

Spring 5-31-2013

Polyaseeker: a computational framework for identifying polyadenylation cleavage site from RNA-seq

Xiao Ling
New Jersey Institute of Technology

Follow this and additional works at: <https://digitalcommons.njit.edu/theses>



Part of the [Bioinformatics Commons](#), and the [Computer Sciences Commons](#)

Recommended Citation

Ling, Xiao, "Polyaseeker: a computational framework for identifying polyadenylation cleavage site from RNA-seq" (2013). *Theses*. 169.
<https://digitalcommons.njit.edu/theses/169>

This Thesis is brought to you for free and open access by the Electronic Theses and Dissertations at Digital Commons @ NJIT. It has been accepted for inclusion in Theses by an authorized administrator of Digital Commons @ NJIT. For more information, please contact digitalcommons@njit.edu.

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use” that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select “Pages from: first page # to: last page #” on the print dialog screen

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

POLYASEEKER: A COMPUTATIONAL FRAMEWORK FOR IDENTIFYING POLYADENYLATION CLEAVAGE SITE FROM RNA-SEQ

**by
Xiao Ling**

Alternative polyadenylation (APA) of mRNA plays a crucial role for post-transcriptional gene regulation. Recently, advances in next generation sequencing technology have made it possible to efficiently characterize the transcriptome and identify the 3' end of polyadenylated RNAs. However, no comprehensive bioinformatic pipelines have fulfilled this goal. The PolyASeeker, a computational framework for identifying polyadenylation cleavage sites from RNA-Seq data is proposed in this thesis. By using the simulated RNA-seq dataset, a novel method is developed to evaluate the performance of the proposed framework versus the traditional A-stretch approach, and compute accurate Precisions and Recalls that previous estimation could not get. It is found that the proposed method is able to achieve significantly higher sensitivity in various scenarios than the A-stretch approach. In further studies, PolyASeeker is applied to human tissue-specific RNA-sequencing data, and through all the polyA sites identified by PolyASeeker and annotated by PolyA DB, special isoform expression patterns among tissues are found. Genes that have a specific 3'UTR expression have also been recognized in the brain. PolyASeeker is also run on an mRNA 3' UTR sequencing dataset and it is found that the software could be quite adapted to the data. Significant isoform shorting events with expression evidences and experimental supports have been found.

**POLYASEEKER: A COMPUTATIONAL FRAMEWORK FOR IDENTIFYING
POLYADENYLATION CLEAVAGE SITE FROM RNA-SEQ**

**by
Xiao Ling**

**A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics**

Department of Computer Science

May 2013

Copyright © 2013 by Xiao Ling

ALL RIGHTS RESERVED

APPROVAL PAGE

**POLYASEEKER: A COMPUTATIONAL FRAMEWORK FOR IDENTIFYING
POLYADENYLATION CLEAVAGE SITE FROM RNA-SEQ**

Xiao Ling

Dr. Zhi Wei, Thesis Advisor	Date
Associate Professor of Computer Science, NJIT	

Dr. Usman Roshan, Committee Member	Date
Associate Professor of Computer Science, NJIT	

Dr. Jason Wang, Committee Member	Date
Professor of Bioinformatics and Computer Science, NJIT	

BIOGRAPHICAL SKETCH

Author: Xiao Ling

Degree: Master of Science

Date: May 2013

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
New Jersey Institute of Technology, NJ, US, 2013
- Bachelor of Science in Bioinformatics,
Huazhong University of Science and Technology, Wuhan, P. R. China, 2011

Major: Bioinformatics

ACKNOWLEDGMENT

Here I want to thank the thesis advisor Zhi Wei, who supervised the progress of the thesis and direct our goal during the study. Also I want to thank other committee members, Usman Roshan and Jason Wang, for offering suggestions toward my thesis. Besides, I am very grateful to Wei Wang, who helped in technical problems and improving the software. At last, I want to thank my parents and Yidong for supporting in my studies; their support gives me the courage to pursue my career goal and overcome the difficulties.

TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION.....	1
1.1 Objective.....	1
1.2 Background Information.....	2
2 POLYADENYLATION SITES IDENTIFICATION	3
2.1 Problem Statement	3
2.2 Approach	3
3 IMPLEMENTATION	5
3.1 Framework of PolyASeeker.....	5
3.1.1 Algorithm for Identifying PolyA Sites	5
3.1.2 Pipeline.....	6
3.2 RNA-seq Dataset Simulation.....	7
4 RESULTS.....	8
4.1 Results of Simulation Study	8
4.2 Results of Real-data Study	10
4.2.1 Human Body Map 2.0 RNA-seq Data	11
4.2.2 Fu 2011 RNA- sequencing Data.....	16
5 CONCLUSIONS	18
REFERENCES	19

LIST OF TABLES

Table	Page
4.1 Genomic locations of polyA sites from Human Body Map 2.0 RNA-seq data.....	11
4.2 Characteristics of polyA sites from Fu 2011 RNA-seq data.....	16

LIST OF FIGURES

Figure	Page
3.1 PolyASeeker pipeline.....	6
4.1 Performance comparison of PolyASeeker vs. A-stretch method in simulation study.....	9
4.2 Sensitivity accumulation under sequencing depth and expressing level.....	10
4.3 Expression characteristics of TPM1, TPM2 and TPM3 difference between tissues.....	14
4.4 Expression characteristics of CDC42 and MAP4 difference in brain.....	15
4.5 Gene expression evidences for 3'UTR switching events identified in Fu's cancer cell dataset	18

CHAPTER 1

INTRODUCTION

1.1 Objective

The objective of this thesis is to develop a computational framework for identifying polyadenylation cleavage sites from RNA-Seq data. The name of this framework is PolyASeeker; it could be downloaded at <http://polyaseeker.sourceforge.net/>. In order to evaluate the performance of this software, PolyASeeker are tested in simulation and real-data studies.

For the simulation test, a simulation tool, FluxSimulator was used to simulate RNA-seq data; it simulated the testing data under varied sequencing scenarios. And a set of Precision and Recall values are calculated based on the performance of PolyASeeker and a traditional 8-A method, the results show that PolyASeeker is superior to the traditional method in Recall values.

For the real data test, two published data RNA-seq datasets are used to evaluate performance. One dataset consist of human tissue-specific RNA-seq data, and the other one include RNA-seq data from two human breast cancer cell samples and one mammary epithelial cell sample. The goal here is to study whether the identified polyA sites could be used to found alternative polyadenylation event, and furthermore whether it could lead to novel biological recoveries.

1.1 Background Information

Alternative polyadenylation (APA) of messenger RNA plays a key role during post-transcription and is a widespread mechanism in higher eukaryotes. The usage of alternative PolyA sites could lead to encode multiple mRNA transcripts for a single gene. In recent years, it has become increasingly evident that the length changes of 3'UTR are versatile in various physiological states and cell types, such as tumor cells, activated T lymphocytes and embryonic cells. Despite of progresses, these studies merely utilized known annotations, for example, PolyA_DB, which is based on series of available database of cDNA/EST, but it reveals incompleteness due to the limitation of sequencing technology.

Recently, the advances of next generation sequencing technology have merged as a powerful tool to interrogate of the transcriptome and provide an opportunity to investigate polyadenylation cleavage sites on an unprecedented scale. A tradition method of identifying potential novel polyadenylation sites by searching and remapping at least four As or Ts among those unmapped RNA-seq reads is commonly used since EST data. However, these simple A-stretch approaches fail to take consideration for sequencing error, which sometimes could lead to false positives for predictions.

CHAPTER 2

POLYADENYLATION SITES IDENTIFICATION

2.1 Problem Statement

In order to correctly identify polyadenylation sites from RNA-seq data, the software is aimed to deal with oligo(dT)/modified oligo(dT)-primed mRNA sequencing data, and it also suppose the input data to have reads that origin from the boundary of 3'UTR and polyA tails. Once the basic requirement is satisfied, the problem have becomes that how to identify those true polyA reads, and how to filter out the false positives from the result. It is also necessary to show the possible applications of polyA sites identified by PolyASeeker.

2.2 Approach

Based on the problems mentioned above, PolyASeeker, a computational framework is designed to align the sequences back to the genome and detect the polyA reads. It used a scoring method to identify the polyA regions. In order to study the performance of this proposed method, a simulated RNA-seq dataset is generated and used to test PolyASeeker. As the true polyA sites are known in simulated dataset, the precision and recall value are measurable in this study. By measuring these values, the best filtering parameter settings are decided. The PolyASeeker is also proved to perform better than a traditional method. Furthermore, two real data applications have been conducted. The goals here is to reveal that the proposed framework could be feasible to analyze most of RNA-seq data; also to prove that PolyASeeker is able to identify significant different

isoform expressions from tissue-specific RNA-seq data, and could find significant 3'UTR shorting event from the cancer cell samples.

CHAPTER 3

IMPLEMENTATION

As is mentioned in Chapter 2, the PolyASeeker is designed for achieving higher accuracy and sensitivity. Also, If the RNA-seq reads is pair-end, then the proposed method should be able to take the fully use of pair information. Therefore the pipeline in PolyASeeker was created to fulfill these demands and become powerful and suitable in analyzing general RNA-seq data.

3.1 Framework of PloyASeeker

Traditionally, PolyA candidate reads in the data set are detected base on their number of As or Ts. Normally, once a read has 8 or more of that, then it is considered contain a polyA tail, However, this method ignores the sequencing quality of each position, which did not take the possibility that the Adenine comes from sequencing error into consideration. Thus, a new evaluation method is developed in PolyASeeker.

3.1.1 Algorithm for identifying PolyA sites

First, in order to detect A-rich regions, different weights are given to discriminate between A and C, G, T in the following manner:

$$S(base) = \begin{cases} 1, & \text{if } base = A \\ -1, & \text{if } base = \bar{A} \end{cases} \quad (3.1)$$

Then, the expectation for each base can be computed by taking account for the sequencing error ε :

$$E(A) = (1 - \varepsilon) \times S(A) + \frac{\varepsilon}{3} \times S(\bar{A}) \quad (3.2)$$

$$E(\bar{A}) = (1 - \varepsilon) \times S(\bar{A}) + \frac{\varepsilon}{3} \times S(A) \quad (3.3)$$

Where ε can be obtained from Phred-scaled base quality score in FASTQ format,
 $\varepsilon = 10^{(Q/-10)}$

Let L be the length of unmapped region, for each aligned reads, the unmapped region can be scored by the summation of expectation for every base in L:

$$Score = \sum_{j=1}^L \pi_j \times E_j(A) + (1 - \pi_j) \times E_j(\bar{A}) \quad (3.4)$$

Where π_j is an indicator if base is A or not,

$$\pi_j = \begin{cases} 1, & \text{if } base = A \\ 0, & \text{if } base = \bar{A} \end{cases} \quad (3.5)$$

3.1.2 Pipeline

The PolyASeeker Pipeline mainly contains three steps: reads alignment, poly-A candidate sites identifying and sites filtering, this pipeline is designed under the consideration of computing efficiency, simplicity and the Recall and Precision values.



Figure 3.1 PolyASeeker pipeline

Figure 3.1 shows the procedures of how PolyASeeker identify polyA reads. First, the short reads in fastq format can be directly mapped by Bowtie2 local model, which is an ultrafast and memory-efficient tool for alignment and supports local alignment that does not require reads to align end-to-end.

Next, for each alignment, the unmapped region can be score by equation (3) and parse the candidate reads above a certain cutoff. By default, PolyASeeker set cutoff=7.8 as default which is approximately equivalent to 8A-strech method.

Finally, PolyASeeker filters false positives and reorganized the result. It contains an internal priming filter and sites clustering function, together with the optional filters including expression filter and supportive reads filter.

3.2 RNA-seq Dataset Simulation

The PolyASeeker was tested on both simulated RNA-seq data and real RNA-seq data. For the simulated data, a recently published RNA-Seq simulation tool, FluxSimulator was used to study the performance of the proposed method. Refseq hg19 gene annotation was used in this simulation. And four scenarios were studied: 100bp paired-end, 75bp paired-end, 100bp single-end, 75bp single-end. For 75bp paired-end and single-end data, FluxSimulator build-in error model is used, while for 100 bp paired-end and single-end data, a custom error model is created. The model of polyadenylation process in FluxSimulator was generated by a Weibull-approximation of the normal distribution with shape=2 and scale=300 to sample random lengths of PolyA tails. Poly-dT priming RNA-seq procedure was performed and different reads depth was also sequenced, including 1M 10M, 25M, 50M, 75M, 100M, 125M, 150M, 175M and 200M for all the four scenarios.

CHAPTER 4

RESULT

4.1 Results of simulation study

In the simulation study, the testing data are all generated by FluxSimulator. The model of polyadenylation process in FluxSimulator was generated by a Weibull-approximation of the normal distribution with shape=2 and scale=300 to sample random lengths of PolyA tails during transcription. Poly-dT priming RNA-seq procedure was performed under four scenarios: 100bp paired-end, 75bp paired-end, 100bp single-end, 75bp single-end. Different sequenced reads depths are also simulated under these four types, ranges from from 1M to 200M. The performance was measured using the following equations:

$$Precision = \frac{TP}{(TP + FP)} \quad (4.1)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (4.2)$$

Figure 4.1 shows that PloyASeeker achieved a much higher Precision in all scenarios than traditional A-stretch approach.

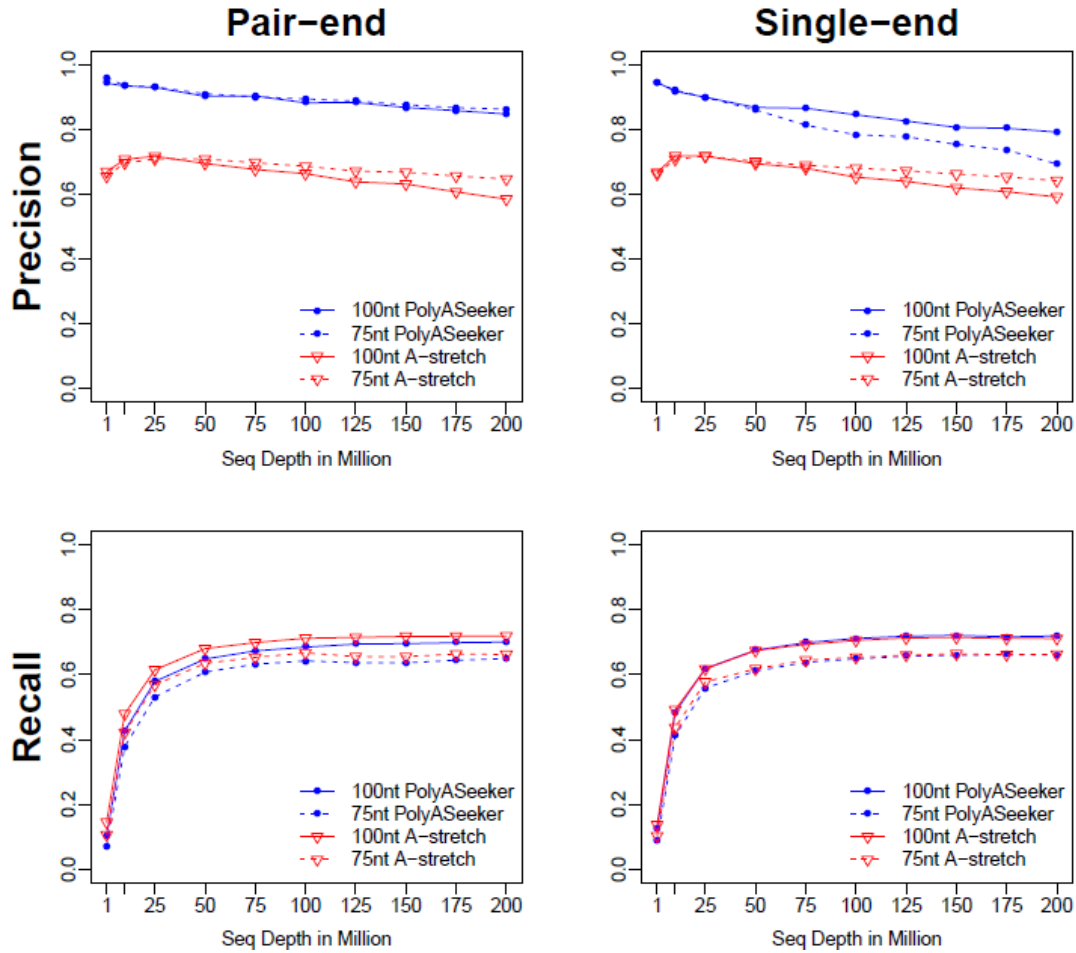


Figure 4.1 Performance comparison of PolyASeeker vs. A-stretch method in simulation study. Two statistics values are shown on each y-axis. Precision denotes the rate of True positive in prediction group. Recall denotes the rate of successful identified polyA sites from the ture polyA sites. Compared to the A-stretch method, the PolyASeeker have identified the equivalent amount of PolyA sites with a more accurate performance.

As to the Recalls, it is observed that these values has a fundamental correlation with the sequencing depth, as was shown in Figure 1, these values would reached climaxes at the Seq-depth ranging from 175 million to 200 million. In order to study the minimum transcription level for a gene to have its polyA sites identified, each simulated transcriptome is split by the genes expressed molecules number. It is found that for the 200 million dataset, above 90% of the genes would have their polyA sites identified if those genes have more than 40 transcript mRNA molecules, as shown in Figure 4.2.

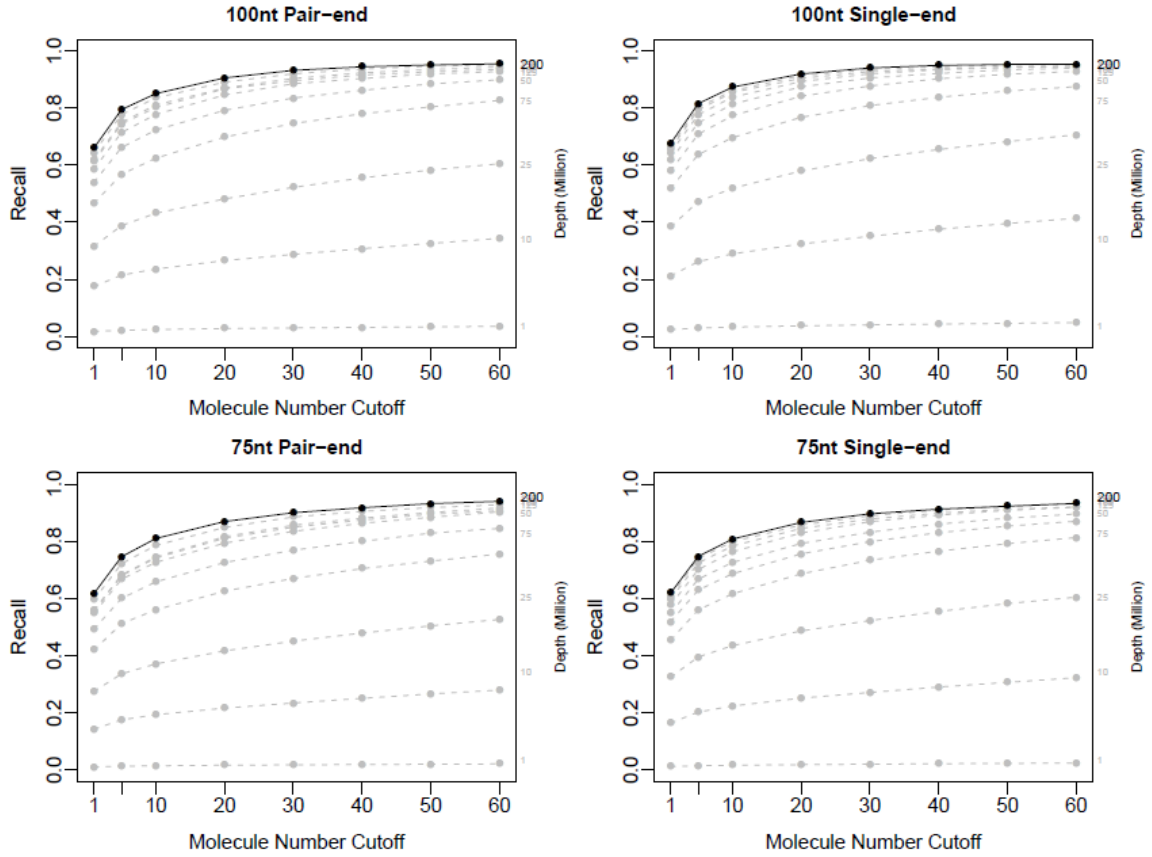


Figure 4.2 Sensitivity accumulation under sequencing depth and expressing level. The growth of sensitivity is shown in four scenarios as the sequencing depth accumulating from 1M(million), 10M, 25M, 50M, 75M, 100M, 125M, 150M, 175M (grey lines) to 200 million (black line). The X-axis denotes what cutoff is used to define the true polyA sites: only those isoforms with a higher transcription numbers than the cutoff will be expected to have a true polyA site at the end.

4.2 Results of Real-data Study

The performance of PolyASeeker is assessed using two real NGS RNA-Seq datasets. The first one is the Illumina bodyMap2 RNA sequencing data, the transcription profiling of individual and mixture of 16 human tissues RNA. The 16 human tissue types includes adrenal, adipose, brain, breast, colon, heart, kidney, liver, lung, lymph, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells. In the analysis PolyASeeker is

applied to the 50 bps paired end data, which with one lane of HiSeq 2000 data per tissue. The second dataset is from a published data set (9) for studying tandem 3'UTR switching in human breast cancer cells. A novel strategy of sequencing APA sites, called SAPAS, was introduced with modified oligo(dT) tags. PolyA reads were reversely sequenced and begin with the linker 5'-TTTTCTTTTTTCTTTTTT-3'. PolyASeeker was applied to two breast cancer cell lines (MCF7 and MB231) in addition to a human normal mammary epithelial cell line (MCF10A) from Illumina GA IIx sequencing platform.

4.2.1 Human Body Map 2.0 RNA-seq data

The Human Body Map 2.0 Project by Illumina generated RNA-seq data for 16 different human tissues (adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph node, ovary, prostate, skeletal muscle, testes, thyroid, and white blood cells). The 50 bps paired end read data is used for analyzing. For each lane of HiSeq 2000 data per tissue, a certain amount of PolyA sites is identified (Table 4.1). And altogether an amount of 12,431 polyA sites is got, for this combined sites, 60% of them are outside from the polyA_DB (Tian. et.al^[18]).

Table 4.1 Genomic locations of polyA sites from Human Body Map 2.0 RNA-seq data

Genome Regions	Number of sites	Distribution percentage
3'UTR	12,703	56%
5'UTR	631	3%
CDS	613	3%
Intron	3,482	15%
Intergenic	6,090	27%
PolyA_DB	9,729	43%

To further explore the application of PolyASeeker in this dataset, a tissue-specific alternative polyadenylation study is conducted. First a 3' UTR annotation is prepared from RefSeq Genes, which contains a combined 3' UTR region from every isoform for each gene. Based on this annotation, the distal and the proximal polyA sites are retrieved from the combined pool of polyASeeker result and polyA DB. By counting an expression ratio of the coverage number at the 200bp window in distal polyA site to the coverage number at 200bp window in proximal polyA site, ratio variance are calculated for all tissue's expression ratio at each 3' UTR region. For those 3'UTRs that contain one more isoforms, their ratio variance are ranked as the value marks the isoform usage variance among different tissues. From the ranking list, several genes are recognized to have different isoform expressions in 3' UTR region among tissues.

TPM1, TPM2 and TPM3, which have been found to have different polyadenylation patterns between the tissues of heart, breast, thyroid, skeletal muscle and the tissues of prostate, ovary, testes, colon, adrenal, adipose, lung, kidney and lymph node, and have low expression in the brain and white blood cells. And the first group trend to express a shorter isoform, where the second group trend to express the longer isoform. Since these genes encode the tropomyosin family of actin-binding proteins which involved in the contractile system of striated and smooth muscles and the cytoskeleton of non-muscle cells, a shorting pattern in skeleton muscles and heart may be reasonable to these tissues, as these two is enriched with striated and smooth muscles, however the similarity shortened pattern in breast, thyroid remains unknown (Figure 4.3).

CDC42 and MAP4, which are ranked in the top 10 list of isoform usage variance between tissues, are found to be both brain-specific genes (Figure 4.4). CDC42, which

have three transcript variants, expresses the short transcript variant 2 in and only in the brain tissue. The case is also the same in gene MAP4, as this gene also has three transcript variant, and the shorter transcript variant 3 is only expressed in the brain tissue have. CDC42 encodes is a small GTPase of the Rho-subfamily, which regulates signaling pathways that control diverse cellular functions including cell morphology, migration, endocytosis and cell cycle progression. MAP4 encoded a major non-neuronal microtubule-associated protein, and the phosphorylation of this protein affects microtubule properties and cell cycle progression. Since the brain tissue composed primarily of neurons and glial cells, which have specific cell structure and long life span, it is not surprising that it has special expression patterns contains in cell cycle related genes. However, how those short transcript variant regulates the morphology and cycle progression in neurons and glial cells remains unclear.

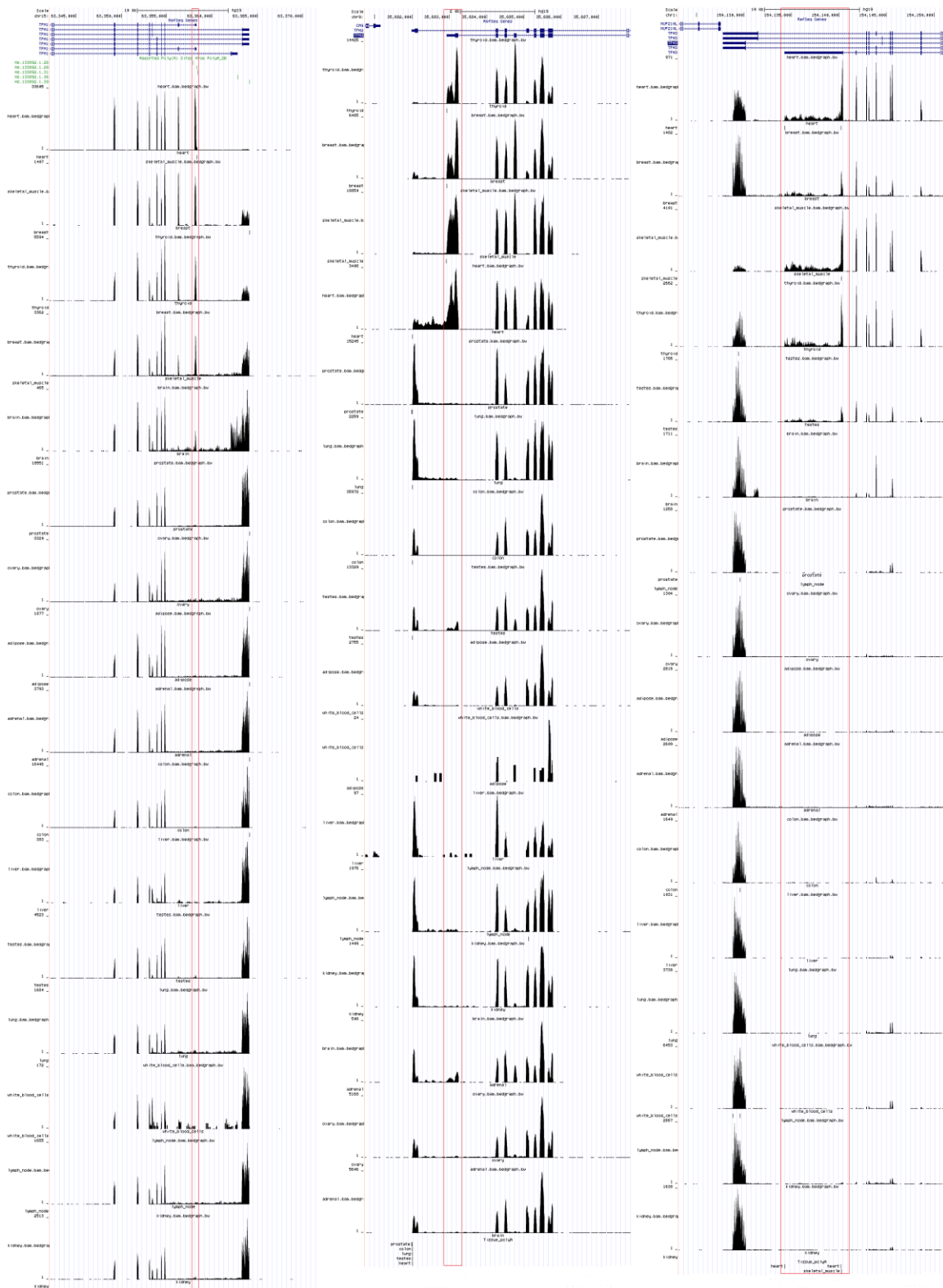


Figure 4.3 Expression characteristics of TPM1, TPM2 and TPM3 difference between tissues.

Source: <http://genome.ucsc.edu/>, accessed March 28, 2013.

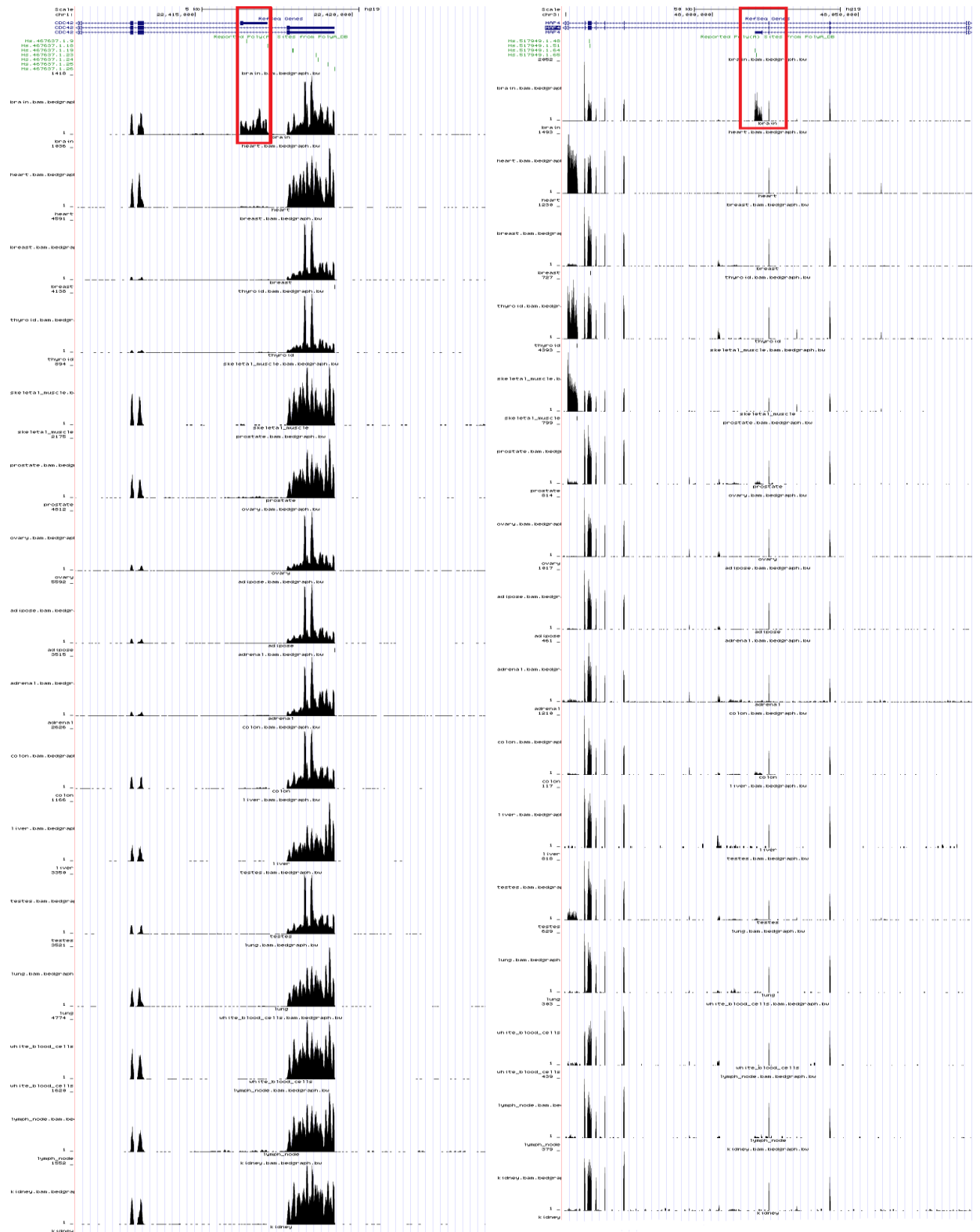


Figure 4.4 Expression characteristics of CDC42 and MAP4 difference in brain.
Source: <http://genome.ucsc.edu/>, accessed April 2, 2013.

4.2.2 Fu 2011 RNA- seq data

To further explore the availability and usage of PolyASeeker, the software is applied to another RNA-seq dataset. This dataset is generated by a 3' ends mRNA sequencing method called SAPAS ^[9]. This method is aimed to complete genome-wide profiling of APA sites and to identify new polyA sites, and its reads covers from an upstream 200-300 bp length window to the 3' end of mRNA, where end with the polyA tail. The samples here include two human breast cancer lines (MCF7 and MB231) and one cultured human mammary epithelial normal cell line (MCF10A). As they use a traditional method to identify polyA sites, the PolyAseeker is employed in this job. The result shows that an equivalent and in some samples slightly more amount of polyA sites have been identified (Table 4.2). The four numbers of supportive reads at the proximal and distal polyA sites are gathered among cancer sample and control sample, and a fisher test is run on these four numbers. For the gene that has significant p-values, it is believed to have a 3'UTR switching event between the corresponding samples. From the result, several experimental verified APA events that have been reported by previous studies have been found (seven of eight genes).

Table 4.2 Characteristics of polyA sites from Fu 2011 RNA-seq data

	Raw reads	Mapped reads	Uniquely mapped to genome (SAPAS)	Best mapped reads with a polyA tail (PolyASeeker)
Combined	31,026,769	-	13,573,367	16,139,436
MCF10A (mammary epithelial cell)	8,319,588	98.57%	4,254,699	5,097,807
MCF7 (breast cancer)	6,755,371	96.66%	3,449,838	4,101,303
MDA231 (breast cancer)	15,951,810	90.44%	5,868,830	6,940,326

	PolyA sites (SAPAS)	PolyA sites (PolyASeeker)	Running time (min)	Known polyA sites :
Combined	89,211	89,475	245	19,683
MCF10A (mammary epithelial cell)	39,246	39,391	69	Novel polyA sites:
MCF7 (breast cancer)	41,184	39,408	50	
MDA231 (breast cancer)	61,812	62,515	126	

DDX5, HSBP1, FAM134A and SEC61A1, which had been identified to have significant 3'UTR switching events in MCF7, all are validated by RT-PCR in previous studies (Fu 2011). For gene, RAB10, ANP32A, DDX5 and RRBP1, which have been found to have significant 3'UTR switching events in MB231, are also validated by qRT-PCR in previous studies (Fu 2011) (Figure 4.5). For the other predicted 3'UTR switching events, some of the result are novel from Fu's study. Since the expression evidences for those sites have been found (Figure 4.6), the results are highly convincing. And their lost may due to the different choice in sequence aligner (Bowtie in Fu, Bowtie2 in PolyASeeker).

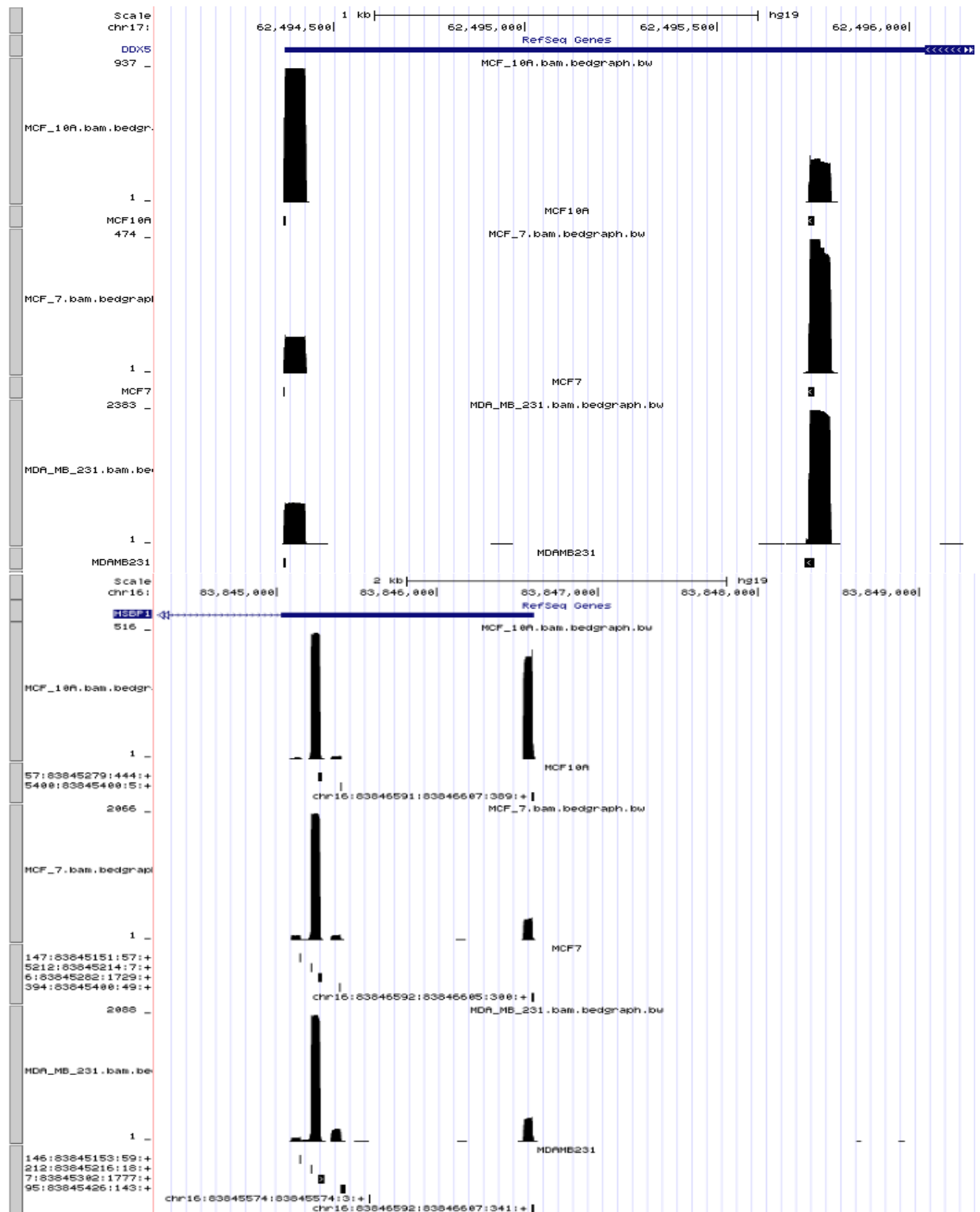


Figure 4.5 Gene expression evidences for 3'UTR switching events identified in Fu's cancer dataset.

Source: <http://genome.ucsc.edu/>, accessed April 29, 2013.

CHAPTER 5

CONCLUSIONS

In this study, the PolyAseeker, a computational framework for identifying polyadenylation cleavage sites in RNA-seq data is proposed. In the PolyAseeker pipelines, an adenine scoring method is applied in polyA reads identification. And several processing, filtering methods have been modified to achieve a higher performance. The Precision and Recall of all the results from simulated dataset is extensively evaluated, PolyASeeker method has consistently reached 0.2 percent or higher in Precision than the 8A-stretch method.

On two real datasets, it has been demonstrated that the method works efficiently and precisely in analysis of RNA-Seq data. And in tissue-specific dataset, genes with different polyadenylation patterns have been identified between tissues. Those genes have created novel topics in studying gene functions and cellular differentiation. In Fu's cancer dataset, genes with significant 3'UTR shortening events have been identified in cancers, leading to a novel potential class of biomarkers and candidates to explain the cancer etiology. Since the applications of PolyASeeker to specific samples have been proved to reveal novel aberrant PolyA sites usage, it is expected that the knowledge to alternative polyadenylation and gene regulation mechanisms will be greatly facilitated as more and more RNA-Seq data become available.

REFERENCES

1. Colgan, D.F. and Manley, J.L. (1997) Mechanism and regulation of mRNA polyadenylation. *Genes & Development*, 11, 2755-2766.
2. Tian, B., Hu, J., Zhang, H. and Lutz, C.S. (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res*, 33, 201-212.
3. Di Giammartino, D.C., Nishida, K. and Manley, J.L. (2011) Mechanisms and consequences of alternative polyadenylation. *Molecular Cell*, 43, 853-866.
4. Mayr, C. and Bartel, D.P. (2009) Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138, 673-684.
5. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. and Burge, C.B. (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*, 320, 1643-1647.
6. Ji, Z., Lee, J.Y., Pan, Z., Jiang, B. and Tian, B. (2009) Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 7028-7033.
7. Lee, J.Y., Yeh, I., Park, J.Y. and Tian, B. (2007) PolyA_DB 2: mRNA polyadenylation sites in vertebrate genes. *Nucleic Acids Research*, 35, D165-168.
8. Jan, C.H., Friedman, R.C., Ruby, J.G. and Bartel, D.P. (2011) Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*, 469, 97-101.

9. Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C. and Xu, A. (2011) Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Research*, 21, 741-747.
10. Derti, A., Garrett-Engle, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M. and Babak, T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Research*, 22, 1173-1183.
11. Shepard, P.J., Choi, E.A., Lu, J., Flanagan, L.A., Hertel, K.J. and Shi, Y. (2011) Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17, 761-772.
12. Oszolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B. and Milos, P.M. (2010) Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, 143, 1018-1029.
13. Lin, Y., Li, Z., Oszolak, F., Kim, S.W., Arango-Argoty, G., Liu, T.T., Tenenbaum, S.A., Bailey, T., Monaghan, A.P., Milos, P.M. et al. (2012) An in-depth map of polyadenylation sites in cancer. *Nucleic Acids Research*, 40:8460-71.
14. Sherstnev, A., Duc, C., Cole, C., Zacharaki, V., Hornyik, C., Oszolak, F., Milos, P.M., Barton, G.J. and Simpson, G.G. (2012) Direct sequencing of *Arabidopsis thaliana* RNA reveals patterns of cleavage and polyadenylation. *Nature Structural & Molecular Biology*, 19, 845-852.

15. Elkon, R., Drost, J., van Haaften, G., Jenal, M., Schrier, M., Vrieling, J.A. and Agami, R. (2012) E2F mediates enhanced alternative polyadenylation in proliferation. *Genome Biology*, 13, R59.
16. Jenal, M., Elkon, R., Loayza-Puch, F., van Haaften, G., Kuhn, U., Menzies, F.M., Oude Vrieling, J.A., Bos, A.J., Drost, J., Rooijers, K. et al. (2012) The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell*, 149, 538-553.
17. Beck, A.H., Weng, Z., Witten, D.M., Zhu, S., Foley, J.W., Lacroute, P., Smith, C.L., Tibshirani, R., van de Rijn, M., Sidow, A. et al. (2010) 3'-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PloS One*, 5, e8768.
18. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y. and Pritchard, J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464, 768-772.
19. Smibert, P., Miura, P., Westholm, J.O., Shenker, S., May, G., Duff, M.O., Zhang, D., Eads, B.D., Carlson, J., Brown, J.B. et al. (2012) Global patterns of tissue-specific alternative polyadenylation in *Drosophila*. *Cell Reports*, 1, 277-289.
20. Mangone, M., Manoharan, A.P., Thierry-Mieg, D., Thierry-Mieg, J., Han, T., Mackowiak, S.D., Mis, E., Zegar, C., Gutwein, M.R., Khivansara, V. et al. (2010) The landscape of *C. elegans* 3'UTRs. *Science*, 329, 432-435.
21. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357-359.

22. Bentley, J. (1984) Programming pearls: algorithm design techniques. *Commun. ACM*, 27, 865-873.
23. Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigo, R. and Sammeth, M. (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40, 10073-10083.
24. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28, 511-515.