BMB Reports - Manuscript Submission

Manuscript Draft

Manuscript Number: BMB-17-019

Title: Analyses of alternative polyadenylation: from old school biochemistry to high-throughput technologies

Article Type: Mini Review

Keywords: alternative polyadenylation; APA; high-throughput; mRNA 3'-end; RNA-seq

Corresponding Author: Jeongsik Yong

Authors: Hsin-Sung Yeh[1], Wei Zhang[2], Jeongsik Yong[1,*]

Institution: [1]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, Minnesota 55455, USA,
[2]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA,

Analyses of alternative polyadenylation: from old school biochemistry to high-throughput technologies

Hsin-Sung Yeh[1], Wei Zhang[2] and Jeongsik Yong[1]*

[1]Department of Biochemistry, Molecular Biology and Biophysics

[2]Department of Computer Science and Engineering

University of Minnesota, Minneapolis, Minnesota 55455

*Corresponding author e-mail: jyong@umn.edu

Abstract

Alternations in usage of polyadenylation sites during transcription termination yield transcript isoforms from a gene. Recent findings of transcriptome-wide alternative polyadenylation (APA) as a molecular response to changes in biology position APA not only as a molecular event of early transcriptional termination but also as a cellular regulatory step affecting various biological pathways. With the development of high-throughput profiling technologies at a single nucleotide level and their applications targeted to the 3’-end of mRNAs, dynamics in the landscape of mRNA 3’-end is measureable at a global scale. In this review, methods and technologies that have been adopted to study APA events are discussed. In addition, various bioinformatics algorithms for APA isoform analysis using publicly available RNA-seq datasets are introduced.

Introduction

Almost all of the protein coding messenger RNAs (mRNAs) in eukaryotic cells, with the exception of histone transcripts, are subjected to polyadenylation. It is a two-step event that occurs towards the completion of transcription in which the nascent transcript is cleaved, followed by the addition of an untemplated stretch of adenine nucleotides to its 3' end. The addition of the poly(A) tail has been shown to be important for the stability, nuclear export and translation of a transcript[1-3].

Polyadenylation is carried out by the coordination of various cis- and trans-acting factors. The poly(A) signal (PAS; AAUAAA hexamer and variants) located 10~30 nucleotides upstream of the cleavage site plays a central role in defining the poly(A) site of a transcript; it is assisted by U-rich upstream elements (USE) and GU-rich downstream elements (DSE), which have been shown to be able to affect the "strength" of the poly(A) signal[4-6]. As these cis-acting elements are transcribed, usually at the end of 3' untranslated region (UTR), they are recognized and bound by various multi-subunit protein complexes. PAS is recognized by CPSF (cleavage and polyadenylation specificity factors), and CSTF (cleavage stimulation factors) bind to DSE. Together, along with several other cofactors including CFI (cleavage factors I) and CFII, they perform cleavage at the poly(A) site. They then recruit poly(A) polymerase to the cleavage site for the synthesis of the poly(A) tail[2].

Most of the transcripts in mammalian cells (~70%) possess more than one PAS, providing the possibility of alternative polyadenylation (APA), that is, one transcript may carry out polyadenylation at two or more different sites[7]. APA can generally be divided into two types based on the location of the alternative poly(A) sites: UTR-APA and CR (coding region)-APA[8]. For UTR-APA, the alternative poly(A) site is located in the 3'UTR, while in the case of CR-APA, the alternative poly(A) site resides mostly in upstream introns. Naturally, the physiological outcomes of UTR-APA differ from that of CR-APA. UTR-APA can lead to the inclusion or exclusion of part of the 3'UTR in a transcript, and since 3'UTR often serves as the binding platform of many regulatory micro-RNAs and RNA binding proteins, UTR-APA is capable of affecting the transcript's localization, translation efficiency, stability, etc[9-11]. On the other hand, when CR-APA occurs, the usage of an upstream intronic poly(A) site results in the exclusion of part of the coding region, leading to the production of a truncated protein, possibly lacking certain functional domains and therefore may exert a different function or be regulated differently than the full-length counterpart[12].

In recent years, new technologies have enabled scientists to study APA at a transcriptome-wide scale. Interestingly, global APA events have been observed to be correlated to various cellular processes such as proliferation and differentiation[13-15]; they also show tissue specificity[16]. Moreover, it has been shown that global APA pattern changes during disease progression, including tumorigenesis[17-20]. These indicate that APA events are finely regulated in cells and that APA is one of the layers of gene expression regulation that control cellular biology. Therefore, the physiological outcomes of APA pattern changes in various cellular contexts and the mechanisms that govern APA have been the focus of many studies.

To study APA systematically, one must first be able to map out APA events at a transcriptome-wide level. Ever-advancing technology has allowed us to achieve this task with ever-improving precision. In this review, we will discuss some important technologies that have been adopted to study APA events.

Early Discoveries of APA

Some of the earliest APA discoveries were reported in the 1980s. For example, a CR-APA event was observed in IgM gene, and DHFR gene showed UTR-APA[21-24]. These cases of alternative processing events were first revealed by the discrepancies of sizes of the same gene in northern-blotting, and western-blotting, in the case of CR-APA. R-loop mapping and restriction mapping were then used to confirm that the differences reside in the 3’ end structure of the transcripts. In the following decade, dozens of APA events were discovered by the similar approach, albeit at a one-gene-at-a-time pace.

As technologies in molecular biology matured and sequencing data accumulated, APA studies were introduced to a more global scale in the 2000s. The first large-scale APA surveys were done by analyzing Expressed Sequence Tag (EST) data of human, mouse, and rat. To search for poly(A) sites in the genomes, Tian *et al.*[25], as well as Yan and Marr[26], first aligned ESTs to the genomes, then singled out 3’ end ESTs by looking for stretches of As and Ts at 5’ or 3’ termini of unaligned EST sequences. These 3’ end ESTs were then validated by searching for the presence of consensus PAS sequence patterns. Their analyses showed that a great proportion of genes (~50% in human and ~30% in murine) have APA. Moreover, many of the APA events are conserved between human and mouse; indicating that APA is a widely employed gene regulation strategy in cellular biology.

Analyzing EST data revealed the presence of APA in genes at a transcriptome-wide scale. However, the dynamics of global APA regulations remained elusive until microarray-based approaches were used to study global APA pattern changes[13,27,28]. In these studies, probes on the microarrays were designed to be APA sensitive. For each APA regulated gene, there are two or more probes specific to only the full-length transcript, and two or more probes specific to both the full-length and the shorter APA product on the microarray chip. After applying fluorescently-labelled nucleic acid library to the chip for hybridization, the ratio of the signals from these probes can then be calculated to measure the APA status of the gene. Microarray data obtained from two different cellular conditions can then be compared to study the APA dynamic changes and its physiological implications. Surprisingly, by adopting this approach it was shown that highly proliferating cells tend to have more 3’ UTR shortening in their transcriptomes; while generating induced pluripotent stem cells, global 3’ UTR lengthening is observed[13,27]. Microarray is a powerful tool to obtain a global picture on APA events in the transcriptome. Nevertheless, it suffers from several drawbacks. First of all, microarray cannot detect novel APA events, for APA-sensitive probe sets can only be designed if an APA events are previously known. Second, it cannot precisely pinpoint where the poly(A) site is, which may be important for studying the physiological functions of APA events. Moreover, if a gene has two or more alternative PAS, probe design and quantification can become quite complicated and challenging[29].

APA Studies in the Second-Generation Sequencing Era

The advent of second-generation sequencing technologies enabled researchers to rapidly obtain a large amount of sequence information at single nucleotide resolution. Technologies such as RNA-seq quickly became commonly used for surveying the transcriptome of various cell types and tissues in different organisms[30-32]. In RNA-seq, poly(A) tail-containing RNA is first isolated from total RNA. They are then either primed with oligo d(T) primer or random hexamers for cDNA synthesis followed by fragmentation. (Alternatively, the poly(A)-containing RNA pool is first fragmented followed by random hexamer priming to generate cDNA pool.) The cDNA pool is then amplified and constructed into library, which can be sequenced by various sequencing platforms, most commonly the Illumina sequencing technology. After mapping the short sequence fragments, or reads, to the corresponding genome, the reads can be piled up for visualization of gene expression profile of the cell or tissue.

The highly quantitative nature of RNA-seq makes it suitable for APA pattern analysis. This may be achieved in a similar way as the APA calculation done in microarray approaches, namely, by taking the ratio of the read density of the long form-only regions and the read density of the regions common to both long and short transcripts. However, since many genes contain isoforms with complicated and overlapping structures, using RNA-seq reads for APA analysis on certain genes can still be challenging. Fortunately, many sophisticated bioinformatics tools have been developed to more accurately analyze APA patterns in transcriptomes.

For instance, a probabilistic transcript quantification method named Kallisto was developed to estimate the expression levels of annotated transcript isoforms[33]. APA dynamics of a gene can then be measured by comparing the expression ratios of its short isoforms over the long isoforms between two biological conditions. This is the general method for CR-APA analysis.

For UTR-APA identification and measurement, many algorithms were written as listed in Table 1. In general, 3'UTR length changes are measured by modeling the RNA-seq read density changes near the 3' end of mRNA transcripts.

Algorithm	Reference	Description
DaPars	34	It first models the RNA-seq-read densities of both tumor and normal as a linear combination of both proximal and distal polyA sites. It then uses a linear regression model to identify the location of the *de novo* proximal polyA site, followed by quantification of the changes in APA between tumor and normal.
ChangePoint	35	It is based on a generalized likelihood ratio statistic for identifying 3'UTR length change in the analysis of RNA-seq data. A directional multiple test procedure is then developed to identifying APA events between two samples.
Roar	36	It is based on Fisher test to detect disequilibriums in the number of RNA-seq reads mapped to the 3'UTRs. Read counts and lengths of fragments are then used to calculate the prevalence of the short isoform over the long one in two biological conditions to identify APA events.
3USS	37	A web-server developed with the aim of giving experimentalists the possibility to identify alternative 3'UTRs between two samples by RNA-seq data analysis.
IsoSCM	38	A method for transcript assembly that incorporates change point analysis by a Bayesian framework to improve the 3'UTR annotation process with RNA-seq data.
KLEAT	39	An analysis tool that uses *de novo* assembly of RNA-seq data to characterize cleavage sites on 3'UTRs through direct observation of poly(A) tails.
GETUTR	40	It first makes a density function of RNA-seq reads aligned to the 3'UTRs using kernel density estimation. A smoothing step is then applied to maintain the biological changes of the 3'UTR. The goal of the method is to estimate the 3'UTR landscape based on these smoothed RNA-seq signal.

Indeed, with the aid of these bioinformatics tools, RNA-seq can be a powerful tool to study the alternative processing of mRNAs. However, when profiling APA patterns, especially when handling genes with multiple isoforms, often times the reads mapped to regions that differentiate isoforms constitute only a relatively small portion of the total reads mapped to the gene, and even less so for 3' end junction reads, making it rather challenging to confidently calculate the expression ratios of different isoform. Moreover, RNA-seq is not particularly accurate when it comes to identifying poly(A) sites, making novel APA isoform identification rather difficult. Therefore, several methods have been developed to address these issues by enriching for 3' end reads in high-throughput sequencing experiments[29,41,42].

The most common way to enrich for 3' end reads (adopted in PAS-seq[43], A-seq[44], 3SEQ[45], SAPAS[46], ect.) is to first fragment the poly(A) tail-containing RNA pool, followed by reverse transcription using oligo d(T) priming. The cDNA pool, which should only contain 3' end junction fragments, is then amplified and sequenced. Alternatively, an oligo d(T) primed cDNA library can be sequenced using oligo d(T) sequencing primer directly. All the sequencing reads should therefore be 3' end junction reads (PolyA-seq[7]). Moreover, direct RNA sequencing technology by Helicos Biosciences has also been used for sequencing the 3' ends of poly(A) tail-containing RNAs[47]. In this method, mRNA molecules are hybridized to a "lawn" of oligo d(T) primers attached to the flow cell and are sequenced directly by synthesis. Compared to PAS-seq and equivalents, Helicos platform is more quantitative as no amplification step is involved; it requires less starting material. However, Helicos platform suffers from higher error rate, shorter read lengths, lower throughput, and the lack of multiplexing capability[29,32].

All of the above mentioned methods use oligo d(T) for priming at some points during the procedures. Internal priming (stretches of As in the middle of transcripts being falsely recognized as poly(A) tails by the oligo d(T) primer), and thus false identification of 3' end junctions, is therefore a major issue for these methods[48]. In an effort to lower the false discovery rate, Jan *et al.*[48] developed a modified version of 3' end sequencing method that avoided the use of oligo d(T) priming, named 3P-Seq. After isolating poly(A) tail-containing RNAs, the first step of 3P-seq is to add a biotinylated double-stranded adapter to the 3’ end of the poly(A) tail through splint-ligation, which eliminates the possibility of internal priming. The mRNAs are then partially digested, and the 3’ end fragments are captured by streptavidin. cDNA synthesis is primed with the adapter itself and reverse-transcribed with dTTP as the only deoxynucleoside triphosphate present, limiting the reverse transcription to the poly(A) site. The RNA fragments immediately upstream of poly(A) tails can then be released and processed for sequencing by RNase H digestion. Since RNase H would only digest RNA strand that is hybridized with a DNA strand, in this case the poly(A) tail region, the RNA fragments released after RNase H digestion should most likely come from poly(A) tail-containing fragments. This method indeed eliminates a great number of false identified 3’ ends, yet it is more labor intensive and involves more enzymatic reactions, which may introduce biases in terms of the quantification of the signals[49].

3’ end sequencing data provides sophisticated knowledge for pinpointing annotated as well as unannotated poly(A) sites in the transcriptome that is under interrogation. As mentioned above, although RNA-seq is highly quantitative, yet it does not provide enough information to accurately identify poly(A) sites. Therefore, by incorporating 3’ end-seq data with RNA-seq data, the quality of APA profiling can be greatly improved. Briefly, by analyzing 3’ end-seq data, potential poly(A) sites and thus isoform structures in the transcriptome can be defined and reported. The expression levels of the isoforms can then be estimated by a maximum-likelihood method that best explains the observed RNA-seq read profiles. Finally, the APA events can be measured by the expression ratios of the isoforms in the gene between two biological conditions.

APA Studies Beyond the Second-Generation Sequencing Era

Second-generation sequencing technologies have revolutionized the research involving transcriptome characterization. However, when it comes to expression profiling of mRNA isoforms, these methods still suffer from their limitation in read lengths. Due to the relatively short read lengths (~100 bp), compared to the lengths of most transcripts, full-length transcript isoforms must be reconstructed via various computational methods. Yet the performances of the reconstruction methods have been shown to be unsatisfactory[50]. Alternative sequencing platforms have been developed to achieve longer read lengths.

For example, the SMRT (single molecule, real-time) sequencing technology by PacBio has achieved average read length of > 10,000 bp[51]. In SMRT technology, DNA polymerase is immobilized to the bottom of a specialized light detecting well called zero-mode waveguide (ZMW). ZMW is designed to only be light sensitive at the bottom of the well, where sequencing by synthesis is performed by the DNA polymerase. A movie that contains the sequencing information can then be recorded as a single DNA molecule is replicated by the DNA polymerase in full-length. PacBio has also developed a protocol specifically for transcript isoform characterization called Iso-Seq. It has been successfully adopted in characterizing the transcriptomes of human and herpesvirus[52,53]. The transcriptome dynamics during lineage commitment of blood cell and the progression of brain tumor have also been characterized by Iso-Seq[54,55]. Last but not least, it was recently used to profile the APA events in sorghum transcriptome[56]. In all of these studies, due to the long read lengths, novel isoforms have been identified and characterized with high confidence.

However, SMRT sequencing still suffers from certain drawbacks. For instance, longer transcripts still cannot be sequenced in full-length in high quality. This is partly due to the limitations in library preparation and the limitation of read length (or movie time), and also the fact that shorter cDNA molecules (~1.5 kb) are more favored by the sequencing platform. Moreover, since shorter transcripts are more favored than longer transcripts during sequencing, the quantitative performance of Iso-Seq is severely affected[51].

To harness the quantitative power of the short-read second generation sequencing and the isoform characterization ability of long-read PacBio sequencing, hybrid sequencing have been developed. By integrating the long read data from PacBio and the short read data from Illumina, Au *et al.* significantly reduced the error rate of long reads in PacBio sequencing[57]. When applying the hybrid sequencing to human embryonic stem cell transcriptome, they demonstrated the higher sensitivity and accuracy of isoform characterization, as well as a better ability to identify novel isoforms, over traditional methods solely using short-read second generation sequencing method[58].

Concluding Remarks

In recent years, APA have been appreciated more and more as a key layer of gene expression regulation mechanism. Since many studies have shown that the expression levels of certain transcript isoforms can have substantial impacts on biology[11,12,17,24,59,60], simply profiling the expression levels of genes in the transcriptomes cannot provide sufficient information regarding the physiological condition of a transcriptome. Indeed, with the methods (and their integrative efforts) discussed above, the transcript isoform expression profiling in transcriptomes can be performed with relatively high confidence. However, several issues remain to be addressed in the future to improve the transcriptome-wide studies of APA events. For example, many of the algorithms developed for APA characterization rely heavily on annotated gene structures. Therefore, a more comprehensive and accurate transcriptome annotation is needed for better performances by these algorithms in APA analysis. Furthermore, as shown in various studies adopting hybrid sequencing methods[61,62], being able to obtain long-read sequencing data by Iso-Seq is highly beneficial to transcript isoform expression profiling as well as novel isoform identification. However, the bias toward shorter transcripts over longer ones of Iso-Seq renders the isoform characterization of long transcripts unreliable. New technologies or sample preparation methods (e.g. the construction of a sequencing library with homogenous length by concatenation followed by fragmentation of the cDNA pool) need to be developed to address this issue for better characterization of APA events in long transcripts by hybrid sequencing methods.

With the advance of modern technologies and the rapid accumulation of the knowledge on the physiological outcomes of APA events in the field, in the near future, transcriptome-wide APA analysis can potentially become as routinely and easily performed as it currently is with gene expression profiling by RNA-seq; more importantly, valuable insights on various biological processes and the pathogenesis of diseases can be obtained by APA analysis.

References

1. Proudfoot NJ. Ending the message: Poly(A) signals then and now. *Genes & Development*. 2011;25(17):1770-1782. doi: 10.1101/gad.17268411.

2. Colgan DF, Manley JL. Mechanism and regulation of mRNA polyadenylation. *Genes & Development*. 1997;11(21):2755-2766. doi: 10.1101/gad.11.21.2755.

3. Zhang X, Virtanen A, Kleiman FE. To polyadenylate or to deadenylate: That is the question. *Cell Cycle*. 2010;9(22):4437-4449. http://www.landesbioscience.com/journals/cc/article/13887/.

4. Danckwardt S, Kaufmann I, Gentzel M, et al. Splicing factors stimulate polyadenylation via USEs at non□canonical 3′ end formation signals. *EMBO J*. 2007;26(11):2658-2669. doi: 10.1038/sj.emboj.7601699.

5. Graber JH, Cantor CR, Mohr SC, Smith TF. Genomic detection of new yeast pre-mRNA 3'-end-processing signals. *Nucleic Acids Research*. 1999;27(3):888-894. doi: 10.1093/nar/27.3.888.

6. Beaudoing E, Freier S, Wyatt JR, Claverie J, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. *Genome Research*. 2000;10(7):1001-1010. doi: 10.1101/gr.10.7.1001.

7. Derti A, Garrett-Engele P, MacIsaac KD, et al. A quantitative atlas of polyadenylation in five mammals. *Genome Research*. 2012;22(6):1173-1183. doi: 10.1101/gr.132563.111.

8. Tian B, Manley JL. Alternative cleavage and polyadenylation: The long and short of it. *Trends Biochem Sci*. 2013;38(6):312-320. doi: http://dx.doi.org/10.1016/j.tibs.2013.03.005.

9. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem*. 2010;79(1):351-379. http://dx.doi.org/10.1146/annurev-biochem-060308-103103. doi: 10.1146/annurev-biochem-060308-103103.

10. Hsin-Sung Yeh, and JY. Alternative polyadenylation of mRNAs: 3â€²-untranslated region matters in gene expression. *Mol Cells*. 2016;39(4):281-285. http://www.molcells.org/journalview.html?doi=10.14348/molcells.2016.0035.

11. Chang J, Zhang W, Yeh H, et al. mRNA 3prime]-UTR shortening is a molecular signature of mTORC1 activation. *Nat Commun*. 2015;6. http://dx.doi.org/10.1038/ncomms8218.

12. Hoque M, Ji Z, Zheng D, et al. Analysis of alternative cleavage and polyadenylation by 3prime] region extraction and deep sequencing. *Nat Meth*. 2013;10(2):133-139. http://dx.doi.org/10.1038/nmeth.2288.

13. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer MicroRNA target sites. *Science*. 2008;320(5883):1643-1647. doi: 10.1126/science.1155390.

14. Sandberg R, Neilson J, Sarma A, Sharp P, Burge C. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science (New York, N.Y.)*. 2008;320(5883):1643â€"1647. http://europepmc.org/abstract/MED/18566288. doi: 10.1126/science.1155390.

15. Ulitsky I, Shkumatava A, Jan CH, et al. Extensive alternative polyadenylation during zebrafish development. *Genome Research*. 2012;22(10):2054-2066. doi: 10.1101/gr.139733.112.

16. Zhang H, Lee J, Tian B. Biased alternative polyadenylation in human tissues. *Genome Biol*. 2005;6(12):R100. http://genomebiology.com/2005/6/12/R100. doi: 10.1186/gb-2005-6-12-r100.

17. Mayr C, Bartel DP. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*. 2009;138(4):673-684. doi: http://dx.doi.org/10.1016/j.cell.2009.06.016.

18. Singh P, Alley TL, Wright SM, et al. Global changes in processing of mRNA 3' untranslated regions characterize clinically distinct cancer subtypes. *Cancer Research*. 2009;69(24):9422-9430. doi: 10.1158/0008-5472.CAN-09-2236.

19. Morris AR, Bos A, Diosdado B, et al. Alternative cleavage and polyadenylation during colorectal cancer development. *Clinical Cancer Research*. 2012;18(19):5256-5266. doi: 10.1158/1078-0432.CCR-12-0543.

20. Lembo A, Di Cunto F, Provero P. Shortening of 3â€²UTRs correlates with poor prognosis in breast and lung cancer. *PLoS ONE*. 2012;7(2):e31129. http://dx.doi.org/10.1371%2Fjournal.pone.0031129.

21. Setzer DR, McGrogan M, Nunberg JH, Schimke RT. Size heterogeneity in the 3′ end of dihydrofolate reductase messenger RNAs in mouse cells. *Cell*. 1980;22(2):361-370. doi: http://dx.doi.org/10.1016/0092-8674(80)90346-3.

22. Alt FW, Bothwell ALM, Knapp M, et al. Synthesis of secreted and membrane-bound immunoglobulin mu heavy chains is directed by mRNAs that differ at their 3′ ends. *Cell*. 1980;20(2):293-301. doi: http://dx.doi.org/10.1016/0092-8674(80)90615-7.

23. Early P, Rogers J, Davis M, et al. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways. *Cell*. 1980;20(2):313-319. doi: http://dx.doi.org/10.1016/0092-8674(80)90617-0.

24. Rogers J, Early P, Carter C, et al. Two mRNAs with different 3′ ends encode membrane-bound and secreted forms of immunoglobulin μ chain. *Cell*. 1980;20(2):303-312. doi: http://dx.doi.org/10.1016/0092-8674(80)90616-9.

25. Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*. 2005;33(1):201-212. doi: 10.1093/nar/gki158.

26. Yan J, Marr TG. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Research*. 2005;15(3):369-375. doi: 10.1101/gr.3109605.

27. Ji Z, Tian B. Reprogramming of 3â€² untranslated regions of mRNAs by alternative polyadenylation in generation of pluripotent stem cells from different cell types. *PLoS ONE*. 2009;4(12):e8419. http://dx.doi.org/10.1371%2Fjournal.pone.0008419.

28. Flavell SW, Kim T, Gray JM, et al. Genome-wide analysis of MEF2 transcriptional program reveals synaptic target genes and neuronal activity-dependent polyadenylation site selection. *Neuron*. 2008;60(6):1022-1038. doi: http://dx.doi.org/10.1016/j.neuron.2008.11.029.

29. Shi Y. Alternative polyadenylation: New insights from global analyses. *RNA*. 2012;18(12):2105-2117. doi: 10.1261/rna.035899.112.

30. Nagalakshmi U, Wang Z, Waern K, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008;320(5881):1344. http://science.sciencemag.org/content/320/5881/1344.abstract.

31. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470-476. http://dx.doi.org/10.1038/nature07509.

32. Ozsolak F, Milos PM. RNA sequencing: Advances, challenges and opportunities. *Nat Rev Genet*. 2011;12(2):87-98. http://dx.doi.org/10.1038/nrg2934.

33. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotech*. 2016;34(5):525-527. http://dx.doi.org/10.1038/nbt.3519.

34. Xia Z, Donehower LA, Cooper TA, et al. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3â€²-UTR landscape across seven tumour types. *Nature Communications*. 2014;5:5274. http://dx.doi.org/10.1038/ncomms6274.

35. Wang W, Wei Z, Li H. A change-point model for identifying 3â€²UTR switching by next-generation RNA sequencing. *Bioinformatics*. 2014;30(15):2162-2170. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103598/. doi: 10.1093/bioinformatics/btu189.

36. Grassi E, Mariella E, Lembo A, Molineris I, Provero P. Roar: Detecting alternative polyadenylation with standard mRNA sequencing libraries. *BMC Bioinformatics*. 2016;17(1):423. http://dx.doi.org/10.1186/s12859-016-1254-8. doi: 10.1186/s12859-016-1254-8.

37. Le Pera L, Mazzapioda M, Tramontano A. 3USS: A web server for detecting alternative 3â€²UTRs from RNA-seq experiments. *Bioinformatics*. 2015;31(11):1845-1847. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4443675/. doi: 10.1093/bioinformatics/btv035.

38. Shenker S, Miura P, Sanfilippo P, Lai EC. IsoSCM: Improved and alternative 3' UTR annotation using multiple change-point inference. *RNA*. 2014. doi: 10.1261/rna.046037.114.

39. Birol I, Raymond A, Chiu R, et al. KLEAT: CLEAVAGE SITE ANALYSIS OF TRANSCRIPTOMES(). *Pacific Symposium on Biocomputing.Pacific Symposium on Biocomputing*. 2015:347-358. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350765/.

40. Kim M, You B, Nam J. Global estimation of the 3′ untranslated region landscape using RNA sequencing. *Methods*. 2015;83:111-117. doi: http://dx.doi.org/10.1016/j.ymeth.2015.04.011.

41. Mangone M, Manoharan AP, Thierry-Mieg D, et al. The landscape of *C. elegans* 3â€²UTRs. *Science*. 2010;329(5990):432. http://science.sciencemag.org/content/329/5990/432.abstract.

42. Fox-Walsh K, Davis-Turak J, Zhou Y, Li H, Fu X. A multiplex RNA-seq strategy to profile poly(A(+)) RNA: Application to analysis of transcription response and 3â€² end formation. *Genomics*. 2011;98(4):266-271. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3160523/. doi: 10.1016/j.ygeno.2011.04.003.

43. Shepard PJ, Choi E, Lu J, Flanagan LA, Hertel KJ, Shi Y. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-seq. *RNA*. 2011;17(4):761-772. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3062186/. doi: 10.1261/rna.2581711.

44. Martin G, Gruber A, 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 Reports*. 2012;1(6):753-763. doi: http://dx.doi.org/10.1016/j.celrep.2012.05.003.

45. Beck AH, Weng Z, Witten DM, et al. 3â€²-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLOS ONE*. 2010;5(1):e8768. http://dx.doi.org/10.1371%2Fjournal.pone.0008768.

46. Fu Y, Sun Y, Li Y, et al. Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Research*. 2011;21(5):741-747. doi: 10.1101/gr.115295.110.

47. Ozsolak F, Platt AR, Jones DR, et al. Direct RNA sequencing. *Nature*. 2009;461(7265):814-818. http://dx.doi.org/10.1038/nature08390.

48. Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of caenorhabditis elegans 3prime]UTRs. *Nature*. 2011;469(7328):97-101. http://dx.doi.org/10.1038/nature09616.

49. Hafner M, Renwick N, Brown M, et al. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA*. 2011;17(9):1697-1712. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3162335/. doi: 10.1261/rna.2799511.

50. Steijger T, Abril JF, Engstrom PG, et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat Meth*. 2013;10(12):1177-1184. http://dx.doi.org/10.1038/nmeth.2714.

51. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*. 2015;13(5):278-289. doi: http://dx.doi.org/10.1016/j.gpb.2015.08.002.

52. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. *Nat Biotech*. 2013;31(11):1009-1014. http://dx.doi.org/10.1038/nbt.2705.

53. O'Grady T, Wang X, HÃ¶nerÂ zuÂ Bentrup K, Baddoo M, Concha M, Flemington EK. Global transcript structure resolution of high gene density genomes through multi-platform data integration. *Nucleic Acids Res*. 2016;44(18):e145-e145. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5062983/. doi: 10.1093/nar/gkw629.

54. Chen L, Kostadima M, Martens JHA, et al. Transcriptional diversity during lineage commitment of human blood progenitors. *Science*. 2014;345(6204). http://science.sciencemag.org/content/345/6204/1251033.abstract.

55. Singh N, Sahu DK, Chowdhry R, et al. IsoSeq analysis and functional annotation of the infratentorial ependymoma tumor tissue on PacBio RSII platform. *Meta Gene*. 2015;7:70-75. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707247/. doi: 10.1016/j.mgene.2015.11.004.

56. Abdel-Ghany S, Hamilton M, Jacobi JL, et al. A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*. 2016;7:11706. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4931028/. doi: 10.1038/ncomms11706.

57. Au KF, Underwood JG, Lee L, Wong WH. Improving PacBio long read accuracy by short read alignment. *PLOS ONE*. 2012;7(10):e46679. http://dx.doi.org/10.1371%2Fjournal.pone.0046679.

58. Au KF, Sebastiano V, Afshar PT, et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proceedings of the National Academy of Sciences*. 2013;110(50):E4821-E4830. doi: 10.1073/pnas.1320101110.

59. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3â€™ untranslated regions and fewer MicroRNA target sites. *Science*. 2008;320. http://dx.doi.org/10.1126/science.1155390. doi: 10.1126/science.1155390.

60. Graham RR, Kyogoku C, Sigurdsson S, et al. Three functional variants of IFN regulatory factor 5 (IRF5) define risk and protective haplotypes for human lupus. *Proceedings of the National Academy of Sciences*. 2007;104(16):6758-6763. doi: 10.1073/pnas.0701266104.

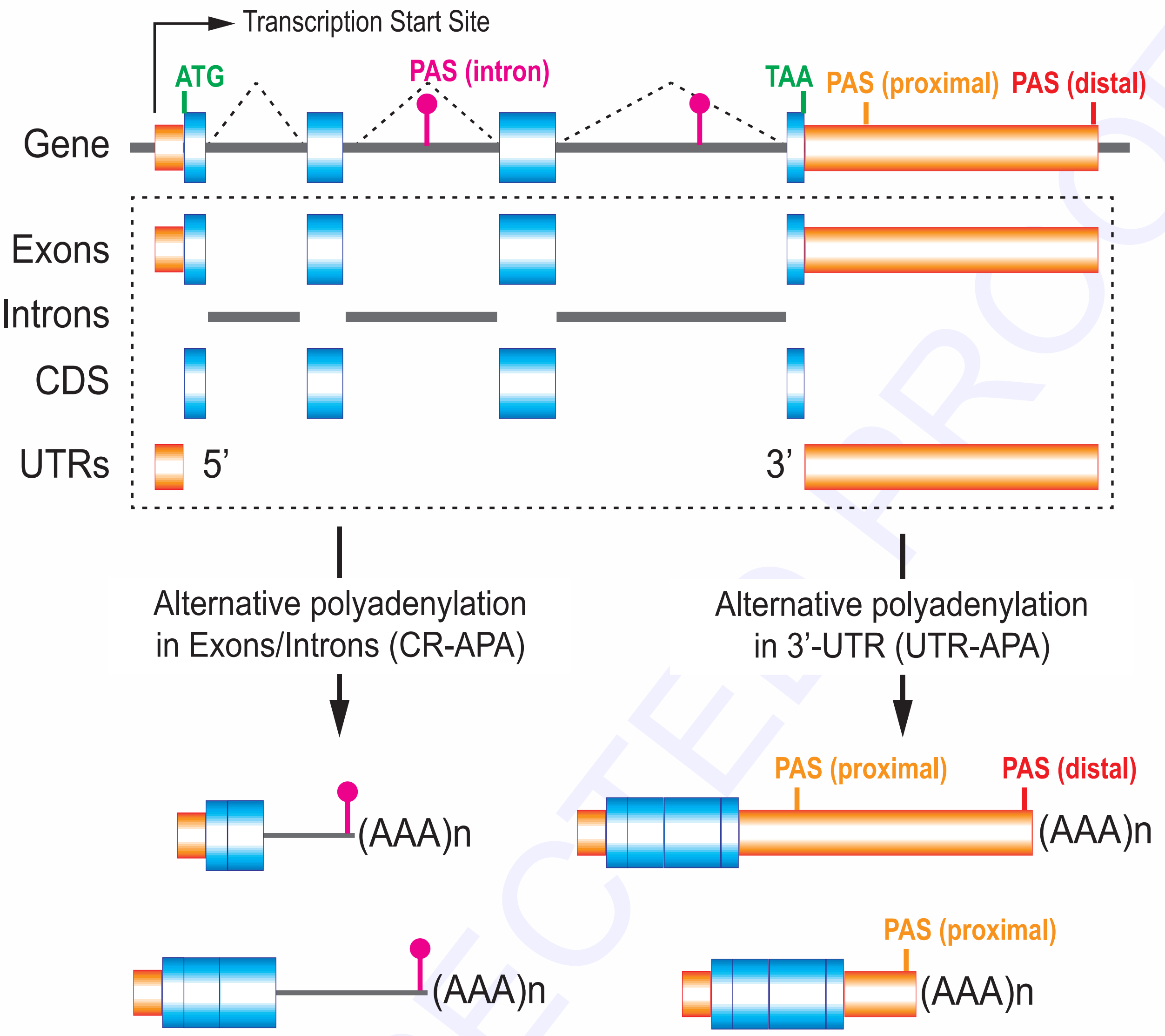
61. Weirather JL, Afshar PT, Clark TA, et al. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res*. 2015;43(18):e116-e116. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4605286/. doi: 10.1093/nar/gkv562.

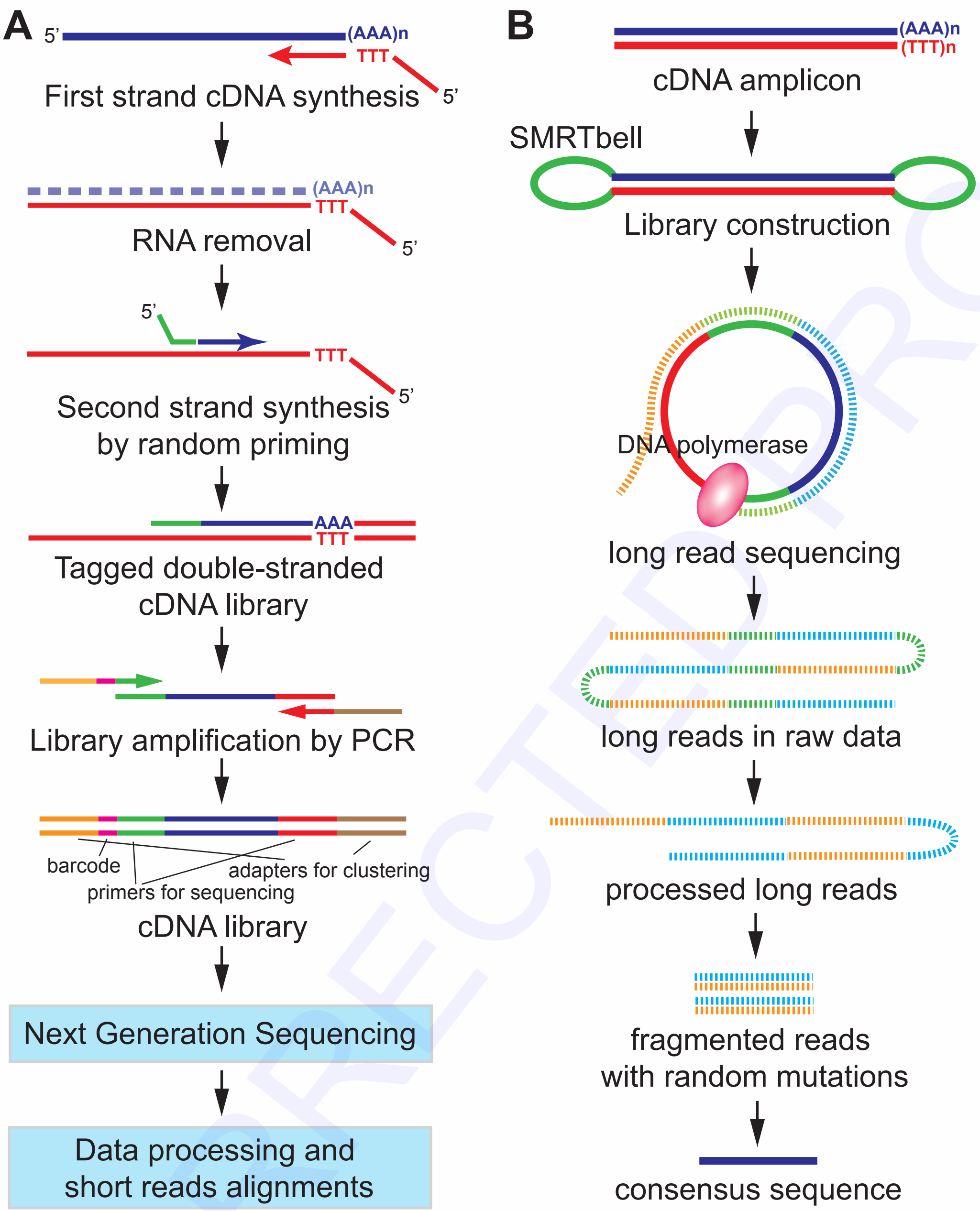
62. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*. 2010;28(5):511-515. http://dx.doi.org/10.1038/nbt.1621.

Figure Legends

Figure 1. Schematic showing a gene structure and alternative polyadenylation. A gene is composed of exons and introns. Exons of a gene divide into coding DNA sequence (CDS) and untranslated regions (UTRs). Alternative polyadenylation can occur within the last exon of a gene (UTR-APA) and/or upstream exons/introns (CR-APA).

Figure 2. A work flow for 3’end-seq and PacBio Iso-seq. (A) Multiple versions of global profiling method for 3’-end sequence of mRNAs are available. An example of 3’end-seq method is shown. A 3’end-seq cDNA library is produced by a series of molecular biology work integrating first strand cDNA synthesis and PCR. Short sequence reads of polyadenylation site are cataloged by conducting RNA-seq using the cDNA library and trimming/aligning sequencing data. (B) The SMRT bell cDNA library for PacBio-seq can be produced from cDNA amplicon which is made by reverse transcription. Concatemerized long reads of insert in SMRT bell cDNA library are produced as raw data. Processed long reads (by eliminating SMRT bell sequences) are aligned to generate a consensus sequence of long reads.


Transcription Start Site
ATG
PAS (intron)
TAA
PAS (proximal)
PAS (distal)
Gene
Exons
Introns
CDS
UTRs
5'
3'
Alternative polyadenylation
in Exons/Introns (CR-APA)
Alternative polyadenylation
in 3'-UTR (UTR-APA)
(AAA)n
PAS (proximal)
PAS (distal)
(AAA)n
(AAA)n
PAS (proximal)
(AAA)n


A
5'
(AAA)n
TTT
5'
First strand cDNA synthesis
(AAA)n
TTT
5'
RNA removal
5'
TTT
5'
Second strand synthesis
by random priming
AAA
TTT
Tagged double-stranded
cDNA library
Library amplification by PCR
barcode
adapters for clustering
primers for sequencing
cDNA library
Next Generation Sequencing
Data processing and
short reads alignments
B
(AAA)n
(TTT)n
cDNA amplicon
SMRTbell
Library construction
DNA polymerase
long read sequencing
long reads in raw data
processed long reads
fragmented reads
with random mutations
consensus sequence