AIMS Genetics

AIMS Genetics, 2(4): 263-280.
DOI: 10.3934/genet.2015.4.263
Received date 17 October 2015,
Accepted date 07 December 2015,
Published date 14 December 2015

http://www.aimspress.com/

Research article

Pros and cons of HaloPlex enrichment in cancer predisposition genetic diagnosis

Agnès Collet [1], Julien Tarabeux [1], Elodie Girard [3,4,5], Catherine Dubois D'Enghien [1], Lisa Golmard [1,2], Vivien Deshaies [3,4,5], Alban Lermine [3,4,5], Anthony Laugé [1], Virginie Moncoutier [1], Cédrick Lefol [1], Florence Copigny [1], Catherine Dehainault [1], Henrique Tenreiro [1], Christophe Guy [1], Khadija Abidallah [1], Catherine Barbaroux [1], Etienne Rouleau [1], Nicolas Servant [3,4,5], Antoine De Pauw [1], Dominique Stoppa-Lyonnet [1,2,6] and Claude Houdayer [1,2,7,]*

[1] Institut Curie, Département de Biopathologie, Paris, France
[2] Institut Curie, Inserm U830, Paris, France
[3] Institut Curie, Paris, France
[4] Inserm U900, Paris, France
[5] Mines ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, Fontainebleau, France
[6] Université Paris Descartes, Sorbonne Paris Cité, Paris, France
[7] Faculté des Sciences pharmaceutiques et biologiques, Université Paris Descartes, Sorbonne Paris Cité, Paris, France

* **Correspondence:** Email: claude.houdayer@curie.fr; Tel: +3-315-624-5837;
Fax: +3-315-310-2648.

Abstract: Panel sequencing is a practical option in genetic diagnosis. Enrichment and library preparation steps are critical in the diagnostic setting. In order to test the value of HaloPlex technology in diagnosis, we designed a custom oncogenetic panel including 62 genes. The procedure was tested on a training set of 71 controls and then blindly validated on 48 consecutive hereditary breast/ovarian cancer (HBOC) patients tested negative for *BRCA1/2* mutation. Libraries were sequenced on HiSeq2500 and data were analysed with our academic bioinformatics pipeline. Point mutations were detected using Varscan2, median size indels were detected using Pindel and large genomic rearrangements (LGR) were detected by DESeq. Proper coverage was obtained. However, highly variable read depth was observed within genes. Excluding pseudogene analysis, all point mutations were detected on the training set. All indels were also detected using Pindel. On the other hand, DESeq allowed LGR detection but with poor specificity, preventing its use in diagnostics. Mutations were detected in 8% of *BRCA1/2*-negative HBOC cases. HaloPlex technology appears to be an efficient and promising solution for gene panel diagnostics. Data analysis remains a major challenge and geneticists should enhance their bioinformatics knowledge in order to ensure good quality diagnostic results.

Keywords: HaloPlex; cancer; predisposition; genetic diagnosis; next generation sequencing; gene panel

1. Introduction

Individuals with hereditary predispositions to cancer are at an increased risk of developing specific cancers compared to the general population. Patients are usually evaluated on the basis of family history and/or individual criteria (age at diagnosis, tumor histology) followed by cascade testing for the most likely genes. However, in negative cases, with a complex family history and genetically heterogeneous diseases, cascade testing by Sanger sequencing appears to be time-consuming and expensive. With constant progress in next-generation sequencing (NGS) technologies and corresponding decreased costs, many diagnostic laboratories have been shifting from Sanger sequencing platforms to higher throughput NGS platforms [1–3]. NGS technologies now allow simultaneous analysis of multiple susceptibility genes for series of patients by gene panel sequencing. Typically, hereditary cancer panels include highly penetrant as well as moderately penetrant genes with known clinical utility and for which clinical guidelines concerning prevention or early detection have been established [4]. These genes are called “actionable genes”. The gene panel approach has obvious advantages compared to Sanger sequencing: increased throughput, decreased delays, optimized molecular diagnosis in patients with a family history suggestive of an inherited susceptibility to cancer.

The need for high-throughput technologies is also increased by the development of personalized medicine, which tries to use targeted therapies with improved selectivity and efficacy in preselected patient cohorts based on gene analysis. One example of molecularly targeted therapy is inhibition of poly (ADP-ribose) polymerase (PARP) enzyme by small molecule inhibitors in tumors harboring *BRCA1* or *BRCA2* mutations. Treatments such as olaparib (which has recently been approved for ovarian cancer therapy by the FDA and European commission in patients with platinum-sensitive, recurrent, high-grade serous ovarian cancer with *BRCA1* or *BRCA2* mutations) [5,6] forces diagnostic laboratories to provide even more rapid diagnostic delivery.

Up until recently, library preparation time really constituted the rate-limiting step in this approach, with the exception of multiplex PCR which is fast but associated with other issues that increase with target size (number of tubes, large genomic rearrangement analysis, and coverage). It is critical to find a method which is sufficiently rapid to allow a suitable waiting time for patients, while at the same time allowing large gene panel enrichment with high diagnostic quality criteria in terms of sensitivity and specificity.

We have consequently designed and tested a HaloPlex (Agilent, Santa Clara, USA) custom gene panel including all 62 genes studied in our laboratory that are involved or suspected to be involved in several diseases: Hereditary Breast and Ovarian Cancer (HBOC), digestive cancer, retinoblastoma, *DICER1* syndrome, ataxia-telangectasia, Fanconi anemia and Bloom syndrome. This panel was composed of actionable genes, moderately penetrant genes but also “research” genes i.e., with no known clinical validity. Libraries were then sequenced on a HiSeq (Illumina, San Diego, USA) sequencer and data were analysed with our academic bioinformatics pipeline and, in some cases, with the NextGENe software (SoftGenetics, State College, USA). The procedure was first tested on a training set of 71 challenging controls samples and then blindly validated on 48 samples. HaloPlex technology was found to be compatible with diagnostic requirements

2. Materials and Methods

2.1. Patients

All patients attended an interview with a geneticist and a genetic counsellor in a family cancer clinic in Institut Curie, Paris, France. Genetic testing was proposed on the basis of the patient’s personal history and/or family history in relation to various clinical presentations: breast and ovarian cancer, digestive cancer, retinoblastoma, *DICER1* syndrome, ataxia-telangectasia, Fanconi anemia and Bloom syndrome. Informed consent was obtained from all patients or their legal guardians. DNA was extracted from leucocytes using the Quickgene 610-L automated system (FujiFilm, Tokyo, Japan) according to the manufacturer’s instructions. A series of 119 patients was studied: 71 as a training set and 48 as a diagnostic set.

The training set was composed of 71 patients previously Sanger sequenced (or by multiplex ligation-dependent probe amplification analysed—MLPA) and harboring 67 representative variations and 98 polymorphisms were used as controls. To adequately address diagnostic issues, this training set was composed of difficult cases (indels, large rearrangements) and at least, one mutation for each gene concerned by our clinical diagnostic activity.

The diagnostic set included 48 consecutive cases from our family cancer clinics who had been previously tested negative for *BRCA1/2* mutations and at high risk of cancer genetic predisposition based on personal or family cancer history.

2.2. Library preparation and sequencing

A custom 62-genes panel was created using Suredesign software (Agilent, Santa Clara, USA) (Table 1). Region of interest (ROI) was defined as coding sequences and flanking splice consensus sequences from genes of interest (padding: −20/+10 bp) with 13,266 different amplicons [7]. Three successive designs were necessary to obtain satisfactory coverage of the complete ROI, especially for *BRCA1* and *BRCA2* coding sequences. Additional genes were introduced in the course of these successive designs (Table 1). Target enrichment was performed according to the manufacturer’s instructions. Briefly, DNAs were fragmented using a cocktail of 8 restriction enzymes, and denatured. A probe library was added and hybridized to targeted fragments. Each probe was an oligonucleotide designed to hybridize to both ends of a targeted DNA restriction fragment, thereby guiding the targeted fragments to form circular DNA molecules. The probe also contained a method-specific sequencing motif and a sample barcode, both incorporated during circularization. HaloPlex probes are biotinylated and targeted fragments can therefore be retrieved with magnetic streptavidin beads. The circular molecules were then closed by ligation, ensuring that only perfectly hybridized fragments were circularized. Only circular DNA targets are amplified, providing an enriched and bar-coded amplification product that is ready for sequencing. All libraries of target-enriched pooled DNA were analysed on LabChip (Caliper, PerkinElmer, Waltham, MA, USA) to assess successful enrichment, demonstrating a smear of amplicons ranging from 50 bp to 500–600 bp with a mean at 200 bp. Following enrichment, samples were sequenced on a HiSeq2500 (Illumina) with the fast module using the 150 paired-end chemistry according to the manufacturer's instructions.

Table 1. Description of the 62-gene panel.

Name	Panel design version at gene inclusion	V1 mean coverage (%)	V2 mean coverage (%)	V3 mean coverage (%)
APC	v1	99.86	99.85	100.00
ATM	v1	98.94	99.69	99.99
ATR	v1	99.83	99.12	99.95
BAP1	v1	100	100.00	100.00
BARD1	v1	100	100.00	100.00
BLM	v1	99.48	99.81	99.99
BRCA1	v1	99.4	99.52	99.79
BRCA2	v1	99.58	99.91	100.00
BRIP1	v1	99.97	99.94	100.00
CDH1	v1	99.98	99.99	100.00
CHEK2	v1	94.03	99.73	100.00
DICER1	v1	99.98	99.98	100.00
EPCAM	v1	100	100.00	100.00
FANCA	v1	99.9	99.92	99.93
FANCB	v1	99.55	99.20	100.00
FANCC	v1	100	100.00	100.00
FANCE	v1	100	99.92	100.00
FANCF	v1	100	100.00	100.00
FANCG	v1	99.39	99.95	100.00
FANCI	v1	99.89	99.94	100.00
FANCL	v1	99.93	97.26	100.00
FANCM	v1	99.9	99.62	99.96
MET	v1	100	99.92	100.00
MLH1	v1	100	100.00	100.00
MRE11A	v1	97.8	99.99	99.96
MSH2	v1	98.11	99.82	100.00
MSH6	v1	100	99.97	100.00
MUTYH	v1	100	99.92	99.92
PALB2	v1	100	100.00	100.00
NBN	v1	99.73	99.29	100.00
RAD50	v1	100	99.96	100.00
RAD51B	v1	94.27	100.00	100.00
RAD51C	v1	98.82	99.01	100.00

RAD51D	v1	99.94	100.00	100.00
RB1	v1	96.44	97.03	95.51
SLX4	v1	99.85	99.94	100.00
SMARCB1	v1	100	100.00	100.00
STK11	v1	98.6	100.00	100.00
WT1	v1	95.81	96.81	98.84
XRCC2	v1	100	100.00	100.00
XRCC3	v1	100	100.00	100.00
ERCC4	v2	-	99.58	100.00
HELQ	v2	-	99.06	100.00
MED4	v2	-	100.00	100.00
MDM2	v2	-	100.00	100.00
BMPR1A	v3	-	-	100.00
CDKN2A	v3	-	-	100.00
FAN1	v3	-	-	100.00
POLD1	v3	-	-	99.98
POLE	v3	-	-	100.00
PTEN	v3	-	-	99.01
RINT1	v3	-	-	99.98
SMAD4	v3	-	-	100.00
TP53	v3	-	-	100.00

Three different versions with an increased number of genes were designed: V1, V2, V3.

Mean coverage observed for the ROI at minimum 30X are reported for each gene. *FANCD2* and *PMS2* are not reported as analysis was not reliable (see text for details).

2.3. Bioinformatics analysis

Both commercial (NextGENe, SoftGenetics) and academic solutions were used as described below. However, NextGENe® v.2.3.1 (SoftGenetics, State College, PA, USA) was available to analyse 2/3 of the training set, making comparison between academic and commercial solutions difficult. Default settings for paired-end Illumina data were used to filter and trim the raw data. The adapter sequence was trimmed at the same time according to a text file. Parameters for alignment and mutation detection were: paired read analysis, “Allowable Mismatched bases” option set to 0, “Allowable Ambiguous Alignments set to 50, 35 bp seed, 7 bases move step, 85% matching base percentage, “detect large indels” option on, 15% mutation percentage.

In addition, we designed our own academic bioinformatics pipeline and applied it on all patients. Index demultiplexing and generation of raw data were performed with Illumina Consensus Assessment of Sequence and Variation (CASAVA) software (v1.8.2, Illumina, San Diego, CA). Adapter trimming was performed with SeqPrep software (v1.0) (https://github.com/jstjohn/SeqPrep). Mapping was performed with Bowtie2 (v2.1.0) [8] on human hg19 reference genome using the sensitive mode. Insert size for valid paired-end alignments was set between 0 and 600 bp. Data were then processed using MPileup (Samtools, v0.1.18) [9] targeted on our region of interest with the following parameters: Do not perform Genotype Likelihood Computation, Do not skip anomalous read pairs in variant calling (-A), Disable probabilistic realignment for the computation of base alignment quality (-B), no mapping quality adjustment for reads containing excessive mismatches (–C 0), Max per-BAM depth was set to 10,000 (-d 10000) to avoid read downsampling. Minimum mapping quality for an alignment to be used was set to 0 and minimum base quality for a base to be considered was set to 12. Point mutation calling was performed using VarScan2 [10] with the following parameters: only bases covered by at least 30 reads are considered, 2 of which must carry the alternative variant. The minimum allelic ratio for a variant to be reported was set to 15%. Variants supported by more than 90% of reads on the same strand were included in the analysis. Variant annotation was performed using Annovar (25/10/2013) [11]. Statistics on the ROI coverage were established using two metrics: the percentage of bases covered by at least 30 reads specified by isoform for each gene and the number and localization of bases covered by less than 30 reads.

Indels of intermediate size (5 bp and larger) were called using Pindel (v0.2.5) [12] with already aligned data. Pindel first extracts reads indicating a potential or already existing indel and uses this position as an anchor, it then splits reads (as well as their potential unmapped counterpart) into several pieces to try to find a better alignment including an indel. Pindel was run with default parameters apart from the following two parameters: in this setting of very deep sequencing, the minimum number of reads supporting an event to be reported was set to 10 and the “insert-size” (defined as the length of sequence between the paired-end adapters) was set to 500bp. Only variants with an allelic ratio greater than 5% were reported.

To manage large genomic rearrangements (LGR) detection, the R package DESeq [13] was used to normalize read counts and estimate fold change per sample and window, fitting a generalized linear model to these normalized counts. In order to distinguish a gene deletion on the X chromosome in a woman from a man with only one copy, the patient’s gender was taken into account for genes located on the X chromosome, dividing the analysis into two sub-analyses. Fold changes thresholds were then estimated on the basis of validated data and were used to highlight potential events.

Analysis filters and pipeline were established and tested using the training set and were then used to call variants in the diagnostic set. All mutations are reported according to the Human Genome Variation Society (HGVS) guidelines. References of coding sequences used for each gene are reported in supplementary Table 1.

2.4. Sanger sequencing confirmation

All point mutations passing these filters in the diagnostic set were confirmed by Sanger sequencing using Big Dye Terminator on the ABI 3500XL (Life Technologies, Carlsbad, USA). No LGR needed to be confirmed in the diagnostic set (none were detected).

2.5. Data availability

All detected mutations were registered to relevant data bases such as UMD data base for breast or colon cancers or LOVD for Fanconi anemia. Sequencing data and reads have already been provided to a national NGS consortium (INCa DGOS/NGS network). At the end of the project, data sould be made available for global community.

3. Results

We studied DNA samples from 119 different patients: 71 previously characterized samples were used to evaluate library preparation and define the appropriate settings for the bioinformatics analysis. Forty-eight clinical samples referred for testing were analysed. Three separate runs containing either 44 or 43 samples were performed on the HiSeq 2500 Sequencing System with the 150 bp paired end sequencing module. Each HiSeq run produced an average of 550 million reads. Run time was 40 hours. The average on-target ratio was 82%. An average of 200 variants were called for each patient i.e. 3 variants per gene.

3.1. Read depth and coverage

No significant decrease in coverage was observed between the three design versions (Table 1) and 99.4% of the target was covered by at least 100X. The average read depth was 3720X with marked heterogeneity between the different genes of our panel with a seven-fold difference between “poor performers” (e.g. *MDM2* and *XRCC2*) and “good performers” (e.g. *FANCA* and *MUTYH*) (Figure 1). Read depth also varies considerably in different parts of the same gene. In *BRCA1*, for example, read depth can easily vary from less than 100X (minimum value: 1X) to more than 6000X (maximum value: 10013X) (Figure 2).

3.2. Training set

The 165 variants present in the training set had been previously identified by Sanger sequencing or MLPA. Sixty-seven mutations had a potential or demonstrated biological effect (Table 2) and 98 were polymorphisms (Supplementary Table S1). Library preparation using HaloPlex technology was successful for all but 1 of the 71 DNA samples of the training set.

All causative point mutations and polymorphisms were identified using the Varscan2 bioinformatics pipeline (Supplementary Table S1).

On a large set of indels ranging from 1 to 50 bp, rearrangements longer than 20 bp could not be detected by either Varscan or NextGENe. The first missed indel was a 23 bp duplication in the *RB1* gene (g.2113_2135dup, p.Pro26Argfs*47) and the second missed indel was a 50 bp insertion in the *BRCA1* gene (c.3729_3730ins50). Pindel with read alignment was therefore used and all indels were detected.

The allelic ratios usually observed with constitutive heterozygous mutations are close to 0.50, but, in this training set, several true constitutive heterozygous mutations were detected with allelic ratios ranging from 0.2 to 0.67 (Table 2). Similarly, several real constitutive heterozygous indels were detected with allelic ratios ranging from 0.18 to 0.8 (Table 2).

A high rate of false recurrent SNVs was observed in our results (recurrence ranging from 25 to 100% of samples and allelic ratios ranging from 1 to 20%). They were always located at the read extremities (in either the forward or reverse reads). Read trimming was then used to avoid these recurrent false positives (see Discussion).

Varscan2, Pindel and NextGENe were unable to detect LGRs. The addition of DESeq in our pipeline allowed the detection of all LGRs, but with poor specificity, preventing its use in diagnostics. For example, in *BRCA1* LGR analysis, we observed 9 false duplications and 10 false deletions in the first training run of 44 samples in addition to the only true *BRCA1* LGR present in this set (exons 3 to 8 duplication).

3.3. Diagnostic set

Gene panel analysis was performed for 48 HBOC patients without *BRCA1/2* mutation. Mutations were detected in 4 (8%) patients in *ATM*, *BRCA2*, *FANCA*, *FANCM* and *PALB2* genes; the *FANCM* and *PALB2* mutations were present in the same patient. Likely deleterious variants were also detected in 4 (8%) patients in *BLM*, *BRIP1*, *CHEK2*, *FANCG* and *NBN*; the *BRIP1* and *CHEK2* variants were present in the same patient (Table 3).

Two *RB1* mutations in the 131 patients were missed by NextGENe, but this problem was resolved by modifying the "allowable mismatch base" from 0 to 2.

4. Discussion

We describe and validate a HaloPlex-based diagnostic pipeline applied to a gene panel. Some technical comments and guidelines for implementation and use based on our experience are discussed below.

The wet lab part of the protocol is easy to implement, as it consists of kits that are easy to handle by any molecular biology laboratory. According to the manufacturer, major protocol steps can be automated. The HaloPlex target enrichment system was able to very rapidly capture and sequence the genomic regions of interest of 62 genes (4 days for a 44-sample library preparation).

DNA capture is performed by DNA fragment circularization with custom probes after enzymatic fragmentation (eight different restriction enzymes). Enrichment by circularization considerably facilitates library preparation for NGS, as sequencing primers can be added to the circularization probe, thereby eliminating the need for any further library preparation steps [14]. This technology achieves high specificity and output for library preparation from small DNA quantities. As a result it could be relevant for PARP inhibitor therapies as *BRCA* sequencing is also performed with DNA extracted from FFPE materials. Nevertheless, optimization would be needed.

However, using Haloplex technology also introduces several biases, especially concerning the homogeneity of gene coverage. These two aspects will be discussed below.

HaloPlex appears to be a specific library preparation technology: 82% of reads were mapped on target. This probably results from the combination of restriction enzyme and circularization by probe hybridization. Three different designs were necessary to achieve satisfactory gene coverage, especially by the addition of probes in *BRCA1* and *BRCA2* genes. Of note, no detrimental impact of the additional probes on the previous design was observed.

However, HaloPlex technology induces considerable read depth heterogeneity between the various sequences of interest due to the use of restriction enzymes for DNA fragmentation with the constraints of corresponding restriction maps. Gaps in coverage will therefore be observed when the distance between two restriction sites is longer than the read length. Inserts designed to be > 300 bases are underrepresented in the sequencing data and this under-representation results in coverage gaps when there is insufficient redundancy over the region [15]. GC content did not impact coverage in this gene panel design.

Table 2. Training set.

Gene	Transcript Reference[1]	Mutation description[2]	Expected consequence[3]	Type	Detectable with our pipeline	AR[d] reported by our pipeline	Identified by NextGene	AR[d] reported by NextGene
ATM	NM_000051.3	c.1402_1403del	p.Lys468Glufs*18	Frameshift deletion	yes	0.36	yes	0.43
BAP1	NM_004656.3	c.639dup	p.Ile214Tyrfs*29	Frameshift insertion	yes	0.49	yes	0.48
BLM	NM_000057.2	c.1544dup	p.Asn515Lysfs*2	Frameshift insertion	yes	0.48	yes	0.45
BLM	NM_000057.2	c.1642C>T	p.Gln548*	Nonsense	yes	0.48	yes	0.48
BLM	NM_000057.2	c.2119C>T	p.Pro707Ser	Missense	yes	0.49	yes	0.49
BRCA1	NM_007294.3	exons 3 to 8 duplication	p. ?	LGR	yes	NA	no	*NA*
BRCA1	NM_007294.3	c.3729_3730ins50	p.His1244Aspfs*8	Frameshift insertion	yes (with Pindel)	NA	no	*NA*
BRCA1	NM_007294.3	exons 3 to 8 duplication	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
BRCA1	NM_007294.3	c.4393A>C	p.Ile1465Leu	Missense	yes	0.49	*NA*	*NA*
BRCA1	NM_007294.3	c.2292_2310dup19	p.Leu771Argfs*3	Frameshift insertion	yes	0.18	*NA*	*NA*
BRCA1	NM_007294.3	c.1016dup	p.Val340Glyfs*6	Frameshift insertion	yes	0.52	*NA*	*NA*
BRCA1	NM_007294.3	c.1487G>A	p.Arg496His	Missense	yes	0.52	*NA*	*NA*
BRCA1	NM_007294.3	c.1961del	p.Lys654Serfs*47	Frameshift deletion	yes	0.34	*NA*	*NA*
BRCA1	NM_007294.3	exons 1 to 7 deletion	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
BRCA2	NM_000059.3	c.1813dup	p.Ile605Asnfs*11	Frameshift insertion	yes	0.47	*NA*	*NA*
CDH1	NM_004360.3	c.586G>T	p.Gly196*	Nonsense	yes	0.46	*NA*	*NA*
CDH1	NM_004360.3	exon 3 deletion	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
DICER1	NM_177438.2	c.1922T>A	p.Leu641*	Nonsense	yes	0.51	yes	0.49
DICER1	NM_177438.2	c.5235del	p.Phe1745Leufs*6	Frameshift deletion	yes	0.47	*NA*	*NA*
FANCA	NM_000135.2	c.1490C>T	p.Pro497Leu	Missense	yes	0.49	yes	0.5
FANCA	NM_000135.2	exons 15 to 21 duplication	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
FANCA	NM_000135.2	exons 7 to 11 deletion	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
FANCA	NM_000135.2	exon 31 duplication	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
FANCA	NM_000135.2	c.3788_3790del	p.Phe1263del	Inframe deletion	yes	0.49	yes	0.47

FANCA	NM_000135.2	c.3764_3765insAGGA	p.Leu1256Glyfs*23	Frameshift insertion	yes	0.28	*NA*	*NA*
FANCA	NM_000135.2	c.3164G>A	p.Arg1055Gln	Missense	yes	0.48	*NA*	*NA*
FANCA	NM_000135.2	c.2574C>G	p.Ser858Arg	Missense	yes	0.51	*NA*	*NA*
FANCB	NM_152633.2	exons 2 to 9 deletion	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
FANCD2	NM_001018115.1	exons 4 to 12 deletion	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
FANCF	NM_022725.3	c.399_407dup	p.Leu136_Arg138dup	Inframe duplication	yes	0.21	no	*NA*
FANCG	NM_004629.1	c.271_280del9insT	p.Asp91_Ala93>Serfs*11	Frameshift indel	yes	0.42	yes	0.44
FANCG	NM_004629.1	c.620del	p.Leu207Profs*2	Frameshift deletion	yes	0.46	yes	0.52
FANCG	NM_004629.1	c.1182_1192delinsC	p.Glu395Trpfs*5	Frameshift deletion	yes	0.8	*NA*	*NA*
FANCL	NM_001114636.1	c.1036-2A>T	p. ?	Splicing defect	yes	0.99	yes	0.99
FANCL	NM_001114636.1	c.1022_1024del	p.Ile341_Cys342delinsSer	Inframe deletion	yes	0.48	yes	0.48
FANCL	NM_001114636.1	c.919-2A>G	p. ?	Splicing defect	yes	0.51	yes	0.51
FANCM	NM_020937.2	c.2586_2589del	p.Lys863Ilefs*12	Frameshift deletion	yes	0.99	yes	1
MET	NM_001127500.1	c.3712G>A	p.Val1238Ile	Missense	yes	0.49	*NA*	*NA*
MRE11A	NM_005591.3	c.424G>A	p.Asp142Asn	Missense	yes	0.45	yes	0.45
MRE11A	NM_005591.3	c.544G>A	p.Gly182Arg	Missense	yes	0.34	yes	0.34
MSH2	NM_000251.2	c.942+3A>T	p. ?	Splicing defect	yes	0.47	*NA*	*NA*
NBN	NM_002485.4	c.330T>G	p.Tyr110*	Nonsense	yes	0.48	yes	0.48
NBN	NM_002485.4	c.1125G>A	p.Trp375*	Nonsense	yes	0.5	yes	0.5
PALB2	NM_024675.3	exon 7 deletion	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
RAD50	NM_005732.3	complete gene deletion	p. ?	LGR	yes (with DESeq)	NA	no	*NA*
RAD51B	NM_133509.3	c.728A>G	p.Lys243Arg	Missense	yes	0.54	yes	0.55
RAD51C	NM_058216.1	c.1026+5_1026+7del	p. ?	Splicing defect	yes	0.42	yes	0.42
RAD51D	NM_002878.3	c.26G>C	p.Cys9Ser	Missense	yes	0.5	yes	0.49
RB1	NM_000321.2	g.2113_2135dup	p.Pro26Argfs*47	Frameshift insertion	yes (with Pindel)	0	no	*NA*
RB1	NM_000321.2	g.2182_2195delinsGCC	p.Asp41Glufs*4	Frameshift deletion	yes	0.27	yes	0.27
RB1	NM_000321.2	c.1463dup	p.Cys489Valfs*4	Frameshift insertion	yes	0.49	yes	0.48
RB1	NM_000321.2	c.1398del	p.glu466Asnfs*12	Frameshift deletion	yes	0.73	no	coverage <15X
RB1	NM_000321.2	c.1613del	p.Ala538Glufs*5	Frameshift deletion	yes	0.47	yes	0.47
RB1	NM_000321.2	c.2288_2289del	p.Arg763Thrfs*31	Frameshift deletion	yes	0.51	*NA*	*NA*

RB1	NM_000321.2	c.575_576del	p.Lys192Serfs*10	Frameshift deletion	yes	0.25	*NA*	*NA*
RB1	NM_000321.2	c.468_472del	p.156_158del	Frameshift deletion	yes	0.3	*NA*	*NA*
RB1	NM_000321.2	c.1498+3A>C	p. ?	Splicing defect	yes	0.48	*NA*	*NA*
RB1	NM_000321.2	c.1147C>T	p.Gln383*	Nonsense	yes	0.98	*NA*	*NA*
RB1	NM_000321.2	c.751C>T	p.Arg251*	Nonsense	yes	0.2	*NA*	*NA*
RB1	NM_000321.2	c.1954del	p.Val654Cysfs*4	Frameshift deletion	yes	0.99	*NA*	*NA*
RB1	NM_000321.2	c.763C>T	p.Arg255*	Nonsense	yes	1	*NA*	*NA*
RB1	NM_000321.2	c.1072C>T	p.Arg358*	Nonsense	yes	0.97	*NA*	*NA*
RB1	NM_000321.2	c.1954dup	p.Val654Serfs*14	Frameshift insertion	yes	0.5	*NA*	*NA*
RB1	NM_000321.2	c.1847A>T	p.Lys616Ile	Missense	yes	0.67	*NA*	*NA*
RB1	NM_000321.2	c.1846_1847insT	p.Lys616Ilefs*37	Frameshift insertion	yes	0.6	*NA*	*NA*
XRCC2	NM_005431.1	c.450C>G	p.Ser150Arg	Missense	yes	0.5	yes	0.5
XRCC3	NM_005432.3	c.448C>T	p.Arg150Cys	Missense	yes	0.46	yes	0.47

Mutations previously identified by sequencing are reported with their corresponding RefSeq.

[1] Nomenclature was numbered on the basis of the following transcripts, [2] Mutation nomenclature according to HGVS recommendations, nucleotide position was numbered with +1 corresponding to the A of the ATG of the translation initiation codon’ [3] Expected consequence on the protein level, [d]Allelic ratios are defined as the ratio of the non-reference allele to the sum of the non-reference allele and the reference allele. Allelic ratio for heterozygous variants should be centered around 0.5. Abbreviations: LGR, large genomic rearrangement; NA, not applicable; AR, allelic ratio

Table 3. Diagnostic set.

ID	Gene	Variant	Variant class	Predicted effect on protein (Align-GVGD class; SIFT prediction)	Personal cancer history (age at diagnosis)	Family cancer history (age at diagnosis)	Allele count in controls (frequency)
1	*ATM*	c.5460del, p.Lys1820Asnfs*8	Frameshift deletion	Unstable or truncated protein	BC (46), OC (57)	Sister, bilateral BC (63,63); sister, BC (50) / Maternal branch: mother, bilateral BC (60,70); aunt, BC (58); uncle, PrC (64); uncle, PrC (80)	-
2	*BLM*	c.2489C>G, p.Thr830Arg	Likely deleterious missense variant	Highly conserved amino acid, Grantham 71 (C65;	BC (52), PaC (52)	Maternal branch: mother, BC (69); aunt, BC (49); aunt, UC (76)	-

				deleterious), in helicase domain			
3	*BRCA2*	c.7617+2T>G	Splicing mutation	Unstable or truncated protein	Bilateral BC (35,35)	Sister, BC (46) / Paternal branch: father, BC (56); grandfather, NHL	-
4	*BRIP1*	c.2469G>T, p.Arg823Ser	Likely deleterious missense variant	Highly conserved amino acid, Grantham 110 (C65; deleterious), in helicase domain	OC (48)	Paternal branch: father, LC (53); uncle, PM (76); grandmother, bilateral BC (31,33) / Maternal branch: uncle, KC (74); aunt, NHL (69); grandmother, BC (56)	-
4	*CHEK2*	c.1399T>C, p.Tyr467His	Likely deleterious missense variant	Highly conserved amino acid, Grantham 83 (C65; deleterious), in catalytic domain	OC (48)	Paternal branch: father, LC (53); uncle, PM (76); grandmother, bilateral BC (31,33) / Maternal branch: uncle, KC (74); aunt, NHL (69); grandmother, BC (56)	3/8600 (0.03%)[1]
5	*FANCA*	c.189+1G>A	Splicing mutation	Unstable or truncated protein	Bilateral BC (31,75), PaC (78)	Sister, BC (46) and her sister's daughter, BC (39); daughter, BC (48) / Maternal branch: mother, OC (74); aunt, BC (63); aunt, OC (58)	-
6	*FANCG*	c.722C>T, p.Pro241Leu	Likely deleterious missense variant	Highly conserved amino acid, Grantham 98 (C65; deleterious)	Bilateral BC (40,62), TC (60), PHNET (62)	Maternal branch: mother, BC (54); aunt, bilateral BC (50,65); aunt, BC (64); aunt, BC (69); cousin, BC (42); cousin, BC (62); grandmother, BC (85); grandfather, IC (68)	-
7	*FANCM*	c.1196C>G, p.Ser399*	Nonsense mutation	Unstable or truncated protein	Male BC (54)	Sister, BC (39); father, CC	-
7	*PALB2*	c.886dup, p.Met296Asnfs*7	Frameshift duplication	Unstable or truncated protein	Male BC (54)	Sister, BC (39); father, CC	-
8	*NBN*	c.643C>T, p.Arg215Trp	Likely deleterious missense variant	Moderately conserved amino acid, Grantham 101 (C0; deleterious)	BC (36)	Maternal branch: mother, BC (55); sister, BC (69) and her sister's daughter, BC (48); sister, BC (70) and her sister's daughter, UC (28); sister, BC (40)	32/8592 (0.4%)[1] or 5/2184 (0.2%)[2]

Mutations and likely deleterious variants detected in HBOC syndrome patients. Allele counting control is from Exome variant Server[1] and 1000 genomes[2].

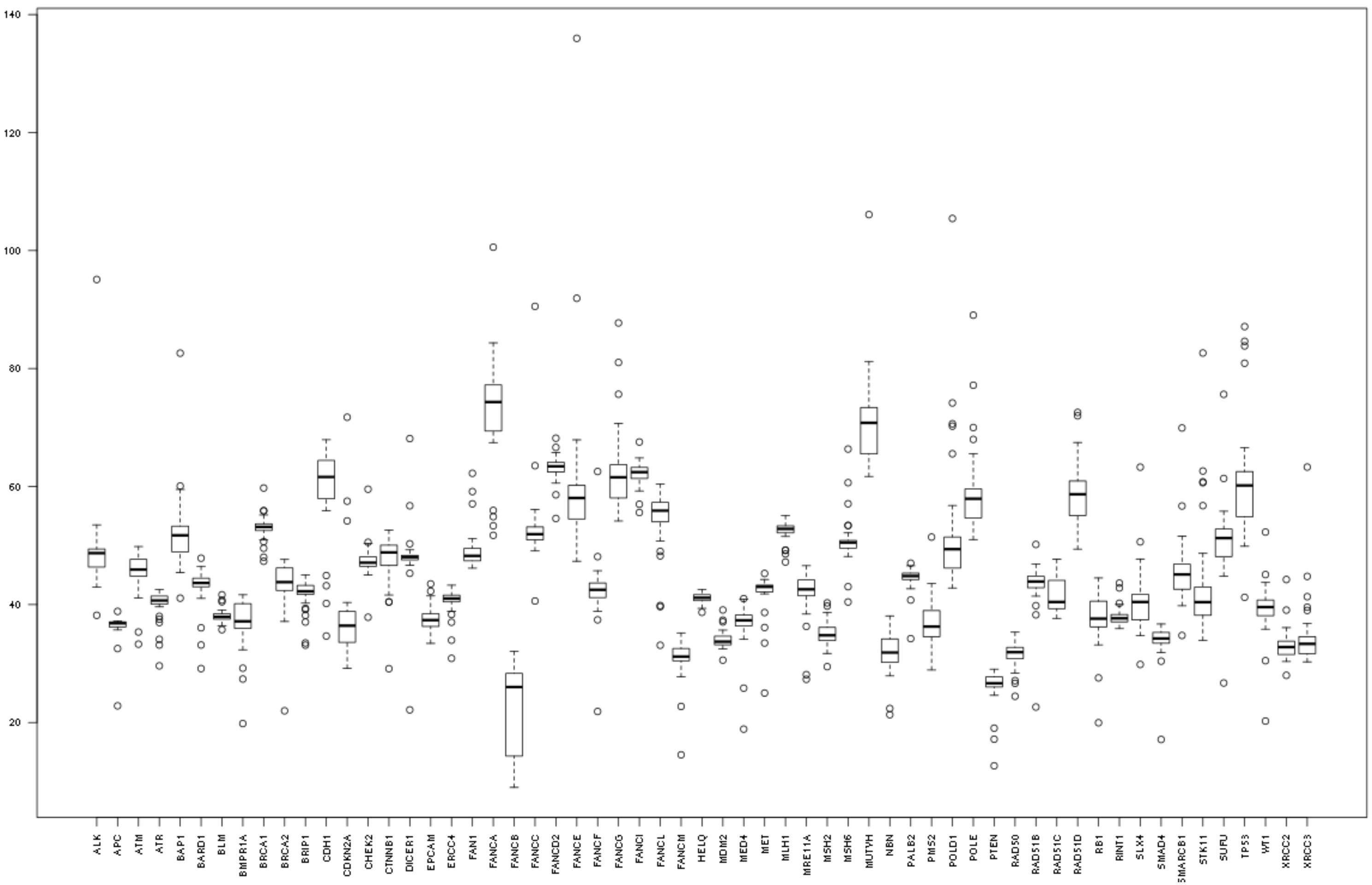

Figure 1. Read depth heterogeneity. Heterogeneity is observed between genes, but also within a gene (e.g. *FANCB*), although to a lesser extend. Boxplots representing the distribution of read counts per gene, normalized by the library size and by the gene size. Each box represents the distribution of values between the first and third quartiles and the median is characterized by a black line. Longer boxplots indicate a higher variability in the normalized read counts distribution and outlier's values are depicted by individual points. x-axis: genes included in the panel, y-axis: number of reads normalized by library and gene sizes.

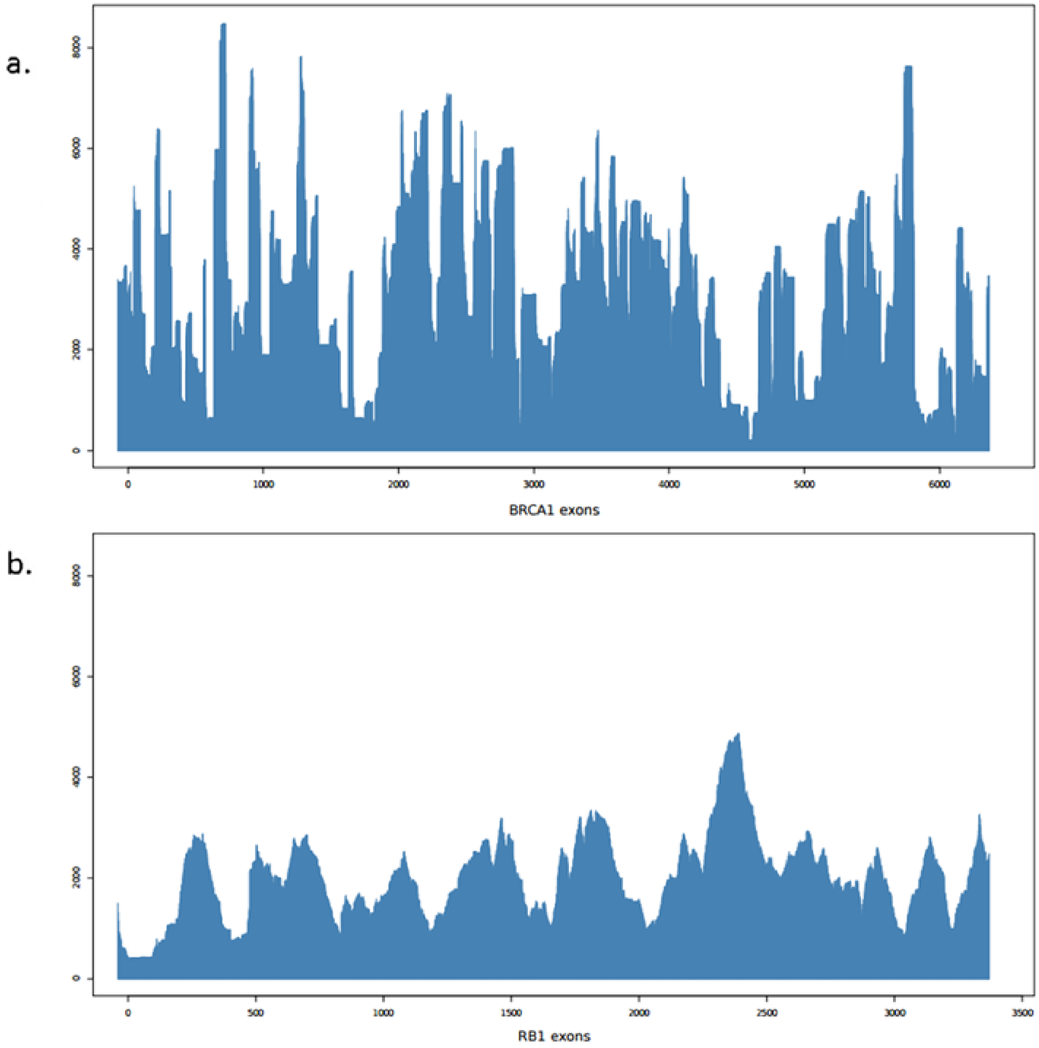

Figure 2. **Coverage heterogeneity in *BRCA1* (a) and *RB1* (b) Average read depth is represented for the entire gene coding sequence at each cDNA nucleotide position (extreme values ranging from 1X to more than 1000X). x-axis: length of the coding sequence, y-axis: average read depth.**

Moreover, probe hybridization regions are limited to a few bases so that, when mismatch occurs, inserts are dropped out and read depth heterogeneity increases at the corresponding target sequences. The HaloPlex design includes redundancy, so targets should be covered even when inserts are dropped out, but marked read depth heterogeneity implies a greater sequencing capacity. A technology that typically has high specificity and uniformity will require less sequencing to generate adequate coverage of sequence data for the downstream analysis, making sequencing more economical. Although HaloPlex provided good specificity in terms of library preparation, coverage heterogeneity constitutes a weak point.

These insert drop outs explain why real constitutive mutations were called at a ratio different from the expected allelic ratio of 0.5, which can range from 0.18 to 0.8, resulting in a marked bias. Moreover, HaloPlex probes hybridize on one strand generating large insert sizes. Consequently, strand biases cannot be used with 150 bp paired-end sequencing to distinguish true and false positive variants [2]. We therefore decided to set the detection threshold in our bioinformatics pipeline to 0.15 for constitutive mutations and 0.05 for mosaicism detection for three specific genes in routine diagnostics (*RB1*, *DICER1* and *APC*) to avoid missing any real mutations. As HaloPlex technology hybridizes only one DNA strand, recurrence criteria were then applied to filter the variants detected with these thresholds.

Our design also included *FANCD2* and *PMS2*, but sequencing quality was too low for diagnostic procedures. Two polymorphisms were indeed missed in *FANCD2* and it can be explained by a well known weakness of bioinformatic analysis i.e., the presence of 2 *FANCD2* pseudogenes [16]. The same holds true for *PMS2*. The reason is that pseudogenes are amplified and sequenced simultaneously and corresponding reads are mixed at the mapping steps. Consequently, variants are called at a low allelic ratio. To avoid this specificity issue, we usually use long-range PCR (7 superamplicons from 2.5 kb to 8.9 kb designed with specific primers) followed by Sanger sequencing to analyse *FANCD2*.

We also tried to exclude pseudogenes from the mapping bed file in order to restrict mapping to *FANCD2*. However, if true variants were correctly found, specificity was too poor, preventing its use in diagnosis. Comparison of the data obtained with this specific approach and *FANCD2* long-range PCR results could be used to define a list of pseudogene variants. Then these variants could be systematically subtracted from NGS analysis in order to focus on specific *FANCD2* variants. Consequently, althought *FANCD2* and *PMS2* are included in the panel design, results were not reported due to low reliability.

Recurrent false-positives were observed at the read extremities (on both forward and reverse reads) and appear to be related to adapter remains, even after using adapter removal software. To improve the quality of the trimmed read, Gr éen et al. proposed further trimming of sequence reads by five bases at the 3' end [17]. We tried another approach consisting of a two-step trimming procedure: adapters were first trimmed, followed by trimming of each single extremity nucleotide of sequence reads. The number of recurrent false-positives was considerably reduced by this procedure with no impact on coverage. Recent optimization of the SureCall software version 1.1.0.15 (Agilent technologies) should also overcome this issue [18].

All point mutations were detected on the training set. However, the same does not apply to larger indels. Small variants can be detected in NGS data by VarScan2 by allowing mismatches and gap opening during read alignment on the reference genome. Each alignment is defined by a score that is calculated on the basis of the number of bases that match correctly on the genome, but also on the number of mismatches and gaps, as well as their size. Inserting a gap in an alignment decreases the alignment score and it is accentuated with its length. If the alignment score decreases below a defined threshold, alignment is not reported. It is therefore very difficult to detect insertions or deletions larger than 20 bp with a tool dedicated to call small variants on certain specific alignment data without dedicated processing steps. In fact, reads that differ excessively (by more than 20% of the sequence) from the reference are usually trimmed. Indels were therefore often partially observed at read extremities, but in insufficient proportions to be detected, which is why we chose to use Pindel in addition to the "VarScan2" bioinformatics pipeline to detect all indels [12]. LGR detection required the use of another bioinformatics tool, which is why we used DESeq to improve our pipeline. Read counts can be compared between samples in order to detect large rearrangements in this particular design, based on the hypothesis that a rearrangement event is unlikely to be recurrent inside a run and that, as this constitutional level, the expected copy number is 2. To increase sensitivity and in order to detect all of the training set LGR, read counts were calculated per exon and large exons were split into 300 bp windows using BEDtools (V2.21), but real variants were then mixed with many false LGR. Poor performers i.e. samples with low coverage below 30X or with low DNA quality, were subsequently removed from analysis, but without resulting in any major improvement. To further enhance the throughput of this technology, the detection capacity for large rearrangements must be developed in the future.

Overall, our study shows that the combination of Varscan2 and Pindel analysis on HaloPlex library is efficient for the detection of small and medium-sized mutations (100% sensitivity on the training set). Actually, "real life" sensitivity is probably a bit lower as 13 genes didn't reach 100% coverage (i.e., 95.5% to 99.9%, see Table 1). LGR detection remains an issue in clinical practice: althought DESEq correctly identified all LGR, lack of specificity would mandate confirmatory MLPA in all patients, prohibiting its use in routine practice.

This process is available in the Galaxy framework [19–21]. It must be stressed that important bioinformatics resources are needed to process and store data compatible with diagnostic practice i.e. with fast turn-around time and secure storage.

As previously described, we found that distinct bioinformatics parameters had a marked impact on the results [22]. Validation of the results must be based on a good understanding and better control of data analysis pipelines. Biologists must adopt a critical approach and mistrust black box solutions.

5. Conclusion

We show that the HaloPlex technology is compatible with oncogenetic diagnostic activity. One single process to sequence all of our genes of interest at the same time represents a major improvement in the laboratory organization by being less time- and personnel-consuming. This process can also be more efficient by proposing actionable gene analysis in addition to analysis of the principal genes analysed in the clinical setting.

Using the whole gene panel in unexplained family cancer histories also represents a major improvement, as we were able to detect 4 new mutations in *BRCA1/2*-negative cases.

This study also demonstrates that data analysis remains a major issue. Geneticists must be actively involved in data analysis based on a good understanding of bioinformatics pipelines to avoid reporting poor quality results.

Acknowledgments

This work was supported by INCa/DGOS (grant NGS and Cancer) and ICGex. NGS projects were also supported by grants from Canceropôle Ile de France and the Region Ile de France (Seq-Start).

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Sie AS, Prins JB, van Zelst-Stams WG, et al. (2015) Patient experiences with gene panels based on exome sequencing in clinical diagnostics: high acceptance and low distress. *Clin genet* 87: 319–326.
2. Tarabeux J, Zeitouni B, Moncoutier V, et al. (2014) Streamlined ion torrent PGM-based diagnostics: BRCA1 and BRCA2 genes as a model. *Eur j hum genet* 22: 535–541.
3. Domchek SM, Bradbury A, Garber JE, et al. (2013) Multiplex genetic testing for cancer susceptibility: out on the high wire without a net? *J clin oncol* 31: 1267–1270.
4. Easton DF, Pharoah PDP, Antoniou AC, et al. (2015) Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *N engl j med* 372: 2243–2257.
5. Audeh MW, Carmichael J, Penson RT, et al. (2010) Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet* 376: 245–251.
6. Guttmacher AE, McGuire AL, Ponder B, et al. (2010) Personalized genomic information: preparing for the future of genetic medicine. *Nat rev genet* 11: 161–165.
7. Houdayer C, Caux-Moncoutier V, Krieger S, et al. (2012) 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* 33: 1228–1238.
8. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat meth* 9: 357–359.
9. Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.
10. Koboldt DC, Zhang Q, Larson DE, et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome res* 22: 568–576.
11. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucl acids res* 38: e164–e164.
12. Ye K, Schulz MH, Long Q, et al. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25: 2865–2871.
13. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome biol* 11: R106.
14. Mertes F, ElSharawy A, Sauer S, et al. (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief funct genomics* 10: 374–386.
15. Coonrod EM, Durtschi JD, VanSant WC, et al. (2014) Next-generation sequencing of custom amplicons to improve coverage of HaloPlex multigene panels. *Biotechniques* 57: 204–207.
16. Claes KBM, De Leeneer K (2014) Dealing with pseudogenes in molecular diagnostics in the next-generation sequencing era. *Methods mol biol* 1167: 303–315.
17. Gréen A, Gréen H, Rehnberg M, et al. (2015) Assessment of HaloPlex Amplification for Sequence Capture and Massively Parallel Sequencing of Arrhythmogenic Right Ventricular Cardiomyopathy-Associated Genes. *J mol diagn* 17: 31–42.
18. Crobach S, Ruano D, van Eijk R, et al. (2015) Target-enriched next-generation sequencing reveals differences between primary and secondary ovarian tumors in formalin-fixed, paraffin-embedded tissue. *J mol diagn* 17: 193–200.

19. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* 11: R86.
20. Blankenberg D, Kuster GV, Coraor N, et al. (2001) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. In: Current Protocols in Molecular Biology. John Wiley & Sons, Inc.; 2001. Available from: http://onlinelibrary.wiley.com/doi/10.1002/0471142727.mb1910s89/abstract
21. Giardine B, Riemer C, Hardison RC, et al. (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome res* 15: 1451–1455.
22. O'Rawe J, Jiang T, Sun G, et al. (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med* 5: 28.