applied biosystems by life technologies™

Microdroplet-based PCR enrichment of genomic regions for targeted sequencing on the SOLiD® System

An effective strategy for identifying common and rare variants in candidate regions of the human genome is critical to understanding the etiology of disease susceptibility. The identification of genetic variants and mutations associated with complex diseases requires a robust and cost-effective approach for systematic resequencing of candidate regions in the human genome. When combined with microdroplet-based PCR enrichment (Figure 1), the scalable throughput of the SOLiD® System facilitates deep sequencing of targeted genomic regions of interest. The method employed by the RainDance Sequence Enrichment Solution enables amplification of regions representing up to 10 Mb of genomic sequence for parallel variant screening in large numbers of genes and samples. Post-enrichment material is purified and incorporated into the SOLiD® System workflow for library generation, templated bead preparation, and ligation-based sequencing (Figure 2). The scalability and specificity of the RainDance targeted sequencing solution coupled with the high throughput of the SOLiD® sequencing platform provides an integrated approach to targeted resequencing that matches the needs of genome-wide association studies (GWAS) and cancer research. This method demonstrates the ability to interrogate genes associated with common cancers to better understand various cancer subtypes.

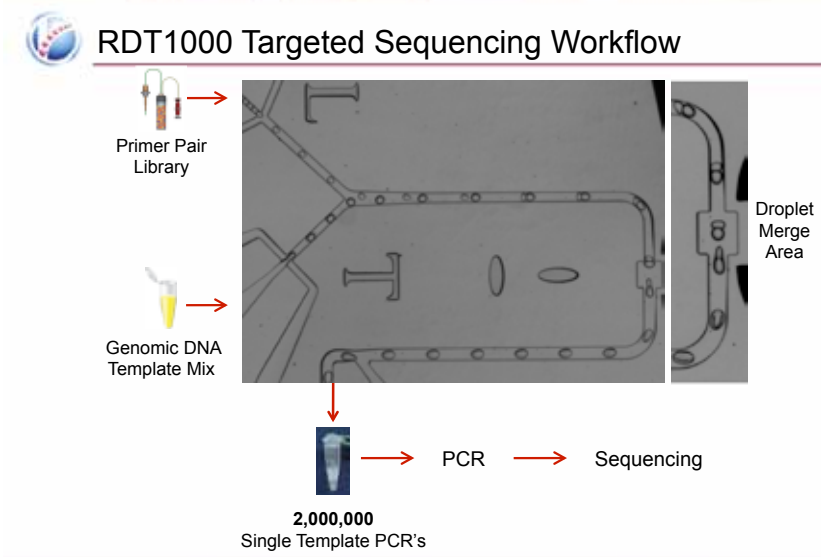


Figure 1. Microdroplet PCR-based targeted enrichment. The process of merging picoliter volume droplets of fragmented genomic DNA with primer pair droplets in a 1:1 ratio on a microfluidic chip to form PCR droplets is depicted. The resulting PCR droplet library, consisting of over 1.5 million droplets, is amplified to enrich for specific regions of the genome. Following PCR, the droplets are destabilized to release the amplicons for purification, library preparation, and sequencing.

life technologies™

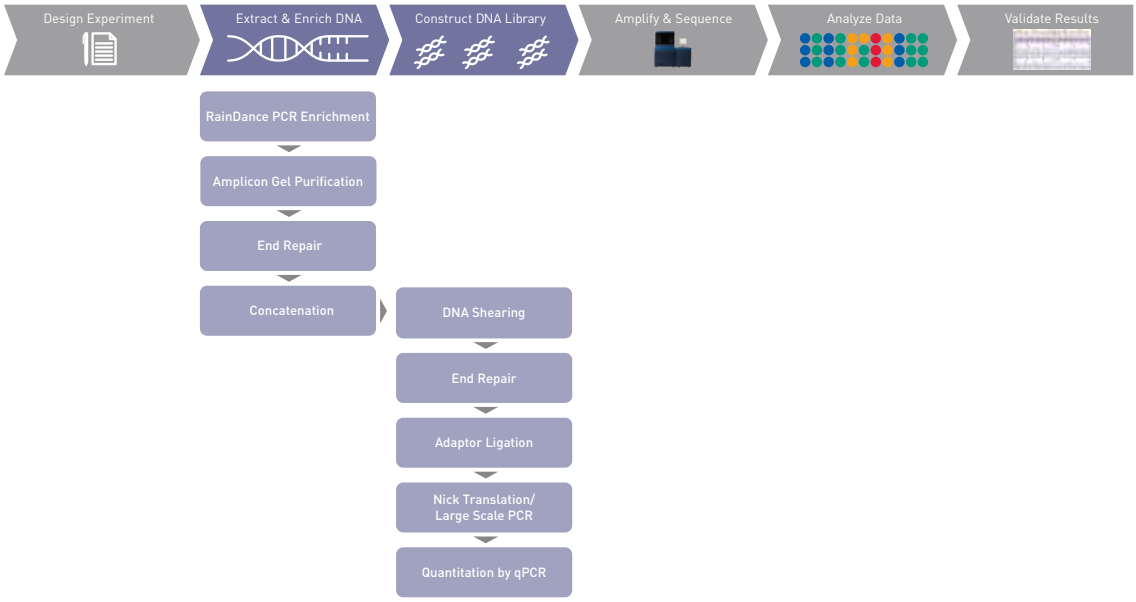


Figure 2. SOLiD® System PCR-based targeted sequencing workflow. Sample preparation is shown with highly parallel singleplex PCR in droplets of desired regions using predesigned primers. Purified amplicons are concatenated to minimize end biases that may result from DNA shearing during library construction. Fragmentation of concatenated DNA, followed by end-repair and ligation to P1 and P2 adaptor sequences, are performed as illustrated. The resulting fragment library undergoes nick-translation, and then amplification to generate enough material for templated bead preparation. Accurate library quantification is determined by quantitative PCR before proceeding to the standard SOLiD® System sequencing workflow. Follow-up validation of novel SNPs or small insertions and deletions is performed by capillary-based sequencing.

Materials and methods

Primer library targets and primer design

The RainDance Oncology Panel consists of 142 genes that harbor potential driver mutations found in a number of common cancers. These regions include coding exons, splice junctions, 5'- and 3'-untranslated regions (UTRs), and promoters of the target genes. Primers were designed for 3,979 amplicons to cover the targets with a total amplicon sequence capture size of approximately 1.5 Mb.

Primer library generation

Individual primer droplets were generated using the RainStorm™ microdroplet-based technology from RainDance, from a primer aliquot containing an equal concentration of both the forward and reverse primers for each of the amplicons in the primer library. Aliquots of the primer droplet library were prepared for use on the RDT 1000.

Targeted enrichment and library preparation

Enrichment of target sequences was performed according to the *RainDance RDT 1000 Sequence Enrichment Assay Manual*. Briefly, for each experimental condition 2 µg of gDNA from HapMap NA18858 (Yoruba DNA) was fragmented to a size range of 2–4 Kb using the Covaris® S2 System (Covaris, Inc.) according to the DNA miniTUBE - Blue protocol (available at http://covarisinc.com/supported-protocols.html).The PCR template mix, containing the fragmented gDNA and PCR reagents, was loaded into the RDT 1000 along with the primer library.

PCR droplets were collected in a 0.2 mL PCR tube and amplified using 55 cycles of PCR. Amplification products were recovered by breaking the emulsions, followed by amplicon purification using a MinElute® PCR Purification Kit (Qiagen).

Quantitative and qualitative analysis of the amplification products was performed on an Agilent® 2100 Bioanalyzer™ (Figure 3B).

Up to 750 ng of enriched DNA was purified using E-Gel® SizeSelect™ 2% Agarose gels (Invitrogen) according to the DNA purification protocol in the E-Gel® Technical Guide. Fractions from electrophoresis were collected at 2 min intervals and those containing the amplicons were pooled. Purification of the pooled, size-selected PCR products was performed using MinElute® columns. To understand the impact of gel purification at this step, a non-purified sample was run in parallel. Purified or non-purified samples were then concatenated for 30 min or overnight as described in the Applied Biosystems® SOLiD® System Amplicon Concatenation Protocol. Standard fragment libraries were generated in accordance with the *Applied Biosystems® SOLiD® 3 Plus System Library Preparation Guide*.

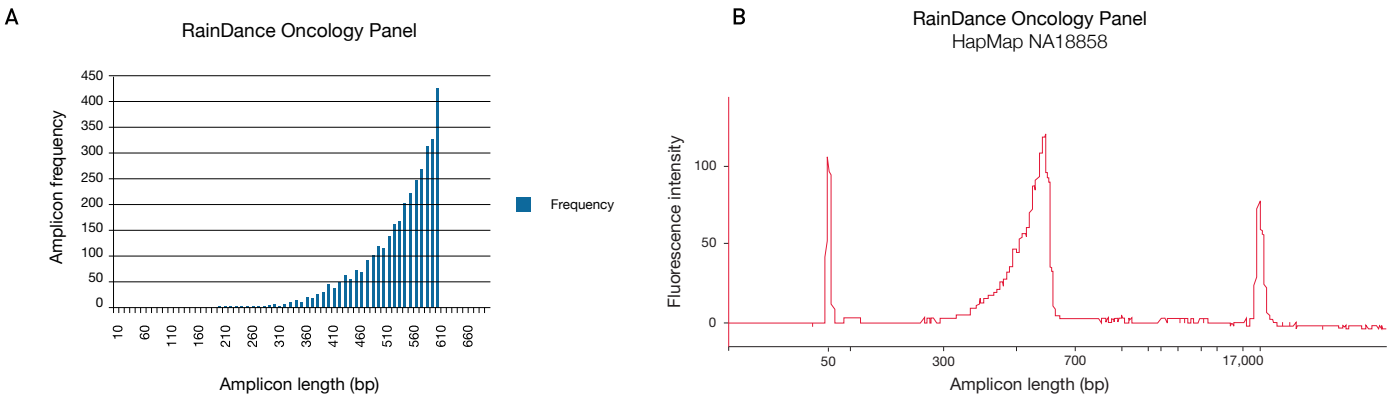


Figure 3. High correlation between expected and actual amplification profiles for the RainDance Oncology Panel. The amplification profiles for the RainDance Oncology Panel were (**A**) calculated or (**B**) measured using the Agilent Bioanalyzer.

Bead preparation and SOLiD® sequencing

Enriched libraries were prepared according to the *Applied Biosystems® SOLiD® 3 Plus System Templated Bead Preparation Guide*. Each templated bead sample was deposited onto an octet of a slide at an average density of 100,000 beads per panel. Sequencing by ligation was carried out to 50 bp on the SOLiD® 3 Plus Analyzer in accordance with the *Applied Biosystems® SOLiD® 3 Plus System Instrument Operation Guide*.

Data analysis

Sequencing reads (50 bp) from each sample were analyzed using the SOLiD® Accuracy Enhancer Tool (available at http://info.appliedbiosystems.com/solidsoftwarecommunity) and aligned to the human genome reference sequence (hg18). Single nucleotide polymorphism (SNP) detection was performed by aligning the reads to the targeted sequences using SOLiD® BioScope™ Software. Enrichment performance was evaluated by calculating the proportion of uniquely mapped reads that aligned to the targeted sequences for each sample.

Results

The SOLiD® System and the RainDance Targeted Sequencing Solution provide a powerful PCR-based approach for detecting genetic variation. Targeted enrichment and sequencing of ~4,000 regions specific to exons, splice junctions, untranslated regions (UTRs), and promoters of 142 genes associated with certain cancers were performed using human DNA samples. Highly specific PCR-based enrichment of a target region of approximately 1.5 Mb was observed, as demonstrated by the amplicon abundance and length (Figure 3), as well as enrichment specificity for all samples (Figure 4).

The sample preparation strategy sought to enhance the existing SOLiD® System PCR-based targeted sequencing workflow by evaluating the effect of amplicon gel purification and concatenation time on enrichment efficiency (Figure 4). Gel purification of the amplicons resulted in increased specificity of the reads to the desired target, as shown by comparing unconcatenated, gel-purified (GP_No Ligation) and non gel-purified (Ori_No Ligation) mapping profiles. The results also show no significant improvement in target specificity when the concatenation reaction time is increased from 30 min to overnight (GP_30minLigation vs. GP_OVNLigation); however, other parameters, such as increasing the ligase concentration in the reaction, could further enhance concatenation efficiency if implemented. Optimal enrichment efficiency was achieved (~90%) by post-enrichment gel purification of the amplification products followed by end repair and concatenation prior to library construction (Figure 4).

For accurate detection of genetic variants, the extent of coverage of the target regions was assessed for all of the enriched samples. Here we show data comparing only 3 of the 6 samples, demonstrating that the optimal parameters include amplicon gel purification and 30 min concatenation time. Evaluation of the average coverage of these samples showed that ≥98% of the target bases were covered by at least 1 read while ≥93% of the target bases were covered by at least 30 reads (Figure 5A). Similar results were observed for the normalized coverage of these samples (Figure 5B). Coverage profile characteristics were highly reproducible when identical sample preparation conditions were used

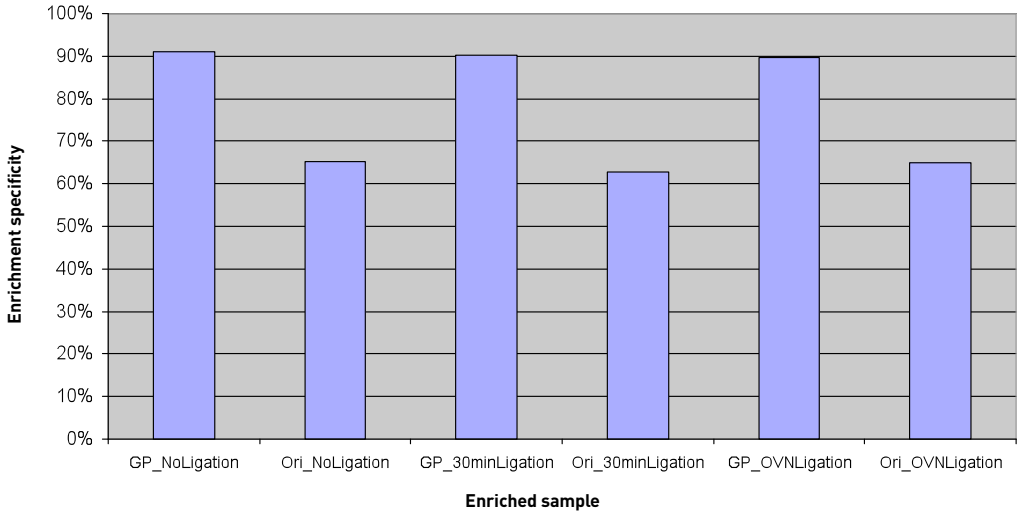


Figure 4. Gel purification increases enrichment specificity. The fraction of uniquely mapped reads for enriched libraries is indicated on the y-axis. Microdroplet-based PCR products were either used directly for library preparation (Ori_NoLigation), gel purified and then used directly for library preparation (GP_NoLigation), gel purified and concatenated overnight (GP_OVNLigation) or for 30 min (GP_30minLigation), or not purified but concatenated overnight (Ori_OVNLigation) or for 30 min (Ori_30minLigation). Uniquely mapped reads are reads that map to a single, unique location. Gel purified samples (GP) show better enrichment specificity compared to samples that were not gel purified (Ori).

(data not shown). Thus, the specificity and sensitivity of the RainDance technology coupled with the accuracy and throughput of the SOLiD® System is ideal for detecting genetic variants in your research samples.

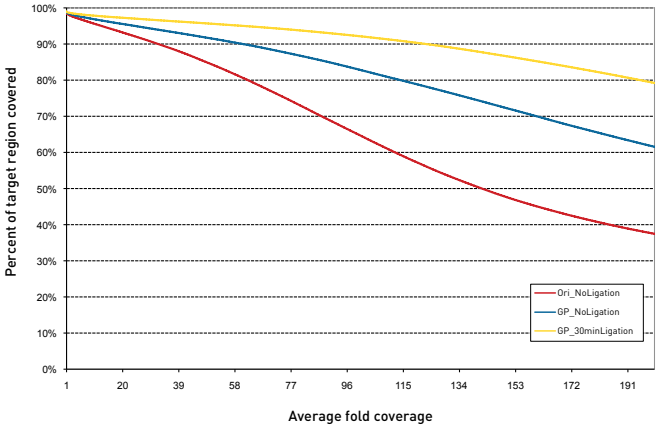
In order to determine sensitivity, specificity, and concordance of SNP detection, SNPs identified in this study were compared to genotypes in the HapMap database for the same sample. Classification of SNPs is outlined in Table 1. The total number of SNPs identified in the targeted regions by the SOLiD® System is shown for each of the 6 samples (Table 1). Approximately 20% of the SNPs identified in the targeted regions of each sample are novel.

Reported values for sensitivity and specificity for the unconcatenated non gel-purified (Ori_No Ligation), unconcatenated gel-purified (GP_No Ligation), and concatenated gel-purified samples (GP_30minLigation) are shown, taking into account the number of undercalls, or true heterozygous SNPs called as homozygous SNPs (Table 2). The GP_30minLigation sample showed the greatest enrichment specificity; sensitivity and specificity values were greater than 98% for homozygous and heterozygous SNPs for this sample. In general, specificity and sensitivity values increase with additional sequence coverage and with more accurate raw reads. Increased sensitivity and specificity is expected when analyzed on the 5500 Series Genetic Analysis Systems, which deliver up to 99.99% accuracy when used with the Exact Call Chemistry (ECC) Module.

Conclusion

Together, the SOLiD® System and the RainDance Sequence Enrichment Solution provide a highly sensitive and specific solution for parallel genetic variant screening in basic and cancer research. The scalability and specificity of the RainDance microdroplet-based enrichment method combined with the throughput and accuracy of the SOLiD® System enable researchers to gain novel insights into tumorigenesis, disease susceptibility, and sample heterogeneity by performing ultra-deep sequencing of specific regions of interest for rare variant discovery.

A



B

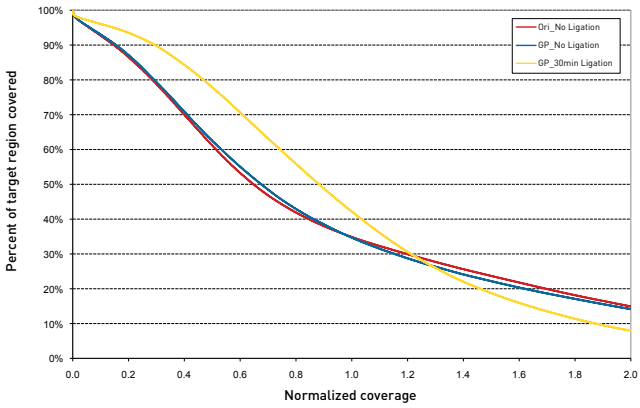


Figure 5. Percentage of bases covered in the target regions vs. depth of coverage for enriched samples. Graphical representations of the percentage of uniquely mapped SOLiD® reads aligning to the target region for each sample based on (**A**) average fold coverage or (**B**) normalized coverage are shown.

Table 1. SNP calls for enriched samples sequenced using the SOLiD® System.

	Number of calls* per sample		
SNP classification	**Ori_NoLigation**	**GP_NoLigation**	**GP_30minLigation**
Heterozygous true positives	394	402	410
Homozygous true positives	2,418	2,424	2,436
Heterozygous false positives	5	5	3
Homozygous false positives	21	15	8
Heterozygous undercalls	19	13	6
Heterozygous uncalled	3	1	0
Homozygous uncalled	26	20	12
Total number of SNPs identified in enriched region	1,502	1,487	1,487
Novel SNPs	331	288	247
Concordance with dbSNP in enriched regions	78.0%	80.6%	83.4%

* Reported values are based on comparisons to HapMap3

Table 2. SNP discovery statistics in enriched samples.

	Ori_NoLigation		GP_NoLigation		GP_30minLigation	
	Homozygous SNPs	**Heterozygous SNPs**	**Homozygous SNPs**	**Heterozygous SNPs**	**Homozygous SNPs**	**Heterozygous SNPs**
Specificity*† (excluding undercalls)	94.9%	99.8%	96.4%	99.8%	98.1%	99.9%
Sensitivity*‡ (excluding undercalls)	99.8%	95.4%	99.8%	96.9%	99.9%	98.6%
Specificity*† (including undercalls)	95.0%	99.8%	96.4%	99.8%	98.1%	99.9%
Sensitivity*‡ (including undercalls)	99.8%	94.7%	99.8%	96.6%	99.9%	98.6%

* Reported values are based on comparisons to HapMap3
† Specificity = (true negatives)/(true negatives + false positives)
‡ Sensitivity = (true positives)/(true positives + false negatives)

Life Technologies offers a breadth of products DNA | RNA | PROTEIN | CELL CULTURE | INSTRUMENTS

Headquarters
5791 Van Allen Way | Carlsbad, CA 92008 USA | Phone +1.760.603.7200 | Toll Free in the USA 800.955.6288
www.lifetechnologies.com

life technologies™