A single workflow for targeted analysis of CNVs, SNPs and mutations – the Ovation® Target Enrichment System

Luke Sherlin, Doug Amorese, Stephanie Huelga, Ashesh Saraiya, I-Ching Wang, Joe Don Heath
NuGEN Technologies, Inc., 201 Industrial Road, Suite 310, San Carlos, CA 94070

Introduction

Much of the effort in targeted resequencing today is focused on studying cancer and inherited disease. The aberrations most important in testing, diagnosis, and treatment include translocation events, (including gene fusions), copy number alterations, and single nucleotide variations. Researchers continue to use independent approaches and techniques in order to interrogate these aberrations. This approach can be complex and costly, requiring expertise in multiple assays and consuming limited sample with multiple, sample-hungry workflows. The NuGEN Ovation® Target Enrichment System is a single, simple workflow suitable for use in the targeted analysis of a wide range of genomic markers.

Here we describe the features and performance of the Ovation Target Enrichment System, a system that can significantly reduce the time and effort required to answer biologically and clinically relevant questions.

Single Primer Enrichment Technology

Single Primer Enrichment Technology (SPET) is a novel approach for targeted resequencing of genomic DNA or cDNA derived from RNA. The streamlined assay enables the interrogation of multiple aberrations and is suitable for a wide range of target sizes from a few kilobases to over 20 megabases. Panels with comprehensive cancer panel content are available, and the system can be customized to contain any customer-defined content.

The technology is highly flexible and can be applied to the targeted analysis of a wide range of genomic markers including mutations, SNPs, indels, gene fusions, alternately spliced transcripts and copy number variants. Multiple mutations can be interrogated in the same sequencing dataset using the SPET approach. The method uses a set of targeting probes that hybridize to the target regions and are extended through the regions of interest. The detailed mechanism is described in **Figure 1**. The approach eliminates the difficulty of designing specific PCR primer pairs and maintains high specificity of recovered target sequences in the final library. In addition, the technology lends considerable flexibility to the regions to be analyzed, as it enables interrogation of unknown regions adjacent to known sequence. This is most relevant for applications where unknown elements are transposed or inserted, such as detection of fusions, translocations, or insertion of transposable elements, trans-genes, or viruses.

Description of SPET workflow

(Refer to Figure 1)

- **Fragmentation to ~500 bp**
  gDNA or cDNA is randomly fragmented to ~500 bp where a proportion of the fragments will contain a target sequence to be enriched (thick blue bar in Figure 1). Fragmentation to 500 bp ensures optimal read density within exons.
- **Ligation of Indexed FWD Adaptor**
  After end repair, forward adaptor sequences are ligated to both ends of the fragments. This effectively prepares each individual fragment as a potential library molecule. In addition, each strand receives a forward adaptor, so that primers that are uniquely designed to either the top or bottom strand can create unique libraries. At this point each sample is barcoded to enable multiplexing.
- **Annealing of Targeting Probes/REV Adaptor**
  Specific targeting probes are hybridized with the adaptor ligated DNA to anneal probes to the 3′ probe landing zone. The enrichment probe pool contains oligonucleotides that anneal selectively to a probe landing zone (red bar in Figure 1) that is within 10–60 bases 3′ of the target region. This design strategy takes advantage of unique sequences adjacent to exons, thus enabling unique interrogation and resolution of homologous genes. Note that probes independently target 3′ probe landing zones for both strands of the target region. The targeting probes have a portion of the reverse sequencing adaptor at their 5′ end.
- **Extension of Probes**
  Annealed probes are extended using a polymerase creating a complementary strand to the target region and continuing through the forward adaptor resulting in library templates containing the target region and both forward and reverse adaptor sequences. The location of the reverse adaptor sequence in relation to the target region is defined by the location of the complementary sequence within

NuGEN
imagine more from less®

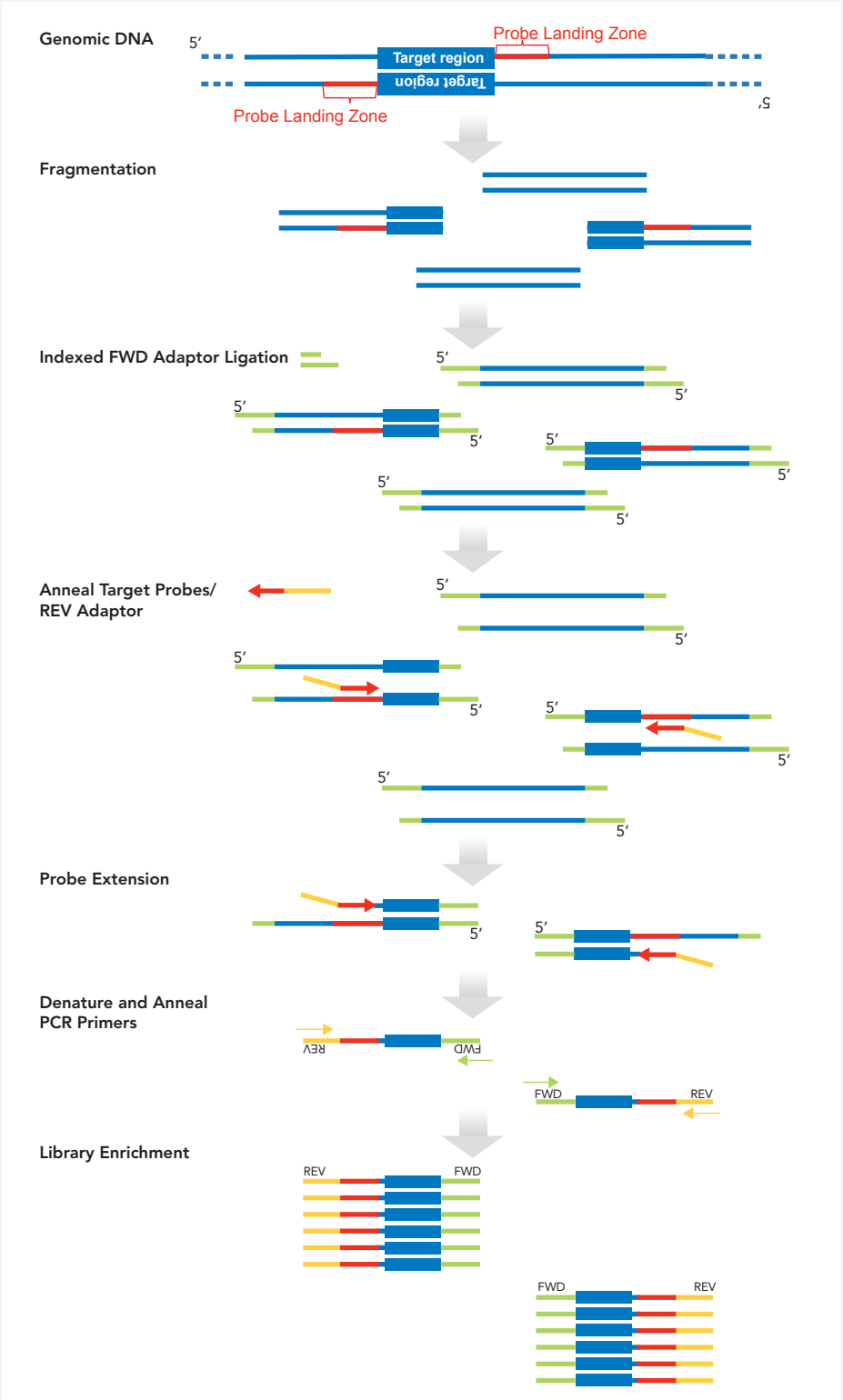

FIGURE 1. Overall mechanism of action of Single Primer Enrichment Technology as applied to a typical exon target enrichment design.

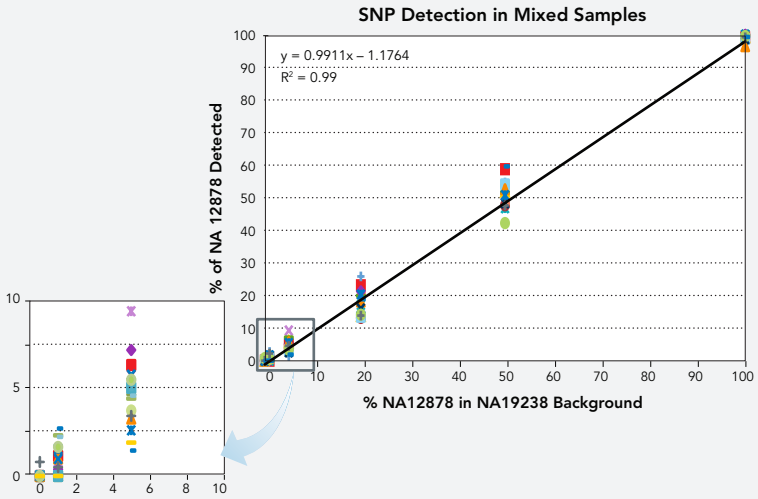

FIGURE 2. Human genomic DNA samples, NA19238 and NA12878, were fragmented by Covaris acoustic shearing and quantified by Qubit dsDNA HS Assay (ThermoFisher Scientific). Fragmented DNAs of NA19238 and NA12878 were blended at 1:99, 1:19, 1:4, and 1:1 ratios. Ovation Cancer Panel target enriched libraries were made with 100 ng of blended DNA samples or 100 ng of fragmented NA19238, and were sequenced on Illumina MiSeq sequencer.

the Probe Landing Zone, while the location of the forward adaptors in relation to the target region is determined by where random fragmentation occurred.

- **Denature and Anneal PCR Primers / Library Enrichment** Following denaturation, PCR primers are annealed completing forward and reverse adaptors and libraries are amplified by PCR creating a target enriched library ready for sequencing.

Sensitivity of Detection

Limits of Sensitivity

Working with clinical samples often requires analyzing DNA from a mixed population of cells. In many cases it is unknown what fraction of the sample is expected to contain an aberration, requiring an assay with a high degree of sensitivity. In order to demonstrate the sensitivity of the Ovation Target Enrichment System, a surrogate system was devised using mixtures of two HapMap samples (NA12878 into NA19238) at various ratios. After enrichment using the Ovation Cancer Panel Target Enrichment System, each of six libraries with a different NA12878:NA19238 ratio were sequenced and analyzed for allelic frequencies f 14 SNPs unique to one of the HapMap samples (NA12878). A sample with 100% NA12878 showed prevalence of all SNPs with a high frequency, as expected (**Figure 2**). As the proportion of NA12878 decreased, the SNPs were all detected reliably down to a level of only 5% NA12878. It is only in the sample with 1% NA12878 where any of the SNPs were no longer detected, indicating a sensitivity of the assay to <5% at a sequencing depth of at least 80X. This demonstrates a high degree of sensitivity to detect mutations even when a low proportion (<5%) of cells contain the aberration.

Input Flexibility: FFPE Samples

The sensitivity of the Ovation Target Enrichment System has also been demonstrated with formalin-fixed paraffin-embedded (FFPE) samples. Libraries from FFPE samples containing qPCR-validated mutations with allelic frequencies ranging from 3% to 33% were prepared with 50 ng and

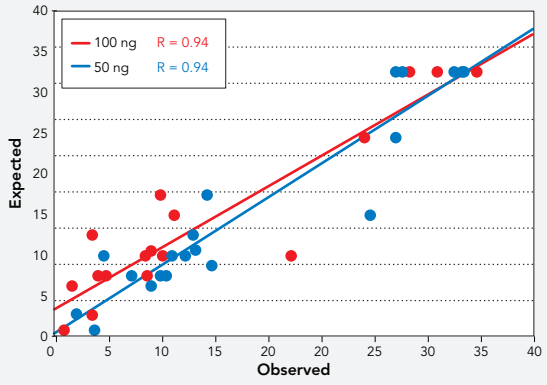

FIGURE 3. FFPE Quantitative Multiplex Reference Standard (Horizon Dx) was fragmented and prepared at inputs of 50 ng and 100 ng using the Ovation Cancer Panel Target Enrichemnt System. Libraries were sequenced using single-end 100 bp reads on a MiSeq, and allelic frequencies of validated mutations were measured and compared to expected frequencies.

Target	Reads	% Aligned	% On Target	% Uniformity	% N6 Duplicates	Mean Coverage
Ovation Cancer Panel 2.0	3 M	98.8	84.1	82.3	16.0	33.9

TABLE 1: Alignment metrics for Ovation Cancer Panel 2.0 Target Enrichment System. Libraries were generated using 100 ng sheared NA12878 genomic DNA. Alignment was performed to hg19 build with default bowtie2 parameters. % Bases On Target = % bases mapped to the target region +100 bp on either side of target. % Uniformity = % of bases covered by at least 20% of the mean. %N6 Duplicates were calculated using the N6 method described in the Supplemental Material. Mean coverage was calculated using the Picard CalculateHsMetrics routine.

TP	FP	FN	TN
1040	1	2	1571531

TABLE 2: SNP calls generated using the GATK default pipeline from data described in Table 1.

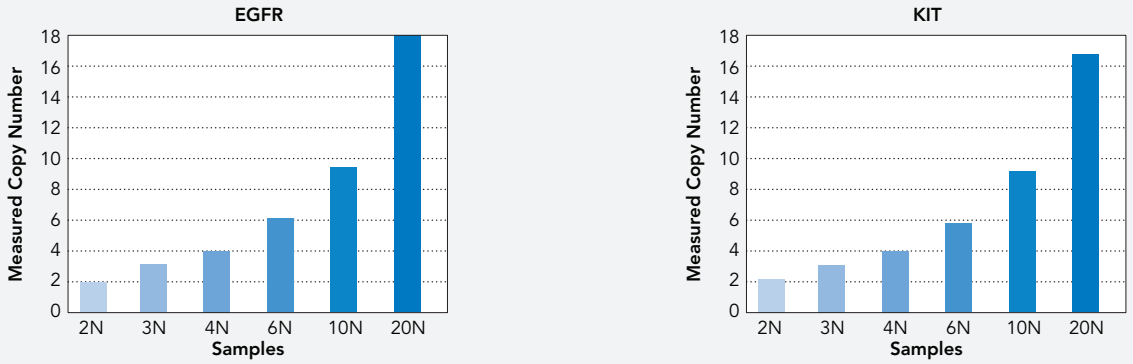

FIGURE 4. Copy number measurements using quantitated BAC clone spike-ins.

100 ng DNA inputs using the Ovation Cancer Panel Target Enrichment System. Correlation between the expected values and the experimentally determined allelic frequencies is shown in **Figure 3**. At both 100 ng and 50 ng inputs, the Pearson correlation coefficient is greater than 0.9, demonstrating the ability to reliably detect important mutations in a heterogeneous sample.

Accuracy of Detection with the Ovation Cancer Panel 2.0 Target Enrichment System

Variant calling accuracy is of the utmost importance in any experiment or assay that uses Next Generation Sequencing. The Ovation Target Enrichment System exhibits a high degree of accuracy. To demonstrate this feature, 100 ng genomic DNA of HapMap sample NA12878 was prepared using the Ovation Cancer Panel 2.0 Target Enrichment System. The enriched libraries were sequenced on a MiSeq and after preprocessing of the sequencing data, including quality trimming and probe trimming, alignment to the human genome and PCR duplicate removal, the GATK analysis package (The Broad Institute) was used to identify variants. Performance metrics with this HapMap dataset showed a high alignment rate, high on target rate, and good uniformity of coverage, as well as a low PCR duplicate rate (see **Table 1**). The variants identified were compared to all target bases that had at least 20X read coverage to the NIST high-confidence variant annotation downloaded from the Genome in a Bottle Consortium (https://sites.stanford.edu/abms/giab). **Table 2** summarizes the concordance results for single nucleotide polymorphisms. Despite modest sequencing depth (mean coverage ~34X), an extremely low rate of false calls were made.

Copy Number Alteration

Accurate Copy Number Detection

In addition to accurate and sensitive mutation detection, the Ovation Target Enrichment Systems enable measurement of gene-level copy number alterations (CNAs) in the same sequencing dataset. To demonstrate the ability to accurately detect various copy number levels, well-characterized commercial samples were spiked with bacterial artificial chromosome (BAC) clones that contained additional copies of the genes KIT and EGFR. This method enabled highly accurate quantitation of the genes by qPCR to act as a point of reference for detection with the Ovation® System. The BAC clones were added to the human male DNA sample to produce 3, 4, 6, 10, and 20 copies relative to the "wild-type" representation of the gene at 2 copies. 100 ng of each mixture was pre-

Blend	Gene	Digital PCR Copy Number	Measured Copy Number	P-Value
1	ERBB2 (HER2)	23.4, gain	24.96	3.84E-32
	MET	1.97	2.64	NS
	EGFR	2.91, gain	3.29	2.88E-09
	AURKA	8.52, gain	9.51	8.93E-09
2	ERBB2 (HER2)	1.8	1.98	NS
	MET	12.8, gain	17.69	7.04E-23
	EGFR	2.28	2.58	NS
	AURKA	2.35	2.85	NS
3	ERBB2 (HER2)	1.36	1.61	NS
	MET	2.02	2.81	NS
	EGFR	10.7, gain	12.74	3.31E-29
	AURKA	3.35, gain	4.05	NS
4	ERBB2 (HER2)	8.88, gain	9.71	6.90E-29
	MET	5.41, gain	8.02	1.17E-18
	EGFR	5.32, gain	5.88	1.68E-22
	AURKA	4.4, gain	5.42	1.81E-06

TABLE 3: Four reference samples with varying copy numbers of the genes ERBB2, MET, EGFR, and AURKA were prepared using the Ovation Cancer Panel 2.0. 100 ng sheared DNA was used as input, and libraries were sequenced using 150 bp single end reads on the MiSeq platform. Bonferroni correction across all genes was applied to the P-value, with a pre-corrected threshold of 0.05 considered significant (NS = not significant).

pared using the Ovation Cancer Panel 2.0 Target Enrichment System, sequenced, and analyzed using the simple counting method (described in detail in Supplemental Material F.) Because of the consistent performance of the Ovation Target Enrichment System probes, a simple probe-based read counting method can be applied to compute copy-number values from the processed sequencing data and compare that to the control sample with no additional copies beyond the native two copies of the gene. **Figure 4** shows the number of copies detected using this method, demonstrating accurate copy number detection from as few as 3 copies to 20 copies.

Copy Number Detection in Engineered Cell Line Samples

To further test the accuracy of the copy number measurements, the Ovation Cancer Panel 2.0 Target Enrichment System was used to produce targeted libraries from 100 ng genomic DNA from 4 Horizon Diagnostics cancer refer-

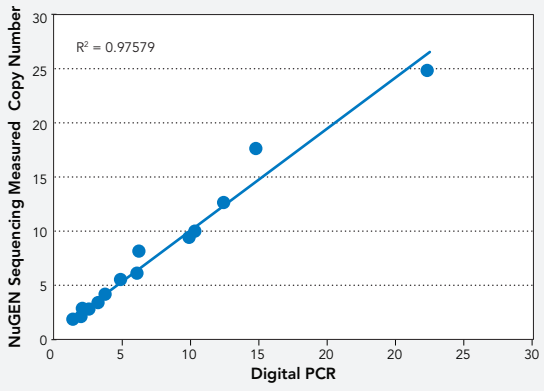

FIGURE 5. Overall correlation between digital PCR and NGS copy number calculation for four characterized reference samples. Copy number changes by sequencing and digital PCR are taken from Table 3.

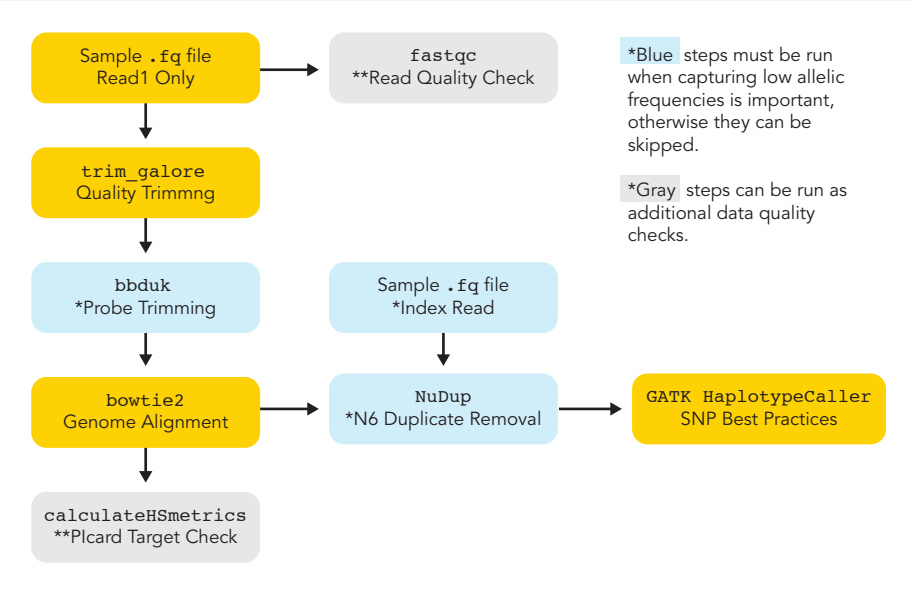

FIGURE 6. General workflow for data generated with the Single Primer Enrichment Technology.

ence samples (www.horizondiscovery.com). These reference samples have validated copy number levels, as measured by droplet (digital) PCR, for 4 genes; ERBB2 (also known as HER2), MET, EGFR and AURKA. **Table 3** shows the results of both the droplet qPCR and sequencing. In all samples, there is excellent correlation between the two methods (**Figure 5**), further demonstrating accurate measurement of copy number changes using the Ovation Target Enrichment System.

Conclusions and Summary

Single Primer Enrichment Technology is a simple and effective means of enriching targets of interest. Simple solutions with minimal manipulation are key to achieving accurate data. Concordance of SNPs levels and copy number changes in mixed samples illustrate the underlying power of the method. Since all of the general attributes of the starting material are maintained, single data sets can be mined for multiple purposes. Rather than running multiple assays to gather SNP and CNV data, the data can be retrieved from the same sequencing run by examining it through an appropriate analysis pipeline. Libraries originally generated to detect SNPs can be reanalyzed to look for changes in copy number.

Supplemental Material

A. Duplicate Identification

The use of a random N6 sequence adjacent to the 8 base sample index facilitates accurate duplicate identification when using only single end reads.

After identification duplicates may be removed before further analysis or used for a highly accurate consensus read of the target region for low frequency events.

B. Data Preprocessing

The unique features of the Ovation Target Enrichment system offer many advantages in bioinformatics analysis. While standard pipelines can be used for analysis, outlined below are best practices for taking advantage of these unique features to enable accurate variant calling and measurement of copy number alterations. The overall best practices workflow is summarized in **Figure 6**.

C. Sequencing and Data Retrieval

Preparation of the data for both variant detection and copy number measurements requires a basic understanding of the library structure. **Figure 7** shows the structure of a typical OTE library, which differs from the typical adaptor structure. In order to best enable use of the N6 random sequence, the standard index priming sequence, barcode,

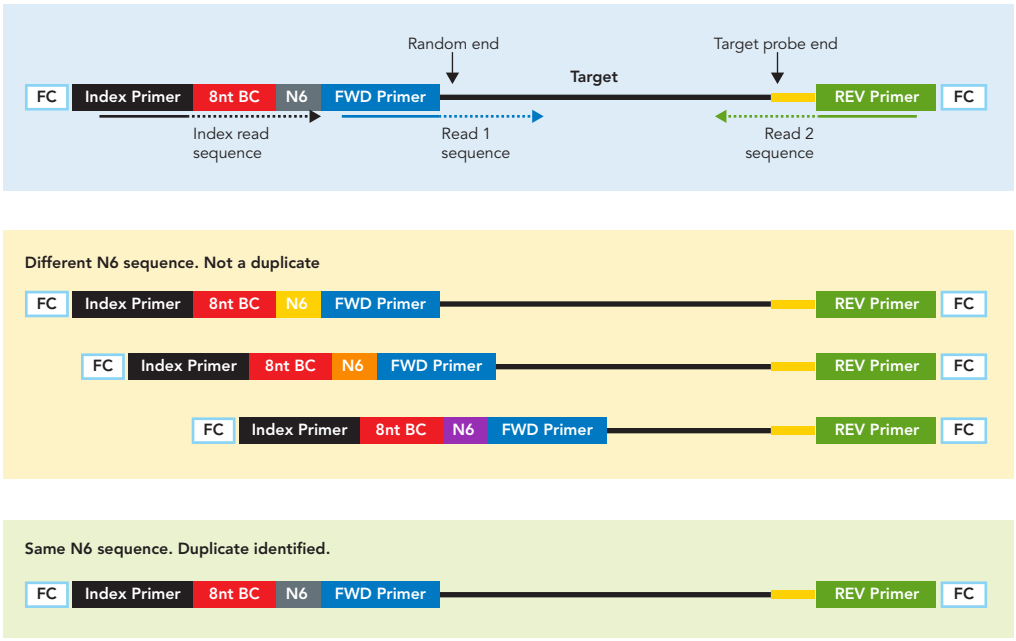

FIGURE 7. Library structure and resolution of PCR duplicate reads using a random N6 sequence.

and N6 random sequence have been placed adjacent to the reverse read priming region. The primary reason for this rearrangement is to enable direct association between the N6 random sequence and the site of ligation. By placing the N6 random sequence adjacent to the barcode, this sequence can be identified with a single index read of 14 bp. This can require modifications to the sequencer setup, and will be platform-dependent. Recommendations are outlined below.

MiSeq Instruments

Parsing multiplex runs using the MiSeq built-in Illumina software replaces the barcode sequence from each library with a numerical substitute. This removes the duplicate information provided by the N6 sequence present after the barcode. To retrieve this information using the MiSeq instrument, we recommend modification of the MiSeq config file to allow generation of an index fastq file during data analysis. This will generate a 14-base index file that is compatible with the NuGEN application. If you are not familiar with editing the config file, we recommend requesting assistance from Illumina Technical Support to make this modification. The steps are as follows:

1. Stop the MiSeq Reporter process.
2. Locate the "MiSeq Reporter.exe.config" file located in C:/Illumina/MiSeq Reporter

3. Open config file and search for a line that reads:

```
"<add key="CreateFastqForIndexReads" value="0"/>"
```

If this line is present, change the value from "0" to "1"

If this line is not present, add the line to the config file with the value set to "1"using the add keys function under the AppSettings tab.

3. Restart the MiSeq reporter process.
4. Requeue the run for data analysis if required.

Other Illumina Sequencers

Use the method described below to generate the read and N6 index fastq files for use with the Ovation Target Enrichment System Data Processing Application using bcl2fastq2 version 2.17.0 (formerly CASAVA).

1. Navigate to the location of the run folder (referred to as RunFolder in this document) and rename the sample sheet (i.e. SampleSheet.csv.bak).
2. To generate the run and index fastq files use the following command:

`/usr/local/bin/bcl2fastq -- runfolder-dir . --output-dir ./Data/Intensities/BaseCalls/ --use-bases-mask y*,y* --minimum-trimmed-read-length 0` (for paired end reads use "`--use-bases-mask y*,y*,y*`").

Note: In order to parse the data during the fastq file generation, modify the sample sheet to remove the six N's located at the end of the barcodes. Run bcl2fastq using the command `/usr/local/bin/bcl- 2fastq --runfolder-dir . --output-dir ./Data/Intensities/ BaseCalls/ --sample-sheet SampleSheet.csv --usebases-mask y*,i8y* --minimum-trimmed-read-length 0` (for paired end reads use "`--use-bases-mask y*,i8y*, y*`"). This command will produce an R1 fastq file with the forward read, an R2 fastq files containing just the N6 information and, if present, an R3 fastq file containing the reverse read.

3. The fastq files will be located in /RunFolder/Data/ Intensities/BaseCalls/ unless specified otherwise.

The generated fastq files can be uploaded to BaseSpace for input into the Ovation Target Enrichment System Data Processing Application.

NextSeq Considerations
When using the NextSeq platform, it is important to take some additional things into consideration, and to consult with a local technical representative to obtain the most current software and reagent recommendations. Due to the unique labeling approach on this platform it is not recommended to use barcodes that begin with the sequence 'GG' when performing low levels of sample multiplexing. Due to the sensitivity of the system to flowcell density, we also recommend targeting a lower load density, particularly when working with small target size designs.

D. Read Trimming

In addition to standard quality trimming, the OTE library structure also requires removal of a standard linker sequence from any reverse reads in order to enable optimal alignment. As described in **Figure 7**, the linker sequence is immediately followed by sequence that is derived from the primers used in the hybridization. For best variant detection results, it is recommended to trim this 50-mer sequence immediately after the linker sequence from any reverse read. It is also recommended to remove this probe-derived sequence from forward reads that are of sufficient length to extend to a primer sequence. These sequences are provided to customers in a FASTA format that can be used for trimming. A variety of different programs are available to perform this type of trimming, and one example is the 'BBDuk' program from the 'BBTools' package, using a line command as follows:

```
bbduk.sh in=R1.fq out=R1_trimmed.fq ref=probeSeqs_ETxxxx.fasta hdist=1 ktrim=r rcomp=f k=31 mink=11 qtrim=r trimq=20 minlen=20
```

E. Alignment & Duplicate Identification

Alignment can be performed using standard alignment tools (such as bowtie2, bwa, or others). In order to identify PCR duplicates using the N6 random sequence, a 14-bp index read must be captured during sequencing (see "Sequencing and data retrieval", above). NuGEN provides two different software tools to identify PCR duplicates with this method. The first, intended for users familiar with bioinformatics tools and analysis, is the script package 'NuDup' (http://nugentechnologies.github.io/nudup/). This tool is designed to perform duplicate identification of aligned reads using the N6 information in the index read. The input files are as follows:

- SAM/BAM alignment file
- Fastq index read file containing the N6 random sequence

The tool is compatible with both single-end and paired-end reads, and generates the following output files:

- Text summary of duplication rate
- BAM alignment file with PCR duplicates marked
- BAM alignment file that retains only a single copy of any PCR duplicates

The second tool is the Ovation Target Enrichment Data Processing Application for BaseSpace®, available in the Illumina BaseSpace environment. The application performs all of the above functions (demultiplexing, trimming, alignment, and PCR duplicate identification) in a single, streamlined application, producing summary alignment metrics and alignment files that can be used in downstream mutation, copy number, and translocation analysis. A summary of the overall workflow is shown in **Figure 8**, and a detailed guide for its use is available www.nugen.com/sites/default/files/M01399v1_Technical_Report_Ovation_Target_Enrichment_BaseSpace_Application.pdf

F. Copy Number Calculations

Copy number alterations can quickly and easily be calculated using single-end or paired-end sequencing data. After duplicate reads have been removed (see methods above), the following steps can be followed to compute gene-level copy number changes.

1. Determine read counts for each probe region

Use of the BEDTools Suite routine 'coverageBed' with the NuGEN-supplied "probePlus300" file will provide quantitation of the number of reads that fall within a 300 bases of the 3´ end of each probe. This routine can be run using a command such as:

```
coverageBed –abam R1_mysample_dedup.bam –b probePlus300.bed > R1_mysample_dedup_coverage.txt
```

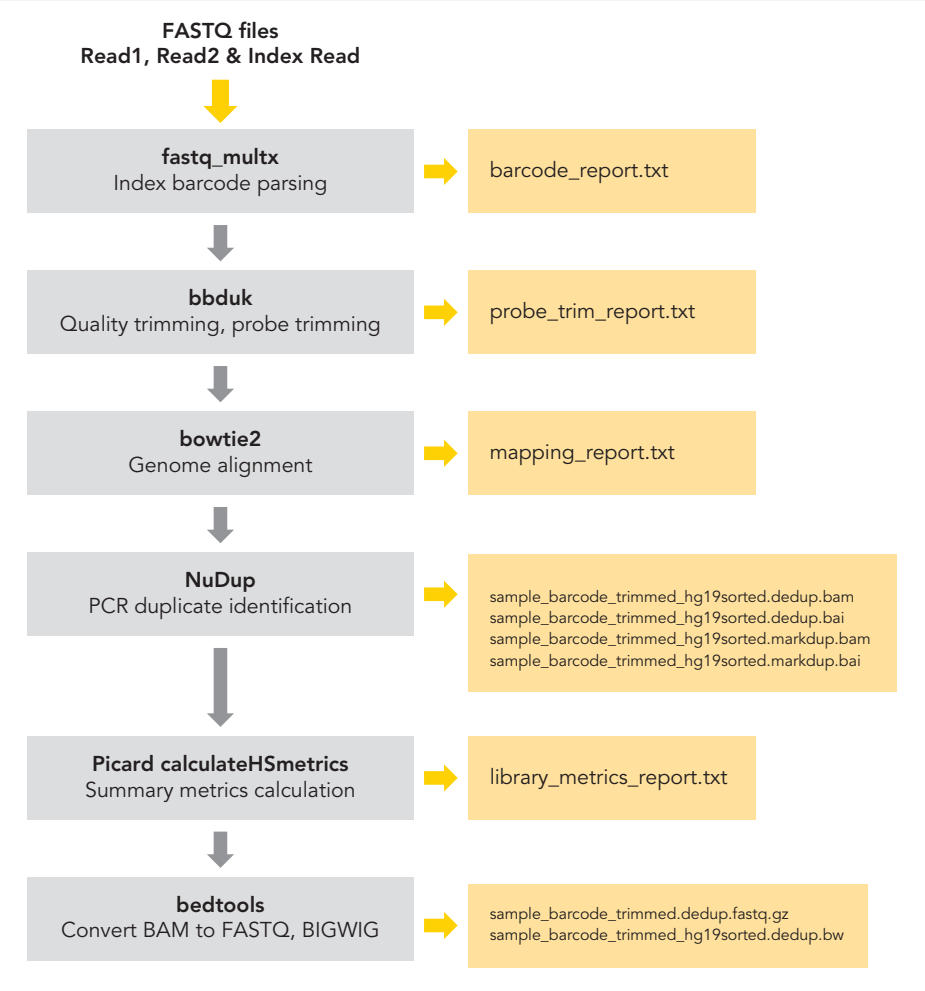

FIGURE 8. Pre-processing workflow and output files.

The output from this routine is the probePlus300 bed file entry with the following appended columns: COL7: The number of reads that overlapped (by at least one base pair) the probePlus300 interval COL8: The number of bases in the region that had non-zero coverage COL9: The length of the region (always 300) COL10: The fraction of bases in the region that had non-zero coverage Values from COL7 are used in the next step of the process.

2. Normalize internal reads

The output of coverageBedmust first be normalized for total reads in the probe regions. This is easily accomplished using a spreadsheet program such as Microsoft Excel. Divide each value in COL 7 by the SUM(COL7) to normalize for fraction of total reads.

3. Normalize probe data

The read-normalized output from Step 2 should then be adjusted using the normalization factors file provided by NuGEN. Normalization is accomplished by dividing the read normalized values by the corresponding probe normalization factor from the normalization factors file. This step provides normalization of probe signal due to small variations in probe efficiency for each probeset. The number can then be multiplied by 2 to obtain a correspond to 2 copy counts.

4. Average normalized probe data

A single copy number value per gene is calculated by averaging the normalized values from Step 3 for each gene. Note that the number of probe regions for each gene can vary depending upon the size of the gene. This step is also easily accomplished using a spreadsheet program or a simple script. Microsoft Excel contains a "Group" feature that can simplify this calculation.

NuGEN Technologies, Inc.

Headquarters USA

201 Industrial Road, Suite 310
San Carlos, CA 94070 USA
Toll Free Tel: 888.654.6544
Toll Free Fax: 888.296.6544
custserv@nugen.com
techserv@nugen.com

Europe

P.O. Box 109
9350 AC Leek
The Netherlands
Tel: +31-13-5780215
Fax: +31-13-5780216
europe@nugen.com

For our international distributors contact information, visit our website

www.nugen.com