Highly accurate detection of low-frequency variants using molecular tags

- Streamlined workflow conserves valuable samples and ensures accurate sample tracking »
- Incorporation of UMIs carefully selected for even representation and high performance »
- Effective elimination of false positives following UMI correction »

Introduction

The ability to confidently detect low-frequency variants from NGS-based assays has been steadily rising in importance. More than ever, researchers in a wide range of fields are interested in pushing the limits of sensitivity and specificity in variant detection, an area that has previously been limited by sample preparation, amplification artifacts, and sequencing errors. Accurate detection of variants has recently been refined by the incorporation of unique molecular identifiers (UMIs) and unique dual indexes (UDIs) on Illumina platforms. As library preparation has a direct impact on the quality of sequencing results, ThruPLEX Tag-Seq HV chemistry has been engineered and optimized to produce highly diverse libraries with reproducible sequencing performance from 5 to 200 ng of input DNA. Its single-tube workflow (Figure 1) is the simplest in the industry and enables the addition of adapters containing UMIs and Illumina-compatible UDIs in three short steps. The sample never leaves the tube, ensuring accurate sample tracking, minimizing handling errors, and preventing loss of valuable samples.

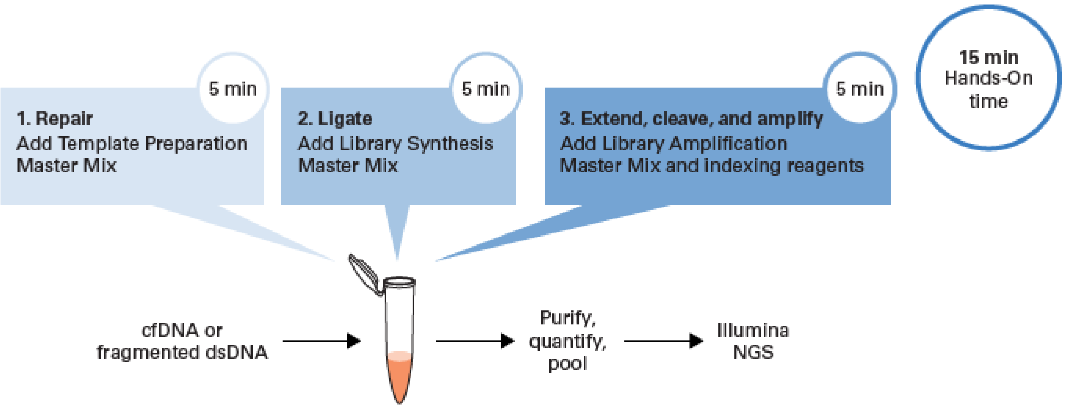

5 min
1. Repair
Add Template Preparation Master Mix
5 min
2. Ligate
Add Library Synthesis Master Mix
5 min
3. Extend, cleave, and amplify
Add Library Amplification Master Mix and indexing reagents
15 min Hands-On time
cfDNA or fragmented dsDNA
Purify, quantify, pool
Illumina NGS

Figure 1. ThruPLEX Tag-Seq HV single-tube library preparation workflow. This protocol consists of three simple steps that take place in the same PCR tube or well, thus eliminating the need to purify or transfer the sample material.

The ThruPLEX adapters have been redesigned to include discrete UMIs (Figure 2, left panel). Each pool of adapters contains 144 unique sequence combinations with a Hamming distance above 6. The adapters were carefully balanced to obtain equal representation. The seven base UMIs are located at the beginning of the reads, ensuring easy demultiplexing of the samples to simplify analysis (see Methods for details). The UMIs used to "tag" DNA molecules are then processed to identify consensus sequences and reduce the false-positive variants introduced by amplification or sequencing errors (Figure 2, right panel).

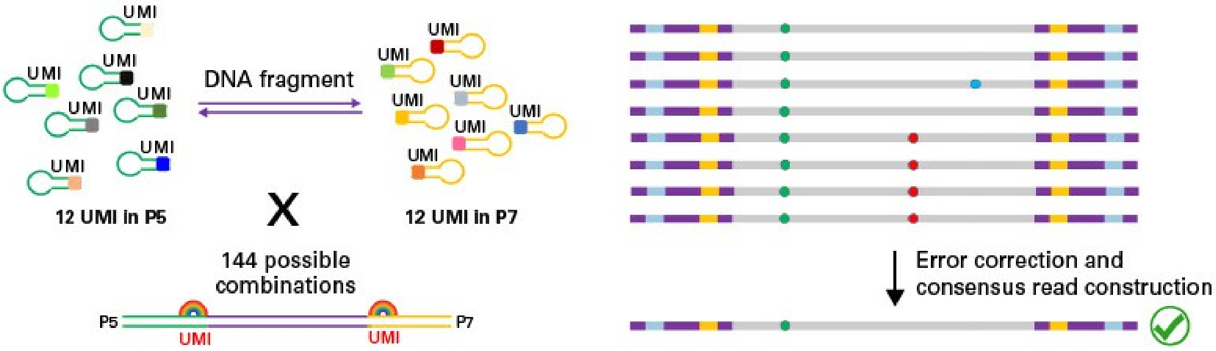
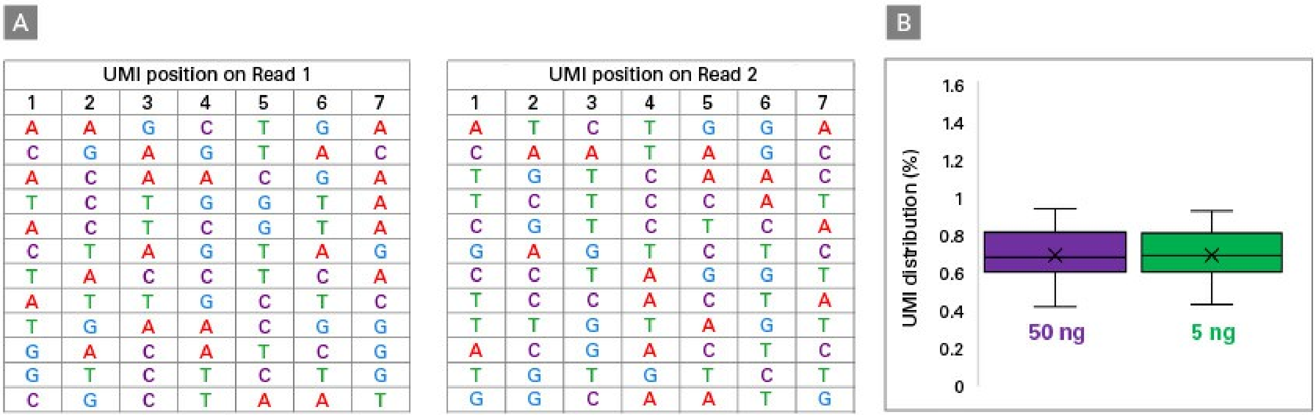

UMI
DNA fragment
12 UMI in P5
X
12 UMI in P7
144 possible combinations
P5
UMI
UMI
P7
Error correction and consensus read construction
✓

Figure 2. ThruPLEX Tag-Seq HV includes 144 discrete UMIs. This workflow is designed to eliminate the ambiguity in variant calling by reducing the false-positive calls resulting from DNA polymerase and sequencing errors.

Results

Design of discrete UMIs for optimal performance

The ThruPLEX Tag-Seq HV UMI sequences were carefully selected for their high Hamming distance, providing high dissimilarity from each other. Since the UMI sequences are known and distinct from each other, it is possible to correct potential sequencing errors before grouping the reads by UMI family. (See Methods for more details.) The UMI sequences were also selected for their color balance in order to enable high-quality sequencing on Illumina platforms (Figure 3, Panel A). The concentration of each UMI adapter was optimized to produce an even representation of the 144 UMI combinations at every DNA input level. The percentage of each UMI combination was calculated from the total number of reads. As expected, the average representation of a given UMI was 0.7% (100%/144 combinations) and deviated less than 50% of the mean (±0.35%) (Figure 3, Panel B).



UMI position on Read 1						
1	2	3	4	5	6	7
A	A	G	C	T	G	A
C	G	A	G	T	A	C
A	C	A	A	C	G	A
T	C	T	G	G	T	A
A	C	T	C	G	T	A
C	T	A	G	T	A	G
T	A	C	C	T	C	A
A	T	T	G	C	T	C
T	G	A	A	C	G	G
G	A	C	A	T	C	G
G	T	C	T	C	T	G
C	G	C	T	A	A	T

UMI position on Read 2						
1	2	3	4	5	6	7
A	T	C	T	G	G	A
C	A	A	T	A	G	C
T	G	T	C	A	A	C
T	C	T	C	C	A	T
C	G	T	C	T	C	A
G	A	G	T	C	T	C
C	C	T	A	G	G	T
T	C	C	A	C	T	A
T	T	G	T	A	G	T
A	C	G	A	C	T	C
T	G	T	G	T	C	T
G	G	C	A	A	T	G

Figure 3. ThruPLEX Tag-Seq HV UMI design strategy. Panel A. UMI sequences were selected to ensure high-quality sequencing on Illumina platforms. **Panel B.** Concentration was optimized to obtain even representation, with DNA inputs of 50 ng (purple) and 5 ng (green) showing the same UMI distribution.

Improved genome coverage

The human genome contains an average of 41% GC content, mostly ranging from 20% to 65%. However, promoter regions are known for their high GC content and their overlap with CpG islands. ThruPLEX Tag-Seq HV has been optimized to capture all regions of the genome, enabling high sensitivity in mutation detection, even in extreme GC content regions. To demonstrate this benefit, the system was used to perform low-pass whole-genome sequencing, and Picard's CollectGCBiasMetrics was used to assess the GC bias. ThruPLEX Tag-Seq HV showed a normalized coverage close to the expected theoretical value of 1, even at AT- and GC-rich windows (Figure 4).

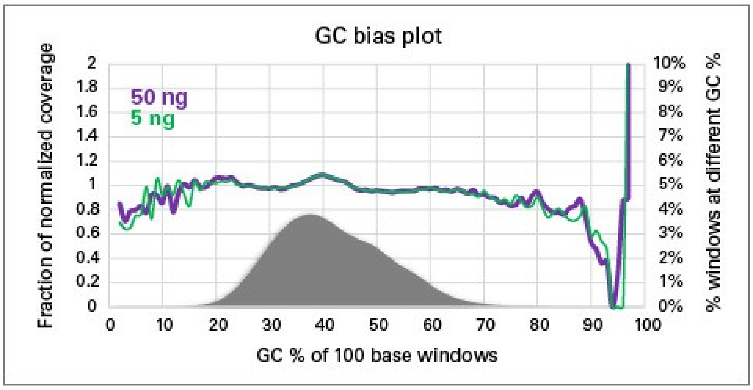


Figure 4. Low GC bias across the human genome. Libraries were prepared from 5 ng and 50 ng of sheared human genomic DNA (Horizon Discovery Quantitative Multiplex Reference Standard, Cat. # HD701) using ThruPLEX Tag-Seq HV. Libraries were sequenced on an Illumina MiSeq® instrument, and reads were aligned to the human genome (HG19) using bowtie2. The alignment metrics and GC bias were calculated using Picard tools.

Reliable detection of variants

Enrichment of regions of interest within the human genome is essential for the detection of low-frequency variants at an affordable cost. Additionally, multiple samples can be processed together in a single enrichment and multiplexed on a sequencer. To test the ability of ThruPLEX Tag-Seq HV to detect low-frequency variants, replicate libraries were produced from reference DNA presenting characterized variants inside cancer-related genes at known allele frequencies. Cancer-related genes were enriched using an IDT xGEN Pan Cancer Panel. The panel targets 800 kb of the human genome—more specifically, 127 genes frequently mutated in solid tumors, including *KRAS*, *EGFR*, and others. The variants were identified from the aligned BAM files before deduplication, based on the read pair start and end coordinates, and by using the UMI consensus reads (see Methods for details). As depicted in Table 1, the average coverage of the targeted regions from the reads deduplicated by coordinates or by UMI consensus was similar and concordant with the PCR duplication reported by Picard MarkedDuplicates. It is also important to note that the UMI representation was unaffected by the hybridization capture process, showing the unbiased ligation of the adapters (Figure 5).

Input (HD701)	Total reads	%reads mapped to human genome	%reads on or near baits	%duplicate reads	Average target coverage before deduplication	Average target coverage after deduplication by coordinates	Average target coverage after UMI collapse
50 ng	20M	98%	50%	17%	1,172	835	216
5 ng	20M	98%	90%	83%	2,095	235	220

Table 1. Hybridization metrics. Libraries were prepared in duplicate from 5 ng and 50 ng of sheared human genomic DNA (Horizon Discovery Quantitative Multiplex Reference Standard, Cat. # HD701), using ThruPLEX Tag-Seq HV. Libraries were enriched using an xGEN Pan Cancer Panel v1.5 (IDT, Cat. # 1056205) and sequenced on an Illumina NextSeq™ 500. The reads were downsampled to 20M reads then aligned to the human genome (HG19) using bowtie2. The reads were deduplicated using Picard MarkDuplicates or collapsed using fgbio bio tools (see Methods for details). Hybridization metrics were determined using Picard HsMetrics.

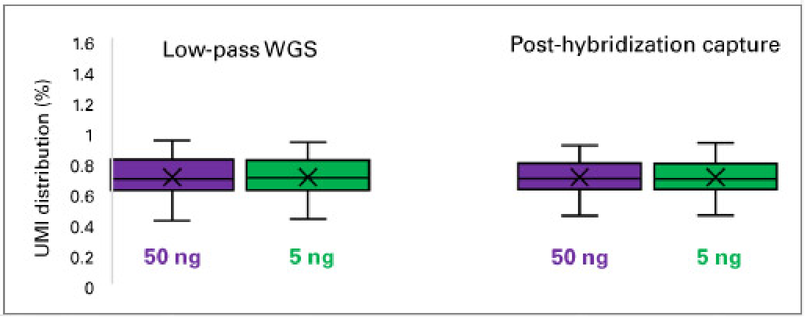


Figure 5. UMI representation before and after hybridization capture. The UMI distribution was unaffected by the hybridization capture process.

All characterized variants were determined from the reads corrected using the UMI algorithm at the expected allele frequencies in the replicate libraries processed with HD701, as depicted in Table 2 below.

Gene	Amino acid change	Expected allele frequency	Observed allele frequency: 50 ng input (HD701)		Observed allele frequency: 5 ng input (HD701)	
ARID1A	P1562fs	33.5%	31.0%	31.0%	34.0%	41.0%
EGFR	G719S	24.5%	25.0%	25.0%	24.0%	28.0%
PI3KCA	H1047R	17.5%	17.0%	15.0%	18.0%	14.0%
KRAS	G13D	15.0%	17.0%	12.0%	18.0%	11.0%
BRAF	V600E	10.5%	7.0%	7.0%	8.0%	5.0%
PI3KCA	E545K	9.0%	5.0%	7.0%	8.0%	7.0%
KRAS	G12D	6.0%	4.0%	8.0%	4.0%	5.0%
EGFR	L858R	3.0%	1.0%	3.0%	2.0%	1.0%
EGFR	T790M	1.0%	0.5%	1.0%	0.3%	0.6%

Table 2. Detection of characterized variants at known allele frequencies. The consensus read sequences were identified using fgbio bio tools (see Methods for details). The variants were called using Vardict and annotated by SnpEff. Only validated substitution variants in the Horizon Discovery Quantitative Multiplex Reference Standard (Cat. # HD701) were reported.

Detection of multiple low-frequency variants from cfDNA surrogates

The sensitivity of the assay was evaluated by calculating the detection rate of low-frequency variants in a reference standard. ThruPLEX Tag-Seq HV libraries were generated from 10 ng of Quan-Plex Patient-Like ctDNA Reference Standard (AccuRef) containing a series of characterized mutations occurring at 0%, 1%, and 5%. Cancer-related genes were enriched using an IDT xGEN Pan Cancer Panel. The hybridization resulted in high enrichment of the targets, as ~79% of the reads were on or near targets for 10-ng inputs. As depicted in Table 3, the average coverage of the targeted regions from the reads deduplicated by coordinates or by UMI consensus was similar and concordant with the PCR duplication reported by Picard MarkedDuplicates.

Input (ARF-1003CT)	Total reads	%reads mapped to human genome	%reads on or near baits	%PCR duplicate	Average target coverage before deduplication	Average target coverage after deduplication by coordinates	Average target coverage after UMI collapse
0%	30M	99%	79%	71%	2,935	594	506
1%	30M	99%	79%	67%	2,879	694	578
5%	30M	99%	79%	67%	2,899	683	583

Table 3. Hybridization metrics. Libraries were prepared in duplicate from 10 ng of human cell-free DNA surrogate (Accuref Quan-Plex Patient-Like ctDNA Reference Standard; Cat. # ARF-1003CT). The ThruPLEX Tag-Seq HV libraries were enriched using the xGEN Pan Cancer Panel v1.5 (IDT, Cat. # 1056205) and sequenced on an Illumina NextSeq 500. The reads were downsampled to 30M reads and then aligned to the human genome (HG19) using bowtie2. The reads were deduplicated using Picard MarkDuplicates or collapsed using fgbio bio tools (see Methods for details). Hybridization metrics were determined using Picard HsMetrics.

Table 4 below displays two replicate measurements of the expected variants at the frequency identified in the reads collapsed using the UMI algorithm detailed in the Methods. The variants were identified around their respective expected allele frequency of 1% or 5%. No variants were detected in the negative control (0%) at the positions of the characterized variants.

Gene	Amino acid change	Expected allele frequency 0%		Expected allele frequency 1%		Expected allele frequency 5%	
EGFR	G719S	0.0%	0.0%	1.4%	1.5%	6.2%	5.9%
EGFR	L858R	0.0%	0.0%	1.0%	1.6%	5.5%	6.1%
EGFR	L861Q	0.0%	0.0%	0.7%	1.6%	2.2%	2.2%
KRAS	G12D	0.0%	0.0%	1.3%	0.5%	4.1%	3.7%
NRAS	Q61K	0.0%	0.0%	2.9%	1.9%	6.3%	4.8%
PIK3CA	E545K	0.0%	0.0%	0.4%	1.0%	4.2%	6.3%

Table 4. Variants identified using UMI algorithm. Variants were called using Vardict and annotated using SnpEff. Only validated substitution variants in the Quan-Plex Patient-Like ctDNA Reference Standard (Cat. # ARF-1003CT) were reported.

A large number of inconsistent variants with allele frequencies above 0.5% are observed in not deduplicated files or files deduplicated solely by coordinates. However, the number of variants identified in the aligned files generated from the consensus sequences from the UMI correction was reproducibly noticeably lower, demonstrating the power of the UMIs in correcting for amplification and sequencing errors. More specifically, a significant number of variants of low or random allele frequencies were reported in near proximity of the expected *EGFR* L858R variant in the files not deduplicated or deduplicated by coordinate only. The consensus sequence reads from the UMI collapsing and filtering cleaned up false-positive variants but retained expected variants.

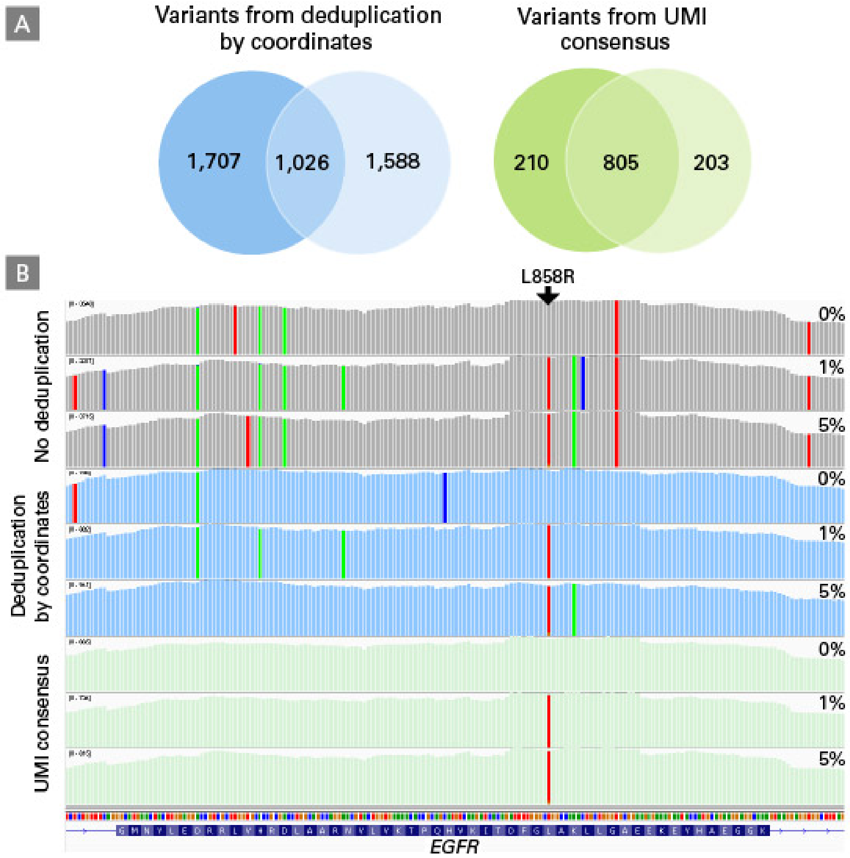


Figure 6. False-positive variants are filtered out when using the fgbio UMI collapsing algorithm. Panel A. The number of variants identified over 0.5% in files deduplicated by coordinates and reads collapsed using UMI algorithm. **Panel B.** Visualization of the variants above 0.5%.

The low GC bias observed in ThruPLEX Tag-Seq HV empowers the even coverage of GC- and AT-rich regions. The 5'UTR of *CEBPA* is known for its very high GC content overlapping a CpG island. As depicted in Figure 7, the coverage is not affected by the varying GC content from replicate 10-ng DNA inputs.

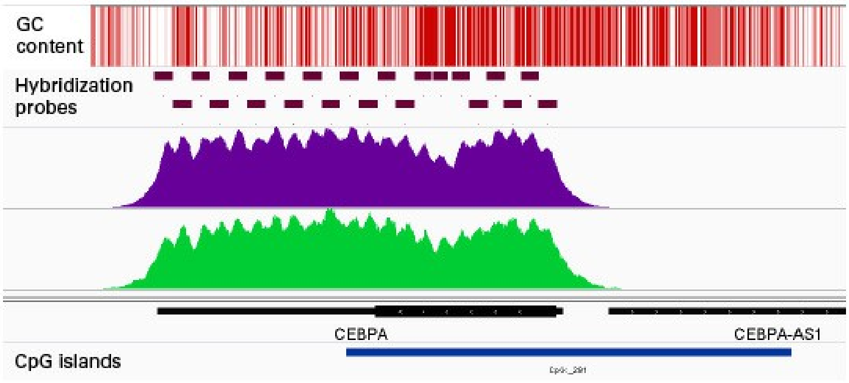


Figure 7. Even coverage of GC-rich regions. Constant read coverage is observed along the *CEBPA* gene. The BAM files were visualized using the Integrated Genome Viewer (IGV, Broad Institute).

Conclusions

The ThruPLEX Tag-Seq HV kit has been engineered and optimized to generate DNA libraries with high molecular complexity and balanced GC representation from input volumes of up to 30 µl. The entire three-step workflow takes place in a single tube or well, enabling the conservation of valuable samples while ensuring accurate sample tracking. No intermediate purification steps and no sample transfers are necessary, thus preventing handling errors and loss of valuable samples. With the incorporation of unique molecular identifiers, ThruPLEX Tag-Seq HV offers added power in correcting for polymerase and sequencing errors—key advantages when detecting low-frequency variants. The UMIs included in the system have been carefully selected for even representation and the best possible performance both in error correction and sequencing results on Illumina platforms.

Methods

DNA preparation

Horizon Discovery (HD701) reference gDNA was sheared at 250 bp using a Covaris M220. Sheared input material was evaluated for size distribution using an Agilent 2100 BioAnalyzer. The concentration of these samples was measured using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific).

Library preparation and hybridization capture

Libraries were prepared following the ThruPLEX Tag-Seq HV kit user manual. Library size profiles were assessed on the Agilent 2100 BioAnalyzer and quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific). The ThruPLEX Tag-Seq HV libraries were sequenced on an Illumina MiSeq for low-pass whole-genome sequencing or enriched by hybridization capture using the xGEN Pan Cancer Panel v1.5 (IDT, Cat. # 1056205) according to the manufacturer's instructions, and then sequenced on an Illumina NextSeq 500.

Data analysis

Illumina adapter trimming was performed using Trimmomatic, and reads were downsampled using setk. The reads were aligned to the human genome assembly HG19 using bowtie2. Duplicates, alignment metrics, insert size, and GC bias were calculated using Picard MarkDuplicates, CollectAlignmentSummaryMetrics, CollectInsertSizeMetrics, and CollectGcBiasMetrics, respectively. Hybridization capture metrics were determined using Picard CollectHsMetrics. Consensus reads based on UMI algorithm were obtained using fgbio tools, and variant calling was performed using Vardict. Additional information can be found here.

Related Products

Cat. #	Product	Size	License	Quantity	Details
R400742	ThruPLEX® Tag-Seq HV	24 Rxns		*	
ThruPLEX Tag-Seq HV uses a simple, three-step workflow to generate high-complexity DNA libraries with unique molecular tags from standard samples and challenging sample sources such as FFPE and cell-free plasma DNA. This product contains reagents for 24 reactions and includes unique dual indexing (UDI) primers. Documents Components					
R400743	ThruPLEX® Tag-Seq HV	96 Rxns		*	

Add to Cart