# Combining NGS technologies for biomarker identification and confirmation research

## ABSTRACT

Advances in next generation sequencing (NGS) technologies have provided immense research opportunities for identification of molecular markers of disease. Highly parallel and multiplex biomarker characterization in disease areas such as cancer can be achieved by combining multiple targeted NGS enrichment approaches, including hybridization capture and amplification-based sequencing using multiplex PCR. These complementary, independent techniques can provide a high level of sensitivity and specificity when trying to detect low levels of variant allele frequencies (VAF) in somatic cancer detection.

IDT has developed 2 unique, high-performing targeted NGS approaches* to maximize researchers' ability to confidently identify and confirm biomarkers in germline and somatic mutations. Here, we describe a follow-up, supplemental study to a previously published application note [1]. Biomarkers identified in cell-free DNA (cfDNA) lung cancer samples show high correlation using the IDT xGen™ Prism Library Kit and Lockdown™ Probe system for target capture and the IDT rhAmpSeq™ amplicon sequencing system. Both systems showed sensitivity down to 1% VAF in cfDNA samples starting from low quantities of DNA input, with the xGen hybrid capture system slightly outperforming the rhAmpSeq system at identifying somatic mutations. The benefits and caveats of each approach are discussed.

## INTRODUCTION

In our earlier study [1], we described how NGS sample preparation with the use of the novel IDT xGen Prism Library Kit and Lockdown Probe capture system can provide a powerful research tool for cancer biomarker discovery and molecular profiling. Using a set of matched normal tissue DNA, formalin-fixed, paraffin-embedded (FFPE) tumor DNA, and cfDNA samples, we showed how the hybridization capture system can lead to high-quality somatic and germline mutation analysis in potentially challenging lung cancer trio samples. The ability of that approach to increase the conversion rate of input DNA molecules to sequenceable libraries provides a significant advantage over other traditional library preparation methods. Nevertheless, such discoveries of previously unknown or uncharacterized biomarkers often need additional confirmation using independent orthogonal techniques.

Whereas capture-based targeted NGS approaches traditionally fish out DNA sequence targets using hybridization probes (thus, sometimes called "baits"), amplification-based approaches use PCR primers and a multiplexed PCR reaction to amplify stretches of DNA and add sequencing adapters to characterize the insert sequence between the primers. The IDT rhAmpSeq technology was developed as a robust amplicon sequencing system, and we demonstrate its effectiveness alongside the xGen capture solution as a complementary approach for highly

correlated results of biomarker characterization. While the overall workflow of the rhAmpSeq system has advantages in its ease of use and cost-effectiveness, the xGen capture workflow has its own advantages with regards to its ability to use unique molecular identifiers (UMIs) for accurate quantification of low-frequency allelic variants.

We studied VAFs as low as 1% in previously uncharacterized cfDNA samples obtained from a commercial vendor and derived from individuals with lung cancer. This cfDNA was part of 3 sample trios (i.e., normal tissue, FFPE tumor, and cfDNA) that were used in the xGen capture study [1]. For this study, we prepared the same cfDNA samples with our rhAmpSeq amplicon sequencing solution, compared the results across FFPE and cfDNA using both hybrid capture and amplicon sequencing, and observed a very high correlation of identified mutations. The variants that were identified in both xGen and rhAmpSeq experiments ranged from 96 to 98% across the 3 cfDNA sample types compared to the FFPE tumor samples from the same individuals assessed previously with the xGen target capture method. As expected, high-frequency mutations such as homozygous or heterozygous variants were identified in both systems at nearly 100%, with only a handful of low-frequency (i.e., somatic) mutations uniquely identified in either xGen- or rhAmpSeq-prepared samples. Some of the differences in the technologies may account for slight differences in the calculated allele frequencies. These differences are discussed.

# RESULTS

Matched FFPE tumor, adjacent fresh-frozen normal, and plasma samples from 3 donors were sourced from a commercial biobank. Standard quality control methods were used to assess the quality of the samples, including fluorometric quantification (Qubit™ dsDNA BR Assay Kit, Thermo Fisher Scientific), capillary electrophoresis (Bioanalyzer High Sensitivity DNA Kit, Agilent), or qPCR (KAPA hgDNA Quantification and QC Kit, Roche), depending on the sample (Table 1).

### Table 1. Sample quality control.

| | Frozen normal gDNA | | | FFPE tumor DNA | | | cfDNA from plasma | |
|---|---|---|---|---|---|---|---|---|
| | Conc. (ng/µL) | $Q_{129}/Q_{41}$ | DIN | Conc. (ng/µL) | $Q_{129}/Q_{41}$ | DIN | Conc. (ng/µL) | Bioanalyzer* |
| Sample 1 | 69.13 | 0.08 | 9.7 | 53.95 | 0.75 | 5.5 | 3.11 | Minimal HMW DNA |
| Sample 2 | 37.38 | 1.15 | 7.2 | 202.5 | 0.76 | 5.9 | 2.33 | Minimal HMW DNA |
| Sample 3 | 37.62 | 0.6 | 6.3 | 32.39 | 0.56 | 4.2 | 2.72 | Minimal HMW DNA |

* HMW = high molecular weight

Sequencing libraries were prepared from 100 ng of tumor and normal samples and captured using a pan-cancer research panel designed to simultaneously detect copy number variations, insertion/deletions (indels), rearrangements, and microsatellite instability across 532 oncogene targets. Variants present in FFPE tumor but absent in matched fresh-frozen normal samples were defined as tumor-associated variants. In addition, germline variants present in these samples were identified.

As circulating tumor DNA (ctDNA) has been correlated to disease progression [2], we wanted to determine if donor-specific variants could be identified in the matched cfDNA samples. We designed custom xGen cancer research panels to identify these tumor-associated mutations using 25 ng of cfDNA and presented our results in the IDT application note: Biomarker discovery—Cancer molecular profiling. Here, we focused on the validating the cfDNA-associated mutations using the IDT rhAmpSeq amplicon sequencing system as an orthogonal target enrichment approach.

The rhAmpSeq Design Tool can be used to create custom, highly multiplexed, amplicon sequencing panels in multiple animal and plant species for a wide range of applications, including hotspot genotyping and CRISPR indel screening. We used the genotyping feature to design donor-specific panels targeting single nucleotide variants (SNVs) in the hg19 genome with a 70–100 bp insert range for the assays. We obtained a high design success rate of >95% across all panels. Libraries were generated with 10 ng of cfDNA using subject-matched custom rhAmpSeq panels following the standard rhAmpSeq library preparation protocol. On-target rates are >90% for all 3 panels (Figure 1A). High coverage uniformity, calculated as the percent of targets with coverage ≥0.2X of the mean, was observed for all 3 panels (Figure 1B). Altogether, high on-target rates and uniform coverage using the rhAmpSeq system ensures high sensitivity across the target space.

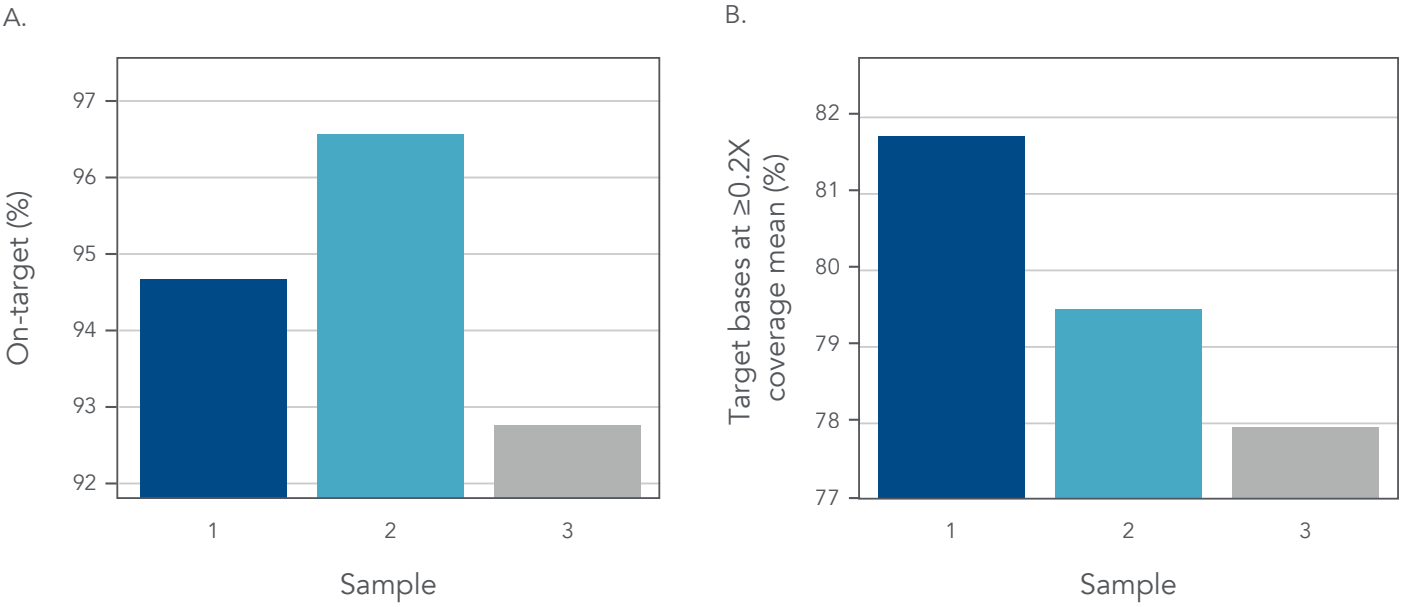


**Figure 1. High-quality sequencing data from cfDNA using the rhAmpSeq amplicon sequencing system.**

We next wanted to validate whether the SNVs identified in tumors were detected in cfDNA samples investigated by the rhAmpSeq system. Using matched cfDNA, both rhAmpSeq sequencing and xGen hybridization capture were able to identify ≥96% of variants (out of the 173, 165, and 178 variants in trios 1–3 respectively) amongst those originally scored in tumor FFPE samples using the xGen platform (Figure 2). As expected, when examining higher frequency homozygous and heterozygous germline variants in cfDNA, both technologies showed almost 100% overlap in the mutations they were able to identify, with only 2 out of 495 total variants scored across the 3 individuals not intersecting between rhAmpSeq and xGen technologies (Table 2). Moreover, we observed high correlation between the observed allele frequencies in cfDNA across the 2 methods for germline variants (Figure 3).

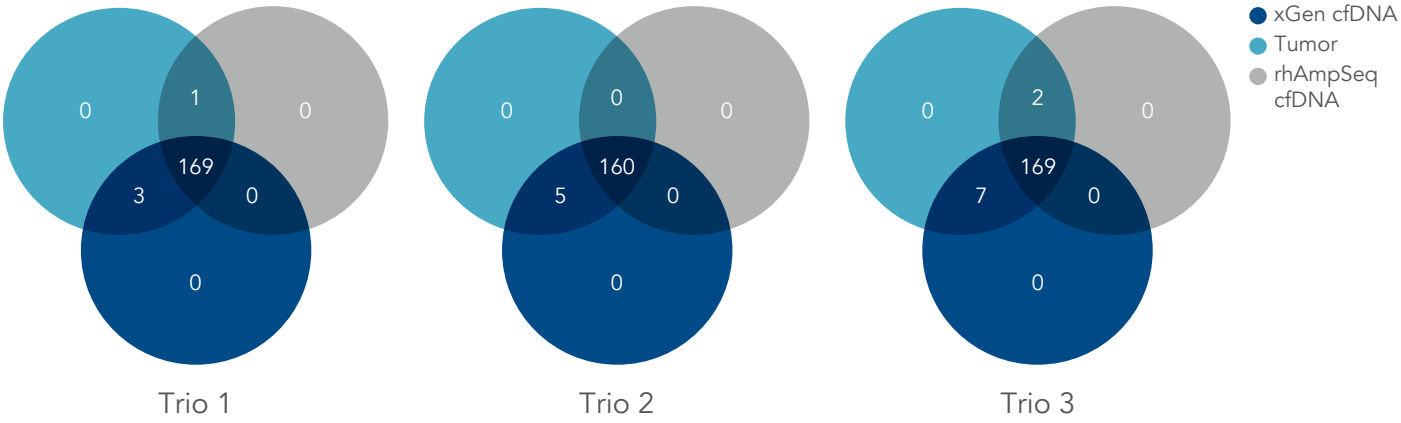


**Figure 2. Concordance between SNVs detected in matched tumor and cfDNA samples.**

## Table 2. Variants identified with the xGen Prism capture workflow and the rhAmpSeq system.

| Description | xGen | | | rhAmpSeq | | |
|---|---|---|---|---|---|---|
| | Trio 1 | Trio 2 | Trio 3 | Trio 1 | Trio 2 | Trio 3 |
| Targeted variants in tumor | 173 | 165 | 178 | 173 | 165 | 178 |
| # of identified homozygous variants | 160 | 147 | 153 | 160 | 146 | 153 |
| # of identified heterozygous variants | 8 | 14 | 13 | 8 | 14 | 12 |
| # of identified somatic variants (AF <40%) | 4 | 4 | 10 | 2 | 0 | 6 |

Variants identified from cfDNA often represent multiple tumor mutational landscapes; however, cfDNA samples contain very low proportions of tumor-derived DNA. Thus, there were a few low-frequency, tumor-associated variants that were uniquely identified in cfDNA by xGen hybridization capture but not detected by rhAmpSeq, and vice versa (Table 2). Among low-frequency somatic variants, defined using our standard of <40% VAF, we saw higher discrepancies between the observed allele frequencies in cfDNA across the 2 methods (Figure 3). These were tied to lower coverage, lower quality scores (from Vardict), or strand bias (mutation detected in either the forward or the reverse strand), in line with low proportions of tumor-derived DNA in cfDNA samples.

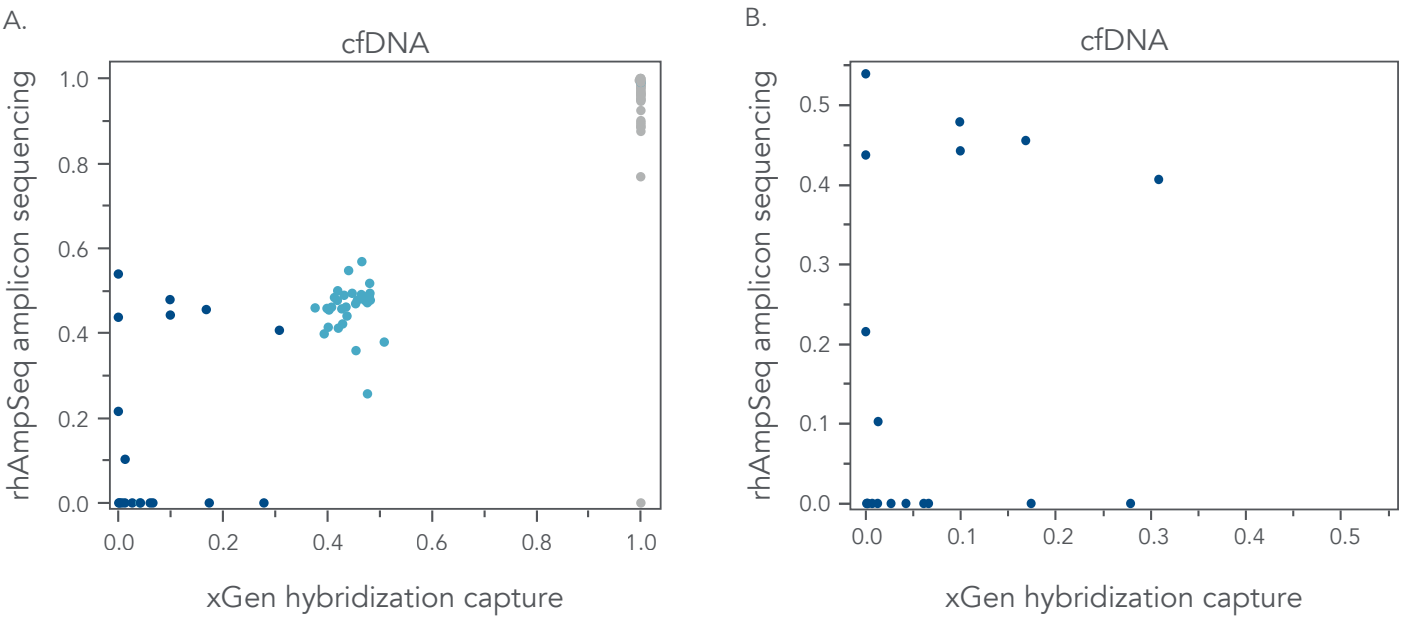


**Figure 3. Concordance between allele frequencies of variants detected in matched cfDNA samples using 2 orthogonal target enrichment approaches.** (**A**) All variants (homozygous germline in gray, heterozygous germline is light blue, somatic in dark blue). (**B**) A close-up view of only the somatic variants.

# DISCUSSION

The use of ctDNA in blood to detect cancer holds tremendous potential for non-invasive screening and monitoring approaches. Identifying DNA biomarkers from these cfDNA matrices and correlating with their original tumor signatures is a critical step to creating a molecular tool for monitoring disease progression. In this study, we used 2 different targeted NGS approaches to characterize DNA variants in cfDNA samples compared to FFPE tumor samples from the same individuals to evaluate the correlation of markers between the different sample types across technologies. Hybridization capture using the xGen platform relies on traditional library prep via adapter ligation to input DNA followed by specific target annealing with biotinylated probes, enrichment using streptavidin beads, and stringent washing to remove non-specific material. The IDT xGen Prism DNA Library Prep Kit has the unique added benefit of reducing adapter dimers and increasing overall conversion rate for maximum sensitivity and sequencing efficiency. Amplicon sequencing using the rhAmpSeq platform uses a primer design and multiplex PCR approach to amplify target regions of interest and a second round of PCR to add Illumina recognition sequences. The unique chemistry of rhAmpSeq technology also has benefits of reducing primer dimers and other non-specific amplification for optimum on-target coverage and specificity.

Both technologies were able to identify ≥96% of variants between cfDNA and tumor FFPE samples. Only 2 out of 495 total germline-level variants did not overlap between xGen and rhAmpSeq results. Both of these variants were called in the xGen analysis but missed in the rhAmpSeq analysis due to low coverage (data not shown). It is important to note that with the selection of a 70–100 bp insert size for the rhAmpSeq protocol, the target binding site for the primers could have been affected by the fragmented nature of cfDNA samples in general; thus, possibly decreasing the insert size to smaller inserts could compensate for the low coverage assays.

When we looked at the lower frequency somatic variants, defined using our standard of <40% VAF, the xGen platform identified a total of 18 variants while the rhAmpSeq platform called only 8 somatic variants across the 3 individuals. The increased sensitivity from the xGen-prepared libraries could be caused by several possible factors. One is the difference in starting input DNA amounts between the 2 preparations. The xGen-prepared libraries started with 25 ng of cfDNA, while the rhAmpSeq libraries started with 10 ng. Given that roughly each nanogram of gDNA contains 300 genome equivalents, the ability to detect low-frequency mutations down to 1% in cfDNA might present additional challenges owing to the unknown proportion of tumor-associated DNA or contamination of gDNA diluting biomarker representation.

Another factor providing added sensitivity for the xGen platform was the use of unique molecular identifiers (UMIs) in the adapters to remove potential errors generated in the NGS workflow. The use of UMI technology in NGS is gaining popularity and becoming a standard for ultra low-frequency allele quantification due to its ability (during bioinformatics data analysis) to correct for PCR errors in both the library prep and sequencing steps of the workflow. This error correction enables the reduction of false positives in variant calling as the limit of detection decreases. Not surprisingly, during development of the rhAmpSeq system, we observed high confidence allele detection down to roughly 5% VAF; below this level, the incidence of false positives increased without any error correction method (internal IDT data not shown). Nevertheless, the combination of rhAmpSeq and xGen systems creates a robust pair of complementary technologies for researching biomarker characterization and confirmation that may help accelerate development of molecular screening tools for human disease.

# REFERENCES

1. Integrated DNA Technologies (2020) **Biomarker discovery research—Cancer molecular profiling**. Online. [Accessed 24 Jun 2020].

2. Bettegowda C, Sausen M, et al. (2014) **Detection of circulating tumor DNA in early- and late-stage human malignancies**. Sci Transl Med 6(224):224ra224.

Technical support: applicationsupport@idtdna.com