Published online 17 February 2010

Nucleic Acids Research, 2010, Vol. 38, No. 10 ***e116***
doi:10.1093/nar/gkq072

Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries

Michal Mokry[1], Harma Feitsma[2], Isaac J. Nijman[1], Ewart de Bruijn[1], Pieter J. van der Zaag[2], Victor Guryev[1] and Edwin Cuppen[1,3,*]

[1]Hubrecht Institute and University Medical Center Utrecht, KNAW, Uppsalalaan 8, 3584 CT Utrecht, [2]Philips Research Laboratories, High Tech Campus 12a, 5656 AE Eindhoven and [3]Department of Medical Genetics, University Medical Center Utrecht, Universiteitsweg 100, 3584 GG Utrecht, The Netherlands

Received October 10, 2009; Revised December 10, 2009; Accepted January 25, 2010

ABSTRACT

Microarray-based enrichment of selected genomic loci is a powerful method for genome complexity reduction for next-generation sequencing. Since the vast majority of exons in vertebrate genomes are smaller than 150 nt, we explored the use of short fragment libraries (85–110 bp) to achieve higher enrichment specificity by reducing carryover and adverse effects of flanking intronic sequences. High enrichment specificity (60–75%) was obtained with a relative even base coverage. Up to 98% of the target-sequence was covered more than 20× at an average coverage depth of about 200×. To verify the accuracy of SNP/mutation detection, we evaluated 384 known non-reference SNPs in the targeted regions. At ~200× average sequence coverage, we were able to survey 96.4% of 1.69 Mb of genomic sequence with only 4.2% false negative calls, mostly due to low coverage. Using the same settings, a total of 1197 novel candidate variants were detected. Verification experiments revealed only eight false positive calls, indicating an overall false positive rate of less than 1 per ~200 000 bp. Taken together, short fragment libraries provide highly efficient and flexible enrichment of exonic targets and yield relatively even base coverage, which facilitates accurate SNP and mutation detection. Raw sequencing data, alignment files and called SNPs have been submitted into GEO database http://www.ncbi.nlm.nih.gov/geo/ with accession number GSE18542.

INTRODUCTION

The need for detection of SNPs and mutations in large genomic segments is increasing rapidly, partially as a result of genome-wide association studies (GWAS) that have pinpointed many genomic loci of interest for specific diseases and disease susceptibilities (1–5). Furthermore, the vastly increased throughput of massively parallel next-generation sequencing technologies enables the interrogation of unprecedented numbers of genes in a single analysis, permitting, for instance, the investigation of the complete protein-coding transcriptome (6). Enrichment of genomic loci by microarray hybridization followed by massively parallel sequencing has become an important method for targeted re-sequencing. The approach is based on hybridization of fragmented and adapter-ligated DNA to capturing probes printed on microarray slides (7–11), present in solution (12,13) or PCR products immobilized on filters (14) that are specifically designed for the regions of interest. After hybridization, non-targeted fragments are washed away and only the captured fragments are eluted for deep sequencing on any of the next-generation sequencing platforms (7–10). Generally, DNA fragment libraries with 500-bp fragment size are recommended and used for optimal efficiency because shorter fragments were reported to increase the number of off-target reads (8). However, many re-sequencing projects are focused on exons of protein coding genes. Because the median size of a human exon is only 120 bp (with 70% of all exons shorter than 200 bp) (Figure 1), long-fragment libraries can have severe limitations. Most importantly, many of the specifically captured DNA from long-fragment libraries will consist of sequences derived from introns flanking the exons of interest, which decreases the effective

*To whom correspondence should be addressed. Tel: +31 0 30 212 18 00; Fax: +31 0 30 254 64 64; Email: e.cuppen@hubrecht.eu

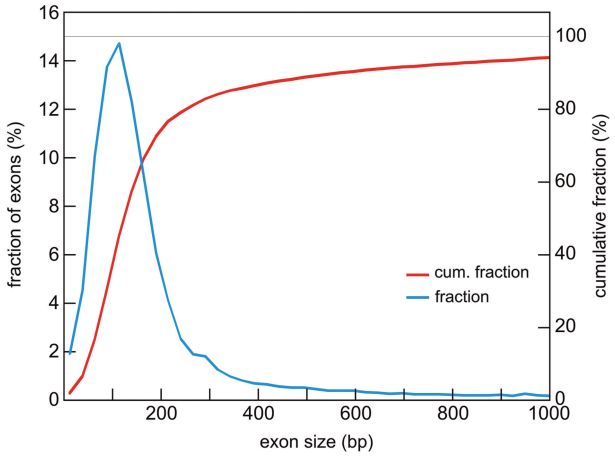

Figure 1. Size distribution of human exons. The median size of human exons is only 120 bp (with 70% of all exons shorter than 200 bp). Therefore many of the specifically captured DNA from long-fragment libraries will consist of sequences derived from introns flanking the exons of interest, which decreases the effective sequencing yield.

sequencing yield. To address this issue, we have explored the efficiency of enrichment of DNA fragment libraries with much shorter fragments (85–110 bp). We used human genomic DNA and an exon-centric capturing design on commercially available custom microarray slides. We developed an effective probe design strategy with tiled oligonucleotides for capturing both genomic strands, resulting in targeting of 1.69 Mb of exonic sequences on a 244 K array. To validate the performance of the microarray-based enrichment strategy, we have tested the specificity of enrichment and evaluated the effectiveness of retrieval of known SNPs that were present in the targeted regions. When applying highly stringent filtering criteria, we were able to evaluate 93% of the target regions without any false-negatives. Moreover, independent verification experiments on newly discovered polymorphisms revealed an overall false-positive rate of less than 1 per 200 kb. These results indicate that the use of short fragment libraries results firstly in highly efficient enrichment with relatively even base coverage over the targeted region and that secondly the sensitivity and specificity of SNP/mutation calling is very high.

MATERIALS AND METHODS

Enrichment array design

Exonic sequences of 1621 genes were selected from hg18 build of the human genome spanning a total length of targeted sequence of 1.69 Mb. Genes were selected based on the potential presence of exonic SNPs with presumed functional effects and thus potential clinical relevance (15). Sixty nucleotide long probes were designed with an average tilling density of 10 bp for both negative and positive strands. The probe selection strategy was set using the following rules: all possible 60-mer probes starting in a 10 bases long window were collected and a single probe with the lowest penalty score (see below) was selected. This procedure was repeated for every 10 nt bin in the region of interest, which presented all coding exons of selected genes. Penalty scores were calculated as follows: 4 points if T_m is $<77^\circ$C or $>81^\circ$C with T_m defined as:

$$T_m = (64.9 + 41 \times (n_G + n_C - 16.4/N))\ [\mathrm{C}] \quad (1)$$

where n_G is total number of guanidines, n_C is total number of cytosines and N represents oligonucleotides length, 2 points per homo-polymer longer than 5 bp, 1 point per each base over or below the limit (C or G fraction $<15\%$ and $>25\%$, A or T fraction $<25\%$ and $>35\%$). To exclude potentially repetitive elements from the design, all probes were compared to the reference genome using BLAST and those returning more than one hit (as defined by E-value cutoff <0.01) were discarded from the design. Probes were synthesized on custom 244 k Agilent arrays with randomized positions.

Library preparation

DNA was fragmented for 6 min using a Covaris S2 sonicator (6×16 mm AFA fiber Tube, duty cycle: 20%, intensity: 5, cycles/burst: 200, frequency sweeping). After fragmentation, fragments were blunt-ended and phosphorylated at the 5′-end using End-it Kit (Epicentre) according to the manufacturer's instructions, followed by ligation of double-stranded short adapters (adapter 1: pre-annealed duplex of 5′-CTA TGG GCA GTC GGT GAT-3′ and 5′-ATC ACC GAC TGC CCA TAG TTT-3′ and adapter 2: pre-annealed duplex of 5′-CGC CTT GGC CGT ACA GCA G-3′ and 5′-GCT GTA CGG CCA AGG CG-3′; all oligo's were acquired through Integrated DNA Technologies (Coralville, IA, USA) and pre-annealing was done by mixing complementary oligonucleotides at 500 µM concentration and running on thermocycler with the following program: 95°C for 3 min, 80°C for 3 min, 70°C for 3 min, 60°C for 3 min, 50°C for 3 min, 40°C for 3 min and 4°C hold). Ligation was performed using Quick ligation kit (New England Biolabs) with 1 µg of fragmented DNA, 750 nM adaptor 1 and adaptor 2, 150 µl of 2× Quick ligation buffer and 5 µl Quick Ligase in a total volume of 300 µl. Samples were purified on Ampure beads (Agencourt) and run on a native 6% polyacrylamide gel. Fragments ranging from 125 to 150 bp were excised; the piece of gel containing fragments was shredded and dispersed into 400 µl of Platinum PCR Supermix with 750 nM of both amplification PCR primers (provide sequence of amplification primers), 2.5 U of Pfu DNA polymerase (Stratagene) and 5 U Taq DNA polymerase (Bioline). Before ligation-mediated amplification, the PCR sample was incubated at 72°C for 20 min in PCR mix to let the DNA diffuse from the gel and to perform nick translation on non-ligated 3′-ends. After eight cycles of amplification, the library DNA was purified on Ampure beads and the quality was checked on a gel for the proper size range and the absence of adapter dimers and heterodimers. This library served as a stock for all subsequent hybridization experiments.

Enrichment hybridization and elution

Prior to hybridization, 50 ng of stock library was amplified using 10 cycles in 1000 µl of Platinum PCR Supermix with 750 nM of both amplification PCR primers to produce a sufficient amount of library DNA necessary for enrichment (Table 1). Amplified library DNA was subsequently purified using a MinElute Reaction Cleanup Kit (Qiagen). Amplified DNA was mixed with 5 × weight excess of human Cot-1 DNA (Invitrogen) and concentrated using a speedvac to a final volume of 12.3 µl. DNA was mixed with 31.7 µl Nimblegen aCGH hybridization solution and denatured at 95°C for 5 min. After denaturing, the sample was hybridized for 65 h at 42°C on a 4-bay MAUI hybridization station using an active mixing MAUI AO chamber (MAUI). After hybridization, the array was washed using the Nimblegen Wash Buffer Kit according to the user's guide for aCGH hybridization. The temperature of Wash buffer I for Library 2 was 42°C instead of room temperature. Elution was performed using 800 µl of elution buffer (10 mM Tris pH 8.0) in an Agilent Microarray Hybridization Chamber at 95°C for 30 min. After 30 min, the chamber was quickly disassembled and elution buffer collected into a separate 1.5 ml tube. Microarray slides were dipped into re-distilled water and stored for re-use. Eluted library DNA was concentrated in a speedvac to a final volume of 50 µl and amplified with a limited number of PCR cycles (12–14 cycles) with full-length primers (amp-P1: 5′-CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG AT and amp-P2: 5′-CTG CCC CGG GTT CCT CAT TCT CTN NNN NNN NNN CTG CTG TAC GGC CAA GGC G, where N represent unique barcode sequence for each library) to introduce barcode sequences as well as adapter sequences required for SOLiD sequencing.

Table 1. Enrichment statistics

	Library1	Library2	Library3	[c]Library1
Length of sequenced tags	35	35	35	50
Microarray slide	new	new	reused	new
Amount of hybridized DNA (µg)	3	3	6.5	3
Washing temperature	RT	42°C	RT	RT
Mappable tags (millions)	6.64	18.88	17.05	14.10
Uniquely mappable tags (millions)	4.97	12.70	13.98	10.78
Mappable sequence on target (%)	56	40	61	53
Uniquely mappable sequence on target (%)	67	60	75	69
Bases covered 1× (%)	99.59	99.38	99.88	99.97
Bases covered 10× (%)	92.49	93.18	98.08	99.37
Bases covered 20× (%)	83.17	86.77	95.46	98.13
Bases covered with 10% of average coverage[a] (%)	95	90	95	98
Bases covered with 25% of average coverage[a] (%)	86	77	86	92
Bases covered with 50% of average coverage[a] (%)	69	60	69	76
Evenness score E (%)[b]	70.2	62.4	70.1	74.8

[a]The percentage of bases covered with a given percentage of the average coverage is a better measurement for comparison of coverage evenness than the percentage of bases covered with a certain depth, because it is independent on overall depth of sequencing.
[b]Evenness score represents the fraction of sequenced bases that do not have to be redistributed from above-average coverage to below-average coverage positions to obtain completely even coverage for all targeted positions. This is a measurement that is relatively independent on sequencing depth (see text).
[c]The last column gives the results of an experiment in which the library resulting from a first enrichment experiment was sequenced using the Solid V3 update, which provides 50-bp read lengths.

SOLiD sequencing

To achieve clonal amplification of library fragments on the surface of sequencing beads, emulsion PCR (emPCR) was performed according to the manufacturer's instructions (Applied Biosystems). A total of 600 pg of double stranded library DNA was added to 5.6 ml of PCR mix containing 1× PCR Gold Buffer (Applied Biosystems), 3000 U AmpliTaq Gold, 20 nM emPCR primer 1, 3 µM of emPCR primer 2, 3.5 mM of each deoxynucleotide, 25 mM MgCl2 and 1.6 billion SOLiD sequencing beads (Applied Biosystems). PCR mix was added to SOLiD emPCR Tube containing 9 ml of oil phase and emulsified using ULTRA-TURRAX Tube Drive (IKA). The PCR emulsion was dispensed into 96-well plate and cycled for 60 cycles. After amplification, the emulsion was broken with butanol, beads were enriched for template-positive beads, 3′-end extended and covalently attached onto sequencing slides. Four physically separated samples were deposited on one sequencing slide and sequenced using SOLiD system version 2 to produce 35-base long reads. Library 1 has been additionally sequenced using the SOLiD version 3 system to produce 50-base long reads in a barcoded experimental setup.

Mapping of sequencing data and SNP calling

Sequencing reads were mapped against the reference genome (hg18 assembly, NCBI build 36) using the Maq package (16), which allows mapping in SOLiD color space corresponding to dinucleotide encoding of the sequenced DNA with following settings: number of maximum mismatches that can always be found −n 3, threshold on the sum of mismatching base qualities −e 150. Raw variant positions were called by the Maq package and filtered using custom scripts (available upon request). For stringent SNP calling, we used the following filtering settings: (i) positions with <20× and >5000× coverage were excluded, (ii) each of non-reference alleles had to be supported by at least three independent reads (as determined by different read start positions) separately on positive and negative strand with quality >10, (iii) the non-reference allele should account for at least 20% of the reads covering the polymorphic position, and (iv) the ratio between + and – strand reads should be between 1/9 and 9 (Table 2). Positions that passed these filtering settings were considered as SNPs. A SNP was qualified as homozygous when the fraction of non-reference alleles

Table 2. SNP calling statistics

	Library 1	Library 2	Library 3	Library 1
Length of sequenced tags	35	35	35	50
Average coverage	67×	148×	204×	213×
SNP positions validated	384	384	384	384
SNP positions filtered due to low coverage (<20×) (%)	18.8	25.8	3.4	1.8
SNPs with enough coverage filtered out for other reasons (%)[a]	26.5	21.1	27.9	6.0
SNPs identified after filtering (with at least 20× coverage) (%)	54.7	53.1	68.7	92.2
False-negative discovery rate (%)	45.3	46.9	31.3	7.8

[a]Due to low base quality, or large strand bias.

was above 95% and heterozygous when the fraction of non-reference alleles was between 20% and 95%.

Calculation of evenness score

A crucial parameter for assessing the effectiveness of any enrichment method is the evenness of coverage (12). Here, we introduce a dedicated parameter to represent the evenness of coverage score, *E*. This score intends to describe the uniformity of base coverage over targeted regions. Together with the percentage of sequenced bases on target, which determines the enrichment level, this score can be used as an objective measure to compare different enrichment experiments as the parameter *E* is quite insensitive to sequencing depth (Figure 2C). The evenness score, *E*, represents the fraction of whole-sequencing throughput that is correctly distributed. Consequently, 1-E represents the fraction of the (whole) sequencing output that still has to be redistributed from positions with coverage above average to positions with coverage below average (by better enrichment) to get the ideal even coverage over all targeted positions. The more even the coverage, the higher the evenness score: *E* will be 100% for completely uniform coverage of every base in the targeted regions and approaches 0% in case of extreme non-uniform distributions. From Figure 2B, one can appreciate that *E* is equivalent to the area under the curve. Hence, a formula for *E* can be readily arrived at by summing for all percentage positions up to the normalized coverage of 1, as in the ideal case all positions will have a coverage at least equal to the average coverage. Thus the evenness score, *E* is defined as:

$$E = \left\{ \sum_{i=1}^{C_{\text{ave}}} \frac{P_i}{C_{\text{ave}} \cdot N_{\text{TP}}} \right\} \cdot 100\ \% = \left\{ \frac{1}{C_{\text{ave}} \cdot N_{\text{TP}}} \cdot \sum_{i=1}^{C_{\text{ave}}} P_i \right\} \cdot 100\ \% \quad (2)$$

Where P_i is defined as number of targeted positions with *at least* coverage C_i, C_{ave} is defined as the average coverage through all targeted positions and N_{TP} is defined as a total number of targeted positions. This formula can be rewritten in a form which is numerically more attractive, as C_{ave} may not be an integer number:

$$E = 100\% \cdot \int_0^1 F(i)di \quad (3)$$

Where $F(i)$ is the fraction of positions with normalize coverage of at least $C(i)/C_{\text{ave}}$. This fraction equals P_i/N_{TP}, where P_i is percentage of position with a coverage of at least $C(i)$ and N_{TP} is the total number of targeted positions. C_{ave} is the average coverage over all targeted positions. This integral corresponds to the area under the curve of the graph between a normalized coverage of 0 and 1 (Figure 2B).

The relative independence of the evenness score *E* on sequencing coverage is brought about by the normalization to the average coverage. Hence, *E* solely reflects the quality of targeted genome selection.

Illumina SNP genotyping

The DNA sample that was used in this study was genotyped using an Illumina HumanHap550+ Genotyping BeadChip through 23andMe services (http://www.23andme.com). A total of 384 genotyped SNPs, which are either heterozygous or homozygous non-reference in the sample, are located in the 1.69 Mb region of our interest. These positions were used as a reference set for identifying false negatives in our sequencing dataset.

RESULTS AND DISCUSSION

Specificity of enrichment

Short fragment (85–110 nt) libraries were made using focused acoustic fragmentation (Covaris) and used for enrichment on custom-designed 244 K Agilent microarrays. The specificity of the enrichment was determined by sequencing on an ABI SOLiD sequencer version 2 with a quadrant slide capacity per sample (loaded at various densities). This resulted in 219–676 millions of mappable bases out of which 40–61% mapped directly to the targeted regions (Table 1). Increasing the stringency by raising the washing temperature from room temperature to 42°C (Library 2) did not increase the enrichment efficiency (Table 1). In contrast, the evenness of coverage E did decrease appreciably from 70% to 62 % (Table 1 and Figure 2A), suggesting selective loss of specific target regions. Increasing the amount of DNA for hybridization (Library 3) had no effect at all.

The percentage of on-target bases increases to 60–75% when only taking into account bases from reads that could be placed completely uniquely on the genome. Our results contrast with previously reported results (8), where the use of 100–200 bp fragments resulted in only 29% percent of reads mapping to targeted exons. The observed difference could possibly be explained by differences in probe design strategy and/or array platform (Nimblegen versus

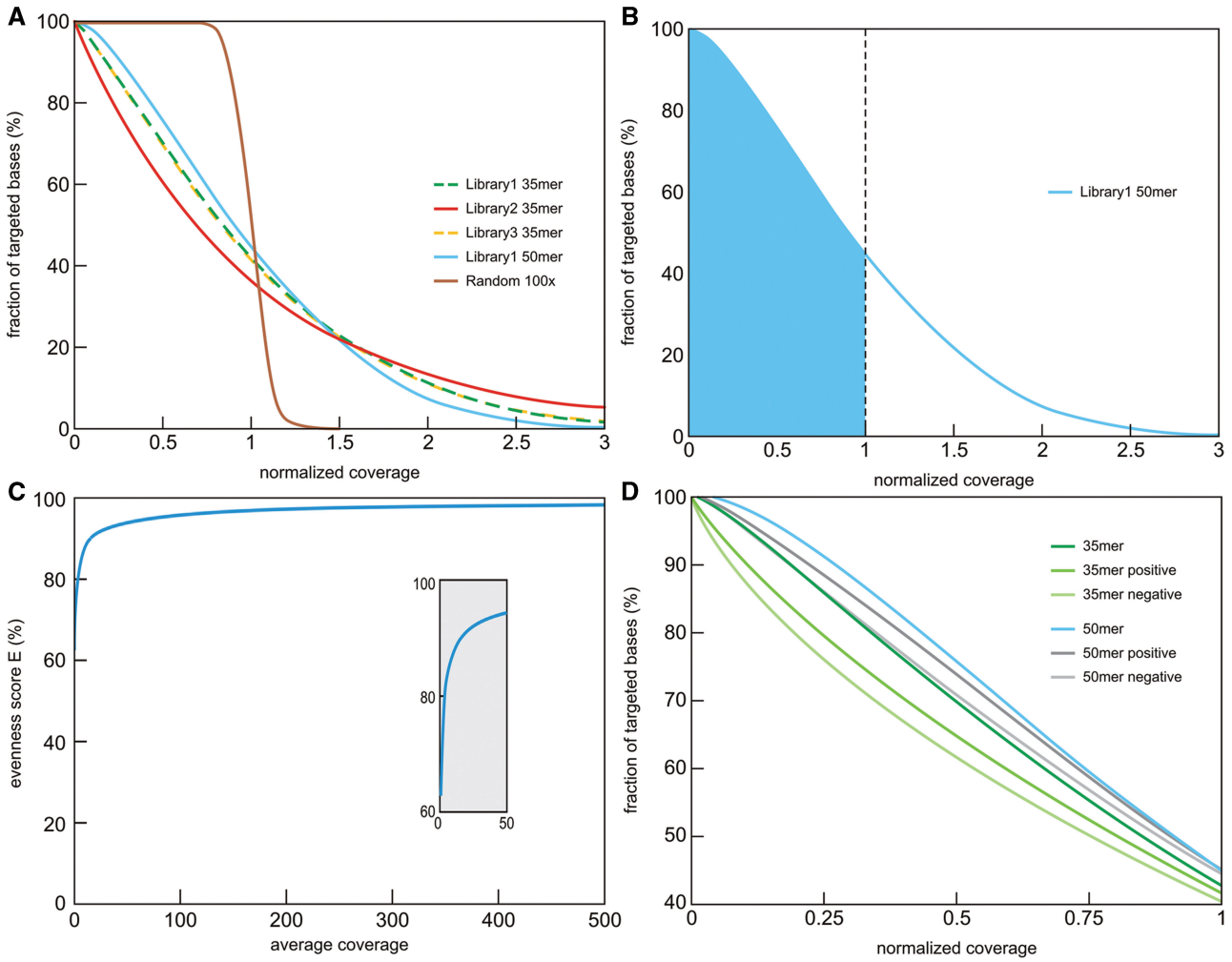

Figure 2. Comparison of sequence coverage evenness after enrichment. The fraction of target positions with at least that coverage was as compared to the average coverage. (**A**) Comparison of various enrichment (washing temperature and input DNA) and sequencing (35- versus 50-mer) conditions. Library 1 sequenced by 50-mer reads results in the most even coverage compared to other libraries. The brown curve depicts the best possible evenness for an ideal evenly enriched sample with 100 × average coverage, where the unevenness is purely caused by statistical randomness in the coverage assuming a Poisson distribution of the sequencing reads. (**B**) The evenness score, *E*, represents the fraction of whole sequencing throughput that is correctly distributed (marked area below the curve). Consequently, 1-E represents the fraction of the (whole) sequencing output that has to be redistributed from positions with coverage above average to positions with coverage below average (by better enrichment) to get the ideal even coverage over all targeted positions. The more even the coverage, the higher the evenness score. (**C**) Correlation of evenness score E for randomized sets to the sequencing depth. In this simulation, the unevenness of these datasets is purely caused by the random distribution of reads and fits a Poisson distribution of sequence coverage. When the discrete character of the data is reduced by sufficient depth of coverage, E changes only slightly with increasing average coverage and thus can be characterized as relatively independent of sequencing depth. (**D**) Comparison between + and – strand coverage for 35- and 50-mer reads. In the case of 50-mer reads, the coverage is more even with fewer positions covered by extremely low (or high) numbers of sequencing tags. This difference is more prominent when the coverage is determined separately for the positive or negative strand. Independent strand coverage is better for 50-mer than for 35-mer sequencing.

Agilent). A different approach for enrichment using 170-nt long biotinylated RNA probes (12), resulted in 42–50% of sequenced bases mapping directly to targeted exons. An alternative solution-based approach used molecular inversion probes (MIPs) (13) for selective capturing of 55 000 human exons and resulted in >99% of all reads mapping to the targeted regions. However, since the first 20 bp of each sequenced read are always coming from the MIPs only the remainder of the sequence is informative for genotyping and a substantial proportion of the overall sequencing output thus has to be discarded as non-informative. Moreover, exons longer than the maximum read length of the sequencing platform will need multiple probe designs for capturing and sequencing. As no commercial solutions are available for cost-effective synthesis of high-quality, long oligonucleotides (>100 nt), which are required for this approach, widespread implementation of this method is questionable.

Evenness of coverage

During enrichment not all DNA hybridizes to capture probes with the same efficiency. As a result, targeted sequences are covered unevenly with sequencing reads. For a limited number of regions there even is no coverage at all. Generally, the more even the coverage distribution is,

the less overall sequencing depth is required for variant detection. In our experiments 60–76% of targeted regions are covered with >50% of average coverage, 77–92% of targeted regions are covered with >25% of average coverage and 90–98% of reads have >10% of average coverage (Table 1 and Figure 2A). In addition, we covered 99.38–99.97% of targeted positions with at least one read suggesting significantly better evenness of coverage compared to other studies, where 82% (10); 95% (12) or 99.21% (17) of targeted positions were covered with at least one read.

For better comparison of coverage evenness, we introduce a new parameter, the evenness score E (see section 'Materials and Methods' section and Figure 2B). E is relatively independent of sequence coverage and thus enables the comparison of different libraries with varying sequencing depth. The relative independence of the evenness score E to sequencing depth is shown in Figure 2C. In this figure, the correlation of E for a randomized read sets assuming a Poisson read distribution, is calculated as a function of the sequencing depth. Figure 2C shows that once the sequencing coverage is sufficiently high (>50×), the evenness score E changes only slightly with increasing average coverage and thus can be characterized as relatively independent of the sequencing depth. The evenness score represents the area under the curve in a fraction of the positions with at least that coverage vs normalized coverage at $X = 1$ (see in Figure 2B). In other words, E represents the fraction of sequenced bases that do not have to be redistributed from above-average coverage to below-average coverage positions to obtain completely even coverage for all targeted positions. The evenness score, $E = 100\%$ for perfectly uniform coverage and approaches 0% in cases of extreme non-uniform distributions. By using short-fragment libraries, we achieved even distributions with evenness scores (E) ranging from 62.4% (Library 2) to 74.8% (Library 1 sequenced with 50-nt read length). Sequencing of the same enriched library on SOLiD V3 with 50-nt long reads instead of 35-nt long reads with SOLiD V2 resulted in a better evenness score with E, rising from 70.2% to 74.8% (Table 1 and Figure 2A). The most obvious explanation for this observation would be that the longer reads resulted in improved bridging of regions with lower capture efficiency due to fragments captured by well-performing flanking probes. In addition, due to low complexity of genomic regions, 35-mer tags cannot be mapped uniquely to a substantial part of the genome and this fraction is reduced with 50-mers.

To illustrate that the described approach universally results in high evenness, we analyzed additional experiments that were performed with the same experimental protocol, but with different array designs and/or species (Supplementary Table S1 and Supplementary Figure S1). We do find a consistent high evenness score, even between organisms (human, rat and *Arabidopsis*). Moreover, we reanalyzed publically available datasets from recently published genomic enrichment experiments (6,11,18) showing that our experiments result consistently in more even coverage, especially when considering strand-specific coverage (Supplementary Table S1 and Supplementary Figure S1). The latter is especially important, as we observed in our dataset that proper coverage on both strands is instrumental for reducing false positive heterozygote SNP calling. This is most likely due to systematic errors that are introduced by the sequencing process due to platform-specific biases in sequencing chemistry, which is in all cases context dependent and thus different for the + and – strand.

Sequencing of positive and negative DNA strand

We found that having sequencing data mapping to both positive and negative strand is an important factor in reducing false positive variant calls. Therefore, we evaluated the evenness of coverage with respect to DNA strand. A substantial part of the targeted sequence was covered by sequencing tags coming from only one strand in case of libraries sequenced as 35-mers with SOLiD V2 chemistry (Figures 2D and 3). Increasing the read length to 50-mers improved double-stranded coverage markedly, in line with the observed overall base coverage (Figures 2D and 3).

Improvements of sequencing coverage

Since the contribution of the random character of sequencing to unevenness of coverage was minor (Figure 2C), further improvements in the evenness of the sequencing coverage could be obtained by improvement of the enrichment procedure. The strategy used in our array design resulted in overlapping probes and did not provide much opportunity for redesigning probes for poorly enriched regions. Therefore, we included various probes at variable quantities in our test design. We found a very strong correlation between the number of probes and eventual base coverage (Figure 4). Spotting more copies of the same probe for underperforming regions therefore seems an effective strategy for improving E. Furthermore, these results also indicate that a limiting factor for DNA yield after elution could be the number of probe molecules and their saturation after 65 h of hybridization and not the depletion of targeted library molecules. This is supported by the observation that increasing the amount of library DNA during hybridization had no effect on enrichment efficiency and coverage. However, we cannot exclude that the efficiency of mixing during hybridization is limited and that local depletion of target sequences occurred, without saturating the capturing probes. Presence of capture probes at physically separated locations (which is the case in the random design used in these experiments) would in such case also improve capturing efficiency.

Yet another possibility for improvement of sequence coverage can be expected from further increasing sequencing read length, in line with our results obtained for 35- versus 50-mer reads.

Identification of polymorphic positions

To determine the accuracy of SNP detection, we compared SNPs called from sequencing data obtained from enriched short-fragment libraries and those genotyped by an Illumina HumanHap550+ Genotyping

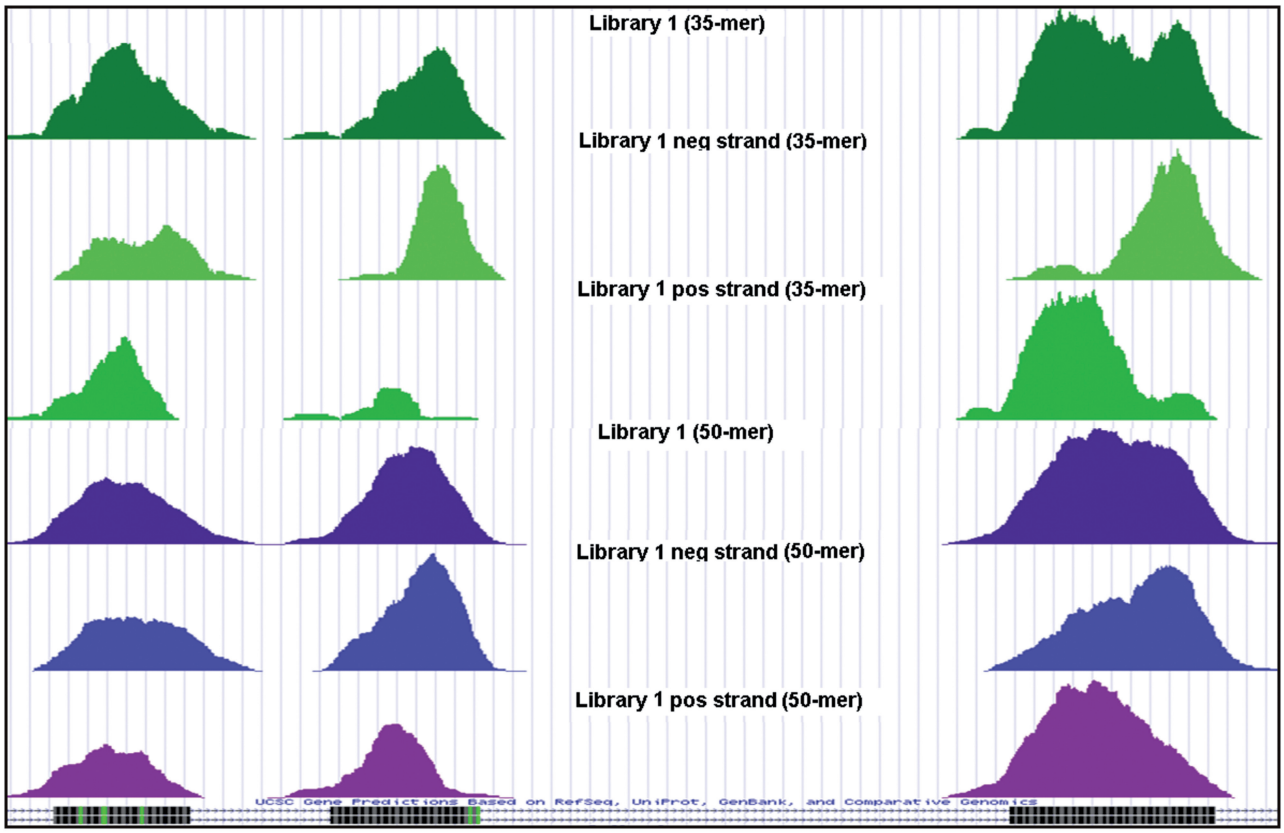

Figure 3. Exemplary representation of target coverage after enrichment. Sequencing results of Library 1 are shown for 35-mer (green) and 50-mer (purple) sequencing. Total, positive and negative strand reads are shown independently. Coverage is more equal and better represented by both strands for the longer sequencing reads.

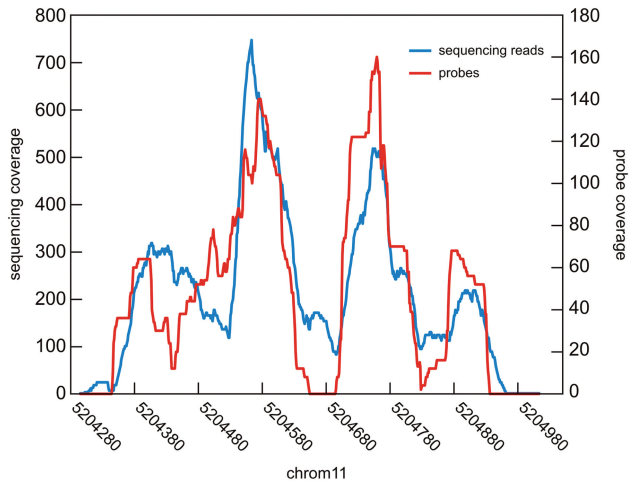

Figure 4. Correlation of probe density and sequencing coverage. Each genomic region was represented on the array with a variable number of capture probes throughout the region. The sequencing coverage per base (blue line) linearly correlates with probe density (red line).

BeadChip. A total of 384 SNP genotyped positions were different from the reference allele (either heterozygous or homozygous) and were located in the targeted regions. From those 384 SNP positions, 53.1– 92.2% passed the stringent criteria of our SNP filtering pipeline, which required sufficient high-quality base coverage on both DNA strands. In concordance with evenness of coverage, Library 1 sequenced on SOLiD V3 with 50-nt long reads gave the best results with only a 7.8% false negative discovery rate over the complete targeted region. From this total of 7.8% false-negative SNP positions, 3.6% had not been covered with (i) at least 20 reads, and (ii) at least 3 reads from each of the strands or (iii) did not have a coverage ratio from both strands within the limits set at 1/9 and 9. Consequently, this part of the targeted regions could already be marked as not surveyed, even prior to SNP calling, since this part would not pass our minimal requirements of SNP calling. Such a prediction will be important for clinical diagnostic purposes because this enables one to predict which regions have not been sequenced sufficiently deep for reliable SNP calling. Taken together, at ~200× average sequence coverage, we were able to survey 96.4% of 1.69 Mb of genomic sequence with only 4.2% of false negative calls, while 3.6% of targeted regions had to be marked as unsurveyed.

The better performance of Library 1 could be explained by better evenness of coverage (more positions have sufficient base coverage for reliable SNP calling) as well as by better strand balance where more positions have good coverage coming from tags mapping to both negative and positive strands (Figures 2D and 3). Another explanation for the better performance of longer reads could be

a reduction of mappability bias. While mapping sequencing reads, non-reference alleles are more likely to be discarded due to low mapping quality, since they already have one mismatch (two mismatches in SOLiD color space) compared to reads coming from reference alleles. Additionally, capture of non-reference DNA molecules to reference capture probes may be slightly less effective. Indeed, the overall frequency of non-reference allele reads for heterozygous positions was shifted downward form the 50% position, although not dramatically (Figure 5). The mappability issue is even more serious for SNPs that have additional linked SNPs in close vicinity. This bias is less prone for longer reads since one mismatch contributes less to overall mapping accuracy in 50-mers compared to 35-mers.

Detection of novel variants

We analyzed the results of Library 1 (50-mer reads) for the presence of polymorphisms. A total of 1197 SNPs were identified within the targeted regions plus 30-bp flanking intronic regions using our SNP detection pipeline. This set included the 384 previously genotyped SNPs as well as 759 other polymorphisms that were already present in dbSNP129 or the Ensembl database. We considered these 1143 (95.5%) SNPs as validated and set out to validate the remaining 54 SNPs by PCR-based dideoxy resequencing. We failed to develop working assays for 10 candidate SNPs, most likely due to the repetitive nature of the genomic environment. We found that only eight of the remaining candidates, all heterozygote scores, were identified as false positives. Altogether, our results indicate a false positive rate of less than one per ~200 000 bp (0.0005%). All eight false positive SNPs tended to have a lower than average percentage of non-reference allele reads and/or low overall base coverage compared to true positives (Figure 6). Although we only used planar microarrays in our experiments, we believe that the characteristics of short-fragment libraries described here will be equally applicable to any hybridization-based approach, including in-solution methods.

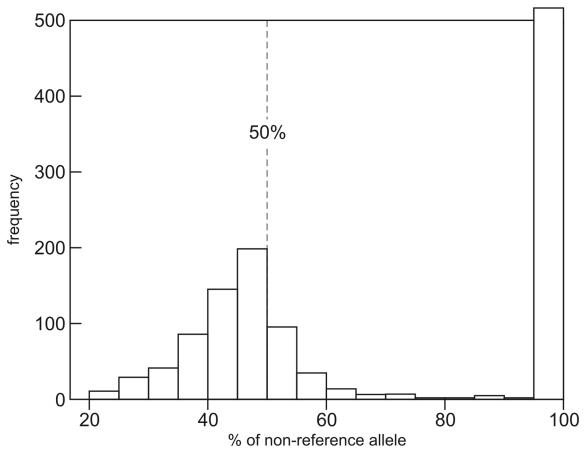

Figure 5. Distribution of non-reference allele reads. The percentage of non-reference allele reads was calculated for every heterozygous and homozygous non-reference allele position in the targeted region ($n = 1197$) and is represented in bins of 5%. For heterozygous calls, the distribution is skewed towards reference allele reads.

Input DNA requirements

Relatively high amounts of DNA are normally used for enrichment procedures, which can be a limiting factor for many clinical applications. In our experiments, we used only 1 µg of genomic DNA for all experiments shown. The fragment library was made and amplified with only eight PCR cycles to produce a stock library, which was sufficient for at least 20 independent enrichment procedures as described here. Before enrichment a small proportion (50 ng) of the initial library was amplified with 10 PCR cycles to produce sufficient material for hybridization. After enrichment the eluted library was amplified by an additional 13 cycles to produce amounts sufficient for accurate quantification and sequencing. By using these logistics, the stock library could be used multiple times for different enrichment experiments, taking away the need for re-isolating DNA or re-preparing sequencing libraries. In addition, most of the amplification cycles (18 out of 31 in total) were done before hybridization. Potential biases in coverage caused by PCR could, in theory, be normalized again during the hybridization step. However, this requires that capturing probes, rather than target molecules, are the limiting factor in this step.

Detailed analysis of our deep-sequencing results revealed no unexpected clonality bias due to the amplification steps.

CONCLUSIONS

Short-fragment libraries provide highly efficient enrichment characteristics for exonic targets. The relative even base and strand coverage facilitates accurate SNP and mutation retrieval and discovery. To measure the

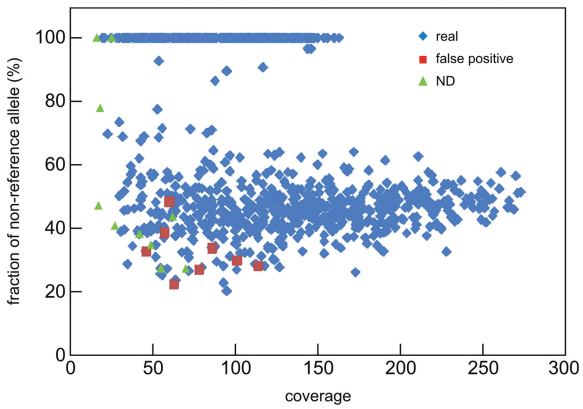

Figure 6. Sequencing coverage and percentage of non-reference allele distribution for validated and non-validated SNPs. All polymorphic and non-reference positions that were identified by the SNP detection pipeline are plotted as a function of total base coverage versus non-reference read frequency. Validated SNPs (either by their presence in dbSNP or by resequencing) are indicated in blue, non-validated SNPs are shown in red and positions for which no working validation assay could be designed in green. False-positive SNPs tend to have a lower percentage of non-reference allele reads and/or low overall coverage.

evenness of coverage, which is relatively independent of sequencing depth, we have introduced the parameter *E* [see Equations (2 and 3) and Figure 2]. This score can be applied to any genomic enrichment experiment and in combination with the percentage of reads on target it provides the possibility to compare the efficiency of different approaches.

ACCESSION NUMBER

Raw sequencing data, alignment files and called SNPs have been submitted into GEO database http://www.ncbi.nlm.nih.gov/geo/ with accession number GSE18542.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Wim Verhaegh, Anja van de Stolpe and Dennis Merkle for their useful comments and help during manuscript preparation.

FUNDING

Funding for open access charge: Hubrecht Institutes will pay for page charges from its primary funding from the Royal Dutch Academy of Sciences (KNAW).

Conflict of interest statement. None declared.

REFERENCES

1. Zheng,W., Long,J., Gao,Y.T., Li,C., Zheng,Y., Xiang,Y.B., Wen,W., Levy,S., Deming,S.L., Haines,J.L. *et al.* (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.*, **41**, 324–328.
2. Song,H., Ramus,S.J., Tyrer,J., Bolton,K.L., Gentry-Maharaj,A., Wozniak,E., Anton-Culver,H., Chang-Claude,J., Cramer,D.W., DiCioccio,R. *et al.* (2009) A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat. Genet.*, **41**, 996–1000.
3. Papaemmanuil,E., Hosking,F.J., Vijayakrishnan,J., Price,A., Olver,B., Sheridan,E., Kinsey,S.E., Lightfoot,T., Roman,E., Irving,J.A. *et al.* (2009) Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat. Genet.*, **41**, 1006–1010.
4. Kathiresan,S., Voight,B.F., Purcell,S., Musunuru,K., Ardissino,D., Mannucci,P.M., Anand,S., Engert,J.C., Samani,N.J., Schunkert,H. *et al.* (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.*, **41**, 334–341.
5. Ahmed,S., Thomas,G., Ghoussaini,M., Healey,C.S., Humphreys,M.K., Platte,R., Morrison,J., Maranian,M., Pooley,K.A., Luben,R. *et al.* (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.*, **41**, 585–590.
6. Ng,S.B., Turner,E.H., Robertson,P.D., Flygare,S.D., Bigham,A.W., Lee,C., Shaffer,T., Wong,M., Bhattacharjee,A., Eichler,E.E. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
7. 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.
8. 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.
9. Okou,D.T., Steinberg,K.M., Middle,C., Cutler,D.J., Albert,T.J. and Zwick,M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, **4**, 907–909.
10. Bau,S., Schracke,N., Kranzle,M., Wu,H., Stahler,P.F., Hoheisel,J.D., Beier,M. and Summerer,D. (2009) Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. *Anal. Bioanal. Chem.*, **393**, 171–175.
11. Summerer,D., Wu,H., Haase,B., Cheng,Y., Schracke,N., Stahler,C.F., Chee,M.S., Stahler,P.F. and Beier,M. (2009) Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Res.*, **19**, 1616–1621.
12. 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.
13. 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.
14. Herman,D.S., Hovingh,G.K., Iartchouk,O., Rehm,H.L., Kucherlapati,R., Seidman,J.G. and Seidman,C.E. (2009) Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat. Methods*, **6**, 507–510.
15. Cariaso,M. and Lennon,G. SNPedia (or Promethease). http://www.SNPedia.com/ (Accessed January 2009).
16. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
17. Hodges,E., Rooks,M., Xuan,Z., Bhattacharjee,A., Benjamin Gordon,D., Brizuela,L., Richard McCombie,W. and Hannon,G.J. (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protocols*, **4**, 960–974.
18. Okou,D.T., Locke,A.E., Steinberg,K.M., Hagen,K., Athri,P., Shetty,A.C., Patel,V. and Zwick,M.E. (2009) Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. *Ann. Hum. Genet.*, **73**, 502–513.