Biomarker discovery research–Cancer molecular profiling

ABSTRACT

Research in the discovery and identification of new, targetable biomarkers is driven by comprehensive tumor profiling using next generation sequencing (NGS). However, converting tissue samples into NGS libraries is often challenging due to the low quantity and quality of DNA in such samples. Here, we present sensitive and accurate detection of low-frequency variants by combining the xGen™ Prism DNA Library Prep Kit, optimized for low-input and degraded samples, with IDT xGen hybridization capture reagents.

The workflow features a proprietary single-stranded ligation strategy that maximizes conversion, virtually eliminates adapter-dimer formation, and reduces chimera rates. Since dimer formation is negligible, a fixed concentration of adapter can be used, and aggressive size selection is no longer required post-ligation. The xGen Prism DNA Library Prep Kit yields high coverage and library complexity, enabling highly sensitive detection of low-frequency variants. We demonstrate detection of tumor-associated variants in matched formalin-fixed, paraffin-embedded (FFPE) and cell-free DNA (cfDNA) samples with a proof-of-concept experiment using archived lung cancer trio samples.

INTRODUCTION

Most cancers are associated with mutated genes. Gene mutations can be inherited or can occur from environmental exposure. Mutations that lead to cancer can be triggered by viruses and can occur in regulatory elements that control gene expression. Studying genomics, genes, and gene function gives researchers and clinicians insight into how mutated genes impact cancer symptoms, tumor progression, treatment response, and health outcomes. Molecular characterization of a variety of cancers has been bolstered by the establishment of The Cancer Genome Atlas (TCGA) and has helped deepen our understanding of cancer [1].

TCGA is a dataset compiled over a period of 12 years drawing from samples of over 1000 subjects. In addition to expanding our knowledge of cancer molecular and genomic mechanisms, TCGA has also revolutionized cancer classification by revealing the heterogeneity of molecular signatures, or biomarkers, of tumors between tissue types and individual patients. TCGA identified cancer subtypes and revealed how tumors can be targeted for treatment by exploiting biomarkers, paving the way toward personalized medicine.

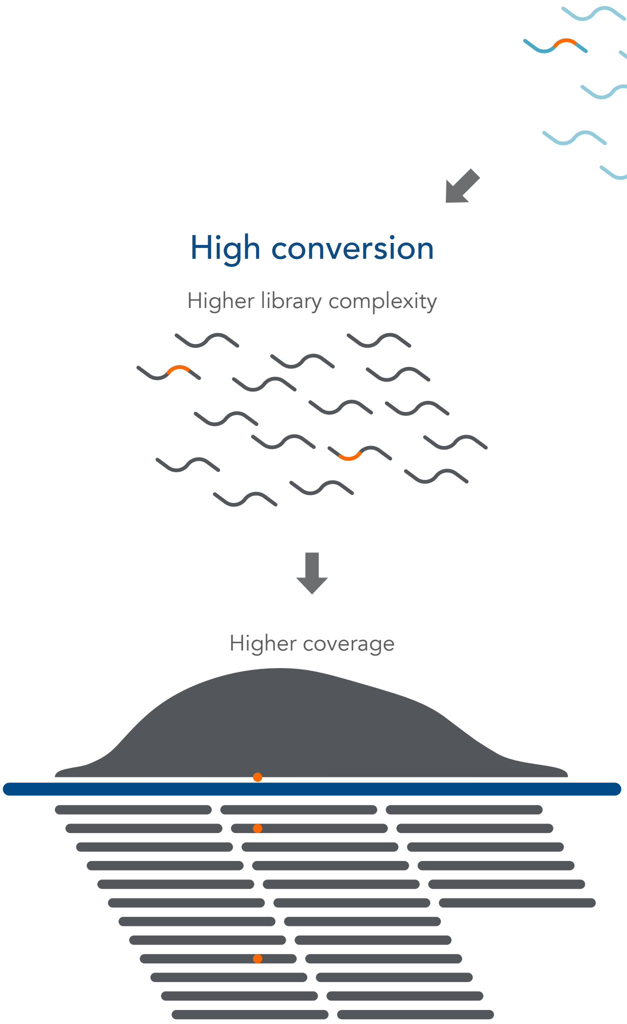
Identifying new biomarkers can aid in personalizing medicine by targeting tumors for treatment. Biomarkers indicate a disease or outcome, often because they are involved in the biological mechanism. Variants, or mutations, in biomarker genes can impact disease risk, affect a subject’s response to treatment, or lead to genetic dysregulation that results in disease. Molecular characterization of tumor types can identify biomarkers through comprehensive tumor profiling with NGS. When matched normal and tumor samples are compared, tumor-specific mutations and their functions can be identified. One way to compare samples is by using a biopsy, a sample of tissue or cells taken from the body. A liquid biopsy can be saliva, blood, or other bodily fluid.

Blood-based tests are commonly used to monitor disease progression in retrospective studies and present an attractive alternative to traditional biopsies due to their noninvasive nature and the ease of collecting multiple samples over time. These results from subject-specific tumor profiling can be used to study the expansion of specific tumor lineages and identify minimum residual disease (MRD). Even without matched tumor profiling, oncology researchers are beginning to use liquid biopsies to detect cancer early in healthy populations and to predict how a subject will respond to a specific treatment [2].

The challenge with these approaches is that tumor sample DNA is often low quality and can be difficult to convert to NGS libraries. DNA can come from samples that are formalin-fixed and paraffin-embedded (FFPE) to preserve proteins and tissue structures, or it can be cell-free DNA (cfDNA), DNA obtained through liquid biopsy. cfDNA may represent tumors that cannot be biopsied or can be used to assess multiple tumors at once. However, using cfDNA for analysis is often limited by small amounts of material and very low proportions of tumor-derived DNA. Successful library conversion is paramount when working with these types of samples.

Why is conversion important?

Conversion is the transformation of input material into library molecules that can be sequenced. When the library is prepared, genetic material is fragmented and ligated to adapters to prepare it for sequencing. The ability of a library prep kit or workflow to convert the input material into a sequenceable product is called the conversion rate. High conversion results in higher yields with fewer PCR cycles, lower bias, an increased sensitivity to detect low-frequency mutations, and an NGS library that accurately represents the input sample. In contrast, low conversion yields lower library complexity (the number of unique molecules in a library), lower coverage, and lower sensitivity, which results in low-quality data and inaccurate results (Figure 1).



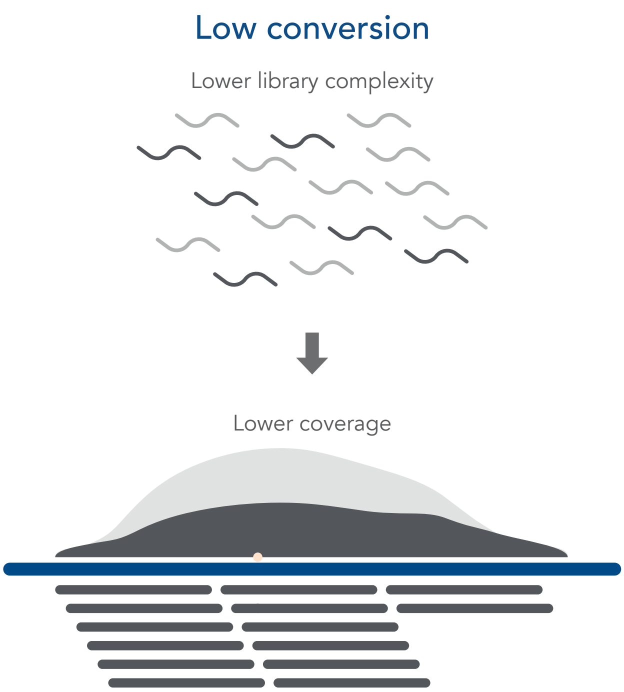


Figure 1. Comparison of sequencing results with high vs. low library conversion. The variant (orange) is lost in the library on the right with low conversion leading to low sensitivity. The light grey sequences and light orange variant (dot) on the right side indicate missing sequences and missing coverage.

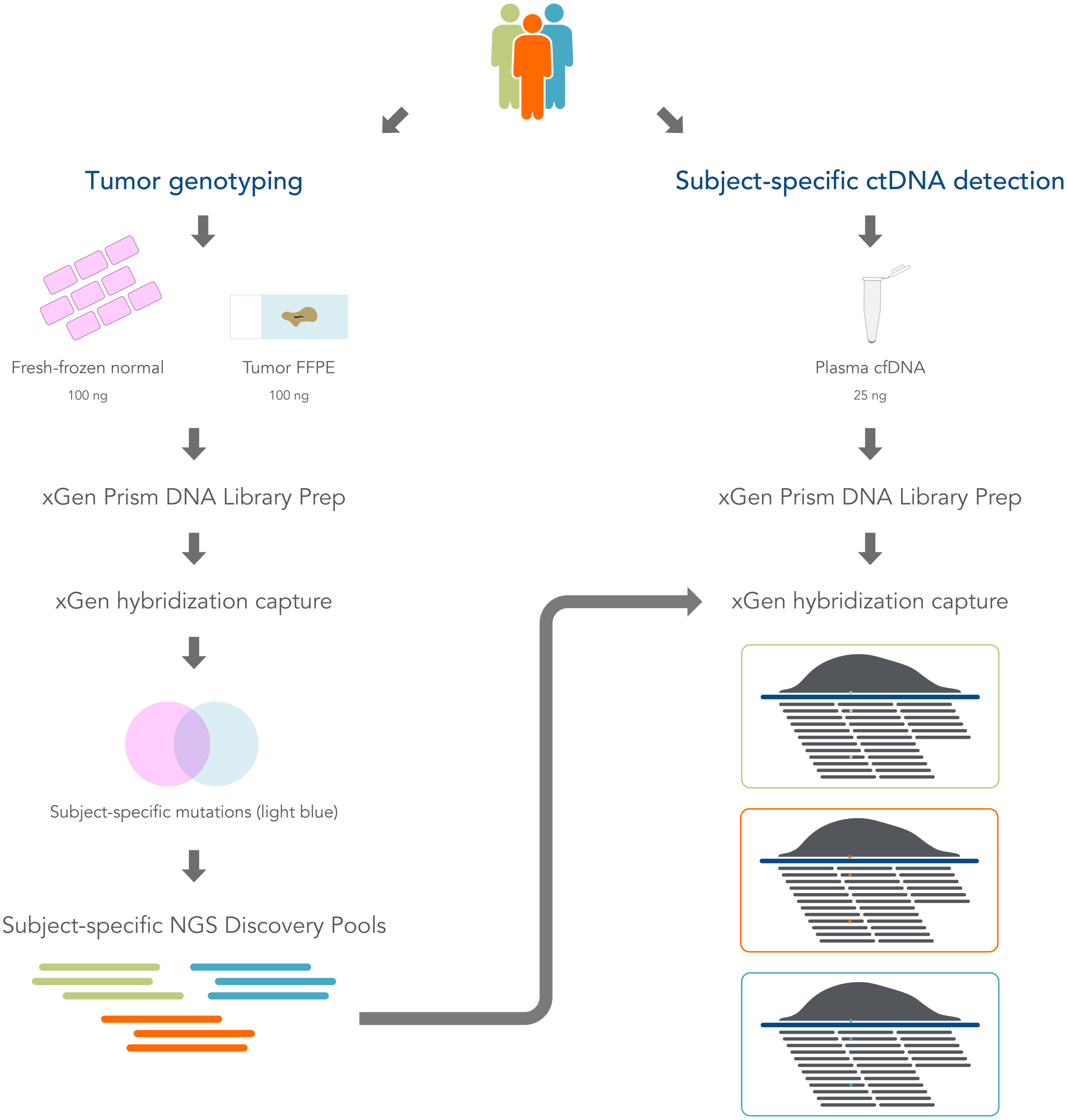

Tumor genotyping
Fresh-frozen normal
100 ng
Tumor FFPE
100 ng
xGen Prism DNA Library Prep
xGen hybridization capture
Subject-specific mutations (light blue)
Subject-specific NGS Discovery Pools
Subject-specific ctDNA detection
Plasma cfDNA
25 ng
xGen Prism DNA Library Prep
xGen hybridization capture

Figure 2. Overview of study. Fresh-frozen normal tissue, tumor-derived FFPE tissue, and plasma cfDNA were extracted from 3 lung cancer subjects. The fresh-frozen normal and FFPE tissues were used for hybridization capture to identify tumor-associated variants. The results from that sequencing were used to design subject-specific NGS Discovery Pools for use in targeted deep sequencing of the plasma cfDNA.

Here, we analyze matched archival lung cancer samples to show how IDT's NGS products enable sensitive and accurate variant detection in difficult FFPE and cfDNA samples (Figure 2). To demonstrate detection of circulating tumor DNA (ctDNA) in a subject's sample, tumor-associated variants in matched FFPE and cfDNA samples are identified in archived lung cancer trios.

RESULTS

Great research performance from low-quality FFPE samples

Matched FFPE-tumor, adjacent fresh-frozen normal, and plasma samples from 3 donors were sourced from a commercial biobank. DNA was extracted from these samples with the AnaPrep FFPE DNA Extraction Kit (PN: Z1322009, Biochain) and the cfPure® V2 Cell-Free DNA Extraction Kit (PN: K5011610-V2, Biochain), respectively. 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 HS DNA chip, Agilent), or qPCR (KAPA hgDNA Quantification and QC Kit, Roche), depending on the sample (Table 1). Quality scores (Q scores) depict DNA quality as the ratio of 129 bp vs. 41 bp reads. An ideal Q score is 1, indicating equivalent amplification of both reads and, therefore, higher DNA integrity. The DNA Integrity Number (DIN) ranges from 1 to 10 where 1 is degraded DNA and 10 is intact DNA.

Table 1. Sample quality control.

	Fresh-frozen normal			FFPE tumor			cfDNA	
	Conc.* (ng/µl)	Q129/Q41	DIN	Conc. (ng/µl)	Q129/Q41	DIN	Conc. (ng/µl)	Bioanalyzer†
Trio 1	69.13	0.82	9.7	53.95	0.75	5.5	3.11	Minimal HMW DNA
Trio 2	37.38	1.15	7.2	202.5	0.76	5.9	2.33	Minimal HMW DNA
Trio 3	37.62	0.6	6.3	32.39	0.56	4.2	2.72	Minimal HMW DNA

* Conc. = concentration
† HMW = high molecular weight

Sequencing libraries were generated with 100 ng of DNA extracted from FFPE samples and adjacent fresh-frozen normal samples from all 3 donors. Regardless of input sample quality, the xGen Prism DNA Library Prep Kit generated high-yield libraries (Figure 3). These libraries were captured in singleplex with a custom xGen cancer panel and sequenced. Picard was used to evaluate library preparation and hybrid capture performance including HS library size, duplicate rate, and coverage after standard start-stop deduplication (Figure 4). Since the goal of this study was to identify as many variants as possible, we chose to use start-stop deduplicated data. This approach sacrifices confidence in the veracity of the variants identified but can retain higher sensitivity. However, if the goal is to identify high-confidence, low-frequency variants, the libraries could be sequenced more deeply to reach the recommended duplication rates (>70%) for single read family error correction. In this study, higher coverage combined with high sensitivity was sufficient to identify tumor variants incorporated in subject-specific hybridization capture panels used to assess cfDNA for ctDNA.

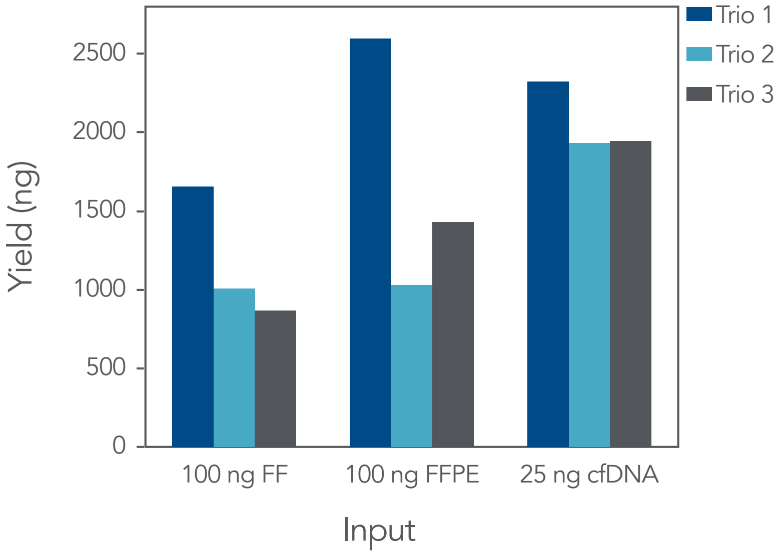


Figure 3. The xGen Prism DNA Library Prep Kit delivers high-yield and complex libraries from FFPE samples across a range of sample qualities. Libraries derived from fresh-frozen normal (FF) and FFPE-tumor samples were generated with 100 ng of input material and PCR amplified with 8 and 9 cycles, respectively. cfDNA libraries were generated from 25 ng of input and amplified with 8 PCR cycles. Library yield was measured with a Qubit fluorometer using the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific).

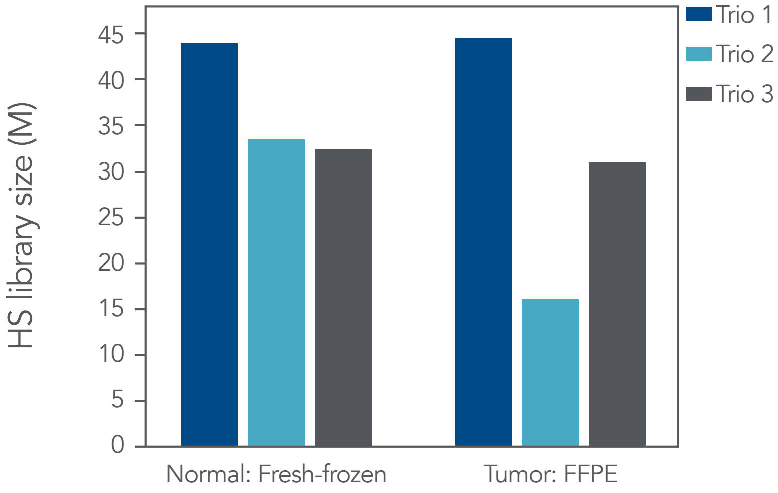


Figure 4. High-quality sequencing libraries from tumor and normal samples. Libraries derived from fresh-frozen normal and FFPE-tumor samples were generated with the xGen Prism DNA Library Prep Kit from 100 ng of input material. Libraries were captured in singleplex with a custom 2.2 Mb xGen cancer panel. Libraries were pooled and sequenced on an Illumina NextSeq® 500 instrument. Reads were downsampled to 140 M reads per library and mapped using BWA (0.7.15). Libraries were then deduplicated based on start-stop position using Picard (2.18.9) or fgbio (0.7.0), as described in the xGen Prism Analysis Guide. HS library size was calculated using Picard [3].

Variants, including single nucleotide polymorphisms (SNPs), insertion/deletions (indels), and mutations, were called from the matched fresh-frozen normal and FFPE-tumor samples using Vardict and Mutect2. Variants present in FFPE-tumor samples, but absent in matched fresh-frozen normal samples were defined as tumor-associated variants. In addition, germline variants present in these samples were identified (Table 2). This workflow identified approximately 250 variants per subject. The variants were targeted using NGS Discovery Pools designed using IDT's design algorithm for 2X tiling.

Table 2. Variants identified in FFPE-tumor samples and sample-specific panel design.

Descriptions	Trio 1	Trio 2	Trio 3
SNPs in both Vardict & Mutect2 >1% AF*	25	34	45
Indels in both Vardict & Mutect2 >10% AF	4	38	12
Mutations overlapping *EGFR, KRAS, ERBB2* >1% AF	26	45	31
Germline indels >90% AF overlapping exons	6	8	5
Germline SNPs >90% AF overlapping exons	166	154	158
Total # of mutations	227	279	251
# of probes in BED file	276	363	305
Length of probe	27 kb	34 kb	30 kb

* AF = allele frequency

xGen Prism DNA Library Prep Kit enables variant detection in cfDNA with high sensitivity

High-yield libraries were generated with 25 ng of cfDNA from each of the trios (Figure 3). The libraries were captured with subject-matched custom NGS Discovery Pools (IDT). Incorporation of unique dual indexes (UDIs) ensured accuracy and prevented sample misassignment. Despite the small size of these panels, we obtained high on-target rates and achieved a sequence depth sufficient to reach duplication rates of >80%, which is recommended for collapsed read analysis to enable error correction (Figure 5A). Collapsing reads uses unique molecular identifers (UMIs) to remove sample-prep, library-prep, and sequencing errors, allowing accurate variant calling of ultra low-frequency variants. Mapped reads were used to generate collapsed single and combined read families, as outlined in the xGen Prism Analysis Guide. The combination of the xGen Prism DNA Library Prep Kit with xGen hybridization capture resulted in high conversion rates, complexity, and coverage for cfDNA (Figure 5B). Consequently, analysis of combined read family error-corrected coverage was able to identify low-frequency variants (Figure 6).

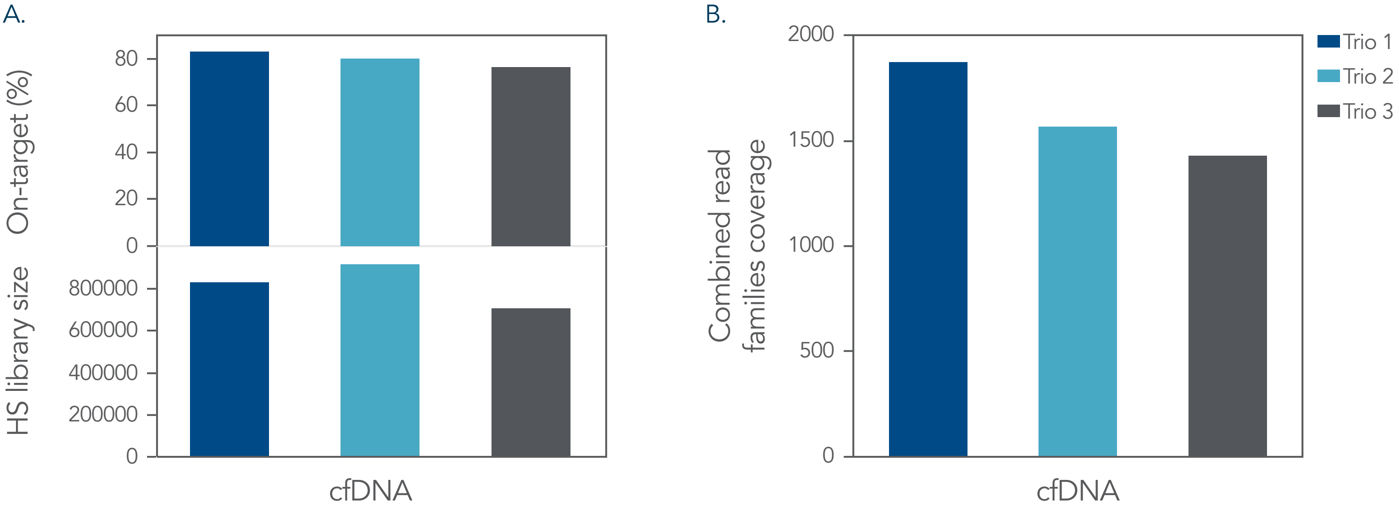


Figure 5. High complexity and coverage sequencing data from cfDNA. Libraries were generated with the xGen Prism DNA Library Prep Kit from 25 ng of cfDNA material. Libraries were captured in singleplex with custom subject-specific NGS Discovery Pools. Libraries were pooled and sequenced on an Illumina NextSeq® 500 instrument. Reads were downsampled to 40 M reads per library and mapped using BWA (0.7.15). Libraries were then deduplicated based on start-stop position using Picard (2.18.9) or error-corrected with combined read families using fgbio (0.7.0), as described in the xGen Prism Analysis Guide. (A) On-target rate was calculated using Picard percent selected bases. HS library size, and (B) coverage were also calculated using Picard [3].

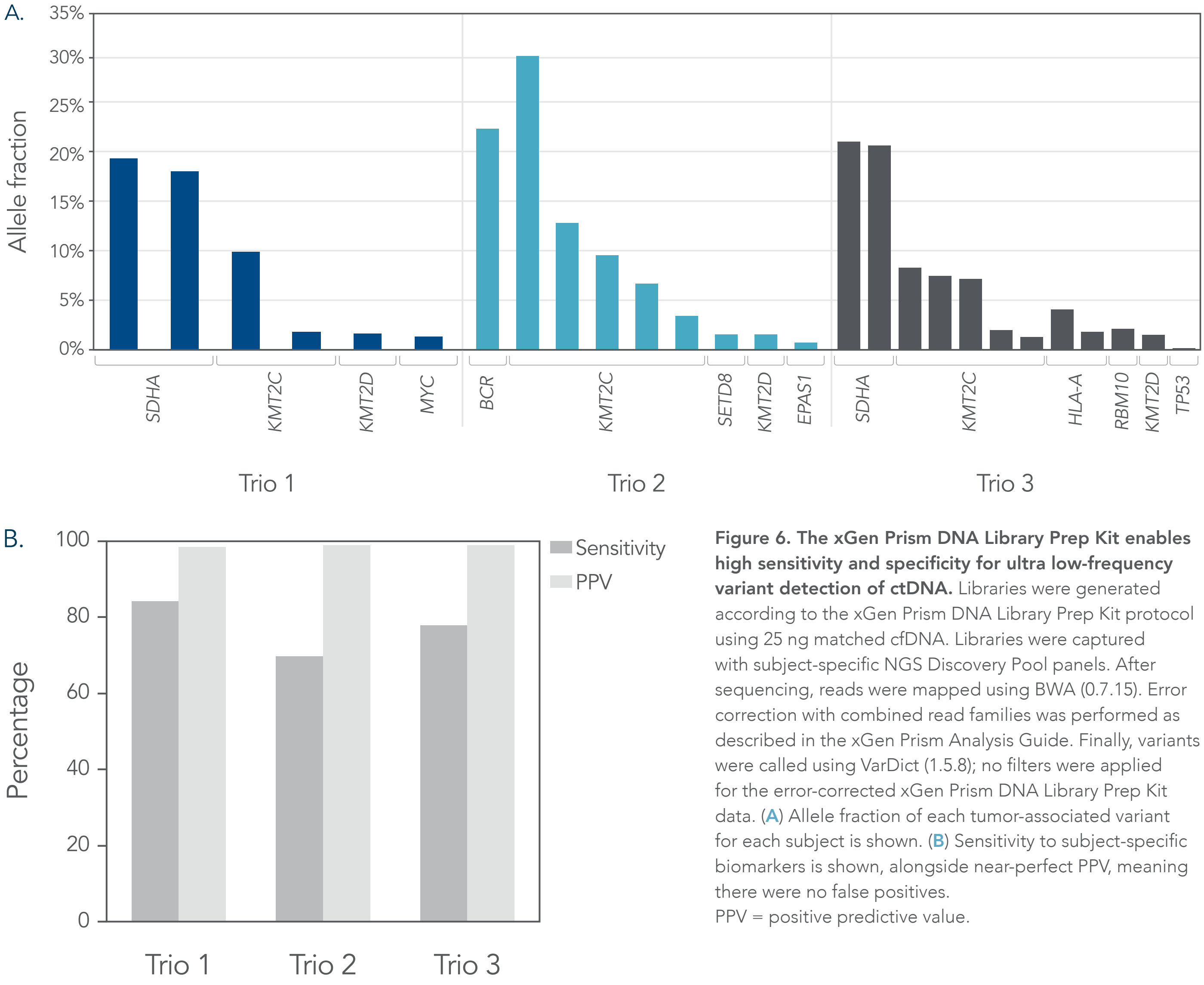


Figure 6. The xGen Prism DNA Library Prep Kit enables high sensitivity and specificity for ultra low-frequency variant detection of ctDNA. Libraries were generated according to the xGen Prism DNA Library Prep Kit protocol using 25 ng matched cfDNA. Libraries were captured with subject-specific NGS Discovery Pool panels. After sequencing, reads were mapped using BWA (0.7.15). Error correction with combined read families was performed as described in the xGen Prism Analysis Guide. Finally, variants were called using VarDict (1.5.8); no filters were applied for the error-corrected xGen Prism DNA Library Prep Kit data. (**A**) Allele fraction of each tumor-associated variant for each subject is shown. (**B**) Sensitivity to subject-specific biomarkers is shown, alongside near-perfect PPV, meaning there were no false positives.
PPV = positive predictive value.

CONCLUSIONS AND FUTURE DIRECTIONS

Here, we present sensitive and accurate detection of low-frequency variants in matched tissue and cfDNA samples by combining IDT NGS products in the field of reearch. Using an archived lung cancer trio, fresh-frozen normal tissue and tumor-derived FFPE samples were sequenced starting with the xGen Prism DNA Library Prep Kit, which is optimized for low-input and degraded samples, and enriched using IDT xGen hybridization capture. NGS Discovery Pools were designed to target tumor-associated variants identified by hybridization capture. These subject-specific variants were identified with targeted deep sequencing of the subjects' plasma cfDNA and detected ctDNA variants. Use of the xGen Prism DNA Library Prep Kit resulted in

- High conversion rates resulting from novel ligase and highly modified adapters
- High-complexity libraries, which enable detection of variants at ≤1% variant allele frequency (VAF)
- High yield and library complexity from severely degraded, FFPE samples
- Minimal errors using UMIs and minimized risk of sample misassignment using UDI primers
- Analysis of tumor-associated variants in trios using a single, streamlined workflow for cfDNA and FFPE

The technologies presented here are typically applied longitudinally to research cancer treatment response over time. The subject's response is quantified by testing for the minimal residual disease. MRD is the small amount of cancer cells that remain in the subject during or after treatment. Cancers evolve and change over time as mutations are acquired. Clonal evolution is a model for tumor progression which suggests that when cancer cells divide and replicate, variants are generated in oncogenes and tumor-suppressor genes, increasing the genetic heterogeneity of the tumor and, consequently, the evolutionary fitness and growth potential of the tumor [4]. The sensitivity and accuracy of this workflow suggest potential for longitudinal applications that research tumor evolution.

REFERENCES

1. Wang Z, Jensen MA, et al. (2016) A practical guide to The Cancer Genome Atlas (TCGA). Methods Mol Biol 1418:111–141.
2. Chen M and Zhao H (2019) Next-generation sequencing in liquid biopsy: cancer screening and early detection. Hum Genomics 13(1):34.
3. Institute B (2019) Picard Toolkit. Broad Institute, GitHub repository http://broadinstitute.github.io/picard/.
4. Cahill DP, Kinzler KW, et al. (1999) Genetic instability and Darwinian selection in tumours. Trends Cell Biol 9(12):M57–60.

Technical support: applicationsupport@idtdna.com

For more than 30 years, IDT's innovative tools and solutions for genomics applications have been driving advances that inspire scientists to dream big and achieve their next breakthroughs. IDT develops, manufactures, and markets nucleic acid products that support the life sciences industry in the areas of academic and commercial research, agriculture, medical diagnostics, and pharmaceutical development. We have a global reach with personalized customer service.

> SEE WHAT MORE WE CAN DO FOR **YOU** AT **WWW.IDTDNA.COM**.

For Research Use Only.

IDT INTEGRATED DNA TECHNOLOGIES