Reports

Creation and application of immortalized bait libraries for targeted enrichment and next-generation sequencing

Robert Querfurth, Axel Fischer, Michal R. Schweiger, Hans Lehrach, and Florian Mertes
Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

BioTechniques 52:375-380 (June 2012) doi 10.2144/0000113877
Keywords: next-generation sequencing; targeted enrichment; emulsion PCR; cancer sequencing; SW480; SNP; variation detection; bait library; in-solution capture

Supplementary material for this article is available at www.BioTechniques.com/article/113877

Since the introduction of next-generation sequencing, several techniques have been developed to selectively enrich and sequence specific parts of the genome at high coverage. These techniques include enzymatic methods employing molecular inversion probes, PCR based approaches, hybrid capture, and in-solution capture. In-solution capture employs RNA probes transcribed from a pool of DNA template oligos designed to match regions of interest to specifically bind and enrich genomic DNA fragments. This method is highly efficient, especially if genomic target regions are large in size or quantity. Diverse in-solution capture kits are available, but are costly when large sample numbers need to be analyzed. Here we present a cost-effective strategy for the design of custom DNA libraries, their transcription into RNA libraries, and application for in-solution capture. We show the efficacy by comparing the method to a commercial kit and further demonstrate that emulsion PCR can be used for bias free amplification and virtual immortalization of DNA template libraries.

The introduction of next generation sequencing (NGS) has revolutionized research in many areas (1,2), especially affecting our fundamental understanding of the genome and its subparts (3). A multitude of protocols exist for specialized nucleic acid preparations for NGS, DNA sequencing, RNA sequencing (4) and ChIP-seq (5).

Despite advances in technology, whole genome sequencing for large genomes is still associated with tremendous cost and workload. If the research is focused only against a subset of the whole genome, genome partitioning methods may be used to selectively enrich for the region of interest (6). Targeted enrichment is employed in many areas of genetic research like whole exome sequencing (7), sequencing of causal disease genes (8), and extensive resequencing for large cohorts (9).

There are various approaches for targeted enrichment available. Most commonly used techniques are based on hybrid capture, PCR, and molecular inversion probes (10). For large target regions, hybrid capture has turned out to be the most efficient. A main advantage of this approach is enrichment in-solution (11) rather than on microarrays (12); this provides easier handling and requires less DNA. In-solution capture often applies biotinylated RNA bait molecules transcribed from DNA template oligo libraries, which are the key component and main cost.

In this study, our goal was to reduce enrichment costs by omitting repeated synthesis of DNA template libraries for recurrent generation of RNA baits. This will be most beneficial for projects that require large target regions and large sample numbers to be enriched for sequencing. We set up a simple strategy for the design of DNA template libraries with two main characteristics: unique target sequences that are relatively short (40bp) and tiled along both strands of the target region in an alternating manner; and universal primers that flank each of the baits for library amplification.

To illustrate this method, we designed a DNA template library for 966 cancer associated genes. The library was ordered from MYcroarray, transcribed into RNA baits, and used for enrichment. The efficacy of our approach was demonstrated by comparison with the SureSelect Kit from Agilent. Both systems were used to enrich cancer associated genes of the cell line SW480 (13), an important model in colorectal cancer research.

Our next step was to amplify the DNA template library by water-in-oil emulsion PCR to prevent the introduction of amplification biases (14,15). Using this approach, we show evidence of bias free amplification and virtual immortalization of a DNA template library. For recurrent analyses in cases such as diagnostics, enrichment costs may be reduced significantly by using a short immortalized DNA template library (SIMLY), which is described here.

Materials and methods

Bait library design

For targeted enrichment, we compiled a list of cancer associated genes from different sources. The list included 408 genes from the Cancer Gene Census catalog (CGC) and 383 genes from the COSMIC database (v47). The Cosmic database lists 120,000 published mutations within 4200 genes of 100,000 different human cancer samples. We selected all genes that were reported as mutant more than three times and were not already covered by the CGC data set. Another 175 genes were selected because they were known to be associated with cancer (e.g., BBC3, PSEN1), to be involved in cancer relevant pathways (e.g., IRAK1), or have been identified in cancer genome sequencing projects running at the Max Planck Institute for Molecular Genetics (e.g., RAPGEF1). In total 966 genes were targeted.

In addition, 481 ultra-conserved regions were targeted, since highly conserved regions are likely to be functionally relevant in processes

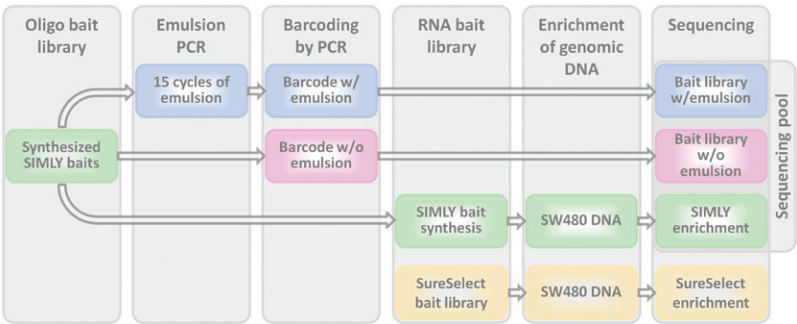

Figure 1. Flowchart summarizing the experimental design for bait library sequencing and targeted enrichment of genomic DNA by SIMLY and SureSelect bait libraries.

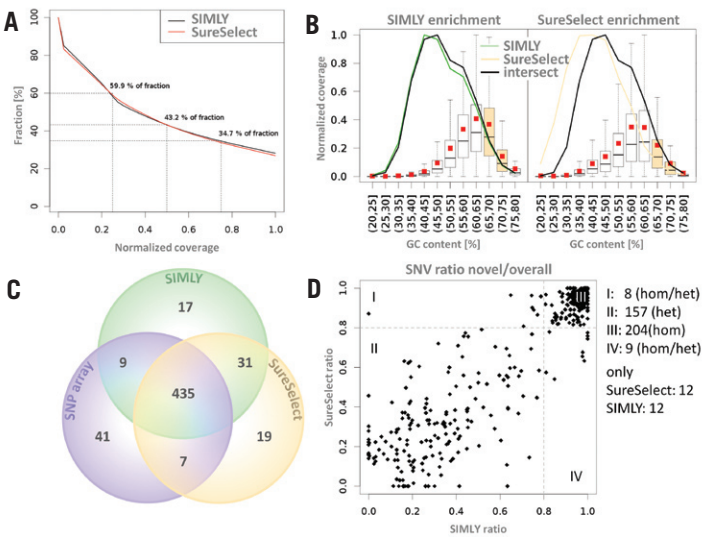

Figure 2. Assessment of SIMLY enrichment and SureSelect whole exome enrichment in terms of normalized coverage, SNP detection, and GC content. For the comparison, only the overlapping data set of 15,517 loci was used. (A) Comparison of SIMLY and SureSelect enrichment. Normalized coverage was plotted against a fraction of the target region. Dashed lines indicate the percent fraction of bases that gave rise to 0.75, 0.5, and 0.25 normalized coverage, respectively. (B) Boxplots showing normalized coverage for the intersecting region of SIMLY and SureSelect targeted enrichment according to GC content. Boxplots of greatest coverage divergence are highlighted. Solid squares indicate mean values for each data point. GC content of the bait libraries as well as the intersection region are given as solid lines. (C) Venn diagram of single nucleotide polymorphisms identified in the overlapping fraction of SIMLY, SureSelect and Affymetrix Human SNP Array 6.0. (D) Single nucleotide variant analysis by normalized ratio of wild type to novel allele for the intersecting region of SIMLY and SureSelect. The heterozygote to homozygote cut-off ≥ 0.8 is indicated by dashed lines, subdividing the plot area into heterozygote (II), homozygote (III), and homozygote-heterozygote (I and IV) SNVs.

such as long-range transcriptional regulation (16). The corresponding probes made up about 3% of the total number of probes.

For gene identification, we used Entrez Gene IDs, while probe design was based on UCSC gene annotation. Entrez Gene IDs were mapped to UCSC gene annotations using the UCSC mapping table "knownToLocusLink" based on hg19. All UCSC exons and ultra conserved regions were intersected to obtain a non-redundant and non-overlapping set of target regions. Probes of 40 bp were tiled along the full length of each target region with gaps of 10 bp in between. Probes were designed to target the plus and minus strand in an alternating fashion, enabling binding of both strands of the prepared fragment library. For short target regions, we designed more probes to improve efficiency, with 3 probes for every ≤50 bp target region and 4 probes for 50–200 bp target regions. To minimize enrichment of repeating elements, repeat masking was performed and probes with repeating elements were excluded. Probe design for the target region resulted in 89,909 oligos, of which 87,119 oligos targeted coding exons of proven and presumed cancer genes while 2790 probes targeted ultra-conserved regions. The total size of the target region was 3.7 Mb with 18,713 targeted loci.

The individual oligos were comprised of a 40 bp unique target sequence plus a universal T7 promoter 5′ sequence (5′-TAATACGACTCACTATAGGG-3′) and a universal 3′ sequence (5′- GCACTGCAAAAAGCAGGCTC-3′), with a total length of 80 bp. The universal sequences allow library amplification and transcription into biotin labeled RNA baits. The DNA template library was ordered as custom synthesis from MYcroarray (Ann Arbor, MI, USA) and resuspended to 50 ng/µL after delivery.

Amplification and tailing of bait library by emulsion PCR

Amplification and PCR tailing of the template library was performed by standard PCR and water-in-oil emulsion PCR with Phusion Taq (NEB/Finnzymes, Frankfurt, Germany) containing 10 ng template library in 50 µL PCR reactions (10 µL 5x HF buffer, 0.5 mM dNTPs, 1 mM each forward and reverse primer, 0.5 µL Phusion Taq; 95°C for 1 min, [98°C for 5 s, 55°C for 10 s, 72°C for 20 s]x15 cycles, 72°C for 2 min). Library amplification was performed with the universal primers (T7-for: 5′- TAATACGACTCACTATAGGG-3′ and uni-reverse: 5′-GCACTGCAAAAAGCAGGCTC-3′; all oligos were synthesized by Metabion, Munich, Germany) in emulsion as described (15). In brief, 1x PCR master mix was emulsified with 6x oil mix. After 15 PCR cycles, products were cleaned by emulsion breaking and column purification. Tailing of the library with barcoded P1 primers (P1-tag1: 5′-CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATCTCTAATACGACTCACTATAGGG-3′ and P1-tag2: 5′-CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGATGAGTAATACGACTCACTATAGGG-3′) and P2 primer (5′-CTGCCCCGGGTTCCTCATTCTGCACTGCAAAAAGCAGGCTC-3′) was performed for 5 cycles of PCR, enabling SOLiD sequencing.

Target library preparation and enrichment

DNA was prepared from SW480 cells and fragment libraries were prepared according to the SOLiD fragment library protocol with truncated P1 (P1-A: 5′-TCTATGGGCAGTCGGTGAT-3′ and P1-B 5′-ATCACCGACTGCCCATAGATT-3′) and P2 (P2-A: 5′-CCGGGTTCCTCATTCTCT-3′ and P2-B: 5′-AGAGAATGAGGAACCCGGTT-3′) adaptors. Prior to enrichment, library amplification was performed with the primer pair P1-A/P2-A in 100 µL (20 µL 5x HF buffer, 0.5 mM dNTPs, 1 mM each forward and reverse primer, 0.5 µL Phusion Taq; 95°C for 1 min, [98°C for 5 s,

Table 1. Target region size and sequencing results for SIMLY and SureSelect.

	SIMLY	SureSelect
Target region size	3.7 Mb	35.0 Mb
Number of target loci	18,713	165,144
Number of bait oligos	89,909	314,276
UMO	35.6 Mio reads 1.78 Gb	33.0 Mio reads 1.65 Gb
UMO on target	19.4 Mio reads 0.97 Gb	31.7 Mio reads 1.59 Gb
UMO on target (ext. ±100)	24.2 Mio reads 1.21 Gb	31.9 Mio reads 1.60 Gb
Mean coverage	280x	43x
	Overlapping dataset	
Target region size	2.7 Mb	
Number of target loci	15,517	
Number of SNPs (SNP Array)	1371	

Comparison of SIMLY and SureSelect enrichment in terms of target region size, number of baits, and targeted loci, as well as sequencing results expressed as unique mappable output (UMO). Additional target region parameters are given for the overlapping data set of SIMLY and SureSelect, along with the Affymetrix Human SNP Array 6.0.

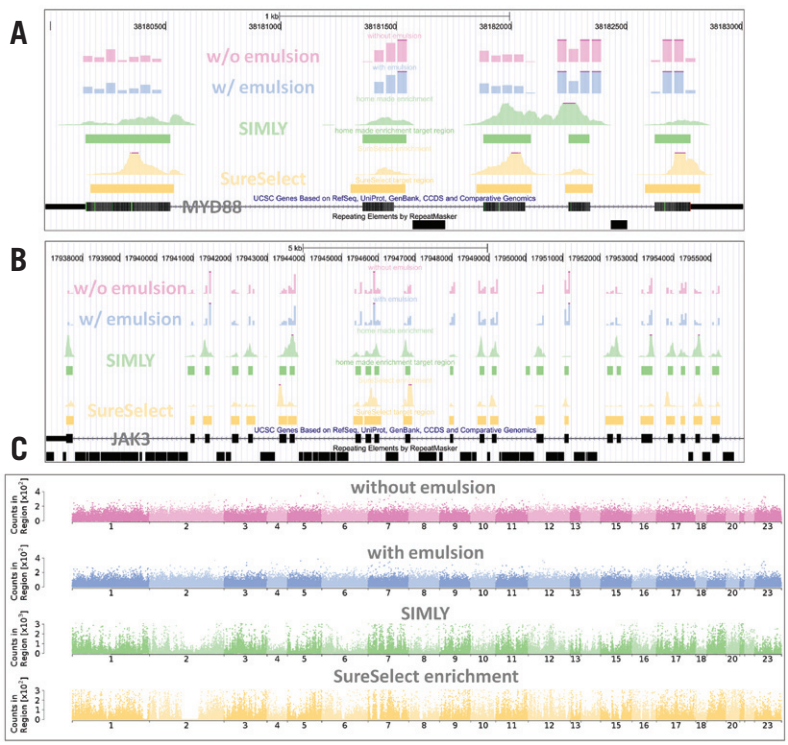

Figure 3. Sequence coverage plots for two genes, MYD88 (A) and JAK3 (B) and the complete targeted region condensed over all chromosomes (C). Coverage is shown for both sequenced bait libraries with (w/) and without (w/o) emulsion amplification and the genomic DNA enriched by SIMLY and SureSelect. Gene tracks are shown in wiggle format on the UCSC genome browser with cut-off values set to 2/3 the maximum value for the respective intervals shown. (A) and (B) Horizontal rows from top to bottom show coverage of the sequenced template library without emulsion amplification; coverage of sequenced template library with emulsion amplification; coverage of SIMLY enrichment; coverage of SureSelect enrichment; refseq gene annotation; repetitive elements. (C) Line plot of the complete 3.7 Mb target region over all chromosomes with counts per target region. The top two tracks show the sequenced bait libraries (with and without emulsion amplification); the bottom two tracks show the sequencing counts for the genomic DNA samples enriched by SIMLY and SureSelect, respectively.

52°C for 10 s, 72°C for 20 s] x 8 cycles, 72°C for 2 min). Size selection was performed after PCR in a size range of 150–200 bp by agarose gel purification. DNA template libraries were in vitro transcribed into biotin labeled RNA bait library probes with the Ambion MEGAscript T7 Kit (Invitrogen, Darmstadt, Germany) according to the manufacturer by replacing 20% of dUTP with biotin labeled dUTP (Biotin-16-dUTP, Roche, Mannheim, Germany). In a single reaction, 500 ng were transcribed for 90 min at 37°C with subsequent DNase (NEB, Frankfurt, Germany) digest and RNeasy (Qiagen, Hilden, Germany) column cleanup. Hybrid capture was performed with equal amounts of fragment library and RNA bait library (each 250 ng) and corresponding blocking oligos (P1-A and P2-A) in 26 µL at 65°C over night according to the protocol described (17). After capture of the enriched fragment library by streptavidin beads (Dynabeads M-280, Invitrogen, Darmstadt, Germany) and purification, the enriched fraction was amplified as described above with full length SOLiD P1 (5′-CCACTACGCCTCCGC TTTCCTCTCTATGGGCAGTCGGT GAT-3′) and P2 (5′-CTGCCCCGGGTT-CCTCATTCTCT-3′) primers for 14 cycles, purified, and quantified by real time PCR for later SOLiD sequencing.

Sequencing of enriched sample and bait library

Sequencing of the enriched sample and barcoded template libraries was performed according to the SOLiD V4 protocol (Applied Biosystems, Darmstadt, Germany). Briefly, 10 million beads for each of the two barcoded template libraries and 100 million beads for the enriched fractions of genomic fragment library were prepared. The beads were combined and sequenced on a single quad of a flowcell with a 50 bp SOLiD 4 fragment run.

Read mapping and SNP calling

Data analysis was performed with the Applied Biosystems Bioscope v1.3.1 package (Applied Biosystems, Darmstadt, Germany) and a custom barcode deconvolution. To map the bait library, all 50mer reads were aligned to the probe sequences, including the T7 sequence and barcode (CTC/GAG), using the Bioscope Alignment module in classic mode and allowing for 5 mismatches. When the enrichment results were mapped, all 50mer reads mapped to hg19. The Bioscope Alignment module was used in seed and extend mode, using the first 25 bp of the reads as seeds for the first round and 25 bp starting at the 15th base in the second round, allowing 2 mismatches in both rounds and a mismatch penalty score of -2 for extension. The attached T7-tag included in the two bait libraries

prevented probe reads from mapping to hg19. After mapping, the maToBam plugin was used to filter out all non-uniquely positioned reads in the genome.

Single nucleotide variants (SNV) were called with the Bioscope DiBayes SNP module. Stringency parameters were set to medium and het.skip.high.coverage set to 0, allowing the algorithm to call heterozygous SNVs for targeted resequencing approaches.

Results and discussion

While developing a method for targeted enrichment based on a short immortalized library, we selected 966 common cancer genes. We performed targeted enrichment and sequencing of genomic DNA derived from the human colon cancer cell line, SW480. Furthermore, we used water-in-oil emulsion PCR to amplify the bait library and demonstrated bias free amplification by sequencing the un-amplified and amplified bait libraries.

A schematic of the experimental design is shown in Figure 1. Herein, the sequencing of both bait libraries with (w/) and without (w/o) emulsion amplification is depicted, as well as the targeted enrichment by SIMLY and SureSelect bait libraries.

Targeted enrichment by SIMLY

This protocol was designed for generating a bait library made of in vitro transcribed and biotin labeled RNA from a DNA template oligo pool made by in situ synthesis on an array from MYcroarray. The template library was designed to comprise 89,909 individual probes with a 5′ T7 adaptor and a 3′ universal adapter (5′-TAATACGACTCACTATAGGG-N_{40}-GCACTGCAAAAAGCA-GGCTC-3′, where N_{40} denotes the specific probe sequence with 40 bp). The attached T7 Promoter has the advantage of direct RNA transcription, which allowed us to omit the second tailed PCR required by other methods when introducing the promoter (11). Prior to enrichment, target fragment libraries of 150 to 200 bp in length were prepared. To minimize nonspecific binding during hybridization, truncated SOLiD adaptors were used for library preparation with the biotin labeled RNA bait library. After enrichment, the full length primers (P1 and P2) were used to re-amplify the targeted sequences.

By applying SIMLY targeted enrichment, we were able to retrieve approximately 70 percent on-target sequences after read mapping. To accommodate for off-target reads in the vicinity of targeted loci due to the fragment library protruding further into off-target regions, target loci were enlarged by 100 bp for on target statistical calculations. The mean depth of coverage obtained by 24.2 Mb uniquely mapped reads on the targeted region was 280x. The distribution of normalized coverage and percent fraction of the targeted region is shown in Table 1 and Figure 2A. Normalized coverage was calculated by base wise coverage normalization against the average base coverage of all exons with at least one read.

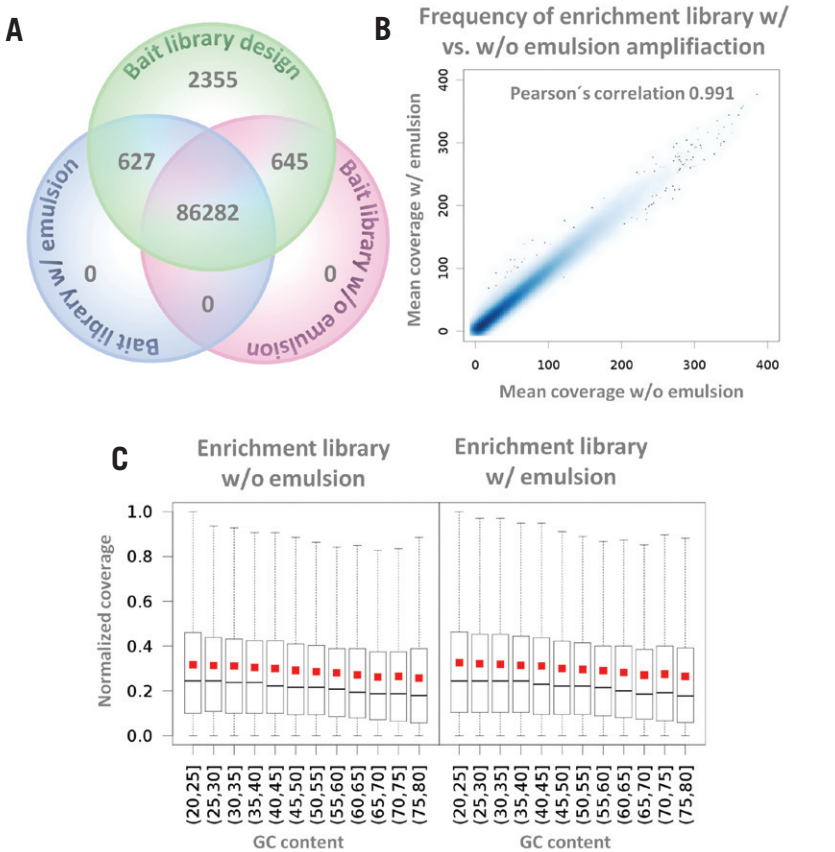

Figure 4. Comparison of template library amplification with and without emulsion PCR. (A) Venn diagram of identified baits after sequencing the amplified template library with and without emulsion PCR. (B) The plot shows mean coverage frequency for the amplified template library with and without emulsion PCR. (C) Boxplots showing normalized coverage for the amplified template library with and without emulsion PCR according to GC content. Solid squares indicate mean values for each data point.

Comparison of SIMLY and SureSelect enrichment

The Agilent SureSelect Whole Exome Kit is one of the most commonly used kits for targeted enrichment. We used this kit for enrichment with the same cell line (SW480) and SOLiD sequencing. Normalized coverage was calculated as described above.

With SIMLY enrichment, we obtained a coverage distribution almost identical to SureSelect Whole Exome enrichment, depicted in Figure 2A. The fraction of the enriched target region over normalized coverage matched for both methods and showed no significant difference over the full range of normalized coverage. For example, almost 60% of the target regions showed a normalized coverage of 0.25 with both methods, while 43% of the regions showed 0.5 and 35% showed 0.75. This high concordance shows that despite differences in library design, bait library generation, and enrichment procedure, both methods have equal efficiency in overall enrichment.

We then generated an overlapping data set containing only the shared fraction of bases for both targeting experiments (in total 15,517 loci covering 2.7 Mb) and similarly calculated a normalized coverage for each intersection locus. Based on this, we compared sequencing coverage for each locus and found the correlation to be 0.37 (see supplementary Figure 1A). We did not expect to see high correlation since bait design is known to strongly influence enrichment efficacy and this differed between the two experiments (18). Similarly, the bait library template abundance did not correlate with the sequencing coverage (Pearson´s correlation: -0.12, see supplementary Figure 1B).

It is well known that the guanine-cytosine (GC) content of the target region is skewed after targeted enrichment, which especially has a negative impact on targeted enrichment with high GC contents (19). Regardless, for our library design, we did not consider GC-content. To address this issue, we compared normalized coverage of

the intersecting regions depending on GC-content (Figure 2B) and found that normalized coverage of SIMLY and SureSelect enrichments were similarly biased toward higher GC-contents. The SIMLY approach shows higher coverage for targets with GC contents >55%, while SureSelect shows higher coverage for targets with <55%. This is likely caused by differences in probe length and melting temperatures (Tm). The GC-rich probes of the SureSelect Kit (120 bp) are more likely to bind and pull down off-target sequences through partial annealing than the shorter SIMLY probes (40 bp). On the other hand, SIMLY probes are less likely to pull down AT-rich targets because of their lower Tm and binding strength. These observations suggest that equalizing probe Tm by generating probes of different lengths may result in better coverage distributions of GC-rich and AT-rich targets.

Identification of genetic variants such as single nucleotide polymorphisms (SNPs) is a major application of targeted enrichment. Therefore, we compared genotype concordance between sequencing results from SIMLY and SureSelect enrichments with previously obtained data from the Affymetrix Human SNP Array 6.0 (Figure 2C). The intersecting region comprises 493 polymorphic loci. 435 (88.4%) of the loci were called identically by all three methods. SIMLY and SureSelect matched for 94.7%, SIMLY and Affymetrix for 90.2%, and SureSelect and Affymetrix for 89.8% of the SNP calls. Accordingly, higher concordance was observed between the independent sequencing data sets than between the sequencing and array results. The median coverage of matching SNP calls in SIMLY and SureSelect was 112x and 23x respectively, while average coverage in mismatched calls dropped to 77x and 11x, respectively. Technical issues therefore are less likely to be the cause of the <95% concordance rate of SIMLY and SureSelect results than the low sequencing depth of the SureSelect experiment.

The shorter SIMLY probes might have altered performance in regions of mismatches. To further address probe performance, we analyzed 402 single nucleotide variations (SNVs) that were identified within the overlapping target region by both methods and had a coverage of ≥ 30x (Figure 2D). The normalized ratio of wild type to novel alleles was plotted for each method. By setting a homozygozity cut-off at ≥ 0.8, we found 208 homozygote and 157 heterozygote non-wild type loci in both methods in concordance, while 41 loci were not called in concordance. Eight loci were called homozygous by SureSelect but heterozygous by SIMLY; and nine homozygous calls from SIMLY were heterozygous in SureSelect. For both methods, locus drop-out affected the same number of loci. If SIMLY probes in SNV regions had demonstrated a performance loss, it would be indicated by an increase in loci drop-out when compared with SureSelect, as well as a decrease in the heterozygozity rate. Since no such changes were observed, we conclude that SNVs do not affect SIMLY probe performance.

When addressing the relevance of the detected SNVs in the target region, we compared all SNVs causing amino acid changes to dbSNP, a database holding all known single nucleotide polymorphisms, and the COSMIC database. The majority (~70%) of the SNVs were common SNPs found in ≥ 1% of the dbSNP samples, while ~30% were not present in dbSNP, indicating that they were potentially novel. In particular, we found several non-synonymous SNVs overlapping COSMIC mutations or SNPs that were reported to associate with clinical symptoms. Those genes, including *APC, ASXL1, BRCA1, ERBB2, KIT, KRAS, MPL, NOTCH2, SYNE1* and *TP53*, are directly linked to colorectal cancer. (All non-synonymous SNVs in SW480 that have been identified in this work are listed in supplementary Table 1.) We did not detect any mutations within the 481 ultraconserved regions.

Figure 3 gives an example of the coverage for the enriched genomic DNA samples of two genes (Figure 3A and 3B). Figure 3C shows the complete target region and the usual fluctuation in evenness generally obtained by hybrid capture enrichment. Note that the sequencing gap on

chromosome 2 of the SureSelect enrichment is due to the TTN gene (~104 kb) that is not included in the whole exome library.

Water-in-oil emulsion amplification of bait libraries

Next, we investigated the possibility of amplifying the bait library to allow enrichment of many samples from the same synthesized template library stock. To impede amplification bias introduced by PCR cycling, the template library was amplified for 15 cycles by water-in-oil emulsion PCR. Subsequently, the original template library and the emulsion amplified replica were re-amplified for 5 cycles. During this step, the tailed primers P1/P2 SOLiD were used, including a custom barcode sequence for later identification. Both libraries were SOLiD sequenced.

In total, we obtained 3.65 and 3.66 million reads for the emulsion amplified library and original template library, respectively. Of the 89,909 individual oligos, approximately 97% of the baits could be identified after mapping the reads against the designed oligos. Therefore, 2355 oligos were not amplified or were missing from the synthesized bait library regardless of the amplification process (Figure 4A). To assess the composition of the template library with and without emulsion amplification, mean coverage frequencies were plotted against each other and a Pearson's correlation coefficient of 0.991 was calculated between the two data sets (Figure 4B). Accordingly, 98.2% of all baits were identically represented in both pools, either with or without emulsion amplification.

High GC content is known to hinder efficient amplification during PCR. However, the GC-coverage profiles of emulsion amplified and the original library are almost identical and show that no significant amplification bias was caused by GC content (Figure 4C).

Figure 4A and 4B illustrate the high concordance of the original and emulsion amplified libraries for two target genes. In both cases, exon coverage patterns are almost identical.

In this work, the aim was to establish a cost efficient method for targeted enrichment based on in-solution capture. We succeeded in targeting 966 cancer associated genes with a total target region size of 3.7 Mb.

We also demonstrated amplification of the synthesized template library using water-in-oil emulsion PCR. We did not find evidence of amplification biases and concluded that this strategy enables copious enrichments from one synthesized bait library stock.

By applying the SIMLY enrichment procedure, costs for targeted enrichment can be reduced by an order of magnitude compared with current commercial kits. In detail, costs for the described enrichment were as follows: template library synthesis cost 1000 USD; targeted enrichment per sample cost about 30 USD, including T7 transcription, enrichment with beads, and PCR or purification. To perform a targeted enrichment experiment with 100 samples, the total cost for enrichment would be 4000 USD. In comparison, a targeted enrichment for a similar target region size performed with SureSelect would cost about 35,000 USD. Final costs for targeted enrichment largely depend on the number of samples to be enriched, because upfront expenses for template library synthesis need to be averaged between all samples.

In conclusion, NGS and targeted enrichment has proven to be a very powerful tool in genetic research and is anticipated to grow further into diagnostic applications. Being able to reduce costs for targeted enrichment might even strengthen these efforts and help open targeted enrichment to an even greater portion of the research community.

Acknowledgments

We are grateful to Uta Marchfelder, Anna Kosiura, and Stefan Boerno for their excellent technical assistance. The research leading to these results received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 201418 (READNA) and support by grants of the German Federal Ministry of Education and Research BMBF (n° 01GS08105 Mutanom, n° 01GS08111 Modifier and n° 0315428A Predict).

Competing interests

The authors declare no competing interests.

References

1. **Fuller, C.W., L.R. Middendorf, S.A. Benner, G.M. Church, T. Harris, X. Huang, S.B. Jovanovich, J.R. Nelson, et al.** 2009. The challenges of sequencing by synthesis. Nat. Biotechnol. *27*:1013-1023.
2. **Metzker, M.L.** 2010. Sequencing technologies - the next generation. Nat. Rev. Genet. *11*:31-46.
3. **Mardis, E.R.** 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. *24*:133-141.
4. **Wang, Z., M. Gerstein, and M. Snyder.** 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*:57-63.
5. **Park, P.J.** 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*:669-680.
6. **Summerer, D.** 2009. Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. Genomics *94*:363-368.
7. **Ng, S.B., E.H. Turner, P.D. Robertson, S.D. Flygare, A.W. Bigham, C. Lee, T. Shaffer, M. Wong, et al.** 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature *461*:272-276.
8. **Hoischen, A., C. Gilissen, P. Arts, N. Wieskamp, W. van der Vliet, S. Vermeer, M. Steehouwer, P. de Vries, et al.** 2010. Massively parallel sequencing of ataxia genes after array-based enrichment. Hum. Mutat. *31*:494-499.
9. **Li, Y.R., N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, T. Jiang, H. Jiang, A. Albrechtsen, G. Andersen, et al.** 2010. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat. Genet. *42*:969-972.
10. **Mertes, F., A. Elsharawy, S. Sauer, J.M. van Helvoort, P.J. van der Zaag, A. Franke, M. Nilsson, H. Lehrach, and A.J. Brookes.** 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics *10*:374-386.
11. **Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E.M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, et al.** 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. *27*:182-189.
12. **Albert, T.J., M.N. Molla, D.M. Muzny, L. Nazareth, D. Wheeler, X. Song, T.A. Richmond, C.M. Middle, et al.** 2007. Direct selection of human genomic loci by microarray hybridization. Nat. Methods *4*:903-905.
13. **Leibovitz, A., J.C. Stinson, W.B. McCombs 3rd, C.E. McCoy, K.C. Mazur, and N.D. Mabry.** 1976. Classification of human colorectal adenocarcinoma cell lines. Cancer Res. *36*:4562-4569.
14. **Griffiths, A.D. and D.S. Tawfik.** 2006. Miniaturizing the laboratory in emulsion droplets. Trends Biotechnol. *24*:395-402.
15. **Schutze, T., F. Rubelt, J. Repkow, N. Greiner, V.A. Erdmann, H. Lehrach, Z. Konthur, and J. Glokler.** 2011. A streamlined protocol for emulsion polymerase chain reaction and subsequent purification. Anal. Biochem. *410*:155-157.
16. **Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, and D. Haussler.** 2004. Ultraconserved elements in the human genome. Science *304*:1321-1325.
17. **Blumenstiel, B., K. Cibulskis, S. Fisher, M. DeFelice, A. Barry, T. Fennell, J. Abreu, B. Minie, et al.** 2010. Targeted exon sequencing by in-solution hybrid selection. Curr Protoc Hum Genet *Chapter 18*:Unit 18 14.
18. **Hedges, D.J., T. Guettouche, S. Yang, G. Bademci, A. Diaz, A. Andersen, W.F. Hulme, S. Linker, et al.** 2011. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. PLoS One *6*:e18595.
19. **Tewhey, R., M. Nakano, X. Wang, C. Pabon-Pena, B. Novak, A. Giuffre, E. Lin, S. Happe, et al.** 2009. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol. *10*:R116.

Received 13 February 2012; accepted 18 May 2012.

Address correspondence to Florian Mertes, **Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany.** Email: mertes@molgen.mpg.de