Multiplex Exome Enrichment from Pooled Barcoded Libraries Yields Efficient SNP and Indel Detection on the SOLiD™ System

Gavin D Meredith, Christopher Adams, Gary Bee, Loni Pickle, Miroslav Dudas, Jennifer Kilzer, Adam Harris, Marie Callahan, George Marnellos, Yongming Sun, Vrunda Sheth, Clarence Lee,
Tanya Sokolsky, Chris Clouser, Kim Mather, Maryam Shenasa, Andrew Hutchinson, Jeffrey Ichikawa, Timothy Harkins, and Rob Bennett, Life Technologies Corp., 5791 Van Allen Way, Carlsbad, CA, 92008

ABSTRACT

The identification of genetic variation associated with human disease requires the development of a robust and cost-effective approach for systematic resequencing of candidate regions in the human genome. Even though the cost of sequencing a human genome continues to drop, the demand for increased sample throughput continues to increase. Higher sample throughput is considered necessary to enable larger patient cohort studies which hold the key to identifying rare disease-related alleles. Thus, scalable and automatable workflows for target enrichment and sequencing are needed to facilitate cancer and other genetic disease research. Described here is a targeted resequencing workflow that employs pooled barcoded fragment libraries, multiplexed exome enrichment, and multiplexed sequencing on the Applied Biosystems™ SOLiD™ System. To validate the performance of this multiplexed workflow, barcoded fragment libraries were made from HuRef gDNA using the new 5500 SOLiD™ fragment library protocol. Resulting libraries were then pooled in multiples of 4 for exome capture with the Agilent SureSelect™ Human All Exon 50 Mb Kit. The 4-plex data obtained from 2 quads of SOLiD™ 4 fragment sequencing yielded an average depth of coverage over the targets of 23.1X. The 8-plex data from a full slide yielded average depth of 29.6X. Overall, good barcode balance, similar mapping efficiencies and similar SNP/indel calls were observed for 4-plex and 8-plex exome capture samples. The percentage of on target reads varied from 71.2% to 74.0% which is comparable to numbers reported by others. For the 8-plex samples, the concordance of SNP calls (average of 31,474 SNPs) to dbSNP was 98.7% (sd=0.1%) for homozygous and 90.3% (sd=0.3%) for heterozygous variants and the concordance of small indels (average of 1560 indels) was 55.9% (sd=0.9%). Of particular note, sequencing a single (38 Mb) exome on a single lane of a 5500xl SOLiD™ System flow-cell yielded an average coverage of 66.9X (76.9% of target bases covered at >= 20X depth) with only 5.1% of target bases left uncovered. The combination of multiplexed exome enrichment and multiplexed on the SOLiD™ System provides an efficient and economical solution for the high-throughput detection of genetic variation in multiple human genomes.

INTRODUCTION

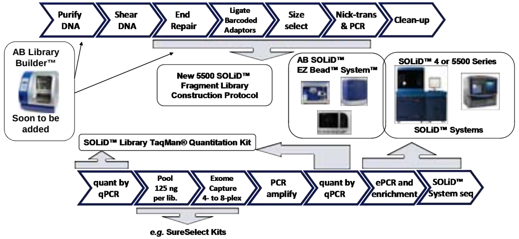
Next-generation sequencing technology has brought high throughput sample processing to genome sequencing, but an accompanying solution for high throughput target enrichment is still lacking. Target enrichment is a term used to describe the ability to selectively enrich and sequence specific regions of a genome. The method employed by the Agilent SureSelect Human All Exon 50 Mb Kit extracts target regions from genomic libraries by hybridization to in-solution biotinylated cRNA probes, or “baits.” Post-enrichment material is amplified and used directly for downstream steps, including emulsion PCR (ePCR) and sequencing on the SOLiD™ System (Figure 1). The inherent scalability and flexibility for automation of the SureSelect in-solution enrichment system coupled with the ultra-high throughput of the SOLiD™ sequencing platform provides an integrated approach to targeted resequencing. The new Agilent SureSelect Human All Exon 50Mb Kit builds upon previous exon products with additional validated novel content developed by the Wellcome Trust Sanger Institute. The new design encompasses coding exons annotated by the GENCODE project and also includes all exons annotated in the consensus CDS (CCDS – March 2009) databa. In addition, the content contains small non-coding RNAs from miRBase (v.13) and Rfam

MATERIALS AND METHODS

HuRef genomic DNA, purchased from the Coriell Institute for Medical Research, was fragmented to a mean length of ~200 bp with a Covaris® S2 System, then 3 µg amounts were used for library construction using a new protocol that included the use of 5500 SOLiD™ System compatible barcoded adaptors. After nick-translation, libraries were PCR amplified for 6 cycles, quantified and pooled in 4-plex (BC1-BC4) or 8-plex (BC1-BC8) using 125 ng of each library- based on Bioanalyzer estimates of average size and qPCR determinations of concentration of amplifiable molecules. The pooled 500 ng (for 4-plex) or 1 µg (for 8-plex) of library DNA was mixed with adaptor blockers, dried-down, and handled as described in the Agilent protocol, using 1X capture probes for 4-plex and 2X probes for 8-plex. After hybridization, capture, elution, and clean-up, the enriched libraries were amplified by 10 more cycles of PCR. Standard steps were taken thereafter to create enriched, templated beads for SOLiD System sequencing. The beads were sequenced as 50-color fragment tags (F3) and the data was progressively mapped in color-space and target enrichment and variant calling statistics were generated with the Targeted Resequencing pipeline in SOLiD™ BioScope™ 1.3 software.

RESULTS

Figure 1. Workflow for Multiplex exome capture with SOLiD™ System sequencing



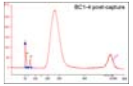
An exome enrichment workflow that permits pre-capture pooling of barcoded libraries and incorporates many of the most recent SOLiD™ System innovations has been developed. A working protocol has been established and may be available upon request.

Table 1. 4-plex library pooling prior to dry-down

Barcode	Nanodrop ng/uL	Bioanalyzer ng/uL*	qPCR ng/uL**	volume (uL) for 125 ng	pooled volume (uL)
BC1	73.2	82.3	43.1	2.9	12.7
BC2	79.1	45.4	40.8	3.1	
BC3	72.0	31.7	40.7	3.1	
BC4	69.1	35.7	35.2	3.6	

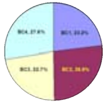
Different quantitative methods yield differing estimates of library yield. Best balance has been achieved by using the average library molecule size from Bioanalyzer* traces and SOLiD™ Library Taqman® Quantitative Kit qPCR** measurements (see Fig. 5 legend).

Figure 2. Exome-enriched 4-plex library



Bioanalyzer traces of library molecules after post-capture amplification confirm that an appropriate distribution is recovered.

Figure 3. 4-plex barcode representation



Good balance between barcodes is observed upon sequence analysis.

Table 2. 4-plex mapping and enrichment stats (SureSelect 50 Mb exome)

Barcode	Number of mapped reads	Percent on target	Fold Enrichment	Percent target bases not covered	coverage >=1x	coverage >=5x	coverage >=10x	coverage >=20x	average coverage depth
BC1	34,421,207	74.4%	44.7	15.6%	84.3%	71.6%	60.8%	42.2%	21.3
BC2	39,463,644	74.9%	45.0	14.9%	85.1%	73.5%	63.9%	46.9%	24.7
BC3	33,740,286	75.9%	45.6	15.9%	84.1%	71.3%	60.5%	42.0%	21.3
BC4	40,992,057	73.2%	44.0	15.4%	84.6%	72.5%	63.0%	46.9%	25.0
Total	148,617,194								
average	37,154,299	74.6%	44.8	15.4%	84.5%	72.2%	62.1%	44.5%	23.1
std dev	3,614,173	1.1%	0.7	0.4%	0.4%	1.0%	1.6%	2.8%	2.0

Table 3. 4-plex variant calls and dbSNP132 concordance (SureSelect 50 Mb exome)

Barcode	total SNPs	homo SNPs	homo dbSNP concord	het SNPs	het dbSNP concord	total indels	homo indels	homo dbSNP concord	het indels	het dbSNP concord
BC1	36364	15897	98.7%	20467	90.5%	1575	552	70.8%	1023	49.0%
BC2	37839	16127	98.4%	21712	90.9%	1647	587	70.4%	1060	50.8%
BC3	35819	15758	98.6%	20160	91.1%	1540	545	70.6%	995	49.4%
BC4	37160	15849	98.6%	21311	91.1%	1715	582	71.0%	1133	50.4%
average	36796	15908	98.6%	20913	90.9%	1619	567	70.7%	1053	49.9%
std dev	887	157	0.1%	722	0.3%	78	21	0.3%	60	0.8%

Figure 4. Pre-pooled 8-plex barcoded libraries

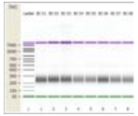
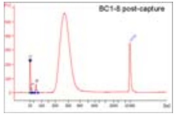


Table 4. 8-plex library pooling prior to dry-down

Barcode	Nanodrop ng/uL	Bioanalyzer ng/uL*	qPCR ng/uL**	volume (uL) for 125 ng	pooled volume (uL)
BC1	73.2	82.3	43.1	2.9	25.0
BC2	79.1	45.4	40.8	3.1	
BC3	72.0	31.7	40.7	3.1	
BC4	69.1	35.7	35.2	3.6	
BC5	77.2	41.3	40.6	3.1	
BC6	78.8	60.0	46.5	2.7	
BC7	75.6	49.8	38.0	3.3	
BC8	74.2	44.4	39.7	3.2	

The protocol yields libraries of consistent size and concentration. Average library molecule size based on Bioanalyzer* traces and molar concentration by qPCR are used to determine the “qPCR ng/uL”**; this value is used to apportion 125 ng of each library into the pool.

Figure 5. Exome-enriched 8-plex library



Bioanalyzer traces of library molecules after post-capture amplification confirm that an appropriate distribution is recovered.

Table 5. 8-plex mapping and enrichment stats (SureSelect 50 Mb exome)

BarCode	Number of mapped reads	Percent on target	Fold Enrichment	Percent target bases not covered	coverage >= 1x	coverage >= 5x	coverage >= 10x	coverage >= 20x	average coverage depth	Balance
BC1	48,326,122	72.5%	43.6	22.4%	77.6%	62.4%	54.4%	44.1%	28.8	12.1%
BC2	53,597,196	72.5%	43.5	21.5%	78.5%	63.6%	55.8%	45.9%	31.9	13.4%
BC3	45,006,332	74.0%	44.4	23.1%	76.9%	61.8%	53.9%	43.3%	27.3	11.3%
BC4	53,581,754	71.2%	42.7	22.4%	77.6%	62.3%	54.4%	44.7%	31.3	13.4%
BC5	54,065,056	72.4%	43.5	21.3%	78.7%	64.0%	56.2%	46.3%	32.1	13.5%
BC6	42,426,062	72.3%	43.4	23.6%	76.4%	60.3%	51.9%	40.9%	25.2	10.6%
BC7	52,462,468	71.8%	43.1	21.9%	78.1%	62.9%	55.0%	45.1%	30.9	13.1%
BC8	49,969,589	72.3%	43.4	22.0%	78.0%	62.7%	54.6%	44.4%	29.7	12.5%
total	399,433,579									
average	49,929,197	72.4%	43.5	22.3%	77.7%	62.5%	54.5%	44.3%	29.6	12.5%
sd	4,365,958	0.8%	0.5	0.8%	0.8%	1.1%	1.3%	1.7%	2.4	1.1%

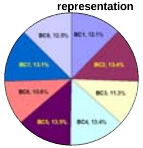
Very similar enrichment statistics were obtained from a full slide (half a run) of SOLiD™ 4 System sequencing on the 8-plex simultaneous exome enrichment sample as compared to those obtained from 2 quads (~40% of a slide) of sequencing on a 4-plex reaction (compare to Table 2). There may have been a minor degree of complexity loss upon scaling to 8-plex based on a slightly lower “on-target” rate and a larger percentage of “target bases not covered”. Nonetheless, both samples have nearly identical numbers of bases (~44.4%) that are covered at 20X depth or greater.

Table 6. 8-plex variant calls and dbSNP132 concordance

BarCode	total SNPs	homo SNPs	homo dbSNP concord	het SNPs	het dbSNP concord	total indels	indel dbSNP concord
BC1	31,437	12,964	98.7%	18,473	90.3%	1590	54.7%
BC2	32,136	13,080	98.7%	19,056	90.3%	1609	55.4%
BC3	30,797	12,799	98.6%	17,998	90.3%	1508	56.0%
BC4	31,218	12,831	99.0%	18,387	90.8%	1552	56.1%
BC5	32,452	13,213	98.7%	19,239	90.2%	1628	56.0%
BC6	30,566	12,854	98.7%	17,712	89.6%	1486	56.5%
BC7	31,724	12,938	98.9%	18,786	90.4%	1614	57.7%
BC8	31,464	13,017	98.7%	18,447	90.5%	1589	55.1%
average	31,474	12,962	98.7%	18,512	90.3%	1559.5	55.9%
std dev	633	139	0.1%	511	0.3%	52.3	0.9%

Again, the variant calls of from this 8-plex exome capture compare well to the 4-plex capture overall (see Table 3), particularly in degree of concordance with dbSNP and number of indels called (96.3% as many); however, there are somewhat fewer total SNPs called (81.5% as many homozygous SNPS and 88.5% as many heterozygous SNPs). Taken along with the differences in enrichment statistics, this suggests, not surprisingly, that there is a trade-off made between the degree of multiplexing one can do and the degree to which one confidently identify variants with exome enrichment.

Figure 6. 8-plex barcode representation



Excellent balance between barcodes is observed upon sequence analysis. There is very little evidence that representation is influenced by the barcodes.

Table 7. SOLiD™ 4 and 5500xl System enrichment on a single exome (SureSelect 38 Mb)

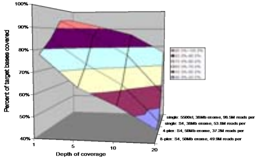
	Total Beads	Percent mapped	Percent on target	fold enrichment	Percent target bases not covered	coverage >=1x	coverage >=5x	coverage >=10x	coverage >=20x	average coverage
SOLiD 4 (1 quad)	83,096,814	64.8%	72.0%	57.9	6.0%	94.0%	86.2%	76.0%	56.9%	33.2
5500xl (1 lane)	123,404,339	79.8%	73.5%	59.2	5.1%	95.0%	91.1%	86.7%	76.9%	66.9

Table 8. SOLiD™ 4 and 5500xl System variant calls on a single exome (SureSelect 38 Mb)

	total SNPs	total homo SNPs	homo dbSNP concord	total het SNPs	het dbSNP concord	total indels	total homo indels	homo indel dbSNP concord	total het indels	het indel dbSNP concord
SOLiD 4 (1 quad)	28,681	12,161	97.9%	16,520	89.3%	600	215	89.8%	385	71.2%
5500xl (1 lane)	32,619	12,433	99.1%	20,186	89.9%	1,406	421	87.7%	985	63.9%

The increased throughput of the 5500 Series SOLiD™ Sequencers permits more exome data to be obtained per run. As an example, one lane of 5500xl sequencing (there are 6 lanes per flow-cell) yields an average coverage (>66X for a 38Mb version of the exome) that is approximately twice that of a quad of SOLiD™ 4 sequencing. This added depth permits more SNPs (~4,000) and indels (~800) to be called as well.

Figure 7. Landscape of exome target base coverage for all data shown



A 3-dimensional surface has been plotted for the percent of target bases covered and the depth of coverage for each of the exome datasets described. This limited study suggests that there is a trade-off between the degree of pre-capture SureSelect exome multiplexing that can be done and the depth of target coverage that will result. This suggests that the enrichment protocol can be further optimized. The direct comparison of SOLiD 4 sequencing and 5500xl sequencing suggests that the higher throughput provided by the new instrument can significantly compensate for this. Multiplexed exome sequencing on 5500 Systems holds great promise.

CONCLUSIONS

A) Barcoded SOLiD™ System libraries can be pooled prior to exome enrichment with the Agilent SureSelect Human All Exon 50Mb Kit; 4-plex and 8-plex simultaneous capture is possible.

B) Multiplex capture yields reproducible results. Good barcode balance is observed and SOLiD™ System barcodes 1-8 do not bias performance.

C) The 5500 Series SOLiD™ System yields the largest amount of high-quality data observed for single exome sequencing. Sequencing 2 exomes per 5500 flow-cell lane can yield >30X average coverage for the SureSelect 38 Mb exome. This is a throughput of ~24 exomes per 5500xl run.

D) Multiplexing exome capture leads to gains in throughput that are balanced against depth of coverage. A full slide of SOLiD™ 4 System sequencing will yield an average depth of coverage of ~30X for 8-plex and >50X for 4-plex for the SureSelect 50Mb exome.

REFERENCES

1. http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/targeted-resequencing.html
2. Hoischen A, et al. Nat Genet. 2010 Jun;42(6):483-85. Epub 2010 May 2.
3. Teer JK and Mullikin JC. Hum Mol Genet. 2010 19:R145-151. Epub 2010 Aug 12.
4. Bell CJ, et al. Sci Transl Med. 2011 3(65):1-14
5. Bainbridge MN, et al. Genome Biol. 2010;11(6):R62. Epub 2010 Jun 17.

TRADEMARKS/LICENSING

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