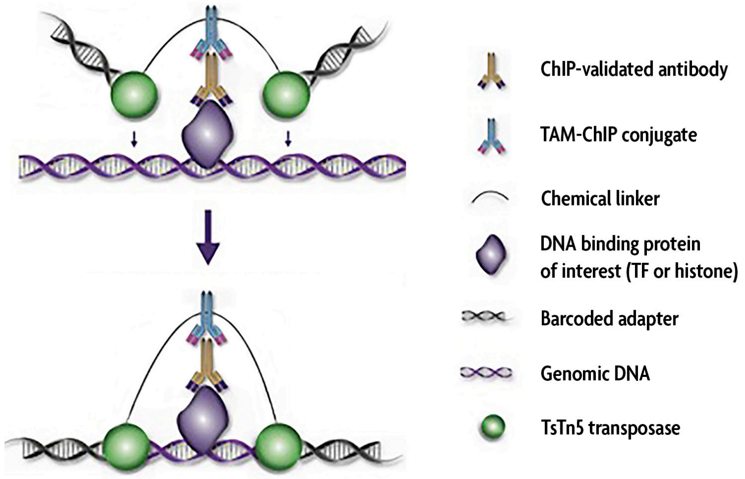
Molecular Identifier (MID) Analysis for TAM-ChIP Single-Read Sequencing

Catalog Nos.: 53126 & 53127
Name: TAM-ChIP antibody conjugate

Description

Active Motif's TAM-ChIP technology combines antibody directed genomic targeting and NGS library preparation in one step. First, a ChIP-validated antibody is used to target a genomic region of interest, such as a histone or transcription factor binding site. Then, a TAM-ChIP anti-species antibody that is conjugated to barcoded sequencing adapters and a TsTn5 transposase is added to each reaction. Activation of the transposase cuts the nearby DNA and pastes the antibody associated adapters into the DNA sequence. Following immunoprecipitation, enriched DNA is ready for library amplification and sequencing using Illumina® platforms.

The inclusion of unique molecular identifiers (MIDs, also known as UMIs) in the TAM-ChIP Index primers enables accurate removal of PCR duplicates from sequencing data. This helps to increase the number of unique alignments for more accurate data sets. The TAM-ChIP antibody conjugate also contains a unique barcode. Through bioinformatic analysis, individual samples are identified and PCR duplicates can be removed from the data set, while fragmentation duplicates are preserved.



Schematic of TAM-ChIP.

Application Notes

TAM-ChIP uses dual index sequencing that can be run as single-read or paired-end and is compatible with Illumina platforms and sequencing reagents. It is recommended to sequence 30 million reads per sample. Different Illumina instruments may require different sample sheet set-up for correct processing. Guidelines are provided below to create a sample sheet template for your specific instrument using the Illumina Experiment Manager (IEM). Modifications will need to be made to the template to add custom i7 and i5 index sequences.

Due to the tethering of the TsTn5 transposase to the TAM-ChIP antibody conjugate, there is a high likelihood that the transposase will insert the sequencing adapters into similar regions of the genome across your sample population. By utilizing the Molecular Identifiers (MIDs), sequencing reads that start at the same position can be identified as biological replicates instead of PCR duplicates. With the MID de-duplication tool from Active Motif, these biological replicates are retained and added back into the sequencing analysis. This increases the total number of sequencing reads up to 3-fold. This document contains instructions on how to prepare FASTQ files for de-duplication of the MIDs that are incorporated into the TAM-ChIP Index primers (i7 index) The Perl script to run the de-duplication can be obtained by contacting Active Motif's technical support team at tech_service@activemotif.com.

Sequencing Recommendations:

- Prepare the sequencing libraries using the TAM-ChIP Index primers provided with the TAM-ChIP antibody conjugate (Catalog Nos. 53126 & 53127) following the instructions provided on the lot-specific product data sheet.
- For sequencing, we recommend 30 million reads on an Illumina® platform.

Table 1: Instrument Selection

Instrument	Max Reads*	Max Read Length	i5 Index Orientation	Application
NovaSeq	3.3 B	2 x 150	Forward	NovaSeq FASTQ only
HiSeq 3000/4000	2.5 B	2 x 150	PE: Reverse Complement SR: Forward	HiSeq FASTQ only
HiSeq 2500	HO: 1.5 B (v3) - 2 B (v4) RR: 400 M	2 x 125 2 X 250	Forward	HiSeq FASTQ only
NextSeq	HO: 400 M MO: 130 M	2 X 150 2 x 150	Reverse Complement	NextSeq FASTQ only

*Output and read calculations are based on single flowcell.

- Each TAM-ChIP Index primer (i7) contains an 8 bp random molecular identifier followed by a 3 bp sample barcode (shown in bold in Table 2 below) for a total of 11 bp.

Table 2: TAM-ChIP i7 Index

	Forward Sequence
TAM-ChIP Index 1	NNNNNNNN**TCG**
TAM-ChIP Index 2	NNNNNNNN**GAC**
TAM-ChIP Index 3	NNNNNNNN**ACG**
TAM-ChIP Index 4	NNNNNNNN**GCT**
TAM-ChIP Index 5	NNNNNNNN**TGA**
TAM-ChIP Index 6	NNNNNNNN**CGT**
TAM-ChIP Index 7	NNNNNNNN**CTA**
TAM-ChIP Index 8	NNNNNNNN**GTA**
TAM-ChIP Index 9	NNNNNNNN**AGT**
TAM-ChIP Index 10	NNNNNNNN**ATC**
TAM-ChIP Index 11	NNNNNNNN**TAG**
TAM-ChIP Index 12	NNNNNNNN**CAG**
TAM-ChIP Index 13	NNNNNNNN**TAC**
TAM-ChIP Index 14	NNNNNNNN**GTC**
TAM-ChIP Index 15	NNNNNNNN**CAC**
TAM-ChIP Index 16	NNNNNNNN**ACA**

- The antibody index (i5) is an 8 bp sequence of which 3 bp are used for analysis (shown in bold in the Table 3 below). Each anti-species TAM-ChIP conjugate has its own i5 index to enable multiplexing of both the anti-rabbit and anti-mouse conjugates within the same sequencing reaction.

Table 3: Antibody i5 Index

	Forward Sequence	Reverse Complement
TAM-ChIP anti-rabbit	GTAAG**GAG**	**CTC**CTTAC
TAM-ChIP anti-mouse	ACTGC**ATA**	**TAT**GCAGT

- We recommend running the sequencing instrument in standalone mode to allow the use of different length sequencing of the i5 and i7 indices (11 bp on i7 and 8 bp on i5).

  Note: BaseSpace® does not allow for different length sequencing so you will need to use 8 bp and 8 bp in this configuration.

Guidelines For Creating a Sample Sheet Template

We suggest using the Illumina Experiment Manager (IEM) to create a sample sheet template (.csv file). IEM can be downloaded as a separate software program from Illumina at: https://support.illumina.com/sequencing/sequencing_software/experiment_manager/downloads.html.

1. In IEM, select "Create Sample Sheet".
2. Select the appropriate instrument.
3. Select the application shown in Table 1 for your instrument (*e.g*. FASTQ only).
4. Fill in the information in the *Sample Sheet Wizard - Workflow Parameters*.
   a. Reagent Kit Barcode is a required field. If you don't know your reagent kit, type any text in order to create the template, this will also serve as the file name.
   b. TAM-ChIP uses dual index sequencing and requires a dual index library prep kit. Select the Nextera DNA selection to create your sample sheet template since this is a dual index library kit.
   c. Select Nextera Index Kit, 24 Indexes if creating a sample sheet template with the Nextera DNA kit. Additional samples can be added later in the Excel template.
   d. Select 1 (Single) Index Reads
   e. Fill out the rest of the experimental details including Read Type and Cycles Read based on your experimental design.
   f. If using the MID de-duplication tool (which is strongly encouraged), uncheck the box for "Adapter Trimming" as the MID de-duping tool will trim the adapter sequences as part of its script.
   g. Select Next.
5. In *Sample Sheet Wizard - Sample Selection*, on the right hand pane, select "Add Blank Row". To create a template you can add 2 rows here and then modify the .csv file later to customize for all the samples. Click the Maximize box in the upper right corner to view a full screen shot of the sample sheet layout.
   a. Fill in a unique name for the Sample ID in each row.
   b. Select unique options from the drop-down menu for Index 1 (i7) and Index 2 (i5) for each row. The actual selections do not matter since this is a template. We will overwrite the data with the custom TAM-ChIP i7 and i5 information in the Excel file.
   c. If the information is entered correctly, the bottom left corner will show Sample Sheet Status: Valid.
   d. Select Finish to complete the Sample Sheet Template and create a .csv file. You can then view and modify the template in Excel.
6. Open the Sample Sheet Template in Excel. Do not modify any of the [Header], [Reads], or [Settings] fields. Only modify the content of the [Data] portion. If other fields need to be adjusted, create a new template using steps 1-5 above.
   a. Overwrite the Excel values for "I7_Index_ID" with the correct TAM-ChIP Index identification. Overwrite the i7 Index sequence using the 11 bp sequence provided in Table 2: TAM-ChIP i7 Index.
   b. Overwrite the Excel values for "I5_Index_ID" with a unique name (*e.g* Rabbit 1, Rabbit2). Overwrite the i5 index sequence with the correct 8 bp antibody conjugate information provided in Table 3: Antibody i5 Index. Use the appropriate i5 Index orientation (forward or reverse complement sequence) based on the recommendations for your instrument as shown in Table 1.
   c. Copy and paste additional rows to your Sample Sheet Template as needed for your sample run. Overwrite data to make sure each row contains unique values.
7. Once the edits to the template are completed, save as a .csv file to your designated folder location in your run directory.

Guidelines to Prepare FASTQ File for MID De-Duplication:

1. Prepare SampleSheet.csv as described above and run sequencing.

   Note: Remember to remove adapter sequences from sample sheet to avoid trimming by bcl2fastq.

2. This protocol is based on using Illumina **bcl2fastq Conversion Software v2.20**. If using another version of the software, please contact Active Motif's technical support team at tech_service@activemotif.com to obtain the necessary changes to the protocol.

   Run bcl2fastq v2.20 to generate de-multiplexed FASTQ files for each sample.

```
bcl2fastq \
        --use-bases-mask Y*,I3n*,I3n* \
        --minimum-trimmed-read-length 0 \
        --barcode-mismatches 0 \
        --mask-short-adapter-reads 0 \
        --no-lanes-splitting \
        -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
```

3. Remove or rename SampleSheet.csv and run bcl2fastq v2.20 again to generate three non-demultiplexed FASTQ files, making sure to change base mask and paths to output directories as shown below.

```
bcl2fastq \
        --use-bases-mask Y*,Y*,Y* \
        --minimum-trimmed-read-length 0 \
        --barcode-mismatches 0 \
        --mask-short-adapter-reads 0 \
        --no-lanes-splitting \
        -R $nextseq_run_dir \
        -o $nextseq_run_dir/FASTQ_concat \
        --interop-dir $nextseq_run_dir/InterOp_concat \
        --reports-dir $nextseq_run_dir/Reports \
        --stats-dir $nextseq_run_dir/Reports/html_concat
```

4. Reformat the three FASTQ files for the non-de-multiplexed data into a single FASTQ file. After Step 3 above, there should be one FASTQ file containing the read sequence data, and two additional FASTQ files containing the i7 and i5 index sequence data, respectively. The first four lines of these three FASTQ files will look similar to the example below:

```
@NS500375:446:HG3Y3BGX2:1:11101:8018:1054 1:N:0:0
GATTATAGGTATCCNCCATCAAGCCCGGCTAATTTTAGTATTTTTGTAGAGATGGGGTNTCNCTGTGTTGGCC
+
AAAAAEEEEE6EEE#EEEEEEEEEEEE/E/EE/EEEEEEAE<EEEEEEEAEEEEAEEE#EA#EEEEEEAEEEE
```

```
@NS500375:446:HG3Y3BGX2:1:11101:8018:1054 2:N:0:0
GATACTAGTAC
+
AAAAAEEEEE6

@NS500375:446:HG3Y3BGX2:1:11101:8018:1054 3:N:0:0
TATGCAGT
+
AAAAAEEE
```

Using simple programs (such as a perl script), extract the sequence of each i7 and i5 index and append these to the end of the first line of the read sequence FASTQ header to create a new output FASTQ file that will look similar to the example below:

```
@NS500375:446:HG3Y3BGX2:1:11101:8018:1054 1:N:0:GATACTAGTAC+TATGCAGT
GATTATAGGTATCCNCCATCAAGCCCGGCTAATTTTAGTATTTTTGTAGAGATGGGGTNTCNCTGTGTTGGCC
+
AAAAAEEEEE6EEE#EEEEEEEEEEEE/E/EE/EEEEEEAE<EEEEEEEAEEEEAEEE#EA#EEEEEEAEEEE
```

5. Reformat the FASTQ files for each sample to incorporate the MID DNA sequence into the FASTQ header.

After Step 4 above, there should be one FASTQ file for each sample, plus one FASTQ for the entire non-de-multiplexed run. The first four lines of the de-multiplexed FASTQ sample file (created in Step 2) will look similar to the example below:

```
@NS500375:446:HG3Y3BGX2:1:11101:8018:1054  1:N:0:GAT+TAT
GATTATAGGTATCCNCCATCAAGCCCGGCTAATTTTAGTATTTTTGTAGAGATGGGGTNTCNCTGTGTTGGCC
+
AAAAAEEEEE6EEE#EEEEEEEEEEEE/E/EE/EEEEEEAE<EEEEEEEAEEEEAEEE#EA#EEEEEEAEEEE
```

The first four lines of the corresponding data from the non-demultiplexed FASTQ sample file (created in Step 4) will look similar to the example below:

```
@NS500375:446:HG3Y3BGX2:1:11101:8018:1054  1:N:0:GATACTAGTAC+TATGCAGT
GATTATAGGTATCCNCCATCAAGCCCGGCTAATTTTAGTATTTTTGTAGAGATGGGGTNTCNCTGTGTTGGCC
+
AAAAAEEEEE6EEE#EEEEEEEEEEEE/E/EE/EEEEEEAE<EEEEEEEAEEEEAEEE#EA#EEEEEEAEEEE
```

Note the difference in the end of the header line, which has truncated the i7 and i5 adapter sequences in the de-multiplexed version compared to the non-de-multiplexed version.

Using a simple program (such as seqtk and perl scripts), extract each de-multiplexed sample from the non-de-multiplexed data (thereby restoring the original non-truncated header line to the de-multiplexed data).

Then modify each header to add the full i7 index sequence to the end of the SEQID with an underscore separator. The final result will look like the example below:

```
@NS500375:446:HG3Y3BGX2:1:11101:8018:1054_GATACTAGTAC  1:N:0:GATACTAGTAC+TATGCAGT
GATTATAGGTATCCNCCATCAAGCCCGGCTAATTTTAGTATTTTTGTAGAGATGGGGTNTCNCTGTGTTGGCC
+
AAAAAEEEEE6EEE#EEEEEEEEEEEE/E/EE/EEEEEEAE<EEEEEEEAEEEEAEEE#EA#EEEEEEAEEEE
```

This is necessary to preserve the molecular ID sequence after the BAM mapping in the BAM QNAME field.

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

```
trim_galore \
--adapter $AdapterSeq \
--path_to_cutadapt $cutadapt_path/cutadapt \
INPUT.fastq
```

7. Map FASTQ data to reference genome to create BAM using tool of choice (*e.g*. BWA).
8. Sort BAM using tool of choice (*e.g*. SAMtools).
9. 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 Perl script.