAAGATC**AAAAAAA**ACACTAGTTTT	7→8	221	10	33%
BRCA1 c.1010del	ACTCCCAGCACAG**AAAAAA**AGGTAGATCTGAATGCTGATCC	7→6	20	12	50%
BRCA1 c.1961delA	AGAGATAAAG**AAAAAAA**AGTACAACCAAA	8→7	76	11	52%

Supp T1: Overview of all sequence variants evaluated on NGS.

substitutions		
DNA level systematic nomenclature (BIC)	*Protein level*	*Suggested classification*
BRCA1		
c.134+3A>C (IVS3+3A>C)	non-coding	Mutation
c.1067A>G (1186A>G)	p.Gln356Arg	Polymorphism
c.1456 T>C (1575T>C)	p.Phe486Leu	Polymorphism
c.1487G>A (1606G>A)	p.Arg496His	Polymorphism
c.2077G>A (2196G>A)	p.Asp693Asn	Polymorphism
c.2082C>T (2201C>T)	p.Ser694Ser	Polymorphism
c.2311T>C (2430T>C)	p.Leu771Leu	Polymorphism
c.2612C>T (2731C>T)	p.Pro871Leu	Polymorphism
c.3113A>G (3232A>G)	p.Glu1038Gly	Polymorphism
c.3179A>C (3298A>C)	p.Glu1060Ala	Polymorphism
c.3548A>G (3667A>G)	p.Lys1183Arg	Polymorphism
c.3661G>T (3780G>T)	p.Glu1221X	Mutation
c.3841C>T (3960C>T)	p.Gln1281X	Mutation
c.4308T>C (4427T>C)	p.Ser1436Ser	Polymorphism
c.4867A>G (4956A>G)	p.Ser1613Gly	Polymorphism
c.4956G>A (5075G>A)	p.Met1652Ile	Polymorphism
c.4987-53 C>T (IVS17-53C>T)	non-coding	Polymorphism
c.536A>G (655A>G)	p.Val179Cys	Polymorphism
BRCA2		
c.10110G>A (10338G>A)	p.Arg3370Arg	polymorphism
c.1113C>A (1342C>A)	p.His372Asn	polymorphism

c.125A>G (353A>G)	p.Tyr42Cys	polymorphism
c.1-25G>A (203G>A)	non-coding	polymorphism
c.1365A>G (1593A>G)	p.Ser455Ser	polymorphism
c.2229 T>C (2457T>C)	p.His743His	polymorphism
c.2971A>G (3199A>G)	p.Asn991Asp	polymorphism
c.3396A>G (3624A>G)	p.Lys1132Lys	polymorphism
c.3807T>C (4035T>C)	p.Val1269Val	polymorphism
c.3851G>A (4079G>A)	p.Ser1284Asn	polymorphism
c.516+1G>A (IVS6+1G>A)	non-coding	Mutation
c.5645C>A (5873C>A)	p.Ser1882X	Mutation
c.5744C>T (5972C>T)	p.Thr1915Met	polymorphism
c.68+62T>G (IVS2+62T>G)	non-coding	polymorphism
c.681+56C>T (IVS8+56C>T)	non-coding	polymorphism
c.7057G>C (7285G>C)	p.Gly2353Arg	polymorphism
c.7242A>G (7470A>G)	p.Ser2414Ser	polymorphism
c.7806-14T>C (IVS16-14T>C)	non-coding	polymorphism
c.8182G>A (8410G>A)	p.Val2728Ile	polymorphism
c.865A>C 1093A>C	p.Asn289His	polymorphism
c.9257-16T>C (IVS24-16T>C)	non-coding	polymorphism
c.9976A>T (10204A>T)	p.Lys3326X	polymorphism
indels		
BRCA1		
c.562-58delT (IVS8-58delT)	non-coding	Polymorphism
c.1010delA (1129delA)	p.Glu337fs	Mutation
c.1016dupA (1135insA)	p.Lys339fs	Mutation
c.1072delC (1191delC)	p.Leu358fs	Mutation
c.1121delC (1240delC)	p.Thr374fs	Mutation
c.1287dupA (1406insA)	p.Asp430fs	Mutation

c.1292dupT (1411insT)	p.Leu431fs	Mutation
c.1319delT (1438delT)	p.Leu440fs	Mutation
c.1504_1508del5 (1623del5)	p.Leu502fs	Mutation
c.1881_1884del4 (2000del4)	p.Val627fs	Mutation
c.1961delA (2080delA)	p.Lys654fs	Mutation
c.2019delA (2138delA)	p.Glu673fs	Mutation
c.2197del5 (2316del5)	p.Glu733fs	Mutation
c.2210delC (2329delC)	p.Thr737fs	Mutation
c.2212_2215del4 (2331del4)	p.Val738fs	Mutation
c.232delA (351delA)	p.Arg78fs	Mutation
c.2380dupG (2478insG)	p.Glu787fs	Mutation
c.2405_2406delTG (2524delTG)	p.Val802fs	Mutation
c.2646_2648delTGC (2765delTGC)	p.Cys882del	Mutation
c.2685_2686delAA (2804delAA)	p.Gln895fs	Mutation
c.2689insA (2809insA)	p.Pro897fs	Mutation
c.2726delA (2845delA)	p.Asn909fs	Mutation
c.2727_2730del4 (2846del4)	p.Asn909fs	Mutation
c.2728delC (2847delC)	p.Gln910fs	Mutation
c.2764_2767del4 (2883delACAG)	p.Thr922fs	Mutation
c.2934T>G (3053T>G)	p.Tyr978X	Mutation
c.2989_2990dupA (3109insAA)	p.Asn997fs	Mutation
c.3329dupA (3448insA)	p.Lys1110fs	Mutation
c.3481_3491del11 (3600del11)	p.Glu1161fs	Mutation
c.3485delA (3604delA)	p.Asp1162fs	Mutation
c.3494_3495delTT (3613-3614delTT)	p.Phe1169fs	Mutation
c.3549AG>T (3668AG>T)	p.Lys1183fs	Mutation
c.3756_3759del4 (3875del4)	p.Leu1252fs	Mutation
c.3770_3771delAG (3889delAG)	p.Glu1257fs	Mutation
c.3820dupG (3939insG)	p.Val1274fs	Mutation
c.3891_3893delTTC (4010delTTC)	p.Ser1297fs	Mutation

c.4165_4166delAG (4284delAG)	p.Ser1389fs	Mutation
c.4391_4393delCTAinsTT (4510delCTAinsTT)	p.Pro1464fs	Mutation
c.4416delTTinsG (4535delTTinsG)	p.Leu1472fs	Mutation
c.4435delG (4554delG)	p.Val1479fs	Mutation
c.4575_4585del11 (4694del11)	p.Gln1525fs	Mutation
c.493del2 (612delCT)	p.Thr164fs	Mutation
c.5030_5033del4 (5149del4)	p.Thr1677fs	Mutation
c.5137delG (5256delG)	p.Val1713fs	Mutation
c.5191+2delT (IVS19+2delT)	non-coding	Mutation
c.5266dupC (5382insC)	p.Gln1756fs	Mutation
c.5329dupC (5448insC)	p.Thr1777fs	Mutation
c.5360_5361delGTinsAG (5479_5480delGTinsAG)	p.Cys1787fs	Mutation
c.5503_5564del62 (5622del62)	p.Arg1835fs	Mutation
BRCA2		
c.462_463delAA (690delAA)	p.Gln154fs	Mutation
c.1310_1313del4 (1538del4)	p.Lys437fs	Mutation
c.1389_1390delAG (1617delAG)	p.Thr463fs	Mutation
c.1705delC (1933delC)	p.Gln569fs	Mutation
c.2150delG (2378delG)	p.Cys717fs	Mutation
c.2584_2590del7 (2812del7)	p.Lys862fs	Mutation
c.2806_2809del4 (3034del4)	p.Lys936fs	Mutation
c.2957dupA (3185insA)	p.Asn986fs	Mutation
c.3269delT (3497delT)	p.Met1090fs	Mutation
c.3453delC (3681delC)	p.Ile1151fs	Mutation
c.3847_3848delGT (4075delGT)	p.Val1283fs	Mutation
c.3866_3867delAA (4094delAA)	p.Lys1289fs	Mutation
c.4171delG (4399delG)	p.Glu1391fs	Mutation
c.4435ins4 (4763 INS4)	p.Ser14797fs	Mutation

c.4449delA (4677delA)	p.Thr1483fs	Mutation
c.4456_4459del4 (4684del4)	p.Val1486fs	Mutation
c.4480dupA (4708insA)	p.Ser1494fs	Mutation
c.469_470delAA (697delAA)	p.Lys157fs	Mutation
c.4936_4939del4 (5164del4)	p.Glu1646fs	Mutation
c.4940delCA (5168delCA)	p.Thr1647fs	Mutation
c.5131delG (5359delG)	p.Val1711fs	Mutation
c.5180delA (5408delA)	p.Glu1727fs	Mutation
c.5213_5216del4 (5441delCTTA)	p.Thr1738fs	Mutation
c.5314delC (4542delC)	p.Val1438fs	Mutation
c.5350_5351delAA (5578delAA)	p.Asn1784fs	Mutation
c.5577_5580del4 (5805del4)	p.Ile1859fs	Mutation
c.5595_5596delAT (5823delAT)	p.Ile1865fs	Mutation
c.5681dupA (5909insA)	p.Tyr1894fs	Mutation
c.5722_5723delCT (5950delCT)	p.Leu1908fs	Mutation
c.5771_5774del4 (5999del4)	p.Ile1924fs	Mutation
c.5964delT (6174delT)	p.Ser1982fs	Mutation
c.6270_6271delTA (6498delTA)	p.His2090fs	Mutation
c.6280_6281delTT (6503delTT)	p.Leu292fs	Mutation
c.6280_6286del7 (6508del7)	p.Tyr2094fs	Mutation
c.634_635delAG (862delAG)	p.Arg212fs	Mutation
c.6351dupA (5579insA)	p.Asn1784fs	Mutation
c.6445delAT (6673delAT)	p.Ile2149fs	Mutation
c.6591_6592delTG (6819delTG)	p.Thr2197fs	Mutation
c.6603_6604delTG (6831delTG)	p.Ser2201fs	Mutation
c.6644_6647del4 (6872del4)	p.Tyr2215fs	Mutation
c.8904delC (9132delC)	p.Val2969fs	Mutation
c.9099_9100delTC (9327delTC)	p.Thr3033fs	Mutation
c.9458delG (9686delG)	p.Gly3153fs	Mutation
c.994delA (1222delA)	p.Ile332fs	Mutation

- GenBank reference sequences NT_010755.15 (4920610–5001764; BRCA1) and NT_024524.13 (13869617– 13953809; BRCA2) cDNA numbering according to reference sequence NM_007294.3 (BRCA1) and NM_000059.3 (BRCA2)
- Nucleotide numbering reflects cDNA numbering with +1 corresponding to the A of the ATG translation initiation codon in the reference sequence. The initiation codon is codon 1.

Supp T2: Specifications of reaction conditions and composition of multiplex reactions

	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	SET 7	SET 8
AMPLICONS	BRCA2_10_4	BRCA1_21	BRCA1_23	BRCA1_10	BRCA1_15_2	BRCA2_27_3	BRCA1_11_1	BRCA1_11_14
	BRCA2 9	BRCA1_22	BRCA1_17	BRCA2_14_3	BRCA2_23_2	BRCA1_18	BRCA1_19	BRCA2_27_2
	BRCA1_11_12	BRCA2_14_2	BRCA1_3	BRCA1_5	BRCA1_6	BRCA2_18_2	BRCA1_15_1	BRCA1_11_8
	BRCA2_11_15	BRCA1 11_2	BRCA2_11_10	BRCA1_24	BRCA2_7	BRCA2_11_12	BRCA2_22_1	BRCA2_15
	BRCA1_8	BRCA1_9	BRCA1_11_15	BRCA2_11_3	BRCA2_25_1	BRCA1_11_18	BRCA2_10_3	BRCA2_11_24
	BRCA2_22_2	BRCA1_11_4	BRCA2_25_2	BRCA2_10_2		BRCA1_11_6	BRCA1_11_13	BRCA2_11_28
	BRCA1_16_1	BRCA1_16_2	BRCA1_13	BRCA2_11_11			BRCA1_12	BRCA2_20
	BRCA2_11_4 BRCA2_11_21	BRCA1_11_16		BRCA1_11_21 BRCA2_11_17			BRCA2_27_4	BRCA2_11_1 BRCA2_10_6
PCR program	*Touch60*	*Touch60*	*Touch58*	*Touch58*	*Touch60*	*Touch58*	*Touch58*	*Touch60*
Taq polymerase	*Titanium Taq*	*Titanium Taq*	*PlatinumTaq*	*PlatinumTaq*	*Titanium Taq*	*PlatinumTaq*	*PlatinumTaq*	*Titanium Taq*

	SET 9	SET 10	SET 11	SET 12	SET 13	SET 14	SET 15	SET 16
AMPLICONS	BRCA1_11_17	BRCA1_11_20	BRCA2_8	BRCA2_12	BRCA2_18_1	BRCA2_11_6	BRCA2_21	BRCA2_2
	BRCA2_27_5	BRCA1_14	BRCA2_10_1	BRCA1_11_7	BRCA2_17	BRCA2_11_20	BRCA2_13	BRCA1_7
	BRCA1_20	BRCA2_11_13	BRCA2_11_7	BRCA2_27_1	BRCA2_3	BRCA1_11_19	BRCA2_11_27	BRCA2_4
	BRCA2_11_14	BRCA1_11_10	BRCA1_11_5	BRCA2_11_9	BRCA2_16	BRCA2_26	BRCA1_11_11	BRCA2_11_23

	BRCA2_23_1 BRCA2_10_5 BRCA1_11_9	BRCA2_11_22 BRCA2_11_8 BRCA2_5/6	BRCA2_11_26	BRCA2_10_7 BRCA2_14_1	BRCA2_19	BRCA2_11_19 BRCA2_11_5 BRCA2_24 BRCA2_11_16	BRCA2_11_25 BRCA1_11_3	
PCR program	*Touch60*	*Touch60*	*Touch60*	*Touch60*	*Touch60*	*Touch58*	*Touch60*	*Touch58*
Taq polymerase	*Titanium Taq*	*Titanium Taq*	*Titanium Taq*	*Titanium Taq*	*Titanium Taq*	*Titanium Taq*	*Titanium Taq*	*PlatinumTaq*

Supp T3: Primer concentrations in all multiplex reactions

SET 1			SET 2		
concentration		**primers/amplicon**	**concentration**		**primers/amplicon**
0.005	µM	BRCA2_10_4	0.009	µM	BRCA1_21
0.002	µM	BRCA2 9	0.001	µM	BRCA1_22
0.004	µM	BRCA1_11_12	0.007	µM	BRCA2_14_2
0.016	µM	BRCA2_11_15	0.002	µM	BRCA1 11_2
0.016	µM	BRCA1_8	0.024	µM	BRCA1_9
0.016	µM	BRCA2_22_2	0.004	µM	BRCA1_11_4
0.008	µM	BRCA1_16_1	0.008	µM	BRCA1_16_2

0.016	µM	BRCA2_11_4	0.016	µM	BRCA1_11_16
0.024	µM	BRCA2_11_21			

SET 3			SET 4		
concentration		**primers/amplicon**	**concentration**		**primers/amplicon**
0.003	µM	BRCA1_23	0.006	µM	BRCA1_10
0.004	µM	BRCA1_17	0.011	µM	BRCA2_14_3
0.008	µM	BRCA1_3	0.032	µM	BRCA1_5
0.008	µM	BRCA2_11_10	0.003	µM	BRCA1_24
0.008	µM	BRCA1_11_15	0.012	µM	BRCA2_11_3
0.008	µM	BRCA2_25_2	0.024	µM	BRCA2_10_2
0.012	µM	BRCA1_13	0.024	µM	BRCA2_11_11
			0.008	µM	BRCA1_11_21
			0.024	µM	BRCA2_11_17

SET 5			SET 6		
concentration		**primers/amplicon**	**concentration**		**primers/amplicon**

0.004	µM	BRCA1_15_2	0.007	µM	BRCA2_27_3
0.004	µM	BRCA2_23_2	0.002	µM	BRCA1_18
0.004	µM	BRCA1_6	0.003	µM	BRCA2_18_2
0.016	µM	BRCA2_7	0.008	µM	BRCA2_11_12
0.003	µM	BRCA2_25_1	0.008	µM	BRCA1_11_18
			0.024	µM	BRCA1_11_6

SET 7

concentration		**primers/amplicon**
0.016	µM	BRCA1_11_1
0.004	µM	BRCA1_19
0.008	µM	BRCA1_15_1
0.008	µM	BRCA2_22_1
0.016	µM	BRCA2_10_3
0.008	µM	BRCA1_11_13
0.008	µM	BRCA1_12
0.032	µM	BRCA2_27_4

SET 8

concentration		**primers/amplicon**
0.007	µM	BRCA1_11_14
0.008	µM	BRCA2_27_2
0.008	µM	BRCA1_11_8
0.005	µM	BRCA2_15
0.004	µM	BRCA2_11_24
0.016	µM	BRCA2_11_28
0.024	µM	BRCA2_20
0.016	µM	BRCA2_11_1

0.024	µM	BRCA2_10_6

SET 9

concentration		primers/amplicon
0.006	µM	BRCA1_11_17
0.008	µM	BRCA2_27_5
0.003	µM	BRCA1_20
0.008	µM	BRCA2_11_14
0.008	µM	BRCA2_23_1
0.024	µM	BRCA2_10_5
0.008	µM	BRCA1_11_9

SET 10

concentration		primers/amplicon
0.006	µM	BRCA1_11_20
0.002	µM	BRCA1_14
0.008	µM	BRCA2_11_13
0.014	µM	BRCA1_11_10
0.024	µM	BRCA2_11_22
0.024	µM	BRCA2_11_8
0.024	µM	BRCA2_5en6

SET 11

concentration		primers/amplicon
0.006	µM	BRCA2_8
0.024	µM	BRCA2_10_1
0.008	µM	BRCA2_11_7

SET 12

concentration		primers/amplicon
0.016	µM	BRCA2_12
0.016	µM	BRCA1_11_7
0.008	µM	BRCA2_27_1

0.008	µM	BRCA1_11_5	0.008	µM	BRCA2_11_9
0.008	µM	BRCA2_11_26	0.016	µM	BRCA2_10_7
			0.008	µM	BRCA2_14_1

SET 13			SET 14		
concentration		**primers/amplicon**	**concentration**		**primers/amplicon**
0.002	µM	BRCA2_18_1	0.008	µM	BRCA2_11_6
0.024	µM	BRCA2_17	0.016	µM	BRCA2_11_20
0.004	µM	BRCA2_3	0.008	µM	BRCA1_11_19
0.024	µM	BRCA2_16	0.016	µM	BRCA2_26
0.008	µM	BRCA2_19	0.016	µM	BRCA2_11_19
			0.008	µM	BRCA2_11_5
			0.016	µM	BRCA2_24
			0.024	µM	BRCA2_11_16

SET 15			SET 16		
concentration		**primers/amplicon**	**concentration**		**primers/amplicon**

0.008	µM	BRCA2_21	0.008	µM	BRCA2_2
0.008	µM	BRCA2_13	0.008	µM	BRCA1_7
0.01	µM	BRCA2_11_27	0.008	µM	BRCA2_4
0.01	µM	BRCA1_11_11	0.024	µM	BRCA2_11_23
0.008	µM	BRCA2_11_25			
0.016	µM	BRCA1_11_3			

Sensitivity						
	Detected variants		False negatives (coverage)		True False negatives	
	#	unique	#	unique		
PoC 1	146	30	16	14	0	0
PoC 2	160	40	12	5	1	1
PoC 3	43	29	12	8	0	0
PoC 4	204	115	2	2	2	2

Supp T4A: Summary of sensitivity data (VIP)

Total	553	144	42	29	3	3

Supp T4B: Summary of specificity data (VIP)

	Specificity (false positives)									
	raw data		Filter 1 (coverage > 38)		Filter 2 (AF >25%)		Filter 3 (Q>30)		Filter 4 (Hp>7)	
	#	unique	#	unique	#	unique	#	unique	#	unique
PoC 1	1145	205	1040	168	382	48	96	27	43	14
PoC 2	2948	693	1955	474	444	112	165	59	161	55
PoC 3	764	377	480	248	145	76	74	36	61	28
PoC 4	213	177	163	140	39	47	13	19	11	7
Total	5070	1452	3638	1030	1010	283	348	141	276	104