RESEARCH ARTICLE

Massive Parallel Sequencing for Diagnostic Genetic Testing of BRCA Genes - a Single Center Experience

Natalya A Ermolenko[1&], Uljana A Boyarskikh[1&], Andrey A Kechin[1,2,6&*], Alexandra M Mazitova[3], Evgeny A Khrapov[1], Valentina D Petrova[4], Alexandr F Lazarev[4], Nikolay E Kushlinskii[5], Maxim L Filipenko[1,3]

Abstract

   The aim of this study was to implement massive parallel sequencing (MPS) technology in clinical genetics testing. We developed and tested an amplicon-based method for resequencing the BRCA1 and BRCA2 genes on an Illumina MiSeq to identify disease-causing mutations in patients with hereditary breast or ovarian cancer (HBOC). The coding regions of BRCA1 and BRCA2 were resequenced in 96 HBOC patient DNA samples obtained from different sample types: peripheral blood leukocytes, whole blood drops dried on paper, and buccal wash epithelia. A total of 16 random DNA samples were characterized using standard Sanger sequencing and applied to optimize the variant calling process and evaluate the accuracy of the MPS-method. The best bioinformatics workflow included the filtration of variants using GATK with the following cut-offs: variant frequency >14%, coverage (>25×) and presence in both the forward and reverse reads. The MPS method had 100% sensitivity and 94.4% specificity. Similar accuracy levels were achieved for DNA obtained from the different sample types. The workflow presented herein requires low amounts of DNA samples (170 ng) and is cost-effective due to the elimination of DNA and PCR product normalization steps.

Keywords: Amplicon sequencing - massive parallel sequencing - diagnostics - BRCA - NGS

Introduction

   Mutations in the BRCA1 and BRCA2 genes increase the risk of HBOC. The detection of BRCA1/2 mutations is commonly performed by Sanger sequencing. However, the large sizes of BRCA1/2 (5592 bp and 10,257 bp, respectively) make this procedure too expensive, and the high cost of testing interferes with the widespread use of this test in the clinic. To date, the analysis of the BRCA genes is assigned primarily to patients with severe family history of breast and ovarian cancer (Euhus et al., 2002). However, almost 50% of HBOC patients are carriers of inherited mutations of BRCA1 or BRCA2 and have no family history of the disease (King et al., 2003). These include, for instance, patients who have inherited mutations paternally and have no female relatives in the paternal line. Currently, these patients are rarely directed to the genetic testing of the BRCA genes.

   Until recently, the diagnosis of BRCA1/2 mutations was essential mainly for the estimation of the individual risk of HBOC and to plan preventive diagnostics. In September 2013, the Astra Zeneca Company initiated the third stage of clinical trials of an inhibitor of the Poly (ADP-ribose) polymerase (PARP) (olaparib), which influences the tumor cells with defective BRCA1 or BRCA2. With the advent of a new drug, BRCA1/2 testing became a powerful predictive tool for targeted therapies. The development of a cost-effective method of BRCA1/2 testing opens the possibility of extending this treatment to a wider class of patients.

   Massive parallel sequencing (MPS) is a technology based on the simultaneous sequencing of spatially separated DNA molecules. MPS-sequencing productivity reaches 10 gb and is characterized by its low cost per nucleotide. Thus, the use of MPS for resequencing of BRCA1/2 in pooled samples significantly reduces the cost of the analysis of mutations in the BRCA genes. The sequencing of the BRCA genes using MPS has been already performed using the platforms 454 FLX (Leeneer et al., 2011), GS Junior (Roche) (Feliubadalo et al., 2013), Genome Analyzer (Illumina) (Morgan et al., 2010; Walsh et al., 2010), SOLiD System, Ion PGM/Ion Proton (Invitrogen) (Chan et al., 2012), and HeliScope (Helicos BioSciences) (Thompson et al., 2012).

   Taking into account the recommendations of earlier studies, we have developed a workflow for the testing of BRCA genes using the MPS platform MiSeq (Illumina) adapted to the diagnostic laboratory setting. The workflow

*[1]Institute of Chemical Biology and Fundamental Medicine, [2]Novosibirsk State University, Novosibirsk, [3]Kazan Federal University, Kazan, [4]Altai Branch of Russian Blokhin Cancer Research Centre, Barnaul, [5]Russian Blokhin Cancer Research Centre, Moscow, [6]Laboratory of Pharmacogenomics, Institute of Chemical Biology and Fundamental Medicine, Lavrentiev av, Russia [&]Contributed equally *For correspondence: a.a.kechin@gmail.com*

presented herein requires low amounts of DNA samples and is cost-effective due to the elimination of the laborious DNA and PCR product normalization steps. In the present study, we describe our use of the MPS method in the routine diagnostics of BRCA mutations in 96 HBOC patients under current observation in the Altai Krai Oncological Clinic.

Materials and Methods

Study participants

The participants were 96 patients with primary invasive breast cancer with one of the following hereditary cancer criteria: (a) at least two cases of breast cancer before age 50 in the family; (b) at least three cases of breast cancer in the family; (c) bilateral breast cancer; (d) male cancer; cases of breast or ovarian cancer in the family; or (e) early breast cancer (before age 40). The family histories and ages of the patients are listed in Table 1. The study participants were diagnosed in the Altai Krai Oncological Clinic between 2012 and 2014. During that period, all patients with breast cancer were tested for hot-spot mutations in BRCA1 (5382insC, C61G, 4154delA, 185delAG, 2080delA, 3819del5, 3875del4) and BRCA2 (6174delT, 9318del4, 1528del4). For the study, patients without hot-spot BRCA1/2 mutations were selected. Written informed consent was obtained from all study participants.

DNA extraction

DNA was obtained from the following different sample types: EDTA-treated peripheral blood leukocytes (n=75); whole blood drops dried on paper (hereinafter "*blood print*," n=14); and buccal wash epithelia (n=7). DNA was extracted from the peripheral blood leukocytes using an in-house method comprising a cell lysis step using with 10% SDS, proteinase K treatment, protein extraction using phenol-chloroform, and ethanol precipitation of the DNA. DNA was extracted from the blood prints and buccal wash samples using the QIAamp® DNA Blood Mini Kit (Qiagen). All DNA samples were quantified using PicoGreen (Promega). There were approximately 25 variations in the sample amounts. We did not equalized the DNA concentration for subsequent PCR.

PCR-based target amplification

We developed primer sets covering the entirety of the BRCA1 and BRCA2 coding regions and splice sites (at least 15 nucleotides). The gene-specific primers contained a universal 5'-end tail (forward primer: 5'-acacgacgctcttccgatct-3' and reverse primer: 5'-gacgtgtgctcttccgatct-3'). The primer structures are available upon request. Per sample, 86 singleplex PCRs with gene-specific primers were performed. PCRs were performed in a total volume of 16 ml. The amplification mixture included 10 мM TrisHCl (pH 8.9), 2.5 мM MgCl2, 55 мM KCl, 200 mM of each dNTP, 1.25 нM Syto13, 300 nM of forward and reverse primers, 0.5 U AmpliTaq Gold (Life Technologies), and 2-50 ng of DNA. The temperature cycling protocol consisted of the following steps: 12 min at 95°C, 30 cycles of denaturation

at 95°C for 10 sec, annealing at 60°C for 10 sec, extension at 72°C during 50 sec, and final extension at 72° for 2 min. PCRs were performed on a CFX384 instrument (Bio-Rad, Hercules, CA). All target PCR products were pooled and purified with magnetic beads. Amount of PCR products was evaluated by Endpoint Fluorescence (EF). Variations between EF value were less 5. We considered such variations are not significant for MPS and thus we did not normalize the samples before pooling. Aliquots of the pooled samples were diluted 100 times and 2 mkl product was used as a template for a second PCR. The second round of PCR was performed using a pair of barcoded primers (30 nM) consisting of a 3'-end universal tail, an 8-nucleotide barcode sequence, a 5'-end bridge adaptor, and a pair of common primers (300 nM) corresponding to the 5'-end adaptor of the barcoded primers. All sequences were provided by Illumina (Diego et al., 2014). The temperature cycling protocol consisted of the following steps: 12 min at 95°C, 15 cycles of denaturation at 95°C for 10 sec, annealing at 60°C for 10 sec, extension at 72°C for 50 sec, and a final extension at 72°C for 2 min. After the second PCR, all 96 patient-specific products were pooled and purified using magnetic beads. DNA concentration of pooled sample was evaluated using q-PCR with primers 5'-aatgatacggcgaccaccga-3', 5'-caagcagaagacggcatacga-3' and TaqMan probe 5'-FAM-tccctacacgacgctcttccg-FQ-3' using PhiX serial dilutions as standard samples.

Resequencing and data analysis

Library sequencing was performed using the Illumina MiSeq 2x250 bp platform according to the manufacturer's instructions. The obtained reads were further processed using the following workflow: the reads were separated by their barcodes using our Python-script and were then mapped to BRCA1 (NC_000017.10 region 41184133..41289677) and BRCA2 (NC_000013.10 region 32876987..32986437) with BWA (Li et al., 2009); InDels were realigned using GenomeAnalysisToolKit (GATK) and variations were called with Samtools (Li et al., 2009) GATK (McKenna et al., 2010), and Freebayes (Garrison et al., 2012) and finally, all variations were annotated using Annovar (Wang et al., 2010) and our Python scripts based on the BIC (http://research.nhgri.nih.gov/bic/), dbSNP135 (http://www.ncbi.nlm.nih.gov/SNP/) and 1000 Genomes (http://www.1000genomes.org) databases.

To determine the best pipeline, we varied the following parameters: SNP and InDel calling programs (Samtools,

Results

Using the MPS approach, we sequenced the BRCA1 and BRCA2 genes in 96 patients. Of these, 16 random samples were characterized with conventional Sanger sequencing and were used to assess the sensitivity and specificity of the workflow presented in this study.

Overall Quality, Tag Sorting, Mapping Reads and Coverage of Sequencing

The BRCA1/2 exons and adjacent splice sites (approximately 40 but not less than 15 nucleotides) were

resequenced by the MiSeq platform. The total number of read pairs was 8,282,483 and the base-call quality ranged from 24 to 39, which corresponded to a maximum of 0.4% probability of a wrong base call. Thus, the quality filters were not applied. Only 70.7% (5854421) of the reads contained barcodes without variants in the nucleotide sequence. The rest of the barcodes contained one (13.22%), two (3.04%) or more variants, or corresponded to different reference sequences (13.06%). Only pairs of reads bearing barcodes without variants were selected for further analysis.

The reads were aligned against the BRCA1 (NC_000017.10) and BRCA2 (NC_000013.10) reference sequences. A high mapping rate (99.2±0.1%) was attained for all samples. The mean coverage of each PCR locus and sample was evaluated. Coverage was high enough (>100×) in all cases. The mean amplicon coverage was 1012 (182-2269) reads. The coverage for various samples

ranged from 240 to 4652 reads (on average, 1019 reads). All DNA samples with the lowest coverage were extracted from the blood prints. The distributions of coverage across the amplicons and samples are shown in Figure 1.

Variant calling

We formed a training set from 16 samples to optimize the variant calling process and assess the accuracy of the MPS method using Sanger sequencing as a reference standard. In total, the training set contained 68 polymorphic variants in BRCA1, including 11 different single nucleotide variants (SNV) and one deletion.

We investigated the identification of the BRCA1 variants using the GATK tool. 103 polymorphic variants of the BRCA1 genes (92 SNV and 11 deletions) were identified after alignment and raw variant calling. Of
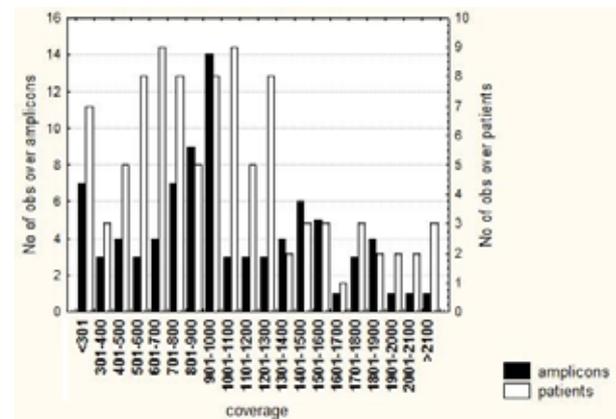
Table 1. Mutational Profile of the 96 Samples

		age	
		< 40	40-50
Family history of disease	2 cases of BC up to 50 years in the family	3	6
	3 cases of BC in the family	-	1
	Bilateral BC	4	7
	Male BC	-	1
	BC and OC in the family	2	4
	-	67	-

Figure 1. Coverage Distributions over Amplicons and Samples

Table 2. Effect of Different Filter Application

Variant calling tool	Coverage cutoff	VF cutoff	TP	FP	FN	Sensetivity/Specifity
FreeBayes	50	10	61	32	7	89.7/85.6
		30	57	3	11	83.8/95.0
	20	10	65	36	3	95.6/64.4
		20	63	22	5	92.6/74.1
		30	61	3	7	89.7/95.3
		40	60	3	8	88.2/95.2
SAMtools	50	10	55	3	13	80.9/94.8
		30	55	3	13	80.9/94.8
	20	10	59	9	9	86.8/95.2
		20	59	3	9	86.8/95.2
		30	59	3	9	86.8/95.2
		40	59	3	9	86.8/95.2
GATK*	50	10	64	6	4	94.1/91.4
		30	56	3	12	82.4/94.9
	20*	10	68	7	0	100.0/90.7
		14*	68	4	0	100.0/94.4
		20	65	3	3	95.6/95.6
		30	60	3	8	88.2/95.2
		40	60	3	8	88.2/95.2

* The optimal tool and filter thresholds

Table 3. Case Details

Gene	Ref. Position	Mutation type	Base change	Variant designation	BIC (Clin. sign.)	1000 Genomes (freq.)	Number of samples
BRCA1	24948	nonsense	G>A	p.Q5385X	-	-	1
BRCA1	38843	missense	C>T	p.M1652I	unknown	0.0093	10
BRCA1	38916	missense	A>G	p.M1628T	unknown	0.0042	1
BRCA1	38962	missense	T>C	p.S1613G	no	0.3327	53
BRCA1	39148R	IVS	T>C	c.4676-25A>G	-	-	1
BRCA1	42356 R	missense	C>A	p.S1512I	no	0.0005	1
BRCA1	50338	synonymous	A>G	p.S1436S	no	0.3085	55
BRCA1	59868	missense	T>C	p.K1183R	no	0.3295	53
BRCA1	60297	missense	C>T	p.S1040N	unknown	0.0121	1
BRCA1	60303	missense	T>C	p.E1038G	no	0.3081	36
BRCA1	60804	missense	G>A	p.P871L	no	0.4909	54
BRCA1	61105	synonymous	A>G	p.L771L	unknown	0.3067	54
BRCA1	61334	synonymous	G>A	p.S694S	unknown	0.3295	49
BRCA1	61339	missense	C>T	p.D693N	no	0.0405	7
BRCA1	61733	synonymous	A>G	p.S561S	-	-	1
BRCA1	62165 R	missense	A>C	p.N417K	unknown	-	1
BRCA1	62349	missense	T>C	p.Q356R	unknown	0.0284	16
BRCA1	67799	IVS	G>A	c.442-34C>T	no	0.098	39
BRCA1	72104 R	frameshift	delGG	p.115_115del	-	-	2
BRCA1	92020 R	IVS	insAT	c.80+118->AT	no	-	2
BRCA1	93145 R	IVS	G>A	c.80+1243C>T	unknown	-	1
BRCA2	16221	IVS	T>A	c.68-7T>A	unknown	0.0023	20
BRCA2	16358 R	synonymous	A>G	p.Q66Q	no	0.0005	1
BRCA2	16383 R	missense	G>C	p.A75P	unknown	-	1
BRCA2	29494	missense	A>C	p.N289H	no	0.0591	9
BRCA2	29607 R	missense	C>A	p.S326R	no	0.0005	1
BRCA2	29743	missense	A>C	p.N372H	-	0.2443	52
BRCA2	29796	synonymous	G>A	p.P389P	-	-	1
BRCA2	29994	synonymous	A>G	p.S455S	no	0.0591	9
BRCA2	30143 R	missense	T>C	p.I505T	no	0.0009	1
BRCA2	33735	synonymous	T>C	p.H743H	no	0.0586	9
BRCA2	34477	missense	A>G	p.N991D	unknown	0.0633	9
BRCA2	34902	synonymous	A>G	p.K1132K	no	0.2834	48
BRCA2	35021	missense	C>T	p.S1172L	unknown	0.0023	1
BRCA2	35022	synonymous	G>A	p.S1172S	no	0.0028	1
BRCA2	35313	synonymous	T>C	p.V1269V	no	0.1819	31
BRCA2	36705	synonymous	C>T	p.S1733S	no	0.0023	8
BRCA2	37165 R	missense	A>G	p.T1887A	-	-	1
BRCA2	37250	missense	C>T	p.T1915M	unknown	0.0116	8
BRCA2	52246	synonymous	A>G	p.S2414S	no	0.2494	35
BRCA2	52256 R	frameshift	delAG	p.2418_2418del	-	-	1
BRCA2	53687 R	missense	C>T	p.T2515I	no	-	2
BRCA2	59660	IVS	T>C	c.7806-14T>C	unknown	0.453	73
BRCA2	91824	IVS	T>C	c.9257-16T>C	unknown	0.0046	2
BRCA2	94217 R	IVS	G>A	c.9648+22G>A	-	-	1
BRCA2	95640	nonsense	A>T	p.K3326X	no	0.0056	2

these variants, 35 were false positives (FP). To reduce the number of FP variants, we tested some additional filters: (i) the variants coverage below 20× or 50×; and (ii) the frequency of reads with variant alleles (VF) below 10-40%. We also compared the three variant calling tools (FreeBayes, SAMtools and GATK). The results of applying these software and the different variant calling filters are shown in Table 2. The optimal tools and filter thresholds were selected to achieve the minimum number of FP variants and the minimum number of false negative variants (FN, i.e., lost true positive variants). In this way, the maximum sensitivity [TP/(TP + FN)] was reached (TP denotes a true positive). The best results were achieved using the GATK tool and application of coverage <20× and VF <14%. Thus, the sensitivity was 100%. The specificity was calculated as a fraction of TP among all positives [TP/(TP+FP)]. This so-called positive predictive value (PPV) gives a more informative value than standard specificity [TN/(TN + FP)] (Zvelebil et al., 2007). Hence, the variant calling specificity was 94.4%.

Next, the variants of all 96 samples were identified. As determined by the analysis of the training set, only the variants with a combined coverage of at least 20 reads and at least 14% VF were considered. In total, we identified 977 variants in BRCA1 and BRCA2 (46 unique SNV and nine indel variants). All identified mutations were tested using Sanger sequencing. As a result, three SNV and six indel variants were not confirmed. We analyzed all FP variants, finding that all three false positive SNVs were directly adjoined to the 3'-end of one of the primers of the first PCR and were detected in only one strand. The use of an additional requirement (the presence of variants in both forward and reverse reads) allowed us to exclude all three false positive SNVs.

All six FP indel variants were located in or close to homopolymeric regions, particularly in the poly(A) tracts. The most likely cause of these FP variants was random errors introduced by the non-proofreading AmpliTaq polymerase (Life Technologies). The library preparation stage of our workflow contained two rounds of PCR; therefore, the probability of random errors was rather high. For example, a 2080delA deletion in the 61,455 position of BRCA1 (exon 11) was determined in 15% of the reads. Sanger sequencing of a corresponding DNA fragment showed the presence of the 2080delA allele if the amplicons were generated using AmpliTaq Gold polymerase, while the reference allele was only present when Pfu-polymerase was used. Thus, it is essential to use the proofreading enzyme in preparing the MPS library. Interestingly, some of FP deletions are present in the BIC database, particularly the 2080delA in BRCA1 and 1806delA deletion in BRCA2. It is possible that some of the BIC database indels were caused by nonrandom sequencing errors. All SNVs and indel variants confirmed by Sanger sequencing are shown in Table 3.

Variant annotation

All variants were categorized by gene location (intronic or exonic) and by their predicted effects (frameshift, synonymous substitution, missense substitution, splice site alteration or nonsense). All variants were annotated for their frequency on the basis of their presence in dbSNP135 or their frequency in the 1000 Genomes Project. We defined a variant as "*rare*" if it was not found in dbSNP135 or if its allele frequency was <0.001 in the 1000 Genomes Project (Abecasis et al., 2010). To evaluate the clinical significance of these variants, we compared all variants to the gene-specific mutation database (BIC, http://research. nhgri.nih.gov/bic/). In addition, we predicted the effects of rare missense SNVs on protein function using the SIFT (Kumar et al., 2009) and PROVEAN algorithms. We defined a variant as "*functional*" if both SIFT and PROVEAN predicted it to be damaging.

As a result, we identified 46 unique and true variants using MPS. Of these, 17 variants were defined as rare. Three of the rare variants (p.2418_2418del in BRCA2; p.115_115del and p.Q5385X in BRCA1) were disease-causing, and the remaining variants consisted of eight splice site alterations (1 c.80+118->AT insertion and seven single-base substitutions), seven missense substitutions and three synonymous substitutions. The role of the 11 remaining rare variants in the development of HBOC is unclear. We defined these variants as likely pathogenic. However, none of them were predicted to be damaging by both SIFT and PROVEAN.

Discussion

The increasing requirement of genetic testing and reduction of analysis turnaround times has led to the need to improve the sequencing technology used in our laboratory. We considered the possibility of using an MPS-based method of target sequencing. Despite a number of studies reporting the successful application of this method in determining the disease-causing mutations, existing MPS platforms are rarely used in clinical practice. In our opinion, one of the key limitations is that the method is not designed for sequencing heterogeneous clinical samples, which often have small concentrations of DNA or consist of poorly preserved samples of fragmented DNA. In the present study, we used the Illumina MiSeq benchtop platform to identify BRCA1/2 mutations in patients with HBOC. We obtained patient DNA samples from various sources and did not use any additional sample preparation methods to standardize them. Using Sanger sequencing as a reference method, we estimated the sensitivity and specificity of the MPS test and the possibility of its application in a routine clinical setting.

The complete BRCA1/2 resequencing workflow consisted of enrichment of targeted regions of interest, targets sequencing, and finally, bioinformatic analysis. Target enrichment significantly reduces the sequencing cost per base and improves the accuracy of variant detection. Currently, several of possible approaches for enrichment have been proposed. The first is the amplicon enrichment approach, which is based on using singleplex, multiplex or long-range PCR amplicons as the MPS template (Morgan et al., 2010; Leneer et al., 2011; Hernan et al., 2012). Singleplex PCR is one of the simplest and unpretentious methods, as it does not require a large amount of starting DNA. However, it has a limited throughput. The main advantage of multiplex

PCR is its higher productivity, though multiplex PCR requires a larger amount of DNA, and the amplification results can vary significantly depending on the reagents and laboratory equipment used. Both of these methods of library enrichment can give rise to nonrandom sequencing errors due to errors introduced by Taq-polymerase during the amplicon synthesis. Long-range PCR uses a mixture of proofreading and non-proofreading DNA polymerases; thus, products of up to 40 kb long can be synthesized. This method is highly productive because the amplification of the target locus is performed using few reactions, and the use of a proofreading DNA polymerase provides a low error rate. The disadvantage of this method is that most of the amplified DNA products fall in the noncoding sequences, significantly increasing the analysis cost.

The second approach is the selection of DNA fragments by hybridization with oligonucleotide probes (Albert et al., 2007; Okou et al., 2007; Gnirke et al., 2009). The disadvantages of this approach include the nonspecific hybridization of homologous sequences (pseudogenes in particular) and the highly variable efficiency with which the probes bind the target fragments (Gnirke et al., 2009). A recent whole exome study has registered wide variations in read coverage ranging from zero to several hundred-fold (Ng et al., 2009).

The third "selector"-based approach involves a process of selector probe capture and ligase-assisted DNA circularization to bring about target enrichment (Johansson et al., 2011). Until recently, this method was not properly optimized. Selector- and hybridization-based approaches generally require 0.8-2 mkg of DNA per sample as the starting material, which may limit their applicability as diagnostic tools.

For the present study, we used an in-house singleplex PCR method of enrichment, which proved to be the most convenient for us because our laboratory has had many years of experience in analyzing BRCA using classical Sanger sequencing. Thus, we were able to use oligonucleotide primers and amplification modes initially optimized for Sanger sequencing. In addition, this method does not require a large amount of DNA; this factor is essential because, in our practice, we must often deal with a small amount of DNA obtained from a blood print on paper or from buccal epithelia.

Target enrichment is a crucial step in sequencing, and an appropriate enrichment method provides complete and uniform coverage of the targeted regions of interest. To date, there is no consensus as to which method of enrichment is preferable. Using our PCR-based approach, we observed reasonably uniform coverage (variation <4) for 74 out of 86 amplified loci. Two remaining loci were over-represented (coverage>2000) and 10 were under-represented (coverage <500), with four- to 12-fold variation in the coverage of these loci. In subsequent studies, we have achieved a more uniform coverage (variation <5) by changing the ratio of amplicons in a pool of first universal-tailed PCR products (unpublished data).

The MPS-based testing cost depends not only on the sequencing cost but also on the cost of individual sample preparation prior to sequencing. The sample preparation process was significantly simplified by eliminating the DNA quantification and normalization steps. In this paper, we have used samples with different amounts of DNA (2-50 ng) as a matrix for the first PCR, which all provided comparable amounts of PCR products (variation <5). For all of the samples, a coverage sufficient for reliable detection of mutations was obtained. However, some of the samples isolated from blood prints had highly fragmented DNA and a significantly lower average coverage (less than 300 readings).

A key stage in the implementation of all MPS platforms is a bioinformatics analysis. Despite a number of BRCA resequencing study, the generation of a reliable variant list remains the bottleneck of this type of project. To separate the TP and FP variants, the application of a filter set is needed. We used a relatively simple set of filters (coverage>20 reads, VF> 14% and presence in both forward and reverse strands). More severe VF/coverage cutoffs reduced the number of FP variants (which enhanced specificity) but increased the number of FN (which reduced sensitivity). Other studies have recommended a VF cutoff ranging from 10% (Morgan et al., 2010) to 25% (Feliubadalo et al., 2013) and a minimum coverage ranging from 10x (Morgan et al., 2010) to 50x (Walsh et al., 2010). Some studies have used additional filters, such as a quality score>30 (Leeneer et al., 2011; Feliubadalo et al., 2013). We preferred weaker filters, as the sensitivity is more important than the specificity for a diagnostic test. A large number of FP variants only increases the load on the conformational Sanger sequencing, but does not affect the results of the test. The present results indicate that the workflow we have developed allows for the pinpointing of single nucleotide variants and indels with a specificity of 94.4% and a sensitivity of 100%.

In addition to small indels and single-base substitutions, large genomic rearrangements (LGRs) have been identified in HBOC families and account for a significant proportion of HBOC cases (from approximately 1% (Thomassen et al., 2006) up to 12-25% (Montagna et al., 2003; Walsh et al., 2006) in different populations). Theoretically, the MPS method may be used to identify LGRs using the relative ratios of reads. Two recent studies have reported the accurate identification of large genomic duplications and deletions (Walsh et al., 2010; Thompson et al., 2012). However, these studies used a hybridization-based method for the target enrichment of the DNA library. We used two rounds of PCR to enrich the DNA library with BRCA1/2 sequences. Previous studies (Leeneer et al., 2011; Feliubadalo et al., 2013) have shown this method to be unreliable for LGR detection.

Relying on our previous experience, we estimated the consumables and time spent in using our MPS approach in BRCA testing. After accounting for the additional analyses (Sanger sequencing to confirm the mutations), we found that the overall cost of testing one sample was approximately two times lower using the MPS method compared with Sanger sequencing. The hands-on time and turnaround time were also noticeably reduced.

The cost of testing for one patient is calculated based on the fact that in a single run of MPS, 96 patients can be analyzed. In 2013, there were 957 women and four men identified with newly diagnosed breast cancer

and 218 women with ovarian cancer in the Altai Krai Oncological Clinic. Thus, the analysis turnaround times is approximately one month (provided that testing of inherited BRCA mutations will be administered to all patients regardless of age and family history of the disease). These large time requirements in the analysis initiation phase noticeably reduce the significance of this method for use in clinical diagnostics. However, the addition of paraffin blocks DNA samples to current BRCA NGS analysis workflow for somatic mutation testing to determine the sensitivity of tumors to a new targeted drug (PARP inhibitor) could considerably improve its clinical usability. The time spent in analysis can also be reduced if, in a single run of MPS, patients with various genetic disorders are analyzed. To do this, it is necessary to develop a uniform MPS workflow for different types of genetic testing.

<u>Conclusion</u>: The present study confirms the possibility of using MPS for genetic testing in routine clinical practices while pointing out some limitations of this method. The method presented here demonstrated excellent sensitivity and specificity (100% and 94.4%, respectively) and could be applied to different types of clinical specimens.

References

Abecasis GR, Altshuler D, Auton A, et al (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-73.

Albert TJ, Molla MN, Muzny DM, et al (2007). Direct selection of human genomic loci by microarray hybridization. *Nat Methods*, **4**, 903-5.

Chan M, Ji SM, Yeo ZX, et al (2012). Development of a next-generation sequencing method for BRCA mutation screening: a comparison between a high-throughput and a benchtop platform. *J Mol Diagn*, **14**, 602-12.

De Leeneer K, Hellemans J, De Schrijver J, et al (2011). Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. *Hum Mutat*, **32**, 335-44.

Diego S (2014). Entire document and oligonucleotide sequences© 2007-2013 Illumina, Inc. *All rights reserved*. 1-28

Euhus DM, Smith KC, Robinson L, et al (2002). Pretest prediction of BRCA1 or BRCA2 mutation by risk counselors and the computer model BRCAPRO. *J Natl Cancer Inst*, **94**, 844-51.

Feliubadalo L, Lopez-Doriga A, Castellsague E, et al (2013). Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. *Eur J Hum Genet*, **21**, 864-70.

Garrison E, Marth G (2012). Haplotype-based variant detection from short-read sequencing. 1-9

Gnirke A, Melnikov A, Maguire J, et al (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, **27**, 182-9.

Hernan I, Borràs E, de Sousa Dias M, et al (2012). Detection of genomic variations in BRCA1 and BRCA2 genes by long-range PCR and next-generation sequencing. *J Mol Diagn*, **14**, 286-93.

Johansson H, Isaksson M, Sorqvist EF, et al (2011). Targeted resequencing of candidate genes using selector probes. *Nucleic Acids Res*, **39**, 8.

King M-C, Marks JH, Mandell JB (2003). Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, **302**, 643-6.

Kumar P, Henikoff S, Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, **4**, 1073-81.

Li H, Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, et al (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-9.

McKenna A, Hanna M, Banks E, et al (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**, 1297-303.

Montagna M, Dalla Palma M, Menin C, et al (2003). Genomic rearrangements account for more than one-third of the BRCA1 mutations in northern Italian breast/ovarian cancer families. *Hum Mol Genet*, **12**, 1055-61.

Morgan JE, Carr IM, Sheridan E, et al (2010). Genetic diagnosis of familial breast cancer using clonal sequencing. *Hum Mutat*, **31**, 484-91.

Ng SB, Turner EH, Robertson PD, et al (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272-6.

Okou DT, Steinberg KM, Middle C, et al (2007). Microarray-based genomic selection for high-throughput resequencing. *Nat Methods*, **4**, 907-9.

Thomassen M, Gerdes AM, Cruger D, Jensen PKA, Kruse TA (2006). Low frequency of large genomic rearrangements of BRCA1 and BRCA2 in western Denmark. *Cancer Genet Cytogenet*, **168**, 168-71.

Thompson JF, Reifenberger JG, Giladi E, et al (2012). Single-step capture and sequencing of natural DNA for detection of BRCA1 mutations. *Genome Res*, **22**, 340-5.

Walsh T, Casadei S, Coats KH, et al (2006). Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA*, **295**, 1379-88.

Walsh T, Lee MK, Casadei S, et al (2010). Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc Natl Acad Sci USA*, **107**, 12629-33.

Wang K, Li M, Hakonarson H (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, **38**, 164.

Zvelebil M, Baum J (2007). Understanding Bioinformatic. Garland Science, New York, **356**.