Comparison of Whole Exome Capture Products – Coverage & Quality vs Cost

B Marosy, J Gearhart, B Craig, KF Doheny

Center for Inherited Disease, Johns Hopkins Genomics, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD

Introduction

Next Generation Sequencing technology has been a driving force for reducing sequencing costs and enhancing genome technology, allowing investigators and clinicians the opportunity to better delineate an individuals' genetic makeup. For facilities providing genetic services, maintaining the balance of generating high quality data at the lowest cost is a continuous effort requiring recurrent evaluation and optimization of protocols. In preparation for higher throughput of exome sequencing using the NovaSeq, we evaluated uniformity, coverage and cost of sequencing to achieve 90% on target at 10x, 20x and 30x depth across four different exome capture products:

- SureSelect Human All Exon v7 (Agilent Technologies)
- xGEN Exome Research Panel v1.0 (Integrated DNA Technologies)
- Prime Exome (Roche Sequencing)
- Human Core Exome (Twist Bioscience)

Methods – Library Preparation

To reduce experimental variables and minimize bias between libraries, all libraries were prepared using the Kapa Hyper prep library prep kit. Table 1 summarizes the parameters used. 50ng of DNA from 8 different HapMap samples was sheared using the Covaris LE220. End Repaired/A-Tailed libraries were ligated with IDT dual indexed adapters and amplified using the Kapa HiFi DNA Polymerase for 8 cycles. Agilent captures were hybridized as single sample reactions using 750ng of library as input. IDT, Roche and Twist captures were hybridized as pools of 8 samples using 750ng, 625ng, 187.5ng of library input, respectively. All Hybridization and Post-hyb capture & washes were performed according to each respective manufacturer's protocol. Post Hyb PCR was performed using the Kapa HiFi DNA Polymerase for 10 cycles, regardless of capture product.

Table 1 – Library Prep Methods

Vendor	DNA Input	Library Prep	Pre Capture PCR Cycles	Capture Pooling	Hyb Input	Post-Capture PCR Cycles	Post-Capture PCR Enzyme
Agilent	50 ng	Kapa Hyper Prep	8	Single	750 ng	10	Kapa HiFi
IDT	50 ng	Kapa Hyper Prep	8	8-plex	750 ng	10	Kapa HiFi
Roche	50 ng	Kapa Hyper Prep	8	8-plex	625 ng	10	Kapa HiFi
Twist	50 ng	Kapa Hyper Prep	8	8-plex	187.5 ng	10	Kapa HiFi

Methods – Sequencing & Data Analysis

Captured libraries were clustered and sequencing was performed on the Illumina HiSeq2500 platform using on-board clustering and SBS chemistry with either 2x100 or 2x125 read lengths. Sequencing data was trimmed to 100bp, if applicable, then downsampled (Picard; DownsampleSam) to 5 Gigabases (Gb) of raw data yield and three different BED files (UCSC coding exons, vendor specific, intersection of all 4 vendor BED files) were applied independently to ensure equal comparison between methods and capture content. Samples that fell below 5 Gb were processed 'as is' but removed from the final analysis. Subsequent data analysis was performed using CIDRs in-house Research Whole Exome pipeline.

Results – Design Coverage

Coverage summaries were generated using Galaxy v1.0.0 coverage tool and summarized in Table 2. Each vendors target BED file and the 4-way intersection was used to determine design coverage across the UCSC Coding exons by region (exonic coordinates) and by base. The total target footprint for each vendor is also listed.

Table 2 - UCSC Exon Coding Coverage Summary by Capture Product

UCSC Coding Exon (35.20 Mb) BED File Coverage Summary	Agilent (49.48 Mb)	IDT (50.84 Mb)	Roche (37.24 Mb)	Twist (33.05 Mb)	4-way Intersection (26.91 Mb)
% Regions not covered	1.74	3.72	5.76	5.20	6.69
% Regions covered <50%	1.03	0.46	0.89	0.67	1.93
% Regions covered >50% & <100%	38.19	3.35	3.97	2.71	37.97
% Regions covered 100%	59.04	92.47	89.38	91.43	53.42
% Bases not covered	7.77	5.02	7.89	6.49	13.95
% Bases covered	92.23	94.98	92.11	93.51	86.05

Results – Sequencing Data QC

Table 3 compares QC metrics generated during sequencing when using vendor specific target BED files. Data quality (TiTv, Percent SNV OnTarget SNP138 and Count SNV OnTarget) across captures is consistent. Insert size, duplication rates and library complexity vary between capture products when using the same library prep methods and amplification, indicating that capture also impacts these metrics.

Table 3 – Sequencing QC metrics based on vendor specific target BED files.

Sequencing QC Metric	Agilent	IDT	Roche	Twist
n = Downsampled to 5 Gb	6	3	4	3
Percent Selection	73.0	87.0	85.5	82.7
Percent Duplication	4.1	4.3	7.6	2.7
Estimated Library Size (millions)	317	316	157	429
Mean Insert Size	250	241	287	216
Percent Total Concordance	99.75	99.82	99.72	99.60
Sensitivity 2 Het	98.82	99.13	99.24	99.38
Percent SNV on Target dbSNP138	98.55	99.14	99.35	99.23
Count SNV on Target	41072	39114	27483	26453
Known TiTv Ratio	3.07	3.07	3.09	3.11
Known TiTv Count	21668	20370	21261	22743

Results – Target Coverage

Figure 1 plots coverage vs percent on Target by 10x, 20x and 30x trellised by BED file. At 10x coverage all captures are effective in capturing targets >90%, with only 5 Gb of sequencing data. As depth increases, with the same amount of Gb, coverage at 90% drops across products. Captures with higher efficiency are able to maintain coverage as depth increases with the same amount of sequencing yield.

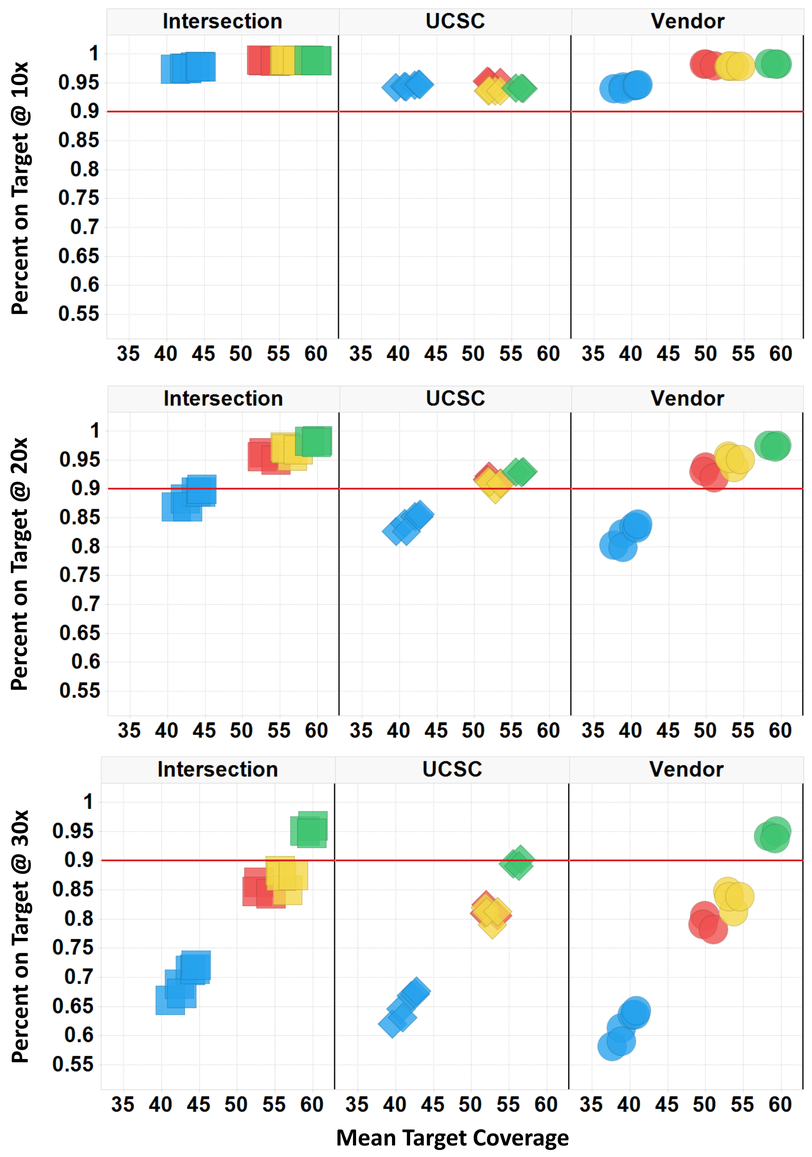

Figure 1a-c: Plots Mean Target Coverage (x-axis) vs Percent on Target (y-axis) with increasing depth (10x=top; 20x=middle; 30x=bottom) for each sample. All samples are downsampled to 5 Gb and each panel is trellised by BED File (4-way Intersection = left; UCSC coding exons = middle; vendor specific targets = right). Points are colored by Capture product (Blue=Agilent, Red=IDT, Yellow=Roche, Green=Twist).

Results – Coverage Uniformity

Coverage uniformity was analyzed in several ways. Figure 2 represents sample statistic summaries generated from GATK - Depth of Coverage (DOC). UCSC exon coding and vendor specific BED files were used, however, the 4-way intersection BED file results were comparable to the UCSC plot and not shown here. Tight/narrow peaks at the highest depth represent captures with the most uniformity. In addition to DOC metrics, we analyzed GC/AT dropout which measures how under covered a low/high %GC region is relative to the mean (Figure 3). Regions of low coverage in clinically relevant genes were viewed in IGV and are shown in Figure 4.

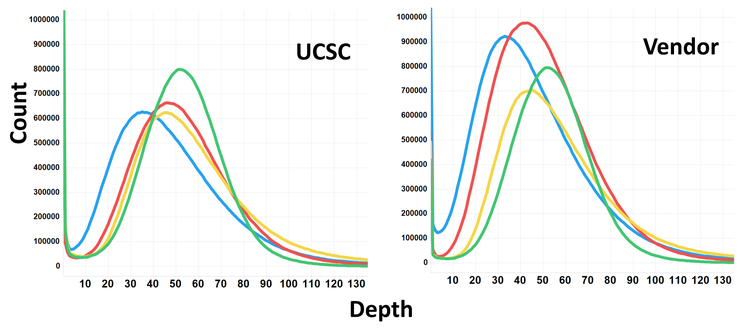

Figure 2: Depicts uniformity across capture products. Data generated from Depth of Coverage is plotted as the count of bases (y-axis) at a given depth (x-axis) for all bases captured from NA12878. Each panel varies by BED File. Lines are colored by capture product (Blue=Agilent, Red=IDT, Yellow=Roche, Green=Twist).

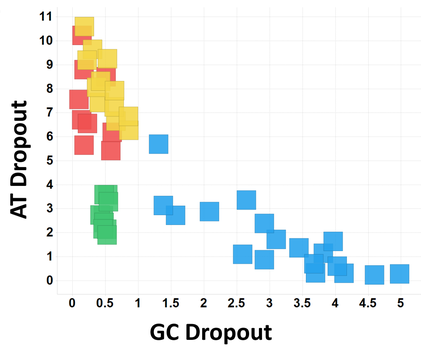

Figure 3: GC dropout (x-axis) vs AT dropout (y-axis) as generated by Picard, which measures how under covered a low/high % GC region is relative to the mean. Points are colored by Capture product (Blue = Agilent, Red = IDT, Yellow = Roche, Green = Twist). Vendor specific target BED files were used.

Results – Pooling Balance

Obtaining 'even' pooling is a critical step in the pre-capture workflow. If a sample fails at sequencing it is more difficult to re-queue since samples which may require additional sequencing cannot be broken out of the pool. This would require re-hybridization of the sample, thus increasing sample costs. Figure 5 depicts the pooling evenness observed between post-capture and pre-capture pooling. Post-capture pooling more consistently produces even pools than pre-capture pooling.

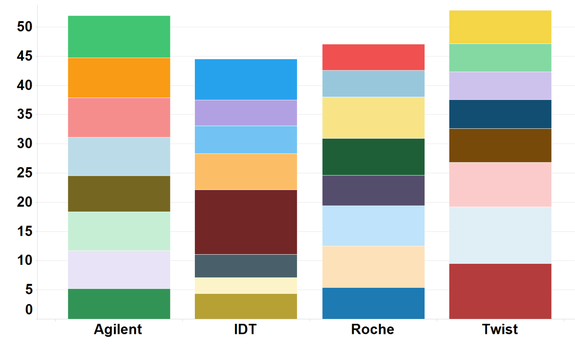

Figure 5 – Pooling Balance. Compares the 'evenness' of pools in the sequencing data between samples (prior to downsampling) that underwent pre-capture pooling (IDT, Roche & Twist) vs samples that were post-capture pooled (Agilent). Colored stripes within each bar denote an individual sample within the pool, size of each bar represents amount of raw sequencing data in Gb.

Results – Coverage Uniformity cont'd

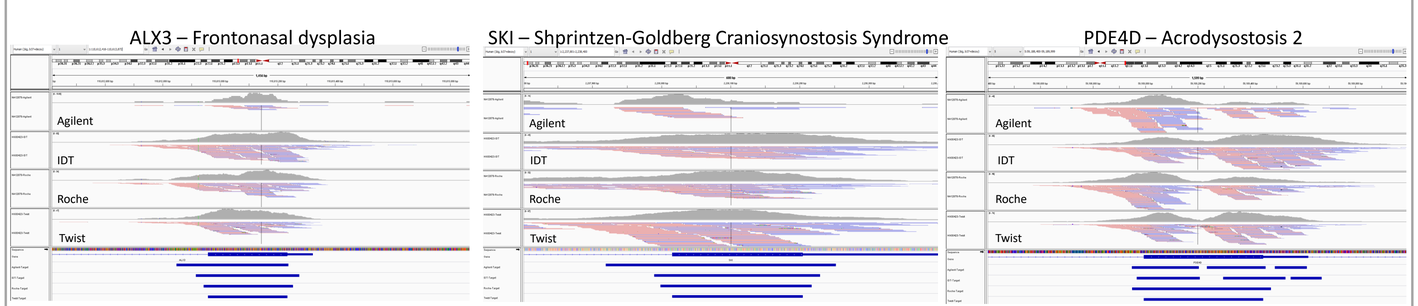

Figure 4a-c : Integrative Genomics Viewer (IGV) plot of NA12878 at regions of low coverage identified within clinically relevant genes. In lower panel: gene track, target BED files for Agilent, IDT, Roche and Twist, respectively.

Discussion & Conclusions

- Captures with higher efficiencies are able to maintain coverage as depth increases with the same amount of sequencing yield. Determining which capture product to use may depend on project specific needs (i.e. coverage of specific targets, depth requirements, available automation etc.) and cost requirements.
- Library Prep Methods – for purposes of this experiment to minimize bias, we utilized a single library prep method that has been optimized in our facility. Vendor specific library prep reagents/protocols may provide additional optimizations that are not addressed here.
- GC/AT rich & Low coverage regions – PCR enzyme performance is typically considered a 'culprit' of these difficult to sequence regions. However, capture products may also influence coverage of these regions.
- Pre-capture pooling can provide a ~2.5 fold reduction in reagent costs. However, the unbalanced pools obtained post capture increases the redo rates as much as ~30%. For higher through-put facilities, this adds considerably to the labor costs and overall time to completion for projects. Capture products that maintain high efficiency will decrease the amount of sequencing required and when combined with the lower cost of sequencing on the NovaSeq will offset the added cost of pre-capture pooling making this a more feasible workflow.