The Journal of Molecular Diagnostics, Vol. 14, No. 6, November 2012

http://dx.doi.org/10.1016/j.jmoldx.2012.05.006

Molecular Analysis of the Breast Cancer Genes *BRCA1* and *BRCA2* Using Amplicon-Based Massive Parallel Pyrosequencing

Geneviève Michils,* Silke Hollants,* Luc Dehaspe,* Jeroen Van Houdt,* Yannick Bidet,† Nancy Uhrhammer,‡ Yves-Jean Bignon,†‡ Joris R. Vermeesch,* Harry Cuppens,* and Gert Matthijs*

From the Center for Human Genetics, University of Leuven, Leuven, Belgium; the Laboratoire d'Oncologie Moléculaire,† Clermont Université, Clermont-Ferrand, France; and the Laboratoire d'Oncologie Moléculaire,‡ Centre Jean Perrin, Clermont-Ferrand, France*

The aim of this study was to implement the massively parallel sequencing technology for diagnostic applications. We evaluated an amplicon-based method for the analysis of the *BRCA1* and *BRCA2* genes on the Roche 454 GS-FLX sequencer, to identify disease-causing mutations in breast and/or ovarian cancer patients. A first evaluation relied on the analysis of DNA fragments containing known mutations. Secondly, the entire coding regions of the *BRCA1* and *BRCA2* genes were interrogated in more than 400 patient samples, using a multiplex PCR-based assay. Variants were filtered on the basis of their frequency (20%) and sequencing depth (>25×). Special attention was given to sequencing accuracy in homopolymers. In the initial evaluation, all known heterozygous mutations were detected. The percentage of mutant reads ranged from 22% to 62%. For the multiplex assay, 95% sensitivity and 91% specificity were obtained. In addition, we were able to reliably distinguish mutations from noise through the analysis of the raw signal intensities in homopolymers. This work presents an evaluation of the next-generation sequencing for use in diagnostics, based on a relatively high number of samples and experiments. We anticipate that the technique would further improve, and would allow reducing the costs per analysis and the turn-around time, to benefit patients who undergo BRCA molecular testing. *(J Mol Diagn 2012, 14:623–630; http://dx.doi.org/10.1016/j.jmoldx.2012.05.006)*

Our aim has been to evaluate and possibly to introduce next-generation sequencing (NGS) technology for the screening of the *BRCA1* and *BRCA2* genes in a diagnostic context. We present an evaluation of amplicon-based sequencing on the 454 GS-FLX Titanium Genome Sequencer (Roche 454 Life Sciences, Branford, CT)[1]as a tool for the molecular diagnosis of hereditary breast and/or ovarian cancer.

This autosomal dominantly inherited cancer syndrome is caused by heterozygous germline mutations in the highly penetrant breast cancer genes, *BRCA1* (MIM 113705) and *BRCA2* (MIM 600185). The great majority of the mutations are nonsense mutations and frameshifts [Breast Cancer International Core (BIC)]. Exonic deletions or duplications have also been reported.

The *BRCA* genes are commonly analyzed by direct sequencing.[2,3] Mutation-scanning methods, such as denaturing high-performance liquid chromatography (dHPLC), denaturing gradient gel electrophoresis (DGGE), or high-resolution melting curve analysis (HRMCA) are also frequently used.[4–6] Exonic deletions or duplications can be detected by multiplex ligation-dependent probe amplification (MLPA),[7–9] Southern blot or (semi-) quantitative (multiplex) PCR.[10]

As only 10% to 20% of the familial cases are attributable to mutations in the *BRCA1* and *BRCA2* genes, no mutation is found in at least 80% of the families.[11] Also, the cost per test is high, especially if Sanger sequencing is used.[12] This explains why in most laboratories, a careful selection of patients and families is made before analysis. The practical approach often consists of a combination of the mutation-scanning strategies, to reduce the economical burden imposed by Sanger sequencing. The NGS technology could change the approach, if it effectively allows reducing the cost. Multiplexing is a way of improving efficiency, especially when there is a great number of amplicons and patients to be screened. Indeed, the development of specific tools for primer design and pooling, that was described by Goossens et al,[13] were recently commercialized.

To date, at least five NGS platforms are commercially available. Their properties differ in many ways, ie, their

Supported by the EU-funded TECHGENE (FP7-223143) and EuroGentest (FP7-261469) projects and by a grant from the National Cancer Plan from the Federal Public Service for Health.

Accepted for publication May 17, 2012.

Address reprint requests to Gert Matthijs, Ph.D., Herestraat, 49 –B 3000 Leuven, Belgium. E-mail: gert.matthijs@uzleuven.be.

sequencing capacity, chemistry, and applications.[14] Specifically for the *BRCA1* and *BRCA2* genes, Walsh et al[15] proposed a capture-based method on the Genome Analyzer (Illumina, San Diego, CA) and showed that mutations could be identified in 10 high-risk breast cancer genes. De Leeneer et al[16] used a multiplex PCR-based approach combined with an in-house data analysis pipeline on the 454 GS-FLX.

We tested the 454 GS-FLX platform with a multiplex PCR design that is different from the one presented by Leeneer et al. We extensively analyzed the NGS data with different software packages and compared the results to Sanger data. In addition, the balance between the coverage and the number of reads with a heterozygous variation was thoroughly examined. Finally, we tested a workflow aiming to improve the base calling and the interpretation of homopolymeric stretches.

Materials and Methods

Sample Preparation

The DNA samples that were used in this study were obtained from referrals for molecular testing of the breast cancer genes. All patients have been genetically counseled at our Center and provided informed consent for testing. Genomic DNA had been extracted from lymphocytes by standard semiautomated methods (Chemagic Magnetic Separator Module I, Chemagen, Baesweiler, Germany; Autopure LS, Qiagen, Hilden, Germany).

The amplicons generated for the GS-FLX System were all purified using MultiScreen PCR Filter Plates (Millipore, Billerica, MA) on a pipetting robot (Janus Automated Workstation, Perkin Elmer, Waltham, MA) using standard procedures. The PCR products were eluted in 30 μL of water. After DNA measurement using the Quant-iT PicoGreen ds DNA Assay kit (Invitrogen, Life Technologies, Grand Island, NY), the amplicons were pooled to generate an equimolar mix. Finally, emulsion-PCR with the GS-FLX Titanium Lib-A emPCR kit and the sequencing reaction with the GS Titanium Sequencing XLR70 kit were performed according to the instructions.

Testing Strategy

For the first validation phase, we evaluated the NGS technology on the basis of a set of amplicons that covered the whole *BRCA1* coding region. For this experiment, we choose to maintain the amplicon design that was originally defined for HRMCA.[6] For the *BRCA1* gene, this represented a set of 45 amplicons, ranging from 172 bp to 370 bp (adaptor and key sequences excluded). For each amplicon, a specific DNA sample with a previously identified variation or a pathogenic mutation was used as template. In total, 37 heterozygous variations (21 substitutions and 16 frameshifts) were selected. These variations were distributed over 34 amplicons; wild-type DNA was used as template in the remaining amplicons. The amplicons were all separately amplified and pooled together to generate an (artificial) control sample (Table 1). The amplicons were amplified using a two-step PCR protocol. In brief, after a first PCR of the individual amplicons with the (exon) specific primers that contained universal tags, a second amplification was performed with primers containing the key and adaptor sequences (Roche, Technical Bulletin 005–2009). The PCR products were then prepared as described above.

Table 1. Heterozygous *BRCA1* Variations Tested for the Evaluation of the AVA and SeqNext Software Packages

Variation cDNA (protein) level	Variant frequency* AVA 2.5	Variant frequency* SeqNext 3.5.0
c.1A>G (p.?)	34	28
c.68_69delAG	47	28
c.133A>C (p.Lys45Gln)	56	62
c.212 + 3A>G (IVS5 + 3A>G)	56	58
c.441G>C (p.Leu147Phe)	55	55
IVS7-34T>C	56	57
c.591C>T (p.Cys197Cys)	47	49
c.692C>T (p.Thr231Met)	47 (2.3)	47 (3.3.2)
c.744C>G (p.Thr248Thr)	49	50
c.1072delC	Not detected	46 (3.3.2)
c.1115G>A (p.Trp372X)	43	42
c.1292delT	Not detected	49
c.1621C>T (p.Gln541X)	55	55
c.2082C>T (p.Ser694Ser)	53 (2.3)	46 (3.3.2)
c.2197del5	24	23
c.2359dupG	Not detected	50
c.2612C>T (p.Pro871Leu)	50	50
c.2728delC	Not detected	44 (3.3.2)
c.2898delT	Not detected	36
c.3113A>G (p.Glu1038GLy)	53	54
c.3119G>A (p.Ser1040Asn)	47	47
c.3331del4	26	22
c.3481del11	58	56
c.3548A>G (p.Lys1183Arg)	59	59
c.3756del4	44	33
c.3891del3	48	27
c.4088C>G (p.Ser1363X)	46	46
c.4165delAG	48	50
c.4308T>C (p.Ser1436Ser)	48	48
c.4327C>T (p.Arg1443X)	50	51
c.4391delCTAinsTT	50	49
c.4535G>T (p.Ser1512Ile)	52	53
c.4956G>A (p.Met1652Ile)	43	43
c.5137delG	Not detected	54
c.5177del4	49	47
c.5266dupC	Not detected	50
c.5278–14C>G (IVS20–14C>G)	45	43
Mean variant frequency (±SD)	47 ± 10	46 ± 10

Variant frequencies are given for each variant.
*Ratio, in percent (%), of the reads containing the variation versus the total number of reads.

In the second part of the study, the *BRCA* MASTR assay v1.2 (Multiplicom, Niel, Belgium) was evaluated in two phases. First, the sensitivity and specificity were calculated by analyzing seven samples for which Sanger data were available for the entire *BRCA1* and *BRCA2* coding regions. Secondly, 17 additional samples that had been previously screened with a combination of HRMCA and dHPLC, were used to generate more NGS data to evaluate the robustness of the method in terms of coverage.

The DNA sample of each patient was amplified in 12 multiplexes through a two-step PCR method according to

Table 2. Testing Strategy and Run Information

Type of experiment	Evaluation software	Evaluation multiplex (1)	Patient screening	Evaluation multiplex (2)
No. samples/run	1	24	32	16
No. runs	1	1	13	2
No. analyzed samples	1	24	400	32
GS-FLX capacity/run	6% (1/16)	50% (1/2)	50% (1/2)	25% (1/4)
Average reads/run	18,710	414,246	486,762	235,905
Amplification design	Simplex	Multiplex MASTR v. 1.2	Multiplex MASTR v. 1.2	Multiplex MASTR v. 2.0
No. amplicons (gene)	45 (*BRCA1*)	169 (*BRCA1/2*)	169 (*BRCA1/2*)	94 (*BRCA1/2*)
Coverage/amplicon	415	97	84	153
% Amplicons >25×	100	89	95	100
% Amplicons >40×	100	79	87	100
Software for mutation analysis*	AVA; SeqNext	AVA; SeqNext	SeqNext	SeqNext

All of the read data were determined by the AVA software.
*See Table 3 for additional details.

the manufacturer's instructions. Briefly, 20 ng of genomic DNA was used as a template in a first PCR that allowed the amplification of the targeted regions with tagged primers. A 1-μL quantity of a 1/1000 dilution of each of these PCR products was subsequently amplified using primers containing the universal tags flanked by the respective multiplex identifiers, ie, "barcode" sequences, the key and adaptor sequences. The 12 final PCR products were pooled per patient using the mixing scheme provided by the manufacturer, purified as described above and prepared for GS-FLX sequencing. Finally, the procedure was extended to 400 new patient samples.

The same strategy was applied to *BRCA* MASTR assay v2.0 (Multiplicom). The amplicon design was adapted for titanium chemistry, allowing the amplification of longer fragments. Overall, 94 amplicons spanning the *BRCA1/2* genes were amplified in five multiplexes. To evaluate this multiplex assay, 32 patient samples with 31 known pathogenic *BRCA1/2* mutations (19 frameshifts and 12 substitutions) were selected and amplified according to the manufacturer's instructions.

A summary of all experiments and validation steps is given in Table 2.

Data Analysis

Reads Filtering

The signals were processed and the reads were filtered and assembled according to the standard parameters (the reader is referred to 454 sequencing system software manual, v 2.5.3, Part B, Roche). The quality of the reads was analyzed with a set of stringent filters. The reads were discarded from further analysis (keypass filter, dot filter, mixed filter) and trimmed back from the 3′ end (primer filter, signal intensity filter, valley filter). Finally, PHRED-equivalent scores were also used to trim reads based on an internal calibrated signal (quality score trimming filter; Roche Diagnostics Belgium, personal communication, 2012).

All filter-passed reads were aligned to the Ensembl files *BRCA1*_ENSG00000012048 and *BRCA2*_ENSG00000139618; the mutations were annotated according to the HGVS guidelines (transcript reference files *BRCA1* NM_007294.3 and *BRCA2* NM_000059.3).

Sequencing Depth

The coverage per amplicon was defined by the amplicon variant analyzer (AVA) software (Roche). A threshold of 25-fold sequencing depth per amplicon was used in the experiments; a 40× cut-off is shown on the figures as an additional indication.

Software Tools

We used AVA, versions 2.3 to 2.5.3, as well as SeqNext software (JSI Medical Systems GmbH, Kippenheim, Germany), versions 3.3.2 and 3.5.0.

Mutation Detection

An extensive evaluation was performed on the ability of the AVA and SeqNext to identify known mutations. Because breast and/or ovarian cancer is dominantly inherited, pathogenic mutations are expected to be heterozygous. A mutation will theoretically be present in 50% of the corresponding reads. However, in reality, the variant frequency (VF) is binomially distributed around this value. In this study, a threshold of 20% was chosen. This cut-off was considered useful to distinguish the signal (ie, true mutations) from the noise (ie, false-positive results). Because this threshold is lower than the one proposed by De Leeneer et al,[16] we decreased the minimum sequencing depth to 25×.

Furthermore, the SeqNext software works with different thresholds simultaneously. The variants are sorted separately into a "distinct" table and an "other" table. To evaluate the sensitivity and specificity of the method, the initial cut-off of 20% has been applied for both tables. This allowed sorting all variants that were sequenced at a minimum of 25x coverage and present in at least 20% of the reads (VF ≥ 20%). After the evaluation phase, new DNA samples from BRCA patient samples were submitted to NGS analysis. To increase the power of detection

Table 3. Results of the Software Evaluation

	Software Version VF threshold	AVA 2.3 to 2.5.3 20	SeqNext 3.3.2, 3.5.0 20
Type of experiment		Artificial control	
BRCA1 37 known mutations 45 amplicons Max sensitivity: 92% (95%CI)	Observed VF range Average VF and SDEV Analytical sensitivity Specificity	24–59 47 ± 10 81% (30/37)* 82% (100–8/45)†	22–62 46 ± 10 100% (37/37) 84% (100–7/45)
Type of experiment		Evaluation multiplex assay v. 1.2	
BRCA1 + *BRCA2* 63 Known mutations 1182 Amplicons Max sensitivity: 95% (95% CI)	Observed VF range Average VF and SDEV Analytical sensitivity Specificity	27–62 49 ± 7 89% (56/63)* 98% (100–25/1182)†	25–63 46 ± 5 100% (63/63) 91% (100–104/1182)

*Indels between 1 and 3 bp were likely not called by the AVA software.
†False-positive results in homopolymers were not reported by the AVA software.

and to avoid false-negative results due to stochastic events, a 20% VF was used for the "distinct" table and a second threshold of 10% was applied to the variants listed in the "other" table. Variants present in homopolymeric stretches (≥7 bp) were listed separately.

Analysis of Homopolymer Sequence Data

Light signal intensities (alias: flow values) in homopolymers were extracted from the dataset using the AVA software and sorted by intervals of 0.1. This was done for homopolymers for which a mutant sample was available (ie, a sample with a true mutation, typically a deletion or a duplication). Theoretically, the signal intensities should correspond to the expected nominal value for the respective stretch (ie, number of bases in the stretch). In practice, however, the signal intensities cluster around this value. By analyzing the number of reads mapping to each 0.1 interval, it is possible to derive relative ratios and to study the distribution of the reads in each stretch.

For example, if 60 reads in total were mapped to a seven-nucleotide stretch, and six of these reads had a raw signal intensity of 7.2, a relative ratio of 10% could be calculated. Those ratios were made for each 0.1 interval, the sum of the ratios for each patient being 100%. To visually present those data, the obtained values were represented in a black-and-white scale (black indicating high, and white indicating low) in a "heatmap." Under perfect circumstances—100% of the reads with a signal intensity of exactly 7—only one black box corresponding to this interval would be present in the plot.

Results

The first experiment was conducted on an artificial set of amplicons containing 37 known heterozygous variations. AVA did not succeed in reporting all variations: the 1 bp deletions or insertions were present in the global alignment but were not reported by the software (results not shown). All variations could be successfully detected by the SeqNext software. The results showed that the variant frequency (mutant reads/total number of reads) of heterozygous variations is 46% ± 10% (range 22 to 62; Table 1).

Multiplexing has been evaluated in 24 samples containing a total of 63 known heterozygous variants including 10 *BRCA1* or *BRCA2* deleterious mutations, using the MASTR 1.2 kit. Seven of the 63 variants tested were missed by AVA because of their nature (seven frameshifts of one and two bp); this gives an experimental sensitivity of 89%. With the SeqNext software, all variations could be detected. Hence, based on these results, an analytical sensitivity of 100% is put forward for the latter (maximum sensitivity: 95% (95% CI)).[17] To calculate the specificity of this assay, data from seven of the 24 samples, which had been completely analyzed by Sanger sequencing, were used. This represented 1183 amplicons (7 × 169), but one amplicon was known to contain a mutation. At a cut-off of 20% and a minimum sequencing depth of 25×, 104 variants were called that were not detected by Sanger sequencing. Hence, the specificity is calculated at 104 false-positive results in 1182 amplicons, or 91%. All 104 false-positive results were located in homopolymer stretches. An overview of the software tools and their specificity and sensitivity is presented in Table 3.

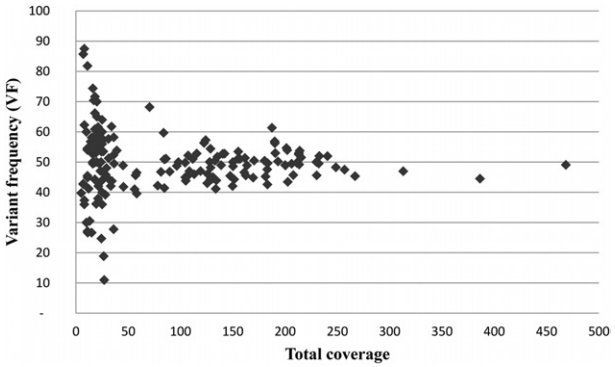

Figure 1. Distribution of the variant frequency (VF) of heterozygous variations versus the total coverage (n = 91).

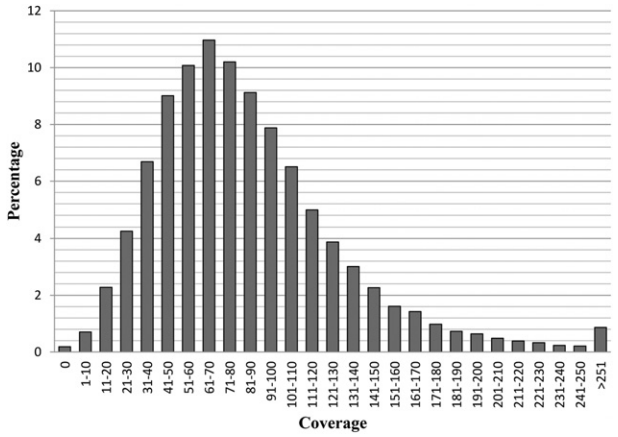

Figure 2. Coverage distribution (up to 250×), ie, the total number of reads (forward and reverse), of each amplicon in the MASTR assay v.1.2 (n = 400 patients). For example, 6.7% of all of the amplicons had a coverage between 31× and 40×. Of the amplicons, 87% had more than 40× coverage, 0.9% of all amplicons had a coverage greater than 250×.

To evaluate the effect of the coverage on the variant frequency, samples were also analyzed in duplicate: once at a very low coverage, and once following a repeated analysis in which the expected average coverage was reached. In Figure 1, the variant frequency of 91 substitutions (27 distinct polymorphisms) was plotted according to the coverage (range 5× to 469×). As expected, the variant frequency came closer to 50% with a higher coverage. The outliers obtained were a VF of 88 with a coverage of 8×, and surprisingly, a VF of 11 associated with a coverage of 27×. In the latter example, detailed analysis showed that the polymorphism was present in only 6% (1/17) of the reads in the forward direction and in 20% (2/10) in the reverse direction.

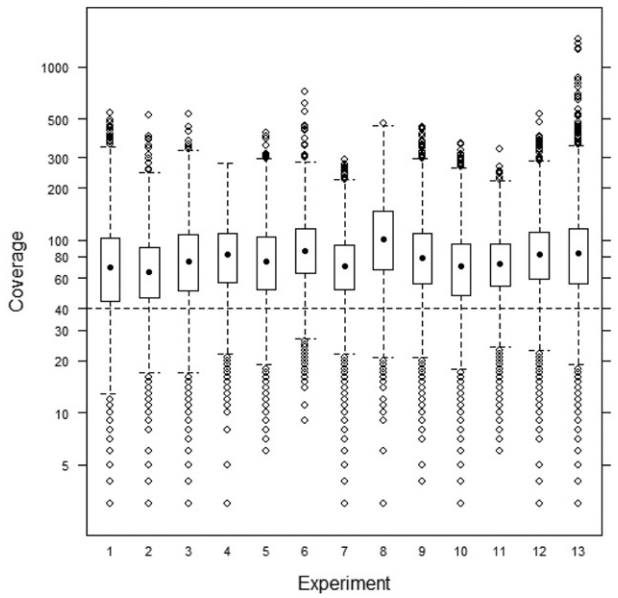

Figure 3. Number of reads (log2 scale) of the *BRCA* MASTR assay v.1.2 (n = 13 experiments and 400 samples). The bottom and top of the box are at the 25th and 75th percentile (the lower and upper quartiles, respectively), and the 50th percentile (the median) is shown in the box. The ends of the whiskers represent the lowest value within 1.5 IQR of the lower quartile, and the highest value within 1.5 IQR of the upper quartile; the outliers are marked by a diamond; 40× coverage is indicated by a dotted line.

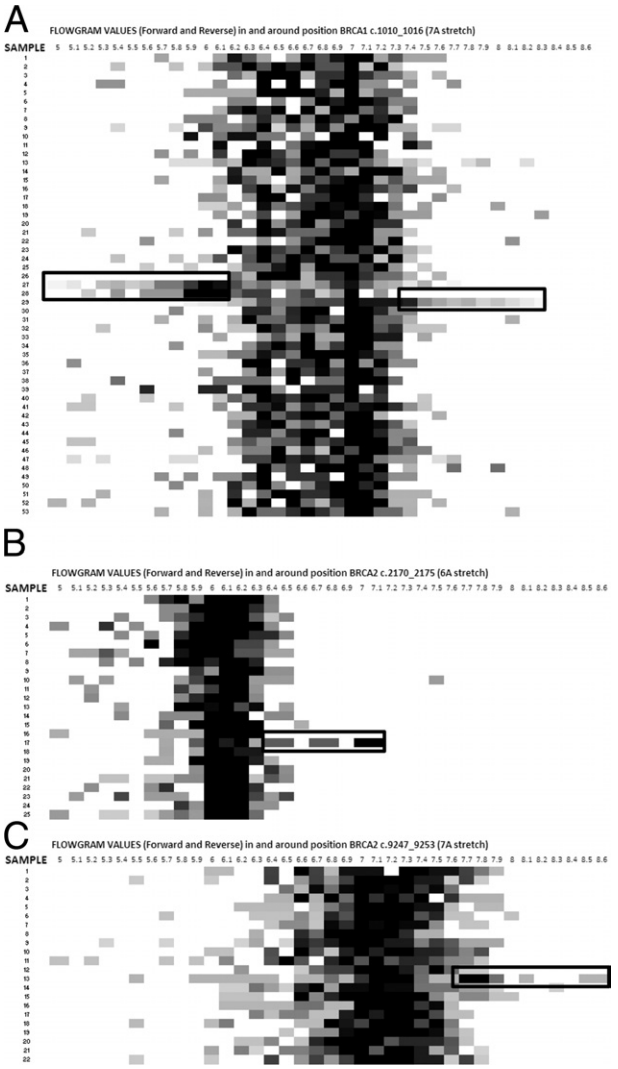

Figure 4. Distribution of the flowgram values in and around the references at positions *BRCA1* c.1010_1016, *BRCA2* c.2170_2175 and *BRCA2* c.9247_9253. Relative ratios (%) were calculated on basis of the number of reads of each interval versus the total amount of reads that were mapped to this position. The values were highlighted in a black-and-white scale (black indicates high, white indicates low). Depending on the stretch, the expected nominal values are 6 or 7 and in the context of a perfect interpretation (100% of reads identical to the wild-type), only black boxes should be present at the respective reference positions. Control samples and five positive controls (**A:** c.1016delA (2 samples) and c.1016dupA; **B:** c.2175dupA and **C:** c.9253dupA) were analyzed. These plots show that, although the distribution of the reads in or around the stretch is sometimes quite broad, the distribution of the flow values in positive controls can be clearly distinguished (see black frames).

Afterward, 400 new samples were analyzed in 13 experiments. These runs were performed in the same conditions (semiautomated workflow, pooling of multiplexes and samples). The coverage distribution showed that a mean coverage of 84× was obtained and that 87% of the amplicons reached a 40× coverage (Figure 2). The analysis of the sequencing depth also showed that the variation, in terms of coverage, was important among and between amplicons.

The coverage data were also plotted per experiment to give an indication of the reproducibility (Figure 3). Over-

Table 4. Variations Located Near or in Homopolymer Stretches, and Results of Mutation Detection with SeqNext Software

Gene	Variation	Sequence context	Observed variant frequency*
BRCA1	c.1016delA	7A >6A	52
BRCA1	c.1016dupA	7A >8A	35
BRCA1	c.3331_3334del4	2 nt after 6A	41
BRCA1	c.3891del3	2 nt after 5T	60
BRCA2	c.927A>G (p.Ser309Ser)	2 nt after 5T	47
BRCA2	c.1281C>T (p.Asp427Asp)	3 nt after 5A	47
BRCA2	c.1804G>A (p.Gly602Arg)	2 nt before 8A	44
BRCA2	c.2175dupA	6A >7A	50
BRCA2	c.4301A>T (p.Lys1434Ile)	in stretch of 5A	49
BRCA2	c.4585G>A (p.Gly1529Arg)	3 nt before 6A	40
BRCA2	c.4935delA	5A >4A	50
BRCA2	c.5353delA	2 nt after 7A	43
BRCA2	c.5645C>A (p.Ser1882X)	C>A, 5A >6A	40
BRCA2	c.7544delC	after 5A	45
BRCA2	c.9253dupA	7A >8A	33

*Ratio, in percent (%), of the reads containing the variation versus the total number of reads.

all, the total number of passed reads in the AVA software was reproducible and the results were in accordance with the manufacturer's specifications (360 10^3–520 10^3 reads per half region). Outliers (amplicons with low (<40×) and high coverage (>200×), were identified in each experiment.

A novel multiplex assay (MASTR 2.0, Multiplicom) containing 94 amplicons instead of 169, was also tested. The major changes were the decrease of the number of amplicons by increasing their length that better fits with the Titanium chemistry and the distribution of the amplicons within the multiplexes. All amplicons were successfully amplified. Furthermore, >99.9% of the amplicons reached the 40× coverage (only two in 3008 amplicons were below). Thus, the MASTR 2.0 assay easily allowed to pool 16 patients samples in one-quarter region of a GS-FLX run (Table 2).

We also focused on mutations in stretches by analyzing the pyrosequencing data in more detail, according to the procedure developed by Y. Bidet and colleagues (personal communication, 2012). A real heterozygous deletion or insertion should generate two populations of read data, whereas for homozygous (normal) reads the signal intensity should cluster around the nominal value.

Figure 4 shows representative examples of the distribution of homopolymer reads in three stretches of adenosines for which positive controls were available: c.1016delA (7A >6A) and c.1016dupA (7A >8A) in *BRCA1*; c.2175dupA (6A >7A) and c.9253dupA (7A >8A) in *BRCA2*. For each homopolymer, "heatmaps" were generated (see *Materials and Methods*) where the proportion of the reads (%) according to their read length was plotted. The results showed that it is possible to distinguish the deletion and the duplications, by a left or right shift, respectively, from the read distribution for the samples containing the mutations. A list of all mutations located near or in homopolymers is presented in Table 4.

Discussion

Sequencing of DNA samples using NGS seems to be easy and straightforward. Nonetheless, numerous parameters affect the quality of the final result. Hence, in view of the accuracy and reproducibility of a diagnostic analysis and the possibility to obtain an accreditation for the diagnostic application of NGS, all of these parameters need to be chosen carefully and maintained. For resequencing the *BRCA* genes on the NGS platform, we have chosen amplicon-based sequencing because this allowed to grossly maintain the current design used for mutation scanning and Sanger sequencing. The major parameters for evaluation were the sequence coverage (number of reads) per amplicon at the end of the procedure and the percentage of mutant versus wild-type reads for heterozygous mutations. As the pathogenic variations in the *BRCA* genes are generally heterozygous (polymorphic variations could be homozygous) and should give a variant frequency of 50%, a good balance between efficiency (affecting the sequencing capacity) and sensitivity has to be chosen to avoid false-positive or false-negative results.

First, a panel of 37 *BRCA1* known variants was tested to evaluate the capacity of the software to accurately call and identify the mutations. The variant frequency obtained in this study was around 50%, as expected, but the range was quite broad (24 to 59 for AVA and 22 to 62 for SeqNext). The discordance between expected and observed values was also described by Bonnal et al.[18] They showed that, for homozygous variations, the expected VF of 100% decreased to 60% for some cases, mainly because of misalignments. Similar observations were made by De Leeneer et al.[16]

Secondly, multiplex amplification was evaluated using the MASTR assays. The coverage results showed a bothersome inter- and intra-sample variation among the amplicons. Nevertheless, the multiplex–polymerase chain reaction approach rendered the sequencing process very efficient, and the sequencing depth has satisfactorily been improved with the newer multiplex design.

The sensitivity of the method was 95% (95% CI). The specificity was determined at 91%. All false-positive results were located in or near homopolymers, and were likely due to pyrosequencing errors. The high rate of

false-positive results in homopolymers is highly similar to that described by De Leeneer et al.[16]

Error rates in homopolymer stretches are well known in pyrosequencing,[19] and in practice, it is very difficult or even impossible to distinguish a mutation from the noise using the current software. We showed that it is possible to detect mutations in stretches by scrutinizing the raw data. We have illustrated this for three stretches, because the comparison can only be achieved when a sample with a true mutation is available. Thus, it would be helpful to collect more information on the reproducibility of these errors to determine the footprint of each stretch. It is unclear whether a variant lying in or near a homopolymer would be missed or whether the result would depend only on other (random) factors. Inclusion of a test such as the one that we presented in the analysis pipeline would greatly improve the calling of insertions and deletions in homopolymers. Until then, the quality of the results may not be accurate enough to reliably use the GS-FLX method in diagnostics. In the meantime, it is advisable to analyze the homopolymers with alternative methods (Sanger sequencing or mutation scanning methods).

For the time being, there is no consensus concerning the minimum coverage required for massively parallel sequencing. De Leeneer et al used a minimum coverage of 38x to accurately detect a heterozygous variation that is present in at least 25% of the reads to obtain a power value of 99.9%, which would correspond to a Phred score of 30.[16] This theoretical calculation is made on the basis of a binomial distribution and our initial cut-off (40×) was based on this calculation. A recalculation showed that a coverage depth of 27× in combination with a filter threshold of 20% is sufficient to obtain a quality value of $P = 99.9\%$.[20] With the implementation of the SeqNext software, we were able to decrease our filter thresholds to 20% ("distinct" table) and to 10% ("other" table), and set a minimum sequencing depth of 25× Indeed, one should not forget that stochastic effects exist and that outliers are always likely, as shown in Figure 1.

For platforms generating higher numbers of reads compared to the GS-FLX system, the required minimal coverage might be higher. For instance, Walsh et al succeeded in sequencing the *BRCA* genes and 19 other breast cancer susceptibility genes on the Illumina Genome Analyzer after capture-based sequencing.[15] The average coverage per base was greater than 1200 reads. Sensitivity and specificity were high. In addition, this approach allowed the detection of exonic deletions and duplications.

Theoretically, it should be possible to detect genomic rearrangements using the GS-FLX platform on basis of the relative ratios of the number of reads. However, because a two-step polymerase chain reaction protocol is used, it is unlikely that the amplifications would still be quantitative, especially at a high number of cycles. Supporting this hypothesis, De Leeneer et al failed to detect exonic deletions in two patients.[16] Although DNA samples with known exonic deletions or duplications were available, we did not include any of those samples in this study. We routinely perform the analysis of *BRCA1* and *BRCA2* genomic rearrangements in patient samples with the MLPA method.

One difficulty in the study was that the methods and software tools are continuously in development. Thus it was not easy, and possibly premature, to establish a final laboratory protocol. Nevertheless, we wish to emphasize that the choice of the software tool is important. At the time of evaluation, the AVA software did not report the 1-bp indels. For the 2- to 3-bp indels, the calls were highly dependent on the sequence context. In addition, because the software tools operate with their own algorithm, there are largely "black boxes" to the user. The greatest risk is that a mutant sequence would be rejected because of poor quality of the alignment, ie, the mismatch or insertion or deletion versus the reference sequence per se. Thus, differences between algorithms might be responsible for important practical issues such as coverage and mutation detection.

For this reason, it was a requisite to evaluate and possibly to include additional software tools in a diagnostic context, whether they are commercial or homemade. With the emergence of the NGS technology, many software packages have been developed for dealing with specific applications. A number of these have been listed in a review by Zhang et al.[14] Clearly, most currently available software tools have been developed primarily for sequence assembly and for the detection of single nucleotide polymorphism substitutions. Much effort has already been expended for the detection of insertions and deletions; however, we believe that the software tools still have to undergo further development. Optimization of software and/or experimental design should preferably be done in collaboration with the manufacturers of the platforms.

In conclusion, *BRCA1* and *BRCA2* resequencing on the Roche GS-FLX System, using an amplicon-based multiplex strategy, works. This study presents an evaluation of the practical issues of NGS for use in diagnostics, on the basis of a relatively high number of samples and experiments. In total, more than 400 DNA samples have been analyzed and 162 variations could be identified (the complete list of the mutations identified in this study is available on request). We are confident about the sensitivity of the method as, indeed, the PCR generated amplicons are efficiently retrieved. Nonetheless, we hope that our study will be the starting point for an international discussion about standards for the validation of NGS-based diagnostic methods.

Practice guidelines for Sanger sequencing analysis and interpretation have (only) recently been published.[21] We also look forward to guidelines that would set the standards both for the analysis and the interpretation of NGS applications in diagnostics, and we believe that the data that we have generated in this study will be useful to set a first, albeit temporary, standard.

Acknowledgments

We thank Tina Overloop, Ilse Van Der Heyden, Liesbet Vliegen, and Vanessa Brys for excellent technical support and Chris Mattocks for advice on the statistics.

References

1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM: Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437:376–380
2. De Leeneer K, Coene I, Poppe B, De Paepe A, Claes K: Rapid and sensitive detection of BRCA1/2 mutations in a diagnostic setting: comparison of two high-resolution melting platforms. Clin Chem 2008, 54:982–989
3. Ford D, Easton DF, Stratton M, Narod S, Goldgar D, Devilee P, Bishop DT, Weber B, Lenoir G, Chang-Claude J, Sobol H, Teare MD, Struewing J, Arason A, Scherneck S, Peto J, Rebbeck TR, Tonin P, Neuhausen S, Barkardottir R, Eyfjord J, Lynch H, Ponder BA, Gayther SA, Zelada-Hedman M and the Breast Cancer Linkage Consortium: Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. Am J Hum Genet 1998, 62:676–689
4. Eng C, Brody LC, Wagner TM, Devilee P, Vijg J, Szabo C, Tavtigian S, Nathanson K, Ostrander E, Frank T; Steering Committee of the Breast Cancer Information Core (BIC) Consortium: Interpreting epidemiological research: blinded comparison of methods used to estimate the prevalence of inherited mutations in BRCA1. J Med Genet 2001, 38:824–833
5. Wagner T, Stoppa-Lyonnet D, Fleischmann E, Muhr D, Pagès S, Sandberg T, Caux V, Moeslinger R, Langbauer G, Borg A, Oefner P: Denaturing high-performance liquid chromatography detects reliably BRCA1 and BRCA2 mutations. Genomics 1999, 62:369–376
6. van der Hout AH, van den Ouweland AM, van der Luijt RB, Gille HJ, Bodmer D, Brüggenwirth H, Mulder IM, van der Vlies P, Elfferich P, Huisman MT, ten Berge AM, Kromosoeto J, Jansen RPM, van Zon PHA, Vriesman T, Arts N, Boutmy-de Lange M, Oosterwijk JC, Meijers-Heijboer H, Ausems HGEM, Hoogerbrugge N, Verhoef S, Halley DJJ, Vos YJ, Hogervorst F, Ligtenberg M, Hofstra RMW: A DGGE system for comprehensive mutation screening of BRCA1 and BRCA2: application in a Dutch cancer clinic setting. Hum Mutat 2006, 27:654–666
7. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G: Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 2002, 30:e57
8. Hogervorst FBL, Nederlof PM, Gille JJP, McElgunn CJ, Grippeling M, Pruntel R, Regnerus R, van Welsem T, van Spaendonk R, Menko FH, Kluijt I, Dommering C, Verhoef S, Schouten JP, van'T Veer LJ, Pals G: Large genomic deletions and duplications in the BRCA1 gene identified by a novel quantitative method. Cancer Res 2003, 63:1449–1453
9. Agata S, Dalla Palma M, Callegaro M, Scaini MC, Menin C, Ghiotto C, icoletto O, Zavagno G, Chieco-Bianchi L, D'Andrea E, Montagna M: Large genomic deletions inactivate the BRCA2 gene in breast cancer families. J Med Genet 2005, 42:e64
10. Puget N, Stoppa-Lyonnet D, Sinilnikova OM, Pages S, Lynch HT, Lenoir GM, Mazoyer S: Screening for germ-line rearrangements and regulatory mutations in *BRCA1* led to the identification of four new deletions. Cancer Res 1999, 59:455–461
11. Stratton MR, Rahman N: The emerging landscape of breast cancer susceptibility. Nat Genet 2008, 40:17–22
12. Sevilla C, Moatti JP, Julian-Reynier C, Eisinger F, Stoppa-Lyonnet D, Bressac-de Paillerets B, Sobol H: Testing for BRCA1 mutations: a cost-effectiveness analysis. Eur J Hum Genet 2002, 10:599–606
13. Goossens D, Moens LN, Nelis E, Lenaerts AS, Glassee W, Kalbe A, Frey B, Kopal G, De Jonghe P, De Rijk P, Del-Favero J: Simultaneous mutation and copy number variation (CNV) detection by multiplex PCR-based GS-FLX sequencing. Human Mutat 2008, 30:472–476
14. Zhang J, Chiodini R, Badr A, Zhang G: The impact of next-generation sequencing on genomics. J Genet Genomics 2011, 38:95–109
15. 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 USA 2010, 107: 12629–12633
16. De Leeneer K, Hellemans J, De Schrijver J, Baetens M, Poppe B, Van Criekinge W, De Paepe A, Coucke P, Claes K: Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. Human Mut 2011, 32:1–10
17. Mattocks CJ, Morris MA, Matthijs G, Swinnen E, Corveleyn A, Dequeker E, Müller CR, Pratt V, Wallace A; EuroGentest Validation Group: A standardized framework for the validation and verification of clinical molecular genetic tests. Eur J Hum Genet 2010, 18:1276–1288
18. Bonnal RJP, Severgnini M, Castaldi A, Bordoni R, Iacono M, Trimarco A, Torella A, Piluso G, Aurino S, Condorelli G, De Bellis G, Nigro V: Reliable resequencing of the human dystrophin locus by universal long polymerase chain reaction and massive pyrosequencing. Anal Biochem 2010, 406:176–184
19. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 2007, 8:R143
20. 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, Coucke P, Vandesompele J, Claes K, Hellemans J: Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. Plos One 2011, 6:e25531
21. Ellard S, Charlton R, Yau M, Gokhale D, Taylor GR, Wallace A, Ramsden SC. Practice Guidelines for Sanger Sequencing Analysis and Interpretation. Copyright Clinical Molecular Genetics Society, Birmingham, UK. Available at *http://www.cmgs.org/BPGs/pdfs%20current%20bpgs/Sequencingv2.pdf*. Approved August 7, 2009