Transcriptome-wide m^{6}A detection with DART-Seq

Kate D. Meyer (kate.meyer@duke.edu)
Duke University School of Medicine https://orcid.org/0000-0001-7197-4054

Method Article

Keywords: RNA, epitranscriptome, RNA methylation, m6A, methyladenosine, DART-Seq

Posted Date: September 23rd, 2019

DOI: https://doi.org/10.21203/rs.2.12189/v1

Abstract

m^{6}A is the most abundant internal mRNA modification and plays diverse roles in gene expression regulation. Much of our current knowledge about m^{6}A has been driven by recent advances in the ability to detect this mark transcriptome-wide. Antibody-based approaches have been the method of choice for global m^{6}A mapping studies. These methods rely on m^{6}A antibodies to immunoprecipitate methylated RNAs, followed by next-generation sequencing to identify m^{6}A-containing transcripts[1,2]. While these methods enabled the first identification of m^{6}A sites transcriptome-wide and have dramatically improved our ability to study m^{6}A, they suffer from several limitations. These include requirements for high amounts of input RNA, costly and time-consuming library preparation, high variability across studies, and m^{6}A antibody cross-reactivity with other modifications. Here, we describe DART-Seq (deamination adjacent to RNA modification targets), an antibody-free method for global m^{6}A detection. In DART-Seq, the C to U deaminating enzyme, APOBEC1, is fused to the m^{6}A-binding YTH domain. This fusion protein is then introduced to cellular RNA either through overexpression in cells or with *in vitro* assays, and subsequent deamination of m^{6}A-adjacent cytidines is then detected by RNA sequencing to identify m^{6}A sites. DART-Seq can successfully map m^{6}A sites throughout the transcriptome using as little as 10 nanograms of total cellular RNA, and it is compatible with any standard RNA-seq library preparation method.

Introduction

General notes:

The DART-Seq approach assumes that APOBEC1-YTH and APOBEC1-YTH^{mut} fusion proteins have been cloned into an expression vector appropriate for the application. We have tested mammalian expression under the CMV promoter, but others will likely work as well.

Prior to starting, pilot studies are recommended to ensure that expression of fusion proteins are not toxic to cells and that adequate expression is achieved.

Practice good RNA handling technique during library preparation.

For standard RNA-seq, we use 1 microgram of total cellular RNA for library preparation with the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs). We have also prepared libraries from as little as 10 nanograms of total RNA with good results. As with any RNA-seq library preparation, lower amounts of starting RNA may require more sensitive library preparation methods. We have used the Single Cell/Low Input RNA Library Prep Kit (New England Biolabs) for our low-input libraries with good results. Other library preparation methods are also compatible with DART-Seq and can be tested by the user.

Equipment

Procedure

A. Transfect and grow cells

The protocol described here has been used for adherent HEK293T cells. Adjustments may be needed for other cell types.

1. Plate 10-cm plates (or other appropriate size) of HEK293T cells by diluting approximately 1:8 and plating.

2. Approximately 12-24 hours later, transfect cells with APOBEC1-YTH fusion proteins using desired method (we have had good success with Fugene HD).

3. Grow cells for 24h at 37°C (5% CO2) to ~70-80% confluency. For HEK293T cells we use Dulbecco's Modified Eagle's Medium (DMEM; Corning) with 10% FBS and 1% Penicillin-Streptomycin (100X; Thermo Fisher). We observe good C to U editing with no cell death after 24h, although shorter/longer time points may be desired and can be tested by the user.

Note: We typically have at least 2-3 biological replicates each for cells expressing: APOBEC1-YTH, APOBEC1-YTH^{mut}, and APOBEC1 alone.

B. Harvest RNA

1. Total RNA is harvested using the RNeasy Plus Mini kit and using either on-column DNase treatment or post-purification DNase treatment (RNase-free DNaseI, New England Biolabs). Alternatively, RNA can be isolated with Trizol followed by digestion with RNase-free DNase.

2. Quantify RNA and assess RNA integrity by running on a gel or with a Bioanalyzer.

C. Next-generation sequencing

A variety of options are available for sequencing library preparation. We use the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs) for 1ug total RNA as starting material or the Single Cell/Low Input RNA Library Prep Kit (New England Biolabs) for samples using < 100ng total RNA as starting material.

Sequencing can be carried out using paired-end or single-read sequencing. We typically use SR 50bp sequencing on an Illumina HiSeq4000 or similar machine. We achieve 30-50 million reads per sample

using this platform.

D. Identify C to U editing events and m^{6}A sites

1. Demultiplex sequencing reads and remove adapter sequences. These steps will depend on the sequencing library prep method and barcode method used. For adapter removal, we use FLEXBAR[3].

2. Collapse exact duplicates using the fastq2collapse command in the CLIP Tool Kit (CTK) suite[4]. Several processing steps utilize various CTK commands; an excellent web resource for these tools can be found here: https://zhanglab.c2b2.columbia.edu/index.php/CTK_Documentation

3. Align reads to the genome. We use Novoalign with the options below or BWA according to the current recommendations of the CTK developers (see the link above).

novoalign -t 85 -d file.nix -f file.fastq -l 16 -s 1 -o SAM -r None > file.sam

4. Remove duplicate reads and prepare file for parsing:

samtools sort file.sam -O BAM -o file.sorted.bam

novosort --markduplicates --keeptags file.sorted.bam -i -o file.uniq.bam

samtools fillmd file.novo.uniq.bam file.fa | gzip -c > file.sorted.md.sam.gz

BAM files can be merged using samtools merge or loaded individually into the genome browser of choice (we use IGV). This enables viewing of mutations at individual sites throughout the transcriptome.

5. Parse SAM file using the CTK program:

parseAlignment.pl -v --map-qual 1 --min-len 18 --mutation-file file.mutation.txt file.sorted.md.sam.gz - | gzip -c > file.tag.uniq.bed.gz

6. Collapse PCR duplicates with CTK:

tag2collapse.pl -v -big -weight --weight-in-name --keep-max-score --keep-tag-name file.tag.bed file.tag.uniq.bed

7. Get mutations in unique tags using CTK:

joinWrapper.py file.mutation.txt file.tag.uniq.bed 4 4 N file.tag.uniq.mutation.txt

8. Merge replicates:

cat file1.tag.uniq.bed file2.tag.uniq.bed file3.tag.uniq.bed > file.tag.uniq.bed

cat file1.tag.uniq.mutation.txt file2.tag.uniq.mutation.txt file3.tag.uniq.mutation.txt > file.tag.uniq.mutation.txt

9. Find C to T transitions

To isolate C to T transitions, we use a script from the miCLIP method[5], which also identifies C to T mutations adjacent to m^{6}A sites:

awk '{if($6=="+" && $8=="C" && $9==">" && $10=="T" || $6=="-" && $8=="G" && $9==">" && $10=="A") {print $0}}' file.tag.uniq.mutation.txt | cut -f 1-6 > file.tag.uniq.C2T.bed

10. Run CIMS with CTK:

CIMS.pl -big -v -n 5 -p -c cache_dir file.tag.uniq.bed file.tag.uniq.C2T.bed file.tag.uniq.C2T.CIMS.txt

11. Identify CIMS with desired FDR:

awk "{if(\$9<=1) {print \$1\"\t\"\$2\"\t\"\$3\"\t$1_\"\$4\"_\"\$9\"\t\"\$5\"\t\"\$6}}" file.tag.uniq.C2T.CIMS.txt | sort -k 9,9n -k 8,8nr -k 7,7n | cut -f 1-6 > file.tag.uniq.C2T.CIMS.p1.bed

12. Filter C to T sites:

This step can be varied to achieve the desired stringency of site detection. Our standard protocol is to take C to T sites identified with the p<1 threshold (Step 11) and keep only those with a minimum of 2 mutations, at least 10 reads per replicate, and a mutation/read (m/k) threshold of 10-60%. We find that adjusting the number of mutations, reads per replicate, and m/k threshold is a good way to increase/decrease stringency of m^{6}A site calls to a desired level. In addition to these filtering steps, known mutations in the genome from the dbSNP database (https://www.ncbi.nlm.nih.gov/snp/), as well as endogenous C to U editing sites identified by sequencing of wild type cells, are also removed. Lastly,

we recommend removing sites identified from cells expressing APOBEC1 alone. These filtering steps can all be completed using bedtools intersect from the bedtools suite (https://bedtools.readthedocs.io/en/latest/).

13. Calculate enrichment over APOBEC1-YTH^{mut}-expressing cells.

To further ensure high-confidence identification of m^{6}A sites, we filter C to U editing sites to include only those with a minimum threshold of mutations per read compared to sites obtained by expression of APOBEC1-YTH^{mut} :

A. First, use bedtools intersect to find C to T transitions that are present only in APOBEC1-YTH samples. Ensure that the YTH and mut .bed files you use have mutations/read (m/k) in column 5. These will be merged back in later.

```
bedtools intersect -s -v -a yth.filt.bed -b mut.bed > ythnotmut.filt.bed
```

B. Next, merge the YTH and mut files for common sites.

```
bedtools intersect -s -wa -wb -a yth.filt.bed -b mut.bed > yth.filt.mergemut.bed
```

C. Then use awk to identify sites that have a m/k ratio that is 1.5-fold greater than in APOBEC1-YTH^{mut} samples:

```
awk '{a=$5/$11;print $0,a;}' yth.filt.mergemut.bed | awk '$13 >= 1.5 {print $1"\t"$2"\t"$3"\t"$13"\t"$5"\t"$6}' > yth.filt.mk1.5overmut.bed
```

The 1.5-fold cutoff can be adjusted by the user to be more/less stringent as desired.

D. Now, add back the sites in YTH that were not in mut:

```
cat yth.filt.mk1.5overmut.bed ythnotmut.filt.bed > yth.mk1.5overmut.final.bed
```

We find that the above filters are a good starting point for identifying m^6A sites with DART-Seq. Further filtering steps can also be implemented (e.g., limiting to sites immediately following an A; limiting to sites in a DRACH motif, excluding sites lost in methyltransferase-depleted cells, etc.).

Troubleshooting

Time Taken

Anticipated Results

References

1. Dominissini, D. *et al.* Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**, 201-6 (2012).

2. Meyer, K.D. *et al.* Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell* **149**, 1635-46 (2012).

3. Dodt, M., Roehr, J.T., Ahmed, R. & Dieterich, C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology (Basel)* **1**, 895-905 (2012).

4. Shah, A., Qian, Y., Weyn-Vanhentenryck, S.M. & Zhang, C. CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics* **33**, 566-567 (2017).

5. Linder, B. *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat Methods* (2015).

Acknowledgements