European Journal of Human Genetics (2014) 22, 535–541

www.nature.com/ejhg

ARTICLE

Streamlined ion torrent PGM-based diagnostics: *BRCA1* and *BRCA2* genes as a model

Julien Tarabeux[1,2], Bruno Zeitouni[3], Virginie Moncoutier[1], Henrique Tenreiro[1], Khadija Abidallah[1], Séverine Lair[3], Patricia Legoix-Né[4], Quentin Leroy[4], Etienne Rouleau[1], Lisa Golmard[1,2], Emmanuel Barillot[3], Marc-Henri Stern[1,2], Thomas Rio-Frio[4], Dominique Stoppa-Lyonnet[1,2,5] and Claude Houdayer*[,1,2,5]

To meet challenges in terms of throughput and turnaround time, many diagnostic laboratories are shifting from Sanger sequencing to higher throughput next-generation sequencing (NGS) platforms. Bearing in mind that the performance and quality criteria expected from NGS in diagnostic or research settings are strikingly different, we have developed an Ion Torrent's PGM-based routine diagnostic procedure for *BRCA1/2* sequencing. The procedure was first tested on a training set of 62 control samples, and then blindly validated on 77 samples in parallel with our routine technique. The training set was composed of difficult cases, for example, insertions and/or deletions of various sizes, large-scale rearrangements and, obviously, mutations occurring in homopolymer regions. We also compared two bioinformatic solutions in this diagnostic context, an in-house academic pipeline and the commercially available NextGene software (Softgenetics). NextGene analysis provided higher sensitivity, as four previously undetected single-nucleotide variations were found. Regarding specificity, an average of 1.5 confirmatory Sanger sequencings per patient was needed for complete *BRCA1/2* screening. Large-scale rearrangements were identified by two distinct analyses, that is, bioinformatics and fragment analysis with electrophoresis profile comparison. Turnaround time was enhanced, as a series of 30 patients were sequenced by one technician, making the results available for the clinician in 10 working days following blood sampling. *BRCA1/2* genes are a good model, representative of the difficulties commonly encountered in diagnostic settings, which is why we believe our findings are of interest for the whole community, and the pipeline described can be adapted by any user of PGM for diagnostic purposes.
European Journal of Human Genetics (2014) **22,** 535–541; doi:10.1038/ejhg.2013.181; published online 14 August 2013

Keywords: Ion Torrent; PGM; diagnostics; *BRCA*1; *BRCA*2

INTRODUCTION

With progress in next-generation sequencing technologies (NGS) and a corresponding decreased cost, capillary sequencing, which has been the norm for clinical diagnosis up until now, is becoming superseded by NGS abilities. As a result, many diagnostic laboratories are shifting from Sanger sequencing platforms to higher throughput NGS platforms,[1–5] already used in research. However, the performance and quality criteria expected from NGS in diagnostic and research settings are strikingly different. As an example, average coverage is usually reported in research settings, a feature irrelevant in diagnosis as the whole region of interest (ROI) must be covered, that is, 100% coverage. In other words, a diagnostic laboratory working with NGS has to provide sensitivity for its favorite genes at least equal to that of routine techniques, such as Sanger sequencing.

The Institut Curie genetic unit in Paris is actively involved in the diagnosis of breast and ovarian cancer predisposition,[6,7] with access to a high-throughput platform equipped with two Ion Torrent's Personal Genome Machine sequencers (PGM, Life Technologies, Carlsbad, CA, USA), two ABI SOLiD v4 platforms (Life Technologies) and one HiSeq (Illumina, San Diego, CA, USA). Consequently, it was tempting to test these novel technologies for diagnostic purposes in order to implement the most appropriate technology.

Following pilot tests combining various enrichment procedures and sequencing platforms (see supplementary data), a PGM-based routine diagnostic procedure for *BRCA1* (MIM 113705) and *BRCA2* (MIM 600185) sequencing was defined and is described below. The procedure was first tested on a training set of 62 control samples, and then blindly validated on 77 samples in parallel with our routine technique. *BRCA1* and *BRCA2* genes constitute a good model, as they are representative of the difficulties commonly encountered in diagnostic settings: these genes are located on autosomes, contain homopolymer regions, segmental duplication with a pseudogene located 5′ to *BRCA1* and a mutational spectrum composed of all kinds of private mutations (single-nucleotide variations (SNVs), insertions/deletions (indels) and large rearrangements of one exon to the entire gene), moreover scattered throughout the coding sequence. For these reasons, we believe that our findings are of interest to the whole diagnostic community, and the diagnostic pipeline described here can be used or adapted by any user of the PGM platform for diagnostic purposes.

[1]Service de Génétique Oncologique, Institut Curie, Paris, France; [2]INSERM U830, Centre de Recherche de l'Institut Curie, Paris, France; [3]INSERM U900, Mines Paris Tech, Centre de Recherche de l'Institut Curie, Paris, France; [4]Next Generation Sequencing Platform, Centre de Recherche de l'Institut Curie, Paris, France; [5]Université Paris Descartes, Sorbonne Paris Cité, Paris, France
*Correspondence: Dr C Houdayer, Service de Génétique Oncologique, INSERM U830, Institut Curie, 75005 Paris et Université Paris Descartes, Sorbonne Paris Cité, Paris, France. Tel: +33 1 44 32 41 17; Fax: +33 1 53 10 26 48; E-mail: claude.houdayer@curie.net
Received 25 January 2013; revised 11 July 2013; accepted 16 July 2013; published online 14 August 2013

PATIENTS AND METHODS

Patients and DNA

All patients attended an interview with a geneticist and a genetic counselor in a family cancer clinic at the Institut Curie, Paris, France. Genetic testing for *BRCA1* and *BRCA2* was proposed to women based on individual and/or family history, as previously published.[7] Patients gave their informed consent for *BRCA1/2* gene analyses. DNA was extracted using the Quickgene 610-L automated system from FujiFilm (Tokyo, Japan) according to the manufacturer's instructions and calibrated to 50 ng/µl by UV spectrophotometric assay (Nanodrop, Thermo Fischer Scientific, Waltham, MA, USA).

Multiplex PCR enrichment and library preparation

Series of 30 patients were PCR-enriched using the BRCA MASTR assay v2.0 (Multiplicom, Niel, Belgium) according to the manufacturer's instructions. Briefly, following a target-specific PCR amplification performed on a Tecan Freedom Evo (Tecan, Männedorf, Switzerland) and a second PCR round with universal primers, amplified products were electrophoresed on an ABI3130XL to control for enrichment efficacy. Universal PCRs were then purified, pooled, and libraries were prepared using the Library Builder (Life Technologies) to obtain 300-bp PCR fragments flanked by adaptor and barcode sequences, allowing sequencing and sample identification, respectively. Libraries for the 30 patients were then pooled and submitted to 10 PCR cycles in order to select and amplify relevant constructions, for example, PCR fragments with correct barcode and adaptor ligation.

Large rearrangement analysis

Large rearrangements were identified at two distinct steps by two distinct analyses, that is, at the bioinformatics step (see the 'Bioinformatic analyses' section) and at the electrophoresis step following Multiplicom enrichment. Large rearrangements were identified by profile comparison, using MLPA-like software analyses.[8] Briefly, electrophoregrams from patients were first superposed, the yield of each amplicon in the various samples was evaluated and deletions/duplications of one or more amplicons were revealed by a 2-fold decrease/1.5-fold increase of the corresponding peak(s), respectively. High-throughput data analysis was made possible by automated profile analysis using Gene Marker software version 1.95 (Softgenetics, State College, PA, USA). Of note, the electrophoresis step was also used to search for large indels, evidenced by supplementary peaks.

Ion Torrent PGM sequencing

Amplified libraries were controlled for primer dimers and size range using LabChip devices (Caliper, PerkinElmer, Waltham, MA, USA) and were then submitted to emulsion PCR with the Ion Xpress template kit using the Ion One Touch system (Life Technologies). Ion sphere particles (ISP) were enriched using the E/S module and were sequenced with an Ion PGM[9] in a 300-bp configuration run using a 318 chip (Life Technologies).

Bioinformatic analyses

The ROI was defined as the coding sequence minus 20 bp and plus 10 bp in the preceding and following intron, respectively, to ensure correct analysis of consensus splice sites.[10] A combination of academic tools (TMAP, diBayes, GATK and DESeq)[11–13] and the commercially available NextGene software from SoftGenetics were used in parallel. For the academic analysis, PGM reads in the SFF file format were aligned onto the reference human genome hg19 with TMAP 0.3.7 using the same parameter set on the Ion Torrent Server 2.2.1. The standalone package of the TorrentVariantCaller 2.2.3 (Life Technologies) was used for calling variants from PGM mapped reads. Whereas the target germline parameters were left as defaults for the SNV Calling with the diBayes tool, a less stringent setting was applied for indels called with GATK (max_alternate_alleles = 3, min-bayesian-score = 0.1, gatk-score-minlen = 900). Candidate variants were obtained by a personal filtering perl script based on the following requirements: a minimum coverage of 30×, a minimum variant frequency of 0.3 for SNVs and 0.2 for indels, and a minimum strand coverage ratio of 0.2. SNVs, and indels were then functionally annotated with Annovar.[14] For large rearrangement detection, the normalization calculation and the comparison procedure of count data were performed separately for each of five PCR multiplexes. Exonic rearrangements were then retrieved if both the adjusted *P*-values were <0.2 (Benjamini and Hochberg procedure)[15] and the log2 ratios (sample/control) were <-0.85 (ie 1.8-fold decrease) or >0.4 (ie 1.3-fold increase), indicating loss or gain of exon(s), respectively (supplementary Table 2). This academic pipeline will soon be available via a Galaxy interface.[16]

Default read conversion and mapping parameters with rigorous alignment were used for NextGene analysis of SNV (version 2.3). SNVs were called using default parameters except for balance ratios and homopolymer indel balance indexes, which were inactivated, without subsequent filtering. For NextGene analysis of indels, we used both reads converted by default and reads converted with higher quality parameters, and the same rigorous alignment was used but with a mutation percentage set to 10, balance ratios set to 0.1 and homopolymer indel balance set to 0.15. Another dedicated layer of filtering was then applied as described in Figure 1. Briefly, isolated mutations in the two analyses (ie, found in one or two patient(s)) were considered to be true positives and were validated by Sanger sequencing, but recurrent mutations (found in more than two patients) were compared in terms of homogeneity between samples to focus on outliers, which were subsequently Sanger-sequenced.

All mutations are reported following the Human Genome Variation Society (HGVS) guidelines on the basis of the coding sequences NM_007294.2 and NM_000059.3 for *BRCA1* and *BRCA2*, respectively. All mutations have been submitted to the BRCA part of the Universal Mutation Database (http://www.umd.be/BRCA1/, http://www.umd.be/BRCA2/).

RESULTS

Ion Torrent PGM sequencing training set

A training set of 62 patients, fully genotyped by our routine technique,[7] was used to calibrate the routine procedure, that is, enrichment, library preparation and bioinformatic analysis parameters. To adequately address diagnostic issues, the training set was composed of difficult cases, for example, insertions and/or deletions of various sizes, large rearrangements and, obviously, mutations occurring in homopolymer regions of *BRCA1/2* (Table 1). Diagnostic pipeline parameters were trained in three distinct experiments running different analytical conditions (eg, the 316 and 318 chips were tested with 200 or 300 bp chemistries) and different bioinformatics parameters (eg various filtering stringencies for depth of coverage, quality value (QV) of the reads, strand bias and allelic ratio). The mandatory condition of complete ROI coverage was readily obtained, thus validating the Multiplicom enrichment. To address the issue of minimum depth-of-coverage, we randomly discarded an increasing number of reads from the analysis, then checked the data to see at what level known SNVs and indels would disappear from the final results (Figure 2). This simulation showed that ~100–130× and 20–30× should ensure reliable detection of indels and SNVs, respectively. Unfortunately, strand bias was recurrent in *BRCA1* exon 2, prompting us to deactivate any primary filtering based on this strand bias. Major trimming of low-quality reverse reads might explain this bias. We found that SNVs were confidently called following rigorous mapping (ie, with trimmed reads with QV > 12), but additional levels of filtering were required for indels and homopolymers calling. The rationale was that PGM sequencing produces very few false SNV when reads are trimmed. On the other hand, it produces sequencing artifacts mimicking insertions and/or deletions that the user must filter in order to achieve adequate specificity while maintaining satisfactory sensitivity. In the extreme case of homopolymers, rather than trying to avoid errors, we tried to take advantage of these errors by relaxing the filtering parameters followed by careful triage of the amount of mutations found, as described in Figure 1. It is well known that the theoretical allelic ratio

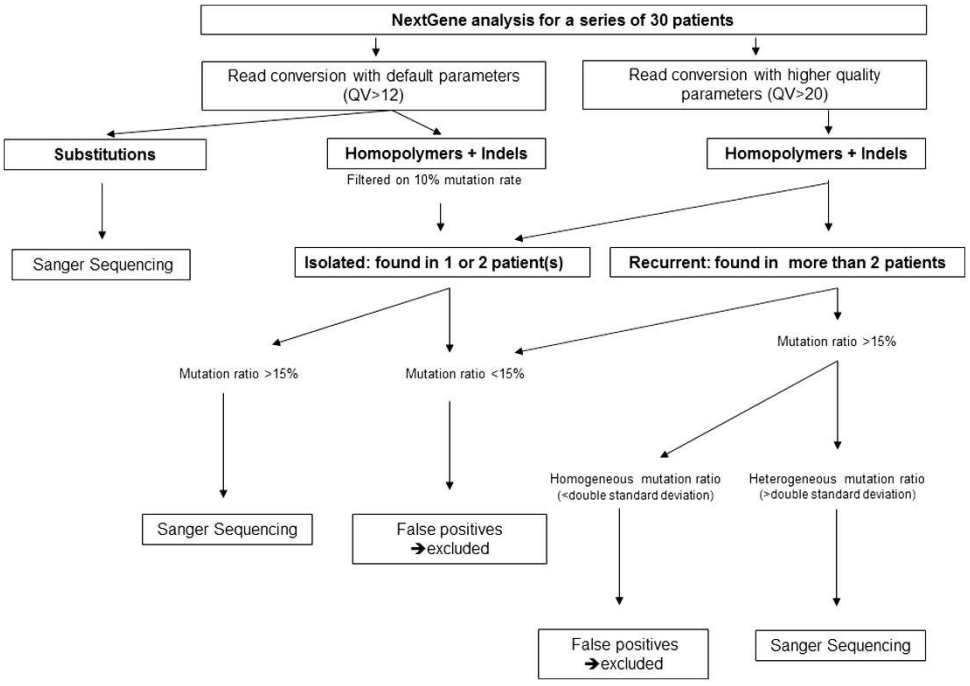


Figure 1 Decision tree for NextGene analysis for series of 30 patients. Two distinct analyses are run with distinct read quality values (QV, ie, a Phred-based score) to address homopolymer regions. True homopolymer variants are found in one patient only and/or in several patients, but as outliers as compared to the false positives (>double standard deviation). Analyses are automatically launched and processed in about 2 h. See text for details.

of 50% is not always obtained,[17] and the lowest value achieved in our training set was 20% for a 1-bp insertion occurring in a 6 T-stretch. Nevertheless, and to ensure maximum sensitivity, the filter was set at 15%.

Regarding large rearrangement detection, our academic pipeline provided promising results, as all rearrangements were detected with low variation between amplicons (Figure 3a and b). NextGene analysis based on coverage, number of reads or reads per kilobase of exon model per million mapped reads was inconclusive, prompting us to use the analysis fragment step of the enrichment procedure for large rearrangement detection. Interestingly, the so-called Portuguese founder mutation (c.156_157insAP003441.3:g105088_105370 also known as c.156_157insAlu close to exon 3)[18] was evidenced by a 50% decrease in *BRCA2* exon 3 peak intensity, indicating that only one allele was amplified (Figure 4).

Ion Torrent PGM sequencing validation set

This optimized pipeline defined on the training set was then blindly applied to a validation set of 77 patients PGM-sequenced on *BRCA1* and *BRCA2*, in parallel with our routine diagnostic strategy. The whole ROI was covered with $>30\times$ (see supplementary Table 3 for minimum requirements). Relative variation between amplicons was <15. All variants previously detected with the routine technique were identified again, but higher sensitivity was obtained with the PGM-NextGene analysis, as four undetected SNVs (validated by subsequent Sanger sequencing) were found, including a *BRCA2* c.68-7 T>A substitution occurring in a homopolymer tract of polypyrimidines. Sensitivity was slightly decreased with the academic pipeline, as three variants were missed, including the *BRCA2* c.68-7dup, occurring in a homopolymer run (see Discussion). On the other hand, the number of false positives was negligible with the academic pipeline compared with Nextgene (see Table 2 and supplementary Table 1). Specificity calculation was difficult for NGS experiments because of the definition of 'true negatives': they could theoretically refer to the number of wild-type bases called, but this would give impressive specificity values unrelated to clinical practice. Various approximations were therefore used, for example, based on amplicon calculation,[17] but keeping in mind the clinical utility, we preferred to rely on the false-positive rate and the number of supplemental Sanger sequencings, which have a major impact on the whole routine process. Using NextGene analysis, complete sensitivity was reached with 113 targeted Sanger sequencing overall, that is, an average 1.5 Sanger sequencings per patient for complete *BRCA1/2* screening. Turnaround time was enhanced, as series of 30 patients were sequenced by one technician, making the results available for the clinician in 10 working days following blood sampling. Costs entailed are obviously site-dependent. However, based on the list of prices, reagent cost for a complete *BRCA1/2* sequencing was 250 euros, that is, 327 US dollars (including DNA extraction, enrichment and library preparation, sequencing and bioinformatics analysis).

DISCUSSION

In this paper, we describe a PGM-based diagnostic pipeline applied to *BRCA* genes, which has a high potential to be implemented for other genes. In this respect, some technical remarks and guidelines for implementation and use, based on our experience, are discussed below.

General comments

In general, the wet lab part of the protocol is easy to implement, as it consists of automated solutions and manufactured kits that are easy to handle by any molecular biology laboratory. Automation is mandatory to ensure adequate throughput in a large-scale diagnostic perspective. Large rearrangement screening is straightforward, as it uses a well-known fragment analysis procedure running with MLPA-like software analysis. Sequence scalability, by using chips of various

Table 1 Training set of nucleotide variations. Nucleotide position was numbered on the basis of the coding sequences NM_007294.2 and NM_000059.3 for *BRCA1* and *BRCA2*, respectively

Gene	*Variant type*	*Description*
BRCA 1	*Large rearrangements*	c.(?_-232)_c.441+? del
		(deletion from 5′ to exon 7)
		c.4676-?_c.5074+? del
		(deletion of exons 16 and 17)
		c.671-?_c.4185+? del
		(deletion of exons 11 and 12)
		c.81-?_c.547+? dup
		(duplication from exon 3 to exon 8)
	Insertions/deletions (homopolymer regions excluded)	c.19_47del
		c.68_69del
		c.1121del
		c.3013del
		c.3416_3427delinsC
		c.3481_3491del
		c.3680_3729dup
		c.3839_3843delinsAGGC
		c.4243_4281dup
		c.5030_5033del
	Insertions/deletions in homopolymer regions	c.5266dup
		c.1016dup
		c.1961dup
		c.1961del
		c.2071del
		c.211del
		c.3926del
		c.2429del
		c.3285del
	SNVs (polymorphisms excluded)	c.736T>G
		c.5471T>A
BRCA2	*Large rearrangements*	c.8332-?_c.8632+? dup
		(duplication of exons 19 and 20)
		c.(?_-227)_(*902_?) del
		(whole gene deletion)
	Insertions/deletions (homopolymer regions excluded)	c.3645_3646delinsTAAAAAG
		c.5835_5842dup
	Insertions/deletions in homopolymer regions	c.161del
		c.994dup
		c.1231del
		c.1593dup
		c.1813dup
		c.1929del
		c.2175dup
		c.2588dup
		c.4284dup
		c.5351del
		c.5351dup
		c.6373dup
		c.7680dup
		c.8207del
	SNVs (polymorphisms excluded)	c.223G>C
		c.4208C>T
		c.5704G>A
		c.8182G>A
		c.8630A>G
		c.9154C>T

Nucleotide numbering reflects cDNA numbering with +1 corresponding to the A of the ATG translation initiation codon in the reference sequence.

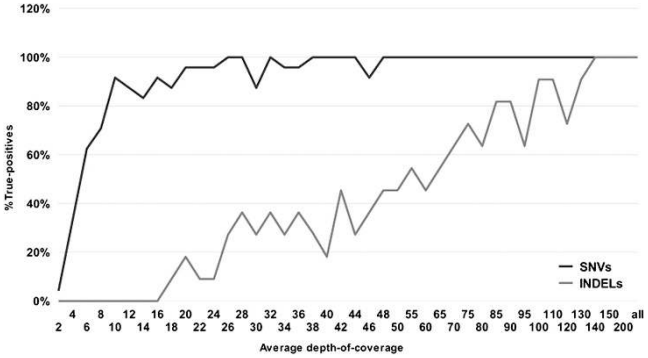


Figure 2 Evaluation of the minimum of coverage required for accurate SNVs and indels detection using the academic pipeline. The percentage of true-positive variants is shown at decreasing depth-of-coverage values obtained by random down-sampling of the total mapped reads. The number of positive SNVs (25) or indels (11) was evaluated at the PGM run level covering the *BRCA1* and *BRCA2* gene sequencing of 16 different patients. Black: SNVs. Gray: indels.

capacity ranging from 50 Mb to 1 Gb combined with short run times are the main advantages of the platform, although the 2-hour run time is only theoretical, as the run time is 5 h for a 318 chip and the whole procedure takes 1 working day, taking into account reagent preparation and PGM initialization.

Patient identification

During NGS analysis, patients are pooled and sequenced together, but remain readily identified by using barcodes ligated to the amplicons during library preparation. Special attention was paid to barcode reliability, as we experienced poor barcode quality in our first SOLiD experiments (see Supplementary data). In our hands, all PGM barcodes worked well with even and similar number of reads for all patients, with $<1\%$ of reads left unattributed.

Bioinformatics

Once mapping and analysis parameters have been set up for routine analysis, quality assessment starts with the number of mapped bases, read length and associated QVs. The number of mapped bases depends on enrichment strategies and the size of the target, but QV should not fall <20 before 150 bp with a 318 chip and 300 bp chemistry.

Main NGS analysis parameters in diagnostics are ROI coverage, QV of the reads, strand bias, allelic ratio and depth-of-coverage. Major genes (such as *BRCA1* and *BRCA2* in hereditary breast and ovarian cancer) obviously need a complete 100% ROI coverage for diagnostic purposes. On the other hand, so-called minor genes with little clinical importance may be addressed in genes packages where the coverage may not be complete. From our experience, a minimum QV of 12 was found to be sufficient for reliable mutation detection, although increasing the value to 20 (ie, selecting higher quality reads) improves homopolymer analysis with fewer false positives. Extreme caution must be taken to avoid strand bias, as it may result from the enrichment step and filtering may therefore exclude *bona fide* variants. On the other hand, we can confidently state that variants with an allelic ratio of $<15\%$ are false positives (somatic analyses excluded), a finding in accordance with those of other platforms.[19,20]

Depth of coverage is another prominent point to consider, because the lower the depth, the higher the risk of missing a variant, by sequencing failure and/or bioinformatics filtering. Conversely, the

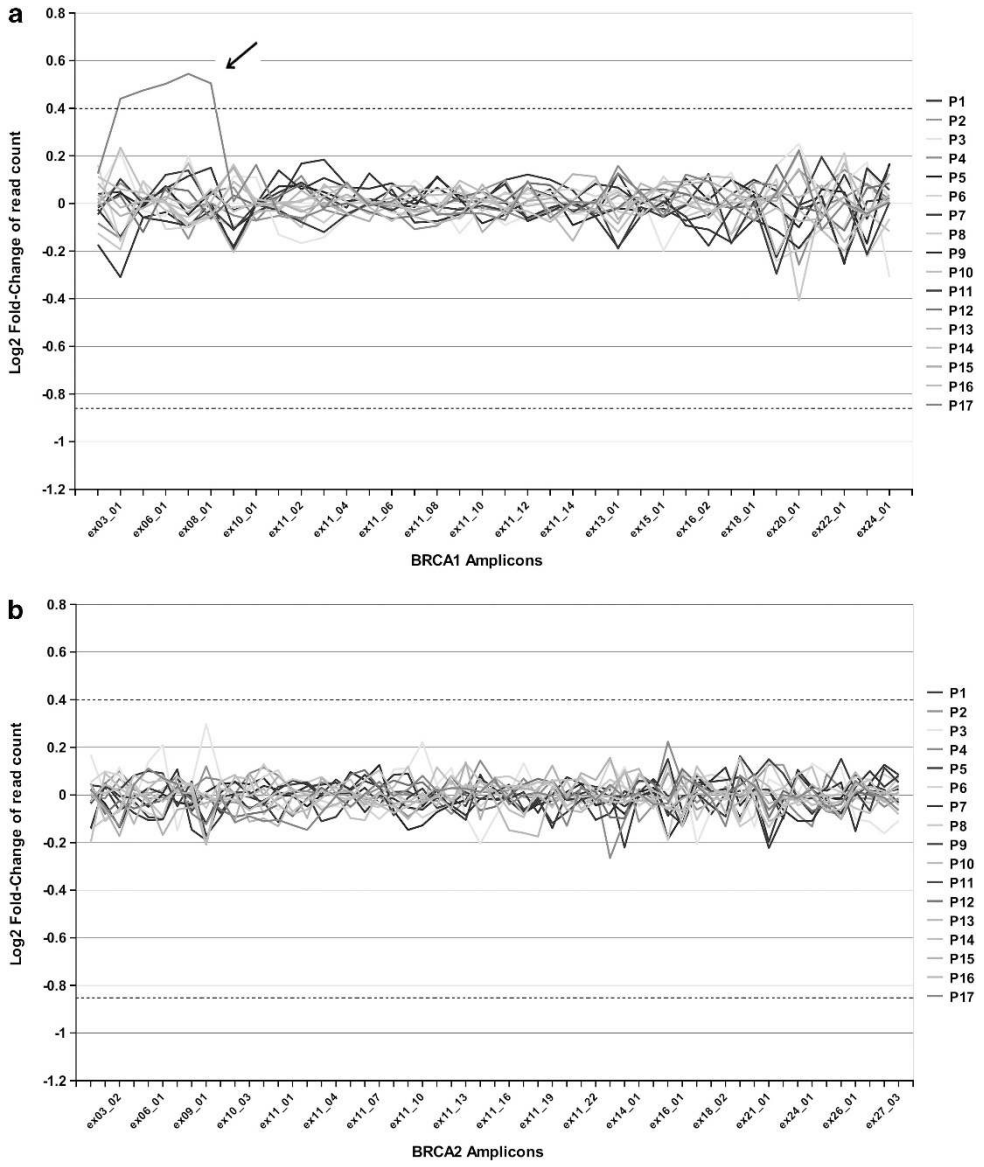
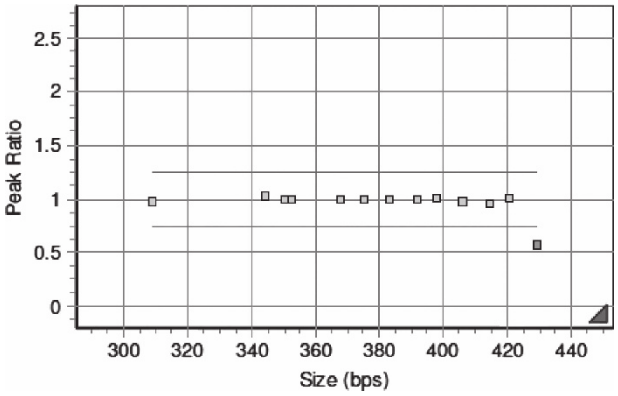


Figure 3 Graphical representation of large rearrangement analysis using the academic pipeline. The differential analysis of read counts in *BRCA* amplicons of 17 different patients is shown. Each line represents a patient. *x* axis: amplicons under study, *y* axis: Log2 fold-change ratios. The thresholds for exonic deletions (< -0.85, ie, 1.8-fold decrease) or duplications (> 0.4, ie, 1.3-fold increase) are indicated by dotted lines. (**a**) Detection of a duplication from exons 3 to 8 of *BRCA1* (arrow). (**b**) Results for the same patients on *BRCA2*: no rearrangements were detected.

higher the depth, the higher the probability of discriminating false positives (ie, random errors) from true positives. As random errors are linked to the sequencer error rate, minimum depth is technology-driven. PGM simulation suggested a minimum range of 20–30 × and 100–130 × for reliable SNV and indel detection, respectively (Figure 2), far from the 20–38 × commonly reported in the literature.[19] Our understanding is that these published values are indicated for SNV and not for indels, which deserve a separate analysis. The reason is that indels recover a wider range of nucleotide modifications, both in nature (insertions, deletions, and insertions–deletions) and size (from one to several bases) compared with SNVs (one-base modification). As a result, minimum depth for indel detection may sound like an oxymoron, because one value should refer to a heterogeneous set of mutations. Another layer of complexity is related to the fact that minimum depth also depends on the quality of the mapped reads. Selecting reads with high QV values (by filtering low QVs) decreases the depth, enhances SNV and simple indel detection by decreasing sequencing artifacts, but is associated with a risk of missing complex indels mimicking these artifacts and thereby decreasing their read QV values. At last, bioinformatics tools have distinct sensitivity/specificity (Table 2). This is illustrated in our simulation with the academic pipeline in which a rapid decrease for indels (< 100–$130 \times$, Figure 2) was linked to the presence of homopolymer mutants and associated lack of sensitivity (see 'Homopolymer' section below). On the other hand, NextGene analysis provided complete sensitivity, and lower depth can be used. This was further illustrated by an experiment in which we depleted the number of ISPs, leading to reduced depth, and nevertheless detected all indels for example, a 68_69del from *BRCA1* at 35 × using NextGene (data available on request). Overall, we can confidently state, in agreement



Probe Name	Size	Number of samples	Mean	Standard Error	ND3R28
BRCA2_exon9_02	309.1	28	0.99	0.00	0.976
BRCA1_exon23_01	343.9	28	1.00	0.00	1.022
BRCA1_exon11_12	350.1	28	0.99	0.00	1.003
BRCA2_exon11_07	352.4	28	1.00	0.00	1.000
BRCA2_exon13_01	368.2	28	1.00	0.00	1.000
BRCA2_exon15_01	375.6	28	0.99	0.00	1.000
BRCA2_exon11_01	383.3	28	1.00	0.00	0.989
BRCA2_exon24_01	391.9	28	1.00	0.00	1.004
BRCA2_exon14_01	398	28	1.00	0.00	1.010
BRCA2_exon11_15	406.2	28	0.99	0.00	0.981
BRCA2_exon11_21	414.7	28	0.99	0.01	0.971
BRCA1_exon11_04	420.7	28	1.00	0.00	1.010
BRCA2_exon3_01	429.1	28	0.98	0.02	0.578

Figure 4 Identification of the c.156_157insAlu (c.156_157insAP003441.3:g105088_105370) founder mutation using the electrophoresis/fragment analysis approach. Left panel, graphical representation, right panel, numerical data. Analysis was performed by GeneMarker software v1.95 (SoftGenetics) using MLPA-like panels. Population normalization was based on all 30 samples. The Alu insertion corresponded to a 50% decrease in exon 3 peak intensity (ratio: 0.578). Note on this panel the redundancy (consequently resulting in higher reliability) for *BRCA1* and *BRCA2* exon 11 (2 and 4 amplicons, respectively).

Table 2 Sensitivity and false positive rates on the validation set

	Academic PGM pipeline		*NextGene pipeline (v2.3)*	
	Sensitivity	*False-positive rate*	*Sensitivity*	*False-positive rate*
Number of expected SNVs = 66	98.5% (65 detected)	1.5% (1 false positive)	100% (66 detected)	0% (no false positives)
Number of expected indels = 9	77.8% (7 detected)	30% (3 false positives)	100% (9 detected)	87.7% (64 false positives)

Sensitivity was calculated as follows: (number of true positives/(number of true positives + number of false negatives)). Specificity was estimated by the using the false-positive rate calculated as follows: (number of false positives/(number of false positives + number of true positives)), see text for details.

with previous literature, that a 30 × minimum depth is sufficient for SNV detection. As discussed above, the situation is less clear for indels and may vary from 30 to 130 × depending on the context (eg, long homopolymers) and bioinformatics.

Finally, detection of long indels depends on read length, because mapping parameter constraints makes indels longer than 45 bp undetectable for a 300-bp read.

Homopolymers

Error rates in homopolymer stretches are well known with pyrosequencers,[17] and the same appears to be true with PGM. Bioinformatic solutions provided a contrasted picture, as the academic solution showed 44% sensitivity for long homopolymers (more than seven identical bases) on the training set, a feature partly explaining the decreased sensitivity observed in the validation set (Table 2). We believe the GATK tool lacks power to properly address this high number of recurrent errors. On the other hand, and as a result of the normal distribution of random errors, a simple rule ensures reliable detection of true positives following NextGene analysis (Figure 1). This was exemplified in our validation set, in which we blindly detected a BRCA2 c.68-7 T>A substitution occurring in a homopolymer tract of polypyrimidines, which was previously missed by our routine method (see Results). This part of *BRCA2* is particularly difficult to sequence, as two PCR species were already obtained following Multiplicom enrichment, that is, the 'true' expected amplicon and another shorter amplicon due to polymerase slippage. Despite that, variants were correctly called in this stretch.

Large rearrangements

Large rearrangements are independently detected twice, at the electrophoresis and bioinformatics steps. Consistency of these two independent analyses ensures reliable mutation detection. Of note, homogeneously prepared, high-quality DNA must be used, as this analysis is based on comparisons between distinct samples, which therefore need to have similar amplification kinetics. Using the electrophoresis method and thanks to the very large redundancy, for example, in exon 11, some poor-performing peaks can be withdrawn to focus on 'good-performers'. Unfortunately, we were unable to obtain reproducible data for *BRCA2* exon 6. Nevertheless, and given the close proximity of the reliable amplicons for exons 5 and 7 (exons 5 and 6 are distant from 91 bp and exons 6 and 7 from 216 bp), we believe that it does not have any impact on diagnostic accuracy, as a large rearrangement impacting exon 6 would undoubtedly also encompass exons 5 and/or 7. Special attention was paid to *BRCA1* exon 9 due to the deletion polymorphism included in the amplicon: the two types of homozygous allelic variants as well as heterozygous allelic variants should be analyzed separately. Performances and procedures are similar for both bioinformatics and electrophoresis methods, but a major advantage of bioinformatics is that all amplicons are taken into account.

Perspectives

Success depends on a good understanding of the entire process by the team, implying extensive training, especially for bioinformatics aspects that require new knowledge and a new way of thinking, which is why we believe that performances will improve as the team

gains experience and knowledge. As an example, we have already achieved complete sensitivity, and our current specificity should rapidly further increase, as the so-called false positives reported here were actually poor-quality variants that will probably no longer be considered for Sanger confirmation in the near future.

NGS is gaining interest in diagnosis due to its huge potential benefits. However, it must be kept in mind that the expected requirements (eg, for coverage) and level of constraint (eg, accreditation) are much higher in routine diagnosis than in research settings. In any case, diagnostic laboratories are at the beginning of their learning curve and, provided industrial partners understand the need for steps/pauses in technological improvements, instead of continuously evolving systems that are incompatible with development and validation of diagnostic procedures, significant progress can be expected both in terms of throughput and number of genes analyzed. Extension of this technology to other genes must obviously be considered in terms of benefits for patients. As diagnostics moves towards exomic approaches, variant interpretation and the risks entailed are becoming major issues, which will mobilize thousands of hours of the biologist's time and will require international collaborations in order to share cosegregation results and expand functional testing. Similarly, bioinformatic resources need to increase accordingly, and diagnostic laboratories should be aware that they cannot rely on competent but 'volatile' post-doc personnel. Last but not the least, the rapid decrease in the cost of data generation has not been matched by a comparable decrease in the cost of computational infrastructures. We therefore believe that raw data should be deleted to store only variant annotation files, as improved technologies promote novel analysis rather than reinterpretation of old data, if 'old' bioinformatics tools are still available for that purpose.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Dr Laurent Castéra for helpful support during the 'technology selection' phase, Anthony Laugé for LIMS optimization, Alain Nicolas for fruitful discussions and the referring clinicians. This work was supported by funding from ICGex, CEST Institut Curie and INCa/DGOS 'Recherche translationnelle sur le cancer, and Séquençage haut debit et predisposition aux cancers du sein'. NGS projects were also supported by grants from Canceropôle Ile de France and the Region Ile de France (Seq-Start).

1 Shanks ME, Downes SM, Copley RR *et al*: Next-generation sequencing (NGS) as a diagnostic tool for retinal degeneration reveals a much higher detection rate in early-onset disease. *Eur J Hum Genet* 2013; **21**: 274–280.
2 Artuso R, Fallerini C, Dosa L *et al*: Advances in Alport syndrome diagnosis using next-generation sequencing. *Eur J Hum Genet* 2012; **20**: 50–57.
3 Ozcelik H, Shi X, Chang MC *et al*: Long-range PCR and next-generation sequencing of BRCA1 and BRCA2 in breast cancer. *J Mol Diagn* 2012; **14**: 467–475.
4 Pern F, Bogdanova N, Schurmann P *et al*: Mutation analysis of BRCA1, BRCA2, PALB2 and BRD7 in a hospital-based series of German patients with triple-negative breast cancer. *PLoS One* 2012; **7**: e47993.
5 Feliubadalo L, Lopez-Doriga A, Castellsague E *et al*: Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. *Eur J Hum Genet* 2013; 864–870.
6 Stoppa-Lyonnet D, Laurent-Puig P, Essioux L *et al*: BRCA1 sequence variations in 160 individuals referred to a breast/ovarian family cancer clinic. Institut Curie Breast Cancer Group. *Am J Hum Genet* 1997; **60**: 1021–1030.
7 Caux-Moncoutier V, Castera L, Tirapo C *et al*: EMMA, a cost- and time-effective diagnostic method for simultaneous detection of point mutations and large-scale genomic rearrangements: application to BRCA1 and BRCA2 in 1,525 patients. *Hum Mutat* 2011; **32**: 325–334.
8 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.
9 Rothberg JM, Hinz W, Rearick TM *et al*: An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 2011; **475**: 348–352.
10 Houdayer C, Caux-Moncoutier V, Krieger S *et al*: Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/*in vitro* studies on BRCA1 and BRCA2 variants. *Hum Mutat* 2012; **33**: 1228–1238.
11 San Lucas FA, Wang G, Scheet P, Peng B: Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* 2011; **28**: 421–422.
12 McKenna A, Hanna M, Banks E *et al*: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.
13 Anders S, Huber W: Differential expression analysis for sequence count data. *Genome Biol* 2010; **11**: R106.
14 Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010; **38**: e164.
15 Hochberg Y, Benjamini Y: More powerful procedures for multiple significance testing. *Stat Med* 1990; **9**: 811–818.
16 Goecks J, Nekrutenko A, Taylor J: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010; **11**: R86.
17 Michils G, Hollants S, Dehaspe L *et al*: Molecular analysis of the breast cancer genes BRCA1 and BRCA2 using amplicon-based massive parallel pyrosequencing. *J Mol Diagn* 2012; **14**: 623–630.
18 Machado PM, Brandao RD, Cavaco BM *et al*: Screening for a BRCA2 rearrangement in high-risk breast/ovarian cancer families: evidence for a founder effect and analysis of the associated phenotypes. *J Clin Oncol* 2007; **25**: 2027–2034.
19 Bell CJ, Dinwiddie DL, Miller NA *et al*: Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 2011; **3**: 65ra64.
20 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.

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