European Journal of Human Genetics (2013) 21, 864–870

www.nature.com/ejhg

ARTICLE

Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of *BRCA1* and *BRCA2* genes

Lídia Feliubadaló[1,6], Adriana Lopez-Doriga[2,6], Ester Castellsagué[1,6], Jesús del Valle[1], Mireia Menéndez[1], Eva Tornero[1], Eva Montes[1], Raquel Cuesta[1], Carolina Gómez[1], Olga Campos[1], Marta Pineda[1], Sara González[1], Victor Moreno[3], Joan Brunet[4], Ignacio Blanco[1], Eduard Serra[5], Gabriel Capellá[1] and Conxi Lázaro*[,1]

Next-generation sequencing (NGS) is changing genetic diagnosis due to its huge sequencing capacity and cost-effectiveness. The aim of this study was to develop an NGS-based workflow for routine diagnostics for hereditary breast and ovarian cancer syndrome (HBOCS), to improve genetic testing for *BRCA1* and *BRCA2*. A NGS-based workflow was designed using BRCA MASTR kit amplicon libraries followed by GS Junior pyrosequencing. Data analysis combined Variant Identification Pipeline freely available software and *ad hoc* R scripts, including a cascade of filters to generate coverage and variant calling reports. A *BRCA* homopolymer assay was performed in parallel. A research scheme was designed in two parts. A Training Set of 28 DNA samples containing 23 unique pathogenic mutations and 213 other variants (33 unique) was used. The workflow was validated in a set of 14 samples from HBOCS families in parallel with the current diagnostic workflow (Validation Set). The NGS-based workflow developed permitted the identification of all pathogenic mutations and genetic variants, including those located in or close to homopolymers. The use of NGS for detecting copy-number alterations was also investigated. The workflow meets the sensitivity and specificity requirements for the genetic diagnosis of HBOCS and improves on the cost-effectiveness of current approaches.

European Journal of Human Genetics (2013) **21,** 864–870; doi:10.1038/ejhg.2012.270; published online 19 December 2012

Keywords: Next-generation sequencing; hereditary breast and ovarian cancer syndrome; *BRCA1*; *BRCA2*; genetic testing; molecular diagnostics

INTRODUCTION

Next-generation sequencing (NGS) is an increasingly used technology that generates up to gigabases of DNA reads at high speed and with low cost per base. This high-throughput technology, based on massively parallel sequencing of spatially separated DNA molecules, is currently used with several available platforms, such as the Genome Sequencer (Roche-454 Life Sciences, Indianapolis, IN, USA), the Genome Analyzer/HiSeq/MiSeq (Illumina-Solexa, San Diego, CA, USA), the SOLiD System, Ion PGM/Ion Proton (Ion Torrent-Invitrogen, Carlsbad, CA, USA), and the HeliScope from Helicos BioSciences (Cambridge, MA, USA).[1,2] In Roche-454 technology, bead-attached DNA fragments clonally amplified in a water-in-oil emulsion (emulsion PCR) are deposited in single-bead capacity wells of a plate over which nucleotides flow sequentially, releasing chemiluminescence only when a nucleotide is correctly incorporated (pyrosequencing). In molecular diagnostics, targeted genomic resequencing of pooled samples from different individuals benefits from the high throughput achieved by using NGS technology. To enrich the target fragments to be resequenced in this type of gene-centric approach, PCR-based methods are generally used.[3,4] *BRCA1* and *BRCA2* are the two main highly penetrant genes that predispose to hereditary breast and ovarian cancer syndrome (HBOCS).[5] Molecular diagnosis of HBOCS is essential for the provision of genetic counseling and to establish preventive screening and therapeutic strategies.[6] Although direct Sanger sequencing is considered the gold standard for the analysis of *BRCA1* and *BRCA2* mutations, their large size (5592 bp and 10257 bp, respectively), and lack of mutation hot spots (see Breast Cancer Information Core database: http://www.research.nhgri.nih.gov/bic/) mean useful prescreening strategies.[7–9] Moreover, large genomic rearrangements (LGRs) of these genes require the use of other complementary techniques.[10,11] The development of cost-effective *BRCA* mutation detection workflows will not only benefit the genetic counseling process for patients with HBOCS but will also enhance the process of selecting patients for personalized treatments, as could be the case of PARP inhibitors, for example. Mutation analyses of *BRCA1* and *BRCA2* using NGS have been already performed for high-capacity NGS platforms, such as the 454 FLX (Roche),[12] the Helicos (Heliscope),[13] the Genome Analyzer (Illumina)[4] and, very recently, the GS Junior instrument.[14] Most of these studies used large-capacity

[1]Hereditary Cancer Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; [2]Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; [3]Prevention Program, Catalan Institute of Oncology (ICO-IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain; [4]Hereditary Cancer Program, Catalan Institute of Oncology (ICO-IdIBGi), Girona, Spain; [5]Institut de Medicina Predictiva i Personalitzada del Càncer (IMPPC), Badalona, Barcelona, Spain
[6]These authors contributed equally to this work.
*Correspondence: Dr C Lázaro, Hereditary Cancer Program, Molecular Diagnosis Unit, Laboratori de Recerca Translacional, Institut Català d'Oncologia (ICO-IDIBELL), Hospital Duran i Reynals, Gran Via 199-203, L'Hospitalet de Llobregat, 08908 Barcelona, Spain. Tel: +34 932607342; Fax: +34 932607466; E-mail: clazaro@iconcologia.net
Received 9 July 2012; revised 28 September 2012; accepted 13 November 2012; published online 19 December 2012

platforms that generally exceed the demand of most mid-sized genetic testing laboratories and whose approaches are difficult to translate to benchtop next-generation sequencers. Only one of the studies used small-scale equipment, the GS Junior, but the number of samples tested is very small and no discussion is offered regarding how to overcome the main problem associated with pyrosequencing, that is, DNA lectures in homopolymeric regions.[14] Here, we present a rigorous sensitivity and specificity analysis of our newly established HBOCS workflow for genetic testing of *BRCA* genes using a small-capacity next-generation instrument. We present data from a Training Set and from a Validation Set of samples. We demonstrate that a combined approach using the GS Junior platform and an specific assay for homopolymeric tracts with a custom bioinformatics pipeline provides accurate results that can be used for genetic diagnosis.

MATERIALS AND METHODS

Samples analyzed

In our unit, a multistep workflow including conformation-sensitive capillary electrophoresis[9] as a prescreening method for analysis of *BRCA* mutations was used (Supplementary Figure 1). A total of 28 DNA samples previously characterized by this workflow were used as a Training Set to setup our NGS workflow, and 14 new DNAs were used as a Validation Set (see Experimental design in the Results section). To properly compare NGS with our workflow, only variants in heterozygosity were considered (as homozygous variants are not detected by conformation-sensitive capillary electrophoresis). This study was approved by our Institutional Review Board.

Multiplex PCR-based target amplification and resequencing

Target amplification of *BRCA1* and *BRCA2* was achieved using BRCA MASTR assays following manufacturer's instructions (http://www.multiplicom.com). Several versions of the kit were used as they were released. Briefly, the assay generates a library of specific amplicons in two rounds of PCR: a first multiplex PCR that amplifies the target sequences; and a second PCR to attach MID (Multiplex Identifier) barcodes and 454 adapters to each amplicon. The barcoded multiplex products were assessed by fluorescent labeling and capillary electrophoresis, and quantified using Quant-iT PicoGreen (Invitrogen). Then, PCRs from different individuals were equimolarly pooled and purified using AgencourtAMPure XP (Beckman Coulter, Beverly, MA, USA) and PicoGreen quantified. Emulsion PCR of the combined purified libraries was carried out using the GS Junior Titanium emPCR Kit (Lib-A) and pyrosequenced on GS Junior following manufacturer's instructions (Roche).

Data analysis

Reads from the GS Junior sequencer were analyzed with the open source software Variant Identification Pipeline (VIP) version 1.4.[15] Using VIP, the reads from each sample were demultiplexed and then aligned against *BRCA1* NG_005905.2 and *BRCA2* NG_012772.1 reference sequences using the BLAT algorithm.[16] Results from VIP were then processed using R (A Language and Environment for Statistical Computing) commands. Specific primers from each amplicon were trimmed and identified variants were annotated according to the Human Genome Variation Society (HGVS) nomenclature recommendations version 2.0 (http://www.hgvs.org/mutnomen/). Two reports were obtained: a coverage report, listing low-coverage fragments indicated for further Sanger sequencing; and a variant report. Intronic variants located deep inside introns (after position +20 of the donor site and before position −50 of the acceptor site) were not included in the variant report. Multiple alignments of reads for each MID and amplicon were visualized with the GS Amplicon Variant Analyzer v2.7 (AVA) software (Roche). Scripts are available upon request (Lopez-Doriga *et al*, manuscript in preparation).

We also evaluated the capacity to detect LGRs. Eight samples with known rearrangements were tested in three different runs. One of the samples was included in the Validation Set, and the other seven were added later. The known LGRs consist of: deletion of exons 1–2, deletion of exons 1–13, deletion of exon 14, deletion of exon 20, deletion of exon 22, and duplication of exons 9–24 in *BRCA1*, and deletion of exons 1–24 and deletion of exon 2 in *BRCA2*. To assess copy number for each amplicon, a methodology described elsewhere was applied.[3] Briefly, the relative read count of an amplicon was determined as the ratio of the read count for that amplicon over the sum of all gene amplicons for the other gene in the specific multiplex to which the amplicon belongs. Hence, to analyze *BRCA1* amplicons, we used the sum of *BRCA2* amplicons from the same multiplex, and vice versa. Next, intersample normalization was performed, dividing each ratio by the average of the control samples in the same experiment (at least three controls were used).

Homopolymer analysis

To treat homopolymers, the BRCA HP v2.0 (Multiplicom, Niel, Belgium) assay was used. This kit targets all *BRCA1*- and *BRCA2*-coding homopolymer stretches of 6 bp or longer by producing 29 PCR products in two multiplex reactions. Fragment length was assessed by capillary electrophoresis (3730 ABi sequencer, Applied Biosystems, Foster City, CA, USA) and visualized with the MAQ-S software (Multiplicom).

Sanger sequencing

All fragments with coverage under 38× and all non-polymorphic DNA variants identified were sequenced by Sanger.

RESULTS

Experimental design

The Training Set (28 samples analyzed in two experiments) contained 23 unique pathogenic mutations and 204 (33 unique) non-pathogenic mutations or mutations with unknown significance DNA variants (Supplementary Table 1) (Figure 1). In the Validation Set, 14 samples were blindly sequenced together with a sample containing a multi-exon duplication in *BRCA1* (Figure 1). To better assess the usefulness of this approach to detect LGR, a set of seven positive samples showing LGRs were also analyzed.

Workflow setup

In experiment 1, 28 samples were amplified with the BRCA MASTR v1.2 kit (170 amplicons, Multiplicom) in four GS Junior runs (R1-R4) (7 patients per run). Only 0.5% of the passed reads was lost, due to short length, low quality or incorrect MIDs or primer sequences, and did not map in the reference sequence. While experiment 1 was being conducted, Multiplicom released a new kit (v2.0, 94 amplicons), which was used in experiment 2 to reanalyze 14 samples from experiment 1 in two runs (R5–R6).

Coverage analysis of the Training Set

The coverage of each run was evaluated (Table 1). In experiment 1, the average mean base coverage was 69 ± 27. The coverage for the various MIDs used (MID1–MID15) did not exhibit any significant difference (data not shown). The number of mapped reads in R5 and R6 was similar to the runs in experiment 1, but coverage was substantially increased (127 ± 53) due to the lower number of amplicons. Of the 24 undercovered amplicons (coverage <38), 14 belonged to amplicon *BRCA1*_exon7 from different patients (Supplementary Figure 2A).

Filters and variant calling in the Training Set

Next, identification of all the variants was investigated. First, each experiment was analyzed alone (data not shown), then the results were combined as the Training Set, incorporating into experiment 2 samples not repeated from experiment 1 (to avoid bias due to duplication of samples). In total, 4260 variants were identified, of

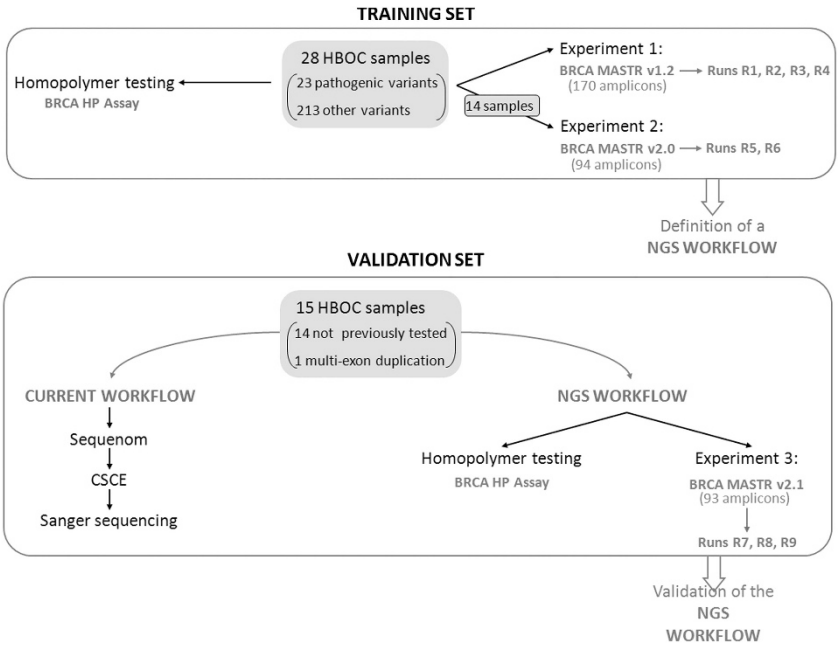

Figure 1 Experimental design. Our study was divided into two parts: the Training Set and the Validation Set. In the Training Set, 28 HBOCS samples, already analyzed by our current diagnostic workflow, were assessed (Supplementary Figure 1). Of this group, 23 samples contained a variety of pathogenic mutations, including challenging insertions and deletions, inside and outside homopolymeric regions, as well as a subset of non-pathogenic variants. The remaining 5 samples belonged to affected individuals from high-risk HBOCS families, in whom no pathogenic mutation had been found after applying our current multistep protocol. In total, this subset of 28 samples contained 23 unique pathogenic mutations and 213 (33 unique) non-pathogenic DNA variants (Supplementary Table 1). The Training Set was subjected to two different experiments: in experiment 1, all 28 samples were amplified using BRCA MASTR v1.2 and sequences in 4 runs; in experiment 2, 14 of the DNAs from experiment 1 were used but they were amplified with the newly released kit (v2.0) and sequenced in two runs. In parallel, homopolymeric regions of all samples were studied with the BRCA HP kit. Thanks to the Training Set experiment, we were able to define an NGS workflow for the genetic analysis of *BRCA* genes in the HBOCS diagnostic routine. In the Validation Set, we assessed a total of 15 HBOC samples, 14 not previously tested and the remaining 1 containing a multi-exon duplication. These samples were analyzed in parallel with our current diagnostic workflow and with the newly designed NGS workflow. In this case, experiment 3 was carried out using the most recent version (v2.1) of the BRCA MASTR kit and samples were sequenced in three runs.

Table 1 Overall coverage results

	Experiment 1 (BRCA MASTR v1.2)				*Experiment 2 (BRCA MASTR v2.0)*		*Experiment 3 (BRCA MASTR v2.1)*		
Run	*R1*	*R2*	*R3*	*R4*	*R5*	*R6*	*R7*	*R8*	*R9*
Samples	7	7	7	7	7	7	5	5	5
Passed reads	106 699	71 391	77 696	98 227	76 860	91 653	89 102	111 668	83 076
BRCA-mapped reads	106 303	70 953	77 339	97 778	76 559	91 421	88 699	110 724	82 718
(% of passed)	(99.6%)	(99.4%)	(99.54%)	(99.54%)	(99.6%)	(99.75%)	(99.5%)	(99.15%)	(99.5%)
Coverage, mean	81.8	50.9	62.7	81.6	115	138	216	269	202
[min, max]	[5,201]	[0,133]	[8,157]	[0,200]	[5,498]	[6,494]	[43,595]	[51,807]	[47,610]
Coverage SD	31.3	21.6	23.77	31.7	49.5	55.8	91.36	107.8	85.26
Coverage fold difference to mean ratio 90%/95%	1.98/2.77	1.95/2.39	1.86/2.34	1.91/2.23	1.81/2.11	1.82/2.43	1.68/2.09	1.49/1.74	1.69/2.04
No. of bases< 38	9430	28 947	15 238	5680	2895	3696	0	0	0
(% of mapped)	(8.2%)	(25.2%)	(13.3%)	(4.9%)	(1.78%)	(2.28%)			
No. of fragments< 38	106	318	178	74	10	14	0	0	0

which 223 were true positives (TP) and 4037 were false positives (FP). The high proportion (95%) of FPs identified by the NGS platform after alignment and raw variant calling means that filters are required. To discard false positives, six filters were assessed as follows (Table 2):

(1) Insertions and deletions covered by the BRCA HP assay. This filter is used to reduce the number of FP of insertions or deletions, caused by HP of 6 bp or longer (targetted by the assay), but also by HP of 5 bp (many of them covered by the BRCA HP assay PCRs). This filter discarded 1730 FP and 11 TP. All these 11 TP, plus one variant not detected by VIP (*BRCA1* c.1961delA, in a homopolymer of 8 As), were found by the HP kit, which demonstrated to be clear and completely reliable detecting length changes.

(2) Variants in regions with coverage below 38 × were considered undercovered and thus Sanger sequenced. This coverage threshold was based on De Leeneer's calculations, according to which this number of reads would allow to find an heterozygous variant for a minimum frequency of 25% with a power of 99.9%. This sensitivity is equivalent to a Phred score of 30.[17] This filter discarded 97 FP and 10

Table 2 Cumulative application of filters

		1[a]*: Ins/del BRCA HP*		*1→2*[b]*: Cov<38*		*(1+2)→3*[c]*: VAF<0.25*		*(1+2+3)→4*[d]*: Fcov=0 or Rcov=0*		*(1+2+3)→5*[e]*: FQ<30 & RQ<30*		*(1+2+3)→6*[f]*: Total Q <30*	
	Before filters	*In*	*Out*	*In*	*Out*	*In*	*Out*	*In*	*Out*	*In*	*Out*	*In*	*Out*
Training Set													
FP	4037	2307	1730	2210	97	512	1698	9	503	228	284	227	285
TP	223	212	11	202	10	202	0	200	2	201	1	200	2
Sensitivity		0.951		0.953		1.000		0.990		0.995		0.990	
Specificity		0.429		0.042		0.769		0.982		0.555		0.557	
Validation Set													
FP	1471	872	599	872	0	168	704	3	165	59	109	59	109
TP	123	122	1	122	0	122	0	121	1	122	0	122	0
Sensitivity		0.992		1.000		1.000		0.992		1.000		1.000	
Specificity		0.407		0.000		0.807		0.982		0.649		0.649	

Variants retained (In) and discarded (Out) by the application of:
[a]filter 1: insertion or deletion covered by the BRCA HP assay.
[b]filter 2: coverage below 38, to variants retained by filter 1.
[c]filter 3: variant allele frequency below 0.25, to variants retained by filters 1 and 2.
[d]filter 4: variant forward coverage or variant reverse coverage equal to 0, to variants retained by filters 1, 2, and 3.
[e]filter 5: variant forward quality and variant reverse quality below 30, to variants retained by filters 1, 2, and 3.
[f]filter 6: total variant quality below 30, to variants retained by filters 1, 2, and 3.

TP in the Training Set, all of them were confirmed by the subsequent Sanger sequencing.

(3) Variants with an allele frequency <25% were disregarded. This filter discarded 1698 additional FP for the Training Set but not any TP.

(4) Variants detected in only one strand. This filter, indicated by VIP as the variant having forward coverage or reverse coverage equal to 0, discarded 503 FP and 2 TP (additionally to filters 1+2+3).

(5) Variants with forward and reverse variant mean qualities below 30.[12] This filter discarded 284 FP and 1 TP (additionally to filters 1+2+3).

(6) Variants with total quality below 30. This filter was very similar to filter 5 but differed in some variants, so it was tested to compare with filters 4 and 5. It discarded 285 FP and 2 TP (additionally to filters 1+2+3).

We observed that the application of the first three filters did not lead to the loss of any true mutation. These filters also lowered the number of FP from 4037 to 512 (Supplementary Figure 3). Filters 4–6 (variants detected in only one strand; variants with variant mean quality in forward and reverse below 30; variants with total quality below 30) resulted in the loss of 1 or 2 TP out of 28 samples, which is not acceptable in a *BRCA* diagnostic setting. If these filters were not used, Sanger sequencing of 512 FP and the 29 TP (23 pathogenic and 6 unknown significance variants, see Supplementary Table 1) would be needed to provide robust results, considerably increasing the cost and time of the workflow. Consequently, we opted for an intermediate strategy that consisted in using filter 4 (variants detected in only one strand) to generate a list of variants for which visual inspection of the aligned region was required. Filter 4 was chosen because it filtered most of the remaining FP (Table 2). Supplementary Figure 4 uses Venn diagrams to show the common and different FP and TP that filters 4, 5 and 6 would discard. Visualization was performed using the Amplicon Variant Analysis (AVA, Roche) software, permiting to discard artifactual variants present only in one strand, while keeping real variants that were wrongly aligned in different positions in both strands. This manual analysis discarded 501 FP and 0 TP, leaving 2FP and 2TP for Sanger sequence analysis (Supplementary Figure 3). Analysis of the HP assay detected all of the insertions and deletions that fall between its primers. Sanger sequencing confirmed that all FPs were pyrosequencing errors.

To summarize, in the Training Set we expected to find 227 heterozygous variants. Considering only the variant calling results from GS Junior with the application of 3 filters, we found 202 TP (none of which were discarded by the blind visual inspection); the HP assay detected 12 more, and Sanger sequencing of low-coverage regions identified the remaining 13 TP variants. As expected, FPs decreased with the correlative application of filters and visualization in our workflow design. Only 11 FP required Sanger sequencing to be discarded. These numbers would correspond to an experimental sensitivity and specificity for point mutations of 100% at the last step of our workflow (Table 3). Consequently, complete analysis of the Training Set enabled us to generate a new NGS-based workflow for genetic testing of *BRCA* genes (Figure 2).

Variants in homopolymer sequences

Pyrosequencing of homopolymers presented a technical limitation, as it was difficult to distinguish FP from TP deletions in homopolymer stretches of 6 bp or longer. Therefore, an HP assay is needed. Examples of homopolymer difficulties are shown in Supplementary Figure 5. Some variants in HP of 6 bp or longer are also detected by VIP but the BRCA HP assay is more reliable.

Validation Set

To validate the usefulness and readiness of the pipeline, 14 consecutive samples received for diagnosis of HBOCS were simultaneously analyzed by separate teams using NGS and our current workflow. A fifteenth sample, which bears a pathogenic *BRCA1* mutation as well as a duplication of exons 9–24 of *BRCA1*, was added to test whether copy-number variation could be detected at this coverage. The library for this Validation Set was created using a new version of the BRCA MASTR kit (v2.1), in which the problem of coverage of *BRCA1* exon 7 was solved. To increase coverage, the 15 samples were sequenced in 3 GS Junior runs (R7–R9), 5 samples per run.

The average mean base coverage was 229 ± 95. The average fold difference to mean ratio was 1.62 at the 10th percentile and 1.96 at the

Table 3 Variant calling results

	Training Set				*Validation Set*			
	GS Junior[a]	*+ Visual review*	*+ HP Kit*	*+ Sanger*	*GS Junior*[a]	*+ Visual review*	*+ HP Kit*	*+ Sanger*
True +	202	202	214	227	122	122	123	123
True −	613 161	613 662	613 662	613 673	347 619	347 783	347 783	347 787
False +	512	11	11	0	168	4	4	0
False −	12	12	0	0	1	1	0	0
Variants in low coverage	13	13	13	0	0	0	0	0
Sensitivity	0.88987	0.88987	0.94273	1.00000	0.99187	0.99187	1.00000	1.00000
Specificity	0.99917	0.99998	0.99998	1.00000	0.99952	0.99999	0.99999	1.00000

[a]After applying filters 1 + 2 + 3.

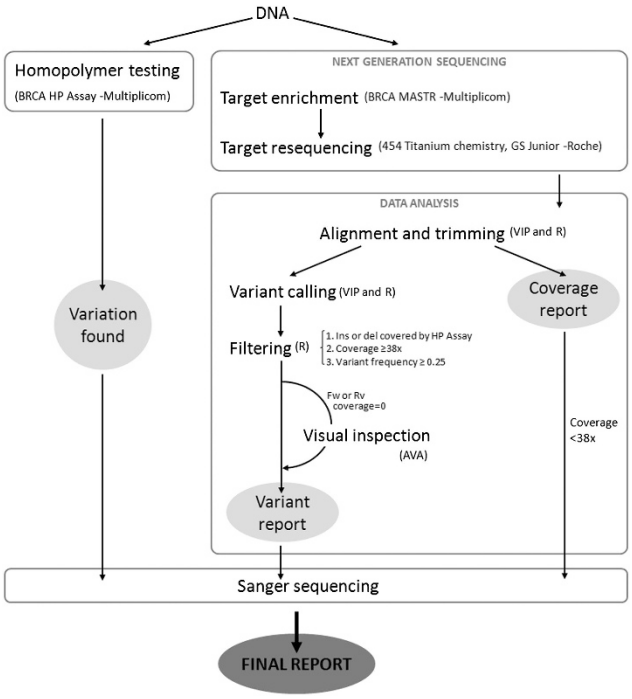

Figure 2 Proposed workflow for analyzing *BRCA1* and *BRCA2* using NGS. A screening using the BRCA HP kit (Multiplicom) allows detection of insertions or deletions located in homopolymers of 6 bp or longer and their surroundings. Sanger sequencing confirms any aberrant pattern found. Simultaneously, DNA samples are analyzed by NGS. *BRCA1* and *BRCA2* coding regions and their intron–exon boundaries are amplified using the BRCA MASTR kit (Multiplicom), adding specific identifiers (MIDs) for each sample to pool them. Sequencing of the enriched regions from pooled samples is performed by using 454 Titanium chemistry in a GS Junior platform (Roche). Data generated by the sequencer are analyzed using the public software VIP and R instructions, which allows us to align all of the sequences generated, trim the surrounding regions of each amplicon (adapters, MIDs and primers) and call putative variants. After filtering the initial variants with filters 1, 2, 3 and 4, a subset (variants with null forward or reverse coverage) is selected for visual inspection of their alignment with AVA, which will discard obvious FPs. All remaining variants are confirmed by Sanger sequencing. As our aim was to integrate this approach into the diagnostic routine, this revision was performed independently by two qualified technicians to generate a common list indicating the decision for any variant under analysis. If a discrepancy arose between the two referees, the most conservative decision was adopted. Regions with low coverage ($<38\times$) are also Sanger sequenced.

5th percentile (Table 1). No bases with coverage under $38\times$ were observed, meaning that Sanger resequencing was unnecessary for low coverage. For example, in experiment R7, all amplicons produced coverage over $50\times$ except amplicon *BRCA1*_ex20.1 in MID1 (Supplementary Figure 2B).

Our analysis algorithm detected 123 heterozygous variants in this set of samples (2 of which were pathogenic). In all, 122 TP (none of which were discarded by the blind visual inspection) were identified by NGS plus filtering, and the remaining TP were detected by the BRCA HP assay. The first three filters reduced FP from 1471 to 168. After the visual alignment review, four FP remained, which were adequately classified after Sanger sequencing. Also for the Validation Set, an experimental sensitivity and an experimental specificity of 100% were achieved by the workflow (Table 3). However, as explained thoroughly in Mattocks *et al*,[18] when the measured sensitivity in the validation of a qualitative test is 100%, a good estimation of the 95% confidence interval should be calculated by the rule of three. As our sample size consists in 123 mutations tested in the Validation Set, our statistical power corresponds to a confidence interval $\geq 97.5\%$.

Large rearrangements detection

A large genomic duplication comprising exons 9–24 of *BRCA1*[19] was included in the Validation Set in run R9. A total of 27 out of 30 amplicons involved in the duplication yielded a dosage quotient value above 1.35, similar to the MLPA results. In addition, the borders of the duplication were quite well defined. To explore the limitations of this analysis in greater depth, we decided to add seven previously identified LGRs showing different deletions and duplications.[19,20] These samples were analyzed in subsequent runs mixed with samples without LGRs. In summary, all LGRs were detected (Figure 3 and Supplementary Figure 6B), duplications showed normalized amplicon values above 1.3 and deletions showed values below 0.7. However, many other amplicons showed values outside these limits (0.7–1.3) representing FPs, which were identified both in control samples (Supplementary Figure 6A) and in other regions of samples showing LGRs. In addition, when very large rearrangements were present in one gene, amplicons from the other gene were affected in the opposite direction due to a bias produced in the normalization process, making it difficult to discriminate real deletions/duplications from FP amplicons.

Cost efficiency

A study of all the consumables and time used, from DNA extraction to obtain the final report, was performed with the aim of comparing

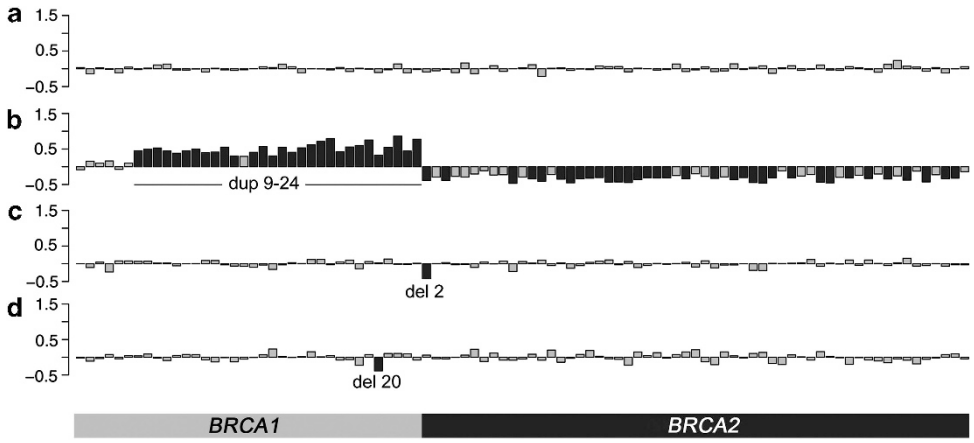

Figure 3 Detection of LGRs using NGS results. Bar plots of the dose of NGS amplicons after normalization. X-axis: NGS amplicons. Y-axis: Count ratio minus 1. Fragments with normalized ratios above 1.3 and below 0.7 are highlighted in black, indicating putative duplications and deletions, respectively. (**a**) Control sample with no alterations. (**b**) Sample with a duplication of the region comprising *BRCA1* exons 9–24. (**c**) Sample with a deletion comprising *BRCA2* exon 2. (**d**) Sample with a deletion comprising exon 20 of *BRCA1*.

our former genetic testing strategy with the new strategy. We found that the overall price of consumables was similar for both approaches (conformation-sensitive capillary electrophoresis + Sanger sequencing *vs* NGS + HP assay + Sanger sequencing), with an estimated cost of €325 in each case. However, the hands-on time and turnaround time were substantially different. By using our proposed NGS workflow, we save 57% of the time cost per technician (down from 14 h/sample to 6 h/sample) and obtain a reduction of ~25% in turnaround time (down from 20 days for 13 samples to 15 days for 14 samples).

DISCUSSION

Here we present a complete workflow for the analysis of the *BRCA1* and *BRCA2* genes, based on the use of a multiplex PCR strategy (Multiplicom) to generate the patient's DNA library followed by pyrosequencing using a benchtop NGS platform (GS Junior) and subsequent bioinformatic analysis based on a combination of three software (VIP, R, and AVA). The analysis of insertions and duplications in homopolymeric regions was performed by an HP assay (Multiplicom). Our results indicate that this workflow achieves an excellent performance for point mutations, with a specificity of 100% and a sensitivity ≥97.5% (95% CI) (Figure 2, Table 2).

Our approach improves previous studies using NGS for BRCA genetic testing in different aspects including: 1) the combination of a Training and a Validation Set, which is the best way to accurately assess the sensitivity of a given approach; 2) the development of a complete algorithm, incorporating the use of the BRCA HP kit, allows us to reach a sensitivity of 100% (≥97.5% with a 95% confidence interval), keeping with an excellent specificity (100%; ≥99.9991% with a 95% confidence interval); and 3) the cost-effective analysis for BRCA analysis in a benchtop NGS platform. Although it seems that improvements on analysis are still needed, the presented results open the door to the identification of large rearrangements, especially those affecting several exons.

The first step when using any NGS platform is to obtain the patient's DNA library for the region/s of interest. We selected a commercial multiplex PCR assay (Multiplicom) because it offers better reproducibility, more straightforward setup and better performance than in-house methods. This assay showed increased efficiency and homogeneity in the amplification of *BRCA* fragments with every new version of the kit released. A crucial step in preparing a DNA library for sequencing is to obtain equimolar proportions of all studied fragments to prevent undercovered regions and avoid the need for high mean coverage, which would generate higher costs. The latest version of the kit achieves an excellent ratio (1.96) between mean coverage and the 5th percentile of coverage (Table 1). This result outperforms the homogeneity previously reported by other groups describing next-generation *BRCA* testing using either long-range PCR,[4] primer-specific direct capture for single-molecule sequencing,[13] or in-house single/multiplex PCR.[12,14] It is also important to note that all of the MIDs used in the present study showed similar coverage results. Overall, this commercial assay allows the generation of a robust library for all the patients under study, maximizing the number of samples analyzed in a run.

Pyrosequencing performance with the GS Junior has been found to be similar to that of the GS FLX system,[12] which also uses Roche-454 technology. The GS Junior offers a more convenient scale for a mid-sized genetic testing laboratory, where the need to pool a large number of samples to use the whole capacity of a GS FLX device would increase waiting lists and, as a result, diagnostic turnaround times. GS Junior offers low entry and operating costs, providing conventional molecular diagnostics laboratories with a means of using NGS. Compared with other NGS technologies, Roche pyrosequencing currently offers the longest reads. This is advantageous for aligning possible mid-size insertions and deletions. In this study, the longest deletion tested (19 bp) was detected without a decrease in the expected allele frequency. The main disadvantage of pyrosequencing relative to other NGS technologies is the accuracy of length determination in homopolymers.[12,17,21] In pyrosequencing, the light-intensity signal observed in each cycle is proportional to the actual number of incorporated nucleotides, which is the base for homopolymer length calling. The accuracy of this method decreases with homopolymer length, which may eventually generate artefactual insertions and deletions in long homopolymers.[22,23] Our workflow circumvents this problem by using the BRCA HP assay.

To analyze the results we designed our own bioinformatic analysis pipeline using a combination of different software. VIP proved to find every variant, when enough coverage, but one deletion in a HP of 8 and has the advantage of being open source, making it preferable to other commercial software packages, which have only a limited capacity for adaptation to particular genes or laboratory needs. The generation of a reliable variant list is one of the most complex parts of the analysis and a key stage in the implementation of all next-generation platforms. The systematic application of a set of

evaluated filters is needed.[12] Ours is a four-filter approach: three run automatically and a fourth filter generates a list of variants that require visual examination or Sanger confirmation. Visual examination took about 3 h per run per revisor, and both revisions provided concordant results. Application of this four-filter approach left 16 fragments per patient requiring visual inspection, after which only 1% of them required Sanger confirmation. The fourth filter was able to remove a substantial proportion of the FPs without losing any TP when compared with other series.[12] The use of the commercial homopolymer kit was paramount for correctly reading sequences containing homopolymer stretches, which often require visual inspection and/or Sanger sequencing. Nevertheless, further development of tools for analysis of HP regions in NGS is needed to improve performance and to reduce the number of results requiring visual inspection.

In relation to the number of samples to be placed in each run, our results indicate that 5–7 is optimal with the new version of the kit. The latest version was experimentally tested using five samples and none of the fragments required resequencing for low coverage. We also carried out an *in silico* simulation of the same test with seven samples in each run instead of the five samples tested experimentally. The simulation was performed by randomly selecting 71% (five sevenths) of reads from each run and following the same analysis pipeline as for the Validation Set. The simulation results indicate that four fragments would have required Sanger sequencing due to low coverage (2 for R7, 0 for R8 and 2 for R9; that is, $\sim$0.2 fragments per sample), maintaining the same specificity and sensitivity as observed in the Validation Set (data not shown).

Although we have been able to detect LGRs, FPs have also been identified both in control and in patient samples, indicating that the specificity is too low for this method to be considered as an alternative strategy for detecting this type of mutations with the current software, kit protocol, and normalization procedures. Hopefully, in the near future, improvements to methodologies will lead to better specificity, allowing this approach to be used for the identification of LGRs in a diagnostic setting.

In a typical clinical setting, it is necessary to study a small number of genes comprehensively with the certainty of covering the whole coding region without any exception, with a sensitivity equal to or greater than that of conventional Sanger sequencing. Few studies have tackled a comprehensive assessment of specificity and sensitivity of NGS in the context of the requirements needed for a clinical diagnosis laboratory. To our knowledge, this is the first time that a NGS-based approach has been developed to perform comprehensive genetic testing of *BRCA* genes, including homopolymer regions, in a benchtop platform. We propose here a workflow that, using the GS Junior platform, allowed the identification of all DNA variants previously detected. A complete methodological process together with a detailed bioinformatic pipeline and validation of filters using open access programs has been critical to this achievement. Our custom-designed NGS workflow for genetic testing of *BRCA* genes meets the sensitivity and specificity requirements for the genetic diagnosis of HBOCS, making it feasible and cost-effective in comparison to current standards.

ACKNOWLEDGEMENTS

We thank Bernat Gel and Anna Ruiz for critical advice and corrections of the manuscript, and Toni Berenguer for statistical advice. We would also like to thank the Spanish Association Against Cancer (AECC) for recognizing our group with one of its awards. Finally, we would like to thank the teams from Multiplicom and Roche for their constant support. We thank contract grant sponsors: Spanish Health Research Fund; Carlos III Health Institute; Catalan Health Institute and Autonomous Government of Catalonia. Contract grant numbers: ISCIIIRETIC: RD06/0020/1051, RD06/0020/1050; 2009SGR290; PI10/01422; CA10/01474.

AUTHOR CONTRIBUTIONS

The project was conceived and the experiments and data analyses coordinated by LF, EC, CL, ES, GC. Samples were genetically characterized by JDV, MM, ET, EM, RC, CG, OC, MP, SG. Bioinformatic analysis was performed by ALD and VM. Samples from patients were obtained from JB and IB. The manuscript was written by LF, ALD, EC, JDV and CL and was discussed and improved by all the authors.

1 Rothberg JM, Hinz W, Rearick TM *et al*: An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011; **475**: 348–352.
2 Voelkerding KV, Dames SA, Durtschi JD: Next-generation sequencing: from basic research to diagnostics. *Clin Chem* 2009; **55**: 641–658.
3 Goossens D, Moens LN, Nelis E *et al*: Simultaneous mutation and copy number variation (CNV) detection by multiplex PCR-based GS-FLX sequencing. *Hum Mutat* 2009; **30**: 472–476.
4 Morgan JE, Carr IM, Sheridan E *et al*: Genetic diagnosis of familial breast cancer using clonal sequencing. *Hum Mutat* 2010; **31**: 484–491.
5 King MC, Marks JH, Mandell JB: Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* 2003; **302**: 643–646.
6 Bermejo-Perez MJ, Marquez-Calderon S, Llanos-Mendez A: Effectiveness of preventive interventions in BRCA1/2 gene mutation carriers: a systematic review. *Int J Cancer* 2007; **121**: 225–231.
7 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.
8 Marsh DJ, Howell VM: The use of denaturing high performance liquid chromatography (DHPLC) for mutation scanning of hereditary cancer genes. *Methods Mol Biol* 2010; **653**: 133–145.
9 Mattocks CJ, Watkins G, Ward D *et al*: Interlaboratory diagnostic validation of conformation-sensitive capillary electrophoresis for mutation scanning. *Clin Chem* 2010; **56**: 593–602.
10 Ewald IP, Ribeiro PL, Palmero EI, Cossio SL, Giugliani R, Ashton-Prolla P: Genomic rearrangements in BRCA1 and BRCA2: a literature review. *Genet Mol Biol* 2009; **32**: 437–446.
11 Sluiter MD, van Rensburg EJ: Large genomic rearrangements of the BRCA1 and BRCA2 genes: review of the literature and report of a novel BRCA1 mutation. *Breast Cancer Res Treat* 2011; **125**: 325–349.
12 De Leeneer K, Hellemans J, De Schrijver J *et al*: Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. *Hum Mutat* 2011; **32**: 335–344.
13 Thompson JF, Reifenberger JG, Giladi E *et al*: Single-step capture and sequencing of natural DNA for detection of BRCA1 mutations. *Genome Res* 2011; **22**: 340–345.
14 Hernan I, Borras E, de Sousa Dias M *et al*: Detection of genomic variations in BRCA1 and BRCA2 genes by long-range PCR and next-generation sequencing. *J Mol Diagn* 2012; **14**: 286–293.
15 De Schrijver JM, De Leeneer K, Lefever S *et al*: Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline. *BMC Bioinformatics* 2010; **11**: 269.
16 Kent WJ: BLAT–the BLAST-like alignment tool. *Genome Res* 2002; **12**: 656–664.
17 De Leeneer K, De Schrijver J, Clement L *et al*: Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics. *PLoS One* 2011; **6**: e25531.
18 Mattocks CJ, Morris MA, Matthijs G *et al*: A standardized framework for the validation and verification of clinical molecular genetic tests. *Eur J Hum Genet* 2010; **18**: 1276–1288.
19 del Valle J, Feliubadalo L, Nadal M *et al*: Identification and comprehensive characterization of large genomic rearrangements in the BRCA1 and BRCA2 genes. *Breast Cancer Res Treat* 2009; **122**: 733–743.
20 del Valle J, Campos O, Velasco A *et al*: Identification of a new complex rearrangement affecting exon 20 of BRCA1. *Breast Cancer Res Treat* 2011; **130**: 341–344.
21 Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM: Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007; **8**: R143.
22 Loman NJ, Misra RV, Dallman TJ *et al*: Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012; **30**: 562.
23 Quinlan AR, Stewart DA, Stromberg MP, Marth GT: Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* 2008; **5**: 179–181.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)