Enabling personalized biomarker discovery in challenging oncology samples by coupling a novel library preparation chemistry with hybridization capture

Karissa Scott, Ushati Das Chakravarty, Hsiao-Yun Huang, Timothy Barnes, Kevin Lai, Jessica Sheu, Ramses Lopez, Lynette Lewis, Anastasia Potts, and Steven Henck
Integrated DNA Technologies
* Corresponding author: kscott@idtdna.com

1. Introduction

Discovery and identification of new, targetable biomarkers in research 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. Here we present sensitive and accurate identification 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.

This workflow features a proprietary single-stranded ligation strategy with a novel mutant ligase 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, while aggressive size selection is no longer required post-ligation. We demonstrate higher coverage and library complexity than conventional TA-ligation–based methods, enabling highly sensitive, low-frequency variant identification. In this internal benchmark study, performance is also demonstrated using lung cancer trios to identify tumor-associated variants in matched formalin-fixed, paraffin-embedded (FFPE) and cell-free DNA (cfDNA) samples.

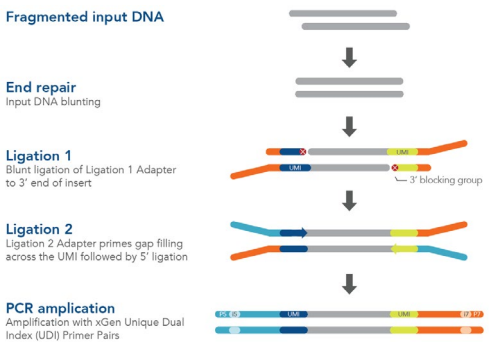


Figure 1. xGen Prism DNA Library Prep Kit workflow using unique molecular identifiers (UMIs).

2. Methods

Standards that mimic cfDNA and FFPE samples were obtained from Horizon Discovery. DNA extracted from matched, archived FFPE tumor, fresh-frozen normal, and plasma samples from individuals with lung cancer were obtained from a commercial biobank. Libraries were generated according to manufacturer's instructions, then captured using the IDT xGen hybridization capture protocol and sequenced on a NextSeq™ 500 instrument (Illumina). After demultiplexing and mapping to *hg19*, deduplication was accomplished using *MarkDuplicates* with Picard (2.18.9), while error-corrected reads were generated when appropriate using *CallDuplexConsensusReads* with fgbio (0.2.1). Coverage and complexity metrics were analyzed with Picard using *CollectHsMetrics*. Relative conversion rates were calculated from mean target coverage at very high duplication rates. Variants were called using Vardict (1.5.8) and Mutect2 (4.1.3.0). Ground truth for calculating positive predictive value (PPV) was built by ultra-deep sequencing of 100% non-mixture samples (Horizon cfDNA standards). Variants present in FFPE tumor, but absent in matched fresh-frozen normal samples, were defined as tumor-associated variants. Personalized NGS Discovery Pools (IDT) and rhAmpSeq™ panels were designed against an individual's tumor-associated variants.

3. High conversion and UMI-based error correction delivers reliable variant calling

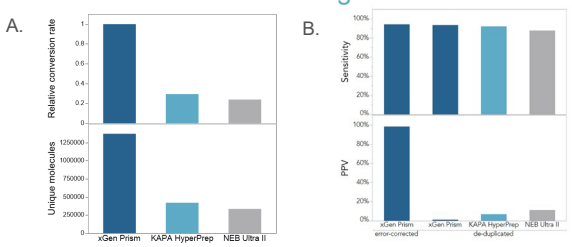
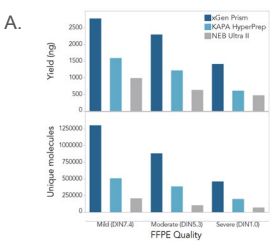


Figure 2. High conversion and UMI-based error correction enables ultra-low variant detection with cfDNA samples. (**A**) Libraries were generated from 25 ng of 0.25% allele frequency (AF) cfDNA reference standard and captured with a 61 kb custom research xGen panel. (**B**) Variants identified with VarDict were filtered with PASS/f0.02 filters for start-stop deduplicated libraries, while no filters were applied for error correction across both strands using xGen Prism libraries.

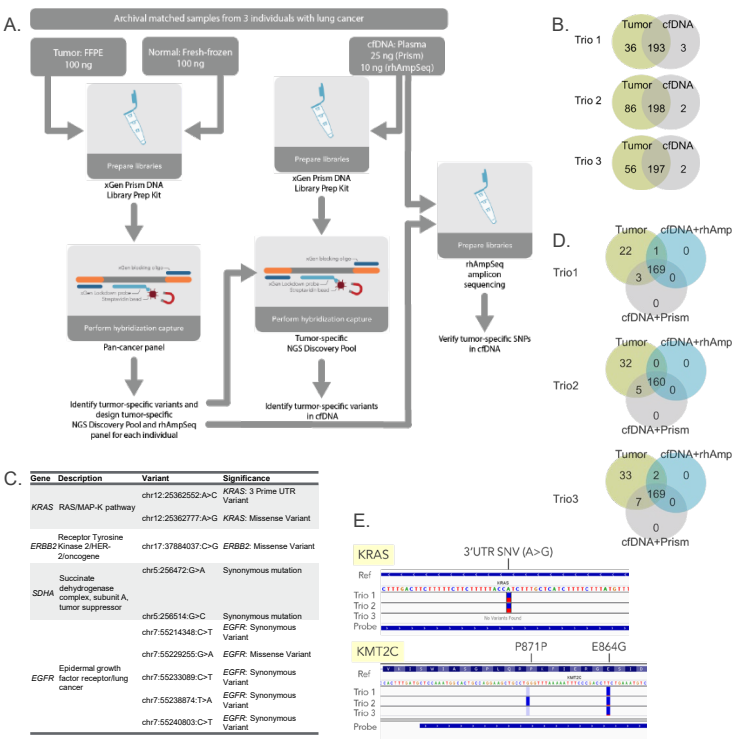


B.

Variants	Expected AF	Input FFPE (DIN 1.0)		
		25 ng	100 ng	250 ng
EGFR:G719S	24.5	22.3 (3/3)	21.8 (3/3)	21.4 (3/3)
PIK3CA:H1047R	17.5	18.3 (3/3)	17.9 (3/3)	18.3 (3/3)
KRAS:G13D	15	13.5 (3/3)	13.2 (3/3)	13.2 (3/3)
NRAS:Q61K	12.5	11.8 (3/3)	10.2 (3/3)	10.3 (3/3)
BRAF:V600E	10.5	9.9 (3/3)	9 (3/3)	9.7 (3/3)
cKIT:D816V	10	9.8 (3/3)	8.6 (3/3)	9 (3/3)
PIK3CA:E545K	9	6.3 (3/3)	5.7 (3/3)	5.6 (3/3)
KRAS:G12D	6	5.9 (3/3)	5.1 (3/3)	5.3 (3/3)
EGFR:L858R	3	3.9 (3/3)	3.4 (3/3)	3.7 (3/3)
EGFR:E746-A750	2	0.6 (3/3)	0.2 (3/3)	0.4 (3/3)
EGFR:T790M	1	1 (3/3)	0.9 (3/3)	0.9 (3/3)

Figure 3. Higher coverage and complexity delivers reliable variant and indel identification in FFPE samples. (**A**) Libraries were generated from 25 ng of FFPE reference standards with various qualities and captured with a 61 kb custom research xGen panel. (**B**) Libraries were generated from a severely damaged FFPE reference standard with variants of defined allele frequencies and captured with a 238 kb custom xGen panel. The average observed variant allele detection reported by VarDict is shown along with the number of samples in which each variant was detected.

4. Personalized biomarker discovery in lung cancer trios



C.

Gene	Description	Variant	Significance
KRAS	RAS/MAP-K pathway	chr12:25362552:A>C	*KRAS*: 3 Prime UTR Variant
		chr12:25362777:A>G	*KRAS*: Missense Variant
ERBB2	Receptor Tyrosine Kinase 2/HER-2/oncogene	chr17:37884037:C>G	*ERBB2*: Missense Variant
SDHA	Succinate dehydrogenase complex, subunit A, tumor suppressor	chr5:256472:G>A	Synonymous mutation
		chr5:256514:G>C	Synonymous mutation
EGFR	Epidermal growth factor receptor/lung cancer	chr7:55214348:C>T	*EGFR*: Synonymous Variant
		chr7:55229255:G>A	*EGFR*: Missense Variant
		chr7:55233089:C>T	*EGFR*: Synonymous Variant
		chr7:55238874:T>A	*EGFR*: Synonymous Variant
		chr7:55240803:C>T	*EGFR*: Synonymous Variant

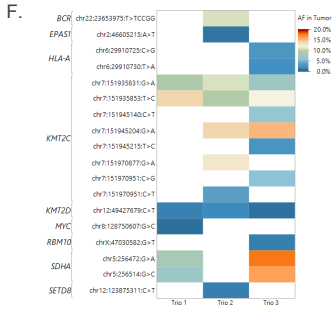


Figure 4. The xGen Prism DNA Library Prep Kit detected tumor-associated variants in matched cfDNA. (**A**) Overview of biomarker discovery workflow. (**B**) Most tumor-associated variants were detected in matched cfDNA. (**C**) Variants identified in this study that are thought to be associated with cancer. (**D**) Most single nucleotide polymorphisms (SNPs) found in both tumor and cfDNA using NGS Discovery Pools were confirmed using the rhAmpSeq workflow. (**E**) IGV screenshots demonstrate a heterozygous, germline, single nucleotide variant (SNV) in *KRAS* (44% AF in tumor) and somatic variants in *KMT2C* (9%, 12% AF in tumor). (**F**) Heat maps demonstrate low-frequency somatic variants (AF<20%) identified in both tumor and cfDNA samples.

5. Conclusions

- Higher conversion rates were achieved using mutant ligase and proprietary ligation adapters than TA-ligation–based methods.
- Higher complexity libraries enable identification of variants at ≤1% variant allele frequency (VAF).
- Higher yield and library complexity were achieved from severely degraded FFPE samples.
- UMIs enable error correction, and unique dual indexing (UDI) primers minimize the risk of sample misassignment.
- A single, streamlined workflow enables analysis of tumor-associated variants in both FFPE and cfDNA.