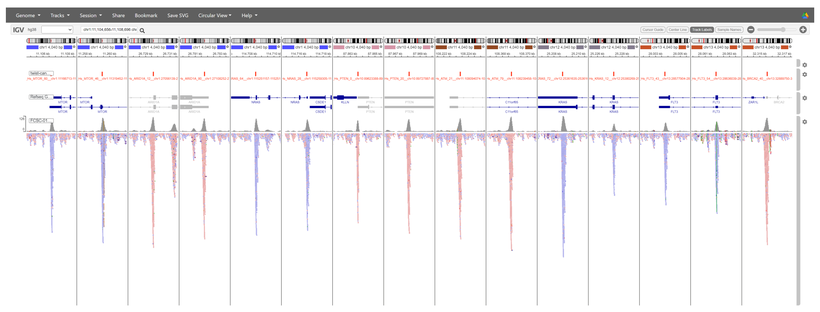
Sensitive analysis of clinically actionable genomic regions with an integrated target enrichment and sequencing workflow

Shawn Levy[1], Kelly Blease[1], Sophie Billings[1], Bryan Lajoie[1], Junhua Zhao[1], Adeline Mah[1], Laurie Edoli[1], Conner Thompson[1], Sinan Arslan[1], and Mike Previte[1]

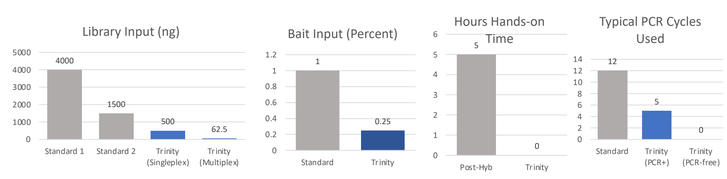
[1]Element Biosciences, San Diego, CA, 92121

Abstract and Introduction

Over the last 15 years, hybridization-based approaches to enrich genomic regions of interest have enabled efficient and sensitive sequencing methods. From sequencing the exome for disease variant discovery to targeting small numbers of genes at high depth in somatic applications, many options exist for translational and clinical analysis. As sequencing has become less expensive, unbiased approaches like genome-wide sequencing have become common in precision health applications. Although genomes have high value and high impact, there are many applications that benefit from targeted enrichment approaches to increase sensitivity in challenging regions of the genome or to improve sensitivity to low-frequency variants, particularly in somatic applications. A significant challenge to targeted workflows is the additional technical burden and time of the hybridization workflow. These steps add several hours to a day to experimental workflows and can be challenging to automate. To address these challenges and create a highly efficient and dynamic workflow, we have created an integrated, on-flowcell target enrichment capability integrated into the standard sequencing workflow on the Element AVITI platform. This workflow leverages unique surface chemistry and does not add time or steps to the sequencing process. It also does not require any library preparation modifications. A single library preparation workflow can be used with a wide variety of target enrichment experiments. We demonstrate efficient target enrichment over regions of interest in oncology and exome. With tunable enrichment levels from 300-fold to over 1,200-fold, our platform can achieve less than 1% allele sensitivity in target regions while also providing either high specificity or uniform background coverage over the sample library at 4x to 30x levels, depending on sequencing depth. Combined, the target enrichment capabilities allow for focused and efficient rare variant detection while uniform background coverage enables variant discovery across a broad range of variant classes, including CNV. Since the enrichment methods are library preparation independent, libraries that allow more complex analyses such as structural variant detection from Hi-C libraries can be used in the protocol. These capabilities allow for advancements using this technology in numerous applications such as cell-free DNA, infectious disease, somatic variant detection and residual disease monitoring.

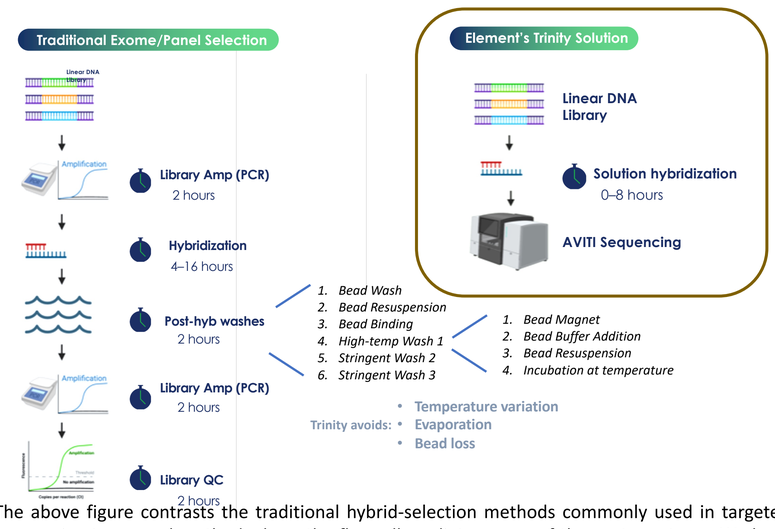


Analysis of a small targeted panel of oncology genes. This IGV plot shows the post-sequencing enrichment levels of 76 targets commonly sequenced in oncology studies. The color of the reads are red and blue representing forward and reverse reads. These targets were assessed in a strand-specific manner to determine the performance of an on-flowcell enrichment protocol. The variant calling performance of this data on a reference control sample is shown below in the Variant Calling Performance section.

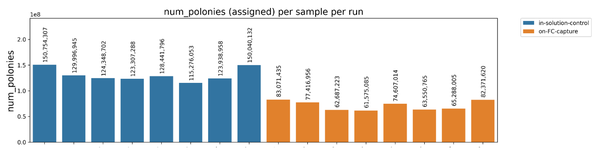


Compared to traditional hybridization-based targeted sequencing methods, the on-flowcell methods described here have relative advantages in library input, bait amount used per hybridization, hands on time, and the number of PCR amplification cycles used in the overall assay. The lower library input and the lower bait amounts provide assay flexibility and cost savings when combined with the lower hands-on time. Avoiding a second round of PCR typical of the bead-based methods appears to increase the complexity of the library as well as retain a slightly larger insert size.

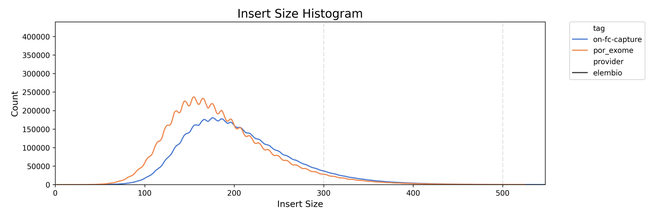
Results



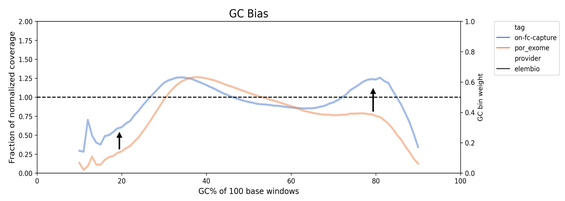
The above figure contrasts the traditional hybrid-selection methods commonly used in targeted sequencing to a novel method where the flowcell mediates many of the capture steps. We have developed a flexible and high-performing method to capture library molecules of interest as part of the sequencing process, eliminating some of the more complex and sensitive steps involved in the targeted sequencing process.



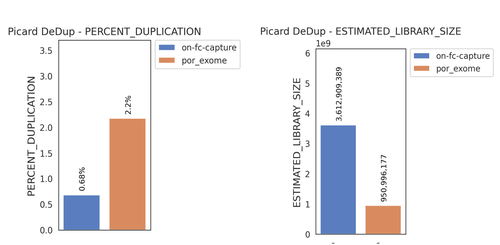
Since multiplex hybridization is a common feature of hybrid selection workflows, we verified that the relative distribution of samples and libraries was not different between an in-solution workflow (blue) compared to the on-flowcell method (orange).



The alignment-based insert size is observed to be slightly larger for the on-flowcell method compared to in-solution methods. This is likely due to the lower PCR cycles and less size selection on the final library for the on-flowcell method.



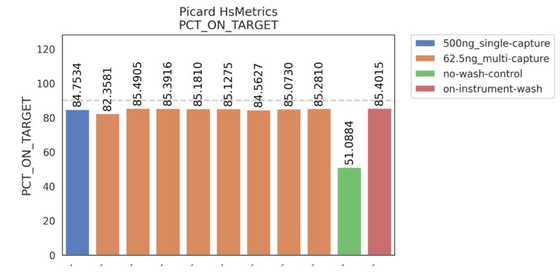
The alignment-based insert size is observed to be slightly larger for the on-flowcell method compared to in-solution methods. This is likely due to the lower PCR cycles and less size selection on the final library for the on-flowcell method.



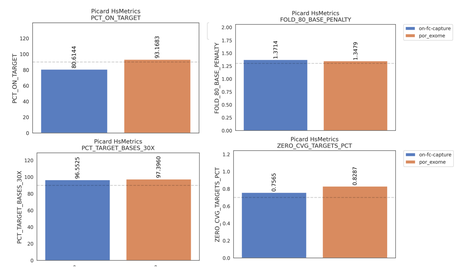
Library duplication rate and library complexity were substantially different improved in the on-flowcell method. The lower duplication rate and higher complexity of the library should allow deeper sequencing and sensitive variant detection for mosaic or somatic variants.

Sample	Duplicate (%)	Zero Target Cov (%)	Bases at 30x (%)	Bases at 50x (%)
Single Capture	1.1675	0.7498	94.4569	72.0591
Multiplex 1	4.2659	0.7491	94.6997	70.6179
Multiplex 2	4.8288	0.747	92.8638	68.042
Multiplex 3	4.8507	0.7403	91.9422	66.5476
Multiplex 4	4.5058	0.7396	92.5965	67.1989
Multiplex 5	4.7241	0.7424	95.253	72.7907
Multiplex 6	4.6502	0.7452	93.93	69.9293
Multiplex 7	4.6811	0.7417	93.177	68.1534
Multiplex 8	4.7808	0.7519	92.6367	67.78
No Wash Control	0.3659	0.7413	65.823	12.5444
On-Instrument Wash	0.8681	0.7234	95.759	69.9207

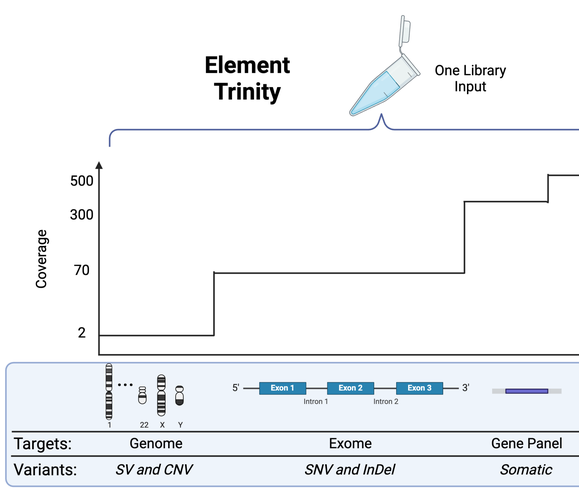
The above table summarizes an experiment using low-input library amounts in a multiplex experiment. 62.5ng of final library was used in each multiplex sample compared to 500ng of input library for the Single Capture sample. Performance and duplication rates remain acceptable with the lower library input.



The above figure compares the percent of reads on-target for a 500ng single plex capture to a low-input multiplex capture. All samples had an on-target rate of greater than 80%. This is slightly lower than the typical on-target rate for an in-solution hybridization of approximately 90%. All analysis was performed with 6Gb data at PE 75 (40M read pairs).

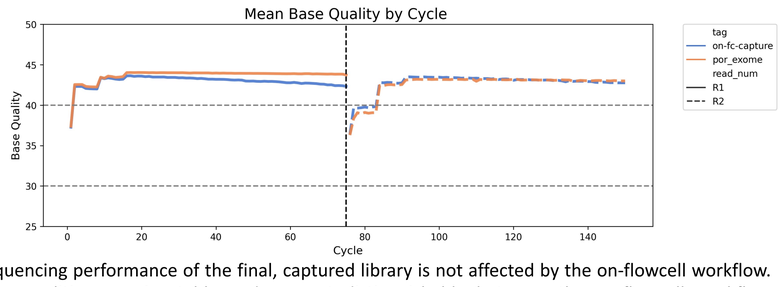


The above figures summarize performance statistics for the same library used in a standard in-solution workflow (orange) compared to that library in the on-flowcell workflow (blue). The on-target rate for the on-flowcell workflow is slightly lower compared to the in-solution workflow, as was observed in the multiplex experiment above. However, the fold-80 and bases at 30x were similar. The zero coverage metric was slightly better for the on-flowcell workflow. All analysis was performed with 6Gb data at PE 75 (40M read pairs).



Conclusions

- **This study demonstrates an improved workflow and performance for hybridization-based sequencing approaches using routinely available reagents. The workflow removes post-hybridization and bead handling steps typical of target enrichment protocols.**
- **The workflow removes the post-hybridization PCR step, avoiding additional amplification and potential reduction of library complexity.**
- **Duplication rate is significant;ly lower and library complexity higher in the Trinity workflow compared to traditional processing.**
- **GC bias is substantially less in the on-flowcell workflow in these experiments. However, there has been some variance observed in this statistic so further replication in needed.**
- **Variant calling is improved over a range of GC content, particularly for InDel variants**

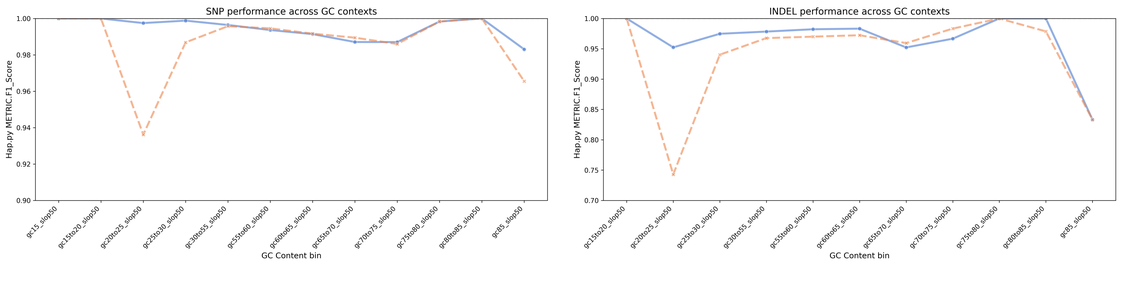


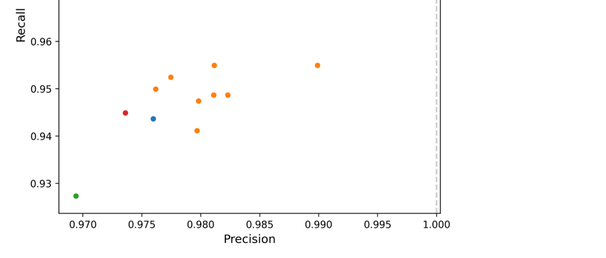
Sequencing performance of the final, captured library is not affected by the on-flowcell workflow. Base qualities remain stable and are not distinguishable between the on-flowcell workflow compared to the standard bead-based methods.

Variant Calling Performance

Probe ID	Genomic Position (GRCh38)	Reference	Alternate	Observed Allelic Frequency (%)	Expected Allelic Frequency (%)	Coverage	Variant Type	Gene
NRAS_NC_000001.11:114713818-114713908_21-61	114713909	G	T	11	12.5	64.80	SNP	NRAS
PIK3CA_NC_000003.12:179218304-179218394_1-41	179218303	G	A	7	9	516.35	SNP	PIK3CA
PIK3CA_NC_000003.12:179234206-179234296_23-63	179234297	A	G	15	17.5	722.68	SNP	PIK3CA
KIT_NC_000004.12:54733156-54733246_35-75	54733155	A	T	11	10	475.50	SNP	KIT
EGFR_NC_000007.14:55173923-55174013_21-61	55174014	G	A	17	24.5	97.05	SNP	EGFR
EGFR_NC_000007.14:55174681-55174771_7-47	55174772	AGGAATTAAGAGAAGC	A	1.20	2	904.30	DEL	EGFR
EGFR_NC_000007.14:55181379-55181469_20-60	55181378	C	T	1	1	270.07	SNP	EGFR
EGFR_NC_000007.14:55191823-55191913_44-84	55191822	T	G	3	3	401.10	SNP	EGFR
BRAF_NC_000007.14:140753245-140753335_40-80	140753336	A	T	12	10.5	680.30	SNP	BRAF
KRAS_NC_000012.12:25245256-25245346_50-90	25245347	C	T	19	15	~141.00	SNP	KRAS
KRAS_NC_000012.12:25245259-25245349_48-88	25245350	C	T	8	6	~141.00	SNP	KRAS

The above table details the observed versus expected variant calling in a reference control sample for somatic variation. The observed versus expected variant frequencies were very close across a range of variant levels from 1% to 25%, including one deletion variant.





SNP and InDel variant calling performance in human reference sample (Hg001) is summarized above by Precision x Recall for the low-input multiplex experiment compared to a single plex experiment. SNP and InDel statistics indicate highly accurate and consistent variant calling, especially for SNP variants.

The two figures to the left detail SNP and InDel F1 rates based on GC content bin. The same library was analyzed with the on-flowcell workflow and standard in-solution workflow. All variant calling and analysis was performed with 6Gb data at PE 75 (40M read pairs). The standard bead-based methods showed lower variant calling accuracy at the extreme ends of the GC range, particularly in the low GC range. Additional experiments are underway to better characterize and validate these observations.

Visit **elementbiosciences.com** for more information.