Video Article

3' End Sequencing Library Preparation with A-seq2

Georges Martin[1], Ralf Schmidt[1], Andreas J. Gruber[1], Souvik Ghosh[1], Walter Keller[1], Mihaela Zavolan[1,2]

[1]Computational and Systems Biology, Biozentrum, University of Basel

[2]Swiss Institute of Bioinformatics, Biozentrum, University of Basel

Correspondence to: Mihaela Zavolan at mihaela.zavolan@unibas.ch

URL: https://www.jove.com/video/56129
DOI: doi:10.3791/56129

Keywords: Molecular Biology, Issue 128, Polyadenylation, poly(A) site, pre-mRNA 3' end sequencing, A-seq2, data analysis software, deep sequencing

Date Published: 10/10/2017

Citation: Martin, G., Schmidt, R., Gruber, A.J., Ghosh, S., Keller, W., Zavolan, M. 3' End Sequencing Library Preparation with A-seq2. *J. Vis. Exp.* (128), e56129, doi:10.3791/56129 (2017).

Abstract

Studies in the last decade have revealed a complex and dynamic variety of pre-mRNA cleavage and polyadenylation reactions. mRNAs with long 3' untranslated regions (UTRs) are generated in differentiated cells whereas proliferating cells preferentially express transcripts with short 3'UTRs. We describe the A-seq protocol, now at its second version, which was developed to map polyadenylation sites genome-wide and study the regulation of pre-mRNA 3' end processing. Also this current protocol takes advantage of the polyadenylate (poly(A)) tails that are added during the biogenesis of most mammalian mRNAs to enrich for fully processed mRNAs. A DNA adaptor with deoxyuracil at its fourth position allows the precise processing of mRNA 3' end fragments for sequencing. Not including the cell culture and the overnight ligations, the protocol requires about 8 h hands-on time. Along with it, an easy-to-use software package for the analysis of the derived sequencing data is provided. A-seq2 and the associated analysis software provide an efficient and reliable solution to the mapping of pre-mRNA 3' ends in a wide range of conditions, from 10^6 or fewer cells.

Video Link

The video component of this article can be found at https://www.jove.com/video/56129/

Introduction

The capture and sequencing of mRNA 3' ends allows the study of mRNA processing and the quantification of gene expression. Due to their poly(A) tails, eukaryotic mRNAs can be efficiently purified from total cell lysates with bead-immobilized oligo-deoxythymidine (oligo(dT)) molecules, which can also prime cDNA synthesis. However, this approach has two drawbacks. First, stretches of A's that are internal to transcripts can also prime cDNA synthesis, resulting in spurious poly(A) sites. Second, homogeneous poly(A) stretches pose specific challenges for sequencing, aside from not being informative for transcript identification. Various approaches have been proposed to circumvent these limitations, such as reverse transcription through poly(A) tails followed by RNase H digestion (3P-seq [1]), use of a custom sequencing primer ending in 20 Ts (2P-seq [2]), preselection of RNA fragments with poly(A) tails of over 50 nucleotides with a CU_5T_{45} primer followed by RNase H digestion (3'READS [3]), and the use of a oligo-dT primer that contains the 3' adapter in a hairpin (A-seq [4]).

The recently developed A-seq2 method [5] aims to bypass sequencing through poly(A) and at the same time to minimize the proportion of dimers that are generated by self-ligation of adapters, particularly occurring when the molar concentration of adapters outweighs the insert concentration. This problem can be eliminated when both adapters are ligated to the same type of polynucleotide ends as in A-seq2, where the 3' adapters are ligated to the 5' end of RNA fragments and the 5' adapters to 5' ends of the cDNAs after reverse transcription. The method is more convenient than our previously proposed A-seq - in which sequencing was in the 5'-to-3' direction thereby requiring precisely controlled RNA fragmentation-, while maintaining a high accuracy of poly(A) site identification. Around 80 % of the sequenced reads in typical samples map uniquely to the genome and lead to identification of over 20,000 poly(A) site clusters, more than 70 % of which overlapping with annotated 3'UTRs.

In brief, the A-seq2 protocol starts with mRNA fragmentation and ligation of reverse-complement 3' adapters to the 5' ends of RNA fragments. Poly(A)-containing RNAs are then reverse transcribed with a 25 nucleotide (nt) long oligo(dT) primer that contains an anchor nucleotide at the 3' end, a dU at position 4 and a biotin at the 5' end, allowing binding of the cDNA to magnetic streptavidin beads. Most of the primer, including the biotin, is removed from the cDNA by cleavage at dU by the USER enzyme mix, containing Uracil DNA glycosylase (UDG) and the DNA glycosylase-lyase Endonuclease VIII. This reaction leaves intact ends for the ligation of a 5' adapter, and the three Ts left after cleavage remain to mark the location of the poly(A) tail. Because both 5' and 3' adapters are attached by ligation to recipient 5' ends, no adapter dimers are generated. Four nucleotide random-mers introduced at the beginning of reads allows cluster resolution on state-of-the-art sequencing instruments and can also serve as unique molecular identifier (UMI) for the detection and removal of PCR amplification artifacts. The size of the UMI can be further increased as done in other studies [6]. The protocol generates reads that are reverse complementary to mRNA 3' ends, all starting with a randomized tetramer followed by 3 Ts. Processing of reads that have the 3 diagnostic Ts at their 5' end starts with the correction of PCR amplification artifacts by exploiting the UMIs, removal of 3' adapter sequences, and reverse complementation. Reads that may have originated from oligo(dT) priming at internal A-rich sites are also identified computationally and discarded. The spurious sites generally lack one of the 18 well-characterized and conserved poly(A) signals which should be located ~21 nucleotides upstream of the apparent cleavage site [7].

The protocol requires about 8 h hands-on time, not counting cell culture and the overnight ligations. The associated read analysis software enables a highly accurate poly(A) site identification. From the poly(A) site clusters created based on 4 samples further highlighted in this manuscript (two biological replicates of control siRNA and si-HNRNPC-treated cells) 84% overlap with an annotated gene, and of these, 75 % overlap with a 3' UTR, and 86 % with either a 3' UTR or a terminal exon. The Pearson correlation coefficient of expression of 3' ends in the replicate samples is 0.92, and values of over 0.9 are typically obtained with the method. Thus, A-seq2 is a convenient method that gives very reproducible results.

Protocol

1. Cell Growth and mRNA Isolation

1. Grow cells according to your experimental design in 6-well plates to ~1 x 10^6 cells per well at 80% confluence.
2. Remove the growth medium and wash the cells once with phosphate buffered saline. Directly lyse the cells on the plate by adding 1 mL of lysis buffer from the mRNA-isolation kit. Transfer the viscous lysate into a 15 mL plastic tube with a 1 mL pipette tip. Use a rubber spatula to completely detach the cell material from the plate surface.
3. Shear the lysate containing viscous DNA with a 1 mL syringe attached to a 23 G hypodermic needle by several vigorous up and down movements of the plunger until the lysate is no longer viscous. Point the syringe needle to the center of the bottom to avoid ejecting the lysate out of the tube.
4. Transfer the lysate into a 1.5 mL tube using the syringe. Spin 5 min at 20,000 x g and 4 °C to remove the debris. Use DNA low bind 1.5 mL vials throughout the protocol.
5. While the centrifuge is running, wash 300 µL of resuspended oligo $(dT)_{25}$ magnetic beads on a magnetic rack with 500 µL of lysis buffer. Mix the tubes 2-3 times on the rack. Remove the buffer after the solution is clear. Collect the clear supernatant from step 1.4 and add to the beads. Resuspend and place tubes on a rotating wheel for 10 min.
6. Place the tubes on a magnetic rack. Remove the clear liquid after 2 min. Add 0.8 mL buffer A from the mRNA-isolation kit. Turn the tube by 180° degrees on the rack, 2-3 times. Repeat this washing step once more with buffer A.
7. Wash the beads 2 times with 0.8 mL of buffer B as described in step 1.6.
8. To elute the bound mRNA from the beads, add 33 µL H_2O and resuspend the beads. Heat to 75 °C for 5 min on a heated block. Immediately spin the tubes for 1 s and place them on the magnetic rack. Transfer the supernatant to a new tube. Samples can be stored at -80 °C until further use.
9. Add 66 µL alkaline hydrolysis buffer to the 33 µL mRNA (step 1.8), mix and heat for exactly 5 min at 95 °C on a heating block. Immediately chill tubes on ice.
10. **Isolate RNA with an RNA cleanup kit.**
    NOTE: Confirm the volume; it should be 100 µL.
    1. Add 350 µL RLT buffer from the kit and 250 µL ethanol. Load onto the column and spin for 30 s at 8,000 x g at room temperature (RT). Wash with 500 µL RPE buffer from the kit. Wash with 500 µL 80% ethanol. Spin for 5 min at 20,000 x g to dry the column. Add 36 µL H_2O to column and spin for 1 min at 20,000 x g. Discard the column and save the eluate.

2. 5' end Phosphorylation and DNase Treatment

1. Add 5 µL polynucleotide kinase buffer, 5 µL 10 mM ATP, 1 µL ribonuclease inhibitor, 1 µL DNase and 2 µL polynucleotide kinase to samples and incubate at 37 °C for 30 min. Optionally prepare master reaction mixes throughout the protocol by mixing 1.1 volumes x n (n = number of samples) of each component.
2. **Change buffer and remove ATP on a spin-column to prevent poly(A) addition in the next step.**
    1. Prespin spin-columns at 735 x g for 1 min. Transfer the columns to new 1.5 mL vials and load the kinase reactions onto the columns. Spin the columns 2 min at 735 x g. Discard the columns and place the tubes with collected reactions on ice or store at -80 °C.

3. Blocking 3' Ends with Cordycepin Triphosphate

NOTE: It is essential to block the 3' ends of RNA fragments to avoid their concatemerization in the subsequent ligation reactions. 3' ends that are not already blocked by a (cyclic) phosphate after hydrolysis are treated by addition of a 3' dATP (cordycepin triphosphate) chain terminator nucleotide with the aid of poly(A) polymerase. Here, yeast poly(A) polymerase (yPAP), that was expressed and purified as described in [8] was used at a concentration of 0.5 mg/mL. Yeast or *E. coli* PAP both have almost the same activity for addition of 3'dATP and can be purchased commercially (see the Table of Materials).

1. Add 13.5 µL 5x concentrated poly(A) polymerase reaction buffer, 2 µL of 10 mM 3' dATP, 1 µL RNase inhibitor and 1 µL poly(A) polymerase to the reaction from step 2.2.1. Mix and spin for 1 s. Incubate at 37 °C for 30 min. Add 32.5 µL H_2O to each reaction. Purify the RNA as in step 1.10.1. Elute the RNA with 14 µL H_2O.

4. Ligation of Reverse 3' Adapters to the 5' End of RNA Fragments

1. Place the reactions in a vacuum concentrator for 10 min to reduce the volume to 6 µL. Add 3 µL 10x T4 RNA ligation buffer, 3 µL 10 mM ATP, 15 µL PEG--8000, 1 µL RNase inhibitor, 1 µL of 0.1 mM reverse complement 3' adapter "revRA3" (see the Table of Materials) and 1 µL high concentration RNA ligase 1, mix.
2. Incubate the reactions at 24 °C for 16 h on a heated mixer with intermittent mixing at 1,000 rpm. Add 70 µL H_2O to each reaction and mix. Purify the RNA as in step 1.10.1. Elute the RNA with 14 µL H_2O. Samples can be stored at -80 °C at this point.

5. Reverse Transcription (RT)

1. Place the eluates in a vacuum concentrator for 3 min to reduce the volume to 11 µL. Transfer reactions to 200 µL PCR tubes. Add 1 µL 0.05 mM RT primer "Bio-dU-dT25". Heat for 5 min at 70 °C in a PCR cycler and leave at RT for 5 min.
2. Add 1 µL 10 mM dNTPs, 4 µL 5x reverse transcriptase buffer, 1 µL 0.1 M DTT, 1 µL RNase inhibitor, and 1 µL reverse transcriptase. Mix and heat the reactions for 10 min to 55 °C and 10 min to 80 °C in a PCR cycler. Keep on ice or at -80° C for longer storage.

6. Digestion with Uracil DNA Glycosylase Enzyme Mix

1. Pipet 100 µL Streptavidin-beads into a 1.5 mL vial, resuspend in 800 µL biotin binding buffer and place on a magnetic rack. Invert tubes 2-3 times. Remove the buffer when clear. Repeat the washing step. Resuspend the beads in 200 µL biotin binding buffer.
2. Add reverse transcription reaction to the beads solution and incubate 20 min at 4 °C on a rotating wheel. Wash the beads 2x with biotin binding buffer as in step 6.1 and 2x with TEN buffer on a magnetic rack. Resuspend the beads in 50 µL TEN buffer, add 2 µL Uracil DNA glycosylase enzyme mix, and incubate 1 h at 37 °C in a mixer with intermittent mixing.
3. Add 50 µL H_2O, 11 µL of RNase H buffer and 1 µL RNase H to the reactions. Incubate at 37 °C for 20 min. Place tubes on a magnetic rack and transfer the liquid containing the cleaved cDNA to a new tube
4. **Purify the cleaved cDNA.**
    1. Add 550 µL of buffer PB from the PCR purification kit to the cleavage reactions. Add 10 µL of 3 M sodium acetate, pH 5.2 to lower the pH. Load the reactions on minimal elution spin columns and spin at 17,000 x g for 1 min.
    2. Add 750 µL buffer PE to columns and spin at 17,000 x g for 1 min. Discard the flow-through. Spin the columns at 17,000 x g for 1 min to dry. Transfer the columns to a 1.5 mL vial, add 16 µL H_2O and spin at 17,000 x g for 1 min. Place the reactions in a vacuum concentrator for 8 min to concentrate to a volume of 7 µL.

7. Ligation of 5' Adapters to 5' Ends of cDNA

1. To the isolated cDNA, add 3 µL 10x T4 RNA ligase 1 buffer, 3 µL 10 mM ATP, 15 µL PEG--8000, 1 µL 50 µM "revDA5" oligo, and 1 µL high concentration T4 RNA ligase 1. Incubate at 24 °C for 20 h. Add 70 µL H_2O to each reaction. Samples can be stored at -20 °C at this point.

8. Pilot PCR, Amplification of Libraries and Size Selection

1. **In a pilot reaction, determine the optimal number of PCR cycles to reach library amplification within the exponential phase.**
    1. Pipet 25 µL DNA polymerase mix, 20 µL ligation reaction, 2 µL H_2O, 1.5 µL 10 µM forward PCR primer (RP1) and 1.5 µL 10 µM reverse PCR index primer into 200 µL PCR tube.
    2. Run the cycler with the following program: 3 min 95 °C, followed by 20 cycles of 20 s 98 °C, 20 s 67 °C and 30 s 72 °C. Collect 7 µL aliquots after 6, 8, 10, 12, 14, 16 and 18 cycles directly from the cycler. Add 1 µL 10x loading buffer (50% glycerol, 0.05% xylene cyanol).Note: Please follow the recommendations of the supplier if using multiplexing when combining barcodes.
    3. Separate products in small slots on a 2% agarose gel in 1x TBE buffer containing a 1:10,00 dilution of fluorescent green dye.
        1. Load aliquots on a 2% agarose gel and run the gel at 100 volts for 15 min. Visualize migration of PCR products on a gel documentation system.

2. **Use the number of cycles at the beginning of exponential amplification in the pilot reaction for a large-scale PCR reaction with twice the volumes as used for the pilot reaction (Figure 2).**
    1. For large-scale PCR reactions, concentrate and desalt the reactions first with a PCR purification kit and separate the products on wide slots on 2% agarose gels in 1x TBE buffer.

3. Cut out gel slices containing 200-350 nt DNA products.Melt the gel in the chaotropic buffer at RT for up to 30 min.Extract DNA from the gel slices with a gel extraction kit. Do not heat to 50 °C to prevent bias in binding of A-rich DNA [9].
4. Submit for sequencing.
   NOTE: Typically, 50 cycles single-read (SR50) are sufficient (see, for *e.g.*, https://www.illumina.com/technology/next-generation-sequencing.html).

9. Data Processing

NOTE: The resulting sequencing data (in fastq format) are processed with software available in the gitlab repository (https://git.scicore.unibas.ch/zavolan_public/A-seq2-processing). The analysis includes four main steps: (1) downloading the git repository, (2) installation of a virtual environment, (3) setting specific parameters in the configuration file and (4) launching the analysis through ‘snakemake’ [10]. The entire analysis done in step 4 requires only one command. A detailed step-by-step description of the analysis can be found in the README file in the gitlab repository and a short description is available below. All individual processing steps are accomplished by the execution of publicly available tools, either from external sources or prepared in-house. The computational pipeline depends on an anaconda-based [11] python 3 virtual environment with the snakemake package available [10]. It runs on machines with Unix-like operating system and was tested in a Linux environment with the CentOS 6.5 operating system installed and 40 GB RAM available.Software dependencies are automatically controlled within the virtual environment. The following publicly available software tools are required and thereby installed together with the environment: snakemake (v3.9.1) [10],fastx toolkit (v0.0.14) [12], STAR (v2.5.2a) [13], cutadapt (v1.12) [14], samtools (v1.3.1) [14,15], bedtools (v2.26.0) [16,17].

1. **Data pre-processing from reads to cDNAs**
   NOTE: The sequencing depth may vary between runs and, depending on the instrument, data from one sample can be split over multiple sequence files. If this is the case, concatenate the files that correspond to one sample into a single input file that is used in the following steps.
    1. Convert the file from fastq to fasta format.
    2. Extract reads with a correct structure (3 thymidines at positions 5, 6 and 7 of the read).
       NOTE: A read that is correctly prepared according to the experimental protocol described above should have the structure (from the 5' end): 4-nucleotide barcode - 3 thymidines - reverse complement of transcript 3' end.
    3. Store the information about the starting tetramer in the description line of the sequence.
       NOTE: The tetramer serves as a unique molecular identifier (UMI) that facilitates the correction of amplification artifacts later in the analysis.
    4. Remove the first seven nucleotides from the read's 5' end.
    5. Correct for amplification artifacts by keeping only one copy of the reads with the same insert sequence and UMI.
    6. Remove the part of the 3' end that matches the adapter sequence and then reverse complement the sequence. Only proceed with reads that have a minimum length (default: 15 nt).
       NOTE: Depending on the length of the original mRNA fragment and the number of sequencing cycles, the 3' end of the read may contain part of the 3' adapter, which is removed in this step.

2. Extract all reads that fulfill the following criteria: maximum 2 unknown nucleotides ('N'), maximum 80 % As, and last nucleotide of the read not A. These reads are considered to be of sufficient quality to be used in the analysis.
3. **Map the reads to the genome with a tool that handles spliced reads and generates an output file in BAM format.**
    1. If STAR is used, create a file with the index of the genome to which the reads should be mapped. For the human genome, this step requires 35 GB of memory (RAM).
    2. Map the reads to the genome.
       NOTE: (STAR-specific notes) Soft-clipping is disabled in order to force the mapping of the 3' end of each read as this is the nucleotide immediately upstream of the cleavage site.

4. Convert the BAM to a BED-file. If a read maps to multiple locations, keep only those with the lowest edit distance.
   NOTE: The copy number of the read mapped at a specific location is used as score. Reads that map to multiple locations are counted fractionally at each location with a weight equal to 1/number of locations to which a read maps.
5. Collapse reads that vary by a likely sequencing error. If two distinct reads map to the same location (start and end position of the mappings are identical) and they share the same UMI, consider them as PCR duplicates and keep only one.
6. Infer all individual pre-mRNA 3' end processing sites.
   NOTE: An individual read provides evidence for a 3' end when its last four nucleotides are mapped to the genome without error. The position to which the 3' end of the read maps is stored as cleavage site.
7. Detect 3' end sites that could have originated from internal priming. Define the site as internal priming artifact when the 10 nt downstream of the cleavage site in the genome satisfy one of the following criteria: contains more than six As, contains six consecutive As, or starts with one of the following tetramers: AAAA, AGAA, AAGA, AAAG.
8. Generate a table of individual 3' end processing sites in BED format.
9. **Identify independently regulated poly(A) site clusters.**
   NOTE: The steps described here follow the procedure that was introduced in a prior publication [5].
    1. Start by collecting individual 3' end processing sites that were obtained in all samples of the study.
    2. Annotate known poly(A) signals [7] in the region of -60 to +10 nucleotides around each individual 3' end processing site.
    3. Identify poly(A) sites expressed above the background in each sample as follows.
        1. Sort the sites by their raw expression within the current sample. Traverse the list of sites from top to bottom, associating lower ranked sites with a higher ranked site if they are located within a predefined distance in the genome (default: 25 nt up- or downstream) from the high-ranking site.
           NOTE: All low-ranking sites associated with a high-ranking site define a cluster whose expression is the number of reads documenting all of these sites.
        2. Sort these clusters by expression and traverse the list of clusters from highest to lowest expression, determining the expression threshold *c* at which the percentage of clusters with an annotated poly(A) signal drop below a predefined threshold (default: 90%).
        3. Discard sites from any cluster below the cutoff.

    4. Cluster closely spaced 3' end sites obtained across samples.
       NOTE: Sort 3' end processing sites first by the number of supporting samples and then by the sum of the normalized read count (reads per million (RPM)) across samples. Traversing the list from top to bottom, associating lower-ranked sites with higher-ranked sites when their distance to the higher-rank site is not larger than a predefined limit (default: 12 nt). Whenever any of the constituting 3' end site overlaps with an annotated poly(A) signal or has a poly(A) signal directly downstream, the corresponding cluster is marked for further inspection to detect internal priming.
    5. Merge poly(A) site clusters.
       NOTE: When a cluster is marked as a putative internal priming candidate, it is either merged into a downstream cluster if the two clusters share their poly(A) signals or retained if the most downstream site in the cluster has a poly(A) signal located at a minimum

distance upstream (default: 15 nt). Finally, closely spaced clusters are merged if: (i) they share the same poly(A) signal(s), or (ii) the span of the resulting cluster does not exceed a maximum (default: 25 nt).

6. Store clusters in BED-file format with the total normalized read count from all 3' end sites in each cluster as score.

Representative Results

Poly(A)-containing RNA was isolated from cultured cells, fragmented by alkaline hydrolysis and cDNAs were made by reverse transcription with oligo(dT) primers. The resulting cDNA was immobilized on streptavidin beads, dU was cleaved in the uracil specific excision reaction, adapters were ligated to 5' and 3' ends of the cleaved fragment and the inserts were sequenced. **Figure 1** depicts a graphical outline of the experiment.

For HeLa and HEK293 cells, 10^6 cells were sufficient to identify poly(A) sites for the vast majority of protein-coding genes at the end of the procedure. However, for other cell types or tissues it may be necessary to test the saturation in the number of identified poly(A) sites as the number of cells used in the experiment increases. Representative results of the pilot PCR step and of the DNA fragment analysis of the sample before sequencing are shown in **Figure 2**.

Figure 3 shows the pre-processing steps of the computational analysis, starting from the fastq file obtained from the sequencer and ending with the quality-checked, adapter-trimmed reads that are ready to be mapped to the genome. **Figure 4** shows the analysis steps that start with the mapping of the reads to the corresponding genome and end with the catalog of mRNA 3' end processing sites that are identified in a particular sample.When multiple samples are analyzed, additional steps are carried out to match the 3' end processing sites that were found in individual samples and report their abundance across samples. These steps are shown in **Figure 5**.

Thus, once samples have been sequenced, the analysis of the resulting sequencing read files (in fastq format) through the available processing pipeline is straightforward. After adding the information about the samples to the configuration file, the execution of the pipeline will result in two main types of output files: 1) BED-files with all 3' end processing sites identified in individual samples (*e.g.* "sample1.3pSites.noIP.bed.gz"), and 2) a BED-file with all poly(A) site clusters (clusters.merged.bed) across all samples of the study. The output also includes the genome coordinates for all reads from each individual sample (*e.g.* "sample1.STAR_out/Aligned.sortedByCoord.out.bam") that can later be viewed in a genome browser like IGV[16]. Visual inspection of the read profile(s) generally provides a first glimpse of the distribution of poly(A) sites in the genome and the changes that occur upon the specific perturbations that were carried out in the study. For example, in **Figure 6** the response of a specific gene to the knock-down of the HNRNPC protein is shown.

Summaries of these genome-wide distributions are also provided (**Table 1**). Specifically, output files in the "counts/annotation_overlap" directory contain fractions of sites that overlap with specific annotated features (from the gtf file provided as input; annotated are: 3' UTR, terminal exon, exon, intron, intergenic). Finally, for each sample, results of individual processing steps are also saved (e.g. "sample1.summary.tsv"). This includes the numbers of:raw reads in each sample, reads that have the expected structure of the 5' end, reads that remain after collapsing full PCR duplicates, high-quality reads according to the criteria defined at step 9.2, reads that map uniquely to the genome (after collapsing those that resulted from sequencing errors, see step 9.5), multi-mapping reads (after collapsing those that resulted from sequencing errors, see step 9.5), raw (not clustered) 3' end processing sites in each sample, raw 3' end processing sites without potential internal priming candidates, unique 3' end processing sites from all samples without internal priming candidates, and final set of poly(A) site clusters.

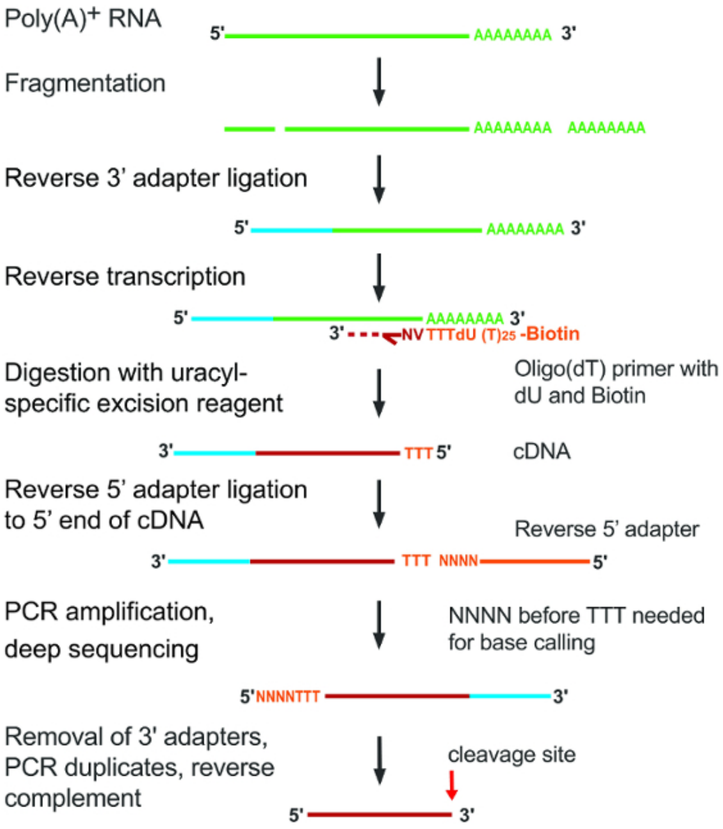

Poly(A)+ RNA
5' AAAAAAAA 3'
Fragmentation
AAAAAAAA AAAAAAAA
Reverse 3' adapter ligation
5' AAAAAAAA 3'
Reverse transcription
5' AAAAAAAA 3'
3' NV TTTdU (T)25 -Biotin
Digestion with uracyl-specific excision reagent
Oligo(dT) primer with dU and Biotin
3' TTT 5'
cDNA
Reverse 5' adapter ligation to 5' end of cDNA
Reverse 5' adapter
3' TTT NNNN 5'
PCR amplification, deep sequencing
NNNN before TTT needed for base calling
5'NNNNTTT 3'
Removal of 3' adapters, PCR duplicates, reverse complement
cleavage site
5' 3'

Figure 1: Main steps of the A-seq2 protocol. Individual steps are indicated on the left side of the figure. Insert RNA fragments are depicted as green lines that turn red for cDNA after reverse transcription; adapters are colored in light blue or in orange. Please click here to view a larger version of this figure.

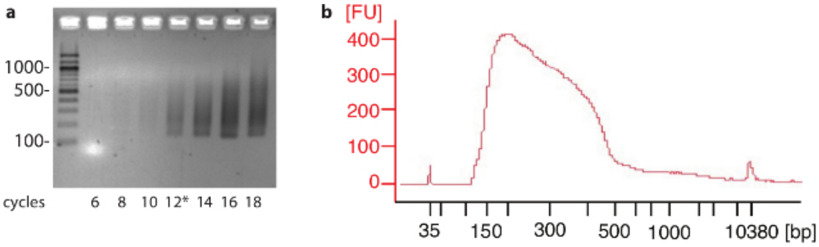
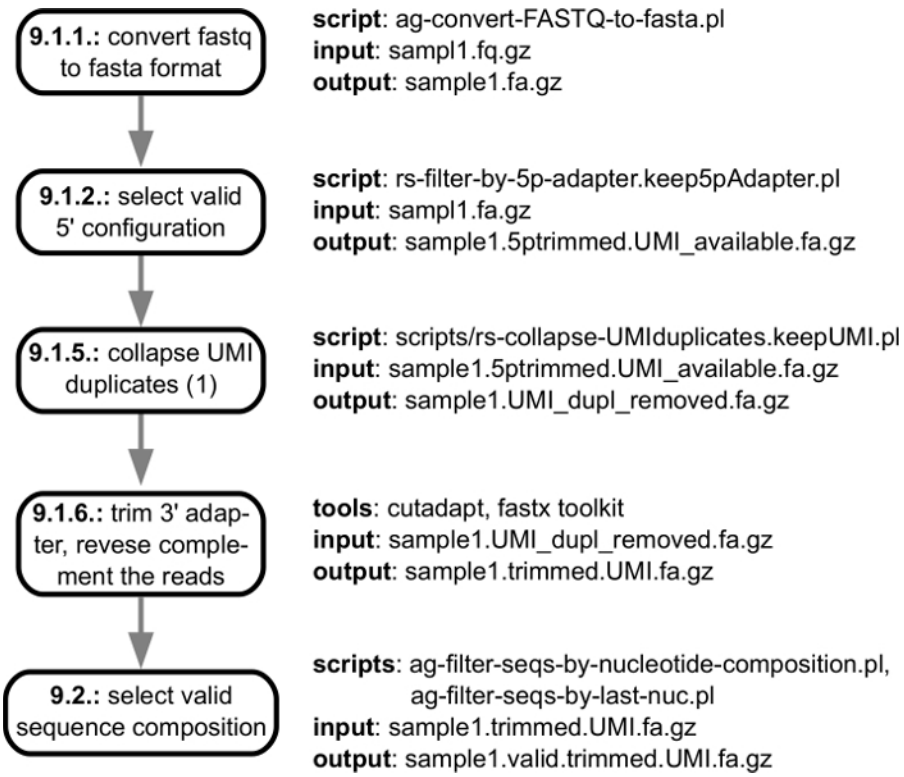


Figure 2: Pilot PCR and final product profile. (**a**) Aliquots from the PCR reaction were collected at different cycles and separated on 2% agarose gels. Numbers to the left indicate size in nucleotides of the respective bands in the DNA ladder. In this experiment 12 cycles (*) were chosen for the large scale PCR reaction. (**b**) Example of a sample after size selection run on a fragment size analyzer revealing an average size of around 280 nucleotides. Numbers to the left [FU] indicate relative signal intensity. Please click here to view a larger version of this figure.


9.1.1.: convert fastq to fasta format
9.1.2.: select valid 5' configuration
9.1.5.: collapse UMI duplicates (1)
9.1.6.: trim 3' adapter, revese complement the reads
9.2.: select valid sequence composition

script: ag-convert-FASTQ-to-fasta.pl
input: sampl1.fq.gz
output: sample1.fa.gz

script: rs-filter-by-5p-adapter.keep5pAdapter.pl
input: sampl1.fa.gz
output: sample1.5ptrimmed.UMI_available.fa.gz

script: scripts/rs-collapse-UMIduplicates.keepUMI.pl
input: sample1.5ptrimmed.UMI_available.fa.gz
output: sample1.UMI_dupl_removed.fa.gz

tools: cutadapt, fastx toolkit
input: sample1.UMI_dupl_removed.fa.gz
output: sample1.trimmed.UMI.fa.gz

scripts: ag-filter-seqs-by-nucleotide-composition.pl,
ag-filter-seqs-by-last-nuc.pl
input: sample1.trimmed.UMI.fa.gz
output: sample1.valid.trimmed.UMI.fa.gz

Figure 3: Outline of the pre-processing of sequencing reads. The fastq files with reads that are generated by the sequencing instrument-associated software are processed to identify high-quality reads that will be mapped to the corresponding genome. The figure shows the input/output specification of individual steps in the pipeline, with links to the individual steps of the protocol described in section "Data processing". Please click here to view a larger version of this figure.

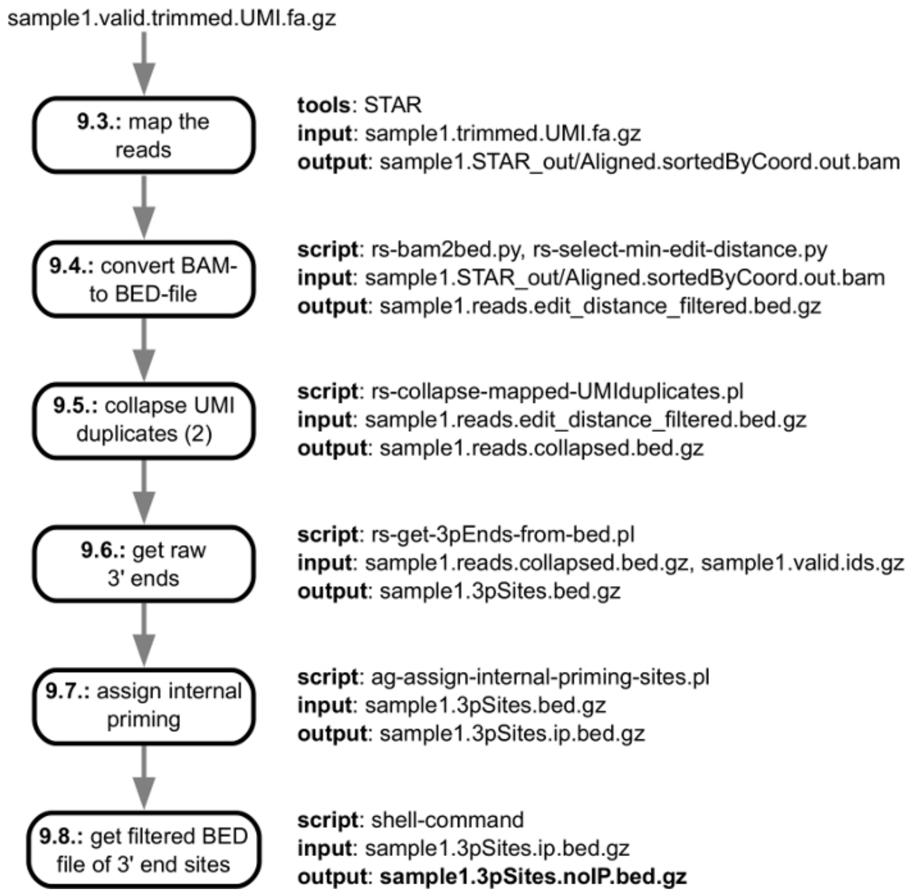
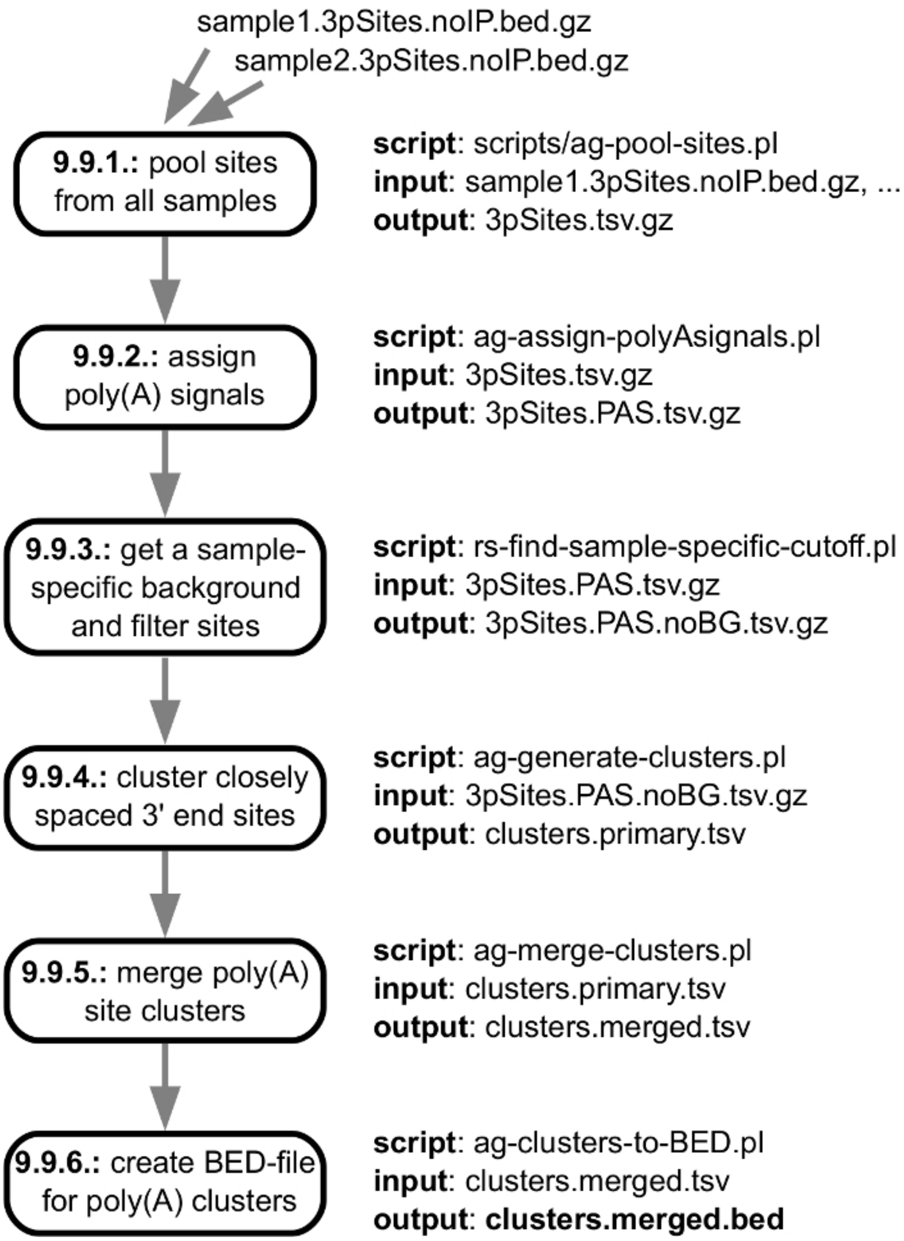

sample1.valid.trimmed.UMI.fa.gz
9.3.: map the reads
tools: STAR
input: sample1.trimmed.UMI.fa.gz
output: sample1.STAR_out/Aligned.sortedByCoord.out.bam
9.4.: convert BAM-to BED-file
script: rs-bam2bed.py, rs-select-min-edit-distance.py
input: sample1.STAR_out/Aligned.sortedByCoord.out.bam
output: sample1.reads.edit_distance_filtered.bed.gz
9.5.: collapse UMI duplicates (2)
script: rs-collapse-mapped-UMIduplicates.pl
input: sample1.reads.edit_distance_filtered.bed.gz
output: sample1.reads.collapsed.bed.gz
9.6.: get raw 3' ends
script: rs-get-3pEnds-from-bed.pl
input: sample1.reads.collapsed.bed.gz, sample1.valid.ids.gz
output: sample1.3pSites.bed.gz
9.7.: assign internal priming
script: ag-assign-internal-priming-sites.pl
input: sample1.3pSites.bed.gz
output: sample1.3pSites.ip.bed.gz
9.8.: get filtered BED file of 3' end sites
script: shell-command
input: sample1.3pSites.ip.bed.gz
output: sample1.3pSites.noIP.bed.gz

Figure 4: Outline of sequence read processing, from the step of mapping to the genome to the generation of individual 3' end processing sites. The figure shows the input/output specification of individual steps in the pipeline, with links to the individual steps of the protocol described in section "Data processing". The main output file that is delivered to the user is marked in bold. Please click here to view a larger version of this figure.


sample1.3pSites.noIP.bed.gz
sample2.3pSites.noIP.bed.gz
9.9.1.: pool sites from all samples
9.9.2.: assign poly(A) signals
9.9.3.: get a sample-specific background and filter sites
9.9.4.: cluster closely spaced 3' end sites
9.9.5.: merge poly(A) site clusters
9.9.6.: create BED-file for poly(A) clusters

script: scripts/ag-pool-sites.pl
input: sample1.3pSites.noIP.bed.gz, ...
output: 3pSites.tsv.gz

script: ag-assign-polyAsignals.pl
input: 3pSites.tsv.gz
output: 3pSites.PAS.tsv.gz

script: rs-find-sample-specific-cutoff.pl
input: 3pSites.PAS.tsv.gz
output: 3pSites.PAS.noBG.tsv.gz

script: ag-generate-clusters.pl
input: 3pSites.PAS.noBG.tsv.gz
output: clusters.primary.tsv

script: ag-merge-clusters.pl
input: clusters.primary.tsv
output: clusters.merged.tsv

script: ag-clusters-to-BED.pl
input: clusters.merged.tsv
output: **clusters.merged.bed**

Figure 5: Outline of the steps that are taken to generate clusters of co-regulated 3' end sequencing sites. The figure shows the input/output specification of individual steps in the pipeline, with links to the individual steps of the protocol described in section "Data processing". The main output file is marked in bold. Please click here to view a larger version of this figure.

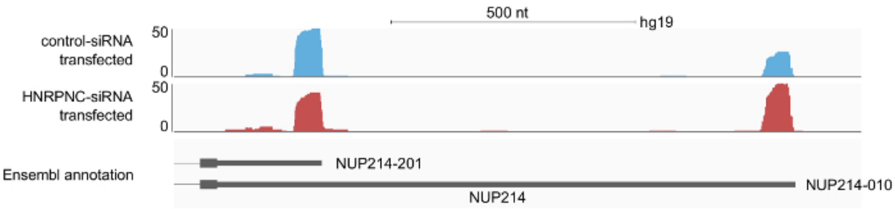


Figure 6: Example results of the profile of 3' end processing reads along the terminal exon of the NUP214 gene, shown in the IGV [16] genome browser. A-seq2 reads were prepared from two samples of HEK 293 cells, treated either with a control-siRNA or with an HNRNPC siRNA. The reads that documented poly(A) sites that were annotated by the analysis pipeline were saved in BAM format that was used as input to the IGV genome browser. The 3' ends of the read peaks map to mRNA 3' ends that are annotated in Ensembl. The profiles indicate an increased use of the long 3' UTR isoform upon HNRNPC knock-down. Please click here to view a larger version of this figure.

	si-Control replicate 1	si-Control replicate 2
	id: 29765	id: 32682
number of raw reads	44210258	68570640
number of valid reads after trimming and filtering	14024538	21211793
number of uniquely mapping reads	6953674	13946436
number of reads mapping to multiple loci	2040646	2925839
number of individual 3' end processing sites	1107493	1710353

Table 1: Example output of the analysis pipeline. Summaries of reads that were obtained at individual steps.

Discussion

The multitude of core and auxiliary factors that are involved in pre-mRNA 3' end processing is reflected in a correspondingly complex polyadenylation landscape. Additionally, polyadenylation is also responsive to changes in other processes such as transcription and splicing. 3' end cleavage sites in pre-mRNAs are typically identified based on the characteristic poly(A) tails that are added to the 5' cleavage products. Most methods use oligo(dT) primers of variable lengths that allow the specific conversion of poly(A)-containing mRNAs to cDNAs in a reverse transcription reaction. A common problem of this approach is internal priming to A-rich sequences resulting in artifactual cleavage sites. Two methods that aim to circumvent this artifact at the stage of sample preparation have been proposed. In the 3P-seq method [1], adapters are specifically ligated to the ends of poly(A) tails with help of a splint oligo followed by partial RNase T1 digestion and reverse transcription with TTP in the reaction as the only deoxynucleotide. The resulting poly(A)-poly(dT) heteroduplexes are then digested with RNase H and the remaining RNA fragments are isolated, ligated to adapters, and sequenced. A simpler and elegant method, 2P-seq, that uses a custom sequencing primer skipping the remaining oligo(dT) stretch in the sequencing reaction was reported by the same authors [2]. In a related method, 3'READS [3], an unusually long primer of 5 Us and 45 Ts, also containing a biotin is annealed to fragmented RNA, followed by stringent washes to select for RNA molecules with poly(A) tails of over 50 nucleotides. Although 3'READS drastically reduces the frequency of internal priming, it does not completely eliminate it [3]. Protocols for direct RNA sequencing have also been proposed, but the resulting reads are short and have a high rate of error and this approach has not been further developed [18,19,20]. The PolyA-Seq and the commercialized Quant Seq protocols combine oligo(dT) based priming with a random priming step for the cDNA second strand synthesis [20]. The use of the template switch reverse transcription reaction with the Moloney Murine Leukemia Virus (MMLV) reverse transcriptase leads to the generation of cDNAs with linkers in a single step and thereby no adapter dimers can appear in the PAS-Seq and SAPAS methods [21,22].

The A-seq2 method presented here stands out in its utilization of a cleavable nucleotide (dU) within a biotinylated oligo(dT) primer. This modification combines the utility of enriching oligo(dT) hybridized, polyadenylated targets with the removal of most of the oligo $(dT)_{25}$ sequence from the isolated fragments before libraries are prepared and the preservation of three Ts, which indicate the prior presence of the poly(A) tail. In contrast, methods that utilize RNase H to remove poly(A) from the RNA molecules randomly leave several As. Since in A-seq2, sequencing is done from the 3' end of the anti-sense strands, cleavage sites are predicted to be located after the NNNNTTT motif at the beginning of raw sequence reads. The randomized tetramers serve not only to allow base calling but also in the elimination of PCR amplification artifacts. Longer UMIs can also be accommodated. The possibility of internal priming remains in A-seq2 and is addressed computationally, first by discarding 3' ends with a genomically-encoded, A-rich downstream sequence and then by discarding 3' end clusters that could be explained by internal priming at the A-rich poly(A) signal itself. A recent analysis of poly(A) sites inferred uniquely by a large number of protocols indicates that the sites that are unique to A-seq2 have the expected nucleotide distribution and location within genes, similar to other 3' end sequencing protocols.

A critical step in A-seq2 is the selection of polyadenylated RNA and removal of ribosomal RNAs and various small RNAs. This is most easily done by an mRNA-isolation kit with oligo $(dT)_{25}$ magnetic beads. In principle, total RNA isolated with phenol containing solutions also gives high quality RNA that can be further subjected to selection by the mRNA-isolation kit or oligo (dT) agarose. A step that can be varied in A-seq2 is the treatment with alkaline hydrolysis which can be shortened or extended to obtain RNA fragments of different sizes. Critical is also that addition of 3'dATP to 3' ends of RNA fragments by the poly(A) polymerase is efficient. In the protocol described here, this treatment is applied to all RNA fragments, to avoid concatemerization during the ligation reaction. Finally, we note that although RNA ligase 1 is normally used as an RNA ligase, it also ligates efficiently single stranded DNA, as we have done here to ligate an adapter to the 5' end of the cDNA molecules.

Thus, A-seq2 is an efficient and easy to implement protocol for the identification of pre-mRNA 3' end processing sites. Future developments could include further reducing the complexity of the protocol and the amount of required material. The associated set of computational data analysis tools further enable the homogeneous processing of 3' end sequencing reads obtained with a wide range of protocols.

Disclosures

The authors have nothing to disclose.

Acknowledgements

The authors thank Mrs. Béatrice Dimitriades for help with the cell culture. This work was supported by the Swiss National Science Foundation grants #31003A_170216 and 51NF40_141735 (NCCR RNA & Disease).

References

1. Jan, C. H., Friedman, R. C., Ruby, J. G., & Bartel, D. P. Formation, regulation and evolution of Caenorhabditis elegans 3'UTRs. *Nature.* **469** (7328), 97-101 (2011).
2. Spies, N., Burge, C. B., & Bartel, D. P. 3' UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* (2013).
3. Hoque, M., Ji, Z., *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. methods.* **10** (2), 133-139 (2013).
4. Martin, G., Gruber, A. R., Keller, W., & Zavolan, M. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Rep.* **1** (6), 753-763 (2012).
5. Gruber, A. R., Martin, G., *et al.* Global 3' UTR shortening has a limited effect on protein abundance in proliferating T cells. *Nat. Commun.* **5**, 5465 (2014).
6. Kivioja, T., Vähärautio, A., *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. methods.* **9** (1), 72-74 (2011).
7. Gruber, A. J., Schmidt, R., *et al.* A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.* **26** (8), 1145-1159 (2016).
8. Lingner, J., & Keller, W. 3'-end labeling of RNA with recombinant yeast poly(A) polymerase. *Nucleic Acids Res.* **21** (12), 2917-2920 <https://www.ncbi.nlm.nih.gov/pubmed/7687347> (1993).
9. Quail, M. A., Kozarewa, I., *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. methods.* **5** (12), 1005-1010 (2008).
10. Rahmann, S. Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics .* **28** (19), 2520-2522 (2012).
11. Analytics, C. *Anaconda Software Distribution.* at <https://continuum.io> (2016).
12. Lab, H. *FASTX-Toolkit - Hannon Lab.* <http://hannonlab.cshl.edu/fastx_toolkit/index.html> (2017).
13. Dobin, A., Davis, C. A., *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics .* **29** (1), 15-21 (2013).
14. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal.* **17** (1), 10-12 (2011).
15. Li, H., Handsaker, B., *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25** (16), 2078-2079 (2009).
16. Robinson, J. T., Thorvaldsdóttir, H., *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29** (1), 24-26 (2011).
17. Quinlan, A. R., & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics .* **26** (6), 841-842 (2010).
18. Ozsolak, F., Platt, A. R., *et al.* Direct RNA sequencing. *Nature.* **461** (7265), 814-818 (2009).
19. Yao, C., Biesinger, J., *et al.* Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. *Proc. Natl. Acad. Sci. U. S. A.* **109** (46), 18773-18778 (2012).
20. Lin, Y., Li, Z., *et al.* An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Res.* **40** (17), 8460-8471 (2012).
21. Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J., & Shi, Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA .* **17** (4), 761-772 (2011).
22. Fu, Y., Sun, Y., *et al.* Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res.* **21** (5), 741-747 (2011).