이학박사학위논문

# 폴리A 꼬리와 RNA-단백질 상호작용에 대한 전사체적 분석

# Transcriptome-wide analysis of poly(A) tail and RNA-protein interaction

2014년 2월

서울대학교 대학원

생명과학부

장혜식

# Transcriptome-wide analysis of poly(A) tail and RNA-protein interaction

Advisor: Professor V. Narry Kim

Submitting a doctoral thesis of phylosophy
October, 2013

Graduate School of Seoul National University
School of Biological Sciences
Hyeshik Chang

Confirming the doctoral thesis written by _____
December, 2013

Chair _____ (seal)

Vice Chair _____ (seal)

Examiner _____ (seal)

Examiner _____ (seal)

Examiner _____ (seal)

# Abstract

# Transcriptome-wide analysis of poly(A) tail and RNA-protein interaction

Hyeshik Chang
School of Biological Sciences
The Graduate School
Seoul National University

RNAs store and transfer information among constituents of the cell. From their biogenesis to processing, transport, translation, catalysis, and decay, many cellular factors are involved to achieve tight regulation. Following the development of high-throughput DNA sequencing, it has become an essential tool to scrutinize RNA molecules in the cell in unprecedented scale and depth. This thesis concerns methodological advances in two aspects of RNA regulation. First, I develop a novel method to survey global status of polyadenylation that takes a fundamentally different approach from the existing techniques. Despite its importance in gene regulation, global investigation of the 3′ extremity of mRNA has not been feasible due to technical challenges associated with homopolymeric sequences and relative paucity of mRNA. The new technique, named as TAIL-seq, allows measuring poly(A) tail length at the genomic scale for the first time. I also discover widespread uridylation and guanylation at the downstream of poly(A) tail. The U-tails are generally attached to short poly(A) tails (<25 nt) while the G-tails are found mainly on longer poly(A) tails (>40 nt), implicating their generic roles in mRNA stability control. Furthermore, TAIL-seq identifies, with a single nucleotide resolution, numerous nucleolytic events involved in microRNA processing and mRNA cleavage. TAIL-seq will enable exploration of unforeseen diversity of RNA processing and modification.

Secondly, I describe an array of new analytic methods to crosslinking, immunoprecipitation, and sequencing (CLIP-seq) to enhance its utility in the investigation of RNA-protein

interactions. CLIP-seq arose as one of the standard techniques to retrieve transcriptome-wide information of RNA-protein interactions in last few years. However, generalized analysis techniques and tools have been missing unlike the other RNA-seq applications. In this study, I generalize analytic workflow for binding site identification by developing new methods. I also provide an open source toolchain that covers most of the common analyses performed for CLIP-seq. In addition, I present *ecliptic*, a fully automated pipeline, and it will speed up the research of RNA-protein interactions and make more information accessible to researchers.

High-throughput experiments are expanding biology by providing unbiased view and leading to unexpected observations. In this thesis, I introduce two types of development for global investigation of poly(A) tails and single nucleotide resolution survey of RNA-protein interactions. By applying these methods, I discover several phenomena at the 3′ end of RNAs and the binding interfaces between RNA and RBPs. Further development and improvement will offer an ample opportunity for the discovery of unforeseen regulatory pathways.

**Keywords:** Transcriptomics; High-throughput sequencing; RNA-protein interaction; Poly(A) tail; Gene regulation

**Student ID:** 2009-30858

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# List of Abbreviations

| | |
|---|---|
| APA | alternative polyadenylation |
| ARE | AU-rich element |
| ART-seq | active mRNA translation sequencing |
| AUC | area under the curve |
| CCD | charge-coupled device |
| CCR4 | carbon catabolite repression protein 4 |
| cDNA | complementary DNA |
| CDS | coding sequence |
| CHART | capture hybridization analysis of RNA targets |
| ChIRP | chromatin isolation by RNA purification |
| CIMS | crosslinking-induced mutation sites |
| CLASH | crosslinking, ligation, and sequencing of hybrids |
| DNA | deoxyribonucleic acid |
| DSE | downstream sequence element |
| eIF4G | eukaryotic translation initiation factor 4 gamma |

| | |
|---|---|
| EM | expectation-maximization |
| ESC | embryonic stem cell |
| FragSeq | fragment sequencing |
| FRT-seq | flowcell reverse transcription sequencing |
| GEO | Gene Expression Omnibus (NCBI service) |
| GMHMM | Gaussian mixture hidden Markov model |
| GMM | Gaussian mixture model |
| GRO-seq | genomic run on sequencing |
| IRES | internal ribosome entry site |
| LINE | long interspersed element |
| LNA | locked nucleic acid |
| lncRNA | long noncoding RNA |
| LTR | long terminal element |
| miRNA | microRNA |
| MPSS | massively parallel signature sequencing |
| mRNA | messenger RNA |
| Mt-tRNA | mitochondrial transfer RNA |
| NCBI | National Center for Biotechnology Information |

| | |
|---|---|
| ncRNA | noncoding RNA |
| NET-seq | native elongation transcript sequencing |
| NOT | negative regulator of transcription (protein) |
| NSR-seq | not-so-random sequencing |
| OLB | off-line basecaller (Illumina® software application) |
| oligo(dT) | oligo-deoxythymidine |
| PABP | poly(A) binding protein |
| PAN2 | PAB-dependent poly(A) specific ribonuclease 2 |
| PAN3 | PAB-dependent poly(A) specific ribonuclease 3 |
| PAP | poly(A) polymerase |
| PARE | parallel analysis of RNA ends |
| PARN | poly(A)-specific ribonuclease |
| PARS | parallel analysis of RNA structure |
| PAS | polyadenylation signal |
| PCR | polymerase chain reaction |
| RACE | rapid amplification of cDNA ends |
| RBP | RNA-binding protein |
| RC | rolling circle (repeat element) |

| | |
|---|---|
| rDNA | ribosomal DNA |
| RIP-seq | RNA immunoprecipitation sequencing |
| RLM-RACE | RNA ligase mediated RACE |
| RNA | ribonucleic acid |
| RNA-PET | RNA paired-end ditag |
| RNAi | RNA interference |
| ROC | receiver operation characteristics |
| rRNA | ribosomal RNA |
| RT-PCR | reverse-transcription and polymerase chain reaction |
| RTA | real-time analysis (Illumina® software application) |
| SAGE | serial analysis of gene expression |
| scRNA | small cytoplasmic RNA |
| SHAPE-seq | selective 2′-hydroxyl acylation analyzed by primer extension sequencing |
| SINE | short interspersed element |
| snoRNA | small nucleolar RNA |
| SNP | single nucleotide polymorphism |
| snRNA | small nuclear RNA |
| SOLiD | sequencing by oligonucleotide ligation and detection |

| | |
|---|---|
| SRA | Sequence Read Archive (NCBI service) |
| srpRNA | signal recognition particle RNA |
| tRNA | transfer RNA |
| UCSC | University of California at Santa Cruz |
| USE | upstream sequence element |
| UTR | untranslated region |
| WC | Watson-Crick |

# 1. Introduction

## 1.1 Post-transcriptional regulation of eukaryotic gene expression

RNA is a dynamic molecule of life. It is continuously generated, processed, used, and degraded in the cell. Nearly every step of RNA metabolism is tightly regulated in mammalian cells. Beginning with the transcription initiation, many mechanisms control expression level and primary structure of RNA. Unlike its cousin, DNA, it forms uncountable types of secondary structures, travels around the cell, and often changes sequence composition.

The earliest widespread regulation in the life cycle of RNA is alternative splicing. It generates various types of isoforms depending on the combination of splicing factors. In human, at least 74% of multi-exon genes are known to have multiple isoforms (Johnson et al., 2003), and the fraction continues to grow with introduction of more sensitive techniques. RNA binding proteins (RBPs) like NOVA, PTB, and FOX2 bind to unspliced pre-mRNAs, and collectively determine inclusion of exons into the mature forms of mRNAs (Keren et al., 2010).

At the other end of the cascade, translation provides an opportunity to control gene expression. mRNA expression levels explain only ~60% of protein expression levels in eukaryotic cells (Maier et al., 2009). Assuming that protein degradation has only minor effects on global deviation of gene expression (Schwanhäusser et al., 2011), the major fraction of the gap between mRNA and protein levels may be explained by the regulations in translation initiation, translation elongation, and mRNA localization. The 5′ UTR of *Fth1* is blocked by iron regulatory protein (IRP) in iron-deficient condition, and iron-dependent release of IRP makes it translatable (Gray & Hentze, 1994). AU-rich element (ARE), known for stability determinant in mRNA sequences, also has roles in both translational up-regulation and down-regulation through interaction with ARE binding proteins (Barreau et al., 2005). MicroRNA-induced silencing complex (miRISC) is another

factor that induces both decay and translational repression of its targets (Fabian et al., 2010; Huntzinger & Izaurralde, 2011; Guo et al., 2010; Bazzini et al., 2012; Djuranovic et al., 2012). Physical position of mRNA in the cell is also subject to post-transcriptional regulations. Type D simian retroviruses escape from nuclear retention by using constitutive transport element (CTE) that recruits nuclear export factors (Braun et al., 1999). The *CaMKIIα* mRNA localizes in the distal dendrites through its *cis*-regulatory element in 3′ UTR, which determines local concentration of its protein product (Mayford et al., 1996).

Cytoplasmic polyadenylation and deadenylation add another complexity of post-transcriptional control of RNA. The cytoplasmic polyadenylation element (CPE), which is found near the polyadenylation signal in the 3′ UTR, is bound by CPE binding protein (CPEB) to extend its poly(A) tail in the cytoplasm (de Moor & Richter, 1999). It is known to promotes translation initiation and stabilize the mRNA during oogenesis, early embryo development, localized translation of *CaMKIIα*, and cyclin mRNAs in cell cycle progression (Mendez & Richter, 2001; Weill et al., 2012; Norbury, 2013). Deadenylation of mRNA is often coupled with its decay. miRISC is known to induce deadenylation of its targets (Huntzinger & Izaurralde, 2011; Djuranovic et al., 2012). Messenger RNAs undergoing nonsense-mediated decay (NMD) are rapidly deadenylated by poly(A) ribonuclease (PARN) before decapping, 5′-3′ and 3′-5′ exonucleases become active (Lejeune et al., 2003). mRNAs with ARE can be deadenylated, depending on the protein partner (Mukherjee et al., 2002).

Post-transcriptional regulation is known to have many layers, and may be even more complex than we currently anticipate. Indeed, there may still exist vast repertoire of RNA-mediated cellular mechanisms yet to be discovered.

## 1.2 High-throughput methods in RNA biology

Development of gene expression profiling techniques enabled simultaneous monitoring of massive number of transcripts. They not only allow measurements of genes of interest but also provide a big picture of physiological status of cells. The first two high-throughput methods in gene expression profiling were cDNA microarray (Schena et al., 1995) and

serial analysis of gene expression (SAGE) (Velculescu et al., 1995). The cDNA microarray technology is an extension of Southern blot into bigger scale. It hybridizes cDNAs labeled with imageable material like fluorophore or silver to oligonucleotide probes attached to a solid surface, then the light from spots is analyzed after imaging (Schena et al., 1995). SAGE is an automated Sanger sequencing-based technique that counts concatamerized short sequence tags generated by treatment of restriction enzyme to double-stranded cDNAs (Velculescu et al., 1995). Many large scale projects have sought genome-wide insights into gene expressions using both technologies for last two decades while microarray has been overwhelmingly popular.

In the late-2000's, the methodology of RNA biology started to face fundamental changes by commercialization of high-throughput DNA sequencing technology (Ronaghi et al., 1998; Bentley et al., 2008). Recent break-through discoveries in the field have heavily relied on high-throughput sequencing methods. The discovery of PIWI-interacting RNAs, long interspersed noncoding RNAs, their action mechanisms, non-canonical processing pathways of microRNA, and global views of splicing regulations are the achievements based on high-throughput sequencing. Unlike the pre-existing techniques which rely on the known sequences and gene structures, RNA sequencing techniques based on high-throughput DNA sequencing delivered far more information. The independence to prior knowledge of gene enabled scientists to discover new RNA molecules, unknown isoforms, RNA modifications, and gene fusions, and to develop applications such as crosslinking, immunoprecipitation, and sequencing (CLIP-seq) and ribosome profiling (also known as ribo-seq or ribosome footprinting). Table 1.1 summarizes some of popular applications of high-throughput sequencing in the field of RNA biology. This section describes some of the techniques related to the main content of this thesis.

### 1.2.1 Transcriptome profiling

RNA-seq was developed as a replacement for cDNA microarray, then soon recognized its powerfulness over the conventional methods. Unlike cDNA microarray, it is free from cross-hybridization, and sensitivity near splice junctions is without parallel to various microarray technologies (Nagalakshmi et al., 2008). When compared to the older sequencing-based

**Table 1.1** List of popular applications of high-throughput sequencing in RNA biology (continued in the following pages).

| Feature | Method | Description | Reference |
|---|---|---|---|
| Expression profiling | RNA-seq | RNA fragmentation, RT-PCR, cDNA sequencing (details often vary) | Nagalakshmi et al. (2008) |
| | DeepSAGE | Concatamers of sequence tags generated from cDNA by restriction enzyme | Nielsen et al. (2006) |
| | FRT-seq | Adapter ligation, RT-PCR on flow cell to lower amplification bias | Mamanova et al. (2010) |
| | NSR-seq | RT with not-so-random primer to avoid rRNAs | Vignali et al. (2011) |
| Transcript structure (with expression profiling) | Targeted RNA-seq | Enrich RNA of interest with oligonucleotide probes | Mercer et al. (2012) |
| | DeepCAGE, nanoCAGE, or CAGEscan | Fragmentation, enrich fragments with 5′ cap or specifically attach 5′ adapter with template switching | Plessy et al. (2010); Kurosawa et al. (2011) |
| | RNA-PET | Mated pair tags with type II restriction enzyme for full-length cDNAs | Fullwood et al. (2009) |
| | Direct RNA sequencing | Hybridize and sequence RNAs directly to flow cell in single molecule sequencer | Ozsolak et al. (2009); Sharon et al. (2013) |
| | 3P-seq, PAS-seq | Enrich 3′ end fragment of 3′ UTRs where poly(A) begins | Shepard et al. (2011); Jan et al. (2011) |
| | 3′T-fill-seq | Skip poly(A) by filling dTTP in dark cycles | Wilkening et al. (2013) |

**Table 1.1** (continued)

| Feature | Method | Description | Reference |
|---|---|---|---|
| Transcription activity | GRO-seq | Short incubation of nucleus for nascent RNA labeling and purify them | Core et al. (2008) |
| | NET-seq | Immunopurify RNAPolII-RNA complex, sequence 3′ ends | Churchman & Weissman (2011) |
| RNA secondary structure | SHAPE-seq | Chemically label for single-stranded regions, probe interrupted reverse-transcriptions | Lucks et al. (2011) |
| | PARS | Treat RNase V1 and S1, analyze accumulated 5′ ends | Kertesz et al. (2010) |
| | FragSeq | Treat RNase T1+A, T4 PNK, or none, analyze accumulated 5′ ends | Underwood et al. (2010) |
| RNA-protein interaction | RIP-seq | Immunopurify RNA-protein complex *in vitro* | Zhao et al. (2010) |
| | HITS-CLIP | Crosslink RNA-protein complex by UV C, treat RNase, immunopurify, and treat protease | Licatalosi et al. (2008) |
| | PAR-CLIP | Incubate cells to label RNA with 4SU or 6SG, and CLIP with UV A | Hafner et al. (2010) |
| | iCLIP | Similar to CLIP, ligate 5′ adapter after reverse-transcription, probe interrupted reverse-transcriptions | König et al. (2010) |

**Table 1.1** (continued)

| Feature | Method | Description | Reference |
|---|---|---|---|
| RNA-RNA interaction | CLASH | Crosslink RNA-protein complex with UV, ligate two nascent RNA molecules to each other, then CLIP | Kudla et al. (2011) |
| RNA editing | m6A-seq | Random fragmentation, immunopurify $m^6A$ containing fragments | Dominissini et al. (2013) |
| Translation activity | Ribosome profiling or ribo-seq | Treat RNase to lysate, purify monosome in sucrose cushion | Ingolia et al. (2009) |
| | ART-seq | Treat RNase to lysate, purify monosome by size-exclusion chromatography | Freeberg et al. (2013) |
| | Polysome profiling by sequencing | Polysome fractionation, purify mono-, di-, tri-, and polysomes separately from fractions | Spies et al. (2013) |
| RNA-chromatin interaction | ChIRP, CHART | Crosslink DNA-RNA-protein complex by formaldehyde. Pull-down chromatin fragments using array of probes complementary to RNA of interest, sequence DNA | Chu et al. (2011); Simon et al. (2011) |
| Endonuclease specificity | Degradome-seq, PARE | Enrich poly(A)$^+$ RNAs, ligate 5′ adapter depending on 5′ monophosphate, analyze accumulated 5′ ends | Addo-Quaye et al. (2008); German et al. (2008) |

techniques like SAGE or , RNA-seq has far more utility not only in generic expression profiling but also in gene structure discovery or in isoform-specific expression profiling thanks to its longer size of sequenced reads (Mortazavi et al., 2008). As the high-throughput sequencing became cheaper and deeper, RNA-seq is now widely used in most modern biology fields.

As a matter of fact, RNA-seq does not point to a specific experimental procedure, but indicates any variant of sequencing techniques starting from RNA when it does not involve a specific biochemical purification except poly(A) enrichment or rRNA depletion. The earliest RNA-seq experiments were performed by conversion of RNAs to double-stranded DNA by RT-PCR, then processed with the regular methods of DNA sequencing library preparation (Nagalakshmi et al., 2008). Later, alternative approaches have been developed to overcome the limitation of the previous method that cannot provide information on the strand of RNA (reviewed in Levin et al., 2010). The three major variants of library preparation methods for RNA-seq are briefly summarized in Figure 1.1. At the time of writing this thesis, the RNA ligation-based method is generally favored over the other methods.

### 1.2.2 RNA-protein interactome analysis

In the global analysis of RNA-protein interactions, methods are separated into two disciplines depending on whether *in vivo* RNA-protein crosslinking is used. RNA immunoprecipitation-sequencing (RIP-seq) is an RNA-seq application that sequences RNA that is co-immunoprecipitated with a protein of interest (Zhao et al., 2010). It is often criticized on its critical drawback that RNA-protein complex can be formed artificially during the experimental process (Riley & Steitz, 2013). HITS-CLIP and its variants are available for more stringent purification. CLIP determines sequences of RNA interacting with a protein by crosslinking RNA and protein using ultraviolet light irradiation (Figure 1.2) (Licatalosi et al., 2008; Licatalosi & Darnell, 2010). Although it excludes *in vitro* artifacts by design, CLIP often misses true targets depending on experimental conditions like UV wavelength, buffer condition, or dozens of minor steps in sequencing library preparation (Singh et al., 2013). It is known that several RNA binding domains are not efficiently crosslinked (Singh et al., 2013). More

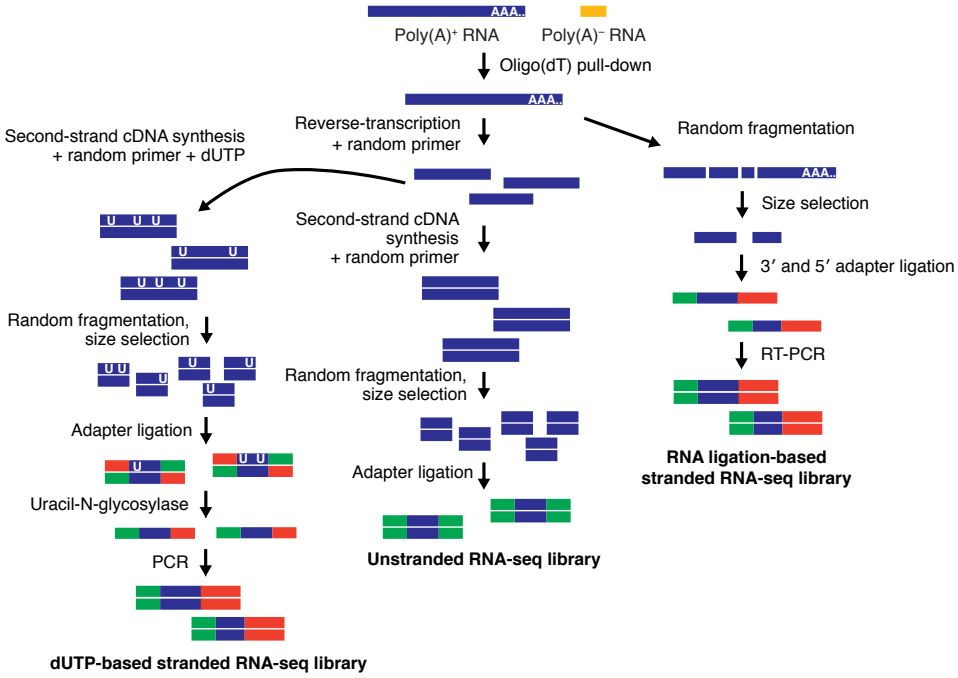



**Figure 1.1** Brief procedure of the three major methods for RNA-seq.

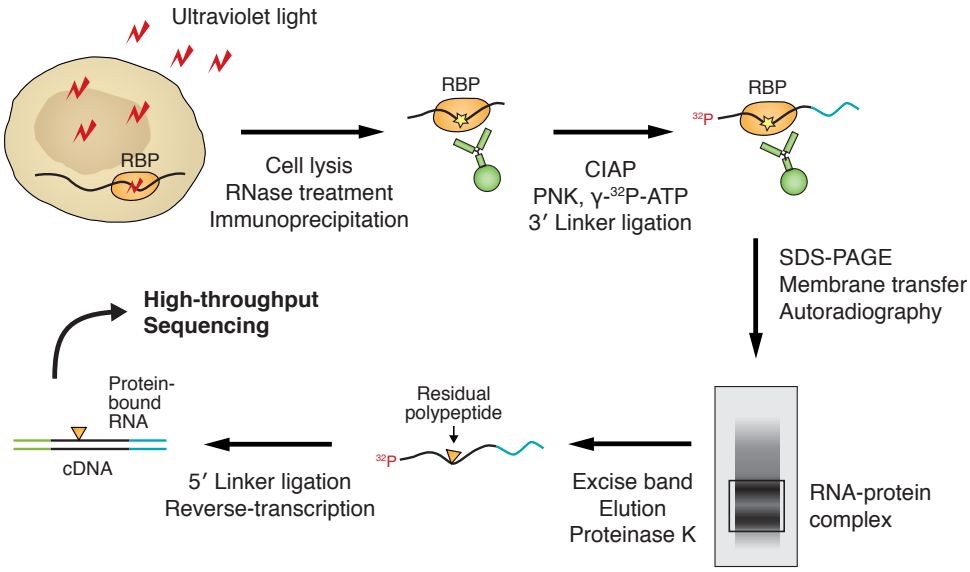



**Figure 1.2** The procedure of a typical CLIP-seq experiment.

details on the CLIP techniques will be covered in Section 3.1.

### 1.2.3 Monitoring transcriptome-wide polyadenylation status

Survey of poly(A) tail length in mRNAs will allow researches to investigate cytoplasmic polyadenylation and deadenylation mechanisms. Cytoplasmic polyadenylation and deadenylation is known to play on important role in late oogenesis, cell cycle progression, microRNA targeting mechanism, and synaptic plasticity (Weill et al., 2012; Norbury, 2013). A sufficiently accurate transcriptome-wide method to measure poly(A) lengths can be a game changer in the field as the tool will enable to find messengers whose tails are regulated. Early attempts were based on oligo(dT) chromatography (Figure 1.3) (Beilharz & Preiss, 2007; Meijer et al., 2007). They were successful to a certain degree, but the method had limitations. It could not distinguish true poly(A) tails from A-rich mRNA body. Longer poly(A) tails could not be subdivided to higher resolution. With the introduction of high throughput sequencing, few groups tried to analyze poly(A) with RNA-seq (Wu et al., 2008; Ulitsky et al., 2012). However, they could not achieve enough accuracy due to insuf-

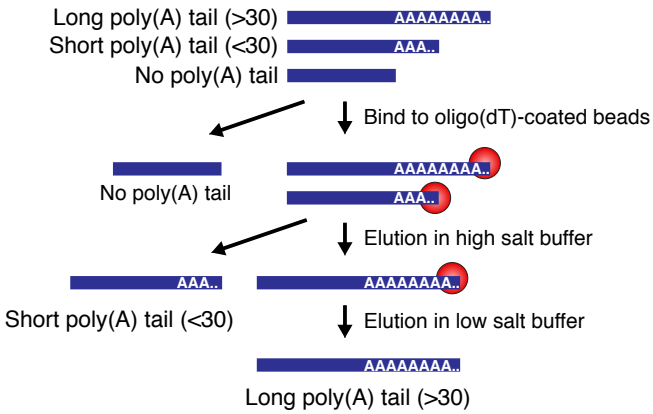



**Figure 1.3** Method to separate short (<30 nt) and long (>30 nt) poly(A) RNAs (Meijer et al., 2007).

ficient dynamic range of signal (Wu et al., 2008) or inaccuracy in Illumina base calling for homopolymers (Ulitsky et al., 2012). In Chapter 2, I introduce a newly developed solution for global investigation of poly(A) tails.

### 1.2.4 Analysis of RNA ends

The termini of RNA are formed and regulated by highly organized mechanisms. The 5′ end of mRNA is where transcription starts. Export, stability, and translation are regulated by 5′ capping (Nevins, 1983). The other end of mRNA is protected by poly(A) tail and it is known to be important for mRNA stability and translation efficiency (Weill et al., 2012). The termini on both ends of tRNA give selectivity in aminoacylation (Schimmel et al., 1993). Stability and processing efficiency of precursor miRNA are controlled by 3′ end nucleotidyl additions (Heo et al., 2009, 2012). Incorrect processing of either end of precursor miRNA can alter the function of mature miRNAs (Vermeulen et al., 2005; Park et al., 2011).

Highly parallel investigation of RNA ends for miRNA is relatively straightforward as the standard protocol gives whole information of the molecules (Park et al., 2011). For longer RNAs, special modifications are required to read either end of the RNA. Rapid ampli-

fication of cDNA ends (RACE) incorporates oligo(dT)+VN primed reverse-transcription for 3′ end of 3′ UTR (3′ RACE), reverse-transcriptase template switching for 5′ end of RNAs (5′ RACE), single-stranded adapter ligation for both ends (RLM-RACE), or circularization for both ends (circular RACE) (Frohman et al., 1988; Liu & Gorovsky, 1993; Scotto-Lavino et al., 2006a,b,c). Modified version of classical RACE is used for parallel investigation of mRNA ends (Olivarius et al., 2009). Another variations of 5′ RACE, parallel analysis of RNA ends (PARE) (German et al., 2008) and degradome sequencing (Addo-Quaye et al., 2008) have shown their another utility on comprehensive identification of endonucleolytic cleavage events. Section 2.9 of this thesis introduces a different variant of these approaches that captures 3′ ends in transcriptome-wide fashion.

# 2. Transciptome-wide profiling for 3′ ends of poly(A)$^+$ RNAs

## 2.1 Background

The 3′ termini of eukaryotic RNAs reflect the history of transcript and play important roles in determining the fate of RNA. The 3′ ends are generated by endonucleolytic cleavage, untemplated nucleotidyl transfer and/or exonucleolytic trimming. In the case of messenger RNAs (mRNAs), the nascent transcripts are cleaved by cleavage and polyadenylation specificity factor (CPSF) and become polyadenylated by canonical poly(A) polymerase (PAP), with an exception of replication-dependent histone mRNAs that lack poly(A) tails (Norbury, 2013). Poly(A) binding proteins (PABPs) not only protect poly(A) tails but also interact with eIF4G bound to the 5′ cap, which is generally thought to facilitate translational initiation (Weill et al., 2012). Despite the importance, the actual sequences of 3′ ends remain unknown for the vast majority of transcripts, and our current knowledge is based on studies of a limited number of individual genes by northern- and RT-PCR/Sanger sequencing-based techniques (Norbury, 2013; Sallés et al., 1999).

Genome scale investigation has been hampered for several reasons. Firstly, current deep sequencing technologies cannot determine homopolymeric sequences of longer than ~30 nt. Although microarray combined with differential elution from oligo(dT) column have been used to roughly estimate poly(A) length (Beilharz & Preiss, 2007; Meijer et al., 2007), the resolution is too low for accurate measurement. Secondly, highly abundant RNAs such as rRNAs and tRNAs dominate cDNA library unless mRNAs are enriched by oligo(dT) capture which inevitably introduces bias towards mRNAs with long poly(A) tails. Moreover, when oligo(dT) is used as a primer for reverse transcription or as an adapter in splint ligation, the sequence information at the very end of RNA is lost in the cDNA library. Thus, global investigation of RNA 3′ end has been largely limited to the mapping

of polyadenylation sites that mark the boundary between mRNA body and poly(A) tail (Beck et al., 2010; Ozsolak et al., 2010; Mangone et al., 2010; Yoon & Brem, 2010; Fu et al., 2011; Jan et al., 2011; Shepard et al., 2011; Derti et al., 2012; Martin et al., 2012; Elkon et al., 2013; Hoque et al., 2013; Wilkening et al., 2013).

## 2.2 Technical difficulties of sequencing poly(A) tails

To the exact sequencing of poly(A) tails by high-throughput sequencing, there are several technical difficulties that cannot be easily handled. This section describes the major limitations of modern high-throughput sequencing technologies on sequencing poly(A) tails.

### 2.2.1 Problems in high-throughput sequencing for long homopolymers

The sequencing technologies without reversible terminator, such as Roche 454 and Life Technologies IonTorrent, report accumulated signals for homopolymers (Metzker, 2010). In their imaging, dynamic range of fluorescence signal is limited within charge-coupled device (CCD) cameras and usually tuned to maximize sequencing performance of regular sequence composition in the genome (Metzker, 2010; Bragg et al., 2013). Therefore, extensively long homopolymers like poly(A) tails often exceed their linear range of signal quantification, or even maximum measurement limits.

Even in technologies adopted reversible terminators, such as Illumina and Life Technologies SOLiD, homopolymers are still one of the most difficult substrates to sequence. The second generation sequencers require PCR amplification of templates to secure enough signal intensity to be detected by fast imaging techniques (Shendure & Ji, 2008; Metzker, 2010). The sequencing reactions from multiple templates are commonly out of sync after several cycles of reaction because each part of the reaction has small chance of failure. Polymerization reaction sometimes fails to incorporate an incoming nucleotide. This type of error and its subsequent effect on sequencing are collectively called *phasing*, and their occurrence is estimated as 0.1% in the modern Illumina equipments (Ledergerber & Dessi-

moz, 2011). Reversely, a single cycle occasionally incorporate more than one nucleotide. The phenomenon, called *pre-phasing*, is known to occur approximately 0.05% of chances. Either type of errors results in mixed-up signals coming from desynchronized templates. Deconvolution using phasing and pre-phasing parameters is essential for sequencing longer reads than 30 cycles (Ledergerber & Dessimoz, 2011).[1] However, their estimations are extremely tricky at the ends of long homopolymers. Correction of the blended signals easily become corrupt as the signal is overwhelmingly uniform inside long homopolymers.

Illumina sequencing technology has another shortcoming called *sticky-T phenomenon* . Each step of sequencing-by-synthesis (SBS) reaction finishes with cleavage and wash out of fluorophores conjugated to nucleotide bases. While most of them are removed from the clusters, minor fractions remain still in the template (Whiteford et al., 2009). The rate of the persistence is specifically higher for T residues, which corresponds to poly(A) tails in the reverse direction. As a result, the persisting fluorophores accumulate over reactions for long homopolymers. T signal lasts for many more cycles after the end of T homopolymeric region, then the subsequent cycles are called as T regardless of their source sequences.

Third generation sequencers are better at these problems by taking merits of single-molecule sequencing. Both Helicos and PacBio are free from phasing and pre-phasing issues (Metzker, 2010; Ozsolak & Milos, 2011). Moreover, the latter doesn't suffer from build-up of fluorophore signals because it utilizes measurement of electrical conductivity, which does not accumulate over time in principle (Metzker, 2010). They could be ideal platforms for sequencing poly(A) tails if they produce enough throughput, but they are still limited to lower throughput by two or three orders of magnitudes when compared to the main stream sequencers (Sharon et al., 2013).

### 2.2.2 Design of library construction

As Illumina was the only platform that provide both long (>300bp) and enough number of reads with base-by-base measurement, our experiment designs were targeted for Illumina

[1]The feasible quality read lengths without corrections of phasing and pre-phasing is usually estimated as 36bp for Illumina and 25bp for SOLiD.

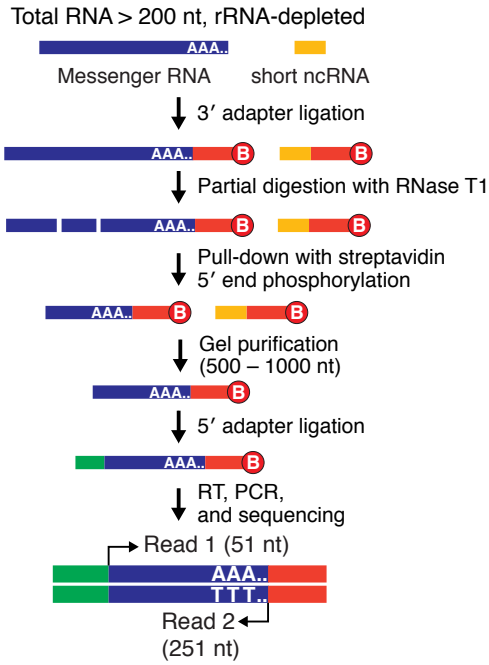



**Figure 2.1** Schematic description of experimental procedure. Horizontal bars represent examples of RNA or DNA molecules in each step. Colors of bars indicate (blue) mRNA or its complementary DNA, (yellow) small non-coding RNAs, such as snRNA, snoRNA or 5.8S rRNA, (red) 3′ adapter or Illumina P7-containing primer, (green) 5′ adapter or Illumina P5-containing primer. B in red circle mark the position of biotins.

chemistry.[2] The experimental procedure is almost identical to the regular paired-end RNA-seq library preparations (Wang et al., 2009; Ozsolak & Milos, 2011). We applied few changes to the conventional schemes to enrich 3′ end of RNAs in resulting library (Figure 2.1). The 3′-most part of RNAs are generally depleted in the conventional RNA-seq due to the different size distribution from the fragments from the middle parts of mRNAs (Stern-Ginossar et al., 2012). In TAIL-seq, 3′ adapter is ligated to the 3′ end of RNA before fragmentation. This enables not only enrichment of 3′-most part but also capturing the sequence information on intact 3′ hydroxyl end of RNA. In addition, we adopted an rRNA

[2]This study was designed in collaboration with Jaechul Lim.

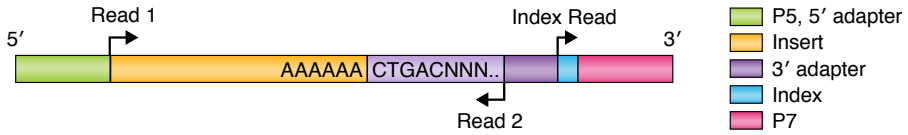



**Figure 2.2** Sequence structure of a complete TAIL-seq tag in sequencing library. P5 and P7 are pre-designated sequences by Illumina for binding and amplification on the flow cell. 'N' stands for a degenerate base.

depletion kit based on antisense LNA probes instead of oligo(dT) pull-down methods to remove rRNA. As one of the major objective of TAIL-seq is to quantitatively profile poly(A) tails, it needs to be independent of affinity to oligo(dT) sequences (Raz et al., 2011). The change also brings an ability to survey 3′ ends of poly(A)$^-$ RNAs.

Poly(A) sequences are challenging for fundamental machinery of sequencing, too. We carefully designed sequences used in the preparation of library to improve the sequencing performance (Figure 2.2). Basically, read 1 provides the identity of the source transcript while read 2 is used for measurement of poly(A) tail length. Read 2 can be also used for investigation of 3′ ends of poly(A)$^-$ RNAs. Index read allows multiplex runs, which sequences multiple samples in the same lane of flow cells. The multiplexing is beneficial not only for a financial reason but also for minimization of lane-to-lane and run-to-run variations in an experiment set.

Sequencers based on optical imaging need to adjust their optics accurately to get high-quality image and sequence. The composition of bases in every sequencing cycle is generally unbiased for the common sequencing libraries. As a TAIL-seq library generally contains significant proportion of poly(A) tails, base composition in the first cycles of read 2 would be greatly biased to T. As Illumina sequencer takes images separately from different bases, getting correct exposure to CCD becomes difficult when the composition is significantly unbalanced (Illumina, Inc., 2011). It is also harder to correctly focus cameras on the plane where reaction occurs (Illumina, Inc., 2011). Moreover, sequence diversity of the first few cycles of sequencing has substantial impact on overall sequencing. Images from the first four cycles are used to identify cluster positions in the field of view on

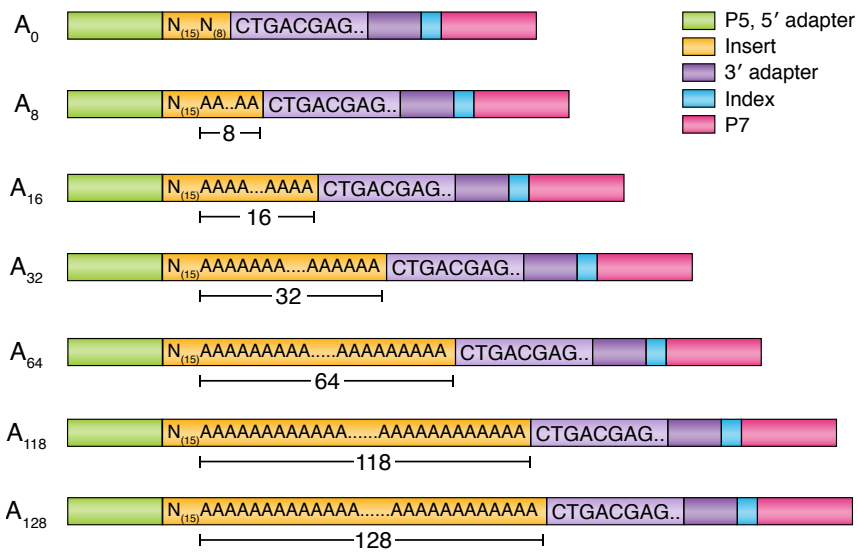



**Figure 2.3** Sequence structure of diagnostic poly(A) spike-in library. See the text for the details of design.

Illumina sequencers (Illumina, Inc., 2011), thus unbalanced cycles are prone to lose true spots or gain false spots. Lack of sequence diversity in poly(A) tails is also problematic for estimation of phasing and pre-phasing matrices as the first 25 cycles serves as its reference data (Illumina, Inc., 2011). To resolve these issues, we added fifteen degenerate bases, which are chemically synthesized from equimolar mixture of dNTPs (Figure 2.2, 'N's in light violet bar). With the complexity region, cluster registration and imaging becomes more stable for highly homogeneous libraries in sequence composition. We also added a pentamer with fixed sequence between inserts and the degenerate bases to distinguish chemically synthesized adapters from sequences from the 3′ end of inserts (Figure 2.2, 'CTGAC' in light violet bar).

Exact measurement requires enough number of references whose quantity is known. We added seven chemically synthesized poly(A) spike-ins to characterize signals from poly(A) tails (Figure 2.3). They are designed similarly as the structure of final cDNA tag for TAIL-seq except that they carry fifteen random bases at the beginning of read 1, and first fifteen bases of read 2 are designated with a fixed sequence (Figure 2.3). The random

region stabilizes optic control and image analysis of read 1 by diversifying the sequence composition. The random region on the side of read 2 are changed to a fixed sequence due to the technical limitation in chemical synthesis of nucleic acids.[3] Characterization of signal, machine learning, and benchmarks using these synthetic poly(A) spike-ins will be covered later in this chapter.

## 2.3 Sequence data processing and acquisition

Signal processing of Illumina sequencing starts with imaging. Cluster spots are isolated from the images, then their signal intensities are quantified for all four channels, A, C, G, and T, over the reaction cycles. The built-in software called real-time analysis (RTA) takes care of these processes including *base calling,* which is to convert the signals into DNA sequences. This section describes the methods of TAIL-seq to process the signals after the initial processing by RTA.

### 2.3.1 Data acquisition and processing

TAIL-seq libraries were sequenced in 51+251 paired-end layout with Illumina HiSeq 2500 or MiSeq. The base calls and signal intensities were acquired from the sequencers after processing by Illumina RTA 1.17.21.3 (HiSeq) or 1.18.42 (MiSeq) (Table 2.1). The base calls were collected and transformed into .qseq files using Illumina off-line basecaller (OLB) 1.9.4. Together with the .qseq files, an in-house script collected cluster intensity matrices from .cif files via Picard 1.91 (http://picard.sourceforge.net/). A new file format with suffix .sqi was designed for efficient storage and random access to the cluster intensity matrices and sequences. Sqi file was defined with seven fields in tab-separated text file (Table 2.2). For the faster access and efficient storage, sqi files are stored as compressed with a random accessible compression format called bgzip (Li et al., 2009) and indexed using Tabix (Li, 2011).

[3] Integrated DNA technologies, Inc., who synthezied these oligonucleotides for us, reported that more degenerate bases make the yield and purity of DNA synthesis worse.

**Table 2.1** List of files output from the Illumina sequencing pipeline that are used in this study. Refer Illumina, Inc. (2011) for more details.

| File type | File name suffix | Description | Use in this study |
|---|---|---|---|
| Cluster intensity | .cif | Raw unprocessed signal intensities | To get original signal intensity of each channel over the cycles |
| Filter | .filter | Quality check result from cluster passing filter | To check if a cluster produces good signals in regard of signal intensity, cluster overlaps, and other factors affecting base calling |
| Control | .control | Flag whether the cluster is control or not | To remove control spots from analysis |
| Position | .clocs or .locs | Geometric positions of clusters in tiles | To find spots in images for case-by-case investigations |
| Offset | .txt | Geometric offsets among images for channels and cycles for a tile | To find spots in images for case-by-case investigations |
| Base call | .bcl | Base calls in DNA sequence | To get sequence information from sequenceable regions (read 1; read 2 for poly(A)$^-$ RNA; and 3′ terminal modification) |
| Thumbnail image | .jpg | Compact summary of original images for diagnostic use | To find spots in images for sample case investigation |

**Table 2.2** Fields and descriptions of sqi file format defined in this study.

| Field name | Type | Description |
|---|---|---|
| Tile | decimal | Name of the tile where the cluster locates |
| Cluster | decimal | Unique identifier of the cluster in tile |
| QC pass | 0 or 1 | Flag indicating whether the cluster passed QC filter |
| Sequence | IUPAC string | DNA sequence from RTA base calls |
| Quality | Phred+33 | Quality score of base call in Phred+33 scale (Ewing et al., 1998) |
| Cluster intensity | base64 (Josefson, 2003) | Raw signal packed in sequence of base64 digrams. Values are adjusted to fit in [0, 4095] by scaling and trimming, then the values are encoded in order of A, C, G, and T, then the unit is repeated over the cycles. |
| Read 2 insert start | decimal | Zero-based inclusive coordinate of the first cycle for 3′ end of insert in read 2 |

The original images of clusters were collected from thumbnail images for diagnostic analyses. First, clusters that are visible in the center magnification window of thumbnail images were selected from the full list of clusters. Then, the positions of clusters were calculated using the position files and the sub-tile offset files. Later image manipulations were performed with Python Imaging Library (PIL) 1.1.7 (http://www.pythonware.com/products/pil/). The thumbnail images were magnified to 10-fold height and width of original size by bicubic spline interpolation to utilize sub-pixel offsets of image alignment. Cluster images were cropped with the window size of 7×7 pixels (70×70 pixels in working buffer). The collected cluster images were stored in raw four channel 8-bit image in a Berkeley hash database.

### 2.3.2 Sequence processing and alignment

The read 1 sequences were aligned to the common contaminants set, which is composed of rDNA repeat units (GenBank accession BK000964.1 for NIH3T3 and U13369.1 for HeLa), PhiX genome (GenBank accession J02482.1), Illumina TruSeq primer sequences, and all sequences for 5S and 5.8S rRNAs of respective species (retrieved from Rfam 11.0 (Burge et al., 2013) of the Wellcome Trust Sanger Institute) using GSNAP 2013-03-31 (Wu & Nacu, 2010) with maximum 5% mismatches allowed. Clusters with any match to the contaminants were removed from the subsequent analyses.

The sequences having completely identical nucleotides in the 21st to 35th cycle in read 1 (representing region of the insert) and the 1st to 15th cycle in read 2 (degenerate bases in 3′ adapter) are deduplicated by leaving only a cluster with the maximum Phred quality sum of read 1. The degenerate and fixed delimiter sequence in 3′ adapter was clipped out from read 2 by searching perfect match of delimiter sequence ('GTCAG' as in the direction of read 2) between the 14th and 16th cycles in read 2. The clusters missing a delimiter sequence or having low diversity in degenerate region (at least two occurrences for all of A, C, G, and T) were removed from further analyses.

The remaining reads after contaminant filter and the first duplication filters were then aligned to the genome sequences (UCSC mm10 for NIH3T3 and UCSC hg19 for HeLa, positions of splicing junctions were processed from the UCSC Genome Browser database for version of Jan 24, 2013) using GSNAP 2013-03-31 (Wu & Nacu, 2010). Three different versions of alignments to genome were used in this study. (1) *R1 alignment*: using only the full read 1 sequences which are 51 nt long. This was used for identification of a cluster. (2) *R2 short alignment*: using only 40 nt right next to the 3′ adapter of read 2. This was used in searching for the poly(A)-free 3′ hydroxyl ends. (3) *paired alignment*: using the full read 1 sequences and part of read 2 sequences trimmed of degenerate bases and delimiter. I filtered out poly(A) stretches encoded from genome using this alignment set. All the alignments were performed with maximum mismatches of 5%, minimum mapping quality of 3. All multi-mapped reads were removed.

PCR artifacts with few mismatches were removed again using the R1 alignment with

15 degenerate bases inside the 3′ adapter region. To detect that kind of artifacts, I clustered the R1 alignments with maximum distance between mapped positions of 10 bp, they were then clustered again within the first cluster using degenerate bases from read 2 of respective reads with CD-HIT-EST 4.5.4 (Fu et al., 2012) (word size=6, sequence identity=0.85). For a set of detected duplicates, I chose a read with maximum sum of Phred quality in read 1 to leave.

### 2.3.3 Sequence annotation and classification

For classification and transcript-level analyses, I compiled reference annotations for human and mouse using NCBI RefSeq (Pruitt et al., 2012), RepeatMasker, gtRNAdb (Chan & Lowe, 2009), Rfam (Burge et al., 2013), and miRBase (Kozomara & Griffiths-Jones, 2011) databases (the first three were downloaded from the UCSC Genome Browser (Kuhn et al., 2013) on Apr 25, 2013; Rfam version 11; and miRBase version 19). The R1 alignments were annotated with intersection with the compiled annotations using BEDTools (Quinlan & Hall, 2010). When multiple annotations were overlapped to an alignment, I chose a class for the statistics requiring exclusive assignment of a genomic source type by the following priority: miRNA, rRNA, tRNA, Mt-tRNA, snoRNA, scRNA, srpRNA, snRNA, lncRNA, RNA, ncRNA, misc_RNA, Cis-reg, ribozyme, RC, IRES, frameshift_element, LINE, SINE, Simple_repeat, Low_complexity, Satellite, DNA, LTR, CDS, 3′ UTR, 5′ UTR, intron, Other, Unknown (higher priority first).

The transcript-level analyses were performed using my custom non-redundant RefSeq (nrRefSeq) transcript set, which is a reduced set retaining only the longest isoform or transcript when regions overlap with each other. The positions of read 1 in nrRefSeq transcripts were positioned with BEDTools intersection between alignments to genome sequences and nrRefSeq annotation set, and then translated to the transcript-level coordination with in-house software.

## 2.4 Processing fluorescence signals for sequencing poly(A) tails

Illumina sequencers produce quantized fluorescence signals in four channel multivariate values. Although the signal intensities reflect the original sequence composition of templates, it is considerably affected by both systematic and random noises. The loss of clarity in signal patterns becomes especially stronger for templates having low complexity like poly(A) tails.

What do signals from poly(A) tail look like? How can they be recognized to measure their length? First, cluster intensity signals from pilot runs of TAIL-seq were analyzed to extract properties of signals that can be used in detection of poly(A) tails.[4] Then, hundreds of clusters were manually inspected whether the original signal contain some clue. Many of poly(A) spike-ins showed remarkable difference in former and later cycles of the designed end of poly(T) stretch (few examples from $A_{64}$ spike-in are shown in Figure 2.4). Although the decrease of T signal intensity was relatively mild and slow, rise of the other signals (A, C, or G) was significant near the borders (Figure 2.4). The transition was more visible for shorter poly(A) spike-ins like $A_{16}$ and $A_{32}$, yet it was detectable enough for longer poly(A) spike-ins like $A_{118}$ and $A_{128}$ (Figure 2.5). Accordingly, I constructed a unified metric that indicates the relative signal intensity to simplify further analyses:

$$U_c = \frac{S_{c,T}}{\sum_{b=A,C,G} S_{c,b}}$$

where $U_c$ is a simplified metric for signal bias to T in cycle $c$, $S_{c,b}$ is the original signal intensity of channel $b$ for cycle $c$. However, the dynamic range of signal intensity of each channel and cluster is differentiated by many factors: inconsistency of chemical environments by physical position of clusters in the flow cell; optical and image processing glitches by cluster's position in view of lens and image sensors; and nucleotide composition or sequence-specific characteristics of cDNA templates. To relieve variability from these factors, I exploited the degenerate bases located in the first twenty nucleotides as normalization factors of the individual channels. The revised formula incorporating normalization

[4] All explorative trials of sequencing in the designing stages of TAIL-seq were prepared and performed by Jaechul Lim.

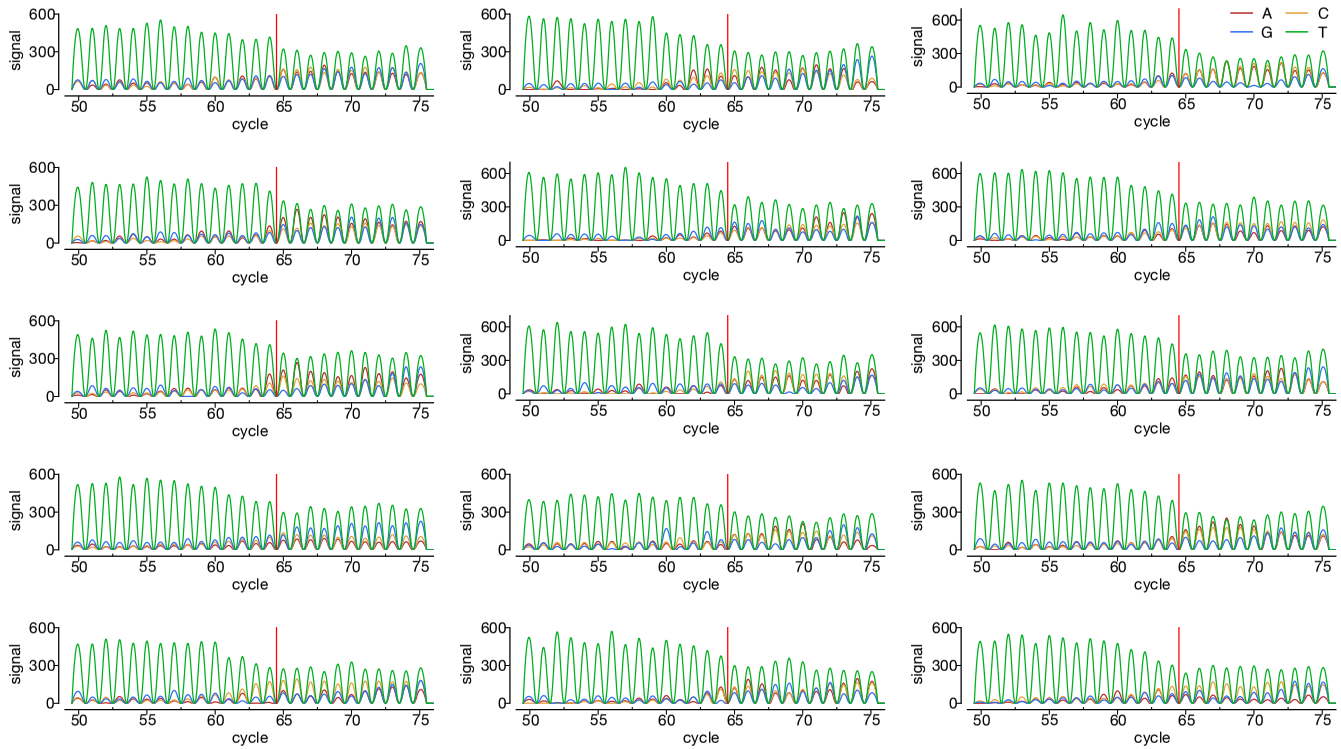



**Figure 2.4** Raw signal intensity values near the designed transition (red vertical lines) from poly(T) stretch to heterogenous sequences of randomly selected $A_{64}$ spike-in clusters. Peak signals of the curves reflect the original signal values. The curves are transformed to resemble oscillating functions via cubic spline interpolation for better legibility.

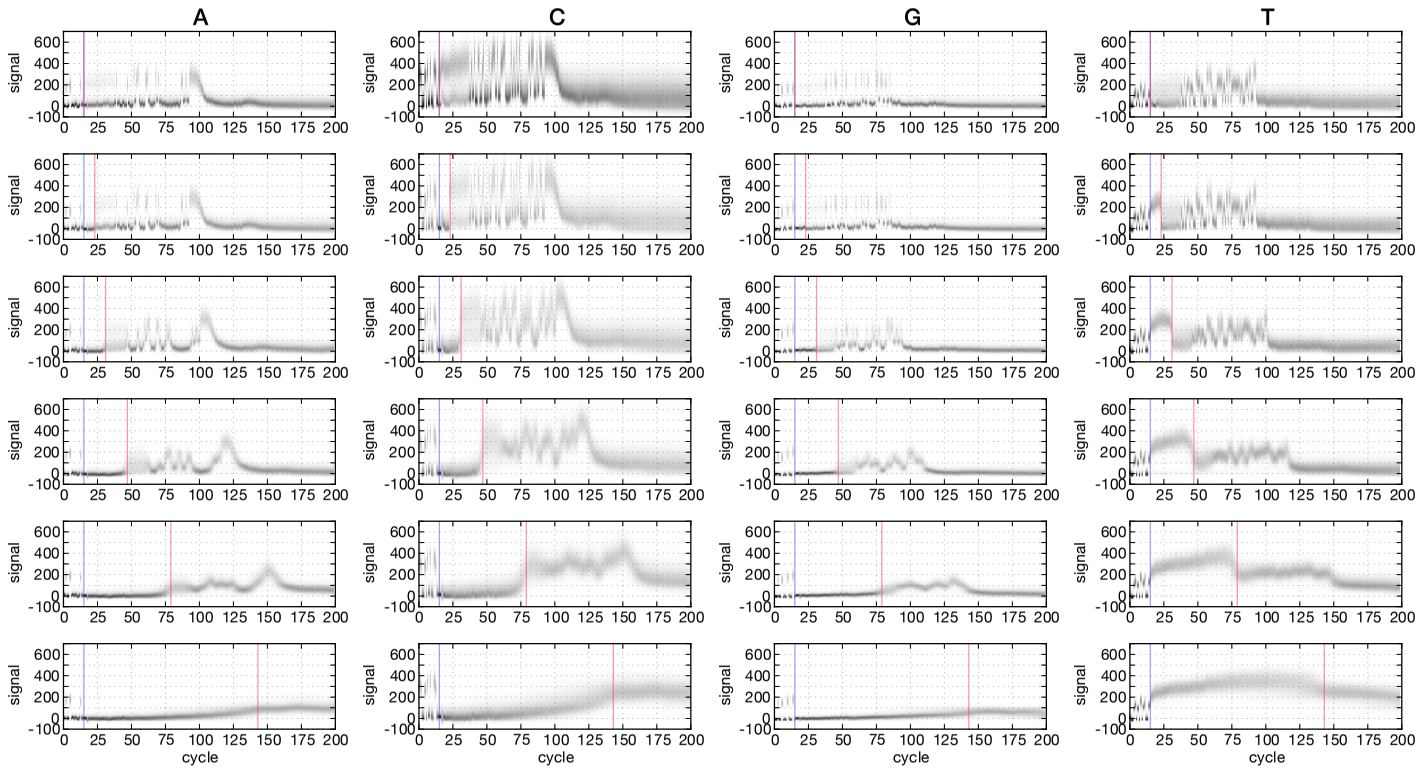



**Figure 2.5** Distributions of signal intensity from four channels for different poly(A) spike-ins. The position of the first cycle inside poly(T) region is indicated with blue vertical line. Red lines show the point of transition from poly(T) stretch to subsequent heterogeneous sequences.

factors becomes:

$$
\begin{aligned}
N_b &= \frac{\sum_{c=R_\alpha}^{R_\sigma} S_{c,b}}{R_\sigma - R_\alpha + 1} \\
F_{c,b} &= \frac{S_{c,b} + \lambda}{N_b + \lambda} \\
T_c &= \log_2 \frac{\lambda + F_{c,T}}{\lambda + \sum_{b=A,C,G} F_{c,b}}
\end{aligned}
$$

where $N_b$ indicates the reference signal of channel $b$, $R_\alpha$ and $R_\sigma$ are the first and last cycles of degenerative bases in the 1-based coordinate, $F_{c,b}$ is an individual signal normalized by the reference for $c$-th cycle of channel $b$, $\lambda$ is a pseudo count number to avoid zero division, and the final metric $T_c$ is called "relative T signal" hereafter. The random samples of relative T signals from poly(A) spike-in samples show visually detectable edges near the expected transition points (Figure 2.6).

## 2.5 Machine learning for detection of poly(A) tail lengths

The relative T signal described in the previous section provided enough information, which simple heuristic method may give a satisfactory solution that matches to experienced human recognition. However, it was not that easy due to variation changes of signals across sequencing cycles. At the early cycles, transitions from T to non-T signals are very steep (Figures 2.5 and 2.6). The signal drop after poly(T) stretch becomes weaker and weaker as T stretch lengthens (Figures 2.5 and 2.6) due to sticky-T phenomenon (Whiteford et al., 2009). In addition, the full signal transition takes much more number of cycles in later cycles in read due to phasing and pre-phasing (Ledergerber & Dessimoz, 2011). Missing data, spot noises, run-to-run variation, and dependency of signal distributions on platforms[5] add more complexity on automated analysis of signals from poly(A) tails. Lastly, A-rich regions near the 3′ end of 3′ UTR can't be easily distinguished from poly(A) tails using context-free algorithms. I describe the design and assessments of methods based on several different approaches later in this section.

[5] HiSeq had more than four-fold wider dynamic range of MiSeq at similar signal-to-noise ratio in our sequencing runs (data not shown).

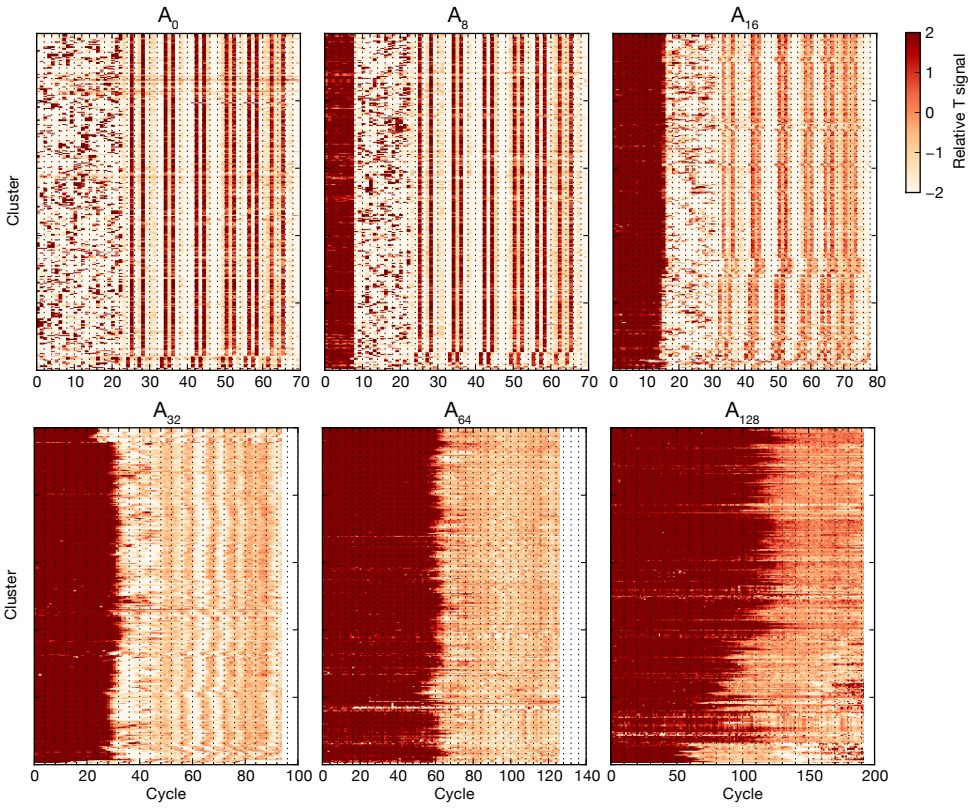



**Figure 2.6** Examples of relative T signals from poly(A) spike-ins. Each panel shows a thousand clusters randomly sampled during the signal processing. Cycle numbers shown on $x$-axes are adjusted to start from 1 where the first cycle of poly(T) stretch.

### 2.5.1 Homogeneous sampling for training set

Poly(A) spike-in samples are highly variable in quality of signals and purity over the sequencing runs. What is the most concerning is variance of poly(A) length in the original template itself. Both chemical synthesis (Hecker & Rill, 1998) and enzymatic amplification (Schlötterer & Tautz, 1992) tend to produce shortened oligonucleotides. Our poly(A) spike-ins showed the significant variability of length in long homopolymeric regions (Figure 2.6). In addition, HiSeq often failed to sequence index reads with high quality (data not shown), which makes samples mixed up. To minimize run-to-run variations and enable automated machine learning of routine analyses, training data set had to be purified before subsequent steps. In this study, I used an outlier filter based on robust Mahalanobis distance implemented in the R *mvoutlier* package 1.9.9 (quan=0.5, alpha=0.025, applied after fifteen-fold downsampling of relative T signals). As a result, majority of outliers was filtered out (Figure 2.7), which is enough for providing homogeneous examples to learn poly(T) to mRNA body transitions.

### 2.5.2 Methods for poly(A) length measurement

The overall design of expected procedure to poly(A) length measurement is to use relative T signals to predict the original state of template sequence and measure the count of consecutive poly(A) states (Figure 2.8).

#### Multivariate Gaussian mixture model

The junctions between poly(T) stretch and heterogenous sequences show steep change of relative T signal in surrounding cycles (Figure 2.6). As the existence of poly(T) stretch and its 3′-most position can be easily detected by sequence analysis, a model of relative T signals near the junctions can reveal the most probable position of the transition. It is first modeled as a multivariate Gaussian mixture model (GMM) of relative T signals from several consecutive cycles. The parameters for the first trial were chosen with empirical estimations from previous observations (Table 2.3). For the better modeling of the signal,

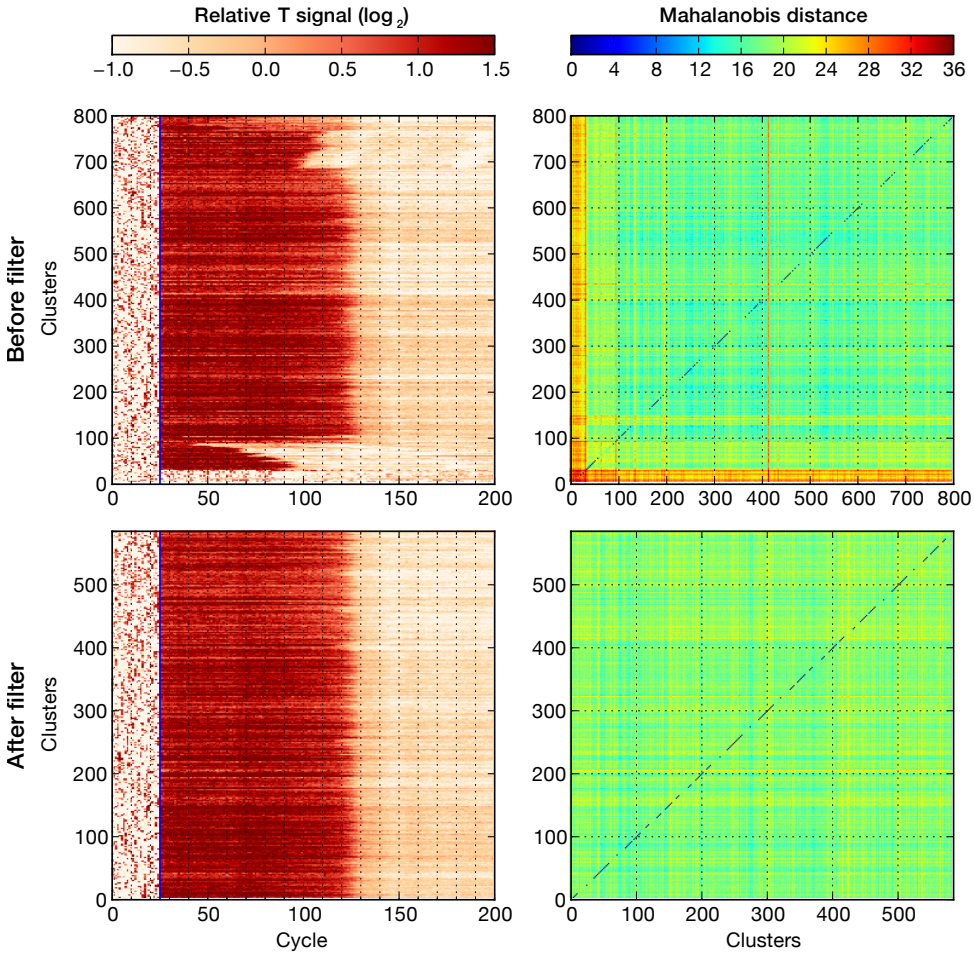



**Figure 2.7** Example of relative T signals of clusters which were used in training poly(A) length measurement algorithms. Left panels indicate normalize T signals while right panels are their internal Mahalanobis distance matrices. The random samples from original poly(A) spike-ins are shown in top panels, the data after *mvoutlier* filtering with Mahalanobis distance are shown in bottom panels. Note that this example is from a pilot run that adopted old design of poly(A) spike-in with 103 As.

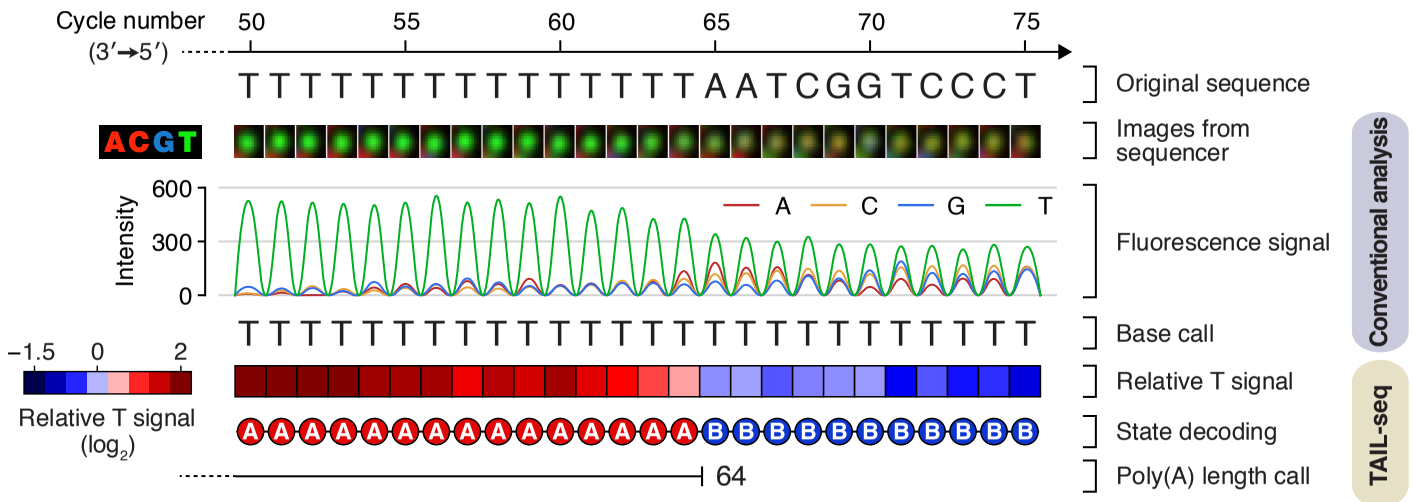



**Figure 2.8** An example of the analysis procedure for poly(A) length measurement. Shown is a spike-in ($A_{64}$) cluster from cycles corresponding to the 50th to 75th nucleotides from the 3′ end. 'Images from sequencer' indicates serial pictures of a cluster taken in each sequencing cycle (red for C, green for T, blue for G; red also reflect A signal due to innate crosstalk between fluorophores). 'Fluorescence signal' is the scaled signal intensity measured from the images. 'Base call' shows the sequence determined by built-in software (Illumina RTA). 'Relative T signal' indicates the T signal divided by the sum of other signals (A, C, and G, see Figure S1A for details), which was then used for machine learning to judge whether or not the cycle is from poly(A) region ('State decoding').

**Table 2.3** The initial parameters of a Gaussian mixture model for relative T signals of adjacent cycles around the end of poly(T) stretch. Positions are shown as intervals when ten cycles for each side of the end are modeled in a window.

| Positions | [0, 10) | [10, 20) |
|---|---|---|
| Distribution 1 | N(1.5, 1.5) × 0.95 | N(1.5, 1.5) × 0.25 |
| Distribution 2 | N(-1, 1.5) × 0.05 | N(-1, 1.5) × 0.75 |

the initial parameters were optimized with the expectation-maximization (EM) algorithm (Dempster et al., 1977) to maximize the product of maximum likelihood of outlier-filtered training set prepared as described in Section 2.5.1.

### Gaussian mixture hidden Markov model

The simple Gaussian mixture model cannot easily account the contextual characteristics that poly(T) is a long continuous stretch and it does not appear once the cycle enters the heterogenous region. Gaussian mixture hidden Markov model (GMHMM) can incorporate information of the entire trend into the detection of transition. Initially, the topology of a GMHMM was designed with two states of poly(A) and non-poly(A), but later it was extended to have four states because it appeared that long poly(A) tails required additional transitive states to cover the longer mixed regions of phased and pre-phased templates (data not shown). I trained the HMM in left-to-right topology (Figure 2.9) with empirical initial parameters (Tables 2.4 and 2.5). The poly(A) spike-ins were sequenced and learned to generate a model together with TAIL-seq libraries on every sequencing run to adapt to variable signal characteristics. Then, the model parameters were optimized using Baum-Welch algorithm with the implementation in the GHMM library (http://ghmm.org) (Table 2.6 and 2.7). Although the length of poly(A) tails were known to every poly(A) spike-in, the optimization did not use any of the prior knowledge of expected transition positions because length variability of poly(A) is already significant at the stage of sequencing. As it would be simpler, more explicit, and more powerful for the model to account signal transitions only, I separated considerations of the length variation from this stage. Length

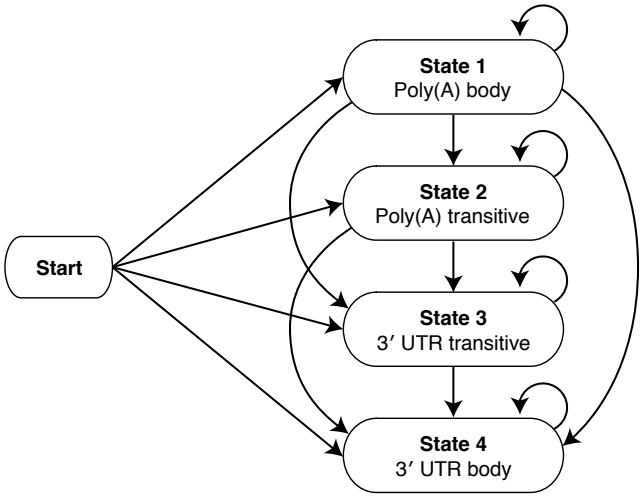



**Figure 2.9** Topology of the hidden Markov model for learning poly(A) signals used in this study.

**Table 2.4** Initial parameters for transition probability matrix of GMHMM of poly(A) signals. 'S' states indicate start or end states, which is inserted between examples to learn.

| From\To | 1 | 2 | 3 | 4 | S |
|---|---|---|---|---|---|
| 1 | 0.94 | 0.03 | 0.01 | 0.01 | 0.01 |
| 2 | 0 | 0.5 | 0.4 | 0.08 | 0.02 |
| 3 | 0 | 0 | 0.6 | 0.38 | 0.02 |
| 4 | 0 | 0 | 0 | 0.95 | 0.05 |
| S | 0.95 | 0.01 | 0.01 | 0.03 | 0 |

**Table 2.5** Initial parameters for emission probability distributions of GMHMM of poly(A) signals. 'S' states indicate start or end states, which is inserted between examples to learn.

| State | Dist. 1 | Dist. 1 weight | Dist. 2 | Dist. 2 weight |
|---|---|---|---|---|
| 1 | $N(1.5, 1.5)$ | 0.95 | $N(-1, 1.5)$ | 0.05 |
| 2 | $N(1.5, 1.5)$ | 0.75 | $N(-1, 1.5)$ | 0.25 |
| 3 | $N(1.5, 1.5)$ | 0.5 | $N(-1, 1.5)$ | 0.5 |
| 4 | $N(1.5, 1.5)$ | 0.25 | $N(-1, 1.5)$ | 0.75 |
| S | $N(1000000, 1)$ | 1 | - | 0 |

**Table 2.6** Example of optimized parameters for transition probability matrix of GMHMM of poly(A) signals. The parameters were fitted to one of our pilot sequencing runs using unsupervised Baum-Welch algorithm.

| From\To | 1 | 2 | 3 | 4 | S |
|---|---|---|---|---|---|
| 1 | 0.972 | 0.019 | 0.009 | 0 | 0 |
| 2 | 0 | 0.958 | 0.015 | 0.027 | 0 |
| 3 | 0 | 0 | 0.981 | 0.007 | 0.012 |
| 4 | 0 | 0 | 0 | 0.981 | 0.019 |
| S | 0.718 | 0.126 | 0.156 | 0 | 0 |

**Table 2.7** Example of optimized parameters for emission probability distributions of GMHMM of poly(A) signals. The parameters were fitted to one of our pilot sequencing runs using unsupervised Baum-Welch algorithm.

| State | Dist. 1 | Dist. 1 weight | Dist. 2 | Dist. 2 weight |
|---|---|---|---|---|
| 1 | $N(5, 0)$ | 0.3973 | $N(3.72, 0.7)$ | 0.6027 |
| 2 | $N(1.77, 0.73)$ | 0.9092 | $N(3.59, 0.83)$ | 0.0908 |
| 3 | $N(1.06, 2.56)$ | 0.2038 | $N(-2.4, 3.01)$ | 0.7962 |
| 4 | $N(-0.42, 0.11)$ | 0.3053 | $N(-0.84, 0.48)$ | 0.6947 |
| S | $N(1000000, 1)$ | 1 | - | 0 |

calling for poly(A) tails was done with the standard Viterbi algorithm (Viterbi, 1967) implemented in the GHMM library. Unlike an algorithm based on base calling (Ulitsky et al., 2012), the newly developed method estimates length of poly(A) tails similar to the expected length (Figure 2.10).

**GMHMM-based method with crosstalk matrix**

Due to innate overlap among emission spectra of fluorophores, raw signal intensity of each channel interferes each other. In the current Illumina chemistry, the strongest interference occurs for C from A and T from G (Ledergerber & Dessimoz, 2011). The phenomenon also confirmed in our data (Figure 2.11, left panel). As the emission spectrum of G overlaps with T, G-rich regions near poly(A) tails could hinder the exact measurement. In this method, I replaced the original raw signals with orthogonalized signals from Illumina OLB 1.19.4 (Figure 2.11, right panel).

**Edge detection using the first or second derivatives**

Edges with sharp gradient can be detected with the first or second derivatives. Most state-of-art edge detectors, such as Canny edge detector and Prewitt operator (Canny, 1986; Prewitt, 1970), use the first derivative of signal to emphasize the boundaries. When a cluster undergo transition from poly(T) region to heterogenous sequences in body, the slope of their relative T signal becomes significantly negative. Due to phasing and pre-phasing, the gradient is much milder in longer poly(A) tails (Figure 2.6), but the spanning width of the negative slope is longer for them. I adopted the Savitzky-Golay filter (Savitzky & Golay, 1964) which is a popular tool to smoothen and differentiate a set of discrete data points simultaneously. The first implementation of this approach finds the end of T stretch by seeking the position where the first derivative of relative T signal is less than zero. Due to pre-phasing and polymerase slippage, the first position where the first derivative turns negative is usually earlier than expected. Therefore, a variant was implemented by extending the T stretch region as long as the first derivate is negative. The relative T signals from long poly(A) tails include substantial amount of noise. I added another variant that

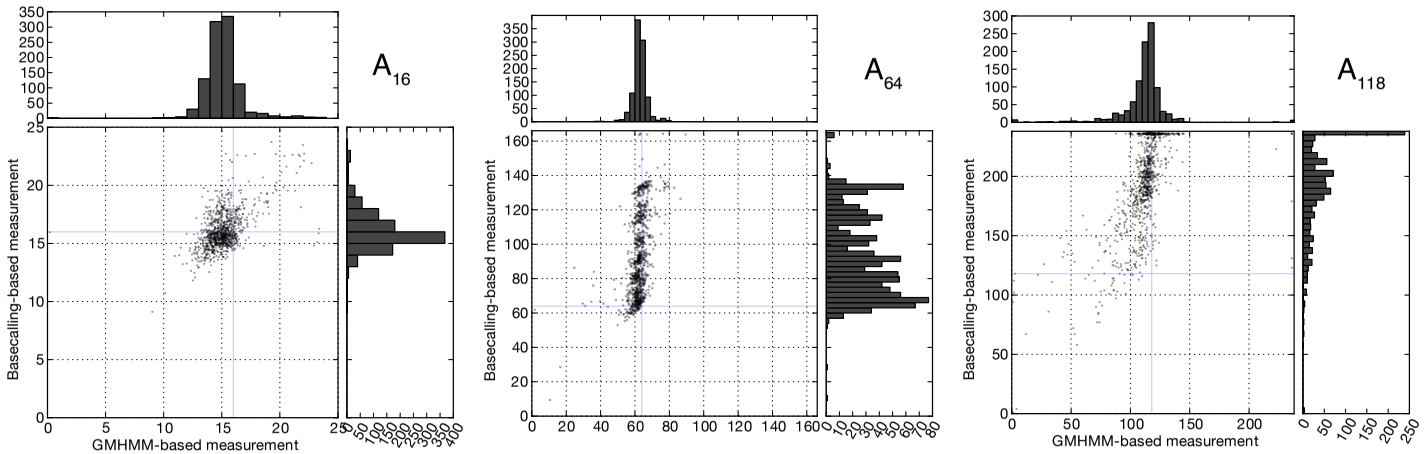



**Figure 2.10** Comparison between a previous method by Ulitsky et al. (2012) based on base calls and the method described in this section. Each dot represents a cluster of spike-in samples. Position in the $x$-axes indicates measurement from GMHMM, and measurements from base calling determines position in the $y$-axes. In the $A_{16}$ plot, Gaussian noise $N(0, 0.5)$ is added to both $x$ and $y$ positions to help recognition of the density distribution.

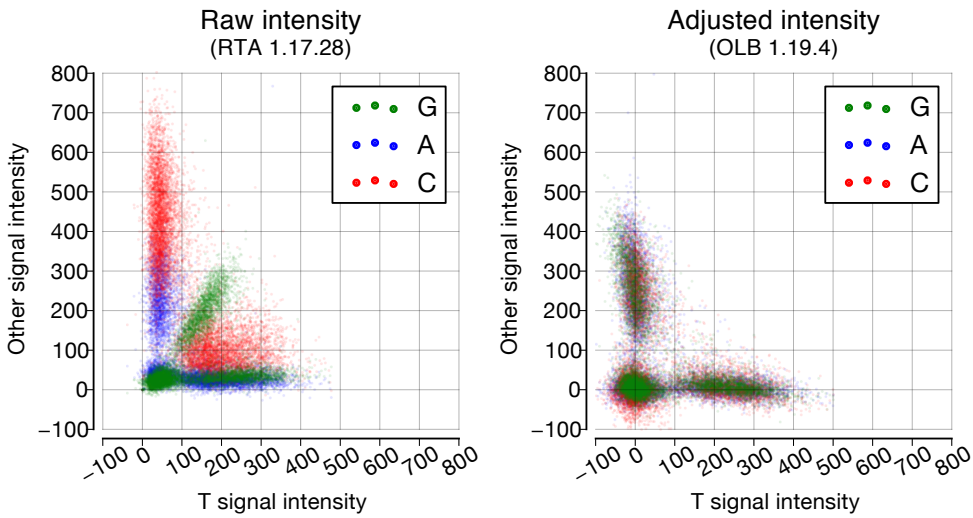



**Figure 2.11** Signal crosstalk between different fluorophores. (left) Correlation between signals from T and the other bases. (right) Correlation between signals after orthogonalization by Illumina OLB 1.19.4.

starts with the cycle where relative T signal is lower than a pre-defined threshold, and extends the region as long as the second derivative is negative.

### Benchmark

The seven approaches mentioned above were assessed by measuring poly(A) lengths for poly(A) spike-ins. In despite of the variability of poly(A) length in the original molecules themselves, an algorithm with better performance would produce more accurate length consistently as designed. Table 2.8 shows representative descriptive metrics from the benchmark. GMHMM without a crosstalk matrix was unanimously the best performer for all poly(A) spike-ins (Table 2.8). Unexpectedly, deconvolution using the crosstalk matrix (*GMHMM 2* in Table 2.8) was less accurate than the original signal (*GMHMM 1* in Table 2.8). It is not clear how the difference results this. The original signal may be clear enough to be recognized by the model. The methods based on the simple Gaussian mixture model (*GMM 1* and *GMM 2* in Table 2.8) performed remarkably worse than the

**Table 2.8** Benchmark result of methods to measure poly(A) lengths described in this section. Tests were performed with outlier-filtered spike-in reads from one of pilot data sequenced with MiSeq. RMSE, root mean squared error. *GMM 1,* multivariate Gaussian mixture model with empirically designed initial parameters (Table 2.3) of 20 cycle wide window. *GMM 2,* multivariate GMM with EM-optimized parameters for 100 iterations. *GMHMM 1,* the method based on Gaussian mixture hidden Markov model with the original signals. *GMHMM 2,* with deconvolved input signal described earlier in this section. *SG 1,* the first position with negative slope from Savitzky-Golay filter. *SG 2,* extending poly(T) until slope become non-negative. *SG 3,* starting with the first position with lower relative T signal than a threshold (0 in this benchmark), then extended the length as log as the second derivative is negative.

| Method | $A_{16}$ median | $A_{16}$ RMSE | $A_{64}$ median | $A_{64}$ RMSE | $A_{118}$ median | $A_{118}$ RMSE |
|---|---|---|---|---|---|---|
| GMM 1 | 15.0 | 2.451 | 59.0 | 12.075 | 91.0 | 46.704 |
| GMM 2 | 64.0 | 105.462 | 66.0 | 28.268 | 114.0 | 11.926 |
| GMHMM 1 | 16.0 | 1.170 | 63.0 | 4.325 | 116.0 | 9.169 |
| GMHMM 2 | 16.0 | 1.339 | 63.0 | 4.597 | 116.0 | 10.226 |
| SG 1 | 14.0 | 3.365 | 56.0 | 28.762 | 74.0 | 68.798 |
| SG 2 | 16.0 | 2.917 | 61.0 | 26.093 | 108.0 | 45.211 |
| SG 3 | 16.0 | 1.578 | 64.0 | 4.759 | 117.0 | 11.257 |

approaches based on hidden Markov model. The short and long poly(A) tails could not be modeled with a single simple model. The EM optimization with all kinds of poly(A) spike-ins fitted the model to long poly(A) tails only (Table 2.8). Even with parameter fitting to single type of poly(A) spike-ins, GMM was more inaccurate than GMHMM-based methods (RMSE=1.536 for $A_{16}$; RMSE=11.722 for $A_{118}$). Among the methods with numerical differentiation (*SG 1-3* in Table 2.8), the third version that adopts a static threshold of starting position was the most precise. Even for the size of smoothing window that performs best, the signals in poly(T) region was not stable enough for accurate detection of the width of the region (Table 2.8). Hereon, the GMHMM with the original signal *(GMHMM 1)* will be used for the measurement of poly(A) tails throughout this thesis.

### 2.5.3 Combination of measurements and base calls

GMHMM-based measurement outperforms the methods using base calls (Figure 2.10). Despite that, it is worthy to refer base calling because it is more accurate for short poly(A) stretches (< 8 nt) (data not shown), and it gives more information on 3′ terminal nucleotidyl additions like poly(A) uridylation (Rissland et al., 2007; Sement et al., 2013). I designed a simple method that determines poly(A) length and 3′ terminal modifications from base calls (Algorithm 2.1). In addition, a combined algorithm is developed to take benefits from both of base calls and GMHMM-based length measurements (Algorithm 2.2). Indeed, the all subsequent analyses are proceeded with the algorithm integrated base calls and GMHMM with original signal (Algorithm 2.2).

## 2.6 Poly(A) tails of the mammalian transcriptome

TAIL-seq libraries from mouse fibroblast cell line NIH3T3 and human cervical cancer cell line HeLa were sequenced and analyzed for in-depth analyses (29,610,077 and 21,794,337 reads, respectively, after filtering out PCR artifacts and rRNA reads).[6] The tags originate

[6] All data for NIH3T3 and HeLa cells except the miR-1 transfection experiment set used in this section are derived from TAIL-seq libraries by Jaechul Lim.

**Algorithm 2.1** Procedure that determines poly(A) tail length and 3′ end modifications from base calls. Scores were set as T=1, A/C/G=−10, and N=−5 in this study. Maximum length of 3′ end modification (*maxmod*) was assigned as 20.

```
procedure LocatePolyA(seq, seqlen)
  longest_i ← longest_j ← seqlen + 1
  longest_length = -1

  /* find the longest [i, j] with sum of score > 0 */
  for i ← from 0 to maxmod-1
    scoresum ← score of i-th base in seq

    /* if longest interval was not found set it with 1-nt long intv */
    if longest_length < 1 and scoresum > 0 then
      longest_length ← 1
      longest_i ← longest_j ← i
    end if

    /* try all possible end positions */
    for j ← from i+1 to seqlen-1
      add score of j-th base to scoresum

      if scoresum > 0 and j-i+1 > longest_length then
        longest_i, longest_j ← i, j
        longest_length ← j - i + 1
      end if
    end for
  end for

  if longest_length < 0 then return with no polyA found

  i, j ← longest_i, longest_j

  while i-th base ≠ T and i ≤ j, increase i by 1
  while j-th base ≠ T and i ≤ j, decrease j by 1

  return with polyA length of j-i+1, modification in [0, i) of seq
```

**Algorithm 2.2** Combined algorithm that determines poly(A) tail length and 3′ end modifications.

```
procedure FindPolyAAndModification(seq, seqlen)
  get basecall polyA length and mod. seq from LocatePolyA(seq, seqlen)

  if basecall polyA length ≤ 8 then
    return basecall polyA length and mod. seq
  else
    calculate GMHMM polyA length

    /* if GMHMM call short pA, it is more reasonable to ignore it */
    if GMHMM polyA length ≤ 8 then
      return basecall polyA length and mod. seq
    else
      return GMHMM polyA length and basecall mod. seq
    end if
  end if
```

mainly from the 3′ parts of genes although we also find internal tags that reflect endonucleolytic and exonucleolytic activities (Figure 2.12). I could measure the poly(A) length of 4,176 mouse and 4,091 human genes supported by ≥30 poly(A)$^+$ tags. Among the transcripts expressed by more than 50 copies per cell, 79.2% were detected with ≥30 poly(A)$^+$ tags in TAIL-seq (Figure 2.13). I compared our TAIL-seq data with previous results generated by differential elution from oligo(dT) column which separates mRNAs with short tails (<~30 nt) from those with long tails (Meijer et al., 2007) (>~30 nt) (Figure 2.14). Despite the differences between two methods, the long/short tail ratio correlates significantly with our measurements (P=0.0024, Pearson's correlation test; Figure 2.14).

Figure 2.15 presents an example of randomly chosen tags that match to the 3′ end of the *Trp53* mRNA, which encodes the p53 protein. Read 1 is used to identify the gene while read 2 is used to sequence the poly(A) tail of heterogeneous lengths. Various types of interesting information can be extracted from the TAIL-seq data.

### 2.6.1 Steady-state length distribution of poly(A) tails

I first examined the global distribution of poly(A) tail lengths. Overall, the distributions are similar between two cell lines examined (Figure 2.16). When the mRNA tags with

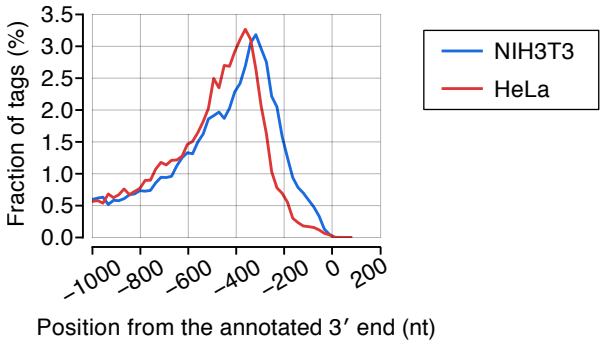



**Figure 2.12** TAIL-seq tags are enriched near the annotated 3′ end of RNA. $x$-axis shows the distance between the 5′ end of read 1 and the 3′ end of annotated transcripts.

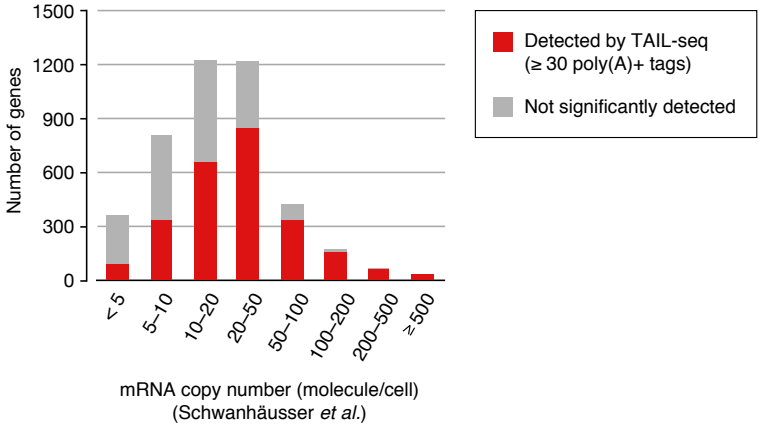



**Figure 2.13** Sensitivity of TAIL-seq according to mRNA level in cell. Number of transcripts that are represented with 30 or more poly(A)$^+$ tags are represented with red columns. The mRNA copy number per cell is based on estimations by Schwanhäusser et al. (2011).

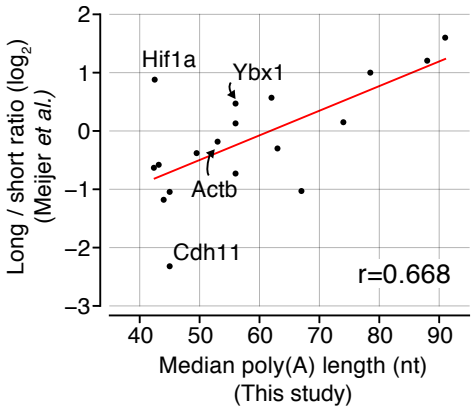



**Figure 2.14** Comparison to a previous estimation of poly(A) tail length in NIH3T3 cells (Meijer et al., 2007).

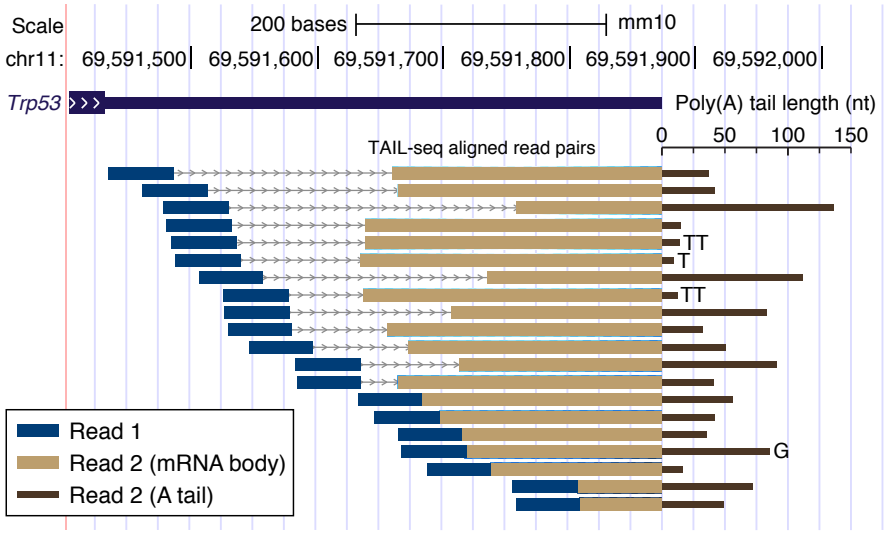



**Figure 2.15** An example pile-up of TAIL-seq tags. Blue bar indicates genome-mapped read 1 while the following light brown bar indicates an inferred region of read 2 corresponding to mRNA body, with untemplated adenine residues shown as dark brown bar. Additional modifications are shown on the right.

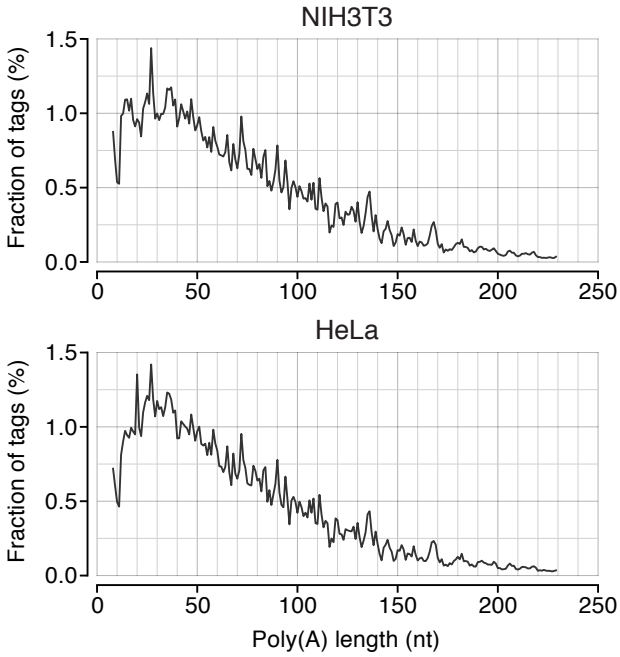



**Figure 2.16** Global distribution of poly(A) tail lengths of TAIL-seq tags.

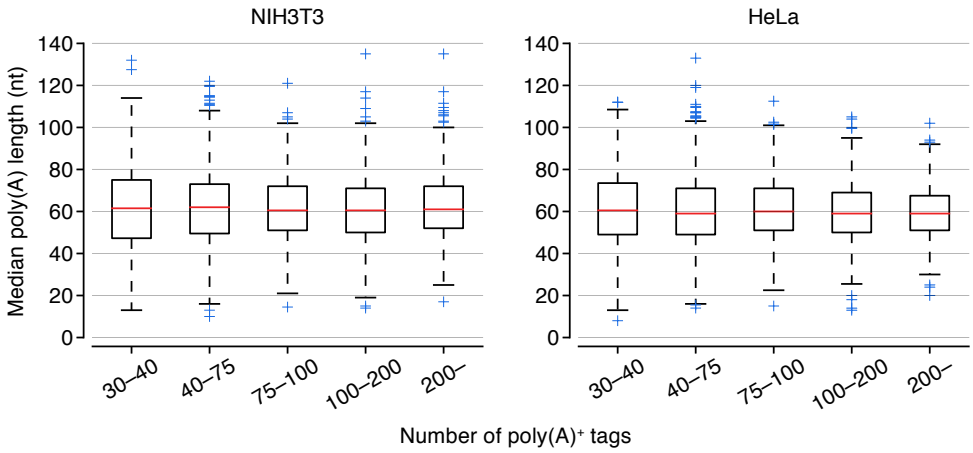



**Figure 2.17** Distribution of median poly(A) tail lengths depending on number of poly(A)-containing tags. Regardless of abundance, the median poly(A) tail lengths was around 60 nucleotides. Box represents the first and third quartiles and the internal bar indicates the median. Whiskers denote the lowest and highest values within 1.5 times the interquartile range of the first and third quartiles, respectively.

poly(A) tails of 8–231 nt are plotted, the median lengths are 60 nt and 59 nt in NIH3T3 and HeLa, respectively. Poly(A) tails over 231 nt could not be counted further due to the limited sequencing cycle but they account for only ~2 % of the total population. Poly(A) tails shorter than 8 nt were excluded from the analyses because the estimation was less accurate with such tags due to the ubiquity of short A stretches in the genome, particularly near polyadenylation sites. Accordingly, poly(A)-free RNAs such as histone mRNAs and decay intermediates were not included in this distribution analysis.

The tags derived from the same gene were clustered to calculate median poly(A) length for each individual gene (4,176 mouse and 4,091 human genes). The distribution of median poly(A) length was consistent over different abundance range of TAIL-seq tags (Figure 2.17). As expected, we found that poly(A) lengths vary widely among different genes (mRNA species) (Figure 2.18). Some mRNA species carry poly(A) tails of ~20 nt while others have long tails of ~100 nt. Based on these median poly(A) lengths for individual genes, transcriptome-wide median length (median of medians) is estimated to be 61 nt

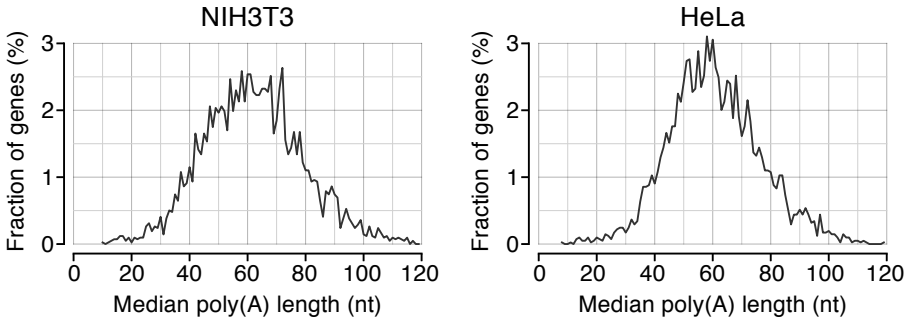



**Figure 2.18** Distribution of median poly(A) tail lengths of individual genes.

and 60 nt in NIH3T3 and HeLa cells, respectively. These values are significantly shorter than what is generally conceived as typical poly(A) tail length in mammals (Elkon et al., 2013). A newly transcribed transcript is known to receive a poly(A) tail of ~230 nt, but they are thought to be gradually shortened by deadenylases PARN, the PAN2-PAN3 complex, and the CCR4-NOT complex (Garneau et al., 2007). There are discrepancies over the poly(A) length in earlier reports based on bulk poly(A)$^+$ RNA or individual genes, which described poly(A) size as ~170 nt in mouse sarcoma polysomes, 100–160 nt in HeLa, and 50–70 nt in rabbit reticulocyte polysomes (Brawerman, 1974). But a recent study suggested that many mammalian mRNAs might have tails of smaller than 30 nt (Meijer et al., 2007). The current work offers an answer to this long-standing question by determining poly(A) tail length at the transcriptome level.

### 2.6.2 Impact of poly(A) tails on gene expression

I next asked whether genes with distinct biological functions tend to differ in poly(A) length distribution, by gene ontology analysis (Figure 2.19). Interestingly, genes associated with regulatory functions such as transcription factors tend to have shorter tails than those with relatively constitutive functions such as ribosomal subunits, which suggests that poly(A) tail of regulatory genes may be under dynamic control.

To understand which step of gene expression poly(A) tail may influence, I first compared the median poly(A) length of each gene with mRNA half-life that was estimated

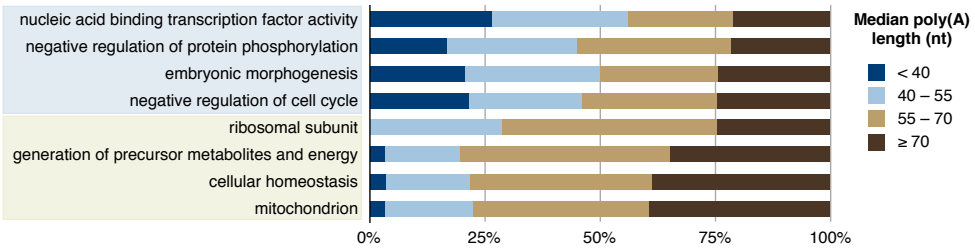



**Figure 2.19** Functional categorization of genes with their median poly(A) tail lengths. Four categories in the upper panel represent genes with relatively short poly(A) tails while the lower four categories represent genes with longer tails.

previously by Schwanhäusser et al. (2011). Overall, there is a modest but significant correlation between poly(A) tail length and mRNA half-life ($P=2.83\times10^{-5}$, Pearson's correlation test) (Figure 2.20). Thus, deadenylation and/or cytoplasmic polyadenylation may affect mRNA stability, as previously shown (Dreyfus & Régnier, 2002; Norbury, 2013). Of note, poly(A) tail length does not correlate significantly with steady state mRNA abundance, as expected (Figure 2.21).

One of the major mechanisms of miRNA action is known to be deadenylation, which had been proposed based on the studies of a few individual genes (Fabian et al., 2010; Huntzinger & Izaurralde, 2011; Bazzini et al., 2012; Djuranovic et al., 2012). This model is tested by examining the global effect of miRNA on poly(A) tail.[7] Synthetic miR-1 mimic was transfected into HeLa cells and subsequently poly(A) length was measured by TAIL-seq. Deadenylation of miR-1 targets was evident 6 hours post-transfection (Figure 2.22, red dots). By 9 hours post-transfection, mRNA level was substantially downregulated, indicating that the deadenylated RNAs were degraded. Consistent with the previous studies, the data indicate that miRNA induce deadenylation of the majority, if not all, of its targets.

I next compared poly(A) length with translation efficiency because it is generally considered that long poly(A) tail is required for effective translation. Unexpectedly, however, poly(A) lengths do not show any meaningful correlation with protein synthesis rates

[7] The preparation of cells and sequencing libraries in this experimental set were performed by Minju Ha.

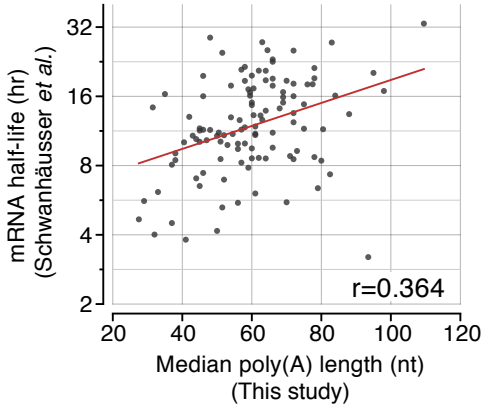



**Figure 2.20** Correlation between median poly(A) length and mRNA half-life, measured by Schwanhäusser et al. (2011). The *r* value refers to Pearson correlation coefficient, which is also applied to all the other scatter plots in this manuscript. mRNAs with more than 200 poly(A)$^+$ tags and with total length ranging from 3000 to 5000 nt were plotted due to the limited labeling of short RNAs in half-life measurement experiment.

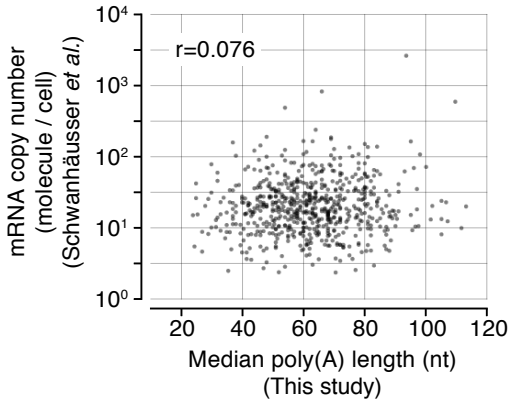



**Figure 2.21** Correlation between median poly(A) tail length and mRNA abundance (Schwanhäusser et al., 2011).

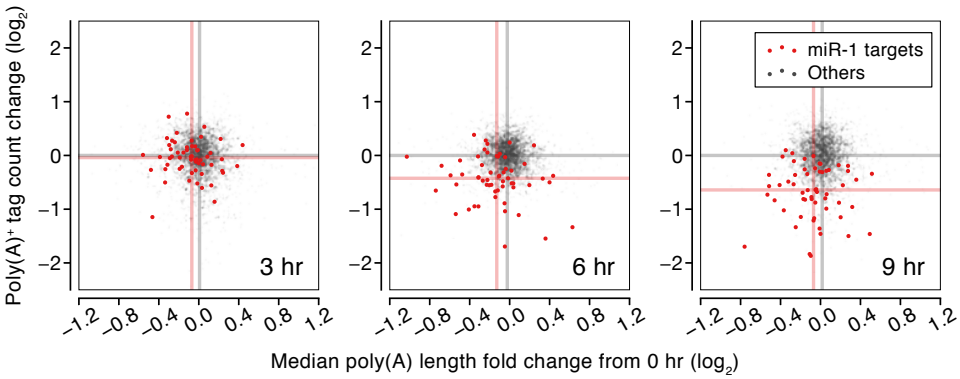



**Figure 2.22** Plots showing the changes of poly(A) tail lengths ($x$-axis) and number of poly(A)$^+$ tags ($y$-axis) after transfection of miR-1. Targets of miR-1 (red dots) are chosen from the list of mRNAs downregulated by more than 30% on 12 hr post-transfection in Guo et al. (2010). Gray dots represent the rest of transcripts. Mean changes are shown in vertical and horizontal lines. P-values from two-sided Mann-Whitney U tests between 26 targets and non-targets are: (3 hr) poly(A) $5.84 \times 10^{-4}$, tag count 0.473; (6 hr) poly(A) $1.87 \times 10^{-5}$, tag count $1.56 \times 10^{-14}$; (9 hr) poly(A) $6.63 \times 10^{-4}$, tag count $2.33 \times 10^{-20}$.

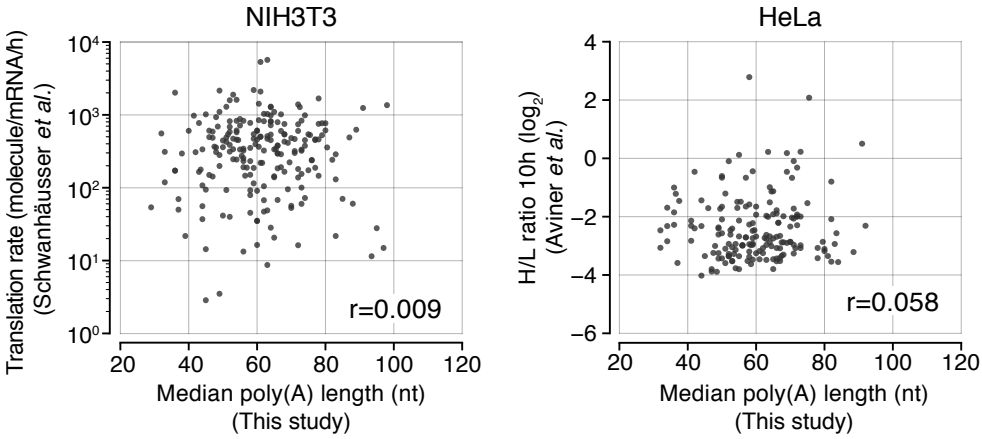



**Figure 2.23** Correlation between median poly(A) tail length and translation rate in NIH3T3, measured by Schwanhäusser et al. (2011), and HeLa cells, by Aviner et al. (2013). mRNAs with more than 200 poly(A)$^+$ tags and with CDS length ranging from 900 to 2,400 nt were plotted, considering the limited labeling of small proteins in translation rate measurement.

(measured by metabolic labeling and mass spectrometry and divided by mRNA abundance) (Figure 2.23; P=0.893 for NIH3T3, P=0.449 for HeLa, Pearson's correlation test). Similarly, when I compared poly(A) length with ribosome density that was determined by ribosomal footprinting (and divided by mRNA abundance) (Guo et al., 2010) (Figure 2.24), there was no detectable correlation, further supporting our conclusion. I did not find any significant correlation even when I used, in place of poly(A) length, the ratio between short and long poly(A) tails employing various lengths as a threshold (data not shown). These results suggest that deadenylation *per se* may not be directly coupled with translational suppression. It does not exclude a possibility, however, that deadenylation may affect translation indirectly and that translation of a subpopulation of mRNAs may be selectively affected by poly(A) length. Regulation of poly(A) tail may play a determining role under specialized conditions such as in neural synapses and early embryos where cytoplasmic polyadenylation is known to induce translation of dormant mRNAs with short tails (D'Ambrogio et al., 2013).

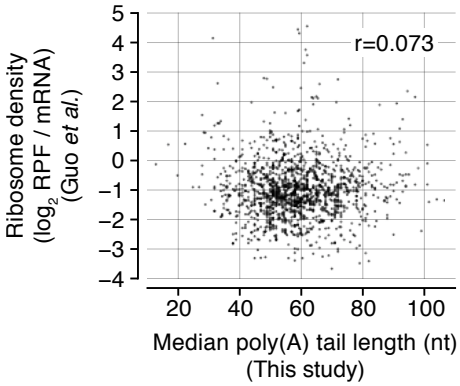



**Figure 2.24** Correlation between median poly(A) tail length and ribosome density (Guo et al., 2010). The *r* values indicate Pearson correlation coefficients.

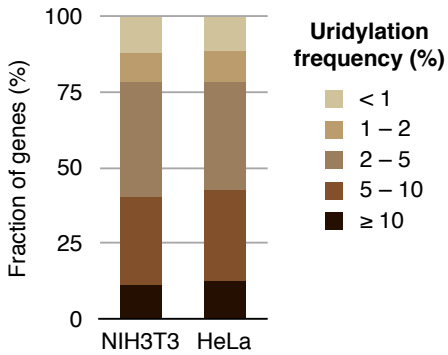



**Figure 2.25** Uridylation frequency of mRNA.

## 2.7 Analysis of 3′ end modification of poly(A) tails

One of the unique strengths of TAIL-seq is its ability to determine the sequences of the very end of RNA and to examine if there is any other sequences apart from simple poly(A) stretches. While looking at the 3′ ends of mRNA reads,[8] I found unexpectedly widespread uridylation in the downstream of poly(A) tail (Figures 2.15 and 2.25). This section describes about the terminal modifications at the 3′ end of poly(A) tails.

[8] All data for NIH3T3 and HeLa cells are derived from TAIL-seq libraries by Jaechul Lim.

### 2.7.1 Method for detection and filtering terminal modifications

As poly(A) tails were initially detected with a constraint that it must begin within the first 30 cycles, so the maximum detectable 3′ end modification of poly(A) tails was limited to the last 30 nucleotides of insert. To exclude A stretches obviously encoded from genomic sequence (with or without 3′ end modifications), I masked detected poly(A) tail ranges with read 2 alignments so that the 3′-most position of alignable (not clipped) is eliminated from poly(A) tail or its 3′ end modifications. All statistics regarding transcript-level modification rates were calculated for transcripts having more than 200 tags with poly(A) tails longer than 8 nt.

### 2.7.2 3′ Terminal uridylation of poly(A) tails

About half of mRNA species carry U-tails at more than 5% frequency; and ~80% of mRNA species are uridylated at a frequency higher than 2% (Figure 2.25). I observed a comparable pattern of uridylation in pilot experiments using a different 3′ adapter (data not shown).

Uridylation detected by TAIL-seq is reminiscent of the observations in fission yeast and *Arabidopsis* where mRNAs bear short U tails (1–2 Us), as analyzed by circularized RT-PCR (Rissland et al., 2007; Sement et al., 2013). Uridyl residues were found mainly on decapped mRNAs which represent decay intermediates and, when the uridylyl transferase (Cid1 in fission yeast) was mutated, mRNA was stabilized (Rissland & Norbury, 2009). These results collectively suggested that uridylation may be involved in mRNA decay in yeasts and plants. In mammals, there are only two known cases of mRNA uridylation. Histone mRNAs are oligo-uridylated, which is required for rapid decay at the end of S phase (Mullen & Marzluff, 2008; Schmidt et al., 2011). Additionally, the 5′ fragments from small RNA-directed cleavage are also uridylated in mammals and plants (Shen & Goodman, 2004). The current observation demonstrates that uridylation is much more pervasive in mammals than previously anticipated and that mRNA uridylation may be an integral part of a generic mRNA decay pathway that is conserved in all eukaryotes.

It is particularly interesting that uridyl residues are found mainly in mRNAs with short

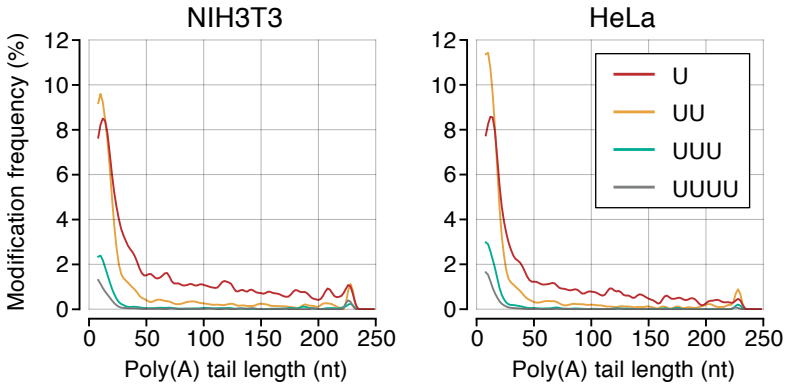



**Figure 2.26** Relationship between uridylation and poly(A) tail length. The density was calculated with 2 nt wide bins, then smoothened with Hanning window (width=7).

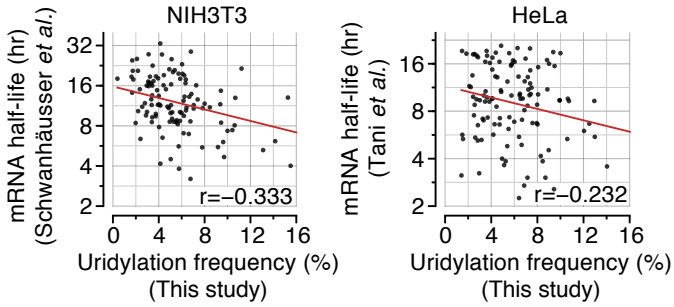



**Figure 2.27** Correlation between uridylation frequency and mRNA half-life (Schwanhäusser et al., 2011; Tani et al., 2012).

poly(A) tails (8–25 nt) (Figure 2.26). This phenomenon is similar to that in *Arabidopsis* where short U (1–2 nt) is added to 10–20 nt poly(A) (Sement et al., 2013). It was proposed that uridylation protects the 3′ end against further deadenylation and promotes decapping and 5′-3′ decay (Sement et al., 2013). In filamentous fungus *Aspergillus nidulans*, a mixture of uridyl and cytidyl residues are added to short poly(A) tails (Morozov et al., 2012) (~15 nt). Uridylation frequency shows a modest negative correlation with mRNA half-life (Figure 2.27), but not with mRNA abundance or translation rate (Figure 2.28). This is intriguing in light of recent reports showing that an oligo-U tail of mRNA serves as a decay marker by interacting with a 3′-5′ exonuclease Dis3L2 in yeast and human (Lubas et al., 2013; Malecki

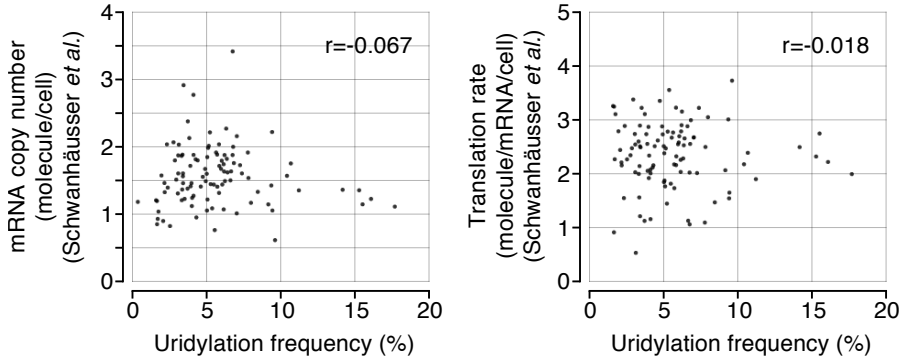



**Figure 2.28** Lack of strong correlation between uridylation frequency and mRNA abundance (left) or translation rate (right) in NIH3T3 as measured by Schwanhäusser et al. (2011). *r* values indicate Pearson correlation coefficients.

et al., 2013) and by recruiting LSM1-7 complex and decapping enzymes (Mullen & Marzluff, 2008; Rissland & Norbury, 2009). In future studies, RNAi of uridylyl transferases and nucleases can be combined with TAIL-seq, so as to elucidate the functional consequence and mechanism of uridylation and decay.

### 2.7.3 3′ Terminal guanylation of poly(A) tails

In addition to uridylation, it was surprising to discover yet another type of modification, that is, guanylation (Figure 2.29). About 20% of mRNA species are guanlylated at the downstream of poly(A) tail at a frequency of higher than 5%; and over 60% of transcripts show G-addition at more than 2% frequency (Figure 2.29). Guanylation was detected in our initial experiments using a different 3′ adapter (data not shown). To my knowledge, this is the first description of RNA 3′ guanylation although it was shown previously that some non-canonical poly(A) polymerases can utilize GTP in vitro (Bai et al., 2011; Heo et al., 2012). In contrast to U tails, terminal G residues are found selectively on longer poly(A) tails (>40 nt) (Figure 2.30). Cytidylation is considerably less frequent and does not show any preference for poly(A) tail size. Mono-guanylation is the prevalent form although the G residue is sometimes followed preferentially by C and subsequently by U

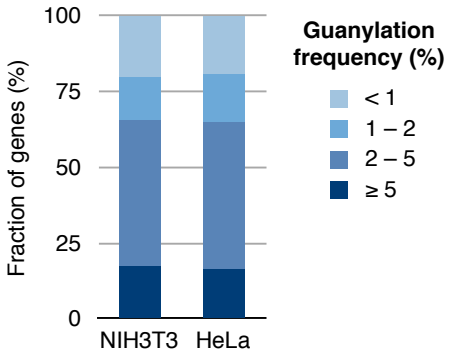



**Figure 2.29** Guanylation frequency of mRNA.

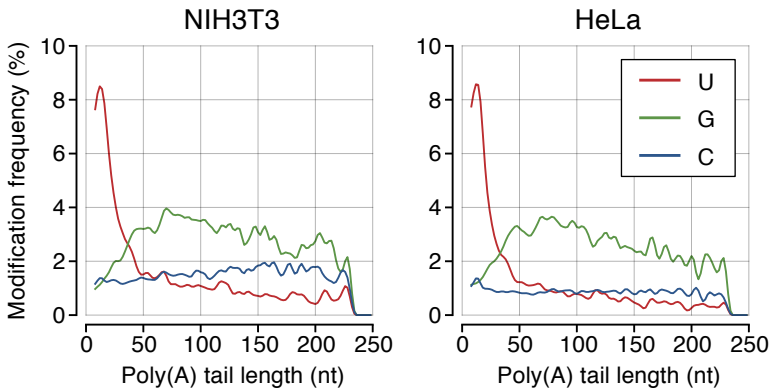



**Figure 2.30** Relationship between guanylation and poly(A) tail length. The density was calculated with 2 nt wide bins, then smoothened with Hanning window (width=7).

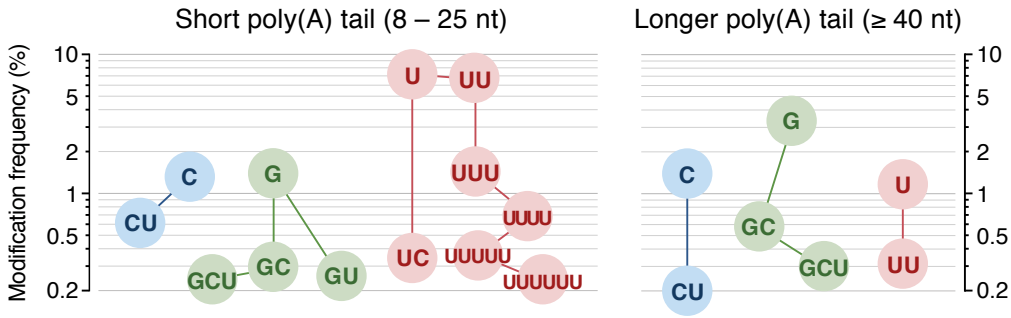



**Figure 2.31** Additional nucleotides attached to either short poly(A) tails (left panel) or longer poly(A) tails (right panel).

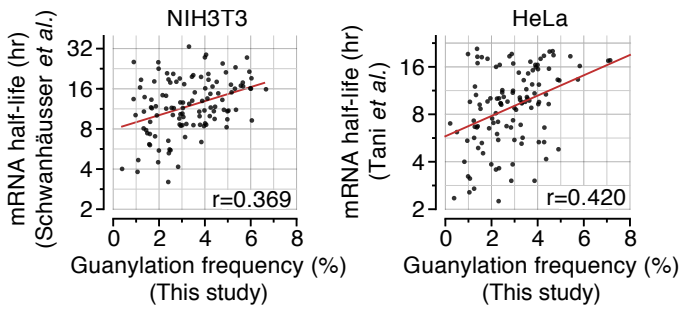



**Figure 2.32** Scatter plots showing the correlation between guanylation frequency and mRNA half-life (Schwanhäusser et al., 2011; Tani et al., 2012).

(Figure 2.31). Because deadenylases PARN and CCR4 are known to have a preference for terminal di-adenosines (Henriksson et al., 2010; Viswanathan et al., 2003) (AA), one can envision that the G addition may block deadenylation to protect mRNAs with long poly(A) tail. I indeed detect a modest positive correlation between guanylation frequency and mRNA half-life (Figure 2.32), but none between guanylation and mRNA level or translation rate (Figure 2.33). Although it would be too early to draw a conclusion, it is tempting to speculate that guanylation may stabilize mRNAs by antagonizing deadenylation. Not mutually exclusively, it is also plausible that G-tailed mRNAs may represent a specific subcellular location and/or a phase of mRNA life cycle.

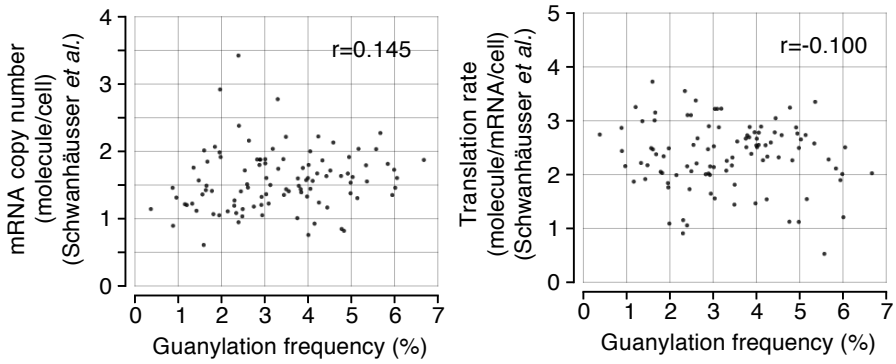



**Figure 2.33** Lack of strong correlation between guanylation frequency and mRNA abundance (left) or translation rate (right) in NIH3T3 as measured by Schwanhäusser et al. (2011). *r* values indicate Pearson correlation coefficients.

## 2.8 Detection of cleavage and polyadenylation sites

Using TAIL-seq data,[9] I could map the poly(A) sites although this was not our primary goal and the depth was lower compared to the other specialized tools developed previously (Beck et al., 2010; Mangone et al., 2010; Ozsolak et al., 2010; Yoon & Brem, 2010; Fu et al., 2011; Jan et al., 2011; Shepard et al., 2011; Derti et al., 2012; Martin et al., 2012; Hoque et al., 2013; Wilkening et al., 2013). In this section, I describe the unique potential of TAIL-seq that could not provided by the existing high-throughput methods for 3′ UTR mapping.

### 2.8.1 Method for polyadenylation site detection

I first selected poly(A)$^+$ tags with 12–20 nt poly(A) tails. The read 2 mappings in paired alignment were processed to remove the unmappable 3′ end modifications including poly(A) tails. Then, I surveyed the 3′ end frequency of genome-mappable spans from the trimmed alignments for all exonic positions of a transcript with 1,000 nt extension to downstream of the annotated 3′ end in RefSeq. The position with the most reads was chosen for the major polyadenylation site. When multiple positions have the same number

[9] All data for NIH3T3 and HeLa cells are derived from TAIL-seq libraries by Jaechul Lim.

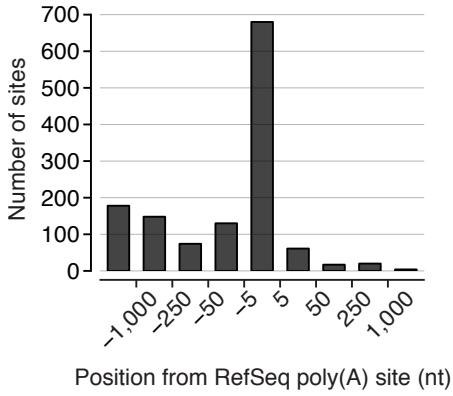



**Figure 2.34** Position of the poly(A) site identified by TAIL-seq, against the RefSeq annotation.

of reads, the 3′-most one was selected. I used the major polyadenylation sites supported by more than 5 reads.

### 2.8.2 Differential poly(A) tail lengths for alternative polyadenylation sites

When compared with the annotated poly(A) sites in RefSeq, the sites detected from our sequencing are significantly enriched at the annotated sites (Figure 2.34). Of note, the 3′ ends detected by TAIL-seq fall predominantly at the upstream of the annotated sites rather than the downstream. The upstream sites may correspond to alternative polyadenylation sites, considering that RefSeq often annotates the most distal sites (Pruitt et al., 2012). The sequences surrounding the detected poly(A) sites show characteristic features of known poly(A) sites (Figure 2.35), including the polyadenylation signal (PAS, AAUAAA and its variants), the U-rich upstream sequence element (USE) and downstream sequence element (DSE), indicating that TAIL-seq detects poly(A) sites accurately. I could also detect alternative polyadenylation (APA) in some genes (Figure 2.36). Notably, certain isoforms differ significantly in their poly(A) length and modification frequency, which is consistent with the notion that APA fundamentally influences mRNA fates (Elkon et al., 2013). For instance, I detected two alternatively processed isoforms from *Bclaf1* gene: one with small 3′ UTR carries a long poly(A) tail and relatively frequent guanylation

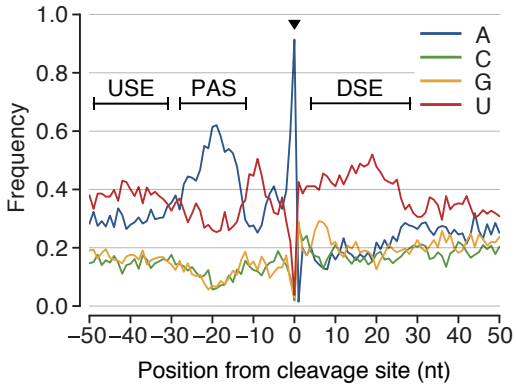



**Figure 2.35** Nucleotide composition of genomic sequences near the detected poly(A) sites. Sequence motifs such as PAS, USE, and DSE are enriched as shown previously (Beck et al., 2010; Mangone et al., 2010; Ozsolak et al., 2010; Yoon & Brem, 2010; Fu et al., 2011; Jan et al., 2011; Shepard et al., 2011; Derti et al., 2012; Martin et al., 2012; Hoque et al., 2013; Wilkening et al., 2013).

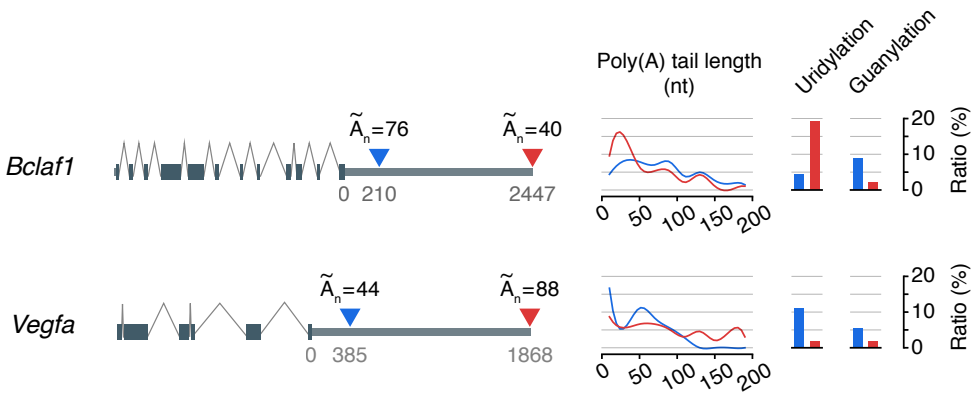



**Figure 2.36** Simultaneous detection of alternative poly(A) sites and their tail structures. $\tilde{A}_n$ refers to the median length of poly(A) tail. Poly(A) tail length distributions are counted in 20-nt wide bins, then shown after cubic spline interpolation.

while another isoform with extended 3′ UTR holds a shorter poly(A) tail with frequent uridylation.

## 2.9 Detection of RNA 3′ hydroxyl ends

Another line of valuable applications of TAIL-seq is to identify the substrates of specific ribonucleases.[10] Endonucleolytic cleavage sites are particularly interesting as they are involved in maturation of many important classes of RNA. This section describes the method and biological findings in search of RNA 3′ hydroxyl ends in the cell.

### 2.9.1 Methods for 3′ end detection

To find specifically enriched 3′ ends from TAIL-seq, we first calculated the frequency of mapped 3′ ends for all positions in the genome. I compared the number of 3′ ends mapped to a specific position (hotspot count, position 0) with the number of 3′ ends mapped to nearby positions within [-50, -2] and [2, 50] (flanking count) for all positions with positive number of 3′ ends. The list of detected 3′ hydroxyl ends were generated for all ends with no less than 10 hotspot tags as well as the number of hotspot tags are at least twice of flanking tags. The statistical significance of a specific hotspot was calculated using a binomial distribution ($p=\frac{1}{99}$ which is a probability when the 3′ ends are positions without preference; n=all 3′ ends mapped to hotspot and flanking region). The p-values from the distribution were adjusted for multiple testing by Bonferroni correction (n=the total length of genome). While the statistical significance was indicated in figures with asterisks and used for genomic source composition analyses, I did not limit the 3′ ends by the statistical significance in sequence motif to avoid sampling bias to highly abundant RNAs. The distance from the closest annotated 3′ ends were calculated against a union of all 3′ ends in NCBI RefSeq transcripts (Pruitt et al., 2012), UCSC known genes (Kuhn et al., 2013), ENSEMBL transcripts (Flicek et al., 2011), miRBase DROSHA cleavage sites (Kozomara

[10] All data for NIH3T3 and HeLa cells used in this section are derived from TAIL-seq libraries by Jaechul Lim.

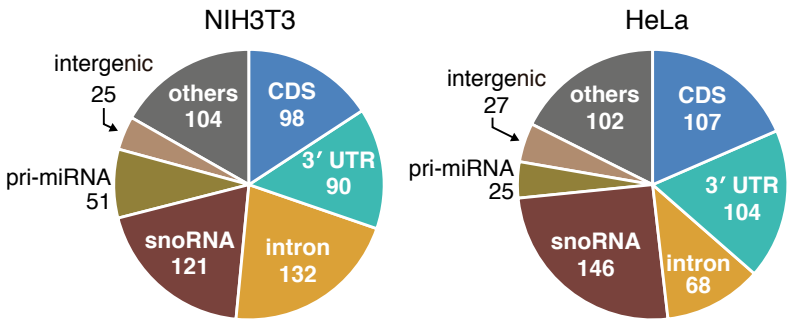



**Figure 2.37** Types of 3′ hydroxyl ends detected by TAIL-seq.

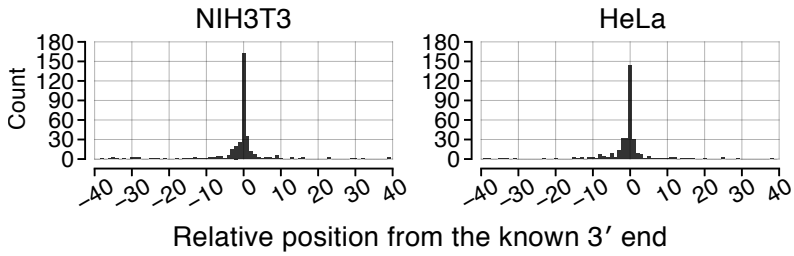



**Figure 2.38** Distribution of detected 3′ ends around the nearest known 3′ ends.

& Griffiths-Jones, 2011), gtRNAdb (Chan & Lowe, 2009) and NCBI RefSeq orthologous transcripts from other organisms (called xenoRefSeq in UCSC Genome Browser).

### 2.9.2 Comparison to the known 3′ ends in transcriptome

I find hundreds of sites from our library, which match to the 3′ ends of transcript sequences in databases. They belong to several distinct classes (Figure 2.37), including coding sequence (CDS), 3′ UTR, intron, and primary microRNA (pri-miRNA). Many of these were found almost exactly matching at the known 3′ end annotations (Figure 2.38). The confirming findings are mostly came from snoRNAs and histone mRNAs, however many of the rest were located in the middle of known transcripts (Figure 2.39). This may suggest transcriptome-wide evidences of endonucleolytic cleavage events. A notable class of such mechanism is pri-miRNA processing sites. I detected 45 sites in NIH3T3 and 22 sites in HeLa (Figure 2.39), which match precisely to the known DROSHA cleavage sites (Figure

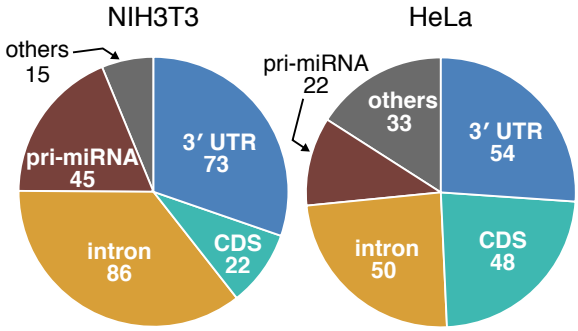



**Figure 2.39** Types of 3′ hydroxyl ends detected by TAIL-seq in the middle of known transcripts.

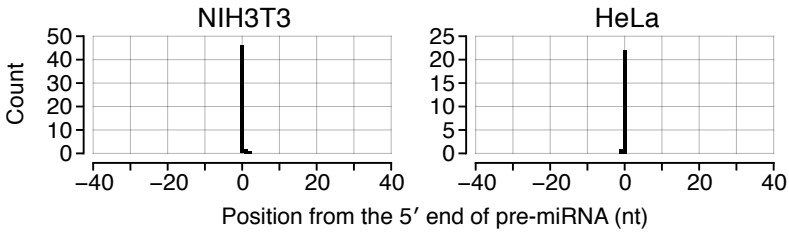



**Figure 2.40** Frequency of detected 3′ hydroxyl ends near DROSHA cleavage sites (5′ end of pre- miRNAs).

2.40). Figure 2.41 shows the miR-17~92 cluster as an example, where all six processing sites are detected by TAIL-seq. It is interesting that the 5′ fragments from DROSHA processing retain intact 3′ ends, suggesting that they may be relatively resistant to 3′-5′ trimming activities. Given that DGCR8 interacts with the basal part of the pri-miRNA hairpin (Han et al., 2006), it is plausible that DGCR8 remains bound to the 5′ product after cleavage reaction. The result suggests that TAIL-seq can be used to map the 5′ border of pre-miRNA even when mature miRNA from the 5p strand is not detected in small RNA sequencing (for instance, mmu-mir-29c, mmu-mir-496a, and hsa-mir-24-1). I can also identify pri-miRNAs that are alternatively processed by DROSHA at more than one site (mmu-mir-214). Alternative DROSHA processing is interesting as it yields multiple mature miRNAs with different targeting activities. Additionally, because it has been proposed that DROSHA may have additional substrates apart from pri-miRNAs, it will be interesting to

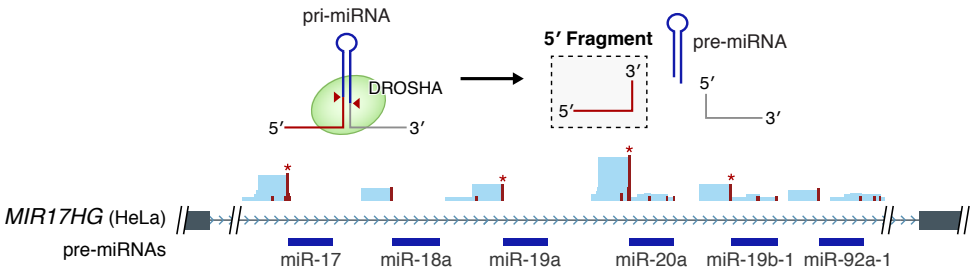



**Figure 2.41** (Top) Schematic illustration of DROSHA processing of pri-miRNA, generating a 5′ fragment (red line) that is detected by TAIL-seq. (Bottom) A histogram showing the tags from the miR-17~92 cluster in HeLa cells. Light blue area shows the accumulated coverage of the 3′-most 40 nucleotides of inserts while red bars indicate the frequency of the 3′ ends of the tags in log scale. Red asterisks mark statistically significant positions (Bonferroni-corrected p-value < 0.05).

search for unknown targets of DROSHA by using TAIL-seq.

### 2.9.3 Newly discovered 3′ ends

Lastly, I searched for putative nucleolytic sites that may be important for mRNA stability control. In this respect, TAIL-seq is complementary to previous degradome studies which mapped the 5′ end of RNA fragments containing 5′ phosphate and poly(A) tail (Addo-Quaye et al., 2008; German et al., 2008; Karginov et al., 2010; Shin et al., 2010). I found 95 and 102 internal sites in NIH3T3 and HeLa, respectively, from mRNA exons which may be potentially involved in mRNA destabilization through endonucleolytic cleavage (Figure 2.39). Figure 2.42 shows such examples where the 3′ ends of multiple tags come from a discrete position, indicative of specific endonucleolytic cleavage or stalled 3′-5′ exonucleolytic activity. Intriguingly, when we searched for a consensus sequence motif from such sites, we detected a trinucleotide motif enriched immediately upstream of the putative cleavage sites (Figure 2.43) that is composed of R (favoring A) - Y (U/C) - H (avoiding G), with the most frequent motif being 'AUU'. To our knowledge, no 3′-5 exonuclease is known to stall at a specific trinucleotide motif. Furthermore, we did not detect any significant secondary structure in the vicinity of the putative sites (data not

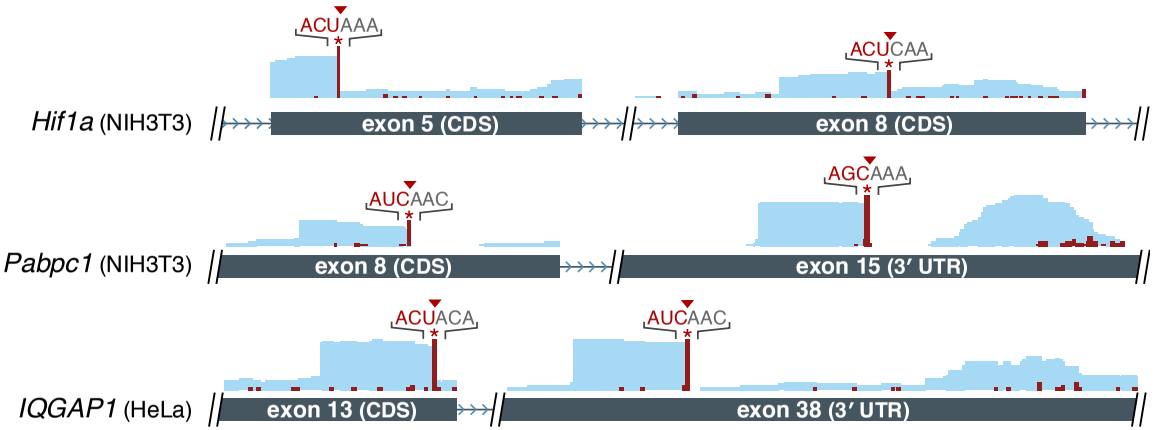



**Figure 2.42** Examples of putative endonucleolytic cleavage sites. Six nucleotides surrounding each statistically significant position are shown with an arrowhead marking the cleavage site.

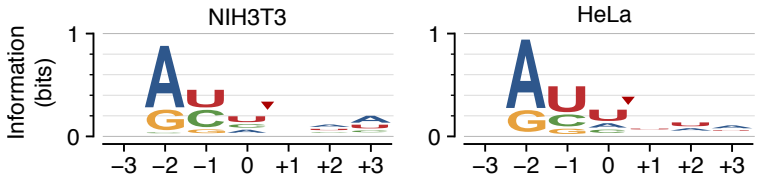



**Figure 2.43** Sequence logos showing enriched motif near the putative cleavage sites found in mRNA exons. Position 0 in $x$-axis indicates the 3′ end of the tag.

shown), which may block the progression of an exonuclease, suggesting that these sites may be targeted by a specific endonuclease. It awaits further investigation as to which factor(s) recognizes this motif and if the factor(s) constitutes a novel pathway for mRNA stability control.

## 2.10 Discussion

TAIL-seq is the first method that allows global survey of poly(A) length and 3′ end modification of mRNA. In designing the current version of TAIL-seq, we[11] aimed to be as comprehensive as possible, which allowed us to discover many new exciting features such as differential poly(A) length control, uridylation, guanylation, and RNA cleavage. There is ample information in TAIL-seq datasets, which remains to be analyzed in future studies. For instance, TAIL-seq determines the 3′ ends of histone mRNAs, post-splicing introns, and various types of noncoding RNAs, which will be interesting subjects to investigate. Because the current version of TAIL-seq covers many classes of RNAs and many different types of modification, it is inevitable that TAIL cannot provide sufficient depth to all the detected features. To study particular types of 3′ ends in greater depth at lower cost, the technology will need to be modified further so as to generate more focused libraries. TAIL-seq is indeed a highly amenable technology that can be modified easily. For instance, one can change the range of size fractionation and/or use RNA extracted from subcellular fractions and immunoprecipitates to enrich for a selective class of RNA.

[11]The current study is designed and performed in collaboration with Jaechul Lim and Prof. V. Narry Kim.

This study raises numerous open questions. It will be of great interest to identify the protein factors involved in each processing and modification discovered in this study, and to understand their mechanisms and functions. To this end, TAIL-seq, combined with systematic RNAi, will serve as a valuable tool. TAIL-seq will also be useful to solve various general issues regarding the relative dynamics of mRNA deadenylation, translation, and decay. In addition, one can examine RNA terminal modifications in diverse physiological and pathological contexts, such as in neural synapse, late oogenesis, early embryogenesis, cellular senescence and inflammation where dynamic control of cytoplasmic polyadenylation is known to play a critical role. The TAIL-seq protocol can be applied to any species and cell types with minor modifications, which will greatly expand the initial observations made in this study.

# 3. Analysis of RNA-protein interactions by high-throughput sequencing

## 3.1 Background

Since the introduction of the HITS-CLIP (also known as CLIP-seq)[1] technique by Licatalosi et al. (2008), it has become one of the most favored methods to gain the transcriptome-wide view of *in vivo* RNA-protein interactions. For the preparation of a CLIP-seq sequencing library, RNA and protein are crosslinked by ultraviolet light in the cell, then the RNA-protein complex is immunopurified following RNase digestion (Licatalosi et al., 2008). The protein portion of the purified complexes is removed by treating non-specific protease so that remaining RNA can be converted to DNA and sequenced in a high-throughput sequencer.

There are variations of the technique called PAR-CLIP and iCLIP. The former uses photo-activatable ribonucleotides such as 4-thiouridine (4SU) or 6-thioguanosine (6SG) under 360 nm UV-A instead of 253 nm UV-C light (Hafner et al., 2010) (Figure 3.1). The another variation called iCLIP delays 5′ adapter ligation to RNA to follow after reverse-transcription so that footprints from incompletely reverse-transcribed cDNAs by physical interference of residual peptide on RNA are captured in the library (König et al., 2010) (Figure 3.2).

Notwithstanding that almost a hundred of studies using any CLIP technique have been published thus far, there is no established standard workflow for the analysis of its data. Nearly every study that utilizes a CLIP-seq technique has designed its own analytic

[1] Both *HITS-CLIP* and *CLIP-seq* are widely used to describe the identical technology with comparable frequency at the time of writing. In this thesis, I will use *CLIP-seq* as an umbrella term that includes *HITS-CLIP* and its all variants.

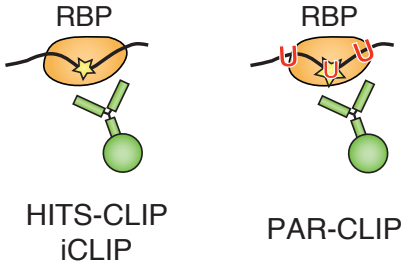



**Figure 3.1** Crosslinking strategy of three CLIP techniques. HITS-CLIP and iCLIP use 253 nm UV light without nonnatural nucleic acid replacements. PAR-CLIP requires cell culture with 4SU or 6SG before 360 nm UV irradiation. Unlike the other variants, PAR-CLIP has limited crosslinking repertoire to U or G depending on the culture medium.

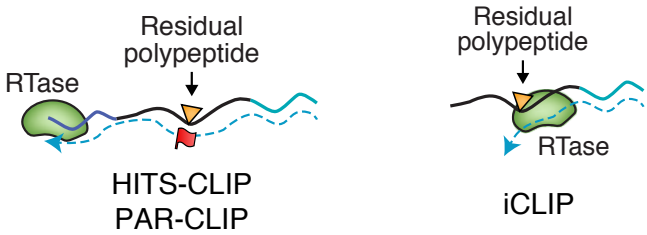



**Figure 3.2** Detection strategy of three CLIP techniques. HITS-CLIP and PAR-CLIP use both tag enrichment level compared to RNA-seq or neighboring positions, and sequence changes accumulated in narrow region of reads. iCLIP detects the accumulated 5′ ends, which are assumed as result of premature termination of reverse-transcription.

methods. In this chapter, I provide an analysis toolchain generally applicable to wide range of HITS-CLIP or PAR-CLIP experiments.[2] Then, I compare dozens of publicly available CLIP-seq datasets and show the results from meta analyses, for the first time for CLIP-seq experiments.

## 3.2 Reference data preparation

For the generalization of data processing and analyses, data from eighty experiments in twenty studies were downloaded from NCBI Sequence Read Archive (SRA) or NCBI Gene Expression Omnibus (GEO) (Table 3.1). The list covers diverse scope of RNA-binding proteins including splicing factors, cleavage and polyadenylation factors, post-transcriptional processors, and translational regulators. It also includes several experiments from PAR-CLIP to compare the different crosslinking techniques.

### 3.2.1 Sequence processing and alignment

The first few steps in sequence analysis were done by using Assaf Gordon's FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). First, the 3′ adapter sequences were removed from reads by using fastx_clipper. The rest was trimmed from the 3′ end so that the remaining reads have Phred quality of 25 or higher. After clipping and trimming, reads of 20 nt or longer were collapsed to generate a set of unique sequences. The sequences were aligned to abundant contaminant sequences (Illumina adapter/primer sequences and ribosomal DNA complete repeating unit, GenBank accession BK000964.1 for mouse and U13369.1 for human) with GSNAP version 2013-03-31 (Wu & Nacu, 2010) with 10% mismatch rate. Filtered reads that do not match to any contaminant and have sufficient sequence complexity (Shannon entropy, at least 0.7 for mononucleotide, 1.5 for dinucleotide) were aligned to the UCSC Genome Browser hg19 (human) or mm10 (mouse) genome assembly with GSNAP version 2013-03-31 (Wu & Nacu, 2010) with options of 10% mismatch rate, no

[2] The software written for this study is released under the MIT license on a *github* repository (http://github.com/hyeshik/ecliptic).

**Table 3.1** List of publicly available datasets used in this study. The accession number starting with "SRA" was downloaded from NCBI Sequence Reads Archive (SRA) (Wheeler et al., 2008), and "GSE" accessions were downloaded from NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2013). When additional RNA-seq libraries are available in the experiment set, they are indicated with *nt* (no treatment), *ctl* (control treatment), *kd* (knock-down), or *exp* (over-expressed). More rows are followed in the next page.

| Accession | Protein(s) | Source | RNA-seq | Technology | Reference |
|---|---|---|---|---|---|
| SRP002550 | NOVA | mouse brain | none | HITS-CLIP | Licatalosi et al. (2008) |
| GSE19323 | PTB | HEK293T | none | HITS-CLIP | Xue et al. (2009) |
| GSE21918 | IGF2BP1, IGF2BP2, IGF2BP3, PUM2, QKI, AGO1, AGO2, AGO3, AGO4, TNRC6A, TNRC6B, TNRC6C | HEK293 | none | PAR-CLIP | Hafner et al. (2010) |
| GSE23694 | HNRNPH | HEK293T | none | HITS-CLIP | Katz et al. (2010) |
| GSE34491 | HNRNPU | HeLa | ctl, kd | HITS-CLIP | Xiao et al. (2012) |
| GSE35800 | HNRPA2B1 | MDA-MB-231 | none | HITS-CLIP | Goodarzi et al. (2012) |
| GSE36987 | PAPD5 | HEK293 | none | PAR-CLIP | Rammelt et al. (2011) |
| GSE37114 | LIN28A | mESC | nt, ctl, kd | HITS-CLIP | Cho et al. (2012) |
| GSE37524 | MOV10 | HEK293 | nt | PAR-CLIP | Sievers et al. (2012) |
| GSE37685 | CIRP | NIH3T3 | ctl, kd | HITS-CLIP | Morf et al. (2012) |
| GSE39086 | DGCR8 | HEK293T | none | HITS-CLIP | Macias et al. (2012) |

**Table 3.1** (continued from the previous page)

| Accession | Protein(s) | Source | RNA-seq | Technology | Reference |
|---|---|---|---|---|---|
| GSE39686 | FMR1, FXR1, FXR2 | HEK293-derived | nt | PAR-CLIP | Ascano et al. (2012) |
| GSE39872 | LIN28A | H9, HEK293-derived | nt, ctl, kd | HITS-CLIP | Wilbert et al. (2012) |
| GSE39911 | MBNL | mouse brain, C2C12 | ctl, kd | HITS-CLIP | Wang et al. (2012) |
| GSE40651 | mTDP43, mFUS/TLS, hFUS/TLS | mouse brain, human brain | ctl, kd | HITS-CLIP | Lagier-Tourenne et al. (2012) |
| GSE40778 | eIF4AIII | HeLa | none | HITS-CLIP | Saulière et al. (2012) |
| GSE42398 | CSTF64 | C2C12 | none | HITS-CLIP | Hoque et al. (2013) |
| GSE42701 | PTB, AGO2 | HeLa | none | HITS-CLIP | Xue et al. (2013) |
| GSE44616 | LIN28B | HEK293 | ctl, exp, kd | PAR-CLIP | Hafner et al. (2013) |
| GSE45148 | FMRP | mouse brain | none | HITS-CLIP | Darnell et al. (2011) |

terminal clipping, and splice site annotations from RefSeq (downloaded from the UCSC Genome Browser on Jan 24, 2013). When an RNA-seq library from the same source is available from the dataset, the sequence alignments from RNA-seq were processed to call single nucleotide polymorphisms (SNPs) with samtools 0.1.19 (Li et al., 2009) (minimum depth=20, substitution mutation rate=0.001, adjusted p-value cutoff=0.01). Since mutations from UV irradiation is indispensable from CLIP experiments, it is preferable to have variant-aware alignments to increase sensitivity near protein interacting sequences, where more mutations are accumulated, and reduce false positives from background variants from tissue or cell line themselves. The sequence reads with related SNP data were aligned with variant-aware indices built with gmap snpindex (Wu & Nacu, 2010). Finally, the alignment results were filtered to leave only the single best hit with minimum edit distance (up to two edits) to obtain a set of single-hit reads. Those with multiple best hits were ignored as repetitive sequences.

### 3.2.2 Sequence annotation and classification

The alignments were annotated with RefSeq (Pruitt et al., 2012), RepeatMasker, miRBase release 18 (Griffiths-Jones, 2010), Rfam (Burge et al., 2013), and GtRNAdb (Chan & Lowe, 2009) by using intersectBed of BEDTools (Quinlan & Hall, 2010). A representative class for a given read was determined as the first matching class from all annotations for all alignments for the read in the following priority: miRNA, rRNA, tRNA, Mt-tRNA, snoRNA, scRNA, srpRNA, snRNA, RNA, ncRNA, misc_RNA, Cis-reg, ribozyme, RC, IRES, frameshift_element, LINE, SINE, Simple_repeat, Low_complexity, Satellite, DNA, LTR, CDS, 3′ UTR, 5′ UTR, intron, Other, Unknown. The annotated representative classes were combined with read counts of previously removed sequences in the first contaminant filtration, and used for CLIP tag classification statistics. For subsequent analyses, the reads classified as rRNA or tRNA were excluded and the rest was used. The alignments for filtered reads were converted to bam format and visualized with the UCSC Genome Browser. The non-redundant RefSeq transcription set was constructed by the identical procedure described in Section 2.3.3.

## 3.3 Binding site detection

Identification of binding sites is usually required before any further analysis of CLIP data. Peak calling of mapped CLIP tags in RNA was used in earlier studies (Licatalosi et al., 2008; Chi et al., 2009). PAR-CLIP and iCLIP can identify RNA-protein interactions in single nucleotide resolution by using T to C transitions and clustered 5′ end positions thanks to biochemical characteristics of their libraries (Hafner et al., 2010; König et al., 2010). This also became possible for HITS-CLIP libraries by crosslinking-induced mutation sites (CIMS) introduced by Zhang & Darnell (2011).

When the list of confident binding sites is prepared, discovery of *cis*-regulatory element or the protein's substrate specificity is a unique benefit of the high resolution of CLIP techniques. The enriched sequence motifs can be easily visualized by simple sequence logo analyses (Crooks et al., 2004; O'Shea et al., 2013) where binding sites were identified in the single nucleotide resolution. Otherwise, statistical overrepresentation of sequences can be used to identify the recognized motif of a protein (Yeo et al., 2009). This section introduces new metrics developed for RNA-protein interactions inducing more substitution errors than deletion errors, and shows how the metrics can be applied to array of proteins.

### 3.3.1 Metrics for crosslinking-induced errors

Although substitution errors[3] are once described near RNA-protein binding sites after UV crosslinking and reverse-transcription (Granneman et al., 2009). So far, the only systematic effort to use the accumulated errors in sequence alignments was made by Zhang & Darnell (2011) for deletion errors on Argonaute proteins and NOVA. In my recent study for LIN28A with Jun Cho (Cho et al., 2012), I found that the protein makes more substitutions than deletions like Nop58 does in *S. cerevisiae* (Granneman et al., 2009). In addition to the

[3] The original term *mutation* used by Kishore et al. (2011) and Zhang & Darnell (2011) may confuse readers to understand that the RNA-protein interactions induce inheritable sequence replication errors. Since the *mutations* are artifacts from UV irradiation during CLIP experiments, I will use *errors* instead with a sense that any mismatch between the reference and sequence reads such as substitution (modification), insertion, or deletion.

previous existing metric, deletion rate, I included more metrics into the regular analysis of crosslinking-induced errors: substitution rate, insertion rate, and substitution and deletion rate.

As a matter of fact, the simple metrics have common false positives coming from background sequence variations. Zhang & Darnell (2011) avoided the problem by ignoring positions with deletion rates higher than 0.9. The workaround is simple and powerful for many cases although it cannot handle heterozygous SNPs. It becomes more problematic in studies using cancer cell lines because they often carry aneuploidy and allele frequency ratio can be virtually any number. Fortunately, RNA-protein interactions often induce substitution errors in non-uniform type of nucleotide changes. Unlike the RNAs including SNPs, which usually have only two types of nucleotides in a site, crosslinking-induced errors generally induce all three types of substitutions and deletions (Figure 3.3). To use this additional information, I introduced Shannon entropy (Shannon, 1948) as a metric for SNP-proof detection of crosslinking-induced errors:

$$C = -\sum_{n} p_n \log_2 p_n$$

where $C$ is the crosslinking-induced error score and $n$ is any type of nucleotide including D for deletion.

### 3.3.2 Error characteristics of different RNA-binding proteins

An earlier study revealed that type of crosslinking-induced errors are different from protein to protein even when the experimental conditions are the same (Granneman et al., 2009). NOVA and Argonautes in mouse (Zhang & Darnell, 2011), Nop1 and Nop56 in *S. cerevisiae* (Granneman et al., 2009), and hnRNP C in human (Sugimoto et al., 2012) are known to induce deletion errors by UV crosslinking more often than in RNA-seq. On the contrary, crosslinking-induced errors by mouse LIN28A is more biased into substitutions (Figure 3.4). Nop58 in *S. cerevisiae* (Granneman et al., 2009) and HuR in human (Kishore et al., 2011) follow the rank of substitution-favoring mode of interactions.

In my meta analysis, most CLIP experiments including both HITS-CLIP and PAR-

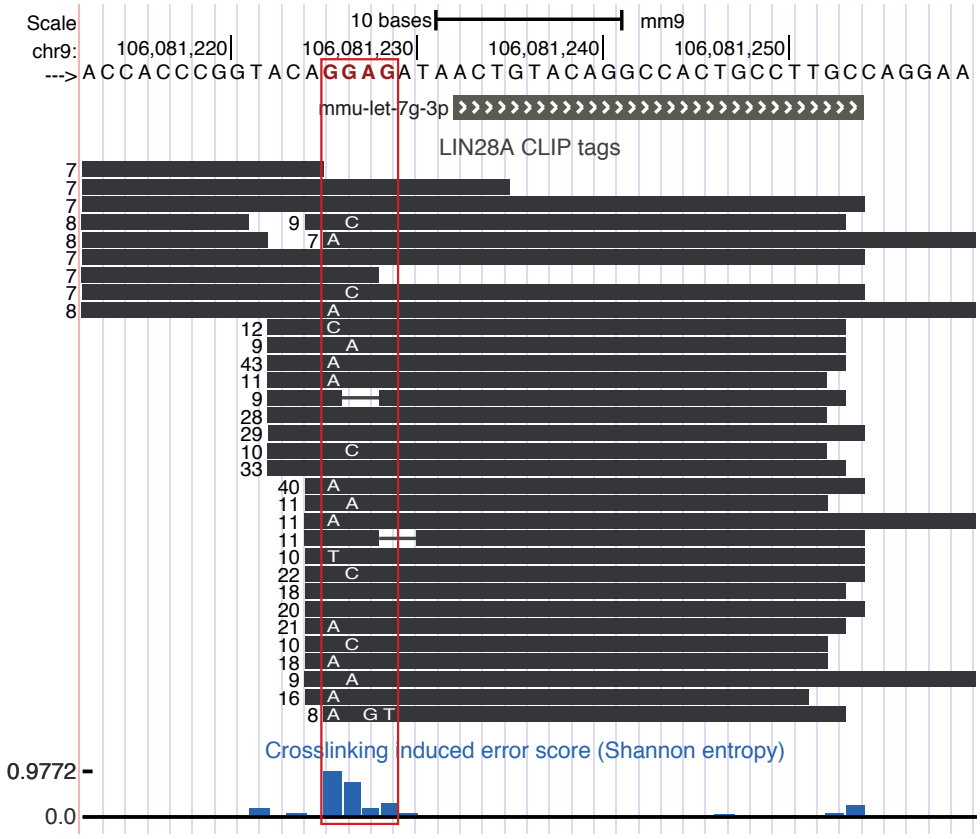



**Figure 3.3** Sequences from a set of HITS-CLIP libraries showing chaotic substitution and deletion errors near expected binding sites (*Mirlet7g* locus from Cho et al. (2012)). The previously known binding site of the protein, the GGAG motif in the terminal loop of precursor let-7g, is marked with a red box. Each unique sequence is represented by a black horizontal bar with the number of reads indicated on the left. Mismatched sequences are shown in white letters. UV crosslinking frequency is quantified by using Shannon entropy and is shown at the bottom with blue bars. Less frequent tags (<7 reads) are omitted to improve visibility.

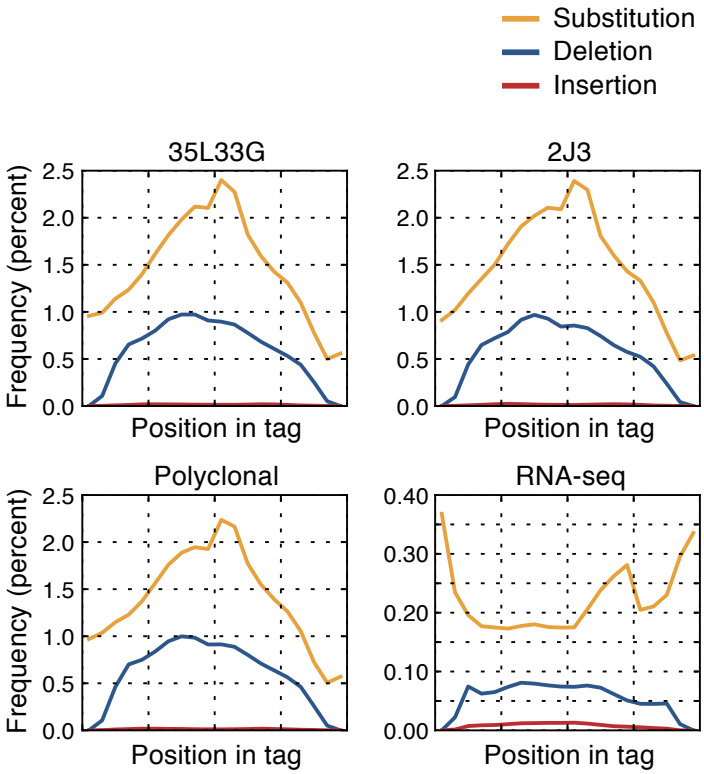



**Figure 3.4** LIN28A example of error frequency profiles as a function of position along the CLIP tags (Cho et al., 2012). Position within the tag was partitioned into 20 bins with the 5′ end of the reads as the leftmost bin ($x$-axis). To avoid underestimation of errors at both ends, I replaced the sequences removed by terminal soft clippings with the original sequences obtained from sequencing. In the case of insertion errors, I assumed that the errors occur only at the left-side of a given base.

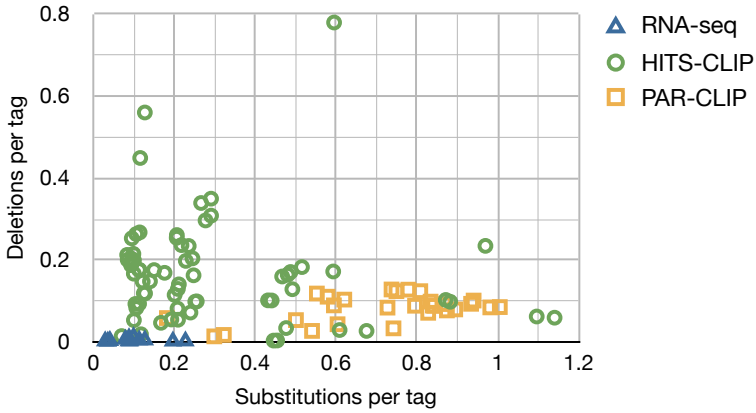



**Figure 3.5** Frequencies of substitution and deletion errors in RNA-seq, HITS-CLIP, and PAR-CLIP libraries.

CLIP showed not only deletion errors but also substitution errors (Figure 3.5). PAR-CLIP generally induced more substitutions as expected (T to C transition), however few HITS-CLIP libraries induced more substitutions than those of PAR-CLIP (Figure 3.5). These were DGCR8, PAPD5 (also known as TRF4-2), and PTB proteins, however it was hard to find the shared factors that distinguishes them from the others. As the meta dataset have several libraries that shares some features, I compared them by contrasting differences (Figures 3.6, 3.7, and 3.8). Presumably, it is confirmed that even if CLIP experiments performed together in a study, the tendency of making substitutions or deletions are unique to the identity of protein (Figure 3.6). Moreover, the substitution preference of a protein was similar in spite of that experiments were performed in different species by different investigators (Figure 3.7). There were, however, significant experimental variances among replicates by same investigators when experimental procedure is changed (Figure 3.8).

### 3.3.3 Statistical analysis of crosslinking-induced errors

Systematic downstream analyses requires statistical significance of detected binding sites. RNA-seq and CLIP libraries are heavily biased by different experimental artifacts. For example, RNA fragmentation is generally known for inducing strong biases under the

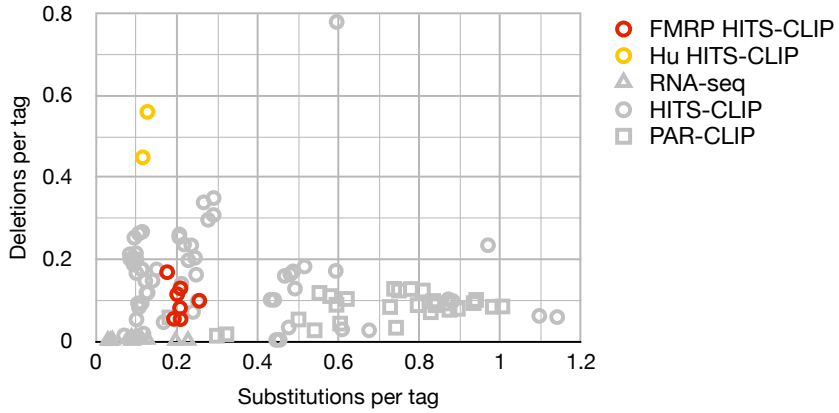



**Figure 3.6** Frequencies of substitution and deletion errors alternatively colored to contrast different samples in an experiment by Darnell et al. (2011).

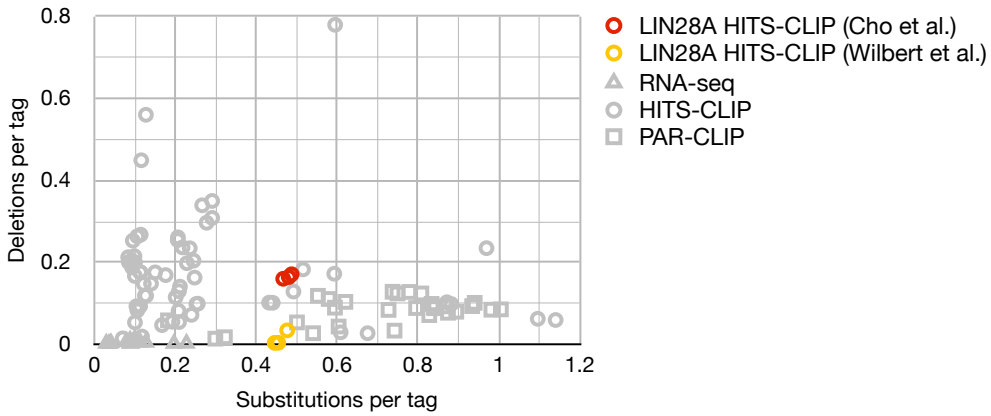



**Figure 3.7** Frequencies of substitution and deletion errors alternatively colored to contrast independent trials for a homologous protein in different species by different investigator (Cho et al. (2012) in mouse embryonic stem cells and Wilbert et al. (2012) in human embryonic stem cells).

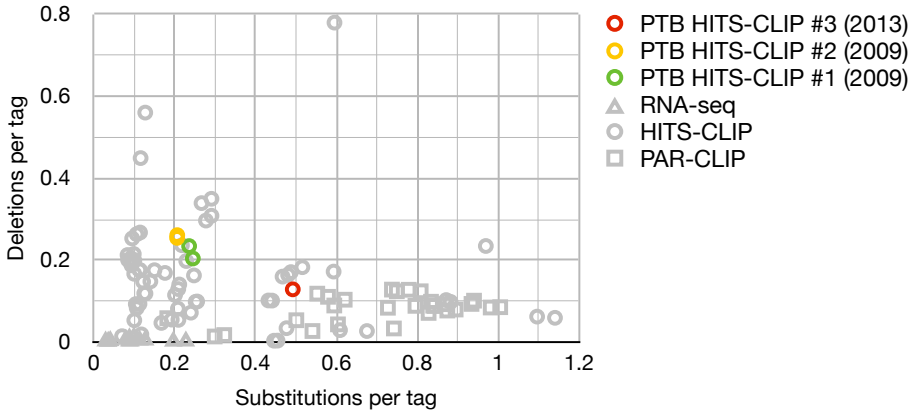



**Figure 3.8** Frequencies of substitution and deletion errors alternatively colored to contrast different trials for same protein and cells by same investigators (Xue et al., 2009, 2013).

effects from RNA secondary structure and sequence composition (Roberts et al., 2011). However, the fragmentation bias in CLIP-seq is totally different from RNA-seq's because RNA-protein complex maintains characteristic molecular structure in solution while RNAs are free from proteins in RNA-seq. Moreover, it is extremely hard to estimate background distribution from neighboring positions in CLIP-seq due to the dispersed nature of most RNA-protein interactions and broad range of mRNA expression levels. Lack of appropriate controls makes harder to model the statistical background distribution of the crosslinking-induced error metrics.

Zhang & Darnell (2011) developed a method for permutation-based estimation of statistical significance of binding sites, called crosslinking-induced mutation sites (CIMS). It estimates the background distribution of deletion rates by switching the deleted bases with randomly chosen base of the same position in read and reference nucleotide in genomic sequence in the other sequence reads (Zhang & Darnell, 2011). While maintaining the basic ideas, I reformed the algorithm for improved scalability and better statistical performance (Algorithms 3.1 and 3.2). Its optimized implementation is included in the toolchain supplementary to this chapter.[4] With the new algorithm and implementation, I evaluated the new metrics introduced in the previous section using HITS-CLIP data for

[4] All codes for this implementation are available from https://github.com/hyeshik/ecliptic.

**Algorithm 3.1** Simplified procedure of one iteration for permutation-based background distribution estimation of crosslinking-induced error metrics. Actual implementation uses slightly altered order for multi-threading, and uses splay tree (Sleator & Tarjan, 1985) for *results* lists for the optimized use of memory.

```
procedure PermutateOnce(readseqs)
  readqueues ← an empty queue for each base (A, C, G, T, D)
  results ← an empty list for each base and read depth levels

  for every sequence in readseqs,
    for every base in sequence,
      append base in read to readqueues[reference base]

  for every queue in readqueues,
    shuffle the queue with Fisher-Yates shuffling

  for every unique alignment in readseqs,
    for i ← from 0 to length of reference of the alignment,
      readcount ← five zeros (for A, C, G, T, D)
      for j ← from 0 to number of duplicated reads of the alignment,
        r ← pop an element from readqueues[reference base][i]
        increase readcount[r] by 1
      end for

      value ← calculate a metric with the readcount
      append value to results[reference base, depth]
    end for
  end for

  return results
```

**Algorithm 3.2** Simplified procedure for setting cutoff that meets given level of FDR.

```
procedure SetCutoffForFDR(real values, permutated values, fdr)
  sort real values in descending order
  sort permutated values in descending order

  match 1 by 1 to perfectly align real values and permutated values
    by adding zeros to either lists

  real_cum ← get cumulative array of real values
  permutated_cum ← get cumulative array of permutated values

  valid_cutoff ← not a number

  for every element in real_cum and permutated_cum,
    if real or permutated count is zero,
      continue to next set of elements /* zero division */

    fdr_calculated ← permutated cumulative count in fraction,
                     divided by real cumulative count in fraction

    /* update valid_cutoff to lower value when satisfy the fdr */
    if fdr_calculated < fdr,
      valid_cutoff ← cutoff value for the current element
  end for

  return valid_cutoff
```

**Table 3.2** Results from the assessment for detection performance of various crosslinking-induced error metrics. Each of 35L33G, 2J3, and 46020 is an experiment using different antibody in mESC LIN28A HITS-CLIP (Cho et al., 2012), and represent respectively each experiment using the antibody here. Number of sites (# sites) shows number of detected binding sites with $\geq$ 50 tags and $<$ 0.01 false discovery rate. AUC values are the areas under the curve in receiver operating characteristics (ROC) curves with an assumption that LIN28A binds only to GGNG or GNG.

| Experiment | 35L33G | | 2J3 | | 46020 | |
|---|---|---|---|---|---|---|
| Metric | # sites | AUC | # sites | AUC | # sites | AUC |
| Deletion | 37,739 | 0.430 | 38,205 | 0.433 | 16,628 | 0.420 |
| Substitution | 50,634 | 0.762 | 53,400 | 0.767 | 50,164 | 0.759 |
| Del. + Subst. | 63,041 | 0.737 | 65,244 | 0.748 | 51,921 | 0.726 |
| Shannon entropy | 46,522 | 0.790 | 47,707 | 0.797 | 39,798 | 0.789 |

LIN28A (Cho et al., 2012) (Table 3.2). Expectedly, deletion errors were not a powerful indicator to detect binding sites of LIN28A (Table 3.2). Substitution rate detected more binding sites with the same level of FDR, but Shannon entropy turned out to be modestly better at picking up the targets with known binding preference. I also compared the error-based crosslinking detection with different metrics against enrichment-based peak callers (Table 3.3). CIMS methods with substitution or Shannon entropy detected binding sites from let-7a-1 *(Mirlet7a-1)* and mir-98 *(Mir98)* loci in addition to the other targets called by Piranha (Uren et al., 2012) or ASPeak (Kucukural et al., 2013) (Table 3.3). However, CIMS performed worse than the enrichment-based analysis methods for experiments with lower depth of reads or lower substitution or deletion rate by UV crosslinking (data not shown). UV crosslinking between LIN28A and its target induced informative substitutions as well as deletions unlike NOVA or Argonautes (Zhang & Darnell, 2011). The more degree of freedom in substitution errors seems to contribute to the better sensitivity of binding target detection in this case.

**Table 3.3** False discovery rates calculated using different approaches to detect crosslinked sites in CLIP. The experiments are from a HITS-CLIP study for LIN28A protein in mESC (Cho et al., 2012). The binding partners of LIN28A shown here are the exhaustive list of let-7 precursors which are well-studied to bind LIN28A (Heo et al., 2009). The values from CIMS are shown in maximum FDRs of the most conservative estimates. FDRs from Piranha (Uren et al., 2012) and ASPeak (Kucukural et al., 2013) are shown in multiple-testing corrected values with the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). '–' indicates the target is not detected using the method.

| Experiment | Target | CIMS subst. | CIMS del. | CIMS entropy | Piranha | ASPeak |
|---|---|---|---|---|---|---|
| 35L33G | let-7a-1 | $< 1 \times 10^{-4}$ | – | $< 1 \times 10^{-3}$ | – | – |
| | let-7d | $< 1 \times 10^{-4}$ | $< 1 \times 10^{-4}$ | $< 1 \times 10^{-4}$ | $3.9 \times 10^{-4}$ | 0.0433 |
| | let-7f-1 | $< 1 \times 10^{-4}$ | – | $< 1 \times 10^{-4}$ | $4.3 \times 10^{-3}$ | $9.0 \times 10^{-4}$ |
| | let-7g | $< 1 \times 10^{-4}$ | – | $< 1 \times 10^{-4}$ | – | $9.0 \times 10^{-4}$ |
| | mir-98 | $< 1 \times 10^{-4}$ | – | $< 0.05$ | – | – |
| 2J3 | let-7a-1 | – | – | – | – | – |
| | let-7d | $< 1 \times 10^{-4}$ | $< 1 \times 10^{-4}$ | $< 1 \times 10^{-4}$ | $2.8 \times 10^{-7}$ | $1.2 \times 10^{-8}$ |
| | let-7f-1 | $< 1 \times 10^{-4}$ | – | $< 1 \times 10^{-4}$ | $1.4 \times 10^{-8}$ | $2.6 \times 10^{-7}$ |
| | let-7g | $< 1 \times 10^{-4}$ | – | $< 1 \times 10^{-4}$ | – | – |
| | mir-98 | $< 1 \times 10^{-4}$ | $< 1 \times 10^{-4}$ | $< 0.05$ | – | – |
| 46020 | let-7a-1 | – | – | – | – | – |
| | let-7d | $< 1 \times 10^{-4}$ | $< 1 \times 10^{-4}$ | $< 1 \times 10^{-4}$ | $2.2 \times 10^{-6}$ | $1.0 \times 10^{-5}$ |
| | let-7f-1 | $< 1 \times 10^{-4}$ | – | $< 1 \times 10^{-4}$ | $1.2 \times 10^{-9}$ | $1.3 \times 10^{-5}$ |
| | let-7g | $< 1 \times 10^{-4}$ | – | $< 1 \times 10^{-4}$ | $2.0 \times 10^{-5}$ | $2.1 \times 10^{-3}$ |
| | mir-98 | $< 1 \times 10^{-4}$ | $< 0.05$ | $< 1 \times 10^{-3}$ | – | – |

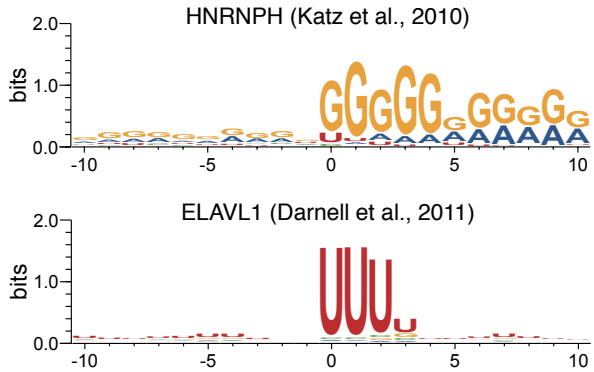



**Figure 3.9** Examples of simple sequence logo analysis for finding sequence motif. Position 0 is where crosslinking-induced errors accumulated.

## 3.4 Recognition motif analysis of binding sites

Binding specificity factors like RNA sequence or secondary structure are valuable information for many cases of RNA binding protein research. For example, poly(A)-binding proteins (PABPs) binds to poly(A) stretches specifically to stabilize the poly(A) tails (Kühn & Wahle, 2004), and LIN28A is known to interact with single-stranded GGNG and GNG sequences on top of hairpin structures (Nam et al., 2011; Cho et al., 2012) for regulation of precursor miRNAs and messenger RNAs. Thanks to the single nucleotide resolution of binding site identification in CLIP, it has became easier to find *cis*-regulatory factors in RNA for the RNA-protein interactions.

### 3.4.1 Sequence motif analysis of binding sites

Sequence motif analysis following the CIMS analysis is relatively easier than in the other techniques that have lower resolution. As crosslinking-induced errors pinpoint the positions where an RBP binds, there is no need for additional sequence cluster and alignment. Simple sequence logo analysis is enough for most cases (Figure 3.9).

When an RBP recognizes RNA sequences with multiple modes or they don't align well, the mixture of enrichment levels can easily fade out the distinct signals. For these cases,

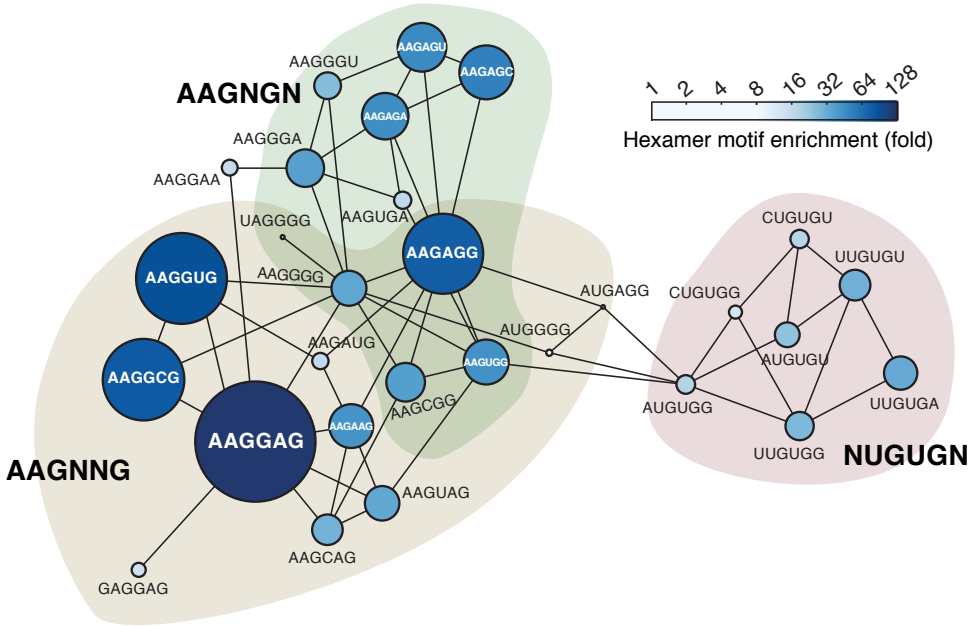



**Figure 3.10** Example of sequence motif clustering by similarity. Each hexamer sequence is an enriched motif in LIN28A HITS-CLIP (Cho et al., 2012). Area and color of each node represent relative enrichment of the hexameric sequence compared to the background frequency from RefSeq transcripts. Any two connected nodes differ by a single nucleotide.

I developed a visualization-based analysis based on clustering by similarity of enriched sequence motifs (Figure 3.10). The alternative visualization shows related sequences in adjacent positions, and naturally reveal the distinguishable groups of sequences that an RBP interacts with. In the example of LIN28A (Figure 3.10), any of simple sequence logo or traditional motif finders for transcription factor binding including MEME (Machanick & Bailey, 2011), PhyloGibbs (Siddharthan et al., 2005), and Trawler_standalone (Haudry et al., 2010) failed to divide the three groups of binding motifs while they are clearly visible in this method (data not shown).

### 3.4.2 Secondary structure motif analysis of binding sites

In addition to the primary sequence, secondary structure is often used for binding of RBPs to RNA. Double-stranded RNA has more stable structure in the cell, whose backbone structures can be a good platform for stable binding (Carlson et al., 2003). By using both dsRNA regions and few bases in single-stranded regions, proteins can build a specific binding domain with relatively small structure. Statistical analysis of secondary structure preference of RBP-bound sequences is often misleading due to limited accuracy of RNA secondary structure predictions. Especially, almost half of ~50 nucleotides long RNA sequences can form certain kind of folds. It causes high false positive rate of secondary structure scanning near the regions of interest and makes weak preferences undetectable. Additionally, setting a size of prediction window is tricky and requires some assumptions.

I developed a new method that is free from RNA secondary structure predictions, but can still detect preferences to small RNA hairpins. The method uses enrichment level of Watson-Crick (WC) co-occurrence compared to that of background frequency between flanking positions of binding sites. It could detect the obvious preference to small hairpins by LIN28A protein in both human and mouse embryonic stem cells (Figure 3.11).

## 3.5 Fully automated pipeline for CLIP-seq analysis

In spite that many alternative approaches and analytic techniques can be applied for thorough analysis of CLIP-seq, there has been no general analysis toolkit except PARalyzer by Corcoran et al. (2011), which is specialized in T-to-C transition analysis of PAR-CLIP. Here, I present *ecliptic*, a elastic, scalable, and yet easy-to-use tool chain package developed for analysis of data from CLIP-seq experiments supporting variety of alternative approaches.[5]

Ecliptic consists of several small programs written in Python and C with a pipeline script for Snakemake (Köster & Rahmann, 2012) to weave them into an automated workflow. Most parts were described earlier in this chapter, the rest are described below.

[5] Ecliptic is available under the MIT license from https://github.com/hyeshik/ecliptic.

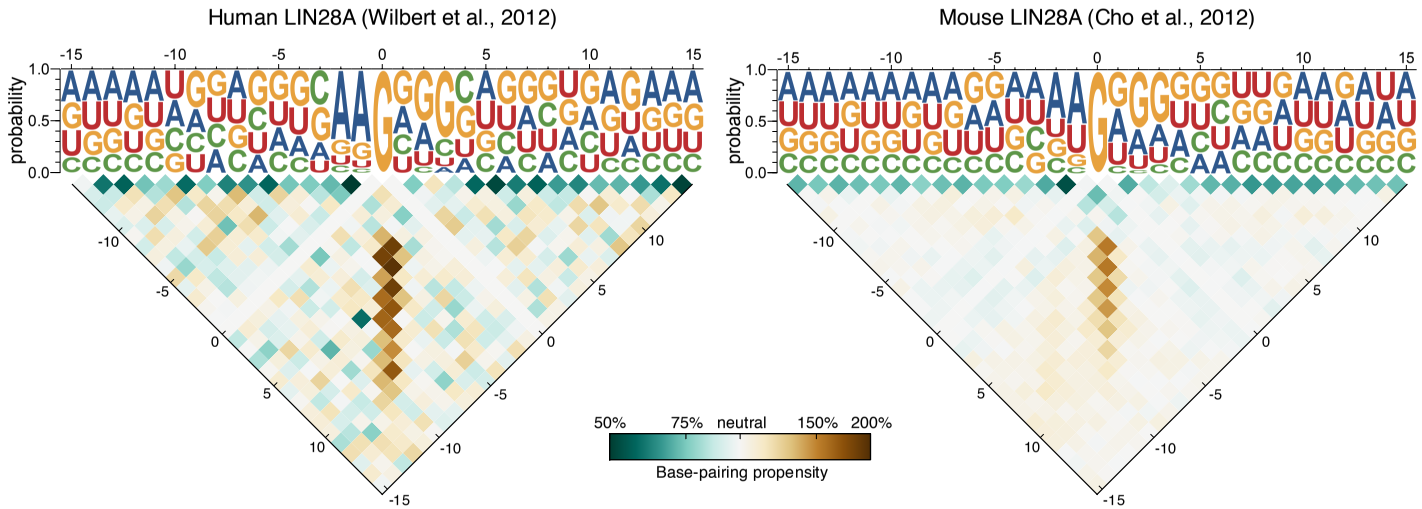



**Figure 3.11** WC-pair preference between two flanking positions around LIN28A-interacting sequences. Sequences are aligned such that the crosslinked bases are centered at zero. Nucleotide frequency is presented above the triangular matrix. Colors in the triangular matrix, indicating WC-pair enrichment or depletion, are normalized to the background frequency.

```yaml
CLIP-35L33G:
    runs:           [C3-091210, C3-110713, C3-111013]
    species:        mmu
    workflows:      [CLIP]
    source:         A3-1
    first_base:     7
    quality_scale:  33
    threep_adapter: ATCTCGTATGCCGTCTTCTGCTTG
    description:    "mESC LIN28A CLIP-seq with 35L33G"
PolyA-1:
    runs:           [P-110922]
    species:        mmu
    workflows:      [RNAseq, SNPreference]
    source:         A3-1
    first_base:     7
    quality_scale:  64
    threep_adapter: ATCTCGTATGCCGTCTTCTGCTTG
    description:    "mESC Poly-A RNA-seq"
```

**Figure 3.12** Example metadata for ecliptic describing a pair of a CLIP and Poly(A)-enriched RNA-seq experiments.

### Configuration and job script generation

Ecliptic accepts metadata describing experiments and comparison pairs in YAML format as shown in Figure 3.12). Several analysis modules written in Snakemake microlanguage (Köster & Rahmann, 2012) are templated and assembled as directed in the metadata. The templates for the modules are processed using jinja2 template engine (http://jinja.pocoo.org/) to make it more flexible and easy to extend the workflow. As the full CLIP-seq analysis includes many computationally expensive parts, most time-consuming tools and the pipeline in ecliptic are implemented with multi-processor and/or multi-node job scheduler support.

### Identification of binding sites

Two major alternative methods for binding site identification are performed by ecliptic. Firstly, crosslinking-induced footprints such as substitution or indel in HITS-CLIP, T-to-C transition in PAR-CLIP, and clustered 5′ ends in iCLIP are scanned throughout the genome, then the list is refined to keep statistically significant footprints only by estimated false

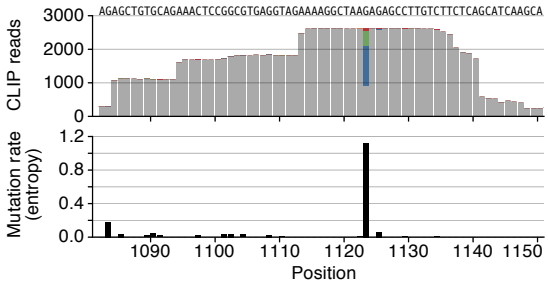



**Figure 3.13** Sequence alignment view around a identified binding site. Upper panel shows coverage of CLIP tags in base positions, and lower panel shows Shannon entropy of sequenced bases in each position.

discovery rate from permutation using my *in silico* CLIP simulator included in ecliptic (see Section 3.3.3 for details). The information near the identified binding sites can be easily visualized by a tool in ecliptic (Figure 3.13 as an example). Alternatively, regions of clustered CLIP tags are examined by one of peak callers like Piranha (Uren et al., 2012), ASPeak (Kucukural et al., 2013), or MACS (Zhang et al., 2008). Ecliptic tries all these methods when the dependent program is available on the system.

### Discovery of substrate specificity factors

To generalize their interaction characteristics at large, it is required to discover what factors affect the specificity between RNA and protein. As many RNA-binding proteins recognizes their substrates by nucleotide sequence and surrounding secondary structure, it is worth to try motif analyses. Ecliptic can automatically prepare inputs and invoke tools for sequence motif discovery like MEME (Machanick & Bailey, 2011), GLAM2 (Frith et al., 2008), PhyloGibbs (Siddharthan et al., 2005), or WebLogo (Crooks et al., 2004) (Figure 3.14). Several kinds of plots are also generated to evaluate secondary structure preference near the binding sites (Figure 3.15). See Section 3.4.2 for background and more details.

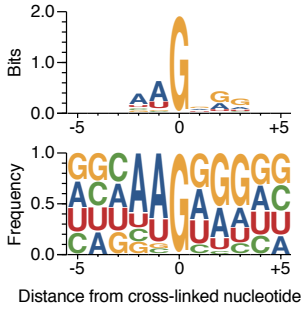



**Figure 3.14** Example of sequence logo showing enriched motif around binding sites for LIN28A in mouse ESC (Cho et al., 2012).

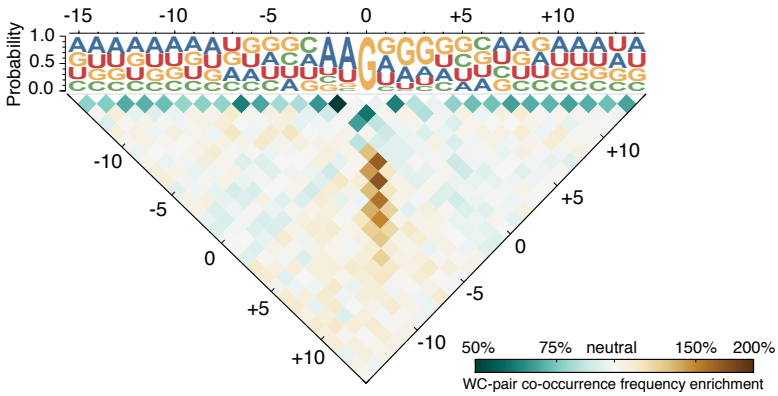



**Figure 3.15** Example of probability matrix showing overrepresented WC-pairing between two bases around binding sites for LIN28A in mouse ESC (Cho et al., 2012).

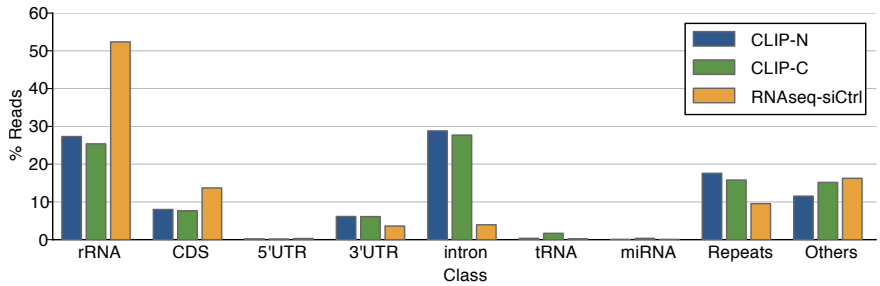



**Figure 3.16** Example of quality check view for transcript source composition in reads for CIRP in mouse (Morf et al., 2012).

### Calling confident target transcripts

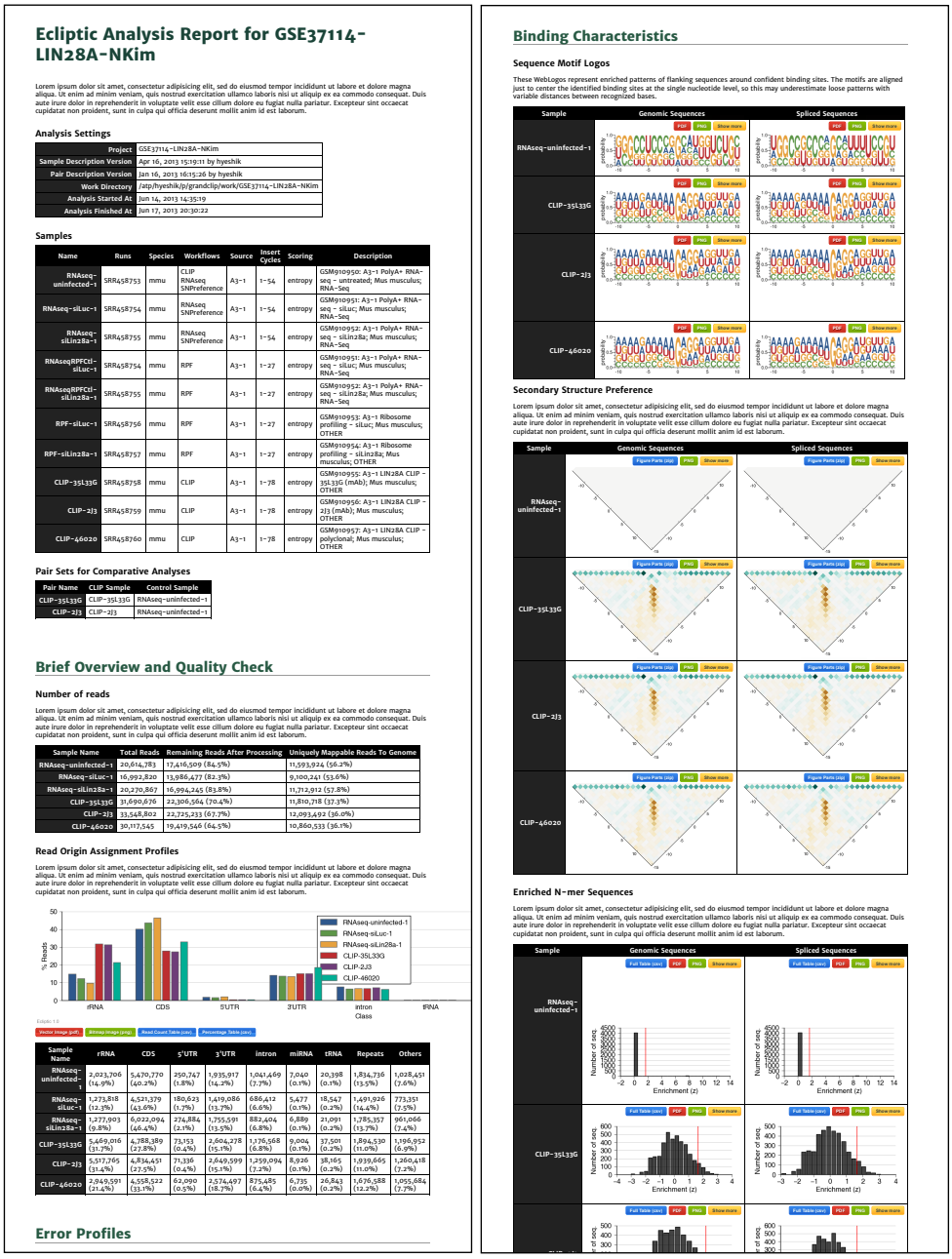
The list of confident target transcripts can lead the study to functional analyses of targets. A number of different lists of potential targets are provided by ecliptic to enable taking a different background or null hypothesis. Simple gene ontology analysis for enriched targets is also produced for the brief overview of results.

### Quality check and miscellaneous statistics

A CLIP experiment often fails. It usually requires several trial-and-errors by tuning various conditions depending on antibody, cell type, mode of RNA-protein interaction and, crosslinking method. To enable more rapid and fast iterations, ecliptic generates basic statistics to check quality of libraries such as sequence diversity, length distribution, read quality, and bacterial contamination (Figure 3.16).

### Reporting

Ecliptic generates a user-friendly report to make primary analyses accessible to researchers who are not familiar with Unix environment (Figure 3.17). It includes all basic information and publication-ready plots for fundamental analyses. As most CLIP-seq experiments eventually need study-specific downstream analyses, raw and intermediate data files in



# Ecliptic Analysis Report for GSE37114-LIN28A-NKim

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

### Analysis Settings

| | |
|---|---|
| Project | GSE37114-LIN28A-NKim |
| Sample Description Version | Apr 16, 2013 15:19:11 by hyeshik |
| Pair Description Version | Jan 16, 2013 16:15:26 by hyeshik |
| Work Directory | /atp/hyeshik/p/grandclip/work/GSE37114-LIN28A-NKim |
| Analysis Started At | Jun 14, 2013 14:35:19 |
| Analysis Finished At | Jun 17, 2013 20:30:22 |

### Samples

| Name | Runs | Species | Workflows | Source | Insert Cycles | Scoring | Description |
|---|---|---|---|---|---|---|---|
| RNAseq-uninfected-1 | SRR458753 | mmu | CLIP RNAseq SNPreference | A3-1 | 1-54 | entropy | GSM910950: A3-1 PolyA+ RNA-seq - untreated; Mus musculus; RNA-Seq |
| RNAseq-siLuc-1 | SRR458754 | mmu | RNAseq SNPreference | A3-1 | 1-54 | entropy | GSM910951: A3-1 PolyA+ RNA-seq - siluc; Mus musculus; RNA-Seq |
| RNAseq-siLin28a-1 | SRR458755 | mmu | RNAseq SNPreference | A3-1 | 1-54 | entropy | GSM910952: A3-1 PolyA+ RNA-seq - siLin28a; Mus musculus; RNA-Seq |
| RNAseqRPFCtl-siLuc-1 | SRR458754 | mmu | RPF | A3-1 | 1-27 | entropy | GSM910951: A3-1 PolyA+ RNA-seq - siluc; Mus musculus; RNA-Seq |
| RNAseqRPFCtl-siLin28a-1 | SRR458755 | mmu | RPF | A3-1 | 1-27 | entropy | GSM910952: A3-1 PolyA+ RNA-seq - siLin28a; Mus musculus; RNA-Seq |
| RPF-siLuc-1 | SRR458756 | mmu | RPF | A3-1 | 1-27 | entropy | GSM910953: A3-1 Ribosome profiling - siluc; Mus musculus; OTHER |
| RPF-siLin28a-1 | SRR458757 | mmu | RPF | A3-1 | 1-27 | entropy | GSM910954: A3-1 Ribosome profiling - siLin28a; Mus musculus; OTHER |
| CLIP-35L33G | SRR458758 | mmu | CLIP | A3-1 | 1-78 | entropy | GSM910955: A3-1 LIN28A CLIP - 35L33G (mAb); Mus musculus; OTHER |
| CLIP-2J3 | SRR458759 | mmu | CLIP | A3-1 | 1-78 | entropy | GSM910956: A3-1 LIN28A CLIP - 2J3 (mAb); Mus musculus; OTHER |
| CLIP-46020 | SRR458760 | mmu | CLIP | A3-1 | 1-78 | entropy | GSM910957: A3-1 LIN28A CLIP - polyclonal; Mus musculus; OTHER |

### Pair Sets for Comparative Analyses

| Pair Name | CLIP Sample | Control Sample |
|---|---|---|
| CLIP-35L33G | CLIP-35L33G | RNAseq-uninfected-1 |
| CLIP-2J3 | CLIP-2J3 | RNAseq-uninfected-1 |

## Brief Overview and Quality Check

### Number of reads

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

| Sample Name | Total Reads | Remaining Reads After Processing | Uniquely Mappable Reads To Genome |
|---|---|---|---|
| RNAseq-uninfected-1 | 20,614,783 | 17,416,509 (84.5%) | 11,593,924 (56.2%) |
| RNAseq-siLuc-1 | 16,992,820 | 13,986,477 (82.3%) | 9,100,241 (53.6%) |
| RNAseq-siLin28a-1 | 20,270,867 | 16,994,245 (83.8%) | 11,712,912 (57.8%) |
| CLIP-35L33G | 31,690,676 | 22,306,564 (70.4%) | 11,810,718 (37.3%) |
| CLIP-2J3 | 33,548,802 | 22,725,233 (67.7%) | 12,093,492 (36.0%) |
| CLIP-46020 | 30,117,545 | 19,419,546 (64.5%) | 10,860,533 (36.1%) |

### Read Origin Assignment Profiles

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.



| Sample Name | rRNA | CDS | 5'UTR | 3'UTR | Intron | miRNA | tRNA | Repeats | Others |
|---|---|---|---|---|---|---|---|---|---|
| RNAseq-uninfected-1 | 2,023,706 (14.9%) | 5,470,770 (40.2%) | 250,747 (1.8%) | 1,935,917 (14.2%) | 1,041,469 (7.7%) | 7,040 (0.1%) | 20,398 (0.1%) | 1,834,736 (13.5%) | 1,028,451 (7.6%) |
| RNAseq-siLuc-1 | 1,273,818 (12.3%) | 4,521,379 (43.6%) | 180,623 (1.7%) | 1,419,086 (13.7%) | 686,412 (6.6%) | 5,477 (0.1%) | 18,547 (0.2%) | 1,491,926 (14.4%) | 773,351 (7.5%) |
| RNAseq-siLin28a-1 | 1,277,903 (9.8%) | 6,022,094 (46.4%) | 274,884 (2.1%) | 1,755,591 (13.5%) | 882,404 (6.8%) | 6,889 (0.1%) | 21,091 (0.2%) | 1,785,357 (13.7%) | 961,066 (7.4%) |
| CLIP-35L33G | 5,469,016 (31.7%) | 4,788,389 (27.8%) | 73,153 (0.4%) | 2,604,278 (15.1%) | 1,176,568 (6.8%) | 9,004 (0.1%) | 37,501 (0.2%) | 1,894,530 (11.0%) | 1,196,952 (6.9%) |
| CLIP-2J3 | 5,517,765 (31.4%) | 4,834,451 (27.5%) | 71,336 (0.4%) | 2,649,599 (15.1%) | 1,259,094 (7.2%) | 8,926 (0.1%) | 38,165 (0.2%) | 1,939,665 (11.0%) | 1,260,418 (7.2%) |
| CLIP-46020 | 2,949,591 (21.4%) | 4,558,522 (33.1%) | 62,090 (0.5%) | 2,574,497 (18.7%) | 875,485 (6.4%) | 6,735 (0.0%) | 26,843 (0.2%) | 1,676,588 (12.2%) | 1,055,684 (7.7%) |

## Error Profiles

## Binding Characteristics

### Sequence Motif Logos

These WebLogos represent enriched patterns of flanking sequences around confident binding sites. The motifs are aligned just to center the identified binding sites at the single nucleotide level, so this may underestimate loose patterns with variable distances between recognized bases.



### Secondary Structure Preference

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.



### Enriched N-mer Sequences

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.



**Figure 3.17** Example pages of analysis report automatically generated by ecliptic.

text-based formats are listed and provided with help texts in the report.

## 3.6 Discussion

The new methods and implementations of analysis of CLIP-seq data allow mapping the binding sites of RBPs on the genomic scale at single nucleotide resolution. They successfully unveiled unknown biochemical properties of RNA binding proteins. In addition, ecliptic is the first full-featured suite for the general CLIP-seq analysis that provides automatic pipelining and modular extension. It performs most widely used analytic methods, and provides not only the final results but also many intermediate data in *de facto* standard formats for more in-depth analyses. It also includes the first publicly available implementation of permutation-based statistical analysis of crosslinking-induced errors in CLIP tags. These will significantly accelerate and lower hurdles of the research of RNA-protein interactions.

# 4. Conclusion

Regulation of RNA plays a pivotal role in diversification of the genetic repertoire, cellular homeostasis maintenance, localized functions, and fine-tuned transitions of cellular status. Being a digital information storage that is relatively easy to read, RNA has been one of the most convenient indicator of cellular regulation status. Through the last two decades, the methodology in RNA biology has been largely moved into top-down approaches with adoption of high-throughput methods. RNA-seq, cDNA microarray, CLIP-seq, and ribosome profiling have been workhorses for a significant fraction of recent researches.

In this thesis, I developed a novel method named TAIL-seq using direct interpretation of fluorescence signals to measure the length of poly(A) tails. It showed a fair level of measurement accuracy and provided the global profile of poly(A) tails for the first time. The analyses of poly(A) tails in NIH3T3 and HeLa cells presented several phenomena that were not described before. They include widespread uridylation and guanylation of poly(A) tails, their preference to short or long poly(A) tails, global view of deadenylation by microRNA targeting, and potential substrates of a newly hypothesized sequence-specific endonuclease. Still, there are plenty of room for improvement of the method. It needs to be more sensitive to transcripts with low quantity and gain technological maturity by improving reproducibility, measurement accuracy, and dynamic range.

The second part of this thesis covers the advances in analytic techniques of CLIP-seq. The RNA-protein interaction profiling method lacked an established workflow of data analysis due to the variability in the characteristics of RBPs. I devised and tested more metrics that quantify crosslinking between RNA and protein to allow sensitive detections for more RBPs. The newly developed software, *ecliptic*, is designed to accelerate the iterations of CLIP experiments and make it more accessible to wet lab scientists. With the optimized implementation of the false discovery estimation algorithm of crosslinked sites, more statistically powerful results will be produced with less computational resources.

In this thesis, I designed novel methods based on high-throughput sequencing, and demonstrated their applications in biological contexts. The newly developed technologies significant improved transcriptome-wide observation of poly(A) tail regulation and RNA-protein interaction. The new observations will enable discovery of unexpected links and mechanisms in RNA-mediated regulation of the cell.

# 국문초록

## 폴리A 꼬리와 RNA-단백질 상호작용에 대한 전사체적 분석

리보핵산(RNA)은 정보를 저장하고 세포 구성 물질 간에 정보를 전달하는 매개 물질이다. RNA의 생성부터 처리, 운반, 단백질 번역, 촉매 작용, 분해까지 세포 안의 많은 구성요소들이 정해진 생리적 현상을 일으키기 위해 RNA을 정확히 조절한다. 대용량 디옥시리보핵산(DNA) 서열분석 기술의 개발에 따라, RNA 조절 분야의 연구자들은 빠르게 이 기술을 받아들여 기존에 없었던 정밀도로 세포 안의 RNA를 관찰하기 시작했다. 이제 해당 분야의 획기적 발견들은 대부분 대용량 서열분석에 강하게 의존하고 있다. PIWI-상호작용 RNA, 단백질을 만들지 않는 긴 RNA (lincRNA), lincRNA의 주요 작용 메커니즘, 기존 경로와 다른 마이크로RNA의 신생성 과정, RNA 접합 조절 등이 모두 대용량 서열분석을 이용한 새로운 발견이다.

이 논문에서 나는 두 가지 종류의 RNA 조절을 연구하는데 사용될 기술을 새롭게 개발하고 그를 응용한다. 첫 번째로, 세포 전체의 아데닐산중합반응 상태를 조사할 수 있는 새로운 방법을 개발했다. 전령RNA의 3′ 말단은 유전자 발현 조절에서 아주 중요하지만, 단독중합체는 서열을 해독하기 매우 어렵기 때문에 전사체 수준에서 분석할 수 없었다. TAIL-seq이라고 명명된 기술을 개발하여, 폴리아데닐산 꼬리 길이를 유전체 전체에서 최초로 잴 수 있었다. 또한, 나는 폴리아데닐산 꼬리 뒷 부분에 유리딘화와 구아닌화가 광범위하게 일어난다는 사실을 발견하였다. 유리딘 꼬리는 보통 25 nt 미만의 짧은 폴리아데닐산 꼬리 뒤에서 관찰되었고, 구아닌 꼬리는 40 nt 이상의 긴 폴리아데닐산 꼬리 뒤에서 주로 발견되었다. 이는 유리딘과 구아닌 꼬리가 전령RNA의 안정성 조절에 관련이 있을 수 있음을 시사한다. 그리고, TAIL-seq 분석 결과 마이크로RNA와 전령RNA의 가공과 분해에 관련된 많은 종류의 핵산 분해 현상이 단일 염기 단위 해상도로 관찰되었다. 따라서, TAIL-seq을 이용하여 RNA 가공과 수식에 관련된 예측하지 못한 현상들을 살펴볼 수 있을 것이다.

두 번째로 나는 자외선 교차결합, 면역침강 후 서열분석법(CLIP-seq)을 개선함으로써 RNA와 단백질의 상호작용을 분석하는데 유용하게 사용할 수 있는 방법들을 개발하였다. CLIP-seq은 RNA와 단백질 상호작용 정보를 전사체 범위로 연구하는 데 있어 필수적인 도구 중의 하나로 떠올랐다. 그러나, 아직 다른 RNA 서열분석법들과 달리 일반화된 방법론과 도구가 정립되지 않았다. 이 연구를 통해 새로 개발된 통계적, 시각적 분석 방법들은 예전에 쉽게 관찰할 수 없었던 새로운 정보를 드러낼 수 있을 것이다. 추가로, 새로이 개발된 CLIP 분석 도구 모음인 ecliptic은 RNA와 단백질 상호작용 연구를 더 빠르게 할 것이며 연구자들이 더 많은 정보에 접근하기 쉽도록 만들 것이다.

이 논문에서는 폴리아데닐산 꼬리의 전체적인 연구와 RNA 단백질 간 상호작용의 특정성에 대한 단일 염기 단위 해상도 조사법에 대해 기술적인 개선을 이루었다. 이로써, RNA의 3′ 말단과 RNA와 RNA 결합 단백질 사이의 인터페이스에 대한 여러 현상들을 발견했다. 대용량 실험 기법들은 편향되지 않은 시각을 제공하고 기존 방법으로 알기 어려웠던 현상을 관찰할 수 있게하여 생물학의 범위를 확장시키고 있다. 기존 기술들의 유연한 변용을 통해서 새로운 발견의 기회가 더욱 늘어날 것으로 보인다.

# Bibliography


Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., & Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the *Arabidopsis* degradome. *Curr Biol*, 18(10), 758–62.

Ascano, Jr, M., Mukherjee, N., Bandaru, P., Miller, J. B., Nusbaum, J. D., Corcoran, D. L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M., Williams, Z., Ohler, U., & Tuschl, T. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. *Nature*, 492(7429), 382–6.

Aviner, R., Geiger, T., & Elroy-Stein, O. (2013). Novel proteomic approach (PUNCH-P) reveals cell cycle-specific fluctuations in mRNA translation. *Genes Dev*, 27(16), 1834–44.

Bai, Y., Srivastava, S. K., Chang, J. H., Manley, J. L., & Tong, L. (2011). Structural basis for dimerization and activity of human PAPD1, a noncanonical poly(A) polymerase. *Mol Cell*, 41(3), 311–20.

Barreau, C., Paillard, L., & Osborne, H. B. (2005). AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res*, 33(22), 7138–50.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, 41(Database issue), D991–5.

Bazzini, A. A., Lee, M. T., & Giraldez, A. J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science*, 336(6078), 233–7.

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., & West, R. B. (2010). 3′-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS One*, 5(1), e8768.

Beilharz, T. H. & Preiss, T. (2007). Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. *RNA*, 13(7), 982–97.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57, 289–300.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R.,

Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., & Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218), 53–9.

Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., & Tyson, G. W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. *PLoS Comput Biol*, 9(4), e1003031.

Braun, I. C., Rohrbach, E., Schmitt, C., & Izaurralde, E. (1999). TAP binds to the constitutive transport element (CTE) through a novel RNA-binding motif that is sufficient to promote CTE-dependent RNA export from the nucleus. *EMBO J*, 18(7), 1953–65.

Brawerman, G. (1974). Eukaryotic messenger RNA. *Annu Rev Biochem*, 43(0), 621–42.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., & Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res*, 41(Database issue), D226–32.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell*, PAMI-8(6), 679–698.

Carlson, C. B., Stephens, O. M., & Beal, P. A. (2003). Recognition of double-stranded RNA by proteins and small molecules. *Biopolymers*, 70(1), 86–102.

Chan, P. P. & Lowe, T. M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*, 37(Database issue), D93–7.

Chi, S. W., Zang, J. B., Mele, A., & Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254), 479–86.

Cho, J., Chang, H., Kwon, S. C., Kim, B., Kim, Y., Choe, J., Ha, M., Kim, Y. K., & Kim, V. N. (2012). LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell*, 151(4), 765–77.

Chu, C., Qu, K., Zhong, F. L., Artandi, S. E., & Chang, H. Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell*, 44(4), 667–78.

Churchman, L. S. & Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330), 368–73.

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., & Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol*, 12(8), R79.

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909), 1845–8.

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res*, 14(6), 1188–90.

D'Ambrogio, A., Nagaoka, K., & Richter, J. D. (2013). Translational control of cell growth and malignancy by the CPEBs. *Nat Rev Cancer*, 13(4), 283–90.

Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y. S., Mele, A., Fraser, C. E., Stone, E. F., Chen, C., Fak, J. J., Chi, S. W., Licatalosi, D. D., Richter, J. D., & Darnell, R. B. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, 146(2), 247–61.

de Moor, C. H. & Richter, J. D. (1999). Cytoplasmic polyadenylation elements mediate masking and unmasking of cyclin B1 mRNA. *EMBO J*, 18(8), 2294–303.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc Ser B*, 39(1), 1–38.

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., Rohl, C. A., Johnson, J. M., & Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Res*, 22(6), 1173–83.

Djuranovic, S., Nahvi, A., & Green, R. (2012). miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science*, 336(6078), 237–40.

Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., & Rechavi, G. (2013). Transcriptome-wide mapping of $N^6$−methyladenosine by $m^6A$-seq based on immunocapturing and massively parallel sequencing. *Nat Protoc*, 8(1), 176–89.

Dreyfus, M. & Régnier, P. (2002). The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. *Cell*, 111(5), 611–3.

Elkon, R., Ugalde, A. P., & Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nat Rev Genet*, 14(7), 496–506.

Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using *Phred*. I. accuracy assessment. *Genome Res*, 8(3), 175–85.

Fabian, M. R., Sonenberg, N., & Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem*, 79, 351–79.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Vogel, J., & Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Res*, 39(Database issue), D800–6.

Freeberg, L., Kuersten, S., & Syed, F. (2013). Isolate and sequence ribosome-protected mRNA fragments using size-exclusion chromatography. *Nat Meth*, 10(5), –.

Frith, M. C., Saunders, N. F. W., Kobe, B., & Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol*, 4(4), e1000071.

Frohman, M. A., Dush, M. K., & Martin, G. R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A*, 85(23), 8998–9002.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150–2.

Fu, Y., Sun, Y., Li, Y., Li, J., Rao, X., Chen, C., & Xu, A. (2011). Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing. *Genome Res*, 21(5), 741–7.

Fullwood, M. J., Wei, C.-L., Liu, E. T., & Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*, 19(4), 521–32.

Garneau, N. L., Wilusz, J., & Wilusz, C. J. (2007). The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol*, 8(2), 113–26.

German, M. A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B. C., & Green, P. J. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol*, 26(8), 941–6.

Goodarzi, H., Najafabadi, H. S., Oikonomou, P., Greco, T. M., Fish, L., Salavati, R., Cristea, I. M., & Tavazoie, S. (2012). Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 485(7397), 264–8.

Granneman, S., Kudla, G., Petfalski, E., & Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci U S A*, 106(24), 9613–8.

Gray, N. K. & Hentze, M. W. (1994). Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to *ferritin* and *eALAS* mRNAs. *EMBO J*, 13(16), 3882–91.

Griffiths-Jones, S. (2010). miRBase: microRNA sequences and annotation. *Curr Protoc Bioinformatics*, Chapter 12, Unit 12.9.1–10.

Guo, H., Ingolia, N. T., Weissman, J. S., & Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308), 835–40.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, Jr, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., & Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1), 129–41.

Hafner, M., Max, K. E. A., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., & Tuschl, T. (2013). Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. *RNA*, 19(5), 613–26.

Han, J., Lee, Y., Yeom, K.-H., Nam, J.-W., Heo, I., Rhee, J.-K., Sohn, S. Y., Cho, Y., Zhang, B.-T., & Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125(5), 887–901.

Haudry, Y., Ramialison, M., Paten, B., Wittbrodt, J., & Ettwiller, L. (2010). Using Trawler_standalone to discover overrepresented motifs in DNA and RNA sequences derived from various experiments including chromatin immunoprecipitation. *Nat Protoc*, 5(2), 323–34.

Hecker, K. H. & Rill, R. L. (1998). Error analysis of chemically synthesized polynucleotides. *Biotechniques*, 24(2), 256–60.

Henriksson, N., Nilsson, P., Wu, M., Song, H., & Virtanen, A. (2010). Recognition of adenosine residues by the active site of poly(A)-specific ribonuclease. *J Biol Chem*, 285(1), 163–70.

Heo, I., Ha, M., Lim, J., Yoon, M.-J., Park, J.-E., Kwon, S. C., Chang, H., & Kim, V. N. (2012). Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. *Cell*, 151(3), 521–32.

Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., Yeom, K.-H., Han, J., & Kim, V. N. (2009). TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell*, 138(4), 696–708.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J. Y., Yehia, G., & Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat Methods*, 10(2), 133–9.

Huntzinger, E. & Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet*, 12(2), 99–110.

Illumina, Inc. (2011). *HCS 1.4/RTA 1.12 Theory of Operation*. Illumina, Inc., San Diego, United States, Rev. May 2011 edition.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924), 218–23.

Jan, C. H., Friedman, R. C., Ruby, J. G., & Bartel, D. P. (2011). Formation, regulation and evolution of *Caenorhabditis elegans* 3′UTRs. *Nature*, 469(7328), 97–101.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R., & Shoemaker, D. D. (2003). Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302(5653), 2141–4.

Josefson, S. (2003). The base16, base32, and base64 data encodings. *RFC 3548*, The Internet Society.

Karginov, F. V., Cheloufi, S., Chong, M. M. W., Stark, A., Smith, A. D., & Hannon, G. J. (2010). Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, Drosha, and additional nucleases. *Mol Cell*, 38(6), 781–8.

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12), 1009–15.

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*, 11(5), 345–55.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., & Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311), 103–7.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., & Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat Methods*, 8(7), 559–64.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., & Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7), 909–15.

Köster, J. & Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520–2.

Kozomara, A. & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39(Database issue), D152–7.

Kucukural, A., Özadam, H., Singh, G., Moore, M. J., & Cenik, C. (2013). ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq. *Bioinformatics*, 29(19), 2485–6.

Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., & Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc Natl Acad Sci U S A*, 108(24), 10010–5.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Brief Bioinform*, 14(2), 144–61.

Kühn, U. & Wahle, E. (2004). Structure and function of poly(A) binding proteins. *Biochim Biophys Acta*, 1678(2-3), 67–84.

Kurosawa, J., Nishiyori, H., & Hayashizaki, Y. (2011). Deep cap analysis of gene expression. *Methods Mol Biol*, 687, 147–63.

Lagier-Tourenne, C., Polymenidou, M., Hutt, K. R., Vu, A. Q., Baughn, M., Huelga, S. C., Clutario, K. M., Ling, S.-C., Liang, T. Y., Mazur, C., Wancewicz, E., Kim, A. S., Watt, A., Freier, S., Hicks, G. G., Donohue, J. P., Shiue, L., Bennett, C. F., Ravits, J., Cleveland, D. W., & Yeo, G. W. (2012). Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat Neurosci*, 15(11), 1488–97.

Ledergerber, C. & Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Brief Bioinform*, 12(5), 489–97.

Lejeune, F., Li, X., & Maquat, L. E. (2003). Nonsense-mediated mRNA decay in mammalian cells involves decapping, deadenylating, and exonucleolytic activities. *Mol Cell*, 12(3), 675–87.

Levin, J. Z., Yassour, M., Adiconis, X., Nusbaum, C., Thompson, D. A., Friedman, N., Gnirke, A., & Regev, A. (2010). Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*, 7(9), 709–15.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27(5), 718–9.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16), 2078–9.

Licatalosi, D. D. & Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*, 11(1), 75–87.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., & Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221), 464–9.

Liu, X. & Gorovsky, M. A. (1993). Mapping the 5′ and 3′ ends of *Tetrahymena thermophila* mRNAs using RNA ligase mediated amplification of cDNA ends (RLM-RACE). *Nucleic Acids Res*, 21(21), 4954–4960.

Lubas, M., Damgaard, C. K., Tomecki, R., Cysewski, D., Jensen, T. H., & Dziembowski, A. (2013). Exonuclease hDIS3L2 specifies an exosome-independent 3′-5′ degradation pathway of human cytoplasmic mRNA. *EMBO J*, 32(13), 1855–68.

Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., & Arkin, A. P. (2011). Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci U S A*, 108(27), 11063–8.

Machanick, P. & Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, 27(12), 1696–7.

Macias, S., Plass, M., Stajuda, A., Michlewski, G., Eyras, E., & Cáceres, J. F. (2012). DGCR8 HITS-CLIP reveals novel functions for the microprocessor. *Nat Struct Mol Biol*, 19(8), 760–6.

Maier, T., Güell, M., & Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Lett*, 583(24), 3966–73.

Malecki, M., Viegas, S. C., Carneiro, T., Golik, P., Dressaire, C., Ferreira, M. G., & Arraiano, C. M. (2013). The exoribonuclease Dis3L2 defines a novel eukaryotic RNA degradation pathway. *EMBO J*, 32(13), 1842–54.

Mamanova, L., Andrews, R. M., James, K. D., Sheridan, E. M., Ellis, P. D., Langford, C. F., Ost, T. W. B., Collins, J. E., & Turner, D. J. (2010). FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods*, 7(2), 130–2.

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., Attie, O., Chen, K., Salehi-Ashtiani, K., Vidal, M., Harkins, T. T., Bouffard, P., Suzuki, Y., Sugano, S., Kohara, Y., Rajewsky, N., Piano, F., Gunsalus, K. C., & Kim, J. K. (2010). The landscape of *C. elegans* 3′UTRs. *Science*, 329(5990), 432–5.

Martin, G., Gruber, A. R., Keller, W., & Zavolan, M. (2012). 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), 753–63.

Mayford, M., Baranes, D., Podsypanina, K., & Kandel, E. R. (1996). The 3′-untranslated region of *CaMKII* alpha is a cis-acting signal for the localization and translation of mRNA in dendrites. *Proc Natl Acad Sci U S A*, 93(23), 13250–5.

Meijer, H. A., Bushell, M., Hill, K., Gant, T. W., Willis, A. E., Jones, P., & de Moor, C. H. (2007). A novel method for poly(A) fractionation reveals a large population of mRNAs with a short poly(A) tail in mammalian cells. *Nucleic Acids Res*, 35(19), e132.

Mendez, R. & Richter, J. D. (2001). Translational control by CPEB: a means to the end. *Nat Rev Mol Cell Biol*, 2(7), 521–9.

Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S., & Rinn, J. L. (2012). Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol*, 30(1), 99–104.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1), 31–46.

Morf, J., Rey, G., Schneider, K., Stratmann, M., Fujita, J., Naef, F., & Schibler, U. (2012). Cold-inducible RNA-binding protein modulates circadian gene expression posttranscriptionally. *Science*, 338(6105), 379–83.

Morozov, I. Y., Jones, M. G., Gould, P. D., Crome, V., Wilson, J. B., Hall, A. J. W., Rigden, D. J., & Caddick, M. X. (2012). mRNA 3′ tagging is induced by nonsense-mediated decay and promotes ribosome dissociation. *Mol Cell Biol*, 32(13), 2585–95.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7), 621–8.

Mukherjee, D., Gao, M., O'Connor, J. P., Raijmakers, R., Pruijn, G., Lutz, C. S., & Wilusz, J. (2002). The mammalian exosome mediates the efficient degradation of mRNAs that contain AU-rich elements. *EMBO J*, 21(1-2), 165–74.

Mullen, T. E. & Marzluff, W. F. (2008). Degradation of histone mRNA requires oligouridylation followed by decapping and simultaneous degradation of the mRNA both 5′ to 3′ and 3′ to 5′. *Genes Dev*, 22(1), 50–65.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881), 1344–9.

Nam, Y., Chen, C., Gregory, R. I., Chou, J. J., & Sliz, P. (2011). Molecular basis for interaction of let-7 microRNAs with Lin28. *Cell*, 147(5), 1080–1091.

Nevins, J. R. (1983). The pathway of eukaryotic mRNA formation. *Annu Rev Biochem*, 52, 441–66.

Nielsen, K. L., Høgh, A. L., & Emmersen, J. (2006). DeepSAGE–digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res*, 34(19), e133.

Norbury, C. J. (2013). Cytoplasmic RNA: a case of the tail wagging the dog. *Nat Rev Mol Cell Biol*, 14(10), 643–53.

Olivarius, S., Plessy, C., & Carninci, P. (2009). High-throughput verification of transcriptional starting sites by Deep-RACE. *Biotechniques*, 46(2), 130–2.

O'Shea, J. P., Chou, M. F., Quader, S. A., Ryan, J. K., Church, G. M., & Schwartz, D. (2013). pLogo: a probabilistic approach to visualizing sequence motifs. *Nat Methods*.

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S. W., Fishilevich, E., Monaghan, A. P., John, B., & Milos, P. M. (2010). Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, 143(6), 1018–29.

Ozsolak, F. & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, 12(2), 87–98.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., & Milos, P. M. (2009). Direct RNA sequencing. *Nature*, 461(7265), 814–8.

Park, J.-E., Heo, I., Tian, Y., Simanshu, D. K., Chang, H., Jee, D., Patel, D. J., & Kim, V. N. (2011). Dicer recognizes the 5′ end of rna for efficient and accurate processing. *Nature*, 475(7355), 201–5.

Plessy, C., Bertin, N., Takahashi, H., Simone, R., Salimullah, M., Lassmann, T., Vitezic, M., Severin, J., Olivarius, S., Lazarevic, D., Hornig, N., Orlando, V., Bell, I., Gao, H., Dumais, J., Kapranov, P., Wang, H., Davis, C. A., Gingeras, T. R., Kawai, J., Daub, C. O., Hayashizaki, Y., Gustincich, S., & Carninci, P. (2010). Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat Methods*, 7(7), 528–34.

Prewitt, J. M. S. (1970). *Picture processing and Psychopictorics*, chapter Object enhancement and Extraction. Academic Press.

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*, 40(Database issue), D130–5.

Quinlan, A. R. & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–2.

Rammelt, C., Bilen, B., Zavolan, M., & Keller, W. (2011). PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif. *RNA*, 17(9), 1737–46.

Raz, T., Kapranov, P., Lipson, D., Letovsky, S., Milos, P. M., & Thompson, J. F. (2011). Protocol dependence of sequencing-based gene expression measurements. *PLoS One*, 6(5), e19287.

Riley, K. J. & Steitz, J. A. (2013). The "Observer Effect" in genome-wide surveys of protein-RNA interactions. *Mol Cell*, 49(4), 601–4.

Rissland, O. S., Mikulasova, A., & Norbury, C. J. (2007). Efficient RNA polyuridylation by noncanonical poly(A) polymerases. *Mol Cell Biol*, 27(10), 3612–24.

Rissland, O. S. & Norbury, C. J. (2009). Decapping is preceded by 3′ uridylation in a novel pathway of bulk mRNA turnover. *Nat Struct Mol Biol*, 16(6), 616–23.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-seq expression estimates by correcting for fragment bias. *Genome Biol*, 12(3), R22.

Ronaghi, M., Uhlén, M., & Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281(5375), 363, 365.

Sallés, F. J., Richards, W. G., & Strickland, S. (1999). Assaying the polyadenylation state of mRNAs. *Methods*, 17(1), 38–45.

Saulière, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonquèze, O., Audic, Y., Paillard, L., Roest Crollius, H., & Le Hir, H. (2012). CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nat Struct Mol Biol*, 19(11), 1124–31.

Savitzky, A. & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem*, 36(8), 1627–1639.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467–70.

Schimmel, P., Giegé, R., Moras, D., & Yokoyama, S. (1993). An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci U S A*, 90(19), 8763–8.

Schlötterer, C. & Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Res*, 20(2), 211–5.

Schmidt, M.-J., West, S., & Norbury, C. J. (2011). The human cytoplasmic RNA terminal U-transferase ZCCHC11 targets histone mRNAs for degradation. *RNA*, 17(1), 39–44.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347), 337–42.

Scotto-Lavino, E., Du, G., & Frohman, M. A. (2006a). 3′ end cDNA amplification using classic RACE. *Nat Protoc*, 1(6), 2742–5.

Scotto-Lavino, E., Du, G., & Frohman, M. A. (2006b). 5′ end cDNA amplification using classic RACE. *Nat Protoc*, 1(6), 2555–62.

Scotto-Lavino, E., Du, G., & Frohman, M. A. (2006c). Amplification of 5′ end cDNA with 'new RACE'. *Nat Protoc*, 1(6), 3056–61.

Sement, F. M., Ferrier, E., Zuber, H., Merret, R., Alioua, M., Deragon, J.-M., Bousquet-Antonelli, C., Lange, H., & Gagliardi, D. (2013). Uridylation prevents 3′ trimming of oligoadenylated mRNAs. *Nucleic Acids Res*, 41(14), 7115–27.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst Tech J*, 27, 379–423.

Sharon, D., Tilgner, H., Grubert, F., & Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol*, 31(11), 1009–14.

Shen, B. & Goodman, H. M. (2004). Uridine addition after microRNA-directed cleavage. *Science*, 306(5698), 997.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol*, 26(10), 1135–45.

Shepard, P. J., Choi, E.-A., Lu, J., Flanagan, L. A., Hertel, K. J., & Shi, Y. (2011). Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17(4), 761–72.

Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., & Bartel, D. P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell*, 38(6), 789–802.

Siddharthan, R., Siggia, E. D., & van Nimwegen, E. (2005). PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7), e67.

Sievers, C., Schlumpf, T., Sawarkar, R., Comoglio, F., & Paro, R. (2012). Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res*, 40(20), e160.

Simon, M. D., Wang, C. I., Kharchenko, P. V., West, J. A., Chapman, B. A., Alekseyenko, A. A., Borowsky, M. L., Kuroda, M. I., & Kingston, R. E. (2011). The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci U S A*, 108(51), 20497–502.

Singh, G., Ricci, E. P., & Moore, M. J. (2013). RIPiT-Seq: A high-throughput approach for footprinting RNA:protein complexes. *Methods*.

Sleator, D. D. & Tarjan, R. E. (1985). Self-adjusting binary search trees. *J. ACM*, 32(3), 652–686.

Spies, N., Burge, C. B., & Bartel, D. P. (2013). 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res*, 23(12), 2078–90.

Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T. K., Hein, M. Y., Huang, S.-X., Ma, M., Shen, B., Qian, S.-B., Hengel, H., Mann, M., Ingolia, N. T., & Weissman, J. S. (2012). Decoding human cytomegalovirus. *Science*, 338(6110), 1088–93.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., & Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol*, 13(8), R67.

Tani, H., Mizutani, R., Salam, K. A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y., & Akimitsu, N. (2012). Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res*, 22(5), 947–56.

Ulitsky, I., Shkumatava, A., Jan, C. H., Subtelny, A. O., Koppstein, D., Bell, G. W., Sive, H., & Bartel, D. P. (2012). Extensive alternative polyadenylation during zebrafish development. *Genome Res*, 22(10), 2054–66.

Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R., & Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods*, 7(12), 995–1001.

Uren, P. J., Bahrami-Samani, E., Burns, S. C., Qiao, M., Karginov, F. V., Hodges, E., Hannon, G. J., Sanford, J. R., Penalva, L. O. F., & Smith, A. D. (2012). Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 28(23), 3013–20.

Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235), 484–7.

Vermeulen, A., Behlen, L., Reynolds, A., Wolfson, A., Marshall, W. S., Karpilow, J., & Khvorova, A. (2005). The contributions of dsRNA structure to Dicer specificity and efficiency. *RNA*, 11(5), 674–82.

Vignali, M., Armour, C. D., Chen, J., Morrison, R., Castle, J. C., Biery, M. C., Bouzek, H., Moon, W., Babak, T., Fried, M., Raymond, C. K., & Duffy, P. E. (2011). NSR-seq transcriptional profiling enables identification of a gene signature of *Plasmodium falciparum* parasites infecting children. *J Clin Invest*, 121(3), 1119–29.

Viswanathan, P., Chen, J., Chiang, Y.-C., & Denis, C. L. (2003). Identification of multiple RNA features that influence CCR4 deadenylation activity. *J Biol Chem*, 278(17), 14949–55.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inform Theory*, 13(2), 260–269.

Wang, E. T., Cody, N. A. L., Jog, S., Biancolella, M., Wang, T. T., Treacy, D. J., Luo, S., Schroth, G. P., Housman, D. E., Reddy, S., Lécuyer, E., & Burge, C. B. (2012). Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell*, 150(4), 710–24.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1), 57–63.

Weill, L., Belloc, E., Bava, F.-A., & Méndez, R. (2012). Translational control by changes in poly(A) tail length: recycling mRNAs. *Nat Struct Mol Biol*, 19(6), 577–85.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., & Yaschenko, E. (2008). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 36(Database issue), D13–21.

Whiteford, N., Skelly, T., Curtis, C., Ritchie, M. E., Löhr, A., Zaranek, A. W., Abnizova, I., & Brown, C. (2009). Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics*, 25(17), 2194–9.

Wilbert, M. L., Huelga, S. C., Kapeli, K., Stark, T. J., Liang, T. Y., Chen, S. X., Yan, B. Y., Nathanson, J. L., Hutt, K. R., Lovci, M. T., Kazan, H., Vu, A. Q., Massirer, K. B., Morris, Q., Hoon, S., & Yeo, G. W. (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Mol Cell*, 48(2), 195–206.

Wilkening, S., Pelechano, V., Järvelin, A. I., Tekkedil, M. M., Anders, S., Benes, V., & Steinmetz, L. M. (2013). An efficient method for genome-wide polyadenylation site mapping and RNA quantification. *Nucleic Acids Res*, 41(5), e65.

Wu, Q., Kim, Y. C., Lu, J., Xuan, Z., Chen, J., Zheng, Y., Zhou, T., Zhang, M. Q., Wu, C.-I., & Wang, S. M. (2008). Poly A- transcripts expressed in HeLa cells. *PLoS One*, 3(7), e2803.

Wu, T. D. & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7), 873–81.

Xiao, R., Tang, P., Yang, B., Huang, J., Zhou, Y., Shao, C., Li, H., Sun, H., Zhang, Y., & Fu, X.-D. (2012). Nuclear matrix factor hnRNP U/SAF-A exerts a global control of alternative splicing by regulating U2 snRNP maturation. *Mol Cell*, 45(5), 656–68.

Xue, Y., Ouyang, K., Huang, J., Zhou, Y., Ouyang, H., Li, H., Wang, G., Wu, Q., Wei, C., Bi, Y., Jiang, L., Cai, Z., Sun, H., Zhang, K., Zhang, Y., Chen, J., & Fu, X.-D. (2013). Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. *Cell*, 152(1-2), 82–96.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D. L., Sun, H., Fu, X.-D., & Zhang, Y. (2009). Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol Cell*, 36(6), 996–1006.

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., & Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2), 130–7.

Yoon, O. K. & Brem, R. B. (2010). Noncanonical transcript forms in yeast and their regulation during environmental stress. *RNA*, 16(6), 1256–67.

Zhang, C. & Darnell, R. B. (2011). Mapping *in vivo* protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol*, 29(7), 607–14.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9), R137.

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., & Lee, J. T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell*, 40(6), 939–53.

# Index