ORIGINAL PAPER

Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays

Stephan Bau · Nadine Schracke · Marcel Kränzle · Haiguo Wu · Peer F. Stähler · Jörg D. Hoheisel · Markus Beier · Daniel Summerer

Received: 7 July 2008 / Revised: 2 October 2008 / Accepted: 6 October 2008 / Published online: 29 October 2008

Abstract We report a flexible method for selective capture of sequence fragments from complex, eukaryotic genome libraries for next-generation sequencing based on hybridization to DNA microarrays. Using microfluidic array architecture and integrated hardware, the process is amenable to complete automation and does not introduce amplification steps into the standard library preparation workflow, thereby avoiding bias of sequence distribution and fragment lengths. We captured a discontiguous human genomic target region of 185 kb using a tiling design with 50mer probes. Analysis by high-throughput sequencing using an Illumina/Solexa 1G Genome Analyzer revealed 2150-fold enrichment with mean per base coverage between 4.6 and 107.5-fold for the individual target regions. This method represents a flexible and cost-effective approach for large-scale resequencing of complex genomes.

Keywords Next-generation sequencing · Sequencing-by-synthesis · Microarrays · Microfluidics · Sequence enrichment · gDNA Library preparation

Introduction

The recent advent of a new generation of DNA-sequencing platforms has massively reduced the cost and effort of sequencing projects and holds promise to transform genetic variation studies to a much more systematic and comprehensive field [1–6].

However, though these platforms have vastly increased throughput compared with conventional Sanger technology, they are still far from being easily applicable to whole-genome resequencing projects of complex, eukaryotic organisms. In most cases, it would be desirable to focus on individual genomic subsets of interest by reducing the sequence complexity of the sample.

Among numerous methods for targeted enrichment of DNA sequences, PCR-based procedures have been the most widely used [7–9]. However, the discontiguous distribution of subsets with high information content, for example exons, and the limitations of PCR product length, specificity, and multiplexing grade make it expensive and laborious to use PCR enrichment in large-scale studies. Moreover, specific PCR demands the use of individual primers for each selected target which can severely impair the flexibility of the approach [8]. Thus, though PCR-based approaches are applicable for limited subsets of target regions, they do not

Stephan Bau and Nadine Schracke contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00216-008-2460-7) contains supplementary material, which is available to authorized users.

S. Bau · N. Schracke · M. Kränzle · M. Beier · D. Summerer (✉)
febit biomed gmbh,
Im Neuenheimer Feld 519,
69120 Heidelberg, Germany
e-mail: daniel.summerer@febit.de

H. Wu
febit Inc.,
99 Hayden Ave,
Lexington, MA 02421, USA

P. F. Stähler
febit holding gmbh,
Im Neuenheimer Feld 519,
69120 Heidelberg, Germany

J. D. Hoheisel
Deutsches Krebsforschungszentrum,
Im Neuenheimer Feld 580,
69120 Heidelberg, Germany

match the potential of next-generation sequencers for analysis of large numbers of genetic loci in massive parallel fashion.

Recently, the targeted selection of sequences by hybridization to microarray capture probes has been demonstrated to have the potential for efficient and selective enrichment of high-complexity sequences from large genomes [10–12]. The use of de novo-synthesized microarrays with custom content thereby allows cost-effective random access to virtually any high-complexity target sequences without limitation regarding spatial distribution within the genome. Further, custom array content facilitates the rapid, experimental prototyping of arrays for targeted subsets, thereby avoiding limitations of probe-calculation algorithms.

It has been shown that desired target sequences can be captured and analyzed by high-throughput sequencing, if starting amounts of ~20 μg of fragmented human genomic DNA (gDNA) per microarray are used and the enriched sample is PCR-amplified before further processing [10–12]. However, because gDNA sample amount obtained from biological material can often be a limiting factor, this demands significant whole-genome amplification (WGA) steps both before and after microarray processing. Because WGA methods are known to be susceptible for the introduction of bias into the amplified library [13, 14], this procedure complicates the use of quantitative analysis approaches such as SNP discovery or detection of somatic mutations and copy-number variations.

Here we report an approach for selective capture of genomic regions using de novo-synthesized microfluidic DNA chips with flexible content. The chip consists of eight individual channels each containing an array of capture probes providing adaptability to different sample numbers and target sizes.

The microfluidic chip architecture allows us to hybridize low sample amounts and enables recovery of sufficient enriched sample for direct sequencing using an Illumina/ Solexa 1G Genome Analyzer. For example, hybridization to one array (~3 μL total volume) requires the typical yield of the standard library preparation protocol from Illumina/ Solexa. Thus, no additional amplification steps have to be introduced into the standard workflow, avoiding unwanted bias in the library to be sequenced. The process is further highly amenable to automation in a closed system with the potential of enhanced reproducibility and parallelism and reduced risk of contamination.

Materials and methods

Microarray design and synthesis

Light-activated in-situ oligonucleotide synthesis was performed essentially as described using a digital micromirror device (Texas Instruments) for light-directed activation on an activated microfluidic array consisting of a glass—silicon-glass sandwich within the Geniom instrument (febit biomed gmbh, Heidelberg, Germany) [15]. Depending on the number of micromirrors used for one feature and for the spacing between features, each chip consists of eight arrays with 6,776 (2×2 mirrors for each feature with 1 mirror spacing), 15,624 (1 mirror for each feature with 1 mirror spacing) or 66,612 (1 mirror for each feature and no spacing) individual DNA oligonucleotide features. This results in a total content of ~54,000, ~125,000, or ~500,000 features. For enrichment of the *BRCA1, BRCA2*, and *TP53* genes, four arrays with 6,776 feature space each were used as enrichment matrix. Probes were distributed over the high complexity target sequence using a combination of two tiling designs of 50 mer probes targeting the sense strand. Two replica sets of probes with 12 bp tiling (total 13,296 probes) and three replica sets with 20 bp tiling (total 12,498) were synthesized together with control oligonucleotides on four arrays. To avoid nonspecific capturing of redundant sequences, probes having a low complexity base content of 80% (according to the Hg18 annotation) or 25 bases in row were excluded from the probe set.

DNA sample preparation

Human genomic DNA (Promega, Madison, WI, USA; 5 μg) was dissolved in 50 μL water and fragmented by sonication to a size distribution of 100–300 bp as judged by agarose gel electrophoresis. Preparation of the adaptor-ligated gDNA library ready for sequencing on an Illumina/ Solexa 1G Genome Analyzer (Illumina, San Diego, CA, USA) was performed according to the manufacturer's standard protocol and the size fraction of 200–400 bp was excised from an agarose gel after the adaptor ligation step. The sample was quantified by UV measurement (Nanodrop 1000; Thermo Scientific, Waltham, MA, USA) and stored in water at −20 °C until use.

Hybridization and elution

Adaptor-ligated gDNA library (6 μg) was dissolved in febit Hybmix-3 (20% formamide, 0.01% Tween-20) in the presence of 20 μg μL^{-1} tRNA, heated to 95 °C for 5 min and placed on ice. SSPE was added to a concentration of 4×, resulting in a final sample concentration of 75 ng μL^{-1}. The microarray was denatured with water at 80 °C and the sample mixture was injected into four connected microfluidic arrays. Hybridization was performed for 16 h at 50 °C with movement of the sample using a febit active mixing device. After hybridization each array was washed twice with 6× SSPE at room temperature and 0.5× SSPE at 45 °C. Each array was subsequently washed with 2 mL each of SSPE-based febit stringent wash buffers 1 (2× SSPE, 0.1% Tween-20, 5% formamide) and 2 (2× SSPE, 0.1% Tween-20, 10% formamide) at room temperature.

For elution of the enriched samples, arrays were filled with 10 µL 90% formamide in water each, incubated at 70 °C for 20 min, and solution was injected into a plastic tube. Sample was placed in a Speed-Vac and dried.

Analysis by sequencing on Illumina/Solexa 1G Genome Analyzer

Quantification of the sample by qPCR using the Illumina/Solexa primers used for PCR within the library-preparation protocol indicated a yield of 1.4 ng eluted material. Sequencing was performed according to the manufacturer's standard protocol. gDNA (1.12 ng, 8.64×10^{-15} mol) from the library with a size distribution of 200–400 bp was loaded on to one chamber of the flow cell and analyzed.

Data analysis

For the three human gene loci *BRCA1*, *BRCA2*, and *TP53*, the reads obtained were filtered for unique reads containing only true bases. The first 32 bp were blasted against target regions that were covered with probes allowing up to two mismatches. In cases of multiple blast hits, hits with the fewest mismatches were selected. Only blast-results passing these criteria were used for the final coverage calculation. Mean coverages for targeted regions were calculated from the sequencing tag numbers obtained. For the human LBR locus, the first 32 bp of reads were mapped using the CLC genomics workbench software applying standard mapping conditions.

Results and discussion

We chose to design a probe set to selectively capture the genomic sequences of the three human gene loci *BRCA1*, *BRCA2*, and *TP53* because of their functional role in cancer development. These discontiguous regions have a total size of 185 kb (81.2 kb, 84.2 kb, and 19.2 kb, respectively). Probes were tiled across the region with a distance of 12 or 20 bp targeting the sense strand, omitting low complexity regions. In total, four arrays of the microfluidic chip having a content of 25,794 features were used for the selected target. The rationale of choosing a specific tiling density and replicate number is thereby based on the target size and the desired amount of target molecules to be received for sequencing. Quantitative PCR on specific markers of the target region carried out with eluates of arrays with varying tiling densities clearly indicated that denser tiling positively affects the number of molecules that can be captured (Electronic Supplementary Material; Fig. S1). On the other hand, dense probe tiling reduces the size of the target region to be captured. However, qPCR data indicate an average loss of captured target molecules of only 35% comparing 12-bp and 24-bp tilings, indicating that capacity can be increased. Moreover, the use of different micromirror setups for light-directed probe synthesis could increase the number of features up to 66.612 (see "Materials and methods"). Using a 12-bp tiling, this results in a theoretical capacity of 800 kb per array or 6.4 Mb per chip.

To test the performance of the capture array, we chose the Illumina/Solexa 1G platform for downstream analysis due to its very high output of sequence information and the resulting potential for large-scale resequencing projects. Human genomic DNA was fragmented to a size of 100–300 bp by sonication, and adaptors for solid-phase amplification and sequencing-by-synthesis were attached according to the supplier's standard protocol. The library obtained was hybridized to the capture arrays and weakly bound fragments were removed by four consecutive washing steps of different stringency. Remaining fragments were eluted and the sample was directly used for processing on the sequencer without further workup (Fig. 1). Typically, between 0.5 and 1.5 ng enriched gDNA was recovered from four arrays. Applying 1.1 ng eluted DNA to one chamber of the flowcell of the Illumina/Solexa 1G, 5.6 million sequence reads of 36 bp or 201.6 Mb of total sequence output were obtained. Of these reads, >96% contained only true bases of which a high fraction (>92%) was unique, potentially reflecting the amplification-free processing after enrichment.

For data mapping using BLAST, only the core region of target sequences covered by capture probes without flanking regions was taken into account. In principle, reads can be obtained from flanking regions at a distance of up to one fragment length from a given capture probe, because of overlap [10–12]. However, mapping to these regions would increase coverage and enrichment data of the target sequence without positively affecting the sequence information originally targeted. Further, array design for gDNA enrichment in our study and in previous studies involves exclusion of probes potentially binding to low-complexity sequences, which results in discontinuous stretches of high complexity target within low-complexity context. Thus, large portions of flanking sequences will usually have low information content resulting in poor reliability of mapping data in these regions for the short reads obtained from the sequencer [11]. We used the first 32 bp of each read for mapping, allowing a maximum of two mismatches to improve BLAST specificity in comparison with previous studies [11]. Analysis revealed that 7% of reads containing only true bases mapped back to the total target region of the three genes. The desired core target sequences were

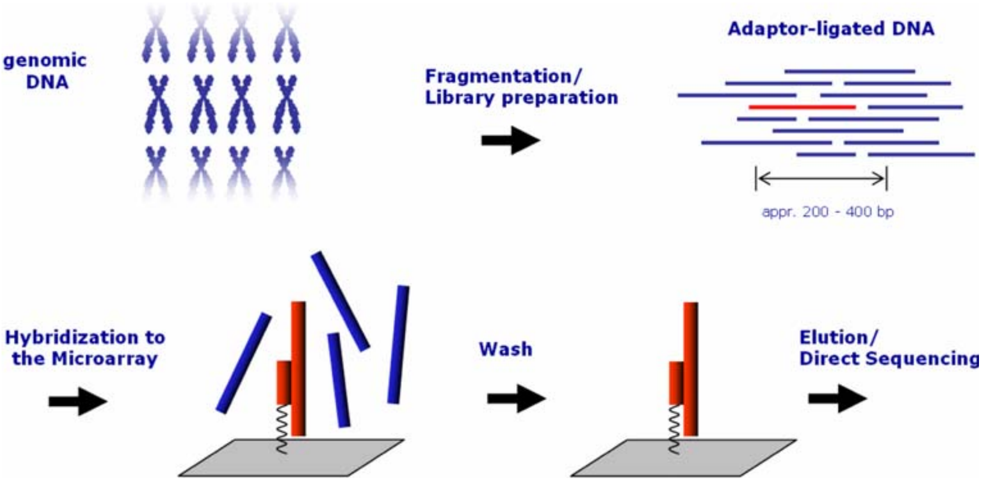
genomic DNA
Fragmentation/ Library preparation
Adaptor-ligated DNA
appr. 200 - 400 bp
Hybridization to the Microarray
Wash
Elution/ Direct Sequencing

Fig. 1 Targeted capture of selected genomic fragments using a microfluidic DNA-array. The gDNA sample is sheared, blunted, and adaptors for subsequent solid-phase PCR on an Illumina/Solexa 1G are attached. Library is injected into microchannels and hybridized to surface-bound capture probes. Weakly bound library fragments are washed away, remaining fragments are eluted and used directly for sequencing

covered by 10,415, 5,193, and 18,186 reads for *BRCA1*, *BRCA2*, and *TP53*, respectively. This corresponds to an overall enrichment factor of 2,150-fold. For the individual genes *BRCA1*, *BRCA2*, and *TP53*, enrichment factors of 2,150, 1,530, and 5630-fold were obtained (Table 1).

The observed variation in coverage observed for the three genes might reflect differences in sequence that affect hybridization, e.g. GC content. However, including properties such as melting temperatures, GC content, or loop rates in future probe designs has the potential to further improve homogeneity of binding. To obtain reliable data for studies such as SNP discovery, detection of somatic mutations or copy-number variations, coverage depth is critical. We obtained mean per base coverages of 4.6, 8.8, and 107.5-fold for *BRCA1*, *BRCA2* and *TP53*, respectively. A close-up of the distribution of sequence coverage for *BRCA1* is shown in the Electronic Supplementary Material, Fig. S2. Though the data obtained for these three individual genes suggest broad applicability to various sequence contexts, we further evaluated the performance of the approach using an additional locus. The human LBR gene, covering a target region of 277 kb, was captured using the described hybridization conditions.

The region was enriched using a 6-bp tiling density with denser tiling of 3-bp resolution for a ~48 kb part. One and a half million reads, or a total sequence output of 54 Mb, was obtained from the Illumina/Solexa 1G instrument. Of these reads, 66,663 mapped to the full target region, corresponding to an enrichment factor of 513-fold. However, 23,904 reads mapped to the 48 kb region corresponding to a 1,067-fold enrichment and indicating an impact of tiling density on the enrichment factor (Table 1). These data illustrate the enormous potential of targeted sequence capture using a microarray-based approach. The ability to select multiple genomic loci for large-scale genomic studies in a single purification step offers significant advantages compared with amplification-based approaches, for example long-range PCR. Besides technical ease and cost-effectiveness, the possibility of selectively targeting multiple unique sequence subsets with high information content within a context of redundant sequence cannot be achieved by amplification of long fragments. Approaches using padlock probes or molecular inversion probes have been reported recently and provide an interesting alternative to circumvent many of the problems associated with long-range PCR. However, these methods exhibited limitations either with respect to uniformity of capture efficiency [16] or the need for optimization of downstream amplification of specific target loci [17].

In comparison with previous studies based on microarray capturing, our results indicate comparable enrichment and coverage depth [10–12]. For example, in the only published

Table 1 Capture data for selected genes *BRCA1, BRCA2* and *TP53*

Gene	Total size (kb)	Size of core region (kb)	Obtained reads core region	Enrichment factor (fold)	Mean coverage per bp (fold)	Target region covered at least once (%)
BRCA1	81.2	42.0	10,415	2,150	8.8	75
BRCA2	84.2	39.4	5,193	1,530	4.6	46
TP53	19.2	7.6	18,186	5,630	107.5	84
LBR	277.2	197	66,663	513	6.1	58

The table shows the size of the target region and the size of core region (region covered by probes) for the captured genes together with the results from high throughput sequencing on an Illumina/Solexa 1G instrument study presenting Illumina/Solexa 1G data for enrichment analysis, mean coverages of 7.7 or 11.8-fold were reported for a 20-bp tiling design for enrichment of an exonic region. However, these data were obtained from much higher read numbers (9.5 or 7.4 million reads, respectively) compared with the data presented here, and coverage is expected to increase linearly with sequencing output.

Previous studies relied on extensive amplification of the target library both before and after the enrichment step, including the potential introduction of sequence bias and negative consequences for quantitative studies [13, 14]. In one study, bias of uniformity of the achieved coverage was described to be evident [10]. Further, PCR can also affect distribution of probe lengths within the library, which could affect performance of downstream sequencing by, e.g., causing inhomogeneous cluster sizes and/or reduced read numbers on a Illumina/Solexa 1G instrument. This could especially cause problems, because the small amounts of sample obtained from microarray hybridizations are difficult to analyze for fragment length distribution. The microfluidic architecture of the programmable array used in our study significantly reduces the volume needed for hybridization; moreover, yields of eluted library are sufficient to routinely apply the sample for downstream sequencing without prior amplification. We are currently transferring the approach to other next-generation sequencing platforms and are convinced that the method presented will find wide application for large-scale sequencing projects.

Acknowledgements We gratefully thank Dr Mark Chee and Dr Stephanie Pond for providing sequencing runs on an Illumina/Solexa 1G Genome Analyzer.

References

1. Bentley DR (2006) Curr Opin Genet Dev 16:545–552
2. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z (2008) Science 320:106–109
3. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Science 316:1497–1502
4. Shendure J, Mitra RD, Varma C, Church GM (2004) Nat Rev Genet 5:335–344
5. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Science 309:1728–1732
6. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Nature 437:376–380
7. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Science 307:1072–1079
8. Mullis KB, Faloona FA (1987) Methods Enzymol 155:335–350
9. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE (2006) Science 314:268–274
10. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA (2007) Nat Methods 4:903–905
11. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR (2007) Nat Genet 39:1522–1527
12. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007) Nat Methods 4:907–909
13. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, Richter DJ, Gabriel S, Herman P, Sasaki H, Altshuler D, Li C, Meyerson M, Sellers WR (2004) Nucleic Acids Res 32:e71
14. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, Egholm M, Rothberg JM, Leamon JH (2006) BMC Genomics 7:216
15. Baum M, Bielau S, Rittner N, Schmid K, Eggelbusch K, Dahms M, Schlauersbach A, Tahedl H, Beier M, Guimil R, Scheffler M, Hermann C, Funk JM, Wixmerten A, Rebscher H, Honig M, Andreae C, Buchner D, Moschel E, Glathe A, Jager E, Thom M, Greil A, Bestvater F, Obermeier F, Burgmaier J, Thome K, Weichert S, Hein S, Binnewies T, Foitzik V, Muller M, Stahler CF, Stahler PF (2003) Nucleic Acids Res 31:e151
16. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, Shendure J (2007) Nat Methods 4:931–936
17. Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M, Bicknell D, Bodmer WF, Davis RW, Ji H (2007) Proc Natl Acad Sci USA 104:9387–9392