PCR-based targeted sequence enrichment for next generation sequencing platform

R. Padilla[1], A.B. Shah[1], L. Z. Pham[2], J. Ni[1], M. Matvienko[1], N. Hoag[1], K. Li[1], J. Ziegle[1]
[1]Life Technologies, Foster City, CA, United States; [2]RainDance Technologies, Inc., Lexington, MA, United States.

ABSTRACT

Many human diseases are associated with genetic polymorphisms. Resequencing candidate regions can provide valuable information about the genetic basis for these diseases. Combining Next Generation Sequencing with a new PCR-based enrichment method generates a robust and cost-effective workflow for deeper interrogation of targeted genomic regions of interest for specific applications. Here, we report the use of Next Generation Sequencing with PCR-based enrichment to extract target regions from Yoruba DNA. We present an optimized and flexible workflow for library construction post PCR enrichment to emulsion PCR and sequencing on a Next Generation Sequencing platform. We demonstrate that this pipeline provides a useful solution for targeted resequencing applications.

INTRODUCTION

The identification of genetic variants and mutations associated with complex diseases requires the development of a robust and cost-effective approach for systematic resequencing of candidate regions in the human genome. When combined droplet-based PCR enrichment approach (Figure 2), the scalable throughput of the SOLiD™ System facilitates deep sequencing of target genomic regions of interest. The method employed by the Raindance Sequence Enrichment Solution can amplify regions representing up to 20 Mb of genomic sequence for parallel variant screening in a large number of genes and large number of samples. Post-enrichment material is then purified and incorporated into the SOLiD™ System workflow for library generation, templated bead preparation, and ligation-based sequencing (Figure 1). The inherent scalability and specificity for the Raindance enrichment solution coupled with the high throughput of the SOLiD™ sequencing platform provides an integrated approach to targeted resequencing that is particularly suited for medical and cancer resequencing as well as for follow-up Genome Wide Association Studies (GWAS).

MATERIALS AND METHODS

Resequencing Targets and Primer Design

The RainDance Oncology Panel consists of 142 genes that contain driver mutations in a number of common cancers. Target regions include the coding exons, splice junctions, 5'-, and 3'-UTRs, and promoters in these genes. 3,979 amplicons were designed to cover the targets at 100% success rate with the total amplicon sequences of 1.5Mb.

Primer Droplet Library Generation

Pooled forward and reverse primers were prepared at equal concentrations for each of the 3,979 amplicons. The primers were reformatted into droplets in a serial process and pooled into a single droplet library (Figure 2). Aliquots of the droplet library were prepared for use on the RDT 1000.

Targeted Enrichment and Library Preparation

Enrichment was performed according to the RainDance RDT 1000 Sequence Enrichment Assay Manual. In short, for each experimental condition, 2 ug 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 from Covaris Inc (www. 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. Droplets were collected into a 0.2 ml PCR tube and amplified using 55 cycles of PCR. Amplification products were recovered by breaking the emulsions and amplicon purification was performed using a MinElute PCR Purification kit (Qiagen). Quantitative and qualitative analysis of the amplification products was performed on an Agilent 2100 Bioanalyzer (Figure 3).

Up to 750ng of enriched DNA was purified using Invitrogen's E-Gel® 2% Size Select™ gel according to the DNA Purification Using E-Gel® SizeSelect™ Agarose Gels protocol in the E-gel® Technical Guide. During target amplicon selection, fractions containing the amplicons were collected during 2 min intervals of electrophoresis and pooled. Purification of the pooled, size-selected PCR products was performed using Qiagen® MinElute columns (Qiagen). 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 minutes 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 1).

Bead Preparation and Massively Parallel Sequencing

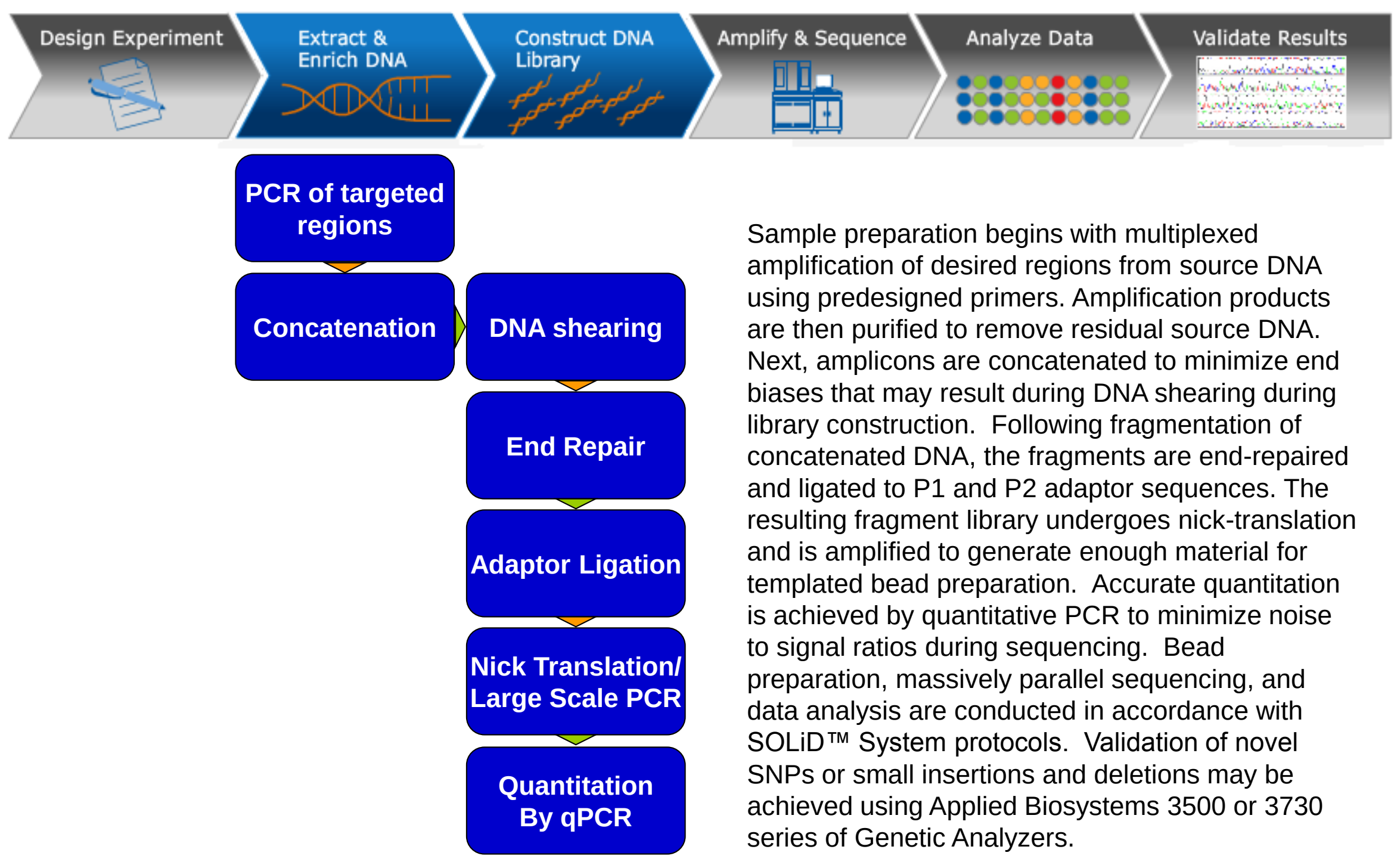
Enriched libraries were prepared based on the Applied Biosystems SOLiD™ 3 Plus System Templated Bead Preparation Guide. Each templated bead sample was deposited on an octet of a slide at an average bead density of 100K per panel. Sequencing by ligation was carried out to 50 bp on the SOLiD™ 3 Plus Analyzer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

Data Analysis

50 bp sequencing reads from each sample were analyzed using the SOLiD™ Accuracy Enhancer Tool (http://info.appliedbiosystems.com/solidsoftwarecommunity) and then aligned against the hg18 human genome reference sequence. SNP detection was performed against the reads aligning to targeted sequences using the SOLiD™ BioScope™ v1.2 Resequencing Pipeline with modified parameters. Enrichment performance was evaluated by calculating the proportion of uniquely mapped reads aligning to the targeted sequences for each sample.

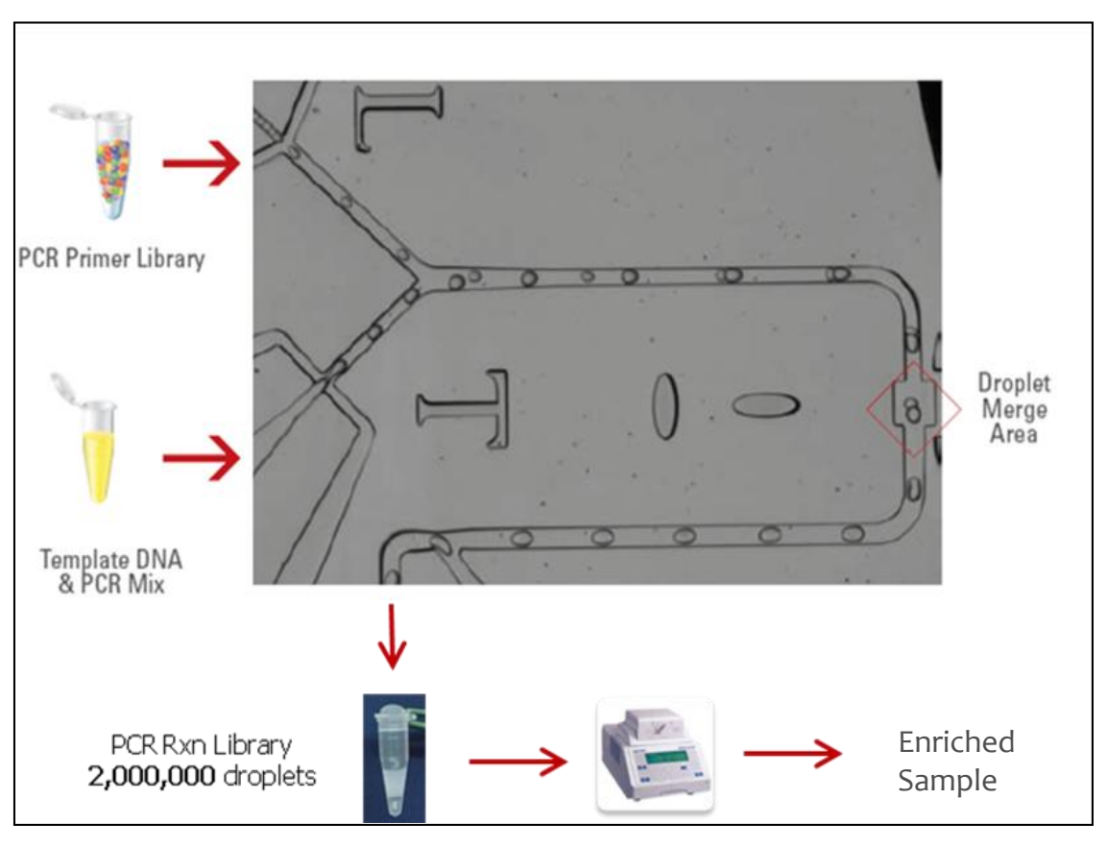
RESULTS

Figure 1. SOLiD™ System PCR-based Targeted Resequencing workflow



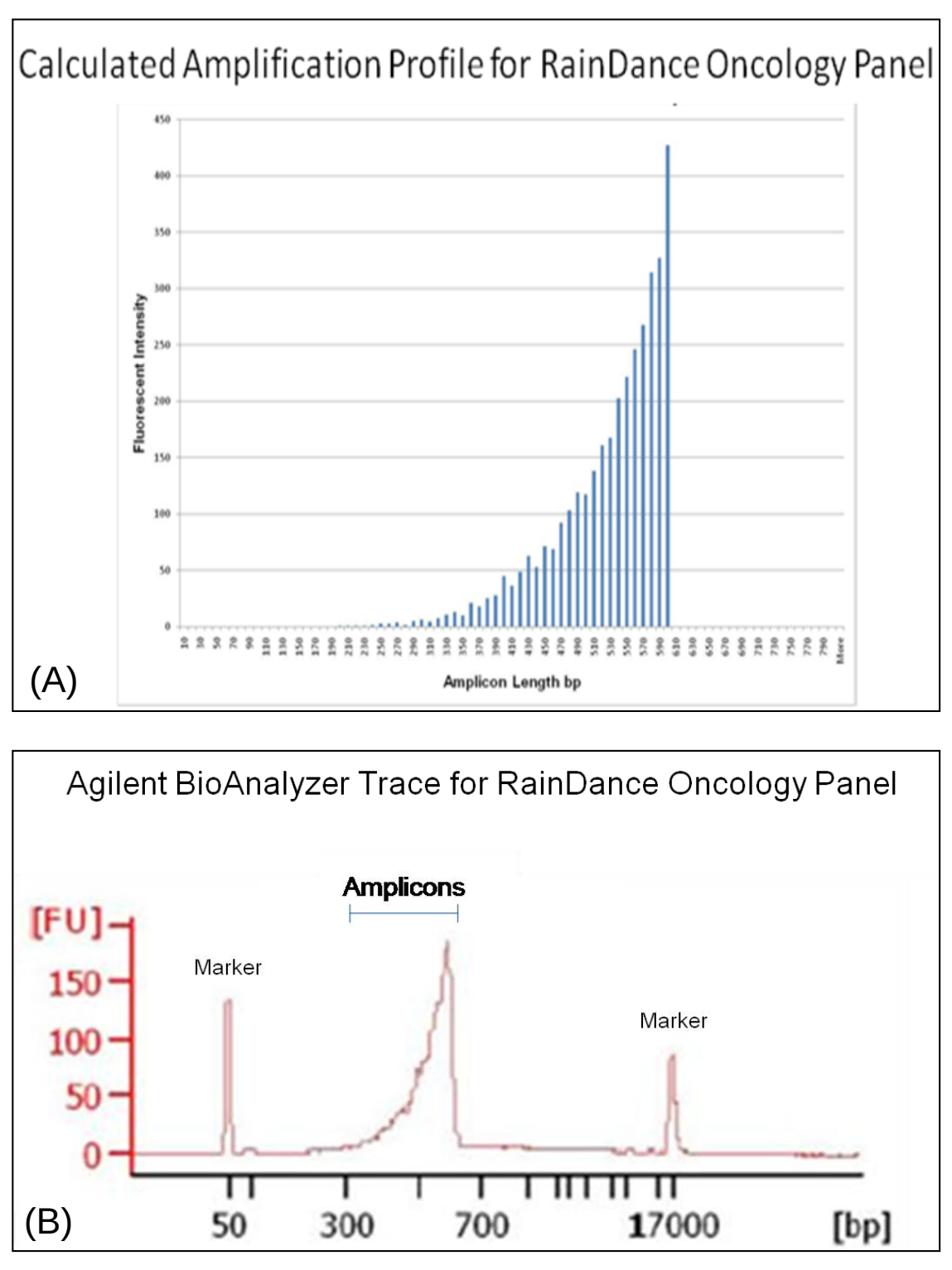
Sample preparation begins with multiplexed amplification of desired regions from source DNA using predesigned primers. Amplification products are then purified to remove residual source DNA. Next, amplicons are concatenated to minimize end biases that may result during DNA shearing during library construction. Following fragmentation of concatenated DNA, the fragments are end-repaired and ligated to P1 and P2 adaptor sequences. The resulting fragment library undergoes nick-translation and is amplified to generate enough material for templated bead preparation. Accurate quantitation is achieved by quantitative PCR to minimize noise to signal ratios during sequencing. Bead preparation, massively parallel sequencing, and data analysis are conducted in accordance with SOLiD™ System protocols. Validation of novel SNPs or small insertions and deletions may be achieved using Applied Biosystems 3500 or 3730 series of Genetic Analyzers.

Figure 2. Microdroplet PCR-based Targeted Enrichment



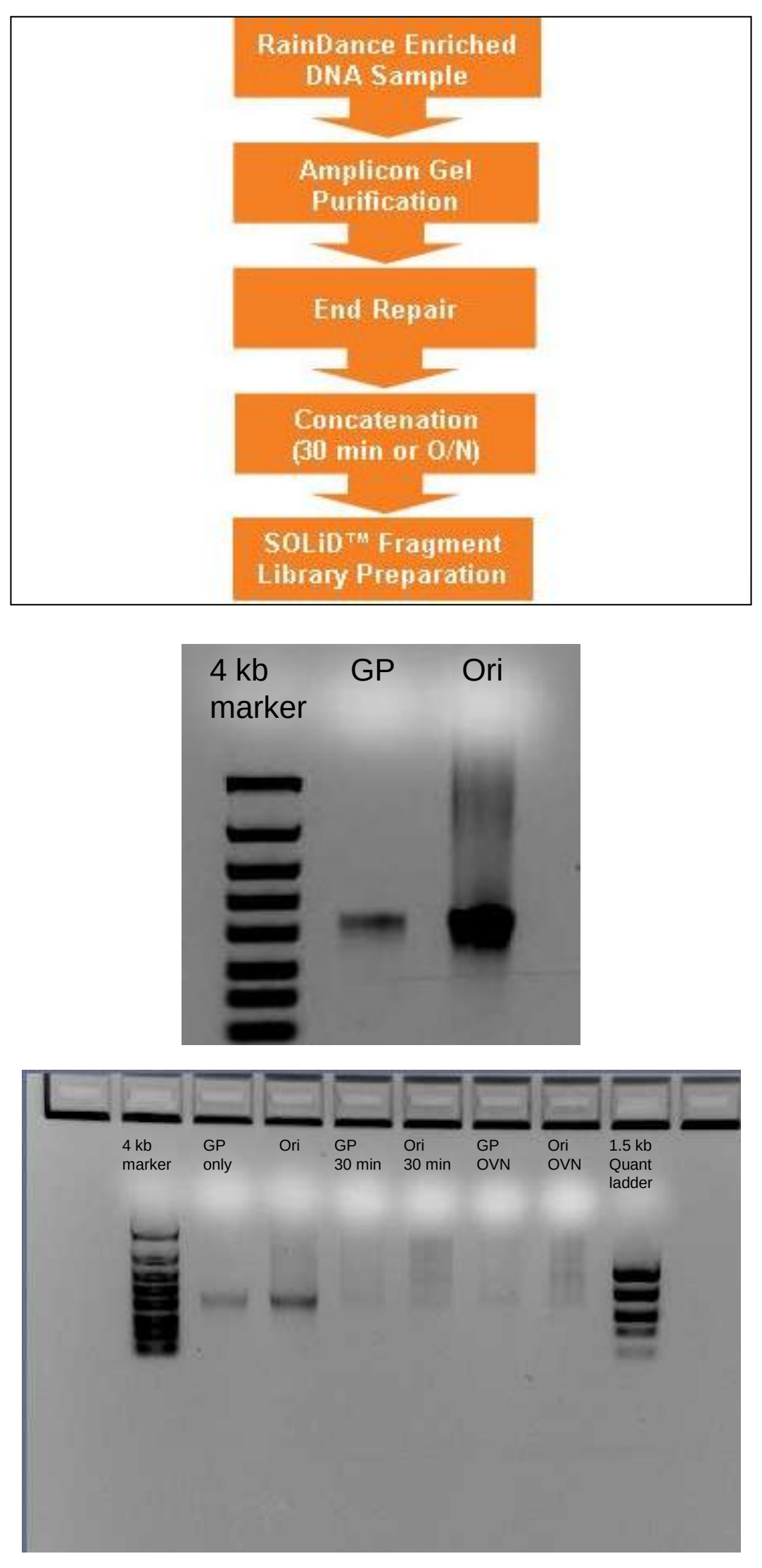
The process of merging picoliter volume droplets of fragmented genomic DNA with primer pair droplets in a 1:1 ratio on a microfluidic chip to form PCR droplets is depicted. The resulting PCR droplet library, consisting of over 1.5 million droplets, is amplified to enrich for specific regions of the genome.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

Figure 3. Relative amplicon abundance and length for Raindance Oncology Panel



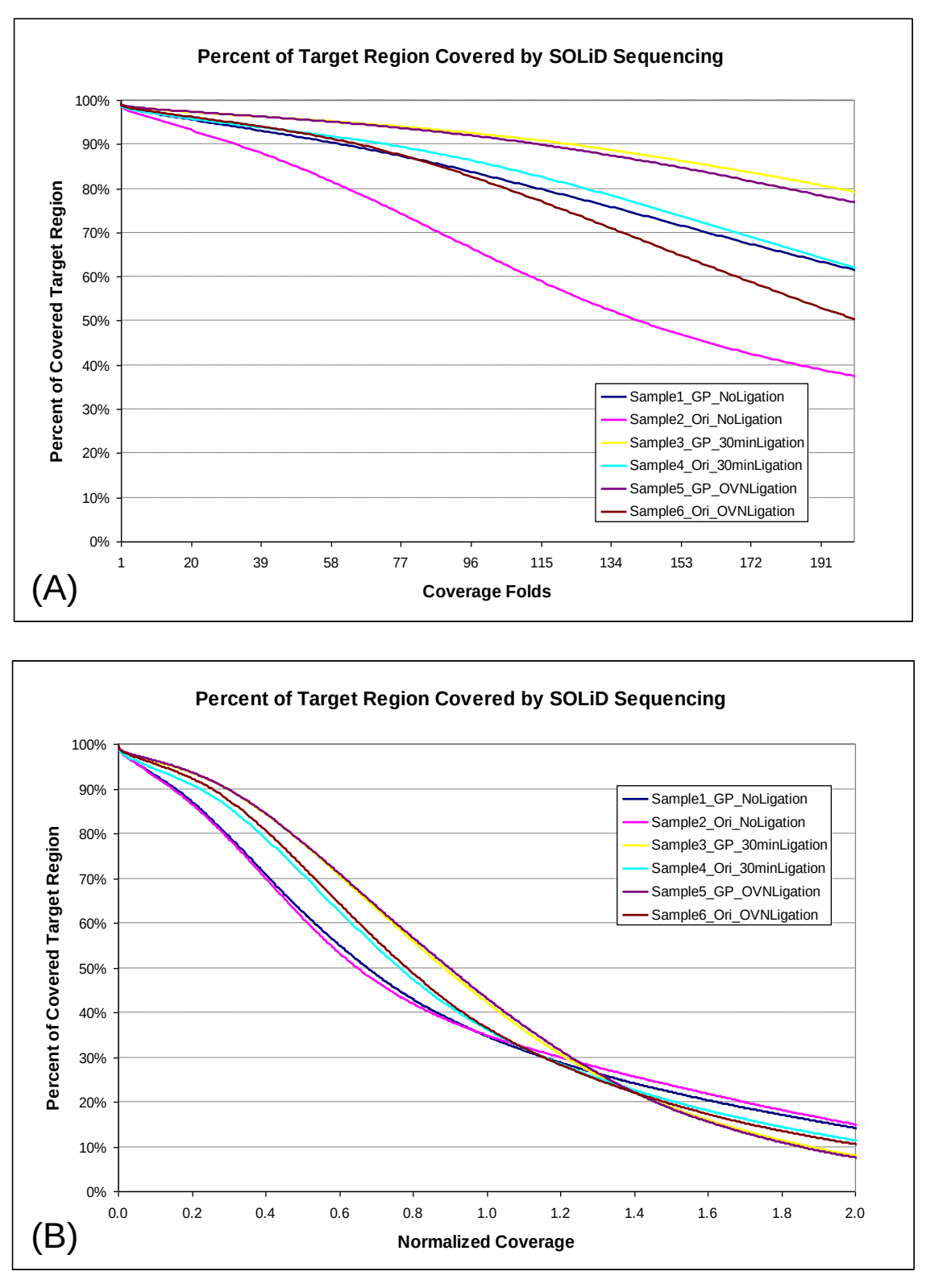
High degree of correlation between predicted (A) and actual (B) amplicons abundance and length of the Raindance Oncology Panel. Fluorescent intensity, an indicator of amplicon abundance, is plotted against amplicon size in base pairs.

Figure 4. Strategy for post-enrichment sample preparation for sequencing sensitivity and specificity


RainDance Enriched DNA Sample
Amplicon Gel Purification
End Repair
Concatenation (30 min or O/N)
SOLiD™ Fragment Library Preparation
4 kb marker
GP
Ori
4 kb marker
GP only
Ori
GP 30 min
Ori 30 min
GP OVN
Ori OVN
1.5 kb Quant ladder

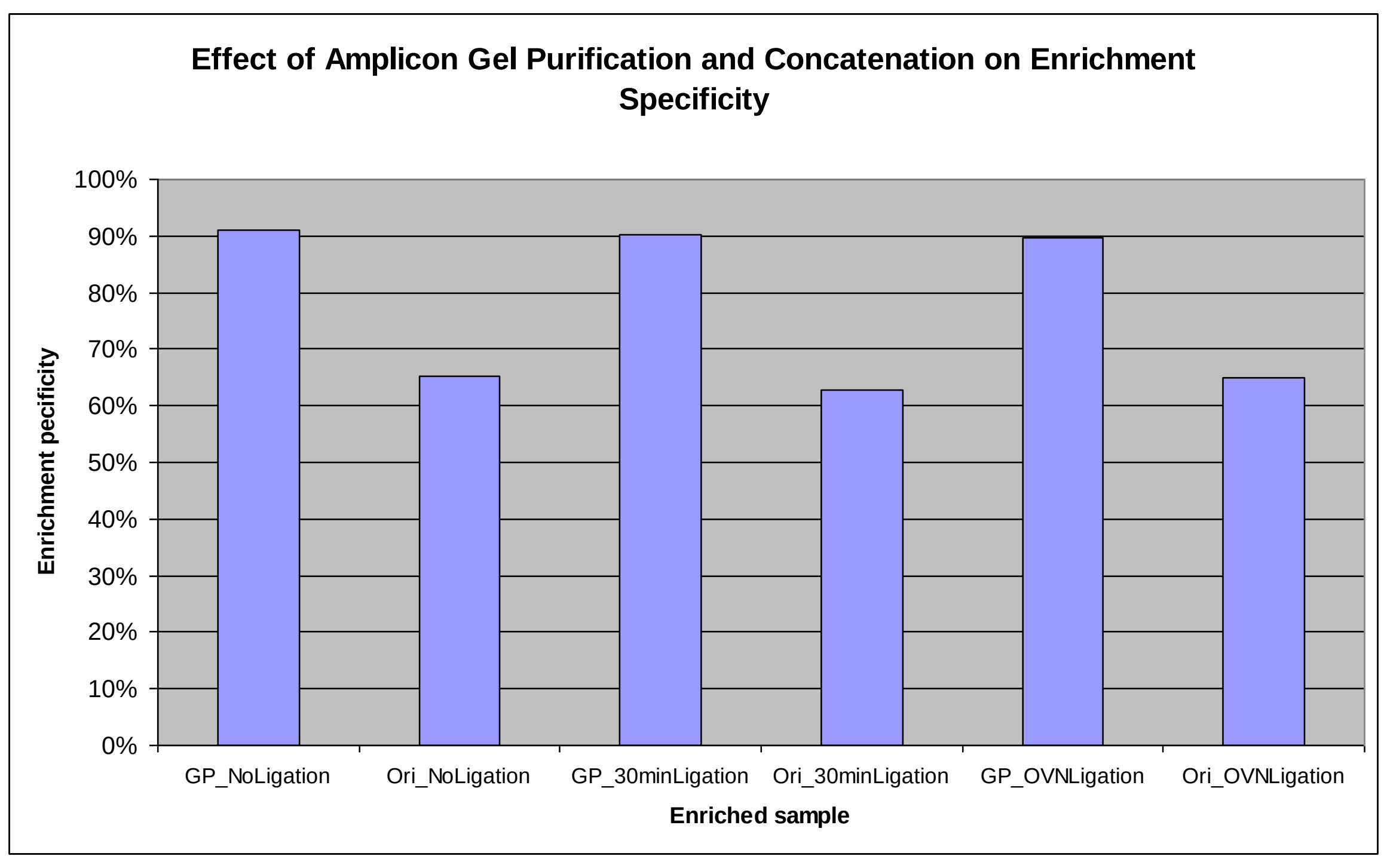
The workflow for effective post-enrichment sample preparation prior to library construction is shown (top panel). Comparison of gel purified (GP) versus non-gel purified (Ori) samples (middle panel); Electrophoretic analysis of concatenation of amplicons (bottom panel).

Figure 5. Percentage of Bases in the Targeted Regions Versus Depth of Coverage for All Enrichment Samples



Graphic representation of the percentage of uniquely mapped SOLiD™ reads aligning to the target region for each sample based on average fold coverage (A) or normalized coverage (B). Samples that were subjected to post-enrichment gel-purification (GP) and/or concatenation with a 30 min (30minLigation) or overnight (OVN) ligation step are indicated.

Figure 6. Effect of amplicon gel purification and concatenation on enrichment performance



Fraction of uniquely mapped reads for enriched libraries produced by microdroplet-based PCR that 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), and non gel-purified but concatenated overnight (Ori_OVNLigation) or for 30 min (Ori_30minLigation). Uniquely mapped reads represent those reads mapping to a single, unique location.

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

Table 1. SNP Calls for Enriched Samples Sequenced by the SOLiD™ System

	Number of calls* per sample					
	GP_NoLigation	Ori_NoLigation	GP_30minLigation	Ori_30minLigation	GP_OVNLigation	Ori_OVNLigation
SNP Classification						
Heterozygous True Positives	402	394	410	405	410	407
Homozygous True Positives	2424	2418	2436	2435	2438	2434
Heterozygous False Positives	5	5	3	3	4	4
Homozygous False Positives	15	21	8	12	8	11
Heterozygous Undercalls	13	19	6	10	6	9
Heterozygous Uncalled	1	3	0	1	0	0
Homozygous Uncalled	20	26	10	11	7	11
Total number of SNPs identified in enriched region	1487	1502	1487	1464	1521	1534
Novel SNPs	288	331	247	260	270	311
Concordance with dbSNP** in enriched regions	80.6%	78.0%	83.4%	82.2%	82.2%	79.7%

* Reported values are based on comparisons to the following version of HapMap: http://hapmap.ncbi.nlm.nih.gov/genotypes/latest/forward/non-redundant/

** Values based on comparison to dbSNP 130

Table 2. SNP Discovery Statistics in Enriched Samples

	GP_30minLigation		GP_OVNLigation	
	Homozygous SNPs	Heterozygous SNPs	Homozygous SNPs	Heterozygous SNPs
Specificity*ƒ (excluding undercalls)	98.1%	99.9%	98.1%	99.8%
Sensitivity*‡ (iexcluding undercalls)	99.9%	98.6%	99.8%	98.6%
Specificity*ƒ (including undercalls)	98.1%	99.9%	98.1%	99.8%
Sensitivity*‡ (including undercalls)	99.9%	98.6%	99.8%	98.6%

* Reported values are based on comparisons to the following version of HapMap: http://hapmap.ncbi.nlm.nih.gov/genotypes/latest/forward/non-redundant/

‡ Sensitivity = (True Positives)/(True Positives + False Negatives)

ƒSpecificity = (True Negatives)/(True Negatives + False Positives)

CONCLUSION

The SOLiD™ System and the Raindance Sequence Enrichment Solution provide a powerful PCR-based targeted resequencing solution for detection of genetic variation. Our results demonstrate highly specific PCR-based enrichment of the approximately 1.5 Mb target genomic region, based on the amplicon abundance and length (Figure 3B) as well as enrichment specificity for all samples (Figure 6).

Our sample preparation strategy sought to enhance the existing SOLiD™ System PCR-based targeted enrichment workflow, by evaluating the effect of amplicon gel purification and concatenation time of enrichment efficiency (Figure 4). Amplicon gel purification results in higher specificity of the reads to the desired target, as shown by comparison of unconcatenated, gel-purified (GP_No Ligation) and non gel-purified (Ori_No Ligation) mapping profiles (Figure 5A and 5B). The findings also show no significant improvement in target specificity when increasing the concatenation reaction time from 30 min to overnight (GP_30minLigation vs. GP_OVNLigation), but we can not rule out the possibility that other parameters, such as increasing the units of ligase, may further enhance concatenation efficiency. Optimal enrichment efficiency was achieved by post-enrichment gel purification of the amplification products followed by end-repair and concatenation performed prior to library construction (Figure 6).

For accurate detection of genetic variants, the extent of coverage for the target regions was assessed for all of the enrichment samples. When all 6 samples were evaluated as a function of average coverage, 96% or more of the target bases were covered by at least 1 read, while 91% or more of the target bases were covered by at least 30 reads (Figure 5A). Evaluation of all 6 samples as a function of normalized coverage produced similar results (Figure 5B). Coverage profile characteristics were highly reproducible under the same sample preparation conditions (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 detection of genetic variants.

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). For all samples, approximately, 20% of the SNPs identified in the targeted regions are novel.

Reported sensitivity and specificity values for the gel purified, concatenated samples are shown (Table 2), as these samples showed the greatest enrichment specificity. For homozygous and heterozygous SNPs for each samples, sensitivity and specificity values were greater than 98% and take into account the number of undercalls, or true heterozygous SNPs called as homozygous SNPs. In general, specificity and sensitivity values increase with additional sequence coverage. Enhanced sensitivity and specificity is expected with 99.99% accuracy of the SOLiD™ PI and 4*hq* Systems***.

The scalability and specificity of the Raindance microdroplet-PCR based enrichment method, combined with the throughput and accuracy of the SOLiD™ System portfolio, enable researchers to perform ultra-deep sequencing of specific regions of interest for rare variant discovery to better understand areas such as tumorigenesis, population diversity, microbial resistance, and disease susceptibility.

ACKNOWLEDGEMENTS

We thank Shan Yang for data analysis.

***These systems are under development and the specifications are subject to change.