Molecular Identifier (MID) Analysis for Paired-End Sequencing

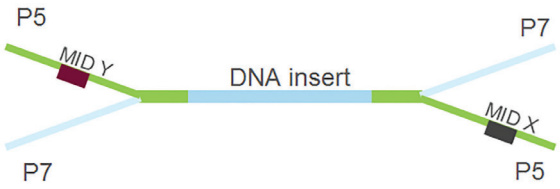
Catalog Nos.: 53216 & 53264

Name: Next Gen DNA Library Kit & Next Gen Indexing Kit

Description

Active Motif's Next Gen DNA Library Kit (Catalog No. 53216) and Next Gen Indexing Kit (Catalog No. 53264) are designed for the preparation of high complexity Next generation sequencing (NGS) libraries from double-stranded input DNA for use with Illumina® platforms. Libraries can be generated from as little as 10 pg DNA, or from as low as 100 ng DNA if preparing PCR-free libraries.

The advantage of the Next Gen DNA Library is inclusion of molecular identifiers (MIDs) to enable accurate removal of PCR duplicates from sequencing data. This helps to increase the number of unique alignments for more accurate data sets. The MID is a 9 base random N sequence that is added with the P5 adapter. Addition of the MID is strand-specific with each dsDNA insert receiving two MIDs (X and Y). The two MIDs cluster and sequence independently. Through bioinformatic analysis PCR duplicates can be removed from the data set, while fragmentation duplicates are preserved.



Schematic of a completed MID tagged and indexed DNA library molecule.

Application Notes

Guidelines are provided to process sequencing data sets using the MIDs. The MID (Reagent B2 MID) is incorporated during Ligation II of the protocol. The standard low throughput (LT) P7 adapters containing a single index for multiplexed sequencing are added during Ligation I of the protocol. Libraries are compatible with Illumina sequencing platforms. Different instruments may require different sample sheet set-up for correct processing. Guidelines are provided below for the NextSeq 500. Adjust as needed for your instrument and desired read length.

The following guidelines are provided for NextSeq 500 paired-end run using Illumina's High Output Kit (75 cycles):

1. Prepare the sequencing libraries using the protocol provided with the Next Gen DNA Library & Next Gen Indexing Kits.
2. Run NextSeq 500 instrument.
3. Prepare SampleSheet.csv (no adapter trimming, 8 bp for both indices, and 38 bases for each read).

   Note: Remove adapter sequences from sample sheet to avoid trimming by bcl2fastq. Libraries allow 9 bp for MID, but the max allowable for 38 bp reads AND having both indices of the same length (required for Illumina basespace) is 8 bp. In stand-alone mode, you can use 7 bp and 9 bp for the indices, respectively, but the bcl2fastq base mask must then be changed.

4. Run bcl2fastq to generate _r1, _r2 and _r3 for each sample. Download bcl2fast1 using the Illumina website.

```
bcl2fastq \
--use-bases-mask Y*,I*,Y*,Y* \
--minimum-trimmed-read-length 0 \
--mask-short-adapter-reads 0 \
-R $nextseq_run_dir \
-o $nextseq_run_dir/FASTQ \
--interop-dir $nextseq_run_dir/InterOp \
--reports-dir $nextseq_run_dir/Reports \
--stats-dir $nextseq_run_dir/Reports/html
```

5. Merge FASTQ lane data for _r1, _r2 and _r3 for each sample.

```
cat sample_L001_r1.fastq.gz \
sample_L002_r1.fastq.gz \
sample_L003_r1.fastq.gz \
sample_L004_r1.fastq.gz > sample_r1.fastq.gz
cat sample_L001_r2.fastq.gz \
sample_L002_r2.fastq.gz \
sample_L003_r2.fastq.gz \
sample_L004_r2.fastq.gz > sample_r2.fastq.gz
cat sample_L001_r3.fastq.gz \
sample_L002_r3.fastq.gz \
sample_L003_r3.fastq.gz \
sample_L004_r3.fastq.gz > sample_r3.fastq.gz
```

6. Merge _r1 and _r2 FASTQ files for each sample to append MID DNA sequences from _r2 to end of FASTQ header in _r1.

   After Step 5 above, there should be three FASTQ files for each sample. They will be named similar to "SAMPLE_r1.fastq", "SAMPLE_r2.fastq, and "SAMPLE_r3.fastq" respectively. The first four lines of the original files will look similar to the example below for SAMPLE_r1.fastq, SAMPLE_r2.fastq, and SAMPLE_r3.fastq respectively:

```
@NS500375:278:HFNC3BGXY:1:11101:5184:1051  1:N:0:ACTTGAA
TCCGANCGTTCTGGTGCCTTGTCCCCCATCAACACTTNCNNCACNCNNNNNNNNNANNNNCCTANCNTCCCTC
+
AAAAA#EEEE6EEAEEAEEEEA/EEEEE<66E66EE#E##EEE#E########E####E6E/#E#E6EEEE


@NS500375:278:HFNC3BGXY:1:11101:5184:1051  2:N:0:ACTTGAA
ACTTCAGGA
+
/AAAAEEEA


@NS500375:278:HFNC3BGXY:1:11101:5184:1051  3:N:0:ACTTGAA
ATGCTGAATCGNNNTCTAATGCGGNNNNNATACCTGATCAGGNNTACGGACTTTANNTTTACAGGAGCAACTC
+
AAAAA#EEEE6EEAEEAEEEEA/EEEEE<66E66EE#E##EEE#E########E####E6E/#E#E6EEEE
```

   Using a simple program (such as a perl script), combine the three files into two new files to result in a modified header in _r1 and _r3. See the example below:

```
@NS500375:278:HFNC3BGXY:1:11101:5184:1051_ACTTCAGGA  1:N:0:ACTTGAA
TCCGANCGTTCTGGTGCCTTGTCCCCCATCAACACTTNCNNCACNCNNNNNNNNNANNNNCCTANCNTCCCTC
+
AAAAA#EEEE6EEAEEAEEEEA/EEEEE<66E66EE#E##EEE#E########E####E6E/#E#E6EEEE
```

```
@NS500375:278:HFNC3BGXY:1:11101:5184:1051_ACTTCAGGA  3:N:0:ACTTGAA
ATGCTGAATCGNNNTCTAATGCGGNNNNNATACCTGATCAGGNNTACGGACTTTANNTTTACAGGAGCAACTC
+
AAAAA#EEEE6EEAEEAEEEEA/EEEEE<66E66EE#E##EEE#E########E####E6E/#E#E6EEEE
```

As shown in the example, the second line of _r2 (ACTTCAGGA) has been added to the header of _r1 and _r3 with the underscore separator. This is necessary to preserve the molecular ID sequence after the BAM mapping in the BAM QNAME field.

7. Trim adapters using tool of choice (example shown):

```
trim_galore \
--adapter $AdapterSeq \
--adapter2 $AdapterSeq \
--path_to_cutadapt $cutadapt_path/cutadapt \
--paired \
SAMPLE_r1.fastq \
SAMPLE_r3.fastq
```

8. Map FASTQ data to reference genome to create BAM using tool of choice (*e.g.* BWA).
9. Sort BAM using tool of choice (*e.g.* SAMtools).
10. Run MID de-duping to mark duplicates in BAM.

    Please contact Active Motif Technical Support a 877-222-9543 or tech_service@activemotif.com to obtain the MID de-duping tool, or for suggestions for use with other Illumina platforms.

Technical Services

If you need assistance at any time, please call Active Motif Technical Service at one of the numbers listed below.

Active Motif North America
1914 Palomar Oaks Way, Suite 150
Carlsbad, CA 92008
USA
Toll Free: 877 222 9543
Telephone: 760 431 1263
Fax: 760 431 1351
E-mail: tech_service@activemotif.com

Active Motif Europe
Avenue Reine Astrid, 92
B-1330 La Hulpe, Belgium
UK Free Phone: 0800 169 31 47
France Free Phone: 0800 90 99 79
Germany Free Phone: 0800 181 99 10
Telephone: +32 (0)2 653 0001
Fax: +32 (0)2 653 0050
E-mail: eurotech@activemotif.com

Active Motif Japan
Azuma Bldg, 7th Floor
2-21 Ageba-Cho, Shinjuku-Ku
Tokyo, 162-0824, Japan
Telephone: +81 3 5225 3638
Fax: +81 3 5261 8733
E-mail: japantech@activemotif.com

Active Motif China
787 Kangqiao Road
Building 10, Suite 202, Pudong District
Shanghai, 201315, China
Telephone: (86)-21-20926090
Hotline: 400-018-8123
E-mail: techchina@activemotif.com

Visit Active Motif on the worldwide web at **http://www.activemotif.com**

At this site:

- Read about who we are, where we are, and what we do
- Review data supporting our products and the latest updates
- Enter your name into our mailing list to receive our catalog, *MotifVations* newsletter and notification of our upcoming products
- Share your ideas and results with us
- View our job opportunities

Don't forget to bookmark our site for easy reference!