Published online 8 November 2010

Nucleic Acids Research, 2011, Vol. 39, No. 2 ***e8***
doi:10.1093/nar/gkq1005

Targeted resequencing of candidate genes using selector probes

H. Johansson[1,2], M. Isaksson[1,2], E. Falk Sörqvist[1,2], F. Roos[1,2], J. Stenberg[1,2], T. Sjöblom[1], J. Botling[1], P. Micke[1], K. Edlund[1], S. Fredriksson[3], H. Göransson Kultima[4], Olle Ericsson[1,2,*] and Mats Nilsson[1,*]

[1]Department of Genetics and Pathology, Uppsala University, Rudbeck Laboratory, SE 75185 Uppsala, [2]Olink Genomics, Dag Hammarskjölds väg 36B, SE 75183 Uppsala, [3]Olink Biosciences, Dag Hammarskjölds väg 54A, SE 75183 Uppsala and [4]Department of Medical Sciences, Uppsala University, Akademiska Hospital, SE 75185 Uppsala, Sweden

Received July 21, 2010; Revised September 22, 2010; Accepted October 7, 2010

ABSTRACT

Targeted genome enrichment is a powerful tool for making use of the massive throughput of novel DNA-sequencing instruments. We herein present a simple and scalable protocol for multiplex amplification of target regions based on the Selector technique. The updated version exhibits improved coverage and compatibility with next-generation-sequencing (NGS) library-construction procedures for shotgun sequencing with NGS platforms. To demonstrate the performance of the technique, all 501 exons from 28 genes frequently involved in cancer were enriched for and sequenced in specimens derived from cell lines and tumor biopsies. DNA from both fresh frozen and formalin-fixed paraffin-embedded biopsies were analyzed and 94% specificity and 98% coverage of the targeted region was achieved. Reproducibility between replicates was high ($R^2 = 0, 98$) and readily enabled detection of copy-number variations. The procedure can be carried out in <24 h and does not require any dedicated instrumentation.

INTRODUCTION

DNA resequencing of whole mammalian genomes can be performed with a range of novel sequencing approaches (1–4). Sufficient coverage to detect the vast majority of genetic variation in a complete genome can be achieved with a single instrument run which makes it a very powerful tool for screening. As discussed below, there are today also cost-efficient approaches for resequencing complete exomes, but as sequencing rapidly becomes cheaper, exome sequencing can be expected to become less rational. However, for many hypothesis-driven or clinical investigations, it is more efficient and rational to direct sequencing to a small fraction of the genome, e.g. to a set of candidate genes. By limiting the field of search, more sequencing capacity can be spent to achieve high coverage of the region of interest (ROI). This is particularly useful for investigations where deep sequencing is required such as sequencing of mosaic tumor biopsies and for clinical applications. Furthermore, it can also facilitate analysis of large patient cohorts which is still very expensive if performed on complete genomes.

Several methods for enrichment of multiple target loci have been presented. The methods are in the most general sense based on either hybridization capture or parallelized PCR amplification of multiple regions (5–17). For more in-depth review of targeted resequencing methodologies, we refer to the reviews by Mamanova *et al.* (18) and Turner *et al.* (19). Assays based on hybridization have proven to be effective in capturing the complete exome, or a large fraction of it. However, relying solely on hybridization for discrimination between target and non-target is difficult with regards to specificity and typically result in capture of 50–70% irrelevant sequences (6). The problem is accentuated when capturing smaller regions that require higher enrichment levels. If the ratio between on- and off-target DNA is low, the risk of misaligning off-target sequences to the target sequence increases, which may lead to false positive, or even false negative results. This lack of specificity in the capture step can to some extent be counteracted in the sequencing step by increasing the read length and/or performing paired end sequencing, however this comes at significant increases in both cost and assay time. To allow for a more

*To whom correspondence should be addressed. Tel: +46 18 471 48 16; Fax: +46 18 471 48 08; Email: mats.nilsson@genpat.uu.se
Correspondence may also be addressed to Olle Ericsson. Tel: +46 18 495 31 22; Fax: +46 18 495 31 21; Email: olle.ericsson@olinkgenomics.com

The authors wish it to be known that, in their opinion, the first two authors and last two authors should be regarded as joint First Authors.

general targeting of fragments, hybridization can be combined with an enzymatic discrimination and amplification step, a strategy which is successfully demonstrated in PCR (20). Since it is difficult to perform highly multiplex PCR reactions with high-success rate, massive numbers of single-locus PCRs have to be applied, requiring compartmentalization in small volumes to be cost-effective. This can be achieved by array based approaches (9) or by emulsion PCR (10), both requiring sophisticated liquid handling and amplification systems.

Two approaches have been presented for efficient single-tube massively multiplexed amplification of genomic loci. Both methods are based on ligase-assisted DNA-circularization reactions that, similar to PCR, rely on enzymatic specificity and dual hybridization recognition. Gap-fill padlock probes (11) are oligonucleotides that hybridize with sequences flanking the ROI (e.g. an exon). The gap between the ends of the probe is filled by a DNA polymerase and sealed by a DNA ligase, creating a circle containing the targeted sequence. With this approach, tens of thousands of probes have been amplified with very low enrichment of off-target material. However, the performance for exon capture is less than ideal: at 322× mean coverage only 75% of the bases were covered at >20× (12). The related selector probes template circularization of restriction fragments and do not include synthetic DNA sequences from the probe in the targeted sequence, thereby reducing the fraction of irrelevant sequence in the captured DNA. A PCR-based version of the selector technique was used to efficiently sequence the exons of 11 cancer genes (13) with excellent specificity but with relatively high-amplification bias and low coverage (70% coverage at >10× coverage at 127 average coverage).

In order to retrieve useful sequence information from the ROI, it is very important that the enrichment is unbiased, since insufficient or irreproducible coverage within the target region will drastically lower the fraction of useful sequence data obtained from an experiment.

In this article, we present a non-PCR-based version of the selector method (21), including a significantly updated and optimized protocol as well as selector design approach. Instead of using PCR an RCA-based multiple displacement amplification (MDA) (22) is used, generating an amplification product which is easily integrated with shotgun library construction for short-read sequencing platforms.

We evaluate the method for enrichment of a relatively small set of exons (501) with high coverage and relatively low-enrichment bias. Amplification of thousands of targeted fragments (1883) in one tube with high coverage is demonstrated in a simple process that does not require genome centers to invest in additional instrumentation.

MATERIALS AND METHODS

Preparation of tumor samples

Matched fresh-frozen tumor/benign lung tissues and FFPE tumor tissue were used in accordance with the Swedish Biobank Legislation and Ethical Review Act (reference 2006/325, Ethical review board in Uppsala). The tissue samples were reviewed by a pathologist and only tumor tissues with a tumor cell fraction >50% were included. DNA was extracted from FFPE and frozen tissue sections using the QIAamp DNA Mini Kit (Qiagen, Hamburg, Germany).

Design and oligonucleotides

Twenty-eight genes (Table 1) known to be mutated in lung and/or colon cancer were chosen for targeted resequencing. A list of coding regions for each gene was downloaded from the consensus coding sequence database, CCDS (build 36.3) and a total number of 501 regions covering 82 kb were targeted. Coding sequences were collected from hg18 (March, 2006 assembly), and processed by the Disperse software (23) to generate a set of restriction fragments using eight combinations of restriction enzymes. A subset of 1883 fragments (Supplementary Table S1) was selected based on length (100–1000 nt), GC-content (20–65%), avoiding repetitive genomic elements in the ends, and to achieve redundant coverage over targeted regions. In the current design, we aimed for a double-fragment redundancy on each targeted

Table 1. Genes included in the design with their corresponding number of exons and the total number of base pairs that was aimed to be covered for each gene (ROI)

Number	Name	Number of exons[a]	Number of ROI bp[a]	Percentage of ROI bp covered[a]
1	AKT3	14	1685	99
2	IDH1	8	1254	100
3	HER4	28	4124	98
4	CTNNB1	14	2472	100
5	PIK3CA	20	3260	100
6	FBXW7	13	2549	96
7	APC	15	8590	99
8	EGFR	30	4198	100
9	MET	20	4192	100
10	BRAF	18	2432	100
11	CDKN2A	4	1074	100
12	PTEN	9	1359	98
13	CCND1	5	888	100
14	MRE11A	19	2431	99
15	ATM	62	9359	98
16	KRAS	5	687	100
17	HER3	28	4357	98
18	AKT1	13	1594	94
19	SMAD3	9	1352	100
20	TP53	10	1330	100
21	NF1	58	8865	98
22	HER2	31	4357	95
23	SMAD2	10	1485	97
24	SMAD4	11	1778	100
25	STK11[b]	9	1370	98
26	CCNE1	11	1357	100
27	AKT2	13	1597	100
28	GNAS	14	2196	98

The last column shows the percentage of the ROI base pairs that were covered in the design.
[a]CCDS 14 April 2009.
[b]No CCDS available, CDS used, 14 April 2009.

base to avoid the risk of fragment dropout due to mutations in the restriction enzyme or probe-binding site. The design achieved ~99% coverage of targeted bases and the missing bases were found to be in or near repetitive elements. Selected restriction fragments are displayed in the supplementary gff-file. Selector probes serving as template for circularization of each chosen fragment were designed using the ProbeMaker software (24). Each selector consists of two Tm balanced sequences of 20–25 nt complementary to the ends of its targeted restriction fragment.

The oligonucleotides (Integrated DNA Technologies) hybridizing to 25 bps on each end of the targeted fragments where pooled in one tube. To add a 3′-biotin to the probes the oligonucleotides where incubated in 2.5 μM total concentration with 1× Tdt buffer (NEB), 1× CoCl2 (NEB), 0.1 mM dUTP-biotin (Roche Diagnostics) and 0.2 U/μl Terminal Transferase (NEB) in a final volume of 50 μl. The reaction was incubated at 37°C for 1 h and inactivated at 75°C for 20 min. To remove unincorporated nucleotides the probes were purified on three consecutive G-50 columns (GE Life Sciences).

Restriction digestion

Of each sample, 100 ng (200 ng for FFPE samples to compensate for DNA fragmentation) was added to eight different restriction reactions containing 1 unit each of two restriction enzymes and their corresponding compatible NEB buffer in 1× concentration and 0.85 μg/μl BSA in a total volume of 10 μl. The eight reactions were SfcI and Hpy188I in NEB buffer 4; DdeI and AluI in NEB buffer 2; MseI and Bsu36I in NEB buffer 3; MslI and BfaI in NEB buffer 4; HpyCH4III and Bsp1286 in NEB buffer 4; SfcI and NlaIII in NEB buffer 4; MseI and HpyCH4III in NEB buffer 4; HpyCH4V and EcoO109I in NEB buffer 4 (New England Biolabs). The reactions were incubated at 37°C for 60 min followed by enzyme deactivation at 80°C for 20 min. After this step the eight reactions were pooled.

Probe hybridization

The 80 μl of digested sample resulting from pooling the reactions were mixed with 10 pM individual concentration of biotinylated selector probes (Integrated DNA Technologies), 0.7 M NaCl, 7 mM Tris–HCl (pH 7.5), 3.5 mM EDTA, and 0.07% Tween-20 in a total volume of 160 μl. The solution was incubated at 95°C for 10 min, 75°C for 30 min, 68°C for 30 min, 62°C for 30 min, 55°C for 30 min and 46°C for 10 h.

Solid phase capture and wash

To remove off-target material, the hybridization solution was mixed with 10 μl M-280 streptavidin coated magnetic beads (3.35×10^7 beads/ml; Invitrogen) in 0.7 M NaCl, 7 mM Tris–HCl (pH 7.5), 3.5 mM EDTA and 0.07% Tween-20 in a final volume of 200 μl, and incubated at room temperature for 15 min. After incubation, the beads were collected using a ring magnet. Following this step, to further remove non-specifically bound DNA, the beads were washed in 1 M NaCl, 10 mM Tris–HCl (pH 7.5), 5 mM EDTA and 0.1% Tween-20 in a total volume of 200 μl at 46°C for 30 min with rotation.

Circularization of targeted fragments

To circularize the genomic fragments, the beads were incubated in 1× Ampligase reaction buffer, 0.25 U/μl Ampligase (Epicentre) and 0.1 μg/μl BSA in a total volume of 50 μl. The ligation mix was incubated at 55°C for 10 min.

Amplification

To enrich for circular molecules, a MDA reaction (Templiphi, GE life sciences) was carried out using the circles as templates. The circularized molecules were separated from the beads into the solution by incubation with 5 μl sample buffer at 95°C for 10 min and thereafter collected by placing the tubes in a ring magnet rack and immediately aspirating the supernatant. To initiate MDA, 5 μl reaction buffer and 0.2 μl enzyme mix was added to the supernatant and incubated at 30°C for 4 h followed by deactivation at 65°C for 10 min.

Sequencing

Of each MDA-amplification product, 300 ng was fragmented using the Covaris S2 system and libraries where constructed from these using the SOLiD (v3) library construction kit (Applied Biosystems). The constructed libraries where sequenced in barcoded sequencing reactions containing 20 different samples on a fourth of a sequencing slide.

Data analysis

All analyses were made on the targeted 501 exonic regions defined by CCDS. SOLiD (v3) sequence data from four specimens, a Yoruba trio (NA1806-8, family Y009), tumor material from a lung-cancer patient and two pairs of matched cell lines derived from normal and breast cancer tissue, were all analyzed by MosaikAligner version 1.0.1388 (Mosaik—The MarthLab Available at: http://bioinformatics.bc.edu/marthlab/Mosaik). MosaikAligner parameters were set using jump database for mapping to the reference genome (hg18, mars 2006 assembly); -hs 15, -mm 6, -ms 6, -mhp 100 and -act 20. The alignment software was set to allow six mismatches per 50-mer read.

Targeted amplification specificity for all samples was calculated as the proportion of reads that uniquely aligned to the genome mapped to the targeted region (Table 2).

SNP array analysis

SNP-array experiments were performed according to the standard protocols for Affymetrix GeneChip® Mapping NspI-250K arrays (Gene Chip Mapping 500K Assay

Table 2. Sequencing results for each sample showing the total number of reads obtained for each sample and the percentage of them that align to the human genome build hg18, and the percentage of the hg18 uniquely aligned reads that aligns to the specified region

Sample	Total number of reads	hg18		Amplified region	Region of interest	Region of interest ± 50 bp
		Percentage of all reads[a]	Percentage of unique reads[b]	Percentage of unique reads[c]	Percentage of unique reads[c]	Percentage of unique reads[c]
NA18506	1653229	51.34	40.02	92.23	66.21	81.41
NA18507	1709623	50.43	40.16	93.31	67.72	82.99
NA18508	1187613	48.25	38.07	92.82	67.62	82.97
HCC1143 Normal	1687199	51.26	41.23	93.45	67.03	82.38
HCC1143 Tumor	1452999	52.86	43.21	93.85	65.80	81.69
HCC1599 Normal	2617647	53.72	42.48	92.20	65.20	80.45
HCC1599 Tumor	1882203	49.44	41.50	95.61	68.70	85.75
Normal tissue fresh frozen	1567773	53.47	44.27	94.63	67.98	83.42
Tumor tissue fresh frozen	1841970	51.17	42.06	94.32	67.30	83.33
Tumor tissue FFPE	1878918	45.79	37.23	93.95	74.11	88.66
NA18506	2008052	49.63	39.41	92.93	66.98	82.07
All samples	19487226	50.76	40.93	93.55	67.62	83.12

[a]Percentage of the total number of reads that aligned to the human genome and selector-induced sequences, counting all reads that align (those that align more than one time is randomly placed and counted only once).
[b]Percentage of the total number of reads that aligned uniquely to hg18.
[c]Percentage of the uniquely aligned reads to hg18 that aligned uniquely to the region specified.

Manual (P/N 701930 Rev2.), Affymetrix Inc., Santa Clara, CA, USA) and the arrays were scanned using the GeneChip® Scanner 3000 7G. Normal samples analyzed at Uppsala Array Platform were used as a reference set to produce log ratios. The rank segmentation algorithm, similar to the Circular Binary Segmentation (CBS) algorithm (25), in the software Nexus from Biodiscovery was used to segment the data across the genome. The significance threshold for segmentation was set at 1×10^{-6}, also requiring a minimum of 40 probes per segment. The results are shown as segmented log2 ratio in Figure 4a.

Copy-number analysis

Copy-number alterations were detected by analyzing the difference in sequencing depth in each of the targeted positions with 20× sequencing depth or more in the normal DNA sample for respective tumor–normal pair. Reference allele ratios were calculated by dividing reference allele call frequency by the total sequencing depth for each position in the targeted region.

To enable comparison of the sequencing and the GeneChip data (Figure 4b), the sequencing depth log2 ratios were derived from the difference in mean hits per gene for the normal–tumor pair. They were based on positions within the amplified ROI with a sequencing depth of at least 20× in the normal DNA sample.

Length dependency analysis

To extract junction-specific reads, a library where two copies of each selector were concatenated to each other was created. The reads were aligned to the library and the reads aligning uniquely over the junction were used as junction specific reads.

Analysis of input DNA amount requirement

A titration of different amount of DNA from 100 to 1600 ng DNA (combined for all eight restriction digestions) was subjected to our enrichment protocol. The amplified DNA was sent to GATC Biotech (Constance, Germany) for Illumina GAII sequencing. The 32-mer reads were mapped to our amplified region as described above, but only allowing for four mismatches (-mm 4). The specificity was calculated as the fraction of unique reads that mapped to our amplified region.

RESULTS

We designed a somatic mutation detection assay for 28 genes frequently mutated in solid tumors such as colon, lung and breast cancer (Table 1), targeting all 501 coding regions according to the CCDS database, defining our ROI. The new version of the selector technique is described in detail in Figure 1a and in brief involves solid-phase enrichment and probe-templated circularization of specific-restriction fragments that are amplified using a circle-specific RCA-based MDA. The overall cost and quality of a targeted-sequencing experiment are determined by the performance of the enrichment technique. Important metrics to consider are *in silico* design coverage, actual coverage, specificity, reproducibility and enrichment uniformity across the targeted loci. To evaluate these parameters, we sequenced a HapMap trio (NA18506, NA18507 and NA18508) using a SOLiD (v3) instrument. The sequencing libraries were bar-coded to allow sequencing of 20 samples on a fourth of a slide. On average, 899 000 reads mapped to the genome, and of the uniquely mapping reads, 94% (range 92–96%) mapped to the amplified region, demonstrating the high specificity of this approach corresponding to an

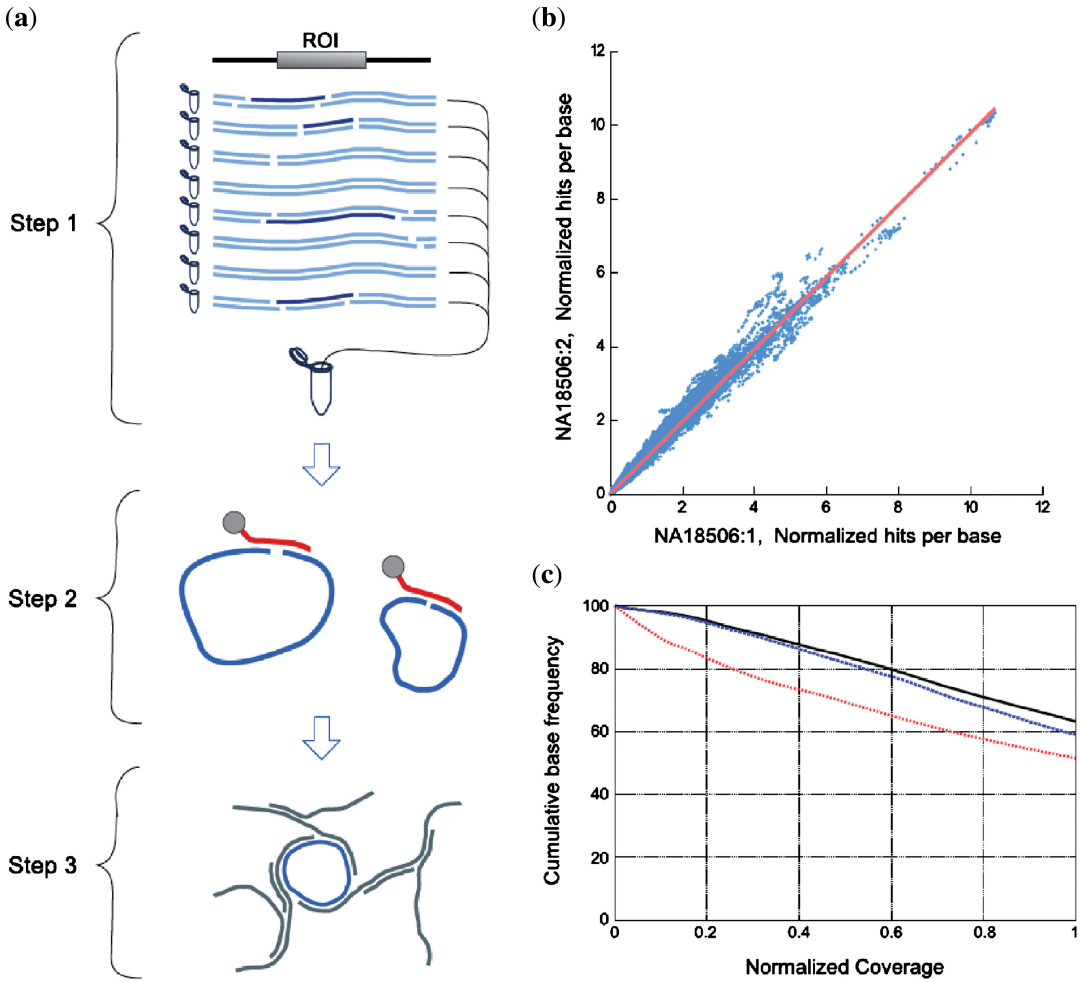

Figure 1. Selector probe technology and performance. (**a**) Overview of the selector-probe amplification procedure. A ROI (e.g. an exon) is targeted by probing several redundant fragments (dark blue) selected from eight separate restriction digestion reactions of a genomic DNA sample (light blue). The restriction-digested DNA samples are pooled and hybridized with biotin-tagged selector probes complementary to the ends of each targeted restriction fragment (red). The targeted fragments are then captured on streptavidin-coated magnetic beads. The fragments are circularized by DNA ligation after removal of non-targeted DNA. Finally, the circularized fragments are released from the beads and specifically amplified using the RCA-based MDA. (**b**) Correlation of the relative coverage of individual bases between two replicate enrichment and sequencing experiments. (**c**) The cumulative fraction of the target sequence covered is plotted as a function of different relative read-depths, indexed to the mean coverage, in three sequencing experiments using DNA from the NA18506 HapMap cell line (black line), fresh-frozen lung cancer tissue (blue broken line) and FFPE lung cancer tissue (red broken line).

enrichment factor [=(sequenced bp in ROI/sequenced bp off ROI)/(size of ROI/size of the genome)] of 200 000 (enrichment statistics for all samples are summarized in Table 2). Some amplified fragments will also contain adjacent off ROI sequence. With a strict coding-sequence target definition, 68% of the reads mapped on target, and 83% on target ±50 bp. The reproducibility in coverage between samples is illustrated in Figure 1b and was calculated to 0.98 (R^2, linear regression). Furthermore, 98% of the targeted bases were covered at >10% of the mean base coverage (mean coverage = 273, Figure 1c and Supplementary Figure S1). To investigate how much input DNA is required for the method, we performed enrichment experiments on different amounts of template DNA (Table 3). We observed that the specificity decreased when <800-ng DNA was added to the reaction, and therefore we chose not to use less DNA than 800 ng per sample in this study.

Table 3. Specificity values obtained for experiments with different amount of input DNA

Input DNA (ng)	Specificity[a] (%)
1600	96.0
800	91.0
400	83.9
200	75.0
100	66.0

[a]Percentage of the uniquely aligned reads to hg18 that aligned uniquely to the region specified.

It is of great importance that an enrichment technique introduces minimal distortion of the original allele ratios in the analyzed sample. To assess the ability of the presented Selector protocol to preserve original allele ratios, we sequenced a HapMap trio

(NA18506, NA18506 and NA18508) and compared our calls with results made available within the HapMap project (26). Reference allele-frequency analysis was performed using MosaikAligner and the result was compared to the genotypes of the 164 available SNPs in the target sequence in the HapMap database which overlapped with the ROI (Figure 2 and Table 4). The concordance was 100% for covered SNPs within ROI and 99% including targeted SNPs outside of ROI (in total 383).

To investigate the utility of this enrichment technique for analysis of somatic mutations in cancer, we sequenced tumor and matched control samples of two breast-cancer cell lines (HCC1143 and HCC1599). By comparing the variant calls from the tumor and the matched normal tissue, it is possible to distinguish germline variants from somatic mutations. The results from comparing allele ratios are plotted in the lower panels of Figure 3a and b and the calls are summarized in Table 5. The difference in allele ratio between the tumor and normal cell line was >0.3 in 14 ROI positions in HCC1143, and 7 ROI positions in HCC1599. The differences in HCC1143 were due to loss-of-heterozygosity (LOH) in eight genes, caused by allelic amplification (CCND1), deletion [MET, CDKN2A, MRE11A, ATM, NF1 and ERBB2 (HER2)] and copy-number neutral LOH (uniparental disomy, APC). An (A→G) mutation was found in the HER3 gene outside the ROI. The allele-ratio differences in HCC1599 were due to LOH in five genes, caused by allelic amplification (CCNE1) and deletions (APC, EGFR, TP53 and NF1). Finally, one (T→A) mutation in the TP53 gene was identified just outside the ROI in association with loss of the normal allele, which confirms a previous study of the somatic mutation spectrum in this cell line (27). In addition to mutation detection, we also compared the relative read depth at different positions in the ROI to detect copy-number alterations. Comparison of relative read depth at each position in the ROI in the cell line pairs HCC1143/HCC1143BL and HCC11599/HCC11599BL are plotted in upper panels in Figure 3a and b, respectively. To exclude that the differences in coverage originated from poor reproducibility of the assay, we also analyzed a replicate sequencing experiment of the NA18506 DNA sample in the same manner (Figure 3c). The analysis revealed considerable relative copy-number variation between the tumor and normal cell line, including amplification, duplication or deletion of the majority of investigated genes.

Finally, we evaluated the technique on DNA prepared from normal and tumor tissue from a lung-cancer patient. We applied the same subtractive approach to analyze the mutational status of the 28 genes as with the cancer cell line samples. We detected copy-number deviations that confirmed previous microarray analysis performed using an Affymetrix 250K SNP chip (Figure 4a, Supplementary Figure S2). The deviations were smaller in this experiment than for the cell lines, which is consistent with a tumor-cell content of ~50% in this sample. The correlation between the array-based copy-number estimate and our sequencing-based estimate proved to be strong ($R^2 = 0.9999$) (Figure 4b). We detected a single base-pair deletion that had previously been identified in the TP53 gene by Sanger sequencing (28) (Figure 4c). The majority of patient tumor samples are prepared and stored as formalin-fixed paraffin embedded (FFPE) specimens. The potential utility of a somatic mutation analysis technique is therefore greatly enhanced if it is compatible with DNA from FFPE tissue samples, which are often severely degraded. To evaluate this potential we amplified and sequenced the 28 genes using DNA from an FFPE sample from the same lung cancer patient. We found the TP53 deletion mutation also in this material (Figure 4d). By quantifying the reads mapping

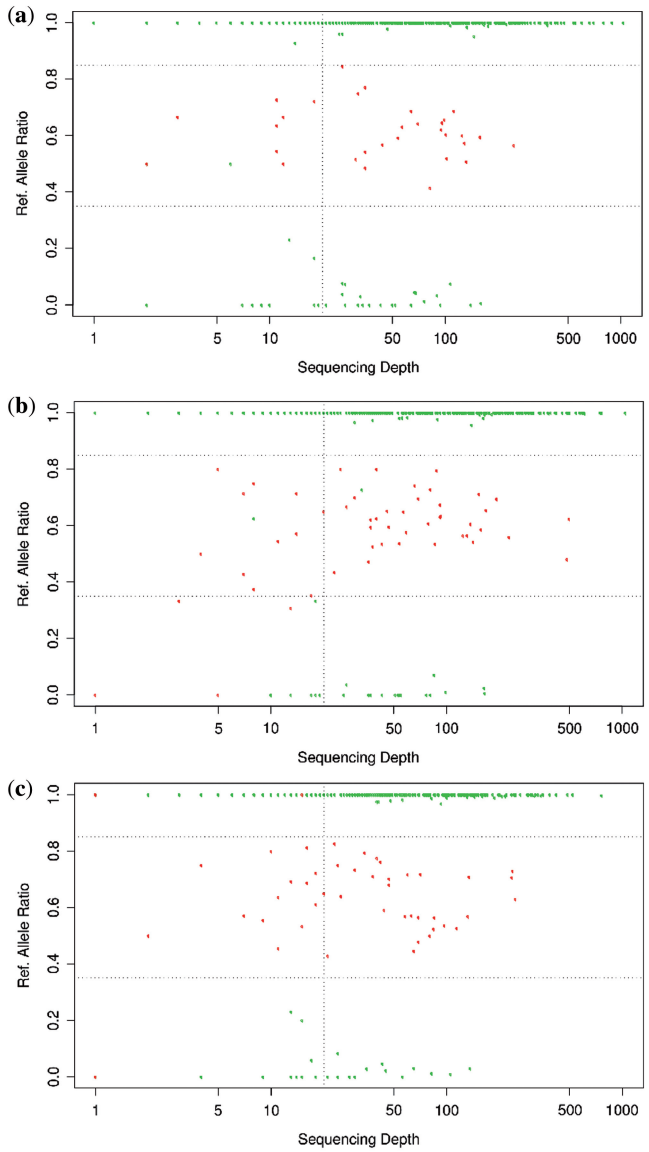

Figure 2. Analysis of concordance with available SNP genotypes of the (**a**) NA18506; (**b**) NA18507; and (**c**) NA18508 DNA samples. The allele ratio of all positions where the SNP genotype is available in the HapMap database are plotted as a function of read depth. Green dots represent homozygous SNP positions in the database and red dots represent heterozygous positions. Heterozygous positions were defined as having an allelic ratio of ≥0.35 and ≤0.85 (dashed horizontal lines). The vertical dashed line indicates a read depth of 20.

Table 4. SNP concordance with data available from three HapMap samples (NA18506, NA18507 and NA18508) at two coverage thresholds: covered at least once and covered at least 20 times

Sample	Relation	Coverage	Region	Number of SNPs[a]	Covered (%)	Homozygote[b]		Heterozygote		Concordance (%)
						Selectors	Hapmap	Selectors	Hapmap	
NA18506	Son	≥1×	ROI	165	100.00	161	161	4	4	100.00
			Amplified region	382	99.21	346	347	33	32	99.74
		≥20×	ROI	165	91.52	147	147	4	4	100.00
			Amplified region	382	75.39	265	265	23	23	100.00
NA18507	Father	≥1×	ROI	164	100.00	151	151	13	13	100.00
			Amplified region	382	99.21	330	326	49	53	98.42
		≥20×	ROI	164	92.07	138	138	13	13	100.00
			Amplified region	382	74.87	248	249	38	37	99.65
NA18508	Mother	≥1×	ROI	165	100.00	151	151	14	14	100.00
			Amplified region	383	98.96	334	329	45	50	98.68
		≥20×	ROI	165	80.61	122	122	11	11	100.00
			Amplified region	383	62.92	211	211	30	30	100.00

Positions are considered as heterozygote if the reference allele ratio is between 0.35 and 0.85.
[a]Number of HapMap SNPs overlaping with region. NN positions omitted.
[b]Reference allele ratio under 0.35 or above 0.85.

to junctions of the circularized fragments, we were able to retrieve information about the performance of individual selectors. We observed worse representation of the longer fragments than the shorter in the FFPE sample compared to the fresh frozen (Figure 5), resulting in more uneven amplification (Figure 1c**)**. However, the difference in coverage between the shortest and the longest fragments in the range 100–300 bp was less than a factor of two, so with a different probe design targeting shorter fragments, evenness of coverage should be improved.

DISCUSSION

In this article, we report on a method for targeted enrichment of a relatively small subset of genes that produces the targeted regions in an unbiased fashion with very low amount of off-target material. For a small set of samples, it would be possible to sequence the same region with PCR and Sanger sequencing with an estimated 384-well plate of reactions for each sample. We are however convinced that the proposed scheme is more practical also for small collections of samples.

Compared to the previously published version of the selector technology (13,29), the current protocol uses MDA instead of PCR. By eliminating the PCR primers in the probes, the oligonucleotide length, and thereby also cost, is significantly reduced. Probe cost has further been reduced through the introduction of a solid-phase purification step to remove excess probe molecules, replacing the enzymatic degradation used in the previous protocol, requiring expensive uracil residues in the probes. The invasive cleavage that was used to generate the majority of fragments in the previous protocol was avoided in this version, since we observed worse bias when it was used (data not shown). Thus in this study, we only used restriction digestion to generate the fragments. We have further increased design redundancy, which should improve coverage. It is difficult to assess whether MDA introduce less or more amplification bias compared to PCR, because of the probe redundancy and since different sequencing platforms were used to sequence the PCR amplicons (454, long read) and the MDA products (SOLiD, short read). In contrast to PCR, the MDA amplification generates a double-stranded high-molecular weight amplification product that is very similar in nature to genomic DNA. We have demonstrated compatibility with the standard sample preparation procedures for both SOLiD and Illumina short-read instruments. In contrast, to be compatible with the short-read-sequencing platforms, a PCR product requires concatemerization by ligation, followed by fragmentation, which is a more complicated scheme. For emulsion PCR, an additional fragmentation step is also required before the PCR resulting in a very laborious procedure. The MDA also avoids clonal propagation of polymerization artifacts by employing rolling circle amplification, a feature valuable for detection of rare variants.

In addition to the importance of maintaining high specificity, only enriching regions that are targeted, it is also important to minimize the number of near-target base pairs that are amplified. The restriction enzyme-based approach of the selector technology provides very high-on-target rate by design, between 50–70% of the amplicons are on exons in the majority of designs (30). This is comparable and even slightly better than published work using parallel PCR on the Raindance platform (10). In this article exons with an average size of 167 bp where targeted with amplicons of between 300 and 600 bp, which will generate a significant proportion of near-exon sequence. The selector amplicons range between 100 and 1000 and can more easily be adjusted to fit the target length when selecting from eight different restriction enzyme reactions (30).

Multiplex targeted sequencing is of particular interest for molecular cancer diagnostics as genetic aberrations in growth-factor-signaling pathways can be decisive

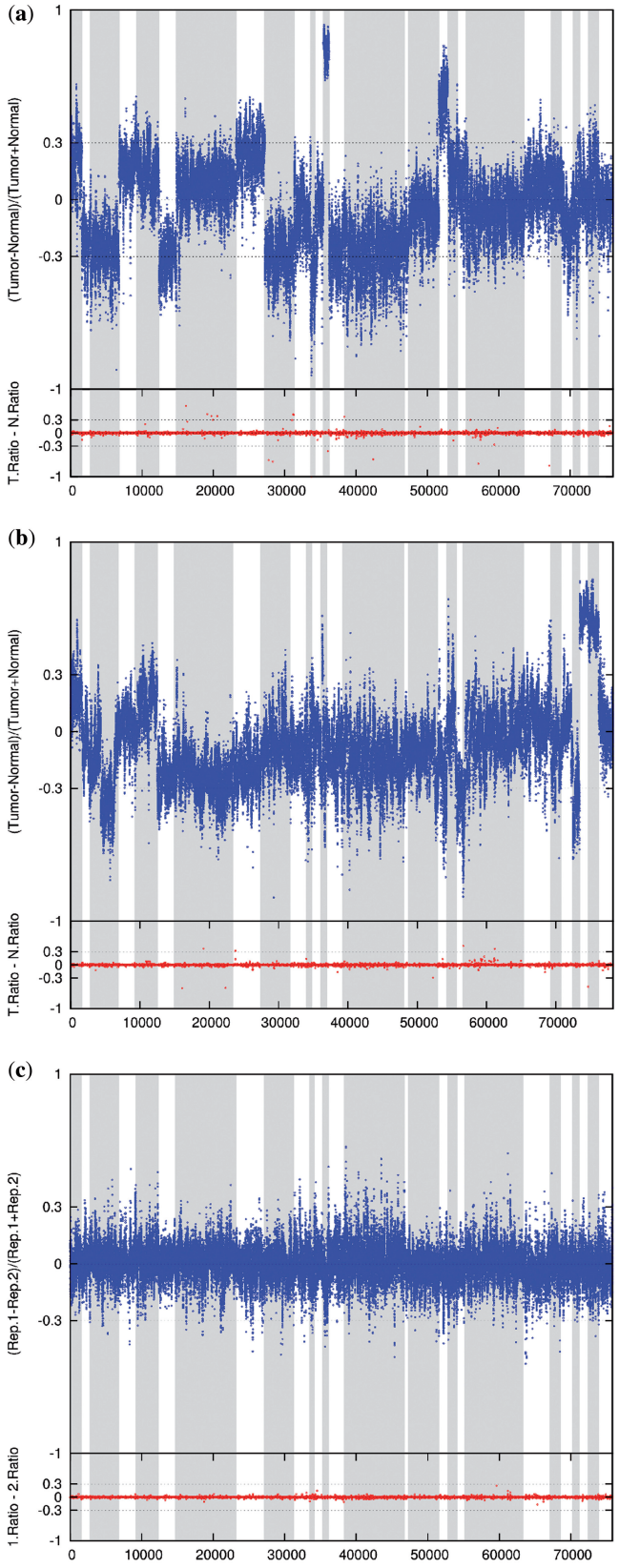

Figure 3. Somatic mutation analysis of two breast cancer cell lines with matched normal controls HCC1143/HCC1143BL (**a**) and HCC1599/HCC1599BL (**b**). In each position in the targeted region, the number of bases called in the matched normal cell line (HCC1143BL/HCC1599BL) is subtracted from the number of bases called in the tumor-derived cell line (HCC1143/HCC1599) and then divided with the sum of called bases for that position in the two samples. The exons of the 28 genes are concatenated in the order specified in Table 1 and gene shifts are demarked by alternating background color. The lower panels show the value obtained from each base when the allelic ratio between the major and minor allele of the normal sample is subtracted from the corresponding allelic ratio of the tumor sample. (**c**) Base-by-base comparison of two replicate sequencing experiments of the cell line NA18506.

for treatment with certain novel drugs. When several genes have to be analyzed separately in an old pipeline using PCR followed by Sanger sequencing, the diagnostic process becomes time-consuming and costly. Also, as tumor material often is limited, e.g. in the form of small biopsies, it is difficult to obtain sufficient DNA for a multitude of separate targeted-mutation assays. The current selector protocol consumes only 800 ng of genomic DNA. We believe that this method can be useful for clinical diagnostics since it can be applied on routinely collected FFPE samples.

The Selector probe set used in this study enriched for nearly all bases within the ROI except for only 1.4% (30 regions) of the bases being unavailable applying the current design criteria. When using targeted sequencing for diagnostic purposes, it is often important to be able to analyze all bases within the ROI that can be uniquely mapped with the read length produced by the sequencing instrument. The bases that we failed to cover with selected fragments either contained repetitive elements or SNPs in the hybridization sites. To achieve complete coverage, manual examination of the failed regions enables us to allow some fragments back into the design that have been previously rejected by the restrictive design filters.

Targeted resequencing is expected to be one of the major applications for next-generation sequencing. The current high-throughput platforms are in some aspects suboptimal for targeted resequencing. To meet the demand of quick and simple sequencing, several scaled down sequencing platforms have been announced (e.g. 454 GS Junior, iScan and Ion Torrent), and to harness the power of these instruments, simple, efficient and high-quality enrichment technologies are needed. We present a technology that offers high-performance targeted analyses using a simple and scalable protocol. To our knowledge, this is the only procedure for scalable multiplex enrichment of targets that achieves high (>95%) coverage and high (>95%) specificity without introducing requirements of dedicated instrumentation. Versions of the technology enabling introduction of sequencing primers and barcodes during amplification are under development. So far, direct library free amplicon sequencing has been limited to the 454 sequencing platform, however, development of existing and new platforms will make this an attractive option for several sequencing technologies.

Table 5. Potential somatic mutations are listed as defined as base positions with an absolute difference in reference allele ratio of at least 0.30 between tumor and normal samples

Sample	Chr (hg18)	Position (hg18)	Gene	Reference sequence	Normal					Tumor					T.Ratio − N.Ratio[a]	Call	Location
					Ratio	A	G	C	T	Ratio	A	G	C	T			
HCC1143	NC_000005.8	112190753	APC	T	0.38	0	0	52	32	1.00	0	0	0	78	0.62	LOH, CNN	ROI[b]
HCC1143	NC_000005.8	112203669	APC	G	0.57	65	85	0	0	1.00	0	160	0	0	0.43	LOH, CNN	ROI[b]
HCC1143	NC_000005.8	112204224	APC	G	0.61	189	292	0	0	1.00	3	597	0	0	0.39	LOH, CNN	ROI[b]
HCC1143	NC_000005.8	112205070	APC	G	0.61	43	68	0	0	1.00	0	144	0	0	0.39	LOH, CNN	ROI[b]
HCC1143	NC_000007.12	116126908	MET	C	0.62	0	0	114	69	0.00	0	0	0	43	−0.62	LOH	ROI[b]
HCC1143	NC_000007.12	116127498	MET	A	0.74	62	22	0	0	0.08	2	21	0	0	−0.65	LOH	ROI[b]
HCC1143	NC_000007.12	116223258	MET	G	0.57	116	152	0	0	0.99	1	114	0	0	0.42	LOH	ROI[b]
HCC1143	NC_000007.12	116223333	MET	G	0.59	75	136	0	0	1.00	0	89	0	0	0.41	LOH	ROI[b]
HCC1143	NC_000009.10	21961039	CDKN2A	G	1.00	0	28	0	0	0.00	0	0	0	0	−1		ROI[b]
HCC1143	NC_000011.8	69172091	CCND1	G	0.49	70	68	0	0	0.08	531	146	0	0	−0.41	LOH, AMP	ROI[b]
HCC1143	NC_000011.8	93865568	MRE11A	C	0.63	0	0	20	12	1.00	0	0	27	0	0.38	LOH	ROI[b]
HCC1143	NC_000011.8	107668697	ATM	C	0.66	0	0	193	100	0.05	0	0	5	87	−0.60	LOH	ROI[b]
HCC1143	NC_000017.9	26577611	NF1	G	0.7	12	28	0	0	0.00	14	0	0	0	−0.70	LOH	ROI[b]
HCC1143	NC_000017.9	35137563	HER2	C	0.75	0	6	18	0	0.00	0	27	0	0	−0.75	LOH	ROI[b]
HCC1599	NC_000005.8	112182868	APC	C	0.68	0	2	393	179	0.15	0	0	32	179	−0.53	LOH	ROI[b]
HCC1599	NC_000005.8	112203516	APC	T	0.62	85	0	0	136	0.99	1	0	0	86	0.37	LOH	ROI[b]
HCC1599	NC_000005.8	112206694	APC	G	0.56	147	187	1	0	0.03	117	4	0	0	−0.53	LOH	ROI[b]
HCC1599	NC_000007.12	55181842	EGFR	C	0.67	1	0	177	87	0.99	0	0	155	2	0.32	LOH	ROI[b]
HCC1599	NC_000017.9	7520197	TP53	G	0.57	0	26	19	1	1.00	0	10	0	0	0.43	LOH	ROI[b]
HCC1599	NC_000017.9	26610220	NF1	A	0.48	29	31	0	0	0.85	23	4	0	0	0.37	LOH	ROI[b]
HCC1599	NC_000019.8	35006506	CCNE1	C	0.65	0	0	108	57	0.16	0	0	75	401	−0.50	AMP	ROI[b]
HCC1143	NC_000002.10	212252269	HER4	G	0.52	14	15	0	0	0.21	19	5	0	0	−0.31		Amp. Region[c]
HCC1143	NC_000007.12	116185999	MET	C	0.62	0	0	34	21	0.05	0	0	1	20	−0.57	LOH	Amp. Region[c]
HCC1143	NC_000011.8	93851696	MRE11A	C	0.69	0	0	27	12	1.00	0	0	23	0	0.31	LOH	Amp. Region[c]
HCC1143	NC_000011.8	93865455	MRE11A	C	0.65	0	0	28	15	0.96	0	0	23	1	0.31	LOH	Amp. Region[c]
HCC1143	NC_000012.10	54763961	HER3	A	0.58	131	0	0	96	0.00	0	0	0	122	−0.58	LOH	Amp. Region[c]
HCC1143	NC_000012.10	54764761	HER3	T	0.97	0	0	1	32	0.62	0	0	3	5	−0.34		Amp. Region[c]
HCC1143	NC_000012.10	54766850	HER3	G	0.46	21	18	0	0	0.00	22	0	0	0	−0.46	LOH	Amp. Region[c]
HCC1143	NC_000012.10	54768447	HER3	T	0.68	0	15	0	32	0.06	0	16	0	1	−0.62	LOH	Amp. Region[c]
HCC1143	NC_000012.10	54778054	HER3	A	1.00	195	0	0	0	0.05	3	60	0	0	−0.95	A → G	Amp. Region[c]
HCC1143	NC_000017.9	26510278	NF1	G	0.48	24	22	0	0	1.00	0	49	0	0	0.52	LOH, CNN	Amp. Region[c]
HCC1143	NC_000017.9	26584058	NF1	C	0.65	32	0	59	0	0.10	47	0	5	0	−0.55	LOH, CNN	Amp. Region[c]
HCC1143	NC_000017.9	26584383	NF1	G	0.64	11	21	0	0	1.00	0	40	0	0	0.36	LOH, CNN	Amp. Region[c]
HCC1143	NC_000017.9	26677419	NF1	T	0.7	0	0	25	57	1.00	0	0	0	96	0.30	LOH, CNN	Amp. Region[c]
HCC1143	NC_000017.9	26679002	NF1	T	0.6	81	4	0	129	1.00	0	0	0	225	0.40	LOH, CNN	Amp. Region[c]
HCC1143	NC_000017.9	35119531	HER2	C	0.53	0	0	75	66	0.00	0	0	0	145	−0.53	LOH, CNN	Amp. Region[c]
HCC1143	NC_000017.9	35122241	HER2	C	0.6	0	0	36	24	0.02	0	0	1	43	−0.58	LOH, CNN	Amp. Region[c]
HCC1599	NC_000001.9	241867813	AKT3	T	0.67	0	0	10	20	0.33	0	0	18	9	−0.33		Amp. Region[c]
HCC1599	NC_000002.10	212134711	HER4	A	0.59	24	0	17	0	0.13	1	0	7	0	−0.46	LOH	Amp. Region[c]
HCC1599	NC_000002.10	212192049	HER4	T	0.67	0	0	7	14	1.00	0	0	0	2	0.33		Amp. Region[c]
HCC1599	NC_000002.10	212252169	HER4	A	0.56	84	65	0	0	0.14	8	48	0	0	−0.42	LOH	Amp. Region[c]
HCC1599	NC_000002.10	212252269	HER4	G	0.71	10	24	0	0	0.00	10	0	0	0	−0.71	LOH	Amp. Region[c]
HCC1599	NC_000002.10	212323573	HER4	C	0.67	0	0	250	122	0.98	0	0	208	4	0.31	LOH	Amp. Region[c]
HCC1599	NC_000002.10	212323761	HER4	T	0.63	0	0	39	66	0.29	0	0	25	10	−0.34	LOH	Amp. Region[c]
HCC1599	NC_000004.10	153464702	FBXW7	C	0.43	0	12	9	0	0.00	0	21	0	0	−0.43	LOH	Amp. Region[c]
HCC1599	NC_000004.10	153551830	FBXW7	T	0.61	0	11	0	17	1.00	0	0	0	26	0.39	LOH	Amp. Region[c]
HCC1599	NC_000007.12	55182141	EGFR	G	0.45	0	14	0	17	1.00	0	8	0	0	0.55	LOH	Amp. Region[c]
HCC1599	NC_000007.12	55187671	EGFR	A	0.66	61	32	0	0	0.97	57	2	0	0	0.31	LOH	Amp. Region[c]
HCC1599	NC_000007.12	55196682	EGFR	G	0.55	94	116	0	0	1.00	0	97	0	0	0.45	LOH	Amp. Region[c]

(continued)

Table 5. Continued

Sample	Chr (hg18)	Position (hg18)	Gene	Reference sequence	Normal					Tumor					T.Ratio − N.Ratio[a]	Call	Location
					Ratio	A	G	C	T	Ratio	A	G	C	T			
HCC1599	NC_000007.12	55227257	EGFR	A	0.66	78	41	0	0	0.03	1	28	0	0	−0.62	LOH	Amp. Region[c]
HCC1599	NC_000007.12	55240320	EGFR	G	0.77	17	56	0	0	0.00	9	0	0	0	−0.77	LOH	Amp. Region[c]
HCC1599	NC_000012.10	25269723	KRAS	T	0.76	0	0	6	19	0.00	0	0	4	0	−0.76		Amp. Region[c]
HCC1599	NC_000012.10	54764729	HER3	A	1.00	20	0	0	0	0.67	2	1	0	0	−0.33		Amp. Region[c]
HCC1599	NC_000012.10	54764863	HER3	T	1.00	0	0	0	25	0.60	0	0	2	3	−0.40		Amp. Region[c]
HCC1599	NC_000012.10	54766915	HER3	G	0.7	17	40	0	0	0.33	24	12	0	0	−0.37		Amp. Region[c]
HCC1599	NC_000012.10	54780089	HER3	A	0.67	65	0	32	0	0.34	22	0	42	0	−0.33		Amp. Region[c]
HCC1599	NC_000014.7	104310237	AKT1	T	0.65	0	0	36	68	1.00	0	0	0	49	0.35	LOH	Amp. Region[c]
HCC1599	NC_000014.7	104329938	AKT1	A	0.51	69	66	0	0	1.00	51	0	0	0	0.49	LOH	Amp. Region[c]
HCC1599	NC_000015.8	65244539	SMAD3	G	0.63	0	98	56	0	0.99	0	124	1	0	0.36	LOH	Amp. Region[c]
HCC1599	NC_000015.8	65244934	SMAD3	C	0.67	8	0	16	0	1.00	0	0	14	0	0.33	LOH	Amp. Region[c]
HCC1599	NC_000015.8	65264504	SMAD3	C	0.64	0	0	32	18	0.97	0	0	37	1	0.33	LOH	Amp. Region[c]
HCC1599	NC_000017.9	7518335	TP53	T	1.00	0	0	0	48	0.10	9	0	0	1	−0.90	T → A	Amp. Region[b]
HCC1599	NC_000019.8	1157823	STK11	A	1.00	25	0	0	0	0.67	4	0	0	1	−0.33		Amp. Region[c]
HCC1599	NC_000019.8	1158238	STK11	G	0.63	0	153	0	90	1.00	0	65	0	0	0.37	LOH	Amp. Region[c]
HCC1599	NC_000019.8	1158280	STK11	C	0.58	0	0	188	134	0.98	0	0	91	2	0.39	LOH	Amp. Region[c]
HCC1599	NC_000019.8	1169523	STK11	G	0.62	0	39	0	24	1.00	0	30	0	0	0.38	LOH	Amp. Region[c]
HCC1599	NC_000019.8	1170274	STK11	G	0.55	13	16	0	0	1.00	0	9	0	0	0.45	LOH	Amp. Region[c]
HCC1599	NC_000019.8	1177772	STK11	C	0.49	0	0	23	24	1.00	0	0	9	0	0.51	LOH	Amp. Region[c]
HCC1599	NC_000019.8	1177901	STK11	G	0.33	0	11	0	22	1.00	0	6	0	0	0.67	LOH	Amp. Region[c]
HCC1599	NC_000019.8	35005416	CCNE1	C	0.77	0	0	20	6	0.13	0	0	3	20	−0.64		Amp. Region[c]
HCC1599	NC_000019.8	35005443	CCNE1	C	0.48	1	12	12	0	0.08	0	66	6	0	−0.40	LOH	Amp. Region[c]
HCC1599	NC_000019.8	45435698	AKT2	G	0.7	25	58	0	0	0.22	169	49	0	0	−0.48	LOH, AMP	Amp. Region[c]

The number of respective base call is shown for both normal and tumor sample, the reference allele ratios, the reference allele ratio and our interpretation for the probable cause for the ratio difference are listed for all positions.

[a]Only base positions with an absolute ratio difference above 0.3 between tumor and normal are listed.

[b]The position is located within the region of interest.

[c]The position is located within the amplified region but not in the region of interest.

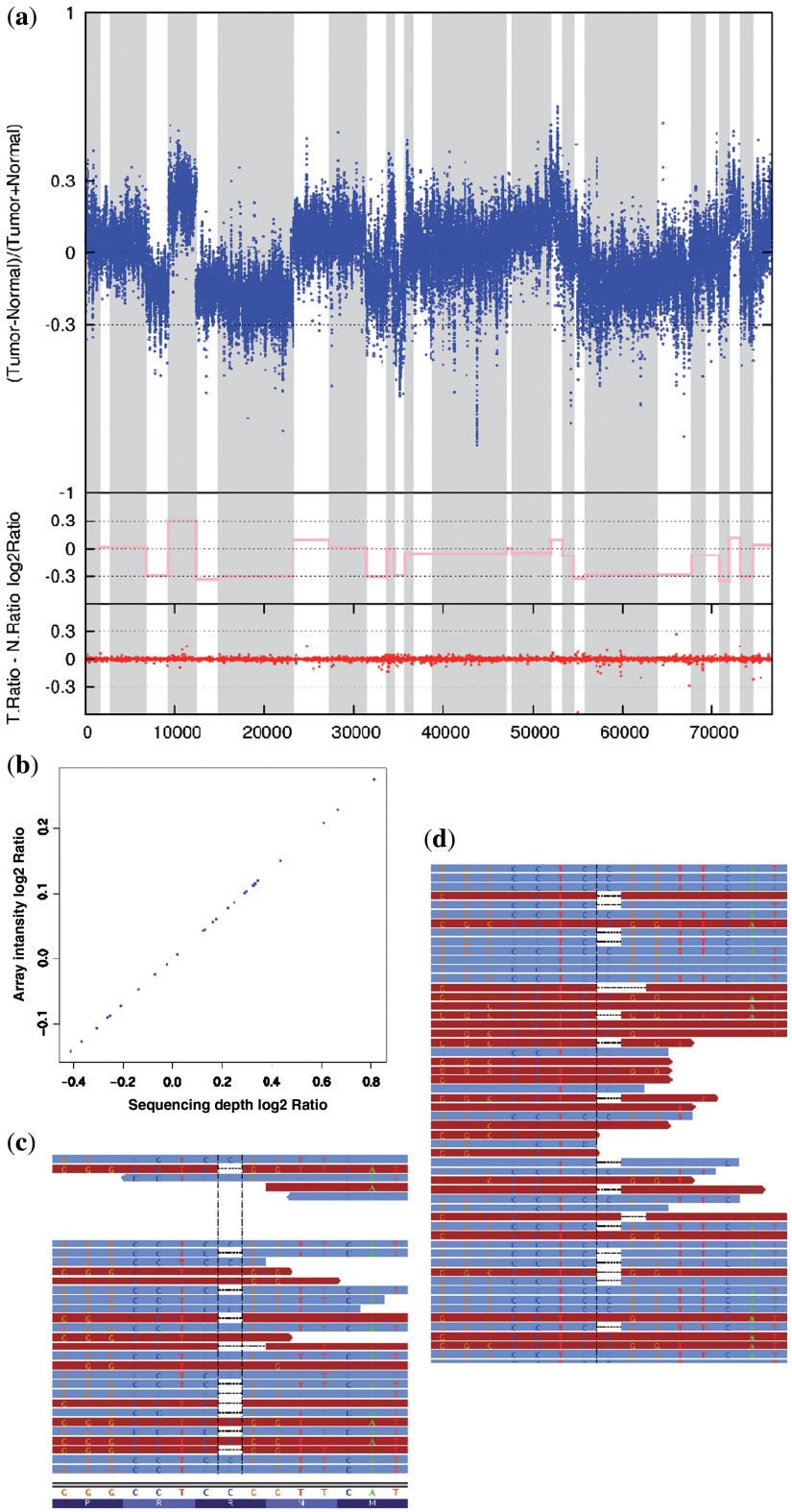

Figure 4. (**a**) Mutation detection and CNV analysis of a lung cancer patient sample with matched control. Upper panel: for each position in the targeted region the number of bases called in the patient-matched normal tissue is subtracted from the number of bases called in the tumor-derived DNA, and then divided with the sum of called bases for that position in the two samples. The exons of the 28 genes are lined up after each other and genes are demarked by alternating background color. Middle panel: the inferred gene copy-number variation in the corresponding genomic loci illustrated by log2 ratios (pink line) derived from SNP array data (Affymetrix Gene Chip Mapping 250K arrays). Middle panel: the log2 ratio (pink line) of the copy-number analysis done on an Affymetrix micro array. Lower panel: the allelic ratio between the major and minor allele at each position is compared between the two samples by subtraction (**b**). A correlation plot between the Affymetrix Gene Chip log2 tumor/normal signal ratio and the log2 tumor/normal sequencing read depth ratio. (**c**) Detection of a single base pair deletion in the TP53 gene. Forward (brown) and reverse (blue) reads are aligned to a 15-bp region of the TP53 gene. Deleted bases are indicated by dashed lines. Alignment visualized in Integrative Genomic Viewer (IGV ver.1.4.2). (**d**) Detection of the same mutation in the FFPE sample from the same tumor.

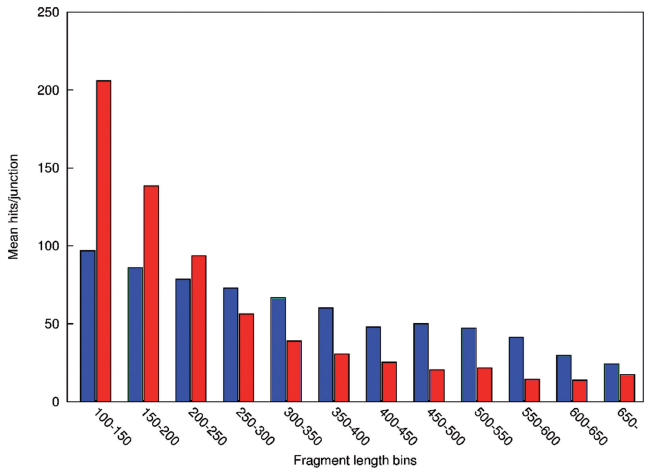

Figure 5. The dependency of the length of selected fragments on amplification efficiency in fresh-frozen and FFPE tumor tissue. Each selector probe in a probe library creates a unique sequence over the ligation junction upon successful ligation and amplification. The junction reads can be used to track the performance of individual probes and fragments. In the graph, the average number of reads spanning a ligation junction of a probe is plotted as a function of the length of the targeted fragments. Fragment lengths are shown in bins of 50 bp. Data from the fresh-frozen and FFPE tumor samples are shown as blue and red bars, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Inger Jonasson at Uppsala Genome Center for the SOLiD sequencing, Yadhu Kumar at GATC Biotech for the Illumina sequencing and Ulf Landegren and Lotte Moens for comments on the manuscript.

FUNDING

Knut and Alice Wallenberg, Göran Gustafsson, and Erik K Fernström foundations; the Swedish Research Council; VINNOVA; EU FP7 project READNA (grant agreement HEALTH-F4-2008-201418); EU FP7 project EUROGENESCAN. Funding for open access charge: EU FP7 project READNA.

Conflict of interest statement. H.J., M.I., J.S., O.E. and M.N. own shares in the company Olink Genomics AB, Uppsala, Sweden, that holds the commercial rights to the technique.

REFERENCES

1. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
2. Margulies,M., Egholm,M., Altman,W.E., Attiya,S., Bader,J.S., Bemben,L.A., Berka,J., Braverman,M.S., Chen,Y., Chen,Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
3. Drmanac,R., Sparks,A.B., Callow,M.J., Halpern,A.L., Burns,N.L., Kermani,B.G., Carnevali,P., Nazarenko,I., Nilsen,G.B., Yeung,G. *et al.* (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*, **327**, 78–81.
4. Smith,D.R., Quinlan,A.R., Peckham,H.E., Makowsky,K., Tao,W., Woolf,B., Shen,L., Donahue,W.F., Tusneem,N., Stromberg,M.P. *et al.* (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.*, **18**, 1638–1642.
5. Hodges,E., Xuan,Z., Balija,V., Kramer,M., Molla,M.N., Smith,S.W., Middle,C.M., Rodesch,M.J., Albert,T.J., Hannon,G.J. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, **39**, 1522–1527.
6. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.
7. Meuzelaar,L.S., Lancaster,O., Pasche,J.P., Kopal,G. and Brookes,A.J. (2007) MegaPlex PCR: a strategy for multiplex amplification. *Nat. Methods*, **4**, 835–837.
8. Varley,K.E. and Mitra,R.D. (2008) Nested patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res.*, **18**, 1844–1850.
9. Weinstein,J.A., Jiang,N., White,R.A., Fisher,D.S. and Quake,S.R. (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science*, **324**, 807–810.
10. Tewhey,R., Warner,J.B., Nakano,M., Libby,B., Medkova,M., David,P.H., Kotsopoulos,S.K., Samuels,M.L., Hutchison,J.B., Larson,J.W. *et al.* (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat. Biotechnol.*, **27**, 1025–1031.
11. Porreca,G.J., Zhang,K., Li,J.B., Xie,B., Austin,D., Vassallo,S.L., LeProust,E.M., Peck,B.J., Emig,C.J., Dahl,F. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nat. Methods*, **4**, 931–936.
12. Turner,E.H., Lee,C., Ng,S.B., Nickerson,D.A. and Shendure,J. (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods*, **6**, 315–316.
13. Dahl,F., Stenberg,J., Fredriksson,S., Welch,K., Zhang,M., Nilsson,M., Bicknell,D., Bodmer,W.F., Davis,R.W. and Ji,H. (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl Acad. Sci. USA*, **104**, 9387–9392.
14. Han,J., Swan,D.C., Smith,S.J., Lum,S.H., Sefers,S.E., Unger,E.R. and Tang,Y. (2006) Simultaneous amplification and identification of 25 human papillomavirus types with Templex technology. *J. Clin. Microbiol.*, **44**, 4157–4162.
15. Albert,T.J., Molla,M.N., Muzny,D.M., Nazareth,L., Wheeler,D., Song,X., Richmond,T.A., Middle,C.M., Rodesch,M.J., Packard,C.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, **4**, 903–905.
16. Burbano,H.A., Hodges,E., Green,R.E., Briggs,A.W., Krause,J., Meyer,M., Good,J.M., Maricic,T., Johnson,P.L.F., Xuan,Z. *et al.* (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*, **328**, 723–725.
17. Summerer,D., Wu,H., Haase,B., Cheng,Y., Schracke,N., Stähler,C.F., Chee,M.S., Stähler,P.F. and Beier,M. (2009) Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Res.*, **19**, 1616–1621.
18. Mamanova,L., Coffey,A.J., Scott,C.E., Kozarewa,I., Turner,E.H., Kumar,A., Howard,E., Shendure,J. and Turner,D.J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, **7**, 111–118.
19. Turner,E.H., Ng,S.B., Nickerson,D.A. and Shendure,J. (2009) Methods for genomic partitioning. *Annu. Rev. Genomics Hum. Genet.*, **10**, 263–284.
20. Saiki,R.K., Scharf,F.A., Faloona,F.A., Mullis,C.T., Horn,H.A., Erlich,H.A. and Arnheim,N. (1985) Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**, 1350–1354.
21. Landegren,U. (1999) US Patent No, 6,558,928.

22. Dean,F.B., Hosono,S., Fang,L., Wu,X., Faruqi,A.F., Bray-Ward,P., Sun,Z., Zong,Q., Du,Y., Du,J. *et al*. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA*, **99**, 5261–5266.
23. Stenberg,J., Zhang,M. and Ji,H. (2009) Disperse–a software system for design of selector probes for exon resequencing applications. *Bioinformatics*, **25**, 666–667.
24. Stenberg,J., Nilsson,M. and Landegren,U. (2005) ProbeMaker: an extensible framework for design of sets of oligonucleotide probes. *BMC Bioinformatics*, **6**, 229.
25. Olshen,A.B., Venkatraman,E.S., Lucito,R. and Wigler,M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
26. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbol,P., Leal,S.M. *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
27. Sjöblom,T., Jones,S., Wood,L.D., Parsons,D.W., Lin,J., Barber,T.D., Mandelker,D., Leary,R.J., Ptak,J., Silliman,N. *et al*. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.
28. Sanger,F. and Coulson,A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol*., **94**, 441–448.
29. Dahl,F., Gullberg,M., Stenberg,J., Landegren,U. and Nilsson,M. (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res*., **33**, e71.
30. Stenberg,J., Dahl,F., Landegren,U. and Nilsson,M. (2005) PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. *Nucleic Acids Res*., **33**, e72.