Published in final edited form as:

Methods. 2017 August 15; 126: 86–94. doi:10.1016/j.ymeth.2017.06.003.

Genome-wide profiling of the 3' ends of polyadenylated RNAs

Piero Sanfilippo^{1,2}, Pedro Miura³, and Eric C. Lai^{1,2,4}

¹Department of Developmental Biology, Sloan-Kettering Institute, New York, New York 10065, USA

²Louis V. Gerstner, Jr. Graduate School of Biomedical Sciences, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA

³Department of Biology, University of Nevada, Reno, Reno, NV 89557, USA

Abstract

Alternative polyadenylation (APA) diversifies the 3' termini of a majority of mRNAs in most eukaryotes, and is consequently inferred to have substantial consequences for the utilization of post-transcriptional regulatory mechanisms. Since conventional RNA-sequencing methods do not accurately define mRNA termini, a number of protocols have been developed that permit sequencing of the 3' ends of polyadenylated transcripts (3'-seq). We present here our experimental protocol to generate 3'-seq libraries using a dT-priming approach, including extensive details on considerations that will enable successful library cloning. We pair this with a set of computational tools that allow the user to process the raw sequence data into a filtered set of clusters that represent high-confidence functional polyadenylation sites. The data are single-nucleotide resolution and quantitative, and can be used for downstream analyses of APA.

Keywords

3' UTR; 3' end; APA; Polyadenylation; Genome-wide profiling

1. Introduction

The final step in the maturation of an mRNA is the recognition of a polyadenylation signal (PAS) at its 3' end leading to cleavage and polyadenylation of the nascent transcript. Although a few genes were known from decades ago to exhibit alternative definition of their 3' termini [1, 2], a process referred to as alternative cleavage and polyadenylation (APA), genome-wide studies eventually revealed this process to be the rule and not the exception. Indeed, a majority of genes in diverse eukaryotic organisms probed to date appear to undergo APA [3–8]. While some of the first examples of APA occur within internal gene regions and affect coding potential [1, 2], most APA sites occur within 3' UTRs [9].

⁴Author for correspondence: laie@mskcc.org, tel: 212-639-5578.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Since 3' UTRs act as hubs of post-transcriptional regulation, APA has substantial implications for determining alternative usage of diverse regulatory regimes. 3' UTRs can affect transcript stability, localization and/or translation, regulatory events that are often mediated via binding of trans-acting regulators such as RNA binding proteins (RBPs) [10] and microRNAs (miRNAs) [11]. Furthermore, the accumulation of alternative 3' UTR isoforms is highly regulated, with isoforms differentially accumulating depending on tissue, developmental and disease states [7, 8, 12–14]. Beyond correlative genome-wide studies, the impact of APA on individual genes can be tangible and substantial. For example, the expression of shorter 3' UTRs on some oncogenes may be associated with transforming properties [15], expression of the long 3' UTR of *α-synuclein* has been shown to be linked to the accumulation and translation of *α-synuclein* transcripts in Parkinson's disease [16], while the expression of the long 3' UTR isoform of *BDNF* leads to transport and translation of *BDNF* transcript in dendrites [17].

Despite great interest in this topic, the underlying mechanisms that lead to the expression of different 3' UTR isoforms still remain to be clarified [9, 18]. Moreover, from the phenotypic viewpoint, much remains to be understood as to the extent that switching of 3' UTR isoforms affects gene regulation. For example, in at least some settings, it was proposed that global 3' UTR isoform modulation has subtle effects on protein outputs [19, 20]. Therefore, there are clearly great needs for ongoing investigations of APA mechanism and biology. Both of these efforts will often need to utilize strategies to profile 3' UTR isoforms in order to examine the regulation, perturbation, and function of 3' UTR isoforms.

Early genome-wide attempts to understand the diversity of transcripts generated by APA involved analyses of cDNA clones [21, 22]. These studies were followed by attempts to infer 3' UTR isoforms expression first by microarrays and later by RNA sequencing. However, these techniques can only infer the genomic location of the cleavage events by either looking at already known events in the case of microarrays [23] or by only recognizing large changes in 3' UTR length by leveraging distinctive changepoints in RNA-seq coverage [12, 24]. Because of these limitations several groups have developed specialized protocol to specifically sequence just the 3' ends of mRNAs [5, 6, 25–27]; see [28] for review of these and other strategies. In the course of our efforts to annotate 3' UTR isoforms in *Drosophila* (in preparation), we also developed methods to sequence and analyze 3' termini of polyadenylated transcripts. We present a detailed experimental protocol and provide bioinformatic tools that permit quantitative, single- nucleotide resolution measurements of sites of cleavage and polyadenylation, allowing APA to be assessed at the transcriptome wide level in a rapid and cost-effective manner.

2. Description of the method

2.1. Overview

3'-seq reports on alternative polyadenylated RNA isoforms by sequencing the junction of the end of the transcript (3' UTR in the case of mRNA) and the polyA tail. This allows for the genome-wide quantification of RNA isoforms that differ at the 3' end (Figure 1). The method can report on any polyadenylated RNAs, including coding and non-coding species. Briefly, the protocol starts by synthesizing cDNA from fragmented total RNA using an RT

primer with a biotin at the 5' end, a part of an Illumina adapter, and oligo-dT with a terminal anchor at the 3' end. The oligo-dT sequence recognizes the polyA tail and the anchor at the end of the primer ensures that the oligonucleotide binds at the junction of the polyA tail with the terminus of the transcript. cDNA is converted to dsDNA and bound on magnetic beads. This step allows for direction specific ligation of the Universal Illumina adapter and ease of washing between different steps. The library is PCR amplified and size selected to enrich for reads that contain the junction between the polyA tail and the end of the transcript. Sequencing of these reads followed by mapping to a reference genome enables determination of 3' ends at single nucleotide resolution, and can be used for differential expression analysis.

2.2. Detailed protocol

2.2.1. Total RNA isolation—This protocol provides a quantitative, genome-wide readout of the 3' ends of all polyadenylated coding and non-coding transcripts. 3'-seq can be performed on any tissue or cell sample, and requires at least 500 ng of total RNA.

- 1 Isolate tissue or cell lines on ice. *It is important to perform this step quickly, keeping isolated material on ice to avoid RNA degradation.*
- 2 Prepare total RNA using Trizol® Reagent (Ambion) according to manufacturer's instructions. *Be sure to completely homogenize samples in Trizol®.*
- 3 DNase treat total RNA to remove DNA contamination according to standard procedures.
- 4 Resuspend total RNA to a concentration of at least 100 ng/μl in nuclease-free water. Store RNA samples at –80 °C.

2.2.2. Total RNA QC

- 5 Determine the quality of prepared total RNA on Bioanalyzer 2100 using the Agilent RNA 6000 Kit. A typical profile for *Drosophila* high quality total RNA is shown, where peaks indicate 18S and 28S rRNA (Figure 2).

2.2.3. Total RNA fragmentation—Total RNA is fragmented using divalent cations under high temperature to obtain small fragments that allow us to pick up the junction between the 3' end of the transcript and the polyA tail upon sequencing. Steps 2.2.3. to 2.2.7. can be carried out either in 1.5 mL tubes when making up to 16 libraries in parallel or in PCR tubes if processing more than 16 samples (see supplies and equipment for details – sample number limitations are due to the size of the magnet used).

- 6 Prepare the fragmentation reaction with 500–2000ng of total RNA and 4 μL 5X FS buffer (SSIII kit, Invitrogen) diluted to a final volume of 10 μL with water.
- 7 Fragment total RNA by placing the reaction on a thermocycler for 10 min at 94 °C.
- 8 Decrease the temperature of the fragmented total RNA to 55 °C to prepare the RNA for 1st strand cDNA synthesis

Note: The fragmentation time used above has been optimized for Drosophila total RNA. The fragmentation time was chosen by performing a time course of the same reaction with species specific RNA. Here the optimal time that leads to small fragments of 150–200 nucleotides in length prior to complete RNA fragmentation was chosen (Figure 3). Fragmentation might need to be optimized for different RNA preparations.

2.2.4. 1st strand synthesis—In this step cDNA of the 3' end of transcripts is synthesized. The anchor on the oligo- dT RT primer as well as the relatively high temperature of 55 °C ensures proper annealing onto the junction between the end of the 3' UTR and the polyA tail. The RT primer also includes a portion of one of the adapters required for Illumina sequencing to which one of the PCR amplification primers in the final step of the protocol will anneal.

- 9 Prepare 1st strand synthesis reaction mix by adding 1 mM dNTPs, 0.8 µM RT primer, 20 mM DTT, 20U of RNaseOUT (Invitrogen) and 200U of SuperScript III (Invitrogen).
- 10 Equilibrate 1st strand synthesis reaction mix to 55 °C on thermocycler.
- 11 Add mix to the fragmented total RNA and mix by pipetting up and down at least 4 times.
- 12 Incubate the reaction for 1 h at 55 °C.
- 13 Inactivate the reaction for 15 min at 70 °C.

Possible stop point. Store at –20 °C.

2.2.5. cDNA cleanup—The cDNA is cleaned up using Ampure XP beads (AGENCOURT BECKMAN) to remove smallest fragments as well as enzymes and buffers.

- 14 Add 1.5 volumes (45 µL) of Ampure XP beads equilibrated at room temperature.
- 15 Allow binding of nucleic acids to beads for 5 min.
- 16 Place on magnetic stand for 5 min or until solution appears clear and remove the supernatant.
- 17 Wash the beads two times with 70% ethanol according to manufacturer's instructions. *Make fresh 70% ethanol on the same day.*
- 18 Air dry the beads for 1 min.
- 19 Elute biotinylated cDNA in 40 µL of 10 mM Tris-HCl pH 8.0 by re-suspending the beads in the elution buffer and place on magnetic stand.
- 20 Transfer eluted cDNA in a new tube.

Possible stop point. Store at –20 °C.

Note: It is important to use 1.5 volumes of beads to cDNA volume to select the right fragment size range. To ensure this, check that the volume has not changed significantly do to evaporation during the cDNA synthesis step.

2.2.6. Second strand synthesis—In this step, double stranded biotinylated cDNA is generated. RNase H is used to nick the cDNA/RNA duplex and *E. coli* DNA polymerase I is used to synthesize the second strand of DNA by nick translation.

- 21 Prepare 2nd strand synthesis reaction mix by adding 0.5 mM dNTPs, 1X NEB2 buffer, 2.5U RNase H (Thermo Scientific), and 20U of *E. coli* DNA polymerase I (NEB) to the above generated cDNA and diluting if necessary to 50 µL with water.
- 22 Incubate the reaction for 2.5 h at 16 °C.

Possible stop point. Store at –20 °C.

2.2.7. ds- cDNA cleanup—ds-cDNA is cleaned up using Ampure XP beads as in step 2.2.5. using 1.5 volumes of beads (75 µL) and eluting ds-cDNA in 50 µL of 10 mM Tris-HCl pH 8.0.

2.2.8. Bind ds-cDNA to magnetic streptavidin beads—Biotinylated ds-cDNA is bound to magnetic streptavidin beads (M-280). This protects one of the ends leading to end specific ligation of the Illumina Universal adapter in the next steps.

- 23 Wash 50 µL M-280 beads (Invitrogen) two times in 2X B&W buffer.
- 24 Re-suspend beads in 50 µL 2X B&W buffer.
- 25 Add the bead mixture to 50µL of the ds-cDNA solution and re-suspend to generate a homogeneous solution.
- 26 Incubate the solution at room temperature for 30 min with rotation to allow binding.
- 27 Wash two times with 1X B&W buffer.
- 28 Wash two times with 1X NEB2 buffer.
- 29 Transfer the solution to a new tube.

2.2.9. ds-cDNA end blunting—In this step Klenow fragment is used to generate blunt ends of the ds-cDNA to ensure blunt-end specific ligation of the ds-DNA adapter in step 2.2.11.

- 30 Prepare reaction mix by mixing 1 mM dNTPs, 1X NEB2 buffer and 5U of Klenow fragment (NEB) and dilute the reaction mix to a volume of 100 µL with water.
- 31 Remove the buffer from the ds-cDNA bound beads from 2.2.8..
- 32 Add the blunting reaction mix to the bead bound ds-cDNA.

- 33 Incubate the reaction for 15 min at 25 °C on a thermomixer using interval mixing of 1400 rpm for 15 s followed by 2 min pause.

2.2.10. Beads clean up—Once the library is bound to the M-280 beads buffers and enzymes are changed using a series of washes that also include a mixture of proteases to ensure enzyme inactivation.

- 34 Wash the beads one time with 1X buffer C.
- 35 Remove buffer C and add 100 µL of cleaning solution (refer to appendix for composition).
- 36 Incubate the cleaning reaction for 15 min at 37 °C to inactivate enzymes on a thermomixer using interval mixing of 15 s 1400 rpm followed by 2 min pause.
- 37 Wash three times with 1X buffer D.
- 38 Wash two times with 1X T4 ligase buffer.
- 39 Transfer beads to a new tube.

2.2.11. dsDNA adapter ligation—In this step the Illumina TruSeq Universal adapter is ligated to the ds-cDNA. The dsDNA adapter fragment has a 5' overhang to ensure direction dependent ligation to the bead bound ds-cDNA.

- 40 Prepare ligation reaction mix by adding 1X T4 DNA ligase buffer, 0.4 µM dsDNA Universal adapter (previously prepared, see Appendix B and C), 2000U of T4 DNA ligase (NEB) and diluting to a volume of 100µL with water.
- 41 Remove buffer from ds-cDNA bound beads and add the ligation reaction mix.
- 42 Incubate the ligation reaction overnight at 16 °C on a thermomixer using a shaking cycle of 1400 rpm for 15 s followed by 2 min pause.

2.2.12. Final beads clean up

- 43 Wash the beads one time with 1X buffer C.
- 44 Remove buffer C and add 100 µL of cleaning solution (make fresh solution).
- 45 Incubate the cleaning reaction for 15 min at 37 °C to inactivate enzymes on a thermomixer using interval mixing of 1400 rpm for 15 s followed by 2 min pause.
- 46 Wash three times with 1X buffer D.
- 47 Wash two times with 1X Phusion HF buffer (Thermo Scientific).
- 48 Transfer beads to a new tube.

2.2.13. Enrichment PCR—A PCR amplification step is carried out to get the library off the beads and add the remainder of the adapter sequences to the library. Each library should be amplified using a different barcoded forward primer (sequences provided in Appendix C), choosing barcode combinations according to Illumina guidelines.

- 49 Prepare PCR reaction mix by adding 1X HF PCR buffer, 6.25 μ M each of universal primer and barcoded primer, 1 mM dNTPs and 1U Phusion High-Fidelity DNA polymerase (Thermo Scientific) and diluting to a volume of 50 μ L with water.
- 50 Remove buffer from the beads and add the PCR reaction mix.
- 51 Perform PCR amplification using this program:
- STEP 1: 98 °C 30 s
 - STEP 2: 98 °C 10 s
 - STEP 3: 63 °C 30 s
 - STEP 4: 72 °C 15 s
 - Cycle to STEP 2 15 times
 - STEP 5: 72 °C 10 min
 - STEP 6: 4 °C HOLD

- 52 Place amplified library on a magnet and transfer supernatant to a new tube.

Possible stop point. Store at -20°C .

2.2.14. Library size selection—In this final step, the library is size selected on a non-denaturing polyacrylamide gel such that a significant portion of the library contains the junction between the polyA tail and the 3' end of the transcript upon sequencing the library in 1X 50bp sequencing mode. If using longer reads, the library should be sized accordingly to the increased read length.

- 53 Precipitate the amplified library with 3X volume of 100% ethanol and 10 μ g of glycogen.
- 54 Incubate for a minimum of 2 h at -20°C .
- 55 Pellet the library by spinning in a bench top centrifuge at 20,800 rcf for 20 min at 4°C .
- 56 Remove supernatant and dry pellet for 2–5 min to remove traces of ethanol avoiding excessive drying of the pellet.
- 57 Re-suspend pellet in 10 μ L of 1X Ficoll loading buffer.
- 58 Load sample flanked by 1 μ g GeneRuler Low Range DNA ladder (Thermo Scientific), 1X Ficoll loading buffer in an 8% TBE polyacrylamide minigel (Invitrogen) and run gel at 200V for 35 min.
- 59 Stain gel for 10 min in 1X SYBR Gold (Invitrogen) in TBE buffer.
- 60 Wash two times in 1X TBE buffer for 10 min.

- 61 Image gel prior to excision of the library and cut the library using a razor blade in the size range between 175bp and 250bp using flanking ladders as size standards (Figure 4).
- 62 Place the gel piece in 400µL Lonza elution buffer and rotate overnight at 4 °C.
- 63 Transfer the eluted library in a new tube.
- 64 Add 2X volume of 100% ethanol, 10 µg glycogen and precipitate at –20 °C for at least 2 h.
- 65 Pellet the library by spinning in a bench top centrifuge at 20,800 rcf for 20 min at 4 °C.
- 66 Wash pellet with 70% ethanol and pellet at 20,800 rcf for 5 min.
- 67 Remove the ethanol solution and let the pellet dry for ~5 min.
- 68 Re-suspend pellet in 6 µL nuclease-free water and store at –20 °C.

Note: Avoid excision below 150 bp, as adapter-adapter products are present at around 120–130 bp and should not contaminate the final library preparation.

2.2.15. Library QC—Validate the library size distribution and quantify library amount.

- 69 Check the quality of 3'-seq library preparation by running 1 µL of sample on a Bioanalyzer 2100 with an Agilent High Sensitivity DNA Kit. A size distribution that reflects the size range extracted from the gel should be obtained (Figure 5). Quantify the library either by using the Bioanalyzer estimate or other methods, such as Qubit fluorometric quantitation.

2.3 Deep sequencing

Sequencing of the library can be done on any Illumina platform that supports TruSeq adapters. Libraries were successfully sequenced using Illumina HiSeq1000 and HiSeq2000 for single end 50 bp reads. Longer reads could be used to increase the proportion of the library that is sequenced into the polyA tail, however using 50 bp is sufficient to have 60–70% of trimmed reads spanning the junction between the polyA tail and the 3' end of the transcript map to a reference genome. Higher depth should be favored for discovery of rare events while lower depth and more replicates should be used for differential expression analysis of 3' UTR isoforms among samples. Barcoded adapters allow for libraries to be multiplexed and sequenced in the same lane. The library should be sequenced with the addition of 5% PhiX spike in control to provide increased complexity because of the large presence of polyA segments in the library.

2.4. Bioinformatics Analysis

3'-seq provides single nucleotide resolution of the 3' ends of polyadenylated RNA transcripts. This is achieved by mapping the reads that encompass the 3' end/polyA junction after trimming of the untemplated polyA. The 3' end nucleotide of those reads represents the nucleotide upstream to the cleavage site generated by the cleavage and polyadenylation machinery. The bioinformatics approach reported here provides commands to use available

software as well as custom made java utilities (available on github) to map reads, filter internal priming and cluster adjacent events to call final 3' end positions. The analysis is designed to run in a Unix environment with >8 Gb of available RAM. Our custom made utilities do not require scripting and can be readily used in a Unix environment to perform the analysis with minimal programming skills. Examples of raw mapped reads and identified 3' ends derived using this pipeline are provided in Figure 6.

2.4.1. De-multiplexing and QC—De-multiplexing of reads can be performed with the Illumina CASAVA utility, which splits the fastQ file by barcode. Standard quality assessment on the fastQ files can be done using FastQC software (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). Over representation of polyA stretches that result from sequencing of polyA tail should be expected.

2.4.2. Mapping to the reference genome—The reads generated with this protocol will map to 3' ends of transcripts. Depending on downstream applications, the reads that map just upstream of the cleavage site can also be used to increase library depth and these reads will be labeled *untrimmed*, and will map to the genome without trimming. The remaining reads require trimming of untemplated polyA stretches to map onto the reference genome. The last nucleotide on the 3' end of these reads represents the nucleotide just upstream to the cleavage site of transcripts. These reads will be labeled *trimmed* and provide us with the nucleotide level resolution of 3' end formation.

2.4.2.1. Mapping the untrimmed reads: The mapping to the reference genome is achieved using GSNAP (v. 2013-03-31) [29].

- 1 Build a genomic index file for a specific reference genome. The reference genome sequence can be downloaded either from UCSC genome browser (<http://genome.ucsc.edu/>) or from the NCBI ftp site (<ftp://ftp.ncbi.nih.gov/genomes/>).

Input:

Reference genome sequence in FASTA format – *genome.fasta*

Name of reference genome – *genome_name*

```
$ gmap_build -D /path/to/genome_index_folder \
-d genome_name genome.fasta
```

Output:

/path/to/genome_index_folder/index_files

- 2 Align the untrimmed reads to the genome. After mapping the reads with GSNAP, the script below will generate a BAM file of mapped untrimmed reads using SAMtools [30]. The folder structure used below will be important for the final step.

Input:

De-multiplexed fastq.gz sequenced reads file – *sample.fastq.gz*

Name of the sample for output – *sample*

```
$ mkdir -p mapped/untrimmed/
$ cd mapped/untrimmed
$ gsnap -D /path/to/genome_index_folder -N 0 --gunzip -A
sam \ -d genome_name sample.fastq.gz | samtools view -bS -
| \ samtools sort - sample
```

Output:

sample.bam

3 Index BAM file.

Input:

BAM file generated above – *sample.bam*

```
$ samtools index sample.bam
```

Output:

sample.bam.bai

2.4.2.2. Trimming of unmapped reads: The unmapped reads include the polyA/3' end junction reads. A fraction of these will also contain parts of adapter sequence. A java tool is provided to trim both the polyA tail and the adapter sequences off the 3' end of the unmapped reads.

Input:

BAM file generated above - */path/to/mapped/untrimmed/sample.bam*

```
$ java -Xmx4g -jar TrimUnmapped.jar trim3p -bam \
sample.bam -out /dev/stdout | gzip - > \
sample_trimmed.fastq.gz
```

Output:

sample_trimmed.fastq.gz

2.4.2.3. Mapping of trimmed reads: The untrimmed reads can be mapped as in steps 2 and 3 of 2.4.2.1 using as input the *sample_trimmed.fastq.gz* file. The resulting bam and bam.bai files should be placed in */path/to/mapped/trimmed* folder for downstream analysis.

Output:

/path/to/mapped/trimmed/sample.bam

/path/to/mapped/trimmed/sample.bam.bai

2.4.3 Internal priming mask—An inherent problem with 3' end sequencing approaches is that stretches of genomically encoded adenosine (A) nucleotides can serve as templates for priming first strand synthesis. This is called "internal priming" as 3' end calls are erroneously made upstream of bona fide termini. A filtering strategy is employed to minimize these erroneous annotations. Provided is a tool that uses a sliding window approach to identify genomic stretches with a certain percentage of A nucleotides provided by the user. The window used below is 16 nt wide and requires at least 9 As in the window. These parameters were derived to filter internal priming events occurring during 3'-seq of *Drosophila* total RNA. The width of window and number of As should be optimized for the particular organism being studied. This optimization can be performed using known alternative 3' ends if a confident set already exists.

Input:

Reference genome sequence in FASTA format – *genome.fasta*

Name of the file for output - *primingMask.bed*

Parameters:

-window – window in which to assess A-richness. The window starts one nucleotide off the 3' of the read (since the first nucleotide after the cleavage site is most often an A)

-maxAdenosine – max number of As in the window. Below this number, in this case 9, a read passes the filter. Every position with equal or above maxAdenosine in the provided window will be reported in the mask.

```
$ samtools faidx genome.fasta (creates genome index)
$ java -Xmx2g -jar ThreeSeqPipeline.jar \
IdentifyInternalPriming -fa genome.fasta -maxAdenosine 9 \
-primingMask primingMask.bed -window 16
```

Output:

primingMask.bed

The internal priming mask will be used in the next step to minimize false positive calls due to internal priming.

Note: An alternative method to identify internally primed events is to add 2 or 3 As to the 3' end of the trimmed reads and map these requiring no mismatches to the reference genome. The reads that map to the genome are likely to come from internal priming events, as the As are template in the genome, and can be omitted from the downstream analysis.

2.4.4. Clustering and 3' end calling—To call the predominant cleavage site, events occurring within a short distance of each other can be clustered. In this step the utility will cluster events occurring in a 25 nt window and call as the cluster end the most abundant

position. The utility takes as input any number of libraries and the final called event is the most abundant amongst all libraries. Parameters for clustering window size as well as read number thresholds can be provided by the user and are outlined in detail in Appendix D. Additionally, the events will be filtered using the internal priming mask generated in the previous step. Finally, each called event will be quantified in each sample provided. Additional parameters and details are outlined in Appendix D.

Input:

Reference genome sequence in FASTA format – *genome.fasta*

Priming mask - */path/to/primingMask.bed*

Directory for gtf output – *gtf_files*

Text file with sample names without file extension, one per line -
sample_name.txt

Parameters:

-minDistinctReads – minimum number of unique reads which are needed to call a specific 3' end.

The script should be run in the folder containing the *mapped* folder created above.

```
$ mkdir htsjdk_tmp
$ java -Xmx4g -Djava.io.tmpdir=htsjdk_tmp \
-jar ThreeSeqPipeline.jar DefineClusters \
-minDistinctReads 3 -inDir mapped -trimmed trimmed \
-untrimmed untrimmed -primingMask /path/to/primingMask.bed\
-outDir gtf_files -baseNames sample_name.txt
```

Output:

atlas.gtf - the 3' end positions are reported in this file. A detailed description is provided in appendix D

gtf_files/ - the output in this folder is described in detail in appendix D

Note: Existing annotation of 3' ends, if available, can be used to quantify events by using the mapped reads from step 2.4.2, using the annotated 3' ends as features to count the reads obtained using 3'-seq.

2.4.5. Basic analysis of 3'-seq data—The analysis steps provided in 2.4.1.–2.4.4. allow the user to annotate all predominant 3' ends of polyadenylated transcripts. Further analysis steps depend on the specific user study application. For example, one can compare changes in 3' UTR expression by taking the ratio of the two most dominant isoforms of 3' UTRs with multiple ends. Comparison between sample will indicate if a certain gene expresses longer or shorter 3' UTR isoforms. Sequence upstream and downstream of the derived 3' ends can be analyzed for the presence of enriched PAS sequences that might be specific to processing in certain tissues or organisms. Furthermore, the output of 3'-seq can be used to quantify

gene expression by deriving gene counts and using gene expression analysis software such as DeSeq2 [31].

Troubleshooting

- *Total RNA QC*: If isolating good quality total RNA proves difficult, revisit the extraction method to minimize the time the sample is not in Trizol[®]. Smaller batches of dissected tissue can be placed in Trizol[®] and subsequently pooled to minimize degradation of RNA.
- *1st strand synthesis*: If the resulting library results in excessive amounts of reads mapping to polyA tail, make sure that the reaction was not allowed to cool down below 55 °C as that will increase priming of the RT primer onto just polyA tail. A certain degree of internal priming and only polyA sequencing is to be expected.
- *PCR amplification*: If PCR duplicates appear to be a problem, the number of cycles can be adjusted or a random k-mer sequence could be added to the barcode to control for PCR duplication.
- *3'-seq reads visualization*: The reads can be visualized using a genome browser such as IGV [32] or the UCSC Genome Browser (<http://www.genome.ucsc.edu>). Visualization of the trimmed reads is sufficient to visualize the 3' end events. The atlas.gtf output can be uploaded to visualize the dominant 3' ends obtained in step 2.4.4. Visual analysis of these files as well as the mapped reads can be useful to determine success of the sequencing run, with reads clustering onto ends of annotated transcript models.

3. Concluding remarks

The recent realization that the majority of genes in eukaryotic organisms undergo APA adds an additional layer of complexity to post-transcriptional gene regulation. The complex pattern of 3' UTR isoforms expression appears to be under substantial regulation as seen in both normal and pathological biological settings. This 3'-seq protocol provides an additional tool to probe this diverse aspect of gene expression.

The mechanisms that lead to the steady state accumulation of different 3' UTR isoforms are just emerging. The genome-wide assessment of 3' UTR isoforms expression during normal as well as perturbed conditions will be an important tool to further our understanding of the regulatory mechanisms that control the expression of different 3' UTR isoforms.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful to Sol Shenker for developing the java utilities. We also acknowledge Sonali Majumdar and Jiayu Wen for insightful comments on the manuscript and the NY Genome Center and the MSKCC Genomics core for discussion.

Funding Sources

Work in E.C.L.'s group was supported by the National Institutes of Health (R01-NS083833 and R01-GM083300) and MSK Core Grant P30-CA008748).

Abbreviations

3' UTR	3' untranslated region
3'-seq	3' end sequencing
APA	alternative polyadenylation
PAS	polyadenylation signal
miRNA	microRNA
A	adenosine
nt	nucleotide

Appendix A – Supplies and equipment

Enzymes and chemicals

TRIzol® Reagent (Ambion, cat. No. 15596026)

SuperScript III Reverse Transcriptase (Invitrogen, cat. No. 18080-093) including:

5X First Strand Buffer (FS buffer)

0.1 M DTT

100 mM dNTPs (Invitrogen, dGTP – cat. No. 10218014, dCTP – cat. No. 10217016, dATP – cat. No. 10216018, dTTP – cat. No. 10219012)

RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen, cat. No. 10777019)

100% Ethanol (Decon, cat. No. 2716G)

DNase/RNase free water

Agencourt Ampure XP (BECKMAN COULTER, cat. No. A63880)

1 M UltraPure Tris-HCl pH 8.0 (Invitrogen, cat. No. 15568025)

DNA polymerase I (E. coli) (NEB, cat. No. M0209S) including:

10X NEB2 buffer

RNaseH (Thermo Scientific, cat. No. EN0201)

Dynabeads M-280 Streptavidin (Invitrogen, cat. No. 11205D)

DNA Polymerase I, Large (Klenow) Fragment (NEB, cat. No. M0210S)

T4 DNA Ligase (NED, cat. No. M0202L)

Phusion High-Fidelity DNA Polymerase (Thermo Scientific, cat. No. F530S)

SYBR Gold Nucleic Acid Gel Stain (Invitrogen, cat. No. S11494)

GeneRuler Low Range DNA Ladder (Thermo Scientific, cat. No. SM1191)

Pronase (Roche, cat. No. 10165921001)

Supplies and Equipment

Benchtop Centrifuge

1.5 mL tubes rotor

Thermocycler

Thermomixer C (Eppendorf, cat. No. 2231000269)

Sterile filter tips

Nonstick, RNase-free Microfuge Tubes, 1.5 mL (Applied Biosystems, cat. No. AM12450)

Dynamag-2 Magnet (Thermo Scientific, cat. No. 12321D)

Novex® TBE Gels, 8%, 10 well (Invitrogen, cat. No. EC6215BOX)

For processing more than 16 samples in parallel

Magnetic Stand-96 (Thermo Scientific, cat. No. AM10027)

SmartBlock™ PCR 96 (Eppendorf, cat. No. 5306000006) *This can be omitted by processing samples in batches using the Dynamag-2 magnet*

PCR Plate, 96-well, segmented, semi-skirted (Thermo Scientific, cat. No. AB-0900)

Appendix B - Solutions

2X B&W buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8.0, 2 M NaCl)

Buffer C (1X PBS, 0.01% Tween 20)

Buffer D (10 mM Tris-HCl pH 8.0, 2 mM EDTA pH 8.0, 0.01% Tween 20)

Pronase stock (10 mg/mL in water – stable 6 mo. at 4 °C)

Cleaning Solution (1X PBS, 1mM CaCl₂, 15 µg pronase)

5X Ficoll loading buffer (18 mM Tris Base, 10 mM Boric Acid, 0.4 mM EDTA, 3% Ficoll Type 400, 0.02% Bromo Blue, 0.02% Xylene Cyanol – stable 6 mo. at 4 °C)

Lonza elution buffer (0.5 mM Ammonium Acetate, 15 μ M Magnesium Acetate, 1 mM EDTA pH 8.0, 1% SDS)

10X Annealing buffer (0.1 M Tris-HCl pH 7.5, 0.5 M NaCl, 10 mM EDTA pH 8.0)

dsDNA Universal adapter (50 μ M Truseq_Universal_Adapter_F (Appendix C), 50 μ M Truseq_Universal_Adapter_R (Appendix C), 1X Annealing buffer. Heat at 95 °C for 5 min. Remove heat block and let cool down to room temperature for annealing of the two primers)

Appendix C – Oligos (attached Appendix C.xlsx)

Appendix D – Java tools and output files description (supplementary materials)

Java tools can be downloaded from github: https://github.com/piotrsan/3seq_javaUtilities

References

1. Amara SG, Jonas V, Rosenfeld MG, Ong ES, Evans RM. Nature. 1982; 298:240–244. [PubMed: 6283379]
2. Alt FW, Bothwell AL, Knapp M, Siden E, Mather E, Koshland M, Baltimore D. Cell. 1980; 20:293–301. [PubMed: 6771018]
3. Oszolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM. Cell. 2010; 143:1018–1029. [PubMed: 21145465]
4. Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V, Attie O, Chen K, Salehi-Ashtiani K, Vidal M, Harkins TT, Bouffard P, Suzuki Y, Sugano S, Kohara Y, Rajewsky N, Piano F, Gunsalus KC, Kim JK. Science. 2010; 329:432–435. [PubMed: 20522740]
5. Jan CH, Friedman RC, Ruby JG, Bartel DP. Nature. 2011; 469:97–101. [PubMed: 21085120]
6. Lianoglou S, Garg V, Yang JL, Leslie CS, Mayr C. Genes & development. 2013; 27:2380–2396. [PubMed: 24145798]
7. Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. Genome research. 2012; 22:1173–1183. [PubMed: 22454233]
8. Smibert P, Miura P, Westholm JO, Shenker S, May G, Duff MO, Zhang D, Eads B, Carlson J, Brown JB, Eisman RC, Andrews J, Kaufman TC, Cherbas P, Celniker SE, Graveley BR, Lai EC. Cell reports. 2012; 1:277–289. [PubMed: 22685694]
9. Tian B, Manley JL. Nature reviews. Molecular cell biology. 2017; 18:18–30. [PubMed: 27677860]
10. Gerstberger S, Hafner M, Tuschl T. Nature reviews. Genetics. 2014; 15:829–845.
11. Bartel DP. Cell. 2009; 136:215–233. [PubMed: 19167326]
12. Masamha CP, Xia Z, Yang J, Albrecht TR, Li M, Shyu AB, Li W, Wagner EJ. Nature. 2014; 510:412–416. [PubMed: 24814343]
13. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. Genome research. 2013; 23:812–825. [PubMed: 23520388]
14. Ji Z, Lee JY, Pan Z, Jiang B, Tian B. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106:7028–7033. [PubMed: 19372383]
15. Mayr C, Bartel DP. Cell. 2009; 138:673–684. [PubMed: 19703394]
16. Rhinn H, Qiang L, Yamashita T, Rhee D, Zolin A, Vanti W, Abeliovich A. Nature communications. 2012; 3:1084.
17. An JJ, Gharami K, Liao GY, Woo NH, Lau AG, Vanevski F, Torre ER, Jones KR, Feng Y, Lu B, Xu B. Cell. 2008; 134:175–187. [PubMed: 18614020]

18. Gruber AR, Martin G, Keller W, Zavolan M. Wiley interdisciplinary reviews. RNA. 2014; 5:183–196. [PubMed: 24243805]
19. Gruber AR, Martin G, Muller P, Schmidt A, Gruber AJ, Gumienny R, Mittal N, Jayachandran R, Pieters J, Keller W, van Nimwegen E, Zavolan M. Nature communications. 2014; 5:5465.
20. Spies N, Burge CB, Bartel DP. Genome research. 2013; 23:2078–2090. [PubMed: 24072873]
21. Gautheret D, Poirot O, Lopez F, Audic S, Claverie JM. Genome research. 1998; 8:524–530. [PubMed: 9582195]
22. Tian B, Hu J, Zhang H, Lutz CS. Nucleic acids research. 2005; 33:201–212. [PubMed: 15647503]
23. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Science. 2008; 320:1643–1647. [PubMed: 18566288]
24. Shenker S, Miura P, Sanfilippo P, Lai EC. RNA. 2015; 21:14–27. [PubMed: 25406361]
25. Pelechano V, Wilkening S, Jarvelin AI, Tekkedil MM, Steinmetz LM. Methods in enzymology. 2012; 513:271–296. [PubMed: 22929774]
26. Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. Nature methods. 2013; 10:133–139. [PubMed: 23241633]
27. Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. RNA. 2011; 17:761–772. [PubMed: 21343387]
28. Elkon R, Ugalde AP, Agami R. Nature reviews. Genetics. 2013; 14:496–506.
29. Wu TD, Nacu S. Bioinformatics. 2010; 26:873–881. [PubMed: 20147302]
30. Li H, Durbin R. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
31. Love MI, Huber W, Anders S. Genome biology. 2014; 15:550. [PubMed: 25516281]
32. Thorvaldsdottir H, Robinson JT, Mesirov JP. Brief Bioinform. 2013; 14:178–192. [PubMed: 22517427]

Highlights

- 3'-seq is a method to map the 3' ends of transcripts genome-wide at single nucleotide resolution.
- 3'-seq provides quantitative measure of 3' UTR isoforms expression.
- Step-by-step protocol to generate multiplexed libraries of the 3' end of polyadenylated RNAs

Figure 1.

Overview of the 3'-seq protocol. IUPAC codes, V=A/C/G; B=C/G/T; N=A/C/G/T.

Figure 2.

Example of high quality total RNA. Total RNA quality is assessed on the basis of the quality of the predominant signal from rRNA (as labeled). The trace is from high quality total RNA from *Drosophila melanogaster* run on an Agilent Bioanalyzer 2100. Notice that insect 28S rRNA dissociates into two subunits of equal size that co-migrate with the 18S rRNA. The migration of rRNA of other organisms will vary and should be taken into account when validating total RNA quality. Bioanalyzer output is shown. On the right the densitometry plot with the standard and the sample are shown. On the left an electropherogram of the sample is plotted with fluorescent intensity (FU) on the red y axis and size (nt) on the x axis. The marker run together with the sample is labeled. STD- standard.

Figure 3.

Optimization of fragmentation time. Agilent Bioanalyzer 2100 traces of the reaction outlined in 2.2.3 stopped at different time intervals. The chemical fragmentation reaction should be stopped when the total RNA peak is around 150–200 nt (10 min) but before the RNA is completely fragmented as shown in the later time points. This step was optimized for fragmentation of *Drosophila* total RNA and should be optimized when using this protocol to determine 3' ends from total RNA of other organisms. Densitometry of the samples run is shown to the left while electropherograms are shown on the right. Samples are labeled above each lane and on each corresponding inset on the right. The marker, present in each sample, is labeled in the top most panel. STD- standard. FU - fluorescent intensity.

Figure 4.

3'-seq library size selection. 3'-seq amplified library from step 2.2.13 was run on an 8% polyacrylamide TBE gel as outlined in 2.2.14. (A) Image of the library before excision of the relevant range. Bands caused by amplification of adapter-adapter sequences are observed below 150 bp (~120 bp). Care should be taken to avoid excision of these bands. (B) Image of the library post excision. The library is excised between 175–250 bp. This range is optimized for sequencing 1 X 50 bp. The range could be extended if sequencing using longer reads. This size range enriches for reads that contain the 3' end/polyA tail junction.

Figure 5.

Example Agilent Bioanalyzer 2100 trace of 3'-seq library. High quality library should result in a size range distribution as expected from the size range extracted from the gel. Shown is a 3'-seq library that falls in the 150–250 bp size range (labeled). Isolation of a fraction of reads below 175bp is to be expected and will still results in reads that can be mapped to the genome. The peak should be above 130 bp to avoid contamination of adapter-adapter reads (~120 bp). On the right side the densitometry plot with the standard and the sample are shown. On the left an electropherogram of the sample is plotted with fluorescent intensity (FU) on the red y axis and size (nt) on the x axis. The markers run together with the sample are labeled. STD- standard.

Figure 6.

Example of 3'-seq raw reads and 3' end calls. Pileup of raw 3'-seq reads from libraries constructed from head, ovary or testis *Drosophila* RNA and 3' end calls obtained using the computational pipeline described in this protocol. *wake* and *CG11791* are examples of genes with 3' end isoforms which lead to the expression of mRNAs that differ in 3' UTR length, while *Nmnat* is an example of a gene which expresses a 3' end isoform through recognition of a polyadenylation site in an intron, which leads to the expression of an mRNA with a different 3' UTR and protein C-terminus. Scale is shown above the tracks and RPM levels are shown for each track. Gene models are from flybase (r6.12). Images were obtained using IGV.