Application Note:

ThruPLEX® Tag-seq Detection Sensitivity and Specificity using Horizon cfDNA Reference Standards

Edward Jan[1], Matt Carroll[1], Jinglan Zhang[2], Hongzheng Dai[2], Richard Yim[3], and Shawn Quinn[4]

Introduction

The ThruPLEX® Tag-seq Kit is a next-generation sequencing (NGS) library preparation kit designed to distinguish low-frequency variants in plasma, tissue, and formalin-fixed paraffin-embedded (FFPE) samples. The kit uses Rubicon Genomics' proprietary ThruPLEX chemistry to prepare molecularly tagged and sample-indexed Illumina NGS libraries from cell-free DNA (cfDNA) and fragmented double-stranded DNA (dsDNA) (Figure 1).

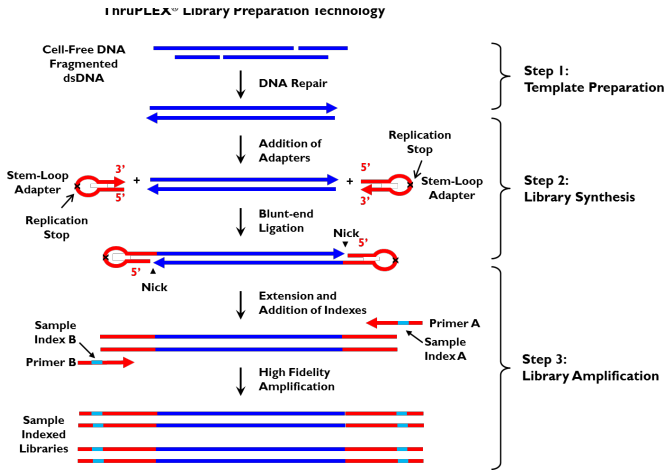

Figure 1. ThruPLEX technology: *A 3-step, single-tube reaction that starts with fragmented, dsDNA or isolated cfDNA. The ThruPLEX stem-loop adapters with molecular tags and blocked 5′ termini are blunt-end ligated to the repaired input DNA. These molecules are extended and then amplified using a high-fidelity polymerase to yield a molecularly tagged and sample-indexed Illumina NGS library.*

ThruPLEX Tag-seq features an extremely simple and fast workflow that can be completed in approximately 2 hours. It requires no intermediate purification or sample transfer steps (Figure 2), which are known to decrease library yield and increase the probability of contamination and handling errors.

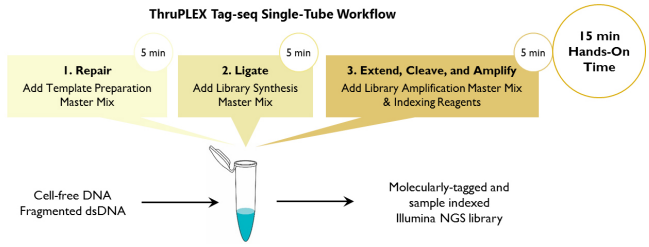

Figure 2. ThruPLEX Tag-seq single-tube library preparation workflow: *The ThruPLEX Tag-seq workflow consists of 3 simple steps that take place in the same PCR tube or well, eliminating the need to purify or transfer the sample material.*

When combined with hybridization-based target enrichment, ThruPLEX Tag-seq libraries can be sequenced at great depth (5,000–10,000X raw coverage) to confidently detect variants present at low frequencies starting from 1–50 ng of DNA. ThruPLEX Tag-seq provides more than 16 million random molecular tags to uniquely label the input DNA fragments. The unique molecular tags (UMTs) are incorporated during the ligation step, allowing subsequent computational correction of errors accumulated during amplification and sequencing. The construction of consensus sequences of molecular duplicates during bioinformatic analysis increases the accuracy of sequencing individual input molecules, enabling the detection of low-frequency variants with high sensitivity and specificity.

To validate the increased accuracy of variant detection using the ThruPLEX Tag-seq Kit, we measured the limits of variant detection using reference cfDNA engineered with variants at different allele frequencies. This effort involved the steps of library preparation, target enrichment, sequencing, and data analysis (Figure 3). It demonstrates the utility of ThruPLEX Tag-seq as a powerful tool for rare allele detection.

Methods

Figure 3: Complete workflow from sample to data analysis.

Horizon reference standards. Horizon Multiplex I cfDNA Reference Standards (Horizon HD780) are derived from human cell lines with six engineered single-nucleotide variants (SNVs). The DNA has been fragmented to an average size of 160 bp to resemble cfDNA extracted from human plasma and is provided with allelic frequencies measured by digital PCR. To generate samples with lower allele frequencies, the Horizon rare-allele samples were diluted with the Horizon wild-type reference DNA.

Library preparation. Molecularly tagged and sample-indexed libraries were prepared from 10 ng or 30 ng of Horizon HD780 DNA using the ThruPLEX Tag-seq Kit with dual indexes. Amplified libraries were purified using AMPure® XP beads and eluted in nuclease-free water for enrichment.

Target enrichment. Hybridization and capture of the pooled, indexed ThruPLEX Tag-seq libraries were carried out using the SureSelect®XT2 or Roche NimbleGen® SeqCap® EZ target enrichment systems, using a 120 kb or 240 kb custom panel. Multiple libraries were pooled for each hybridization reaction. In addition, 1 µL (1 nmol) each of i5 and i7 xGen® Universal Blocking Oligo - TS HT (Integrated DNA Technologies) were added into the hybridization reaction to prevent nonspecific binding.

Sequencing. Target-enriched libraries were sequenced on a HiSeq® 2500 or NextSeq® 500 system to achieve a raw coverage of 5,000X.

Data analysis. Data processing and analysis were performed on the Curio Genomics bioinformatics platform. FASTQ files were uploaded to Curio, and the sequencing data was aligned with unique molecular tag processing. Variant analysis was conducted using the Curio variant detection module.

Results and Discussion

Using the Curio bioinformatics platform, ThruPLEX Tag-seq's UMT reads were grouped into amplification families and then collapsed into family consensus sequences that representunique DNA molecules. Reads were discarded if the amplification family consisted of fewer than 2 members. Sequencing positions were called as ambiguous (N) if fewer than 60% of the reads at the position in the amplification family had the identical sequence. This process removes errors generated during library preparation, sequencing, and base-calling.

Figure 4 illustrates the power of using ThruPLEX Tag-seq to reduce background errors for more confident identification of variants. The aligned raw reads show that the expected EGFR T790M mutation (yellow dot) is obscured by false-positive noise (grey dots) in the 130 bp window centered at the mutation position, making it difficult to distinguish the true mutation from false positives (Figure 4A). In contrast, using family consensus reads, the consensus sequence shows a dramatic reduction in the level of background errors (Figure 4B). This error reduction provides clear separation of the true mutation from the noise.

The signal-to-noise ratio was calculated for each sample by processing the data after removing duplicate reads or by constructing consensus read using UMTs. For each sample and processing method, signal was calculated by taking the average of the allele frequency detected at six expected mutation positions and noise was calculated by averaging the allele frequency detected outside the six positions across the entire captured region. The results (Figure 4C) show a 3X to 6X improvement in signal-to-noise ratio when UMTs were utilized for error correction and consensus building during data processing.

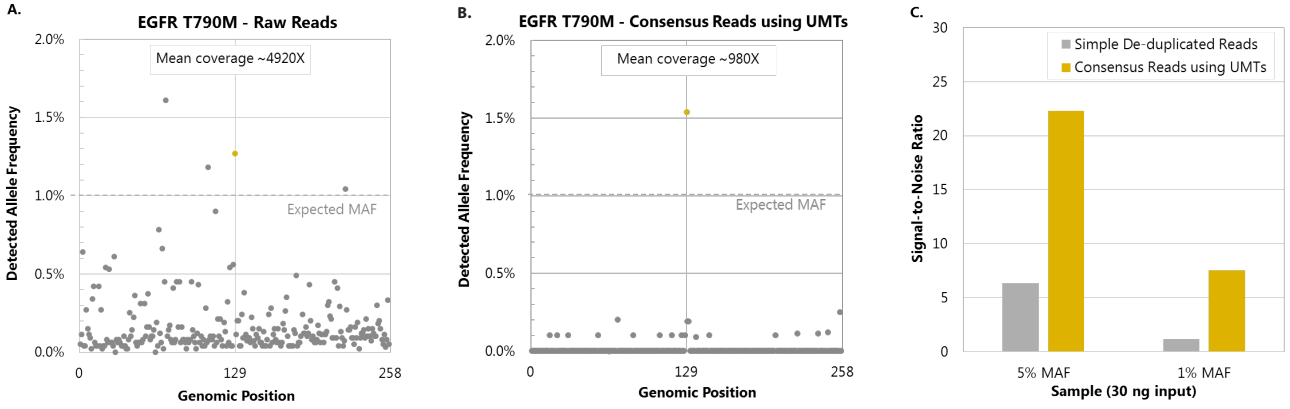

Figure 4. UMTs greatly reduce background signal. *Data from the same sample were processed without (A) or with UMTs (B). A: In conventional library preparation, the true mutation (yellow dot) can be difficult to detect due to background errors (gray dot). B: With the use of UMTs, the background is significantly reduced, and the true mutation can be detected. C: UMTs greatly increase the signal-to-noise ratio: The signal-to-noise ratio is significantly improved at various allele frequencies between conventional library preparation (gray) and library preparation containing UMTs (yellow).*

DNA Input	Enrichment	Sequencing	Sample MAF	Detected Variant Frequency						Sensitivity	Specificity
				EGFR L858R	EGFR T790M	KRAS G12D	NRAS A59T	NRAS Q61K	PIK3CA E545K		
30 ng	110 kb panel, Agilent SureSelect	~1000X mean unique coverage, NextSeq 500	5%	3.7%	5.7%	6.1%	4.4%	5.9%	5.6%	100.0%	99.8%
			1%	0.5%	1.4%	1.5%	1.7%	1.2%	1.0%	100.0%	99.9%
			WT**	0%*	0%*	0%*	0%*	0%*	0%*		
10 ng	240 kb panel, Roche NimbleGen	~500X mean unique coverage, HiSeq 2500	2.5%	2.3%	1.0%	2.5%	3.6%	2.3%	1.6%	100.0%	99.6%
			1%	1.4%	0.6%	1.3%	0.9%	0.4%	1.7%	100.0%	99.8%
			0.5%	1.4%	0.2%	0.9%	0.8%	1.3%	1.1%	100.0%	99.8%

*Not detected
**100% wild type negative control

Table 1. UMTs provide excellent variant detection. *Horizon Multiplex I cfDNA Reference Standards (Horizon HD780) were used as is or titrated using the wild-type reference standard to generate samples at additional allele frequencies. Variants were detected at their expected frequencies (MAF) with high sensitivity and specificity.*

Overall, we tested six Horizon cell-free DNA reference standards ranging from 0% (wild type) to 5% minor allele frequency (MAF). Table 1 shows that all six variants were called at their expected frequencies, with 100% sensitivity and specificity (per base) greater than 99.6%. By combining deep sequencing with the ThruPLEX Tag-seq Kit, it was possible to detect mutations present at 0.5% allele frequency using a starting input of just 10 ng of DNA. Lower detection limits and higher specificity can be achieved, depending on sample quality, input amount, capture efficiency, sequencing depth, and data processing algorithms.

Conclusion

Using 10 ng and 30 ng of Horizon reference standards, we prepared molecularly tagged libraries using the ThruPLEX Tag-seq Kit and performed targeted sequencing to a mean raw coverage of 5,000X. Data processing and analysis were conducted using the ultra-fast Curio bioinformatics platform. We demonstrated that ThruPLEX Tag-seq libraries can be used to detect variants at 0.5% allele frequency with high sensitivity and specificity, using just 10 ng of input DNA.

Equipped with more than 16 million unique molecular tags, ThruPLEX Tag-seq is a powerful tool for confident detection of low-frequency alleles. ThruPLEX Tag-seq's highly efficient chemistry and single-tube workflow work together to preserve molecular complexity, allowing researchers to discover more information from precious samples, using just 1 to 50 ng of DNA.

Researchers have the freedom to use any commercially available capture panels. Alternatively, they can design custom capture panels to interrogate genomic regions of interest that span hundreds of genes and study variants present at low allele frequencies. Lower detection limits can be achieved, depending on sample quality and input amount, capture efficiency, and sequencing depth.

1. Rubicon Genomics, Inc., Ann Arbor, Michigan, United States
2. Baylor Miraca Genetics Laboratories, Houston, Texas, United States
3. University College London, London, United Kingdom
4. Curio Genomics, Dexter, Michigan, United States

Trademarks