Statistical Modeling and Computational Analysis of Ribosome Profiling for the Dissection of Translational Control

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M.Sc. Yi Zhong
aus Deyang, China

Tübingen
2016

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	04.11.2016
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Gunnar Rätsch
2. Berichterstatter:	Prof. Dr. Oliver Kohlbacher

Abstract

mRNA translation is one of the most complex molecular processes that has been developed during the evolution of the cell. It synthesizes proteins that are the building blocks and workhorses of the cells. Translation consists of initiation, elongation and termination. These three steps require hundreds of molecules to participate in a concerted manner. Translational control regulates protein levels in response to intra- and extra-cellular environmental changes. Regulation at the translational level is important in situations where transcription regulation alone cannot satisfy the emergent needs of the cell or when local control over protein abundance is required. Translational control plays an essential role in maintaining cell homeostasis, physiology and modulating cell growth. Dysregulation of mRNA translation or aberrant function of translation machinery can lead to a variety of diseases including metabolic disorders and cancer. Thus, elucidating the mechanisms of translational control is key for understanding how diseases develop.

High-throughput sequencing technologies are widely used to determine and quantify DNA and RNA molecules on a large scale, which has remarkably facilitated our understanding of many biological functions in a system-wide manner. An extension of this technology, ribosome profiling, even allows characterization of ribosome-occupied mRNA fragments. Ribosome profiling, thus, provides an opportunity to globally monitor the translation *in vivo* and study the mechanisms of translational control.

My thesis consists of two main parts: 1) The first part focusses on development of *RiboDiff*, a statistical framework and computational tool for detecting genes under differential translational regulation. *RiboDiff* fits quantitative ribosome profiling and RNA-Seq measurements with a negative binomial based generalized linear model. Subsequently, a statistical test is performed to identify genes under differential translational control between measurements in two conditions. Our experiment demonstrates *RiboDiff* outperformed state-of-the-art existing approaches. 2) The second part establishes a computational pipeline for analyzing ribosome profiling data. Using this pipeline, we studied the translational regulation in leukemia and other conditions. We identified mRNAs that presented distinct translation efficiency and ribosome footprint density in a condition that the cells were treated with a chemical compound Silvestrol. Further analysis of these mRNAs revealed that the guanine quartet $(GCC)_4$ sequence pattern, which forms a G-quadruplex structure, is enriched in the 5' UTR of 280 mRNAs with down-regulated translation. Experimental validations supported our findings and confirmed that the G-quadruplex is an RNA element that represses the translation initiation activity. Applications of the computational approach on other translational researches illustrate the versatility of the proposed analysis methodology.

Zusammenfassung

Die mRNA-Translation ist einer der komplexesten molekularen Prozesse die sich entwickelt haben. In diesem Prozess werden Proteine, wichtige Elemente und Arbeitspferde von Zellen, synthetisiert. Die Translation unterteilt sich in drei Schritte: Initiation, Elongation und Termination. Diese drei Schritte erfordern das abgestimmte Zusammenspiel von Hunderten von Molekülen in einer bestimmten Reihenfolge. Durch Kontrolle der Translation regeln Zellen die Proteinspiegel in Reaktion auf intra- und extrazelluläre Veränderungen der Umwelt. Regulation auf der Translationsebene ist insbesondere in Situationen wichtig für die Zellen, in denen die Transkriptionsregulation allein die auftretenden Bedürfnisse der Zelle nicht erfüllen kann oder die lokale Kontrolle über Proteinmengen benötigt wird. Translationskontrolle spielt ausserdem eine wesentliche Rolle bei der Aufrechterhaltung der Zell-Homöostase, der Physiologie und um das Zellwachstum zu modulieren. Dysregulation der mRNA-Translation oder anomale Funktion der Translationsmaschinerie kann zu einer Vielzahl von Krankheiten, wie zum Beispiel Stoffwechselerkrankungen und Krebs, führen. Um das Entstehen von Krankheiten besser zu verstehen, ist es daher wichtig die Mechanismen der Translationskontrolle zu verstehen.

Hochdurchsatzsequenzierungstechnologien werden häufig zur Bestimmung und Quantifizierung von DNA- und RNA-Moleküle in großem Maßstab verwendet, was unser systemisches Verständnis von vielen biologischen Funktionen erleichtert hat. Eine Weiterentwicklung dieser Technologie, das Ribosome Profiling, ermöglicht sogar die Bestimmung der an Ribosomen gebundenen mRNA-Fragmente. Ribosome Profiling ermöglicht somit die Translation *in vivo* zu betrachten und erleichtert es die Mechanismen der Translationskontrolle zu ergründen.

Meine Arbeit besteht aus zwei Hauptteilen: 1) Sie beschäftigt sich einerseits mit der Entwicklung von *RiboDiff*, einer statistischen Methode und einem Computerprogramm um die Gene zu bestimmen, deren Translation unterschiedlich zwischen zwei Bedingungen reguliert ist. Zu diesem Zweck modelliert *RiboDiff* quantitative Ribosome Profiling und RNA-Seq Messungen mit einem verallgemeinerten linearen Modell das auf der Negativ-Binomial-Verteilung basiert. Anschließend wird ein statistischer Test durchgeführt, um Gene unter differentieller Translationskontrolle, zwischen den Messungen in beiden Bedingungen, zu identifizieren. Unsere Experimente zeigen, dass *RiboDiff* existierenden Ansätzen überlegen ist. 2) Weiterer Gegenstand der Arbeit ist die Entwicklung einer Pipeline für die Analyse von Ribosome Profiling-Daten. Mit dieser Pipeline untersuchten wir die Translationsregulation unter anderem in Leukämie. Mit Hilfe dieser Pipeline konnten wir mRNAs identifiziert, die unterschiedliche Translationseffizienzen und Ribosom-Dichten in Zellen aufwiesen, welche mit dem chemischen Wirkstoff Silvestrol behandelt wurden. Eine weitergehende Analyse dieser mRNAs ergab, dass das $(GCC)_4$ Guanin-Quartett Sequenzmuster, das eine G-Quadruplex-Struktur induziert, in der 5'-UTR von 280 mRNAs mit herunterregulierter Translationseffizienz angereichert ist. Eine experimentelle Untersuchung dieser Motive bestätigte unsere Ergebnisse und zeigte, dass G-Quadruplexe ein Translationsrepressor sind.

Acknowledgements

First, I would like to express my deep gratitude to Gunnar Rätsch who accepted me as a graduate student in his lab in 2011 and supervised me afterwards. I benefit quite a lot from his inspirations, intuitions and suggestions. I also appreciate Oliver Kohlbacher who agreed to be my second advisor and provided guidance and feedback during my study. I also deeply thank Karsten Borgwardt and Daniel Huson who came to U.S. to conduct the coursework exams for me.

During my stay in Gunnar's lab, I have met many people and all of us had an enjoyable time. Philipp Drewe helped me a lot on translational control project. In addition to his guidance, it was fun to invest energy on research together with him. Jonas Behr helped me to learn many technical skills in my first year. Vipin Sreedharan and David Kuo offered me countless help, including coding assistance, life experience sharing *etc.* David also kindly helped for proof readings many documents. I very much thank Theofanis Karaletsos's patient explanation of math and help in writing manuscript. André Kahles and Kjong Lehmann are always happy to discuss and provide suggestions which I benefited a lot. I also would like to thank other people for their help and inspirations throughout the doctoral study. They are Kadeem Ho Sang, Sarah Danes, Quaid Morris, Marius Kloft, Julia Vogt, Darya Karelina, Kana Shimizu, Xinghua Lou, Christian Widmer, Yuheng Lu, Natalie Davidson, Stefan Stark, Stephanie Hyland, Behrouz Tajoddin, Antonija Burcul, Linda Sundermann, Katherine Chan, Yun Yan, Lijie Tu, Xuefan Gao, Achim Walzer and Karin Stark.

In the past four years, our collaborations with wet-lab experimentalists were fruitful due to creative discussion and their hard working. They are Guido Wendel, Kamini Singh, Andrew Wolfe, Viraj Sanghvi, Lionel Ivashkiv, Xiaodi Su and Yingpu Yu. Additionally, I appreciate Volker Tresp, Florence Demenais and Karsten Borgwardt generously offered me the internship opportunity in the Marie Curie ITN network with coordinate assists from Birgit Knapp. I also gained much knowledge in discussions with ITN students: Meiwen Jia, Cristóbal Esteban, Melanie Pradier and all others within the network.

Lastly, I deeply thank my wife, Hongjie Hou, and my little son, Bowen Zhong, who firmly accompany me every day and encourage me, by either words or smiles.

Contents

1 Introduction

1.1 DNA and Genetic Information

DNA, or deoxyribonucleic acid, is a macromolecule composed of the monomers called nucleotide. Each nucleotide unit is made up of a nucleobase, a deoxyribose and one phosphate group [1, 2]. Typically, four different nucleotides exist in the DNA strand. They are adenine (A), thymine (T), guanine (G) and cytosine (C), depending on what nucleobase is attached to deoxyribose. The nucleotides are connected to one another through the chemical bonds between the deoxyribose of one nucleotide and the phosphate of the next. In the cell, the linear sequence of nucleotides along the DNA encodes all the genetic information a cell needs to have during its life time, which holds true for all three domains of life, but not including RNA viruses [3].

The DNA is in a helical shape with double-strand form as its primary structure. The two strands are reverse complementary, following the matching rule between a pair of bases on the two chains: A pairs with T, and C pairs with G, also called as Watson-Crick base pair [4]. In human, each cell contains 3.2×10^9 base pairs of DNA [5]. This genetic blueprint is known as the human genome. It has been estimated that if stretch out the DNA from a single human cell, it is about 2 meters in length [6]. The question is how the DNA is packed small enough to fit the space in the cell. Studies have revealed that, in eukaryotes, the DNA is spirally coiled with histone proteins and forms the basic unit of DNA packaging—nucleosome [7]. The nucleosome is further coiled to form a complex of macromolecule with a loose structure—chromatin [8]. Consequently, the later is tightly packed with other proteins resulting a highly compact but well organized structure called chromosome [9].

In human, the diploid cell contains 23 pairs of chromosomes, one in each pair is inherited from the mother and the other is from the father [10]. Although the somatic cells carry the same genomic DNA, the genetic material is used in many different ways depending on the type of the cells. A gene is a piece of DNA that encodes a biological function. All the cellular processes are under the precise control of the gene expression, namely, the on and off status of a gene. The gene transcription copies the genetic information from the DNA to messenger RNA and further downstream molecules that contribute to the diverse functions of the cell.

1.2 Transcription

Transcription is a cellular process by which the genetic information is transferred from DNA to ribonucleic acid (RNA) by RNA polymerase. This process, also called gene expression, is the first step that cell produces its functional molecule—protein. Therefore, it is under sophisticated regulation that differs from cell to cell, and time to time [1, 11].

In eukaryotes, as the DNA is wrapped around histone proteins and tightly packed as the form of chromosome, the gene transcription firstly starts with the modification of chromosome architecture at the small region where the gene of interest is located, resulting a loose and open chromatin state, known as chromatin remodeling, to facilitate RNA transcription machinery proteins to access [12, 13].

Transcription initiation takes place at the promoter region, typically 25 nucleotides upstream of a gene, where a specific DNA sequence is firstly recognized and bound by transcription factor TFIID. Afterwards, other proteins, such as RNA polymerase II, TFIIA, TFIIB, TFIIE, TFIIF *etc.*, are subsequently assembled into the transcription initiation complex around the promoter and then triggers the transcription [1]. In eukaryotic cell, RNA polymerase II is responsible for transcribing the vast majority of protein coding genes, whereas RNA polymerase I and III transcribe genes encoding ribosomal RNA, transfer RNA and other small RNAs [14, 15, 16].

Once the transcription initiation complex is assembled, RNA polymerase II begins to elongate the RNA transcript. Briefly, RNA polymerase II unwinds the local DNA helix and uses the antisense strand of the DNA as the template to incorporate ribonucleotides from four nucleoside triphosphates (ATP, CTP, UTP, and GTP) as substrate. The incorporated ribonucleotide is determined by the complementary base-pairing to the nucleotide on the DNA template. Note, on the RNA the base uracil (U) instead of thymine (T) matches to adenine (A) in DNA [17]. Roughly, the RNA polymerase moves 20 nucleotides per second for eukaryotic transcription in the direction of 3' to 5' on the antisense DNA strand [1], namely the nascent transcript is generated from its 5' end to 3' end. Therefore, the RNA sequence is the same as the gene on the sense strand of the DNA with T replaced by U.

While RNA polymerase approaches to the end of a gene, two proteins—CstF and CPSF—bind to specific sequences on the RNA molecule [18]. Then the RNA is cleaved at the cleavage site, and up to 200 adenines are added by poly-A polymerase to the 3' end of the cleavage site [19]. Once the transcription ends, the RNA polymerase detaches the DNA template and can be involved in initiating another transcription. It is possible that a gene is under transcription by multiple RNA polymerases to rapidly produce the transcripts to meet the needs of the cell. However, on average 10-15 copies of each transcript present in a single cell [1].

The RNA that is being generated by RNA polymerase is called pre-RNA. The pre-RNA needs to be processed to become a mature RNA. The RNA processing is always coupled with transcription [20]. It includes the capping on the 5' end of the nascent pre-RNA [21], splicing out the noncoding introns and concatenating the exons [22], and polyadenylation of the 3' end [19] as describe previously. The processed RNA is called the messenger RNA (mRNA), which is exported from nucleus to cytoplasm and can guide to produce the proteins. There are also some other mature RNAs directly execute their functions instead of translating into proteins, such as rRNA, tRNA, miRNA, lincRNA, *etc.*

The strength of the transcription depends on a few aspects. For instance, a favorable promoter sequence by the transcription factors; an enhancer element far away from the promoter region; and also some additional activators or repressors can regulate the gene expression and change the concentration of mRNA molecules [1].

Unlike the DNA formed as a double-stranded molecule connected by hydrogen bonds between the two strands, RNA is a single strand molecule with shorter length. However, the ribonucleotides on the linear strand can pair with each other, therefore, fold into variety of secondary structures, such as stem, loop, hairpin, *etc.* In addition, some nonconventional base-pair interactions even result in folding RNA molecule into more stable three-dimensional structures [23].

1.3 Translation

The protein is a workhorse that performs vast majority of the biological functions in the cell, including catalyzing biochemical reaction, cross-membrane transportation, signal transduction, *etc.* In addition, it is also the building blocks of the cell. Translation is the process that decodes the genetic information in mRNA into the amino acid and produces the protein.

The genetic code carried by mRNA includes four types of nucleotides (A, C, G, U). However, 20 different amino acids are commonly found in the natural proteins. It has been proved that three consecutive nucleotides on mRNA comprised of the codon that determines one amino acid type [24]. As the codons are redundant ($4^3 = 64$), some amino acids are determined by more than one triplet [25].

The protein translation takes place on the ribosome in cytoplasm. Ribosome is a large molecular complex with sophisticated structure. It contains two subunits—a 40S small subunit including an 18S RNA and 33 ribosomal proteins, and a 60S large subunit consists of a 5S RNA, a 5.8S RNA, a 28S RNA subunits and 46 proteins [26, 27, 28]. The ribosomal RNAs are responsible for catalyzing the synthesis of the peptide, whereas the ribosomal proteins stabilize the structure of ribosome [1]. The two subunits are assembled in nucleolus [29] and relocated to cytoplasm or attach onto the membrane of endoplasmic reticulum [30]. It has been reported that the number of ribosome in human Jurkat cell is about 2 million [31].

Ribosome itself is not able to match amino acids to the codons on mRNA. This task is performed by a unique type of RNA molecules known as transfer RNAs (tRNAs). The 80 nucleotide long tRNA forms the structure in a shape of cloverleaf with the anticodon on its one end and the other end is a short single-stranded region where the amino acid is attached to [32]. The aminoacyl-tRNA synthetase recognizes the amino acid and covalently links it to the appropriate tRNA that harbors the correct anticodon corresponding to the amino acid [33]. Therefore, both anticodon on tRNA and the aminoacyl-tRNA synthetase ensure the genetic code is converted to amino acid correctly.

In eukaryotic cell, the first step of translation is initiation, which is the rate-limiting step of the entire translation [1]. Briefly, a specific initiator tRNA that carries a methionine is firstly assembled onto the small subunit of ribosome with other translation initiation factors (eIFs) such as eIF2, eIF1 and eIF5. This preinitiation complex is then attached to eIF4E and eIF4G that have already bound to the 5’ cap of mRNA [34]. Next, the complex moves to the downstream along mRNA, scanning the start codon (AUG) with the assist of eIF4A to go through the secondary structures on the 5’ untranslated region (UTR) [35]. Once the small subunit reaches the start codon, the initiation factors are dissociated and the ribosome large subunit is assembled with small subunit to form the complete ribosome. The detail of translation initiation can be found in Figure 1.1.

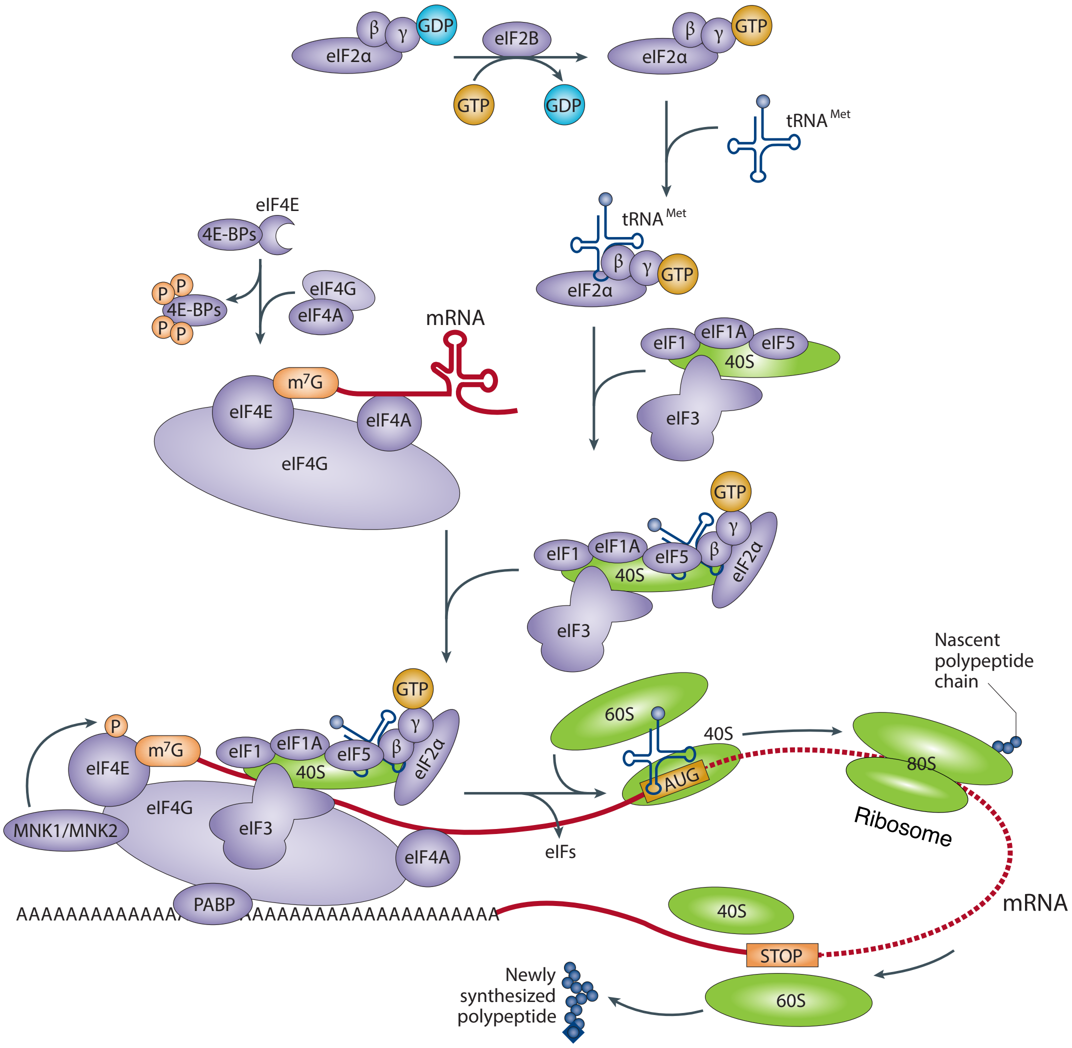

Figure 1.1: Diagram of the cap-dependent eukaryotic translation initiation. This figure is based on the original version in Mamatha Bhat, *et al.* *Nat Rev Drug Discov*, 2015 [36] with modifications.

The ribosome provides three sites to tRNAs while elongates the peptide [1]. The initiator tRNA [37] occupies the ribosomal P-site with its anticodon complimentarily match to the AUG on the mRNA. After the next tRNA carrying the second amino acid occupies the ribosomal A-site, the peptidyl transferase transfers the methionine from the P-site tRNA to the C-terminal of the amino acid on the A-site tRNA [38]. Then ribosome undergoes confirmation change to move mRNA three nucleotide forward, which shifts the two tRNAs to the E- and P-sites and leaves an empty A-site for the third tRNA to enter. During the elongation step, the two factors (EF-Tu and EF-G) remarkably accelerate the translation [39]. It was estimated that, in eukaryotic cell, the ribosome extends two amino acids per second in average [1].

The elongation ends when the stop codons (TAG, TTG, TGA) are encountered at the A-site. The release factor binds to ribosome upon this signal, which triggers the newly synthesized peptide to be released from the ribosome [40]. Once the mRNA is also released, the ribosome separates itself and the two subunits can enter another translation cycle.

Although protein synthesis is the most energy-consuming process in the cell, translational control can provide quicker adjustment directly to cellular alterations than transcription. First of all, the intact cap structure and poly(A) sequence on the two ends of mRNA are the most essential elements that promote the translation [1]. The cap-binding affinity of eIF4E has been recognized having globally impact on translation efficiency. The mTORC signaling pathway can inhibit the protein 4E-BP and free the eIF4E to bind the cap and initiate translation [41]. Other initiation factors have also been reported to play a role in translational control. For instance, the phosphorylation of eIF2α subunit down-regulates translation by blocking the GTP-GDP exchange on eIF2 at the initiation stage [42]. The sequence features and structures on mRNA also provide alternates for translational control. The internal ribosome entry site (IRES) in the 5' UTR or even the coding exons allows ribosome to synthesize a protein in different rate [43]. Other translational controls such as repression of specific genes' translation by their upstream open read frames (uORFs) as well as miRNA involved translational regulation indicated a diverse spectrum of controlling the gene expression at the translation level [44].

1.4 RNA-Seq

In the past 10 years, the revolutionary high throughput sequencing technologies have comprehensively changed the way we think and conduct our biological research. In contrast to the traditional Sanger sequencing method, high throughput sequencing, or deep sequencing, is capable to sequence the DNA or RNA at very high coverage at affordable cost [45]. RNA-Seq is a technology using high throughput sequencing platform to determine the whole transcriptome of the cell samples [46, 47]. It differs from DNA sequencing in how the library is prepared. Currently, RNA-Seq is the state-of-the-art strategy to study transcriptome such as expression quantification, alternative splicing, *de novo* transcript assembly, and small RNA identification, *etc* [46, 48].

Several commercially available platforms can be used to perform RNA-Seq experiment, including FLX pyrosequencing system from Roche 454 [49], Illumina Genome Analyser [50], PacBio System [51] and AB SOLiD system [52]. The first three are based on sequencing by synthesis approach, whereas AB SOLiD system employs sequencing by ligation.

The first step of RNA-Seq is sample preparation. After extract the RNA component from sample cell, the DNA contamination is removed by using DNase, and ribosomal RNA is removed by Poly(A) mRNA selection [53] or hybridization-based rRNA depletion [54]. The remaining RNA molecules undergo fragmentation, then followed by reverse transcribed into cDNA.

The next step is template preparation. This step can be different depending on which sequencing platform is used. Here we take the Illumina Genome Analyser as an example to explain the process. After fragment size selection, the cDNA is ligated with the adaptors containing universal priming sites to both ends of the fragments. Then a PCR reaction with a few cycles is performed to enrich the successfully ligated cDNA. Next, the single-stranded fragments containing adaptors are immobilized onto a solid surface—flow cell—which is coated with the adaptors in advance. Then, the solid-phase amplification is performed to produce 100-200 million spatially separated template clusters [45].

The last step is sequencing and imaging. Illumina Genome Analyser uses the reversible terminator chemistry to determine the nucleotides sequence of the template. Briefly, four types of fluorescently labeled reversible terminator nucleotides are added to synthesize the complementary strand of the template cluster obtained from previous step. Here, the DNA chain is extended one nucleotide at a time. The non-incorporated nucleotides are washed away. A camera takes images of the fluorescently labeled nucleotides. Then the fluorescent dye, along with the terminal 3' blocker, is chemically removed (cleavage step) in order to restore the 3-OH group, allowing for the next cycle to begin [45].

In the end, RNA-Seq platforms generate a file containing millions of the short fragment sequences—reads, and the quality of every nucleotide on each read, which can be used in different way depending on what types of analysis needs to be done. RNA-Seq is a technology that not only provides the number of reads to quantify the gene transcription, but also reveals the nucleotide resolution insight of the transcripts, which facilitates the transcriptome assembly, alternative splicing identification and gene fusion study [46, 55].

1.5 Ribosome Profiling

Ribosome profiling is a deep sequencing based technology that globally monitors the protein synthesis by sequencing the mRNA fragment that is occupied by ribosome. Therefore, it provides tens of millions of quantifiable ribosome footprints as the snapshot of the translatome of the cell sample. This technology was firstly established by Nicholas Ingolia and Jonathan Weissman in 2009 [56]. It filled the technical gap between RNA-Seq and Mass spectrometry which genome-wide quantify the transcriptome and proteome respectively.

Ribosome profiling shares the same sequencing strategy with RNA-Seq, but the preparation of the sequencing libraries differs a lot between the two protocols [58]. The first step of ribosome profiling is to obtain the cell lysates in which the ribosomes are immobilized on the mRNA by translation elongation inhibitor cycloheximide [59]. Because the translation initiation is the rate-limiting step of protein synthesis [1], after treat the cell with cycloheximide, the likelihood to have ribosome freezed at the first 10 codons is relative high. To digest the mRNA that is not occupied by ribosome, nuclease (RNase I) is added into the cell lysate and the ribosome with the protected mRNA fragment are selectively enriched. Next, the ribosomes are separated from the footprints by sedimenting through a sucrose cushion. The footprints are subjected to serials of manipulation including linker ligation, reverse transcrip-

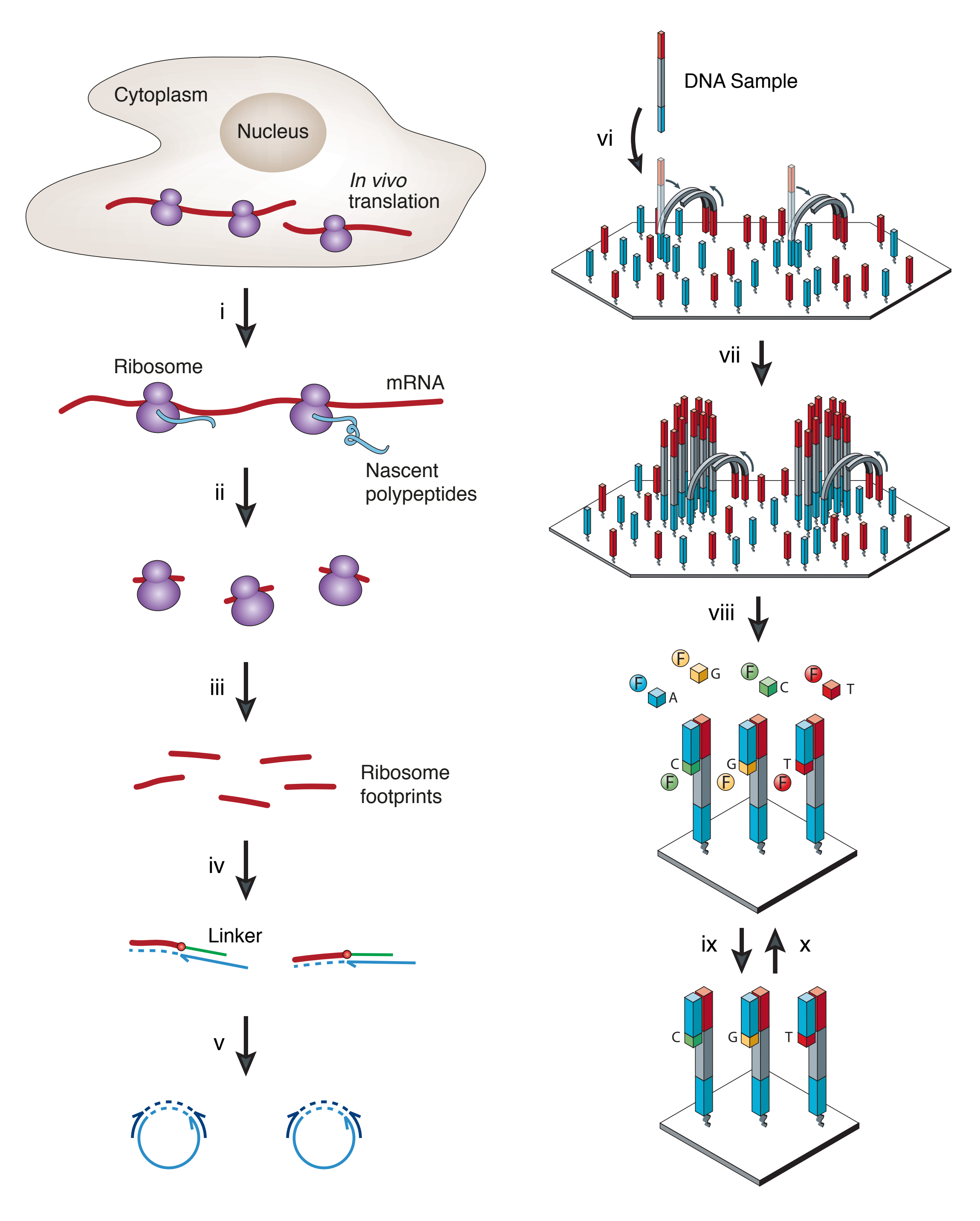

Figure 1.2: Illustration of ribosome footprinting sample preparation and Illumina's high throughput sequencing. i, cell lysate preparation; ii, RNase digestion; iii, Footprint recovery; iv, Linker ligation and reverse transcription; v, Circularization; vi, DNA template immobilization on the flow cell; vii, Bridge amplification to obtain

Figure 1.2: the DNA clusters (Each cluster represents an original DNA template. Up to 200 million clusters are on a single flow cell.); viii, Synthesis DNA by adding one nucleotide labeled with dye; ix, Recording the color of dye, cleaving dye and terminating groups on the nucleotides, wash; x, repeat step viii, ix until the end of sequencing reads. The figure is based on the original versions in Nicholas Ingolia, *Nat Rev Gene* 2014 [57] and Michael Metzker, *Nat Rev Gene* 2010 [45] with modifications.

tion and circularization. Lastly, the ribosomal RNA contamination is removed by hybridization to biotinylated sense-strand oligonucleotides followed by removal of the duplexes through streptavidin affinity. The remaining footprint samples are sequenced by deep sequencing platform such as Illumina HiSeq [60], and the $25 - 35$ bp long single-end footprint sequences are obtained in the end. Figure 1.2 shows the detailed steps of the entire ribosome profiling and deep sequencing workflow.

A few points need to pay attention when analyze the ribosome profiling data. First, always check the quality of sequencing results. Remove any over-represented reads, for instance, the linker sequence. Also, estimating the proportion of ribosomal RNA reads is recommended. Although rRNA depletion is performed before sequencing, rRNA contamination can still exists which reduces the chance to sequence the real footprints. The read length after trimming the linker sequence should be evaluated. Read length shorter than 20 bp needs to be filtered out. As the 30 bp single-end footprint easily results multi-mapping, only using uniquely aligned reads can avoid the ambiguous alignment.

Ribosome profiling provides a quantitative manner to monitor the translation of protein coding genes [57, 61]. Together with RNA-Seq measurement, it sheds lights on the genes translation efficiency [62, 63] and can reveal the translational control independent of transcriptional regulation. As ribosome profiling also provides the footprints at the nucleotide resolution, it reflects the ribosome density along the mRNA. Therefore, by searching for the location where ribosome stalls, we can postulate the biological reason that leads to the change of footprint density distribution [64], for instance, special RNA structures. Furthermore, by treating the cells with translation initiation inhibitors, such as harringtonine [62] or lactimidomycin [65], ribosome profiling facilitates study the alternative start codons under certain conditions as well as identification of short upstream open reading frames (uORFs).

1.6 Probability Distributions

1.6.1 Definitions and Properties

In probability theory, a *variable* is a symbol (x, y, *etc.*) that denotes a mathematical object, which could be any specified set of values, such as a number, a vector or a matrix. If the value of a variable is subject to variations due to effect of randomness, that variable is a *random variable.* A *probability distribution* is an equation that links the value of random variable with its probability of occurrence. Sample distribution and population distribution are the two types of probability distributions. The frequency of random variable occurs in one or a few experiments is a sample distribution, for instance, the read count frequency of ribosome profiling of the genes in one organism. Whereas, the frequency of random variable occurs in infinite experiments is the population distribution. Obviously, read count frequency of ribosome profiling in all organisms is a population distribution.

The random variable can be discrete or continuous. A *discrete random variable* only takes finite number of values. In contrast, *continuous random variable* is defined as an interval where infinite numbers of values in that interval shape the entire random variable. For instance, the read count of RNA-Seq or ribosome profiling is discrete, while the fold change of gene expression is a continuous random variable. The sum of the probability for both discrete and continuous random variable in the sample space X is equal to 1:

- for discrete variable, $\sum\limits_{x_i \in X} f(x_i) = 1$,
- for continuous variable, $\int\limits_{\text{all } x} f(x)\,\mathrm{d}x = 1$.

The *expected value* (*mean* μ) of a random variable X is the sum of all possible values of the product of the random variable and its frequency:

- for discrete variable, $E(X) = \mu = \sum\limits_{x_i \in X} f(x_i) \cdot x_i$,
- for continuous variable, $E(X) = \mu = \int\limits_{\text{all } x} f(x)\,\mathrm{d}x \cdot x$.

The expected value represents the location of the probability distribution. We also need an estimates that represents how spread out the random variable is. To measure the dispersion, the most common used estimate is the *variance*. The variance is calculated as the average squared deviation of each variable from the expected value:

$$\mathrm{Var}(X) = \sigma^2 = E((X - E(X))^2) = E(X^2) - (E(X))^2 \tag{1.1}$$

Therefore,

- for discrete variable, $\mathrm{Var}(X) = \frac{1}{n} \sum\limits_{x_i \in X} (x_i - \mu)^2$
- for continuous variable, $\mathrm{Var}(X) = \int\limits_{\text{all } x} x^2 \, f(x) \, \mathrm{d}x - \mu^2$

Usually, the function used to describe a discrete probability distribution is called a *probability mass function* (*pmf*). And the function for continuous probability distribution is called *probability density function* (*pdf*). In the next section, I will highlight some frequently used probability distributions in biological research.

1.6.2 Discrete Probability Distributions

Binomial distribution If n numbers of statistic experiments are performed, each individual experiment with probability p resulting in outcome A, and $1 - p$ for outcome B, and experiments are independent, then the probability to observe x numbers out of n trails showing outcome A follows a *Binomial distribution*, namely, $X \sim \mathrm{B}(n, p)$. If $n = 1$, the binomial distribution is called a *Bernoulli distribution*.

The probability mass function of Binomial distribution is given by:

$$f(x; n, p) = Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad \text{with} \quad \binom{n}{x} = \frac{n!}{x!(n - x)!}. \tag{1.2}$$

One biological example is if a SNP frequency in the population is 0.3, and we randomly select 10 persons from the population, the probability that half of them will have this SNP can be calculated by Binomial probability mass function.

Poisson distribution In a statistic experiment, if the number of possible outcomes is large, and the possibility for each outcome to appear is small, the probability to observe a certain number of the outcome in a fixed time interval follows a *Poisson distribution* with mean λ: $X \sim \mathrm{P}(\lambda)$.

The probability mass function for Poisson distribution is given by:

$$f(x;\lambda) = Pr(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}, \tag{1.3}$$

where e is the Euler's number; $k!$ is the factorial of k; λ is the expected value of X, namely, $E(X) = \lambda$. In Poisson distribution, the variance is also equal to λ, $Var(X) = \lambda$.

The Poisson distribution can be derived as a limit of the binomial distribution. Therefore, it is used as an approximation to the binomial distribution with parameters n and p, when n gets very large and p is small. In this case the $\lambda = n \cdot p$.

Poisson probability distribution is widely used in Biology. For instance, we know the mutation in the genome follows a Poisson distribution. Also, if we do high throughput exon sequencing, the read count of protein coding exons or the untranslated region are also Poisson distributed.

Negative binomial distribution In serials of statistic experiments, where each experiment has probability p for outcome A, and $1 - p$ for outcome B, and experiments are independent of each other, to obtain outcome A for r times, the number of trails (x) to be performed follows a *negative binomial distribution*, $X \sim \mathrm{NB}(r, p)$.

The probability mass function for negative binomial distribution is given by:

$$f(x; r, p) = Pr(X = x) = \binom{x+r-1}{x} p^x (1-p)^r. \tag{1.4}$$

$\binom{x+r-1}{x}$ can also be written as:

$$\binom{x+r-1}{x} = \frac{(x+r-1)!}{x!(r-1)!} = (-1)^x \binom{-r}{x}. \tag{1.5}$$

This is why the name "negative binomial" is given. In biology related study, the negative binomial distribution is often parameterized in terms of its mean μ and dispersion κ. Then we have

$$\begin{aligned} p &= \frac{\mu\kappa}{1+\mu\kappa} \\ r &= \frac{1}{\kappa} \\ f(x;\mu,\kappa) &= \binom{x+\frac{1}{\kappa}-1}{x}\Big(\frac{\mu\kappa}{1+\mu\kappa}\Big)^{x}\Big(\frac{1}{1+\mu\kappa}\Big)^{\frac{1}{\kappa}} \end{aligned} \tag{1.6}$$

Compare to Poisson distribution, negative binomial has an dispersion parameter that does not equeal to its mean μ. The variance and dispersion relation is given by

$$\sigma^2 = \mu + \kappa \cdot \mu^2. \tag{1.7}$$

Therefore, when the dispersion κ is getting very small, the negative binomial distribution is approaching to Poisson distribution:

$$\text{Poisson}(\lambda) = \lim_{\kappa \to 0} \text{NB}(r,p) = \lim_{\kappa \to 0} \text{NB}(r, \frac{\lambda}{\lambda + r}) \tag{1.8}$$

Later we will discuss the read count data for genes from RNA-Seq and ribosome profiling with a few biological replicates follow the negative binomial distribution.

1.6.3 Continuous Probability Distributions

Normal distribution A continuous random variable X follows *normal distribution*, or *Gaussian distribution*, if it has the following probability density function:

$$f(x;\mu,\sigma) = Pr(X = x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1.9}$$

where μ denotes the expected value; σ is the standard deviation. Normal distribution is the most commonly used continuous probability distribution because of *central limit theorem*, which states that if we subsample a random variable following any probability distribution multiple times, the mean of the obtained subsamples follows a normal distribution.

The normal distribution is symmetric around the point $x = \mu$. When the $\mu = 0$ and $\sigma^2 = 1$, the normal distribution becomes a standard normal distribution. The random variable of a standard normal distribution is called a *Z-score*. Any random variable X from a normal distribution can be transformed into a Z-score by:

$$z = \frac{X - \mu}{\sigma} \tag{1.10}$$

t distribution When the sample size is small, the t *distribution* is used to describe the distribution of a sample that drawn from the normal distributed population. The probability density function of t distribution is given by

$$f(x) = Pr(X = x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{1.11}$$

where Γ denotes the gamma function; ν is the number of degrees of freedom, namely, $\nu = n - 1$, with n representing the number of observations. In other words, the t distributions for different sample sizes are different in their shapes. Compare to the normal distribution, t distribution has heavier tails, although it is a symmetric bell shaped.

The t distribution can be used to infer whether a sample with size n, mean $\bar{x}$ and standard deviation s is from a population distribution with mean μ. In this application, a t score is calculated by

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}. \tag{1.12}$$

Chi-squared distribution The sum of squares of k independent random variable from standard normal distribution follows a *chi-squared distribution* with k degrees of freedom. Namely, let $X = \sum_{i=1}^{k} Z_i^2$, we have $X \sim \chi^2(k)$.

chi-squared distribution is very rarely used in modeling natural phenomena. Instead, it is frequently used to do hypothesis testing, such as chi-squared test of goodness of fit of observed data, likelihood ratio test and log-rank test in survival analysis, *etc.*

The probability density function of chi-squared distribution is given by

$$f(x;k) = Pr(X = x) = \frac{x^{\frac{k}{2}-1}\, e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\,\Gamma(\frac{k}{2})} \tag{1.13}$$

Gamma distribution If a continuous random variable follows the following probability density function, the random variable is drawn from a *gamma distribution* with shape parameter α and scale parameter s:

$$f(x;\alpha,s) = Pr(X = x) = \frac{x^{\alpha-1}\, e^{-\frac{x}{s}}}{s^{\alpha}\, \Gamma(\alpha)}, \tag{1.14}$$

where $x > 0$ and $\alpha, s > 0$, under the assumption $X \sim Gamma(\alpha, s)$.

The skewness is one of the properties of gamma distribution. It only depends on the shape parameter α, and can be calculated as $2/\sqrt{\alpha}$. Accordingly, the gamma distribution approaches to Gaussian when the shape parameter is large enough.

In evolutionary biology, the rate of evolution of amino acid sites within a protein differs among each other due to different natural selection pressures. The distribution of these rates follows a gamma distribution as described already [66, 67].

1.7 Probabilistic Model for Ribosome Profiling Data

Similar to RNA-Seq, ribosome profiling is a deep sequencing-based technology that generates millions of short sequences. The standard data analysis pipeline includes alignment of the short sequences to the genome or transcriptome of interest and counting how many sequences are mapped to each gene or transcript. This is a translation quantification process because the *in vivo* translation strength is proportional to the mRNA abundance and the amount of ribosome associated to mRNA molecule. If a gene is actively translated in the cell, the concentration of ribosome protected mRNA fragment within the cell lysate is relatively high compared to other genes, therefore, it has more chance to be sequenced, resulting more sequencing reads of its own.

For the purposes of comparing differential translation or, even more complicated, evaluating translation efficiency change, the read counts is obtained as the quantification measurement for every gene or sub-gene features. Unlike the microarray data, the quantification of deep sequencing based methods ends up with the discrete variables—count data. One option is to transform and standardize the read counts to approximate as Gaussian distributed data set [68]. However, this transformation has drawback towards the low counts that are far from normal distribution. In addition, the transformation loses the mean-dispersion relationship [69, 70] that count data intrinsically retain, leading to potentially inefficient statistical inference. Therefore, a discrete probability model that accounts for the properties of the read count data is more powerful and sensitive than simply transforming to Gaussian in terms of detecting the differential translation.

An mRNA molecule being translated has its specific ribosome footprint concentration μ within the cell. Assuming every footprint fragment has the same opportunity to be sequenced (for example, no GC content bias), the probability that these footprints are successfully detected by ribosome profiling experiment is proportional to the μ with variations due to randomness from the sequencing process. Assuming the sequencing platform has small technical variation, if we sequence the same cell sample multiple times, the resulting footprint counts for a single transcript follows the Poisson distribution centered around the mean μ.

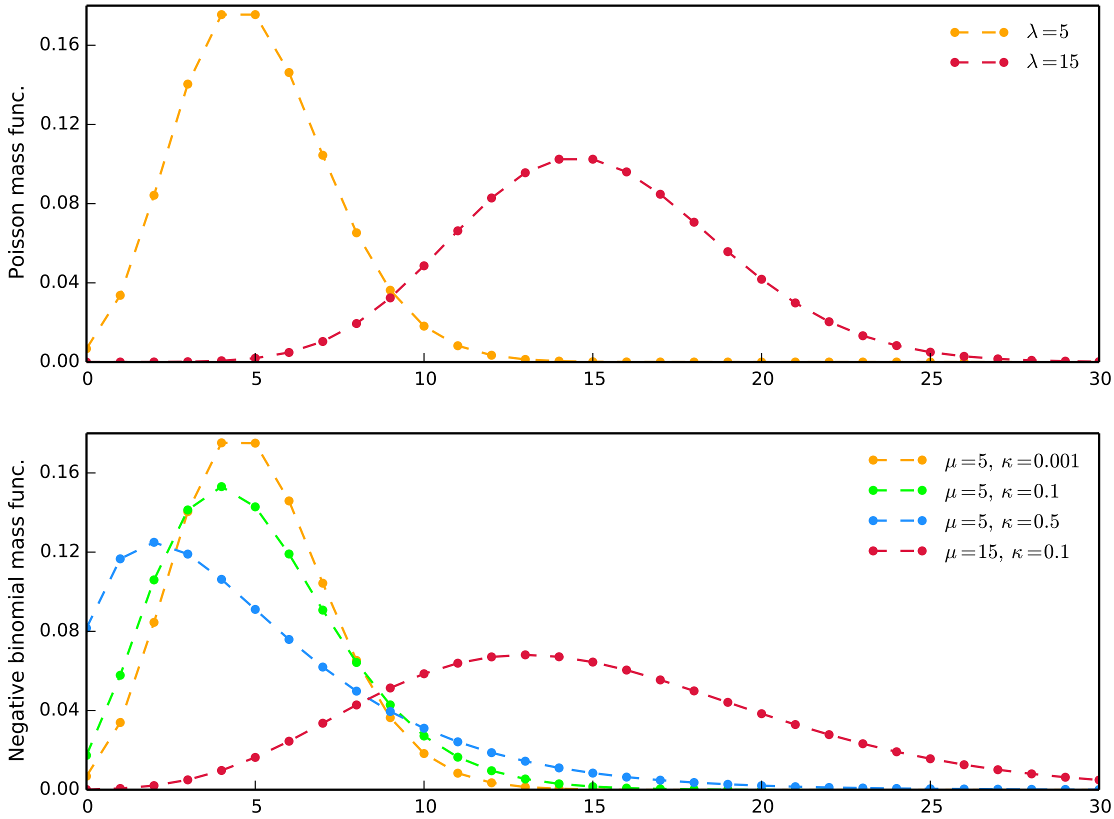

Figure 1.3: Probability mass function of Poisson (top) and negative binomial (bottom) distribution. The mass functions are defined only at integer values on x-axes. The dashed curves connecting each dots are only for easily observation. λ, denotes Poisson mean; μ and κ, denote negative binomial mean and dispersion, respectively.

As described before, in the Poisson model the variance equals to the mean. This feature has limitations when applies Poisson to the discrete count data. Currently, experimentalists usually do two or three biological replicates for each conditions, one can easily observe the large variance (over-dispersion) of the counts between the samples. The variance consists of both biological and technical parts. The observed variance can be very large especially for low read counts and cannot be fully represented by the Poisson variance. Therefore, a more suitable probability model is the negative binomial distribution parameterized by mean μ and dispersion κ. The variance of negative binomial is calculated by adding a second term $\kappa \cdot \mu^2$ to the Poisson mean. Hence, negative binomial model is capable to capture a wide spectrum of dispersion across the replicates and do statistical inference appropriately. In figure 1.3, the probability mass function of Poisson and negative binomial distribution demonstrate the variance of negative binomial can be much larger than Poisson when their means are the same.

1.8 Regression and Generalized Linear Model

Regression In statistical modeling, *regression analysis* is a statistical process that estimates the relationship between the *independent variables* X (or "*explanatory variables*") and the *dependent variables* Y (or "*response variables*"). The estimated relationship is a mathematical function of the independent variables called the regression function. Generally, the regression aims to describing how the mean of the dependent variable $E(Y)$ changes when the independent variable changes, given by $E(Y) = f(X, \beta)$, where β is the unknown parameters to be estimated from the regression system. In regression analysis, it is also essential to characterize the variation of the dependent variable which follows a certain probability distribution. The regression usually falls into two categories: linear model and nonlinear model. In linear model, the regression function is expressed as a linear combination of the one or multiple parameters βs. In contrast, nonlinear model cannot be expressed by the additive form of βs, and numerical optimization algorithms are applied to determine the best fitting.

Linear regression Given data set containing explanatory variable X and response variable Y, if the mathematical relationship between X and Y is linear with Gaussian noise ε added to Y, the relationship is called *linear regression*, $Y = X\beta + \varepsilon$. For instance, the following two expressions are linear regressions because of the linear combination of βs:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{with} \quad i = 1, \ldots, n. \\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad \text{with} \quad i = 1, \ldots, n. \end{aligned} \tag{1.15}$$

In practise, different strategies can be used to estimate the parameter βs from the X and Y, such as ordinary least squares (OLS), generalized least squares (GLS) and total least squares (TLS) *etc.* Among these, ordinary least squares is the most commonly used method. It minimizes the sum of squared residuals, which are the distances between the response variable and its expected value. Linear regression has a closed form of expression for the estimated parameter $\hat{\beta}$:

$$\hat{\beta} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y \tag{1.16}$$

Generalized linear model In the previous section, we introduced the linear regression in which the response variable is assumed to follow the normal distribution. However, in reality, many observed data to be modeled is clearly not normal distributed. For instance, the relationship between two genes in the gene network or metabolic pathway is a binary variable; the read count of mRNA from RNA-Seq experiment is negative binomial distributed. This section introduces the *generalized linear model* (GLM) to accommodate response variable that follows any distribution of exponential family. In GLM, a *link function* that links the response variable to the linear predictor, where the linear combination of βs similarly exists as that in linear regression.

A generalized linear model consists of three elements:

- The response variable follows a probability distribution from the exponential family;
- A linear predictor $\eta = X \cdot \beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$;
- A link function $g()$ which transforms the nonlinear distributed expected value of response variable to the linear predictor, $E(Y) = g^{-1}(\eta) = g^{-1}(X \cdot \beta)$. The canonical link function of some exponential family distributions can be found in Table 1.1.

The most commonly used distributions of exponential family can be expressed in

Table 1.1: Canonical link and response range for the listed distributions.

Probability distribution	Canonical link	Range of Y
Binomial	Logit	0, 1, 2, $\cdots$
Poisson	Log	0, 1, 2, $\cdots$
Negative binomial	Log	0, 1, 2, $\cdots$
Gaussian	Identity	$(-\infty, +\infty)$
Gamma	Inverse	$(0, +\infty)$

the following generalized form:

$$f(y;\theta,\phi) = exp\Big(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big), \tag{1.17}$$

where $a()$, $b()$ and $c()$ are known functions that vary according to different probability distributions. θ is the canonical parameter for the distribution in question. It is related to the mean μ of the distribution. ϕ is the scale parameter. For some probability distributions, it is a fixed value; while in other distributions it is an unknown parameter to be estimated from the data.

The reason for expressing diverse distribution functions in the common exponential form is that general properties of exponential family can be applied to the individual distribution. For instance, the expected value and variance are given by

$$E(Y) = \frac{\mathrm{d}b(\theta)}{\mathrm{d}\theta} = b'(\theta) \tag{1.18}$$

$$Var(Y) = a(\phi)b''(\theta) = a(\phi)\frac{\mathrm{d}^2 b(\theta)}{\mathrm{d}\theta^2} \tag{1.19}$$

where $b'(\theta)$ and $b''(\theta)$ are the first and second derivatives of $b(\theta)$.

The useful feature of a generalized linear model is, for any probability distributions of exponential family, the parameter βs in the linear predictor can be estimated by the same algorithm—*iterative reweighted least squares* (*IRLS*). The goal of IRLS is to minimize the residue for GLM in order to find the best *beta*s:

$$\arg\min_{\beta} \|Y - X\,\beta\| = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - x_i\beta| \tag{1.20}$$

The iterative reweighted least squares includes the following steps:

- First starts with an initial μ^0, and the $\eta^0 = g(\mu^0)$. A simple choice for the initial μ^0 is to set $\mu^0 = Y$;
- At each iteration, a working dependent variable Z is calculated by

$$Z = \eta + (Y - \mu)\frac{\mathrm{d}\eta}{\mathrm{d}\mu} \tag{1.21}$$

- And an iterative weight w is calculated by

$$w = \frac{\phi/a(\phi)}{b''(\theta)(\frac{\mathrm{d}\eta}{\mathrm{d}\mu})^2} \tag{1.22}$$

- Lastly, a $\hat{\beta}$ is obtained by regressing the working dependent variable Z on the predictors X using the weight factor w:

$$\hat{\beta} = (X^{\mathsf{T}}WX)^{-1}X^{\mathsf{T}}WZ, \tag{1.23}$$

  where X is the model matrix; W is a diagonal matrix of weight factor w; Z is the working response variable.

The IRLS procedure is repeated until the algorithm converges at a point where successive estimates βs change less than a specified value.

2 Detecting Translational Control from Ribosome Footprints and RNA-Seq

2.1 Motivation

The deep sequencing based RNA-Seq is a revolutionary technology that tremendously facilitates researchers to quantitatively measure mRNA abundance and infer the transcriptional regulation globally under a given condition [71, 72]. However, the correlation between mRNA and protein abundance is frequently low due to translational regulation and post-translational modification [73]. The recently described ribosome footprinting protocol [58] uses high throughput sequencing to identify mRNA fragments that are occupied by ribosome during protein translation. Therefore, ribosome footprint profiling provides valuable information on protein synthesis. These information includes overall abundance of ribosomes loaded on mRNA, the density of ribosomes at a specific mRNA region and the ribosome pausing events, *etc.* Any alteration of these observable outcomes are the consequences of certain translational regulation. However, to study translational control is nontrivial, because the observed ribosome profiling is fundamentally confounded by mRNA transcriptional activity. To simply illustrate the issue, we can take the overall ribosome occupancy as an example, where the ribosome occupancy can be measured by the footprint read count from original sequencing data. As shown in Figure 2.1C, although the footprint counts in two experimental conditions are drawn from different negative binomial distributions, the transcriptional landscape (Figure 2.1A and B) needs to be considered before making any conclusion on translational control. Briefly, if the confounding factor—transcriptional scenario—shares the similar distribution (panel A) with ribosome profiling (panel C), the probability that translational regulation has observable effect on footprint is low, whereas if the treatment leads to different profiles at transcriptional and translational levels (panel B and C), it indicates a translational regulation exists on top of transcription.

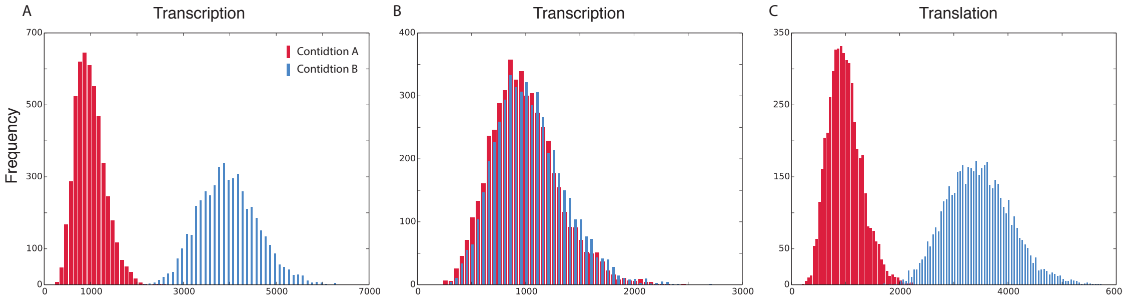

Figure 2.1: Illustration of read counts sampling from different negative binomial distributions at transcription (A and B) and translation (C) levels.

Studies have reported that translational control is essential for cell in response to stresses [74], regulating development [75] and immune reaction [76] *etc.* Remarkable progress has been achieved in demonstrating translational regulation plays a important role in causing disease such as cancer [77, 78, 79, 80]. In order to decipher principles of translational regulation in case-control studies, tools that can reliably detect changes in ribosome footprint taking mRNA activity into account are needed.

In this chapter, I will introduce a statistical framework and analysis tool, *RiboDiff*, that I developed in order to detect genes with changes in translation efficiency across experimental conditions. *RiboDiff* uses generalized linear models to estimates 1) the mean counts from the RNA-Seq and ribosome profiling read counts respectively and 2) the over-dispersion of the biological replicates of these two sequencing protocols separately, and performs a statistical test based on the estimates for differential translation efficiency. Hence, it provides the genes or transcripts that are potentially governed by translational regulation in a given condition.

2.2 Related Methods

2.2.1 Z-score Method

Translation efficiency (TE) is widely used to measure the rate of mRNA translation into proteins within cells. For gene i, it can be calculated by

$$TE^i = \frac{A^i_{RF}}{A^i_{mRNA}}, \tag{2.1}$$

where A represents the true concentration of mRNA and ribosome footprint (RF) fragments in the cell. In the case of high throughput sequencing, the absolute concentration can be approximated by the Reads Per Kilobase of transcript per Million (RPKM) or read counts (y) in the exonic region of gene i obtained from the RNA-Seq and ribosome profiling data [56, 62, 63, 81]. Therefore, the TE fold change (ΔTE) in different conditions is given by

$$\Delta TE^i = \log(\frac{TE^i_{C1}}{TE^i_{C2}}) = \log(\frac{K^i_{RF,A}}{K^i_{mRNA,A}}) - \log(\frac{K^i_{RF,B}}{K^i_{mRNA,B}}), \quad \text{with} \quad K^i_{t,c} = \underset{j}{\text{mean}}(y^{i,j}_{t,c}), \tag{2.2}$$

where t denotes the data type, as $t = \{\text{RF, RNA-Seq}\}$; c denotes the condition, as $c = \{A, B\}$; j indexes the replicates. Therefore, $y^{i,j}_{t,c}$ stands for the t type of read count y of gene i in its j^{th} replicate under condition c and $K^i_{t,c}$ is the mean count of gene i in type t and condition c. In order to identify the candidate genes whose ΔTE remarkably deviate from the mean, a Z-score can be calculated for each gene i by

$$z^i = \frac{\Delta TE^i - \mu_{\Delta TE}}{\sigma_{\Delta TE}}, \quad \text{with} \quad \sigma_{\Delta TE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta TE^i - \mu_{\Delta TE})^2}, \tag{2.3}$$

where $\mu_{\Delta TE}$ is the mean of ΔTE of all genes; N denotes the total gene number; $\sigma_{\Delta TE}$ is the standard deviation. Therefore, the target genes with TE down and up-regulation are defined by identifying those $|z^i| > G$, where G is an arbitrarily chosen cutoff.

2.2.2 Errors-in-Variables Regression

In 2013, Olshen *et al.* developed a tool, *Babel*, which is based on errors-in-variables regression to detect the translational regulation [82]. Similar to the previously published tools for RNA-Seq analysis, *Babel* models read counts using the negative binomial distribution parameterized by mean count and dispersion ϕ.

Specifically, the model treats mRNA abundance as measured with error instead of a fix value. Assume x_g and y_g represent the mRNA and ribosome profiling abundance of gene g respectively and a linear relationship exists between these two variables as $y_g = \hat{\beta} \cdot x_g$, where $\hat{\beta}$ is estimated from a trimmed least squares regression. Next, the genes are split into B bins based on mRNA abundance. The dispersion is then estimated by minimizing the squared error between empirical variance V_b in bin b and the overall dispersion ϕ:

$$\hat{\phi} = \arg\min_{\phi} \sum_{b=1}^{B} (V_b - \phi)^2. \tag{2.4}$$

After mean and dispersion are estimated, the parametric bootstrap is used to gain P-value under the null hypothesis that ribosome profiling follows the expectation from the mRNA-ribosome regression. Finally, the Fisher′s method is used to convert the P-values of each RNA-Seq and ribosome profiling pair in every replicate within a condition into a single consensus P-value, and the changes in translational regulation between conditions is assessed by converting the within-condition P-value to standardized Z-statistics using the Gaussian quantile function.

2.3 RiboDiff

2.3.1 Library Size Normalization

Normalization for deep sequencing data is a critical step before starts doing any further analysis [83]. The systematic biases can come from sample preparation, sequencing, and even downstream pipeline, such as alignment and quantification, *etc.* If the goal of the experimental design is to identify genes that are differentially transcribed or translated in case-control study, the initial mRNA or ribosome occupied mRNA fragment must be prepared in the same amount before sequence the samples. The technical details used in the data parsing pipeline, such as the number of mismatch, supporting splicing reads during alignment, using uniquely aligned or multiple mapped reads to do quantification, can also create biases towards genes. However, the most likely global bias is caused by the different sequencing depths on different flow cell lanes of the high throughput sequencer [83]. As the sample libraries (biological replicates) are usually deposited on different lanes, the coverage of sequencing reads for libraries vary within and between experimental treatments. In order to make the data from different replicates clearly comparable, transformation of the data taking into account the library size that would distort the entire raw data distribution is inevitably needed.

Several deep sequencing-based library normalization methods have been reported [84, 85, 86]. Here we use the metrics similar to [85] with modifications. We calculate the normalization constant (or size factor) for RNA-Seq and ribosome footprinting (RF) libraries separately. The formula is given by:

$$S_T^r = \underset{i:y_T^{i,r}>0}{\text{median}} \left(\frac{y_T^{i,r}+1}{\sqrt[n]{\prod_{j=1}^{n}(y_T^{i,j}+1)}} \right). \tag{2.5}$$

Here, T denotes data type (RNA-Seq or RF); r denotes the r-th sample in data type T which includes replicates of both experimental treatments. $y_T^{i,r}$ is the observed count of type T for gene i in sample r. For all genes in all replicates, we add one to their count value to avoid the degenerate case of setting the geometric mean across all replicates (indexed by j) in the denominator to zero. We calculate the ratios (Ω) of observed counts of all genes in a given sample to the geometric means and determine the median of these ratios whose count is greater than zero as the size factor. Figure 2.2 shows an example from real data (GEO accession: GSE56887) that a large proportion of ratio Ω that is equal to zero in both ribosome footprint and RNA-Seq data. As we know that many genes' transcriptional or translational activities are too low to be detected by deep sequencing, excluding these genes from calculating the size factor can avoid it drifting towards zero.

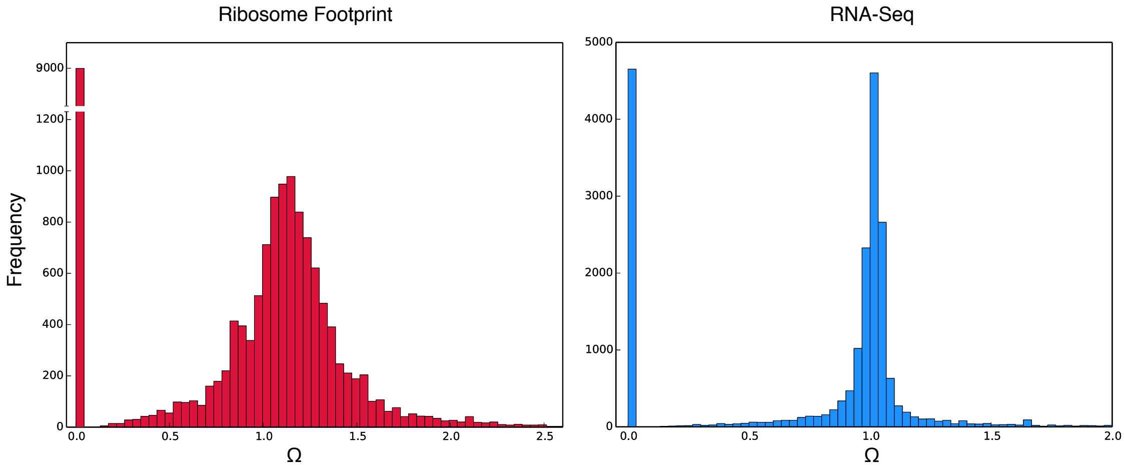

Figure 2.2: Histogram of the ratio Ω in ribosome footprint and RNA-Seq data. The bars at zero indicate the genes without any read count.

With the size factor calculated, the normalized read counts for each gene can be easily obtained by dividing their original counts by the size factor of the corresponding sample.

2.3.2 Estimating the Mean for Count Variables

As described before, the ribosome footprint profiling (RF) is naturally confounded by mRNA abundance. We seek a strategy to compare RF measurements taking mRNA activity into account in order to accurately discern the translational effect in case-control experiments.

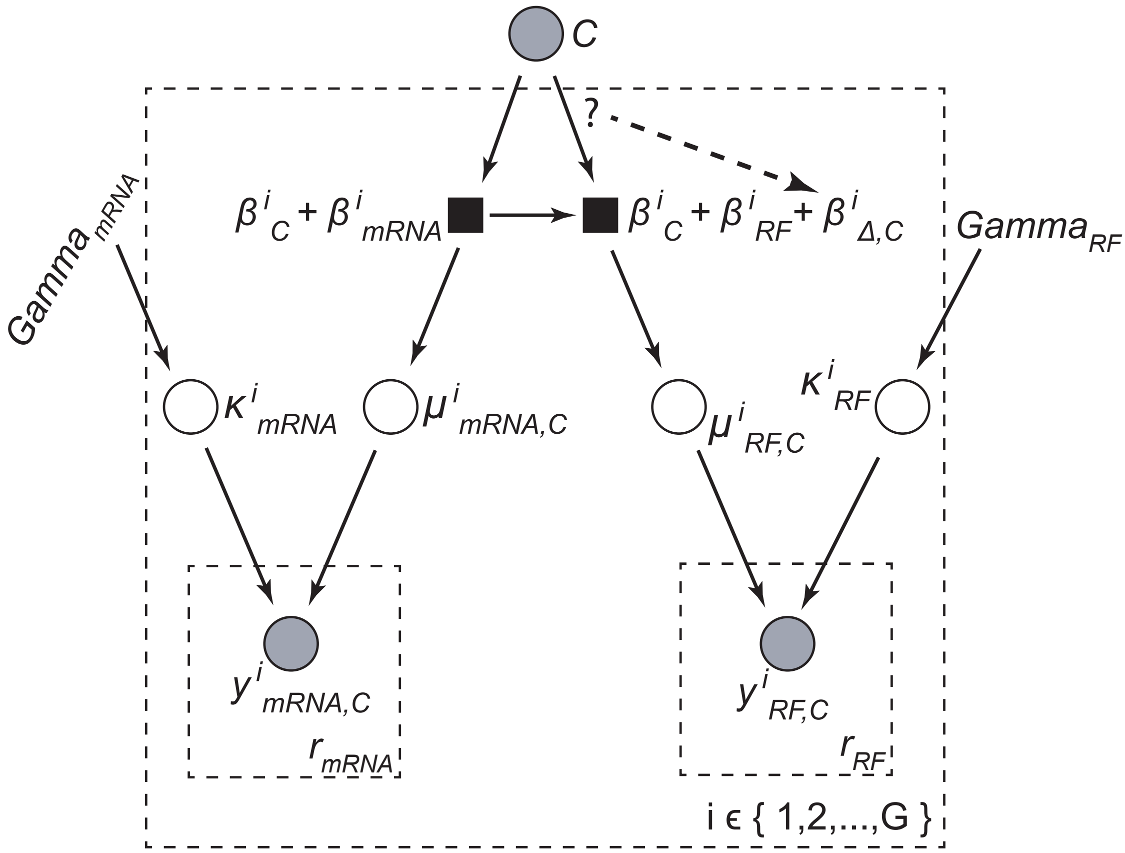

Figure 2.3: Graphical model representing *RiboDiff*. Grey circle: observable variables; empty circle: unobservable variables; black square: functions; arrow: dependency; C: a given experimental condition; G: number of genes; r: biological replicates. The dashed arrow denotes the relationship that we aim to test.

The graphic model of our method, *RiboDiff*, is highlighted in Figure 2.3. Briefly, assuming an experimental treatment C perturbs the transcriptional activity of gene i and the effect of this perturbation is spontaneously passed from transcription (black square on the left side) to the translation (black square on the right side), as shown by the arrow between the two squares. Next, a question we can ask is that whether treatment C affects the translation of gene i directly. In other words, does the dashed arrow from treatment pointing to the translation exist? In this system, we only observe the quantification of gene as the form of read count in each replicate r at mRNA and RF levels. To answer this question, we first estimate the mean μ^i of the count variable y^i, which probabilistically follows the negative binomial distribution [87, 85, 86] as:

$$y^i \sim NB(\mu^i, \kappa^i), \tag{2.6}$$

where κ^i is the dispersion parameter across the biological replicates needs to be estimated later. Here, I am describing how we couple the two biological processes together and perform the statistic inference.

We formulate the problem as a generalized linear model (GLM) with the logarithm as the link function:

$$\log(\mu) = \eta = X \times \beta, \tag{2.7}$$

where η is the linear predictor of GLM. X is the explanatory matrix. β is the coefficients or latent quantities. In particular, at transcription level, we express expectations on read counts as a function of a latent quantity β_C that represents the baseline mRNA abundance in the two conditions ($C = \{0, 1\}$) plus the latent quantity β_{mRNA} that relates mRNA abundance to RNA-Seq read counts:

$$\log(\mu^i_{mRNA,C}) = \beta^i_C + \beta^i_{mRNA} \tag{2.8}$$

We assume that transcription and translation are successive cellular processing steps and that abundances are linearly related. Therefore, the expected ribosome footprint read count $\log(\mu^i_{RF,C})$ is given by:

$$\log(\mu^i_{RF,C}) = \beta^i_C + \beta^i_{RF} + \beta^i_{\Delta,C} \tag{2.9}$$

A key point to note is that β^i_C is revealed to be a shared parameter between the expressions governing the expected RNA-Seq and RF counts. It can be considered to be a proxy for shared transcriptional/translation activity under condition C in this context. The term β^i_{RF} relates mRNA abundance to RF read counts. Then,

$\beta^i_{\Delta,C}$ indicates the deviation from that activity under condition C, with $\beta^i_{\Delta,C} = 0$ for $C = 0$ and free otherwise.

We use a generalized linear model (GLM) to learn the latent quantities βs from the observed count data, and then calculate the means. To control the observed read counts fitting into the GLM system, an $n \times 5$ explanatory matrix X is designed, where n equals to the total number of replicates in both experimental conditions for RNA-Seq and RF. Here we show it in the context of linear predictor η of GLM:

$$\eta = \begin{array}{c} \begin{array}{ccccc} C0 & C1 & mRNA & RF & \Delta_{Eff.} \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \times \begin{bmatrix} \beta^i_{C=0} \\ \beta^i_{C=1} \\ \beta^i_{mRNA} \\ \beta^i_{RF} \\ \beta^i_{\Delta} \end{bmatrix}. \tag{2.10}$$

In X matrix, the first two columns represent the baseline mRNA abundance β^i_C in the two conditions. The third and fourth columns (β^i_{mRNA} and β^i_{RF}) define whether the counts are from RNA-Seq or RF, respectively.[1] The fifth column ($\beta^i_{\Delta,C}$) relates the RF count to the potential translational effect. Each row of X is used to control how the observed count of a specific sample should be decomposed into latent quantities in order to fit the GLM models. In this example, as indicated by the third column, the first four rows (marked in blue) model RNA-Seq counts with *two replicates* for each condition, $C0$ and $C1$, while the last six rows (marked in green) model RF counts with *three replicates* for each condition. Note the first and second columns in X are shared between RNA-Seq and RF counts, where we couple the two different data sets. The linear predictor η then is linked with negative binomial distributed mean $\mu^i_{RF,C}$ and $\mu^i_{mRNA,C}$ through logarithm as the link function, namely $\log(\mu) = \eta = X \times \beta$. The βs are estimated by maximizing the likelihood of GLM

[1] In the implementation, in order to keep full rank of X, we do not include the fourth column β^i_{RF}, as it is linearly dependent with the third column.

[88] (See Chapter 1 for details).

After the β^i is estimated for each gene, the expected RNA-Seq and RF counts, $\mu^i_{mRNA,C}$ and $\mu^i_{RF,C}$ can be obtained. The next step is to estimate the dispersion parameter κ^i by maximizing the negative binomial likelihood function with the observed read counts and the expected counts.

2.3.3 Estimating the Dispersion for Count Variables

The discrete variable of read count from deep sequencing data follows a negative binomial distribution with parameter mean μ and dispersion κ. Therefore, we estimate κ given observed counts and the previously estimated mean μ by maximizing the NB likelihood function.

The probability mass function of the negative binomial distribution is given by:

$$Pr(y^{i,j}) = \binom{y^{i,j} + 1/\kappa^{i,j} - 1}{y^{i,j}} \left(\frac{1/\kappa^{i,j}}{1/\kappa^{i,j} + \mu^{i,j}}\right)^{1/\kappa^{i,j}} \left(1 - \frac{1/\kappa^{i,j}}{1/\kappa^{i,j} + \mu^{i,j}}\right)^{y^{i,j}}, \quad (2.11)$$

where $y^{i,j}$ is the observed RF or RNA-Seq read count of j^{th} replicate of gene i; $\kappa^{i,j}$ is the dispersion parameter of the NB distribution where $y^{i,j}$ is drawn from; $\mu^{i,j}$ is the estimated count of j^{th} replicate. Thus the logarithmic likelihood of negative binomial of gene i is given by

$$\log \ell_{NB} = \sum_{j=1}^{n} \log(Pr(y^{i,j})) - \frac{1}{2}\log(\det(X' \cdot \mathrm{diag}(\frac{\mu^i}{1+\mu^i\kappa^i}) \cdot X)). \quad (2.12)$$

Note that the likelihood function is adjusted by a Cox-Reid term as suggested by Robinson *et al.* [89] to compensate bias from estimating coefficients in fitting GLM. Again, X is the explanatory matrix; n is the total number of RNA-Seq and RF replicates; μ^i is the vector of estimated counts; κ^i is the dispersion vector.

2.3.4 The Mean-Dispersion Relationship

Fitting the GLM consists of learning the parameters β^i and dispersions κ^i given mRNA and RF counts for the two conditions $C = \{0, 1\}$. We perform alternating optimization of the parameters β^i given dispersions κ^i and the dispersion parameters κ^i given β^i, similar to the EM algorithm:

$$\beta^i = \arg\max_{\beta^i} \ell_{glm}(\beta^i|y^i, \kappa^i) \quad \text{and} \quad \kappa^i = \arg\max_{\kappa^i} \ell_{NB}(\kappa^i|y^i, \mu^i). \tag{2.13}$$

The raw dispersion estimated from each gene carries large sampling variance due to limited number of replicates used in the previous two likelihood maximizations. As proposed by Anders *et al.* [70], a systematic trend of dispersions κ_F as a function of the mean is given by:

$$\kappa_F^i = f(\mu) = \lambda_1/\mu^i + \lambda_0 \tag{2.14}$$

To obtain the coefficients λ_1 and λ_0, we regress the raw dispersion κ^i given the mean counts use the GLM with Gamma exponential family distribution. An example of the regression result is shown as the red curve in Figure 2.4.

2.3.5 Finalize Dispersion Estimation

To get the final dispersion κ_S^i, we follow the empirical Bayes Shrinkage approach published by Love *et al.* recently [90]. This approach is based on the observation that the dispersion follows a log-normal prior distribution [91] centered at the fitted dispersion κ_F^i. Moreover, the variance (σ_w^2) of the logarithmic residual between raw dispersion κ_R^i and κ_F^i is comprised of 1) the variance of sampling distribution of the logarithmic dispersion σ_x^2 and 2) the variance of the log-normal posterior distribution σ_p^2. The σ_x^2 can be approximately obtained from a trigamma function:

$$\sigma_x^2 = \psi(\frac{m-d}{2}), \tag{2.15}$$

where m is the number of samples and d is the number of coefficients. Whereas, the σ_w^2 is calculated as the median absolute deviation (mad) of logarithmic residuals between pairs of κ_R^i and κ_F^i:

$$\sigma_w^2 = \operatorname*{mad}_i(\log \kappa_R^i - \log \kappa_F^i). \tag{2.16}$$

Therefore, we can get the σ_p^2 by

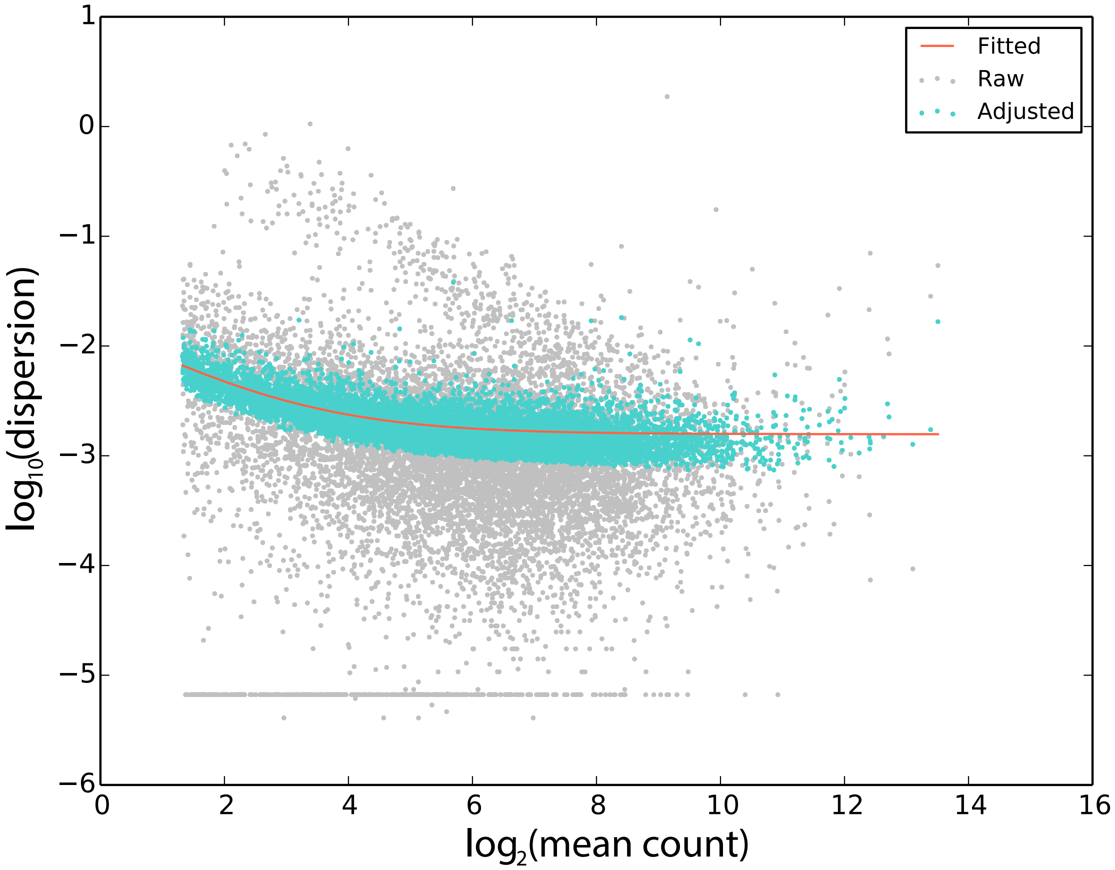

Figure 2.4: Scatter plot of the mean-dispersion relationship. Dispersions smaller than 1×10^{-6} are plotted on the bottom.

$$\sigma_p^2 = \sigma_w^2 - \sigma_x^2, \tag{2.17}$$

and obtain the final dispersion κ_S^i by maximizing the posterior:

$$\kappa_S^i = \arg\max_{\kappa_S^i} \left(\ell_{NB}(\kappa_S^i | y^i, \mu^i) - \frac{(\log \kappa_S^i - \log \kappa_F^i)^2}{2\sigma_p^2} \right). \tag{2.18}$$

In Figure 2.4, the green dots are the finalized dispersions that are shrunken from the raw dispersions towards the fitted dispersions.

2.3.6 Statistical Test

In a treatment/control setting, we evaluate whether a treatment has a significant differential effect on translation efficiency compared to the control. This is equivalent to determining whether the parameter $\beta_{\Delta,1}$ differs significantly from 0 and whether the relationship denoted by the dashed arrow in Figure 2.3 is needed or not. Statistically, the Null model is given by:

$$\begin{aligned} \log(\mu^i_{mRNA,C}) &= \beta^i_C + \beta^i_{mRNA} \\ \log(\mu^i_{RF,C}) &= \beta^i_C + \beta^i_{RF} \end{aligned} \tag{2.19}$$

And the alternative model is given by:

$$\begin{aligned} \log(\mu^i_{mRNA,C}) &= \beta^i_C + \beta^i_{mRNA} \\ \log(\mu^i_{RF,C}) &= \beta^i_C + \beta^i_{RF} + \beta^i_{\Delta,C} \end{aligned} \tag{2.20}$$

We fit the observed count data y^i and the explanatory matrix X into the Null and alternative GLM models in parallel. The difference between the deviances for the two GLM fitting (D_{H0} and D_{H1}) follows an approximate χ^2 distribution. The deviance of a GLM fitting is calculated as:

$$D = -2(\log(\ell(y^i|\beta^i_0)) - \log(\ell(y^i|\beta^i_s))), \tag{2.21}$$

where β^i_0 denotes the coefficients used in this GLM model, while β^i_s denotes the coefficients for the "saturated model" of the GLM system. Therefore, the deviance is -2 times the log-likelihood ratio of the used model compared to the saturated model.

The p-values generated from the χ^2 test are further corrected by multiple testing correction methods, such as Bonferroni correction [92] and Benjamini-Hochberg procedure [93].

2.3.7 Estimating the Dispersion for RNA-Seq and RF Separately

Because RNA-Seq and ribosome footprinting are different sequencing protocols, the properties of the read counts from these two protocols can vary. Here we show an example where estimating κ separately may be needed. The example data are from a recent publication [94].

The empirical dispersion estimates for RNA-Seq and RF counts are calculated from the following equation [85, 87, 90, 69]:

$$\sigma^2 = \mu + \kappa\mu^2. \tag{2.22}$$

Figure 2.5 shows the mean-dispersion relationship. It demonstrates the deviation of the empirical dispersion of RNA-Seq and ribosome footprint data in this experimental setting. The deviation between these two data sets becomes small when read count increases.

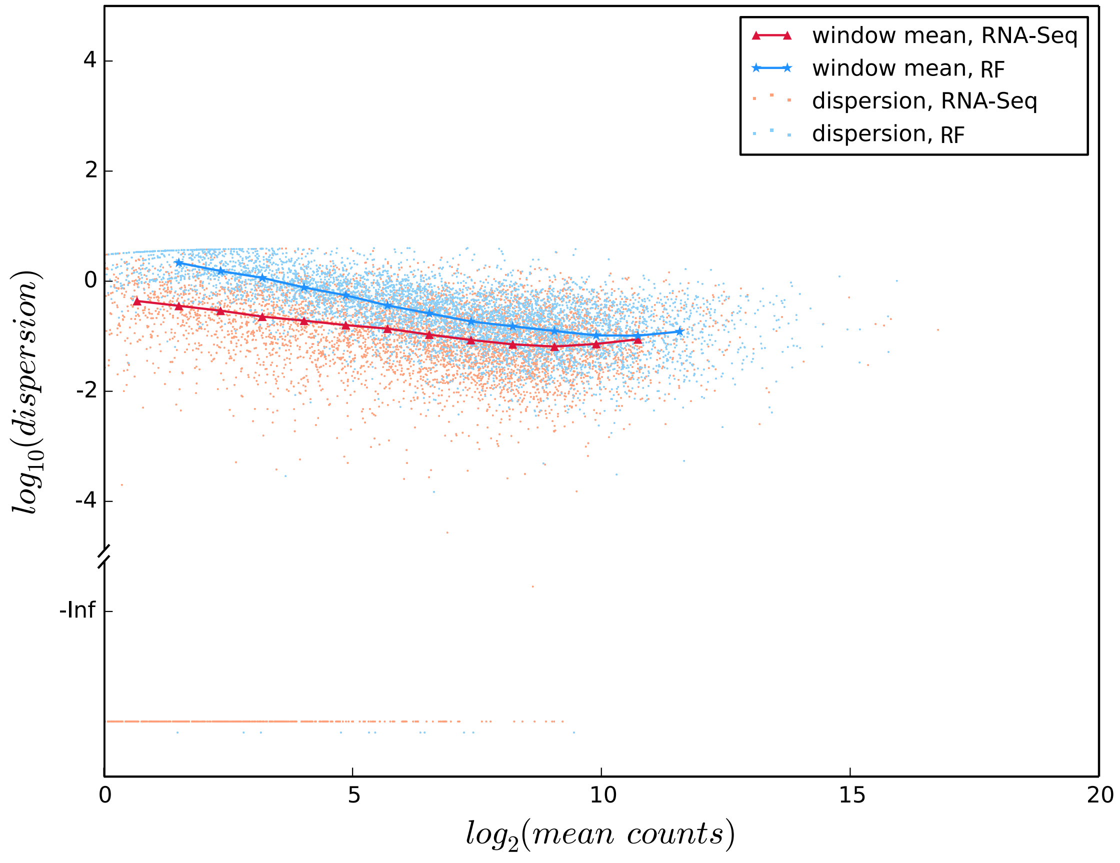

Figure 2.5: Scatter plot of empirical dispersions. The X-axis is split into several bins and the median of κ in each bin is highlighted and connected. The empirical κ smaller than zero are plotted at the bottom of the figure.

In the implementation, enabling *RiboDiff* to infer dispersion parameter κ for different data sources is to replace the scalar κ with a vector variable in the EM algorithm:

$$\beta^i = \arg\max_{\beta^i} \ell_{glm}(\beta^i|y^i, \overrightarrow{\kappa^i}) \quad \text{and} \quad \kappa^i = \arg\max_{\overrightarrow{\kappa^i}} \ell_{NB}(\overrightarrow{\kappa^i}|y^i, \mu^i). \tag{2.23}$$

Additionally, the mean-dispersion regression and the empirical Bayes shrinkage for the dispersions are also performed for RNA-Seq and RF separately.

2.4 Results from Simulated Data

2.4.1 Data Simulation

To test the performance of *RiboDiff*, we simulated the RF and RNA-Seq read count for 2,000 genes with 500 genes showing down regulated translation efficiency (TE) and 500 genes showing up regulated TE. There are three replicates for each of the two conditions (i.e., treatment and control) for RNA-Seq and RF. Therefore, count matrix dimensions are 2,000 × 12.

We first generated the mean counts for two treatments of both RF and RNA-Seq across all 2,000 genes assuming their mean counts are randomly drawn from a negative binomial distribution with parameter n and p, where $n = 1/\kappa$ and $p = n/(n + \mu)$. Then, for each mean count μ^i, we generated three count values as three replicates from a negative binomial distribution with parameter μ^i and κ^i, where κ^i is calculated as $\kappa^i = f(\mu^i) = \lambda_1/\mu^i + \lambda_0$.

To simulate the genes with TE changes in two treatments, we multiply the fold difference to the mean count of the target genes, assuming the fold changes follow a gamma distribution that is observed from real data (GEO accession: GSE56887). The gamma distribution has a shape parameter α and a scale parameter s, and its mean $\mu_G = \alpha \cdot s$. In the following simulation, we fix s and only change α to obtain different means for the two treatments and simulate genes having different fold changes using these two means. The fold increase F_I is obtained by

$$F_I = X_G(\alpha, s) + 1, \tag{2.24}$$

where X_G is a random vector containing 500 elements generated from a gamma density function. And the fold decrease F_D is obtained by

$$F_D = \frac{1}{F_I}. \tag{2.25}$$

Here, we simulated five groups of count data. In each group, 1,000 out of 2,000 genes showing TE changes:

- mean count has a fold change only for RF count, with $\alpha = 0.8$;
- mean count has a fold change only for mRNA count, with $\alpha = 0.6$;
- mean count has a fold change only for RF count, with $\alpha = 1.5$;
- mean count has a fold change only for mRNA count, with $\alpha = 1.5$;
- mean count has a fold change for RF with $\alpha = 0.8$ AND for mRNA with $\alpha = 0.6$, referred as "combined" in Figure 2.6.

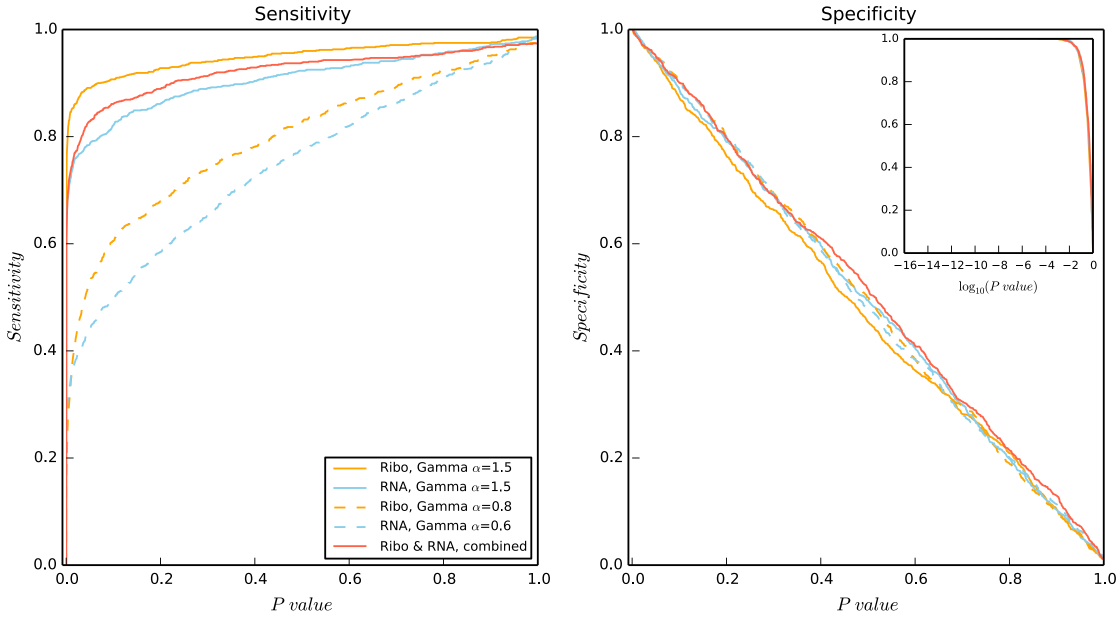

Figure 2.6: Sensitivity and specificity of *RiboDiff* on simulated data.

Note that in the last group, if the gene has fold increase in RF, it must have a fold decrease in RNA-Seq. By doing this, the effect at the mRNA level is added to the TE change outcome instead of offsetting the effect caused by RF. Other simulation parameters are as follow: for all RF and RNA-Seq, $n = 1$, $\lambda_1 = 0.1$, $\lambda_0 = 0.0001$, $s = 0.5$. The parameter p controls the scale of the count. We use 0.008 for RF and 0.0002 for mRNA. We run *RiboDiff* with the five dataset to estimate its sensitivity and specificity (Figure 2.6).

2.4.2 Performance under Different Numbers of Replicates

To evaluate how the number of replicates influences the dispersion estimation, RF and RNA-Seq counts for $5,000$ genes with two to ten replicates for each condition were simulated using the same way as described above. For instance, two replicates for condition A and two replicates for condition B in RF, and the same number of replicates for condition A and B in RNA-Seq. In total, we have 9 data sets. Each of them has a certain number of replicates ranging from two to ten. Next, we run *RiboDiff* on these 9 data sets.

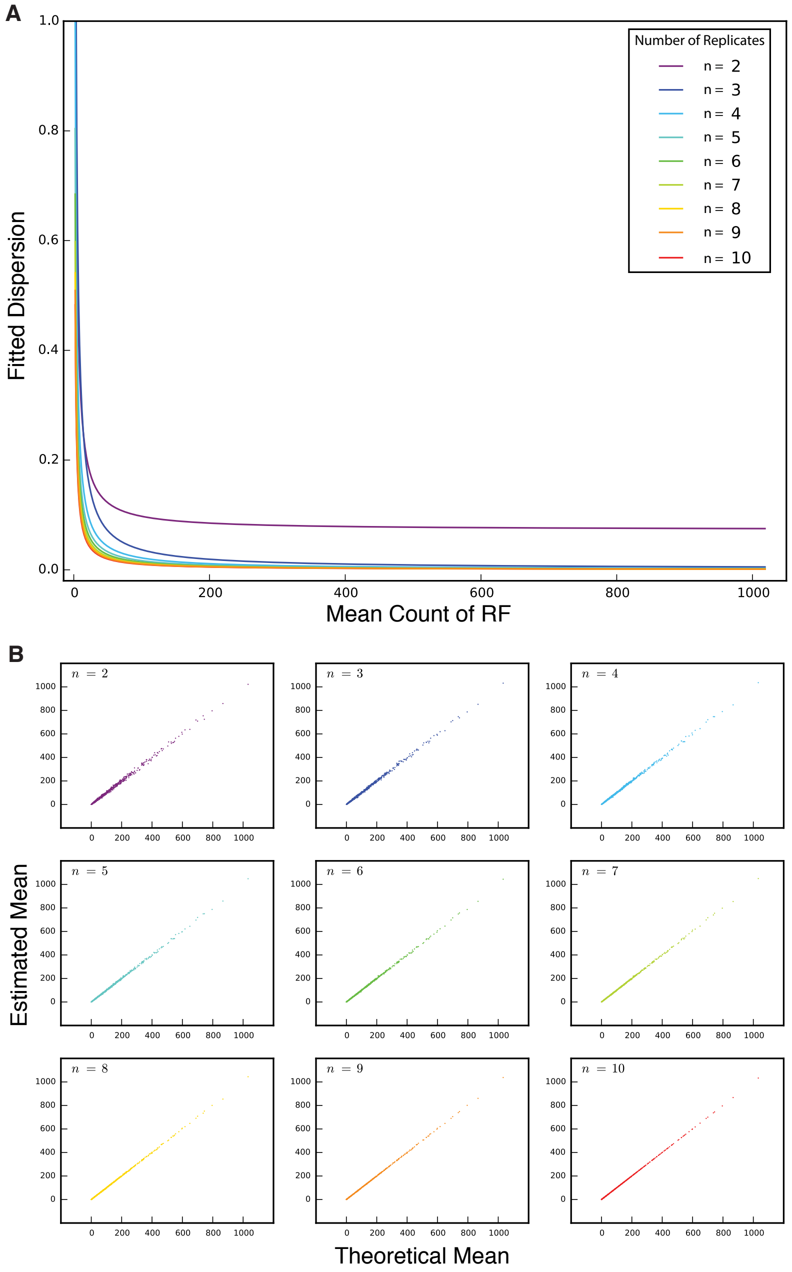

Figure 2.7: Evaluation of *RiboDiff* performance by using different number of replicates. A, Mean-dispersion relationship. B, Comparison between the theoretical and estimated mean.

RiboDiff firstly estimates the raw dispersion κ^i for each gene based on their RF and RNA-Seq counts. Then, a mean-dispersion relationship $\kappa_F = f(\mu) = \lambda_1/\mu + \lambda_0$ is obtained by regressing the raw dispersion κ^i given the mean count μ^i using GLM to learn λ_1 and λ_0. Figure 2.7A shows the mean-dispersion relationship function for different number of replicates. From this plot we see that the estimated mean-dispersion relationships, using three to ten replicates, are rather similar to each other, whereas the result using only two replicates deviates from the rest. This indicates that the raw dispersion κ^i estimated using two replicates is less reliable. We observed that the dispersion estimates of high read count genes are larger if only two replicates are used, which can decreases true positive rate.

We use the same simulated data set to show how the number of replicates affects the latent quantity β. For each gene, there are multiple β's that represent different latent quantities, and these βs are summed up to obtain the estimated counts of RNA-Seq or RF. Hence, we compare the estimated RF count (μ^i_{RF}) of every gene i against their mean counts (theoretical means) that are used to generate the negative binomial counts in the data simulation. In Figure 2.7B, each subplot is the comparison of estimated counts (Y axis) from n replicates against the theoretical means (X axis). As we can see, the theoretical means and the estimated means correlate well in all 9 experiments (all $r > 0.99$).

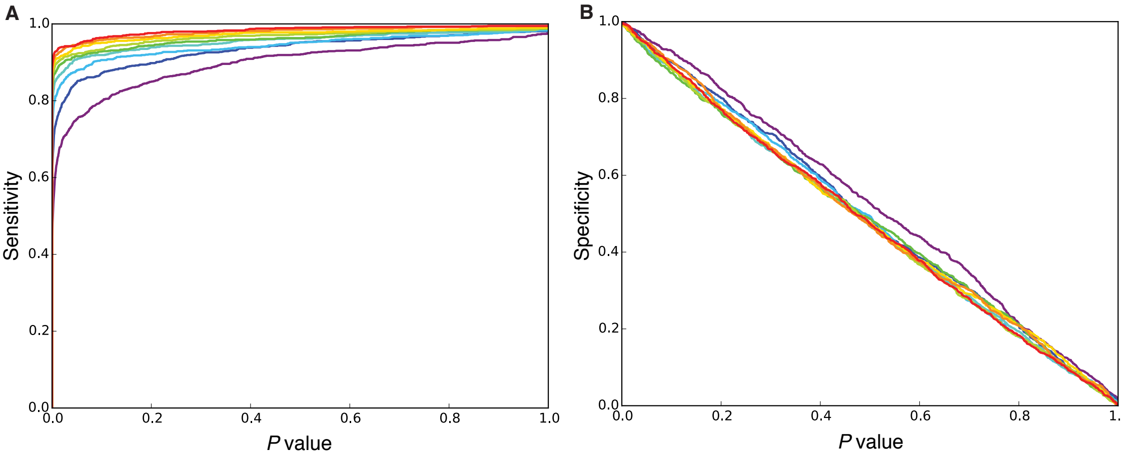

Figure 2.8: The sensitivity and specificity of *RiboDiff* by using different number of replicates. This figure shares the same legend with figure 2.7A.

Figure 2.8 shows how sensitivity and specificity depend on a chosen p-value threshold. For the sensitivity, the area under curves for 2 to 10 replicates increases when the number of replicates increase, whereas the specificities from the same data set do not have large difference among them. This illustrates that the test is well-calibrated and that one can recommend using three replicates to achieve a close-to-best sensitivity.

2.4.3 Comparison to other methods

We simulated data with different dispersions applied to mRNA and RF count to illustrate the performance of *RiboDiff* and compare it with Z-score method and *Babel*. The same simulation strategy was used as we described before with modifications. Briefly, 1,000 out of 2,000 genes were chosen to have ΔTE fold change by altering their mean counts of mRNA and RF. The following parameters were used to generate the mRNA count:

$$n = 1, p = 0.5 \times 10^{-4}, \lambda_1 = 0.1, \lambda_0 = 0.1 \times 10^{-3}, \alpha = 0.8, s = 0.5$$

And for RF count, the parameters were as follow:

$$n = 1, p = 0.1 \times 10^{-2}, \lambda_1 = 10.0, \lambda_0 = 0.01, \alpha = 0.8, s = 0.5$$

For the data, we run *RiboDiff* using the model where dispersions are estimated for the simulated RNA-Seq and RF counts separately. We also run *Babel* using its default parameters. The Z-score for every gene is calculated as described before. The receiver operating characteristic (ROC) curve (Figure 2.9) indicates superior quantitative performance of *RiboDiff* compared to *Babel* and Z-score method.

As *RiboDiff* is able to estimate the dispersion for different sequencing protocols either jointly or separately, the next is to study how different the results of the two approaches are and which one outperforms the other. To illustrate this question, we simulated two other data sets as following: the mRNA dispersion of every gene used above is multiplied by factors of 10 and 100, respectively, and we re-generated the mRNA counts using these different dispersions. Then the three data sets of mRNA counts with gradient differences in the dispersions are paired with the same simulated RF counts from above. In addition to running *RiboDiff* with the two dispersion estimation settings on the three data sets, we also included *DESeq2* [90] to compared the performance to *RiboDiff*. In general, the typical purpose of using *DE-*

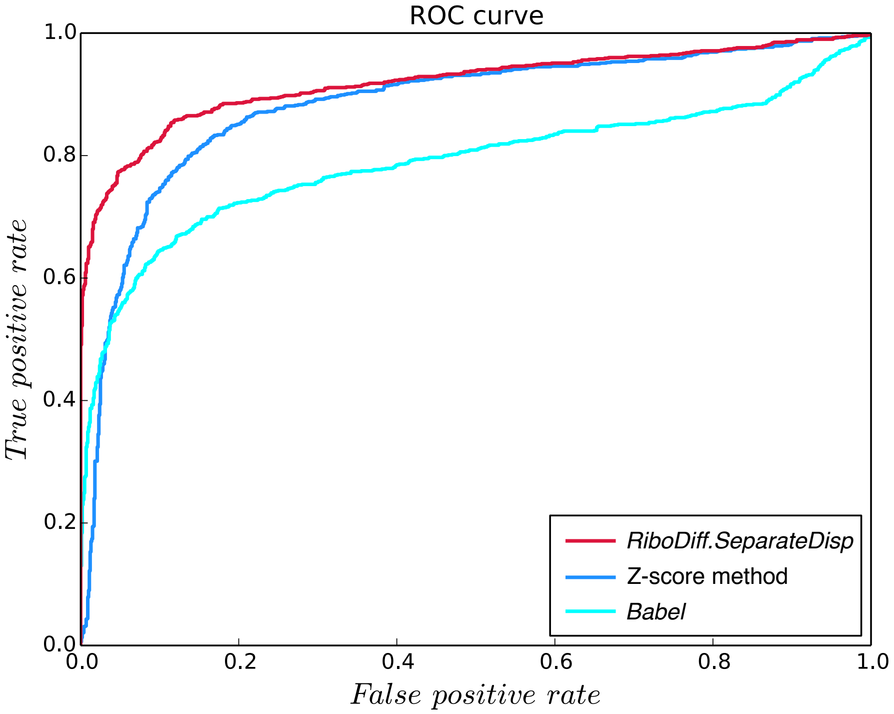

Figure 2.9: The receiver operating characteristic (ROC) curve of *RiboDiff*, *Babel* and Z-score method by using simulated data.

Seq2 is to identify differentially expressed genes in case-control study. Here, we use a specific design formula for *DESeq2*: condition + protocol + condition:protocol. The interaction term between sequencing protocol and experimental condition represents the possible condition differences controlling for protocol type. In Figure 2.10, from the top to the bottom, the three dispersion plots on the left side show the three simulated data sets where mRNA dispersions are approaching to merge with RF dispersions. The ROC curves on the right side are the corresponding performances of *RiboDiff* with joint and separate dispersion estimates and *DESeq2*. Although *RiboDiff* with joint dispersion estimate performs similar to DESeq2, estimating dispersion separately yields better results under the condition of different dispersions of the two protocols.

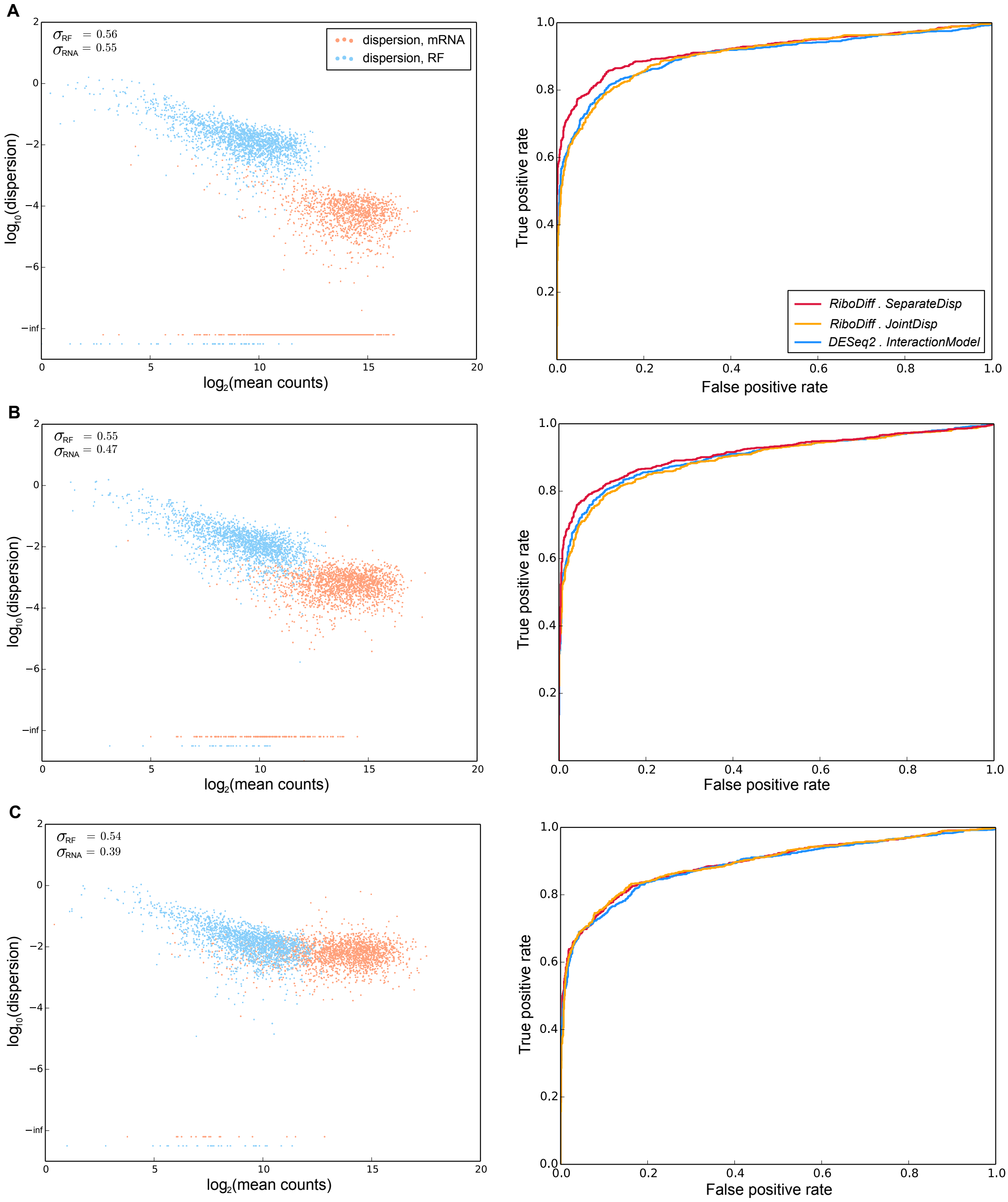

Figure 2.10: Comparison of ROC curves of *RiboDiff* and *DESeq2* using simulated data. (A-C) The left panel are the dispersions of mRNA and RF; the right panel are the corresponding ROC curves. From the top to the bottom, the differences of dispersion are large, moderate and small, respectively.

2.5 Results from Real Biological Data

2.5.1 Data Source

The real biological data we used here is from a recent study where Wolfe *et al.* discovered the translational control of oncoproteins is represented by the stable mRNA structure called G-quadruplex in the 5' untranslated region and the translation initiation factor eIF4A RNA helicase [64]. In this study, there are two biological replicates for both DMSO (control) treated and drug (Silvestrol, 25 nM) treated experiments at ribosome footprint level, and three replicates for the two experimental conditions at RNA-Seq level. The data is from NCBI Gene Expression Omnibus database with series number GSE56887.

2.5.2 Methods Comparison

The sequencing data were processed in a similar way as before [64], which includes trimming the adapter tail in the reads, aligning the reads, filtering the ribosomal RNA contamination, and counting the reads for genes, *etc.*

Here, we compared the results from *RiboDiff*, Z-score method and *Babel*. We ran *RiboDiff* and *Babel* using their default parameters. The translation efficiency change (ΔTE) was calculated as described in Equation 2.2. Then a Z-score for each gene was obtained by using Equation 2.3. The genes with $| z^i | \geq 1.5$ were selected as significant. Figure 2.11 shows the histogram of (ΔTE). The proportion of significant genes identified by *RiboDiff*, Z-score method and *Babel* in each bar are labeled in different colors. As shown in this figure, Z-score method classifies genes with most extreme ΔTE by using the fixed cutoff, whereas *RiboDiff* estimates the significance in a parametric manner, where gene with large ΔTE but low read count are deemed likely to be false positive, because their variances is large compared to other highly translated genes.

Figure 2.12 A and B shows that the overlap of significant genes detected by *RiboDiff* and Z-score based method are limited in both TE down and up regulated gene sets. Further analysis indicates most of the significant genes detected by the Z-score method having their mean RF counts smaller than 100 with only a few exceptional cases. In contrast, the significant genes detected by *RiboDiff* scatter over a wide range of mean RF count (Figure 2.12C and D). It is rational that

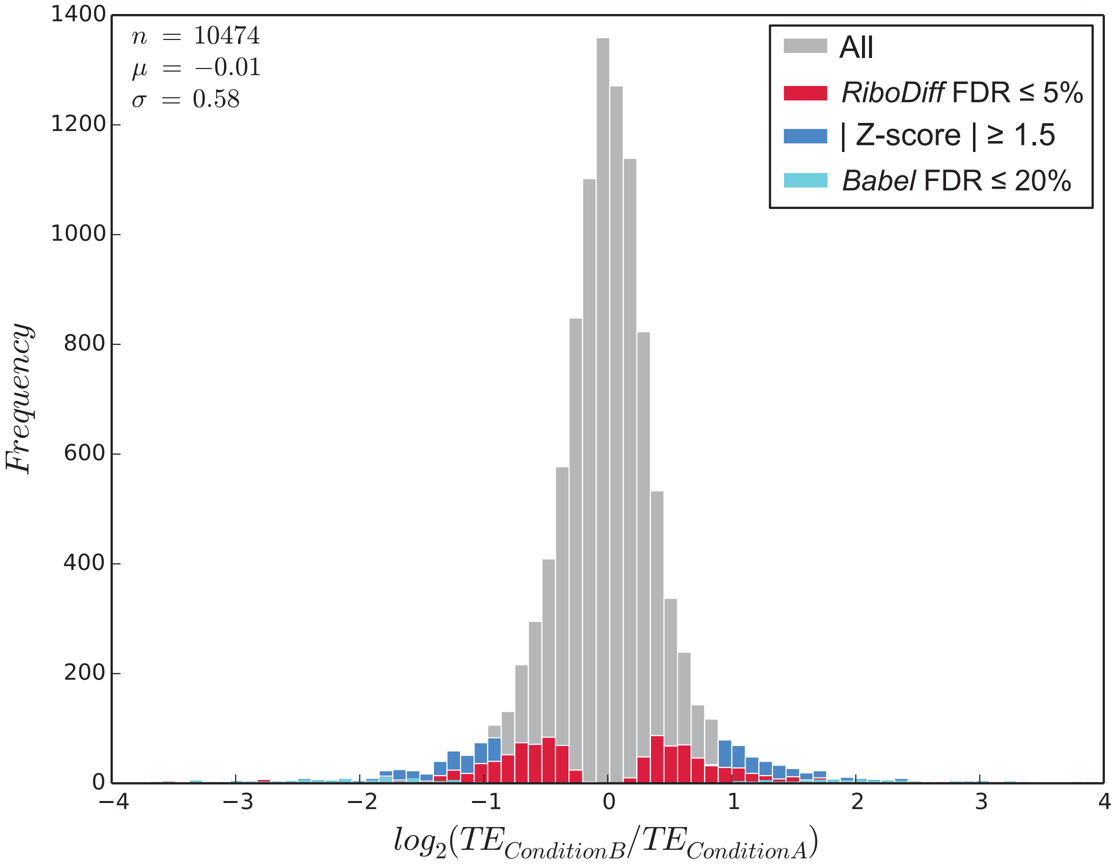

Figure 2.11: Comparison of *RiboDiff*, Z-score method and *Babel* using real biological data.

for highly translated genes, it is more confident to identify significant TE change between two treatments due to enough supported read counts. This is the reason that *RiboDiff* can detect highly translated genes as significant ones even though their absolute value of Z-score are less than 1.5 ($\mid \Delta TE \mid$ below the dashed lines in Figure 2.12 C and D). This comparison indicates *RiboDiff* identifies more sensible hits and is not biased towards genes with low mean count that inherently have more uncertainty rather than statistically significant differences.

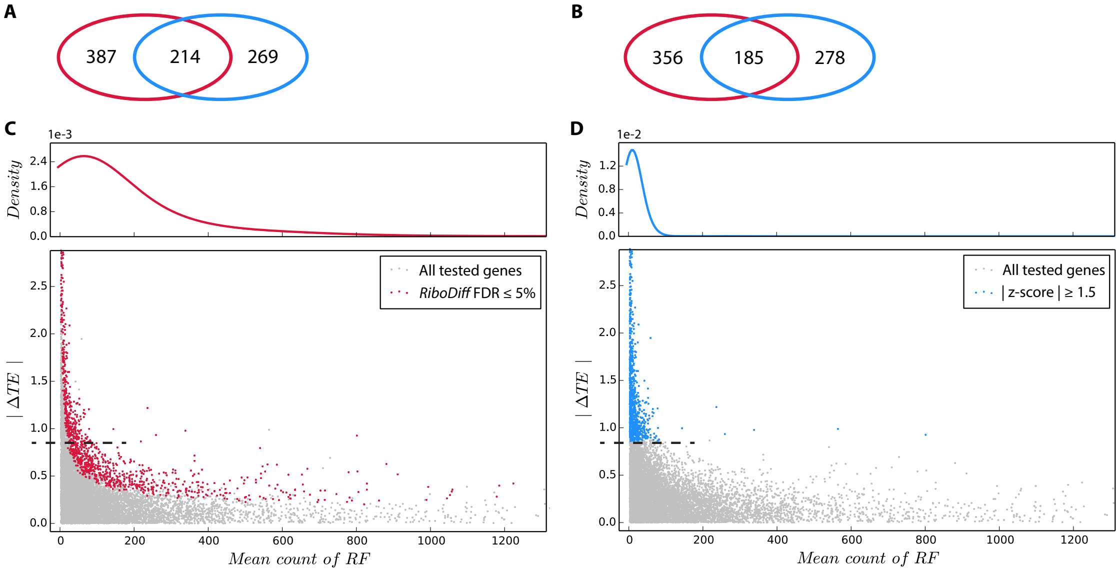

Figure 2.12: Comparison between *RiboDiff* and Z-score method on real biological data. (**A** and **B**) Venn diagrams showing the number of overlapping and self specific genes detected by *RiboDiff* and Z-score method. (**A**) TE down regulated genes. (**B**) TE up regulated genes. Red ellipse: results from *RiboDiff*; blue ellipse: results from Z-score based method. (**C** and **D**) Scatter plot of mean RF count against the $\mid \Delta TE \mid$. (**C**) Result of *RiboDiff*. Significant genes are labeled as red. (**D**) Result of Z-score based method. Significant genes are labeled as blue. The narrow panels above the scatter plots are the estimated density functions of significant genes on x-axes by using non-parametric kernel density estimation.

3 Computational Pipeline for Ribosome Footprinting Data

3.1 Introduction

As described before, the deep sequencing-based ribosome profiling method globally characterizes the mRNA fragments occupied by ribosome during protein synthesis. Similar to RNA-Seq data, the sequencing reads of these footprints provide not only the quantification of mRNA translation but also the ribosome density distribution at the nucleotide resolution. However, many properties of the sequencing results differ between these two protocols.

First, as the ribosome protected mRNA fragment is relative short, usually from 25 to 35 base pair long [62, 64], the ribosome footprint contained in the raw sequencing read at its 5' end is only a fraction of the entire sequence, and the remaining 3' end is the linker introduced in footprint sample preparation followed by a general sequencing adapter sequence. For the same reason, the ribosome profiling is always done in the single-end sequencing approach, whereas paired-end sequencing has become the standard procedure for RNA-Seq. Short read also has difficulties when align it against the reference genome, including the higher likelihood to align it to multiple genomic loci and the lower likelihood to be mapped if it crosses the mRNA splicing site.

Additionally, the ribosome RNA contamination in the footprint data can strongly dominate in the cDNA templates that are going to be sequenced [58]. This is due to the huge amount of ribosomal RNA compared to mRNA concentration in the cell. Although in the ribosome profiling protocol there are steps responsible for rRNA depletion, we still find 25-65% reads in the FASTQ sequencing output belongs to ribosome RNAs. These rRNA reads not only decrease the amount of informative sequences, but also can result in erroneous alignment because of the short read length. Therefore, computationally remove the rRNA reads is needed.

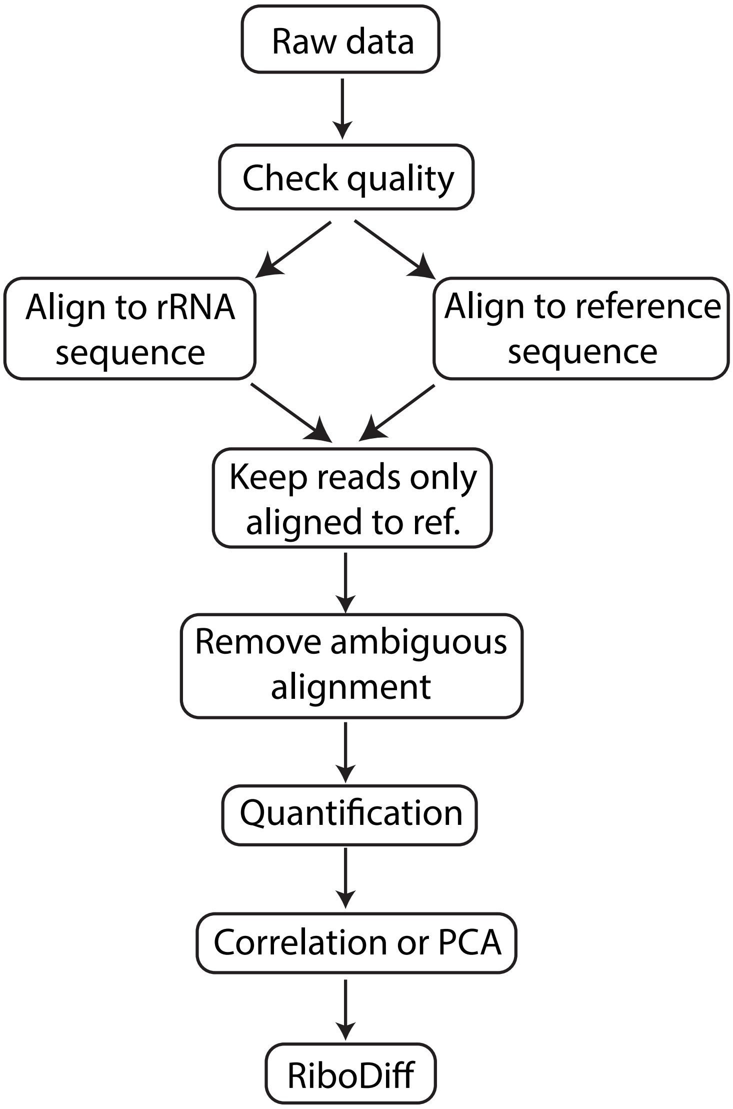
Raw data
Check quality
Align to rRNA sequence
Align to reference sequence
Keep reads only aligned to ref.
Remove ambiguous alignment
Quantification
Correlation or PCA
RiboDiff

Figure 3.1: The flowchart of computational pipeline for analyzing ribosome footprint data.

In this chapter, I will describe the computational pipeline for processing ribosome profiling data together with RNA-Seq result (Figure 3.1). The Shell and Python scripts are included in *RiboDiff* release version. Figure 3.2 shows an example of the ribosome footprint and RNA-Seq read coverage distribution at the locus of gene FKBP4 (Chromosome 12, 2904119-2914577, hg19 assembly) after processing the FASTQ sequencing data.

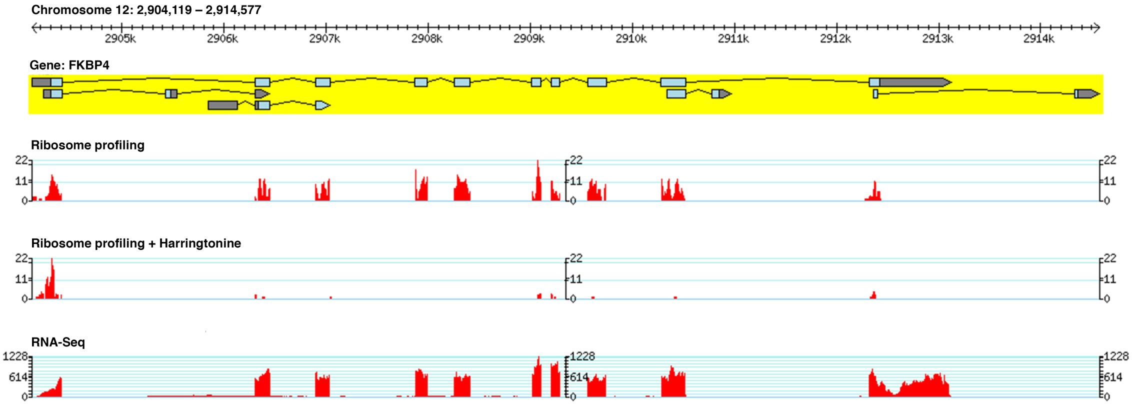

Figure 3.2: Ribosome footprint and RNA-Seq read coverage distribution of gene FKBP4. The gene structure is highlighted in yellow. The skyblue rectangles represent protein coding exons and grey ones represent untranslated regions. The arrows of the last exons indicate this gene is on the plus strand of the DNA. The lines links the exons represent the introns. FKBP4 contains five annotated transcripts. The first read coverage track is the ribosome footprints. Most of the reads locate in the protein coding exons. The second track shows the ribosome profiling treated with the translation elongation inhibitor harringtonine. As a consequence, almost no footprint is observed in the coding region. The third track shows the RNA-Seq read coverage, where we can see the 5' and 3' UTR are mapped as well as the coding region. The height of each peak indicates the depth of the coverage at the corresponding site.

3.2 Checking Sequencing Quality

The FASTQ file generated by a high throughput sequencing machine is a text format containing both nucleotide sequence and the corresponding quality for each base pair. The quality is encoded with a single ASCII character. Below is an example of FASTQ format from the Illumina Genome Analyzer:

```
@HISEQ:220:H9TFBADXX:1:1101:2636:2142 1:N:0:TTAGGC
TGGGAGGAGCAGCAGCAGGGTGGGACTGGGGCGTTCTACATCTCATTCAG
+
=@@B?D?;CDBFFDFDF?AC+A@G=?BBDGIIE55;A=7AAEDBDBBBB>
```

The first line starts with a “@” character and followed by an identifier for the read. The second line is the nucleotide sequence. The third line begins with a “+”

character separating the sequence and the quality information below. And the last line is the corresponding quality of each base pair in the second line. The quality is indicated by an ASCII character whose numerical representation is calculated as:

$$N = -10 \times \log_{10}(p) + 33, \tag{3.1}$$

where p is the estimated probability of the base call being wrong. The term $-10\ \log_{10}(p)$ is called the Phred quality score Q, which is used by traditional Sanger sequencing.

The sequencing quality largely depends on many factors, including the quality of the purified DNA template, the reagents added into the sequencing reaction and the manipulation by the experiment operators. Therefore, it is always worth checking whether the quality of the obtained sequence file is good enough to do the downstream analysis.

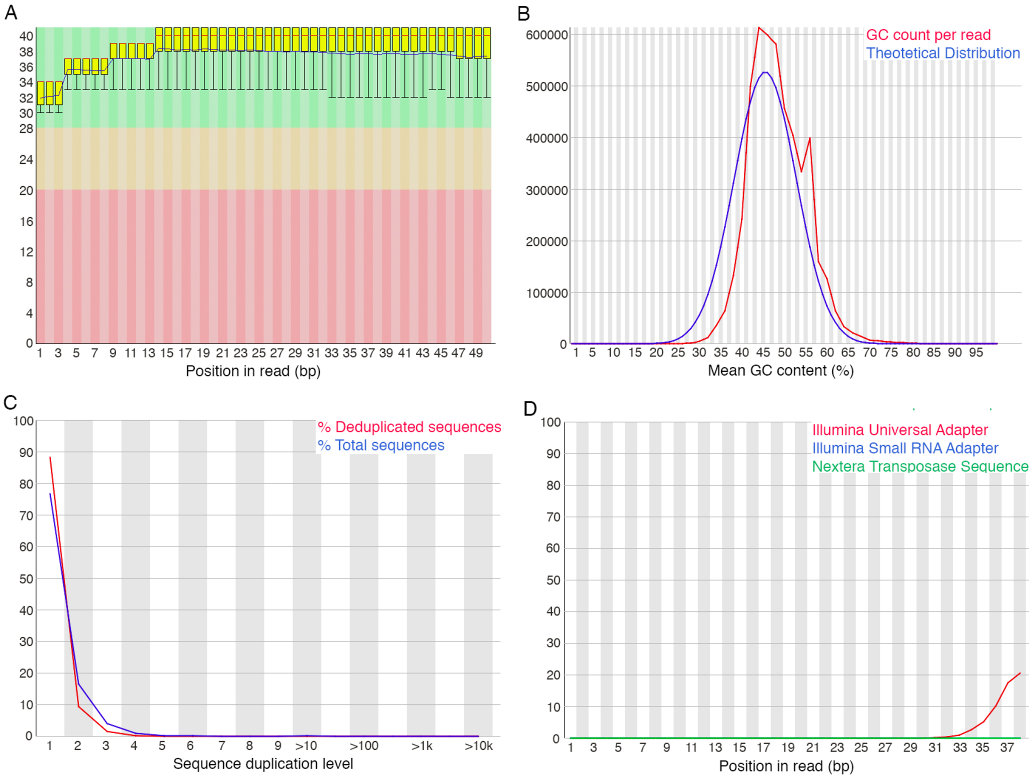

Figure 3.3: An example of selected outputs from FastQC. A, Mean quality scores across all bases from all reads. The Y-axis is the standard Sanger score. B, GC content distribution over all reads. C, Percent of remaining reads after remove duplicates. D, Adapter content.

In our pipeline, we use FastQC [95] to evaluate the qualities of both ribosome profiling and RNA-Seq data. It takes the FASTQ file as input and performs basic statistics such as total number of sequences, number of sequences with poor quality, sequence length and average GC content. In addition, it provides other more useful information, including per base sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, sequence duplication levels, overrepresented sequences, adapter content, and so on. Figure 3.3 shows an example of FastQC run on our unpublished data. The plots indicate the overall base pair quality along the reads is good ($Q > 30$, error rate $< 0.01\%$) and the read duplication level is low. However, a subset of reads with high GC content is indicated by the second peak at 56% on x-axis. The bottom right panel shows where the Illumina universal adapter locates in the reads. Specifically, as in this example we use the ribosome profiling reads, there is a 17-bp linker (CTGTAGGCACCATCAAT) before the universal adapter.

3.3 Remove Ribsomal RNA

Processing massively parallel high throughput sequencing data includes the quality checking, trimming the adapter from the reads, alignment, quantification, *etc.* There are several tools capable of dealing with RNA-Seq data [96, 97, 98, 99]. However, due to the unique features of ribosome profiling data, such as short single-end reads, large amount of rRNA composition, a specific pipeline is needed. Here, I will describe the computational workflow I built up for analyzing the footprint data.

Based on our observation, the rRNA-like reads varies from 25-65% for the single ribosome profiling library. The rRNA is transcribed from ribosomal DNA (rDNA) sequence on the genome. Because the reads are short, it is possible that these rRNA-like reads are (partially) aligned to other genomic loci such as protein coding genes. Besides, as the rDNAs are organized as both repeated operons (18S, 5.8S and 28S rDNA) and the widely scattered 5S rDNA on different chromosomes, any misassemblies of the highly similar rDNA copies lead to erroneous annotation for both coding and non-coding regions. Therefore, failure to remove all rRNA reads from the ribosome profiling data can result in misclassifications during the read counting step for gene translation quantification. Obviously, these issues cannot be solved by removing rDNA annotation from the gtf or gff file which is used to guide the quantification of the translation.

In our computational pipeline, we use an alignment-based strategy to identify rRNA-like reads from the ribosome profiling data and remove them subsequently. Specifically, all the rRNA sequences for a given species are retrieved from SILVA [100], NCBI [101] and Ensembl [102] databases. SILVA is a comprehensive resource that contains quality checked ribosomal sequences for thousands of species. Next, the rRNA sequences are linked one to another with $10,000$ 'N's separating the neighboring two rRNAs. The obtained sequence is further used as the species-specific ribosome reference to which the ribosome profiling reads can be aligned. All the successfully aligned reads are labeled as the ribosomal reads originated from the ribosomal RNA contamination. The rRNA sequences retrieved from the databases should be redundant, because more reference sequences with nucleotide variations improves the read mappability given the same parameters used for the alignment.

We use STAR [103] to do all the alignments for both ribosome profiling and RNA-Seq data in the pipeline. STAR is a splice-aware aligner which searches for a maximal mappable prefix in the reference sequence for each read. It has a significant speed advantage because it uses un-compressed suffix arrays to represent the reference. The suffix array of reference sequence can be created by executing the following command:

```
$ STAR −runMode genomeGenerate −genomeFastaFiles human_rRNA.
    fasta −genomeDir ./ref_rRNA/ −runThreadN 1
```

The argument '-genomeFastaFiles' is for specifying the rRNA reference file in FASTA format. '-genomeDir' is the path where the suffix array and other related files are generated.

Next, the ribosome profiling reads are aligned to the rRNA reference:

```
$ STAR −genomeDir ./ref_rRNA/ −readFilesIn footprint.fastq −
    runThreadN 1 −clip3pAdapterSeq CTGTAGGCAC −clip3pAdapterMMp
     0.1 −outFilterMultimapNmax 1000 −outFilterMismatchNmax 2 −
    alignIntronMax 9998 −seedSearchStartLmax 15 −genomeLoad
    NoSharedMemory
```

The argument '-clip3pAdapterSeq' specifies the linker sequence needs to be trimmed. '-readFilesIn' specifies the FASTQ file to be aligned. '-clip3pAdapterMMp' sets the mismatch rate when searching for the linker. '-outFilterMultimapNmax' sets the number of multiple mapping for each read. '-outFilterMismatchNmax' sets the mismatch nucleotides. '-alignIntronMax' specifies the maximum length of intron for splicing alignment. As we have $10,000$ 'N's separating each rDNA in the reference sequence, to avoid false splicing mapping cross different rDNA, we set 9998 for this argument. '-seedSearchStartLmax' sets the seed length that read is split into pieces no longer than it. Next, a Python script is called to parse the alignment BAM file to obtain the rRNA read IDs, which is serialized into bytes and stored as cPickle file.

The next step is to align the ribosome profiling reads against the reference genome. This is the most essential and fundamental step of data analysis, because all the downstream works are based on the alignment information. As we study the translation regulation in human T-cell acute lymphoblastic leukaemia (described in the next chapter), we downloaded the human genome (assemble version GRCh37) and its corresponding annotation GTF file from Ensembl [102]. To build the suffix array of reference genome, the following command is executed:

```
$ STAR −runMode genomeGenerate −genomeFastaFiles Homo_sapiens.
    fasta −genomeDir ./ref_genome/ −sjdbGTFfile Homo_sapiens.
    gtf −sjdbOverhang 49 −runThreadN 1
```

Here we provide the annotation GTF file such that STAR can use the splicing junction information to guide the aligning in addition to *de novo* splicing mapping. Next, to do the alignment, the pipeline calls STAR again:

```
$ STAR −genomeDir ./ref_genome/ −readFilesIn footprint.fastq −
    runThreadN 1 −clip3pAdapterSeq CTGTAGGCAC −clip3pAdapterMMp
     0.1 −outFilterMultimapNmax 1 −outFilterMismatchNmax 2 −
    alignIntronMax 500000 −seedSearchStartLmax 15 −genomeLoad
    NoSharedMemory
```

Note the argument '-outFilterMultimapNmax' is set to '1' such that only uniquely mapped reads are output in BAM files. This is a trade-off between reducing alignment uncertainty and obtaining large amount of reads to detect lowly translated genes. Subsequently, our computational pipeline removes the alignment entries from the BAM file if the read IDs are identified as rRNA contamination previously. The similar commands are executed for RNA-Seq data to identify and remove rRNA reads. Figure 3.4 shows an example of removing rRNA contamination from both ribosome profiling and RNA-Seq data. In this example, rRNA reads account for 24.54% of originally aligned ribosome profiling reads, whereas only 1.63% of the aligned RNA-Seq sequence are from ribosome contamination.

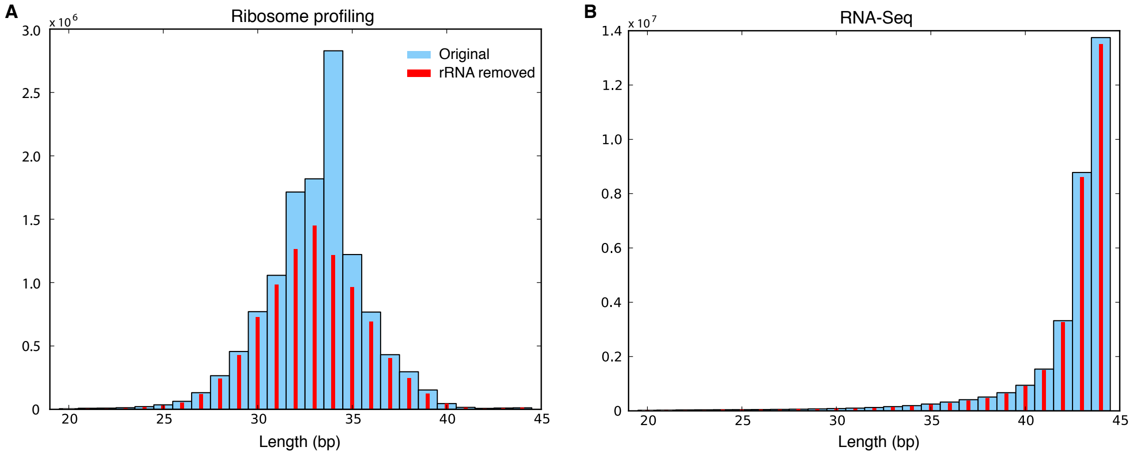

Figure 3.4: A histogram of number of reads before and after removing rRNA contamination. A, ribosome profiling data; B, RNA-Seq data. The GEO data accession number: GSE66810. The average read length of ribosome profiling is slightly longer than from the previously published protocol [58] due to using micrococcal nuclease to digest the ribosome-protected mRNA fragment.

3.4 Ambiguous Alignment

Similar to other read aligners for high throughput sequencing data, STAR trims the read from both 5' and 3' ends in order to find the maximum mappable prefixes on the suffix array of the whole genome. This is very useful for relative long and paired-end RNA-Seq reads. However, as the ribosome profiling read is short (usually 50 bp) and contains a specific linker sequence at the 3' end, additional trimming 3' end of the reads after cutting the linker can result in false aligning. This is because the quality of the bases in the middle part of the read is significantly higher than the two ends. Similarly, over trim the read from the 5' end can also leads to fallacious mapping. As shown in figure 3.3A, although the quality scores of the first few bases are lower than others, the error rate of one base with quality score Q equals to 30 is 0.1% (calculated from $Q = -10\log_{10}(p)$). Obviously, clipping the bases with error rate 0.1% from an extremely short read to increase the mappability must be cautious.

Here we define an alignment is ambiguous if

- for the 3' end, it satisfies:
  - the first 8 bases of trimmed sequence is NOT 'CTGTAGGC' (allowing one mismatch).
- for the 5' end, it satisfies either one of the following three criteria:
  - one or more bases are trimmed if the aligned length is 16-24 bp;
  - two or more bases are trimmed if the aligned length is 25-30 bp;
  - four or more bases are trimmed if the aligned length is 31-50 bp.

Once being identified, the ambiguous alignments are removed from the BAM files. Table 3.1 shows an example of filtering the ambiguous ribosome profiling reads of our unpublished data. It indicates the ambiguous alignment can account for a large amount of the reads if the aligned length (not include the trimmed sequences at the 5' and 3' ends) is short. Further filtering by ribosome footprint length can also be applied to refine the alignment, which has been implemented in our pipeline as well.

Table 3.1: Statistic of filtering the ambiguous alignments of ribosome profiling reads.

Length	rRNA reads removed	Ambiguous alignment removed	Ambiguous aligment (%)
16-20 bp	2,936,366	829,576	71.7%
21-25 bp	1,170,487	796,549	31.9%
26-30 bp	1,852,942	1,740,299	6.1%
30-50 bp	1,318,365	1,167,907	11.4%

3.5 Sample Quality Control

3.5.1 Quantification of Transcription and Translation

After align the ribosome profiling and RNA-Seq reads to the reference genome, we quantify the transcription and translation by counting the reads based on the BAM file given the annotated genomic features. The BAM file is a compressed, binary file containing all the alignment information for each read, such as the chromosome name, position, strand, mismatch/insertion/deletion of the alignment, the read ID, sequence, and the quality, and so on. It can be converted to a human readable text format—SAM file. More details about BAM and SAM files can be found at [104]. The annotation file is usually in GTF or GFF file formats. Take GTF format as an example: the GTF file is a text file contains nine columns separated by tabs. The nine columns are sequence name, data source, feature, start, end, score, strand, frame and attribute. The feature includes 'gene', 'transcript', 'exon', 'CDS', 'UTR' *etc.* More information about GTF and GFF can be found at [105].

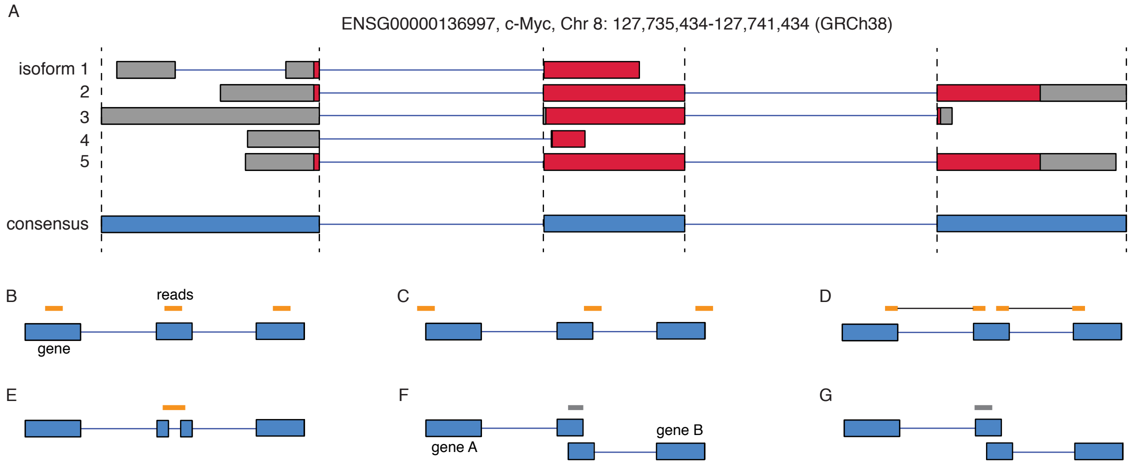
A
ENSG00000136997, c-Myc, Chr 8: 127,735,434-127,741,434 (GRCh38)
isoform 1
2
3
4
5
consensus
B
reads
gene
C
D
E
F
gene A
gene B
G

Figure 3.5: Schematic of counting the reads for gene. A, the transcript structure of gene c-Myc from Ensembl annotation version 75. The rectangles are exons. Red represents coding regions; grey represents 5' and 3' untranslated region; blue represents the consensus exonic region. B-G, cases where reads are counted (B-E) and not counted (F-G).

Figure 3.5 illustrates how we count the reads. In this example, the c-Myc gene has five annotated transcript isoforms. Each transcript isoform consists of different number of exons separated by introns which are spliced out during transcription. The exons have different boundaries as shown in the figure. To quantify the transcription and translation for a gene, we collapse the structure of all exons into one consensus sequence (Figure 3.5A) and count the RNA-Seq or ribosome profiling reads in the consensus exonic region. Specifically, reads that reside within and partially overlap with the exons are counted. Splice junction reads and reads that bridge exons are also counted. However, in the case of one gene overlapping with another, reads entirely mapped to both genes or entirely mapped to one gene but partially to the other are not considered as informative reads, hence are excluded from counting (Figure 3.5B-G). The counting script is originally written in Python by my colleague, Andre Kahles, with modifications.

3.5.2 Sample Correlation

Before performing the downstream analysis, it is always useful to assess the overall similarity between samples. This assessment helps to understand the sample correlations. Generally, the sample correlations within the same experimental condition are higher than that cross conditions. Any aberrant correlations caused by certain sample outliers, which were produced inappropriately during sample preparation or sequencing steps, should be removed for further computational analysis.

Pearson and Spearman correlations are the two frequently used metrics. For Pearson correlation, it is assumed that the relationship between a pair of data set $(X,\ Y)$ is linear. Pearson's correlation coefficient r is the covariance of X and Y divided by the product of their standard deviations σ_X and σ_Y:

$$\begin{aligned} r = \frac{cov(X,Y)}{\sigma_X\,\sigma_Y} &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\,\sigma_Y} \\ &= \frac{\sum_{i=1}^{n}(x_i-\mu_X)(y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_X)^2}\,\sqrt{\sum_{i=1}^{n}(y_i-\mu_Y)^2}}, \end{aligned} \tag{3.2}$$

where E is the expectation. μ_X, μ_Y is the mean of X and Y. The correlation coefficient ranges from -1 to 1. A value equals to 1 indicates the two data sets positively correlate with each other perfectly, namely Y increases as X increases; A value equals to -1 indicates a perfect negative correlation where Y deceases as X increases. If the two data sets do not correlate at all, the value equals to 0.

To calculate Pearson correlation of ribosome profiling and RNA-Seq data, we use the quantitative measurement—read count—of all pairs of genes in two samples. However, a logarithmic transformation is applied to the count data because the Pearson correlation is a measure of the linear correlation. Figure 3.6 shows an example of Pearson correlation of a ribosome profiling experiment, where both the control and treatment contain three replicates. We can see that the third sample in control is less correlated with the other two samples in the same condition. Similarly, the second sample in treatment is also an outlier compared to the other two samples. Therefore, these two samples should be removed from the data set for analysis in the next step.

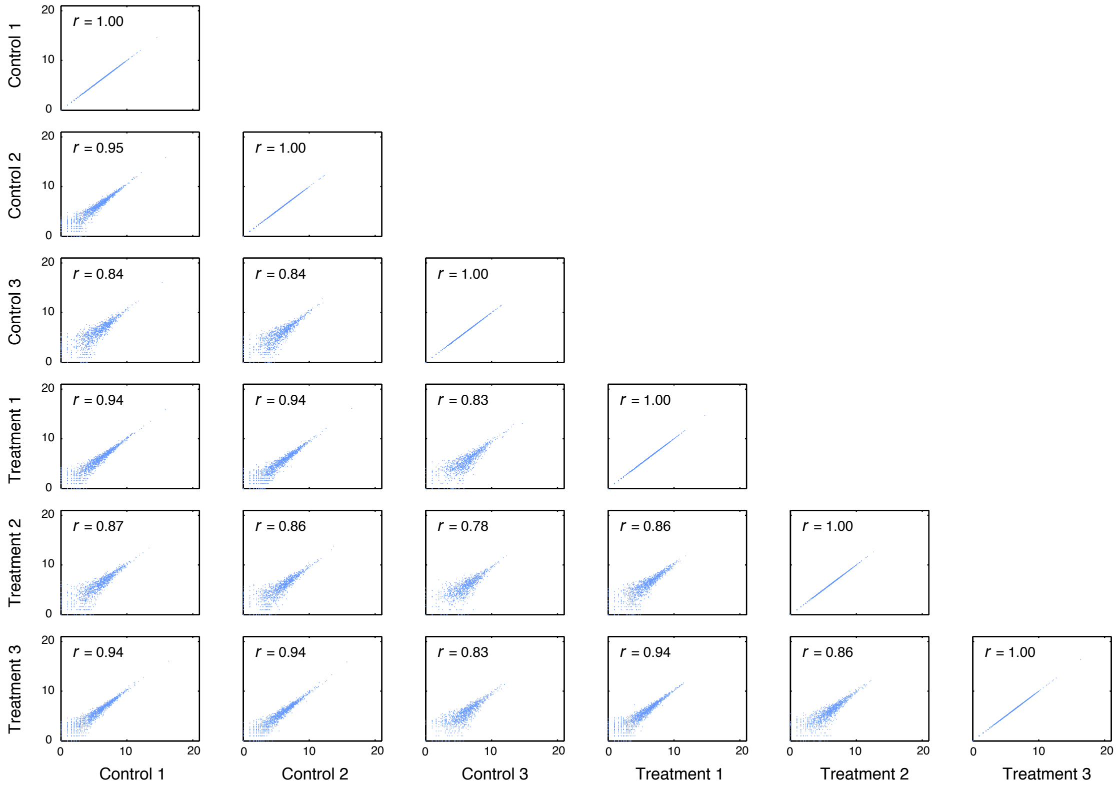

Figure 3.6: Scatter plot of pairs of samples from a ribosome profiling experiment. The data used here are from our collaboration with Wendel lab at MSKCC.

In contrast, Spearman correlation is a measure of the monotonic correlation between X and Y. To calculate the Spearman correlation, instead of taking the values of data sets, X and Y are firstly converted to ranks. Then the correlation ρ is obtained by comparing the ranks of each pair of x_i and y_i:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}. \tag{3.3}$$

As Spearman correlation uses rank information, it can only indicates the monotonicity of the data. In our computational pipeline, we calculate Pearson correlation because the goal of the data assessment is to know the overall similarity between samples, not only the change of the rank in two expression count data.

3.5.3 Principal Component Analysis

Another way to assess the sample similarity is to perform a principal component analysis (PCA). Assuming we have an $m \times n$ data matrix X of ribosome profiling counts, where m is the number of samples and n is the total number of protein-coding genes, the PCA identifies the patterns in the data in a way that the patterns express the data with minimal loss of information. The patterns are called principal components T. Because principal components are much less than the features (genes entries) in the data, PCA is widely used to reduce the dimensions of the datasets.

Briefly, the principal component T is a $k \times m$ matrix. The first component $t_{k=1,m}$ is given by

$$t_{k=1,m} = X \cdot w_{k=1,m} \ , \tag{3.4}$$

where w is called the loadings. It is obtained by

$$w_{k=1,m} = \arg\max_{w} \sum_{i} (x_i \cdot w_{k=1,m})^2 = \arg\max_{w} \frac{w^{\mathsf{T}} X^{\mathsf{T}} X w}{w^{\mathsf{T}} w} \tag{3.5}$$

The next jth component is then calculate by

$$w_{k=j,m} = \arg\max_{w} \frac{w^{\mathsf{T}} \hat{X_j}^{\mathsf{T}} \hat{X_j} w}{w^{\mathsf{T}} w} \quad \text{with} \quad \hat{X_j} = X - \sum_{s=1}^{k-1} X w_s w_s^{\mathsf{T}} \tag{3.6}$$

Therefore, each row in dataset X (the n samples) are projected to a new space with less dimensions while the variance of the data is explained by all the elements in principal components vector $t_1, t_2, \ldots, t_k$ in a descending order.

Practically, the PCA consists of the following steps:

- calculate the mean of each row in X and subtract the mean from the each element of the row;
- Calculate the covariance matrix of the rows in X;
- Calculate the eigenvectors and eigenvalues of the covariance matrix;
- Sort the eigenvectors by decreasing eigenvalues and Choose eigenvectors with the largest eigenvalues;
- Transform the samples onto the new subspace.

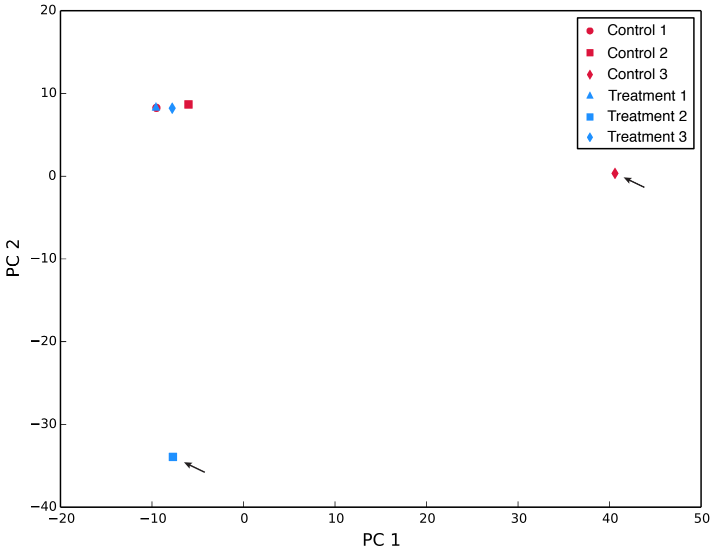

Figure 3.7: PCA of samples from a ribosome profiling experiment. The data are the same as in Figure 3.6. The arrows indicate the two outlier replicates.

Details about the steps can be found at [106, 107]. In out pipeline, we implemented the PCA using scikit-learn. To compare to the Pearson correlation result in the previous section, we computed the principal components using the same data and plotted the first two PCs (Figure 3.7). As shown here, the two outliers deviated far from other four samples are exactly the same as detected by the Pearson correlation.

3.6 Implementation and Software

The computational pipeline is implemented in Shell and Python. *RiboDiff* is implemented in Python. The libraries that *RiboDiff* depends on include Numpy (1.8.0 or higher), Scipy (0.13.3 or higher), Matplotlib (1.3.0 or higher) and Statsmodels (0.5.0 or higher). These requirements can either be installed individually or as a Python distribution that includes all the required packages. Please find more details at `http://www.scipy.org/install.html`

The supplemental information can be found at `http://bioweb.me/ribodiff`. Source code is publicly available at `https://github.com/ratschlab/RiboDiff` under GPL license.

4 EIF4A Promotes Oncogene Translation in Cancer

4.1 Background

In the central dogma, one of the essential step is messenger RNA translation, where the DNA information is decoded to produce the protein. Precise control of this step ensures many aspects in cellular processes, such as normal growth, differentiation, homeostasis and response to enviromental changes *etc.* [74, 75, 76]. Dysregulation of mRNA translation has been observed in many disease development such as cancer [80, 108]. Many tumour suppressors and oncogenes can affect the translation machinery, leading to aberrant translation of a subset of genes in tumor cells [109, 36]. One example is overexpression of the translation initiation factor eIF4E causes malignant transformation in cultured rodent cells [110]. Other studies have shown that eIF4E-binding protein (4E-BP) [111, 112], ribosomal protein S6 kinase (S6K) [113], eIF4GI [114, 115] and eIF3 [116, 117] are also implicated in cell transformation and tumorigenesis. Therefore, the discovery of molecular mechanism of translational change that provokes the cancer development holds promise to design anticancer therapies against the potential targets.

Ribosomal footprint profiling combined with deep mRNA sequencing technology enables precise measurements of changes in mRNA translation [56]. Immediate readouts are the number of ribosome footprints (RFs) at the translated region, which provides a surrogate indication of translational activity for a given gene. Moreover, each RF can be mapped to a specific location and indicate distribution along the transcript.

In this chapter, I will describe our discovery of a novel translational control mechanism represented by translation initiation factor eIF4A promoting the oncogene translation. We found that the RNA helicase activity of eIF4A enables it to unwind the structure called G-quadruplex in the 5' untranslated region in mRNA, which facilitates the ribosome to successfully locate the start codon and initiate the translation. We also show the anticancer effect of a natural compound—Silvestrol—can offset the over translated oncogenes in leukemia cell line by disrupting the eIF4A helicase activity. This work is a collaboration with Dr. Hans-Guido Wendel at Memorial Sloan Kettering Cancer Center. I performed most of the computational work by combining the RNA-Seq and ribosome footprint data to decipher the translational control through the interaction between eIF4A and RNA G-quadruplex.

4.2 Aberrant Translation Causes Leukaemia

4.2.1 EIF4A Accelerates Leukaemia Development

We first created the leukemia mouse model as following (Figure 4.1): 36 T-cell acute lymphoblastic leukaemia (T-ALL) samples were historically collected from paediatric patients at multiple organizations. The mutation analysis was performed according to the literatures [118, 119, 120]. Among these samples, we found PTEN mutations (14%) and deletions (11%), NOTCH1 mutations (56%) and an IL7R mutation (3%). This result agrees with the previous report that Notch signaling pathway is responsible for the development and progression of human malignancies, including leukemia [121]. These mutated NOTCH1 was then cloned into retroviral vector and transduced to hematopoietic progenitor cells (HPCs), which were collected from a pregnant mouse. The HPCs harboring the NOTCH1 mutations were injected to irradiated mice (C57BL/6J females between 6 and 10 weeks of age). After 91.5 days in average, the mice developed T-cell acute lymphoblastic leukaemia.

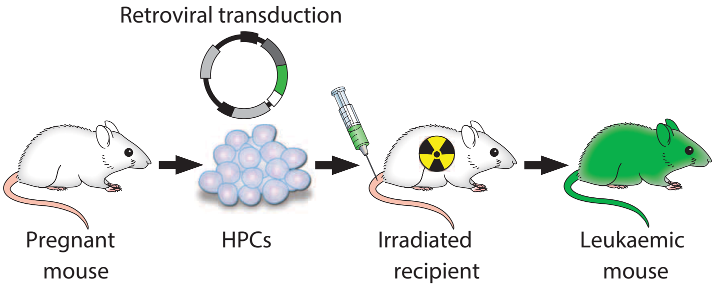

Figure 4.1: Diagram of the intracellular NOTCH1-driven murine T-ALL model. HPCs, hematopoietic progenitor cells.

The leukaemia disease onset was accelerated by the following experiments:

- knockdown Pten by shRNA (short-hairpin RNA) [122];
- expression of the mutant Il7r;
- expression of translation initiation factor eIF4E;
- expression of translation initiation factor eIF4A.

The data were analysed in Kaplan-Meier format using the log-rank (Mantel-Cox) test [123] for statistical significance. Table 4.1 shows the number of days before the leukaemia onset observed in the experimental mice. The leukaemia development after transplantation of HPCs transduced with NOTCH1 and empty vector, eIF4E, eIF4A1, IL7r, sh-Pten shown in Figure 4.2A. Down-regulating of eIF4E by expressing a constitutive 4EBP-encoding allele [124] or eIF4A by constructing the shRNA can both rapidly eliminate T-ALL from a mixture of murine cell population with and without the knockdown of eIF4E or eIF4A (Figure 4.2B).

Table 4.1: Number of days before the leukaemia onset in mice.

Gene	Days before T-ALL onset	Number of mice	p value
sh-Pten	47.1	n = 10	$p < 1 \times 10^{-4}$
IL7r	35.5	n = 4	$p < 1 \times 10^{-4}$
eIF4E	30.75	n = 4	$p < 1 \times 10^{-4}$
eIF4A	33.8	n = 5	$p < 1 \times 10^{-4}$

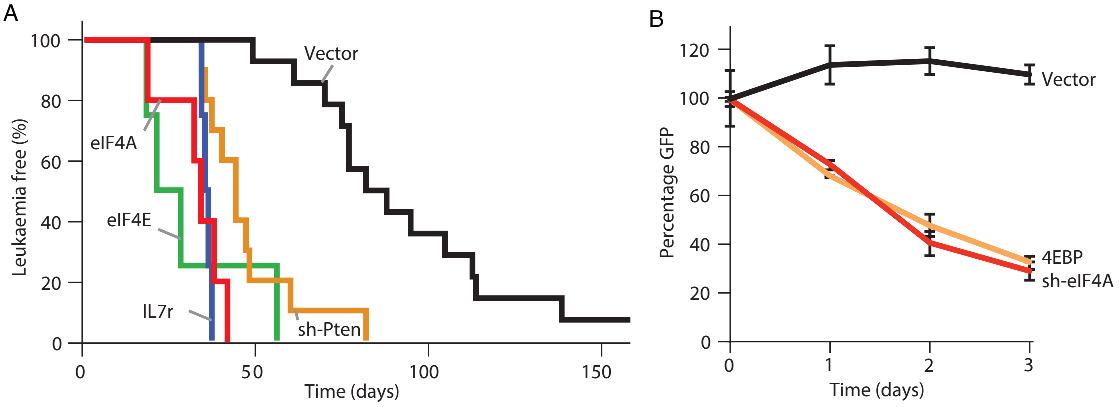

Figure 4.2: EIF4E and eIF4A accelerate the onset of T-ALL (A); This effect can be eliminated by the knowdown of eIF4E and eIF4A (B).

4.2.2 Anticancer Drugs Suppress Leukaemia

It has been reported that several chemical compounds have therapeutic effect against mouse lymphoma model [125, 126] by suppressing the translation initiation of target genes. Silvestrol, a natural compounds isolated from species of the *Aglaia* genus of the Meliaceae plant family, are novel inhibitors of translation initiation [125, 127]. In this study, we use Silvestrol and its synthetic analogue, (±)-CR-31-B (CR, Figure 4.3) to show their anticancer effect. A reporter assay confirms that both drugs preferentially block cap-dependent translation of *Renilla* luciferase compared to *Firefly* luciferase expressed from the hepatitis C virus (HCV) internal ribosome entry site (IRES) (Figure 4.4A).

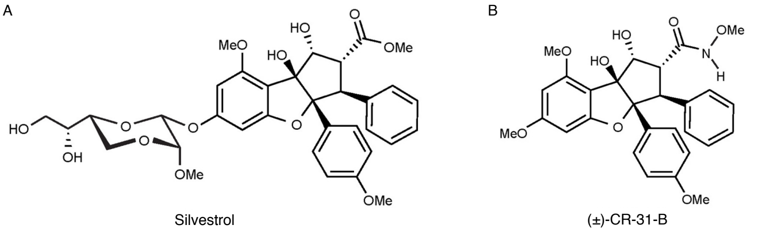

Figure 4.3: Chemical structure of silvestrol (A) and (±)-CR-31-B (B).

As shown in Figure 4.4B, C and D, Silvestrol has excellent single agent activity against T-ALL *in vitro* and *in vivo*. To generate half-maximum inhibitory concentration (IC50) curves, T-ALL cell lines and samples were treated with silvestrol for 48 hours. The IC50s for Silvestrol ranged from 10 nM in DND41 cells to 86 nM in MOLT-16 cells, and for the analogue CR the IC50s were in similar range. Notably, Silvestrol was equally active against PTEN wild type (KOPT-K1) and PTEN mutant (JURKAT, CEM) cells, and the least sensitive line (MOLT-16) carries a c-MYC translocation (Figure 4.4B). To demonstrate the anticancer effect of the drugs, we performed the following xenograft study. Briefly, $5,000,000$ KOPT-K1 cells in 30% matrigel were injected subcutaneously into C.B-17 SCID mice. When tumours were readily visible, the mice were injected with 0.5mg/kg Silvestrol, 0.2 mg/kg (±)-CR-31-B, or vehicle control on 7 consecutive days. Tumour size was measured daily by calliper. *In vivo*, both Silvestrol and CR were effective against xenografted T-ALL cells (Figure 4.4C and D). Treatment of KOPT-K1 tumor (1 cm^3) bearing NOD/SCID mice with systemic administration of Silvestrol (day 1-6) and CR (day 1-7) produced a significant delay in tumor growth (Silvestrol: $n = 7$, $p < 0.001$;

CR: $n = 8$, $p < 0.001$). Pathology on treated tumors showed massive apoptosis by TUNEL and loss of proliferation.

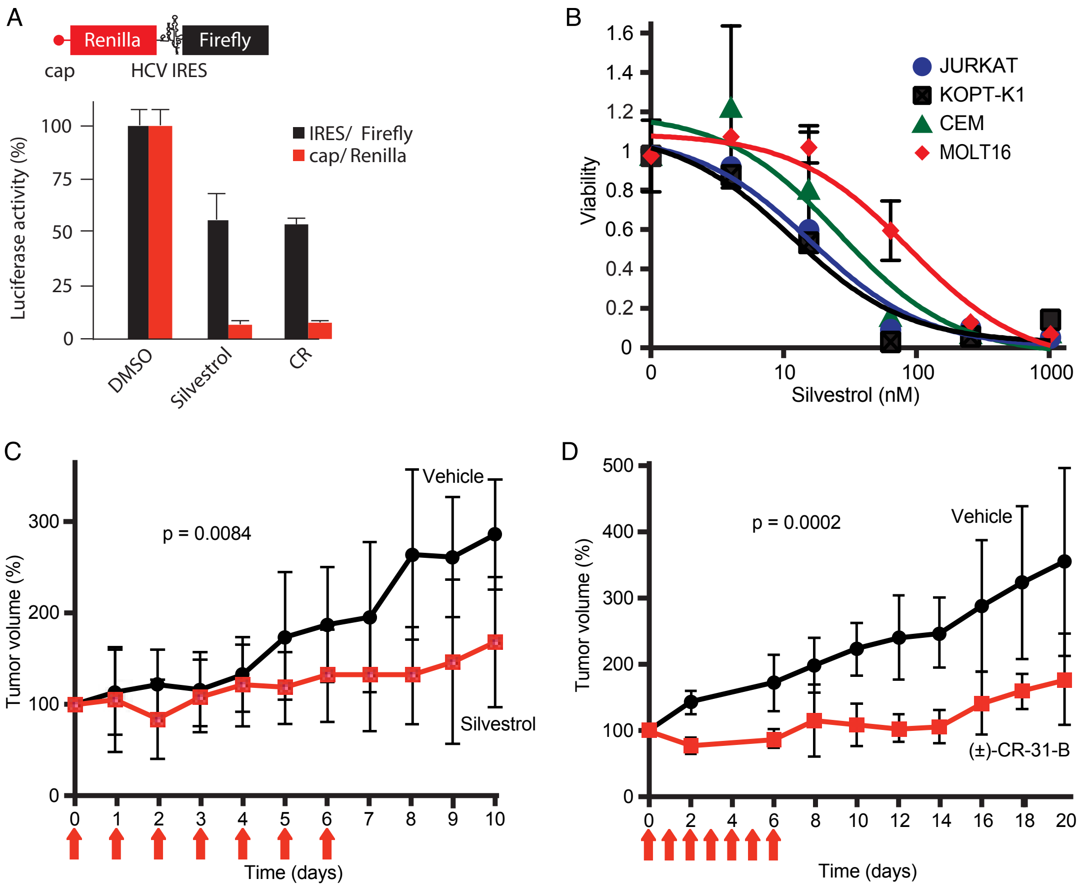

Figure 4.4: Silvestrol has single-agent activity against T-ALL. A, Reporter system with capped *Renilla* luciferase (red) and *Firefly* luciferase under the HCV IRES (black). Bottom, relative levels of *Renilla* luciferase (red) and *Firefly* (black) luciferase upon control (DMSO), Silvestrol or CR treatment. B, Viability of T-ALL cells treated with silvestrol. C and D, Tumour size of KOPT-K1 xenografts treated with Silvestrol and CR or vehicle control on days indicated by red arrows. *P* values were calculated using ANOVA.

Notably, we did not observe severe toxicity, death, or weight loss. CR treatment at therapeutic doses showed a reversible drop in white cell count with a nadir on day 19, and no other changes in blood counts or bone marrow cytology, or serum chemistry. In addition, we observed no changes in intestinal histology. Hence, Silvestrol and the CR analogue are highly effective T-ALL drugs and well tolerated *in vivo*.

Several studies have shown that Silvestrol is an inhibitor of cap-dependent translation through the RNA helicase eIF4A [125, 127, 128]. However, the precise mRNA features that necessitate the eIF4A helicase action are not known. In the next section, we use ribosome footprint profiling and RNA-Seq high throughput sequencing technologies to uncover the mechanism of translational control in T-ALL cell that is targetable by Silvestrol.

4.3 Ribosome Profiling and Computational Analysis

4.3.1 Sample Preparation and Deep Sequencing

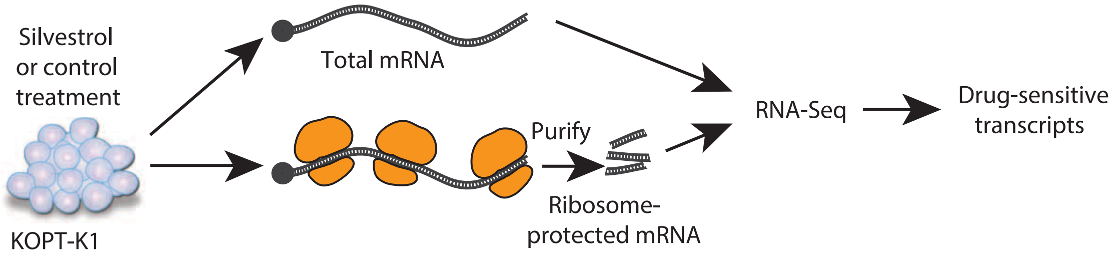

Figure 4.5: Schematic of the ribosome footprinting study.

To understand the effect of Silvestrol on translation in T-ALL, we use human T-ALL cell line KOPT-K1, which is originally from haematopoietic and lymphoid tissue and bears a mutation in Notch1 that contribute to T-ALL induction and maintenance [119], to perform the downstream experiments. KOPT-K1 cells were treated with Silvestrol or Dimethyl sulfoxide (DMSO, here used as the control) for 45 min. This early time point was chosen to capture effects on translation and minimize secondary transcriptional changes. Next, we treated the cells by cycloheximide for 10 min to inhibit the translation [59] and then harvested for total mRNA and ribosome footprint mRNA fragment isolation. Total mRNA was isolated using RNA isolation kit from Qiagen and subjected to RNA sequencing. Ribosome protected mRNA fragments were isolated following published protocol [58]. Briefly, cell lysates were subjected to ribosome footprinting by nuclease treatment. Footprint fragments were purified by one step sucrose cushion and gel extraction. Deep sequencing libraries were generated from these fragments. Both total mRNA and footprint fragment libraries were sequenced on the Illumina HiSeq 2000 platform. Figure 4.5 illustrates the experimental pipeline of sample preparation.

4.3.2 Alignment, Filtering and Quantification

The human genome sequence hg19 was downloaded from UCSC public database (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes). Ribosome footprint (RF) reads were aligned to reference genome hg19 using PALMapper [129]. PALMapper clips the linker sequence (5'-CTGTAGGCACCATCAAT-3') which is technically introduced during RF library construction, and trims the remaining sequence from the 3 end while aligns the reads to reference sequence. Briefly, we set the parameters for PALMapper as following: maximum number of mismatches: 2; maximum number of gaps: 0; minimum aligning length: 15; maximum intron length (splice alignment): 10000; minimum length of a splicing read aligned to either side of the intron boundary: 10. We only use the uniquely aligned reads for further analysis.

To remove ribosome RNA contamination, the footprint reads were also aligned to a ribosome sequence database using PALMapper with the same parameters except allowing splice alignment. We retrieved the human ribosome sequences from BioMart Ensembl [130] and SILVA [100] databases and merged the results into a single FASTA file which was used as reference sequence to align against. The rRNA-aligned reads were filtered out from hg19-aligned reads.

After we removed the rRNA contamination, we still observed a portion of reads that were dominant by linker sequence and Illumina P7 adapter. These reads can also be trimmed during mapping and cause false alignment. Therefore, we searched the nucleotide sequence for possible RF linker from the trimming site (± 2 bp) up to 8 bp to its 3' direction allowing 1 nt mismatch. We removed the read if there was no such linker sequence existed. Figure 4.6 highlights the filtering step for footprint data. Finally, we filtered out reads ≤ 24 bp and ≥ 36 bp, and the remaining reads with aligned length from 25 to 35 bp were used to analyze the translational effect of Silvestrol (Figure 4.7).

Total mRNA sequencing reads were aligned to the hg19 reference using STAR [103]. We performed the splice alignment and only use the uniquely aligned reads with maximum three mismatches. rRNA contaminating reads were also filtered out using the same strategy described before.

For each gene, we quantified their mRNA and RF abundance by counting the aligned reads that were mapped within exonic regions. The genome annotation

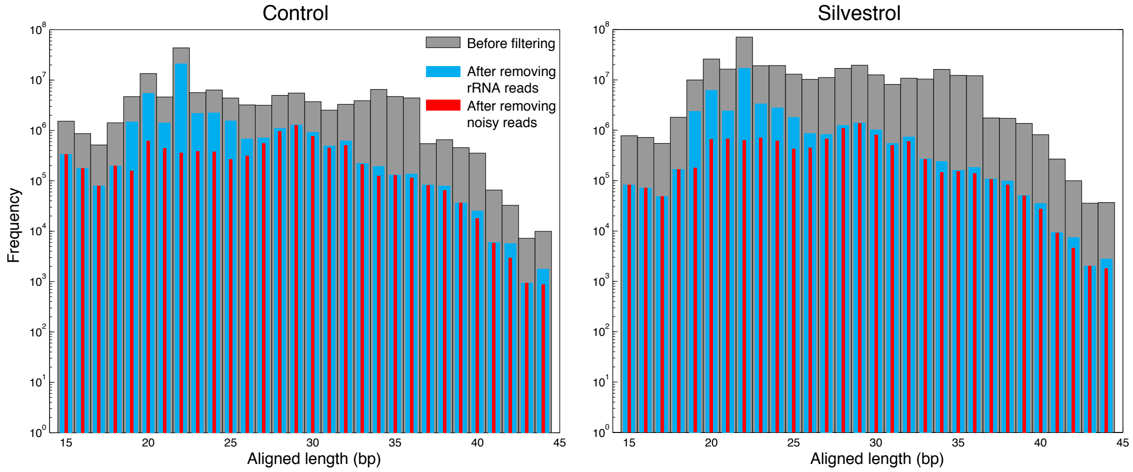

Figure 4.6: Read counts by length of mapped sequence before and after filtering.

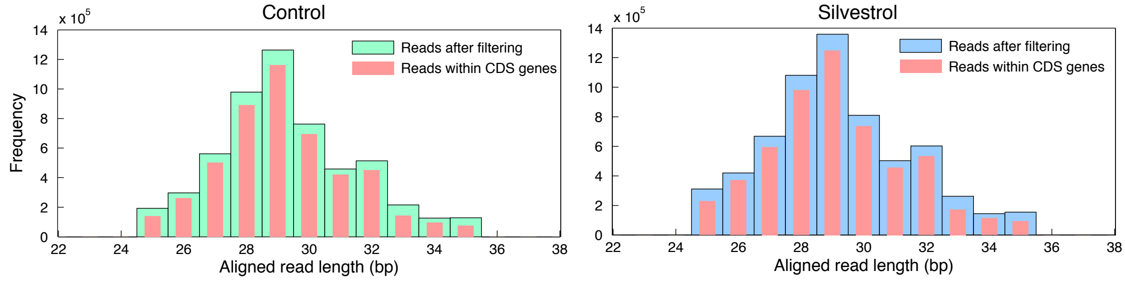

Figure 4.7: Read length histograms of ribosome footprint data after quality control filtering.

downloaded from GENCODE (http://www.gencodegenes.org/releases/14.html) was used to guide the quantification. We then calculated the Pearson's correlation coefficient (r) between a pair of samples using the equation 3.2. Here we excluded genes with zero read count. The heat map of Pearson's correlation coefficient between samples are shown in figure 4.8. For the footprint data, out of six samples we removed two outliers (the third replicate in control, and the second replicate in Silvestrol treatment) and the remaining two biological replicates showed excellent consistency (control, $r = 0.95$; Silvestrol, $r = 0.94$).

4.3.3 Initial Survey of the Data

First, we checked the overall changes at transcriptional and translational levels between control and drug treated samples. For each gene, we calculated the abundance measurements, Reads Per Kilobase per Million mapped reads (RPKM) [71], for both

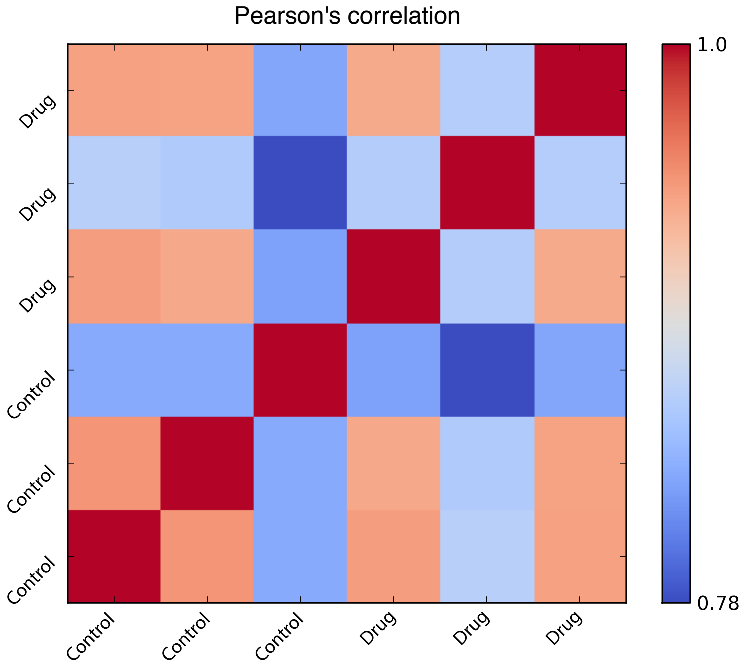

Figure 4.8: Pearson's correlation coefficient between ribosome footprint samples.

RNA-Seq and ribosome footprint data. Figure 4.9A shows the frequency of RPKM fold change of Silvestrol treated against control samples. It indicates the ribosome protected mRNA fragments were fewer in number and showed a wider variation of RPKM fold change than total mRNA, which confirmed that our choice of an early time point of treatment had indeed minimized transcriptional variation. Next, we examined the correlation between control and drug treated footprint data. Figure 4.9B shows a good linear relationship of the data under the two experimental conditions.

However, an assay of metabolic labeling of nascent proteins indicated a broad inhibitory effect on translation. We found that measurements of nascent protein synthesis with non-radioactive L-azidohomoalanine (AHA) labeling across a 4 hour time course revealed a progressive reduction of protein synthesis reaching 60% with Silvestrol and 80% with Cyclohexemide compared to control (Figure 4.9C). This assay was performed as following: after treated with Silvestrol, Cycloheximide or DMSO, cells were incubated in methionine free medium for 30 min prior to AHA labeling for one hour. Cells were then fixed with 4% paraformaldehyde in PBS for 15 min, permeablized with 0.25% Triton X-100 in PBS for 15 min followed by one wash with 3% BSA. Next, cells were stained using Alexa Fluor 488 Alkyne with Click-iT

Cell reaction Buffer Kit. Changes in mean fluorescence intensity as a measure of newly synthesized protein was detected by flow cytometry analysis.

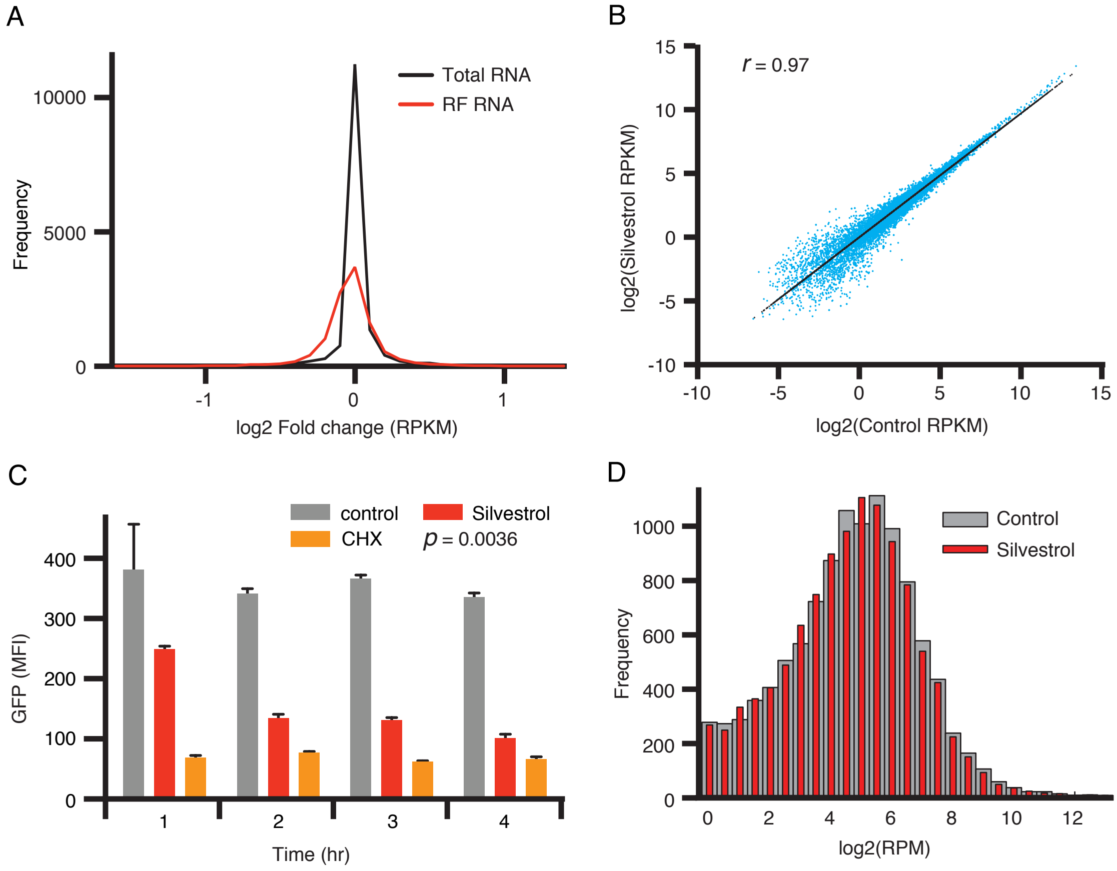

Figure 4.9: The global effect of Silvestrol on translation. A, Silvestrol-induced changes in total mRNA and ribosome occupied mRNA. B, RF abundance (RPKM) for genes across control and Silvestrol samples. C, Mean fluorescence intensity of incorporated AHA in newly synthesized proteins in KOPT-K1 cells treated with DMSO (control), Silvestrol (25 nM), or Cycloheximide (CHX; 100nM) for the indicated time periods. D, Histogram of the RF abundance of all genes (measured as unique RPM) for Silvestrol- and DMSO-treated cells.

Moreover, as shown in Figure 4.9D, the overall RPM (Reads Per Million mapped reads) frequency distribution of ribosome footprint from control and Silvestrol treated samples were largely overlapping, which indicated Silvestrol affected mRNAs were not limited to specific subgroups of mRNAs, e.g. those with especially high or low ribosome occupancy. Hence, Silvestrol produces a broad inhibitory effect on the translation of many transcripts in human T-ALL cells.

The RPKM and RPM was calculated by

$$RPKM = \frac{1}{N}\sum_{j=1}^{N} \frac{10^6 \cdot C^{i,j}}{K^{i,j} \cdot S^j} \quad \text{and} \quad RPM = \frac{1}{N}\sum_{j=1}^{N} \frac{10^6 \cdot C^{i,j}}{S^j} \tag{4.1}$$

where $C^{i,j}$ and $K^{i,j}$ are the read count and exonic length for gene i in sample j, respectively. S^j is the library size factor of sample j. N denotes the total number of samples in either Silvestrol-treated or control experiment.

4.3.4 Identify Gene with TE Change

As described previously, Silvestrol showed a broad spectrum of inhibitory effect on many genes, we sought to identify the targets whose ribosome footprint profiles were significantly changed in response to Silvestrol treatment. We used DEXSeq [70] to perform the statistical test. DEXSeq use a generalized linear model to identify the significant difference by assuming the read count—discrete random variable—follows negative binomial probabilistic distribution. Therefore, it takes into account the large biological variability between replicates, which has been demonstrated to be crucial to avoid a great number of false positives.

Although DEXSeq aims to detect differential exon usage from RNA-Seq data, here we used it in a specific way: for each gene, we fit the footprint and mRNA read counts into DEXSeq framework, in which Silvestrol treatment and control are two experimental conditions, and we tested whether footprint read counts (consisting 2 replicates for each condition) were significantly different between control and drug-treated conditions given the confounding factor mRNA abundance measurements (We split the 3 replicates of RNA-Seq data and recombined them into two pairs such that each of them consists of two replicates). Hence, this statistical framework identifies the gene with change in translation efficiency (TE), where TE is the ribosome footprint (RF) abundance normalized by mRNA expression from RNA-Seq. The analysis of TE change was further visualized by plotting a histogram of the log-ratio of TE in Silvestrol treated samples to controls (Figure 4.10). In the figure, a shift to the left is consistent with a broad inhibitory effect on translation. The significant gene targets were color-highlighted according to the statistical significance. Using a stringent cut-off ($p < 0.03$), we identified 281 transcripts whose TE was the most affected by Silvestrol (TE down), and 190 transcripts that were least sensitive

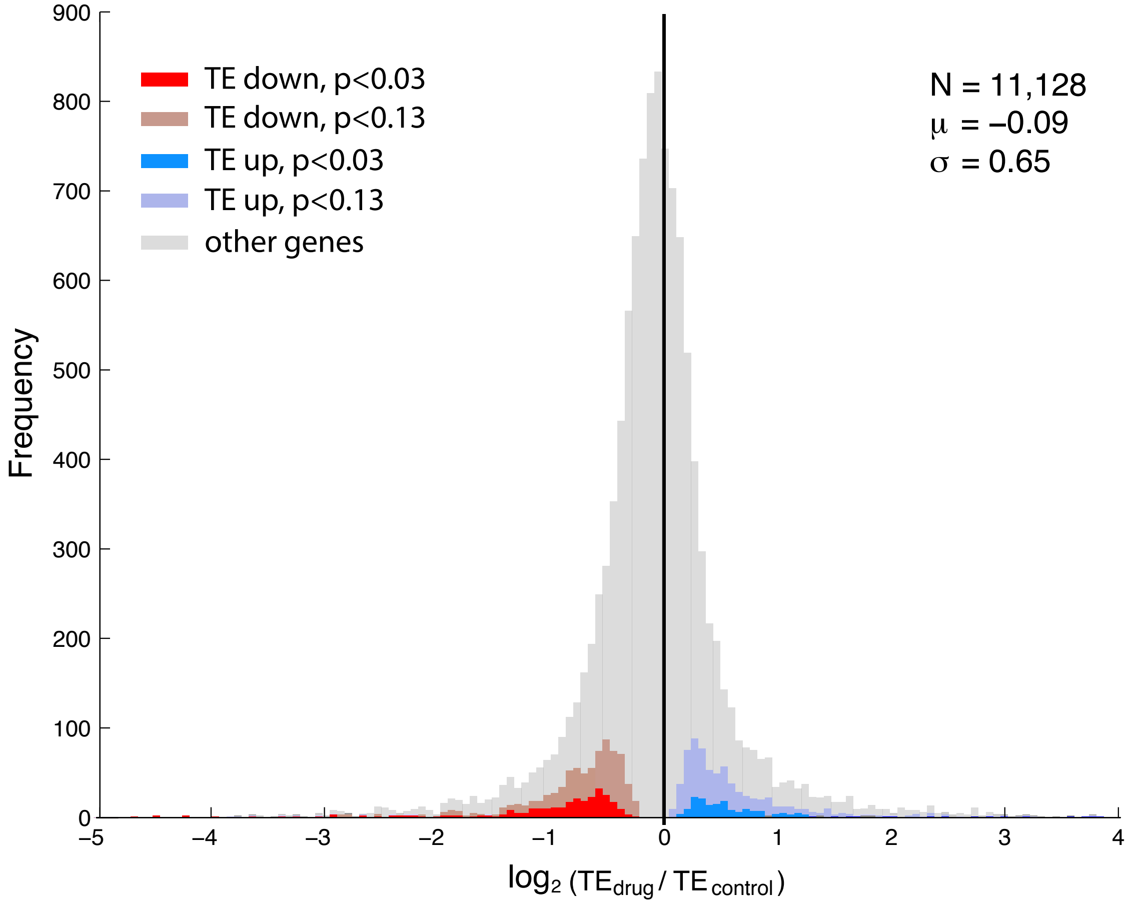

Figure 4.10: Histogram of the TE change. More or less affected genes are identified as TE down (red) and TE up (blue), respectively.

and showed an increase in translation efficiency (TE up).

4.3.5 Identify Gene with RF Density Change

In the previous section, we only considered changes in the abundance of ribosome footprint (RF) per gene as an indication of translation efficiency. However, we reasoned that changes in the density of ribosomes along the gene might provide an additional indication of translational effect of Silvestrol. Note that the RF read alignment provides exact positional information of the ribosome protected mRNA fragment on the gene. We used rDiff [86] to identify any significant changes of RF density across the length of the genes. Briefly, the BAM files containing alignment within exonic region and a GTF file containing the gene annotation information were the two inputs for rDiff. Here, rDiff maps the high dimensional alignment information (A_1 and A_2, denote for two experimental conditions) to a Hilbert space ($\mathcal{H}$) via a function ϕ as $[\phi : A_1 \rightarrow \mathcal{H}]$ and $[\phi : A_2 \rightarrow \mathcal{H}]$. It represents each alignment as one point in the $\mathcal{H}$ space by defining a mean map such that

$$\mu_{A_1} = \frac{1}{N}\sum_{i}^{N}\phi(A_1^i) \quad \text{and} \quad \mu_{A_2} = \frac{1}{N}\sum_{i}^{N}\phi(A_2^i). \tag{4.2}$$

Next, it computes the distance between the two mean maps (Maximum Mean Discrepancy) [131, 132] for the two alignments from Silvestrol treated and control samples:

$$MMD(A_1, A_2) = \parallel \mu_{A_1} - \mu_{A_2} \parallel_{\mathcal{H}} . \tag{4.3}$$

A nonparametric test with $10,000$ permutations was performed to detect distinct RF density. We found that 847 protein-coding genes showed a significant ($p < 0.001$) change in RF distribution. We refer to these as the rDiff positive set. Figure 4.11A, B and C are the three gene examples with different ribosome footprint density under the two conditions. As the RF density varies too much in different genes, we asked weather a common pattern exists if we average the RF density from all the gene in the rDiff positive set. To answer the question, we normalized read coverage for each transcript by the mean coverage of that particular transcript. Then the UTR and coding exon length were normalized in proportion to the total length of the corresponding regions in the rDiff positive genes. Finally all the normalized transcripts were averaged together to plot the RF density. The density was smoothed using "moving average" smoothing algorithm.

As shown in Figure 4.11D, these transcripts presents an high density in the 5' UTR and corresponding loss of coverage across the protein coding region in the Silvestrol treated experiment. This indicates unknown factor in the 5' UTR may lead to the observation.

4.3.6 Motifs of Silvestrol Sensitive Transcripts

Inspired by the observation found in rDiff positive genes, we also performed the similar analysis on TE down and TE up genes identified by DEXSeq. Interestingly, among the TE down genes we saw increased RF density in the 5' UTR and reduction across the coding sequence (figure 4.12A). In contrast, no observable change in the

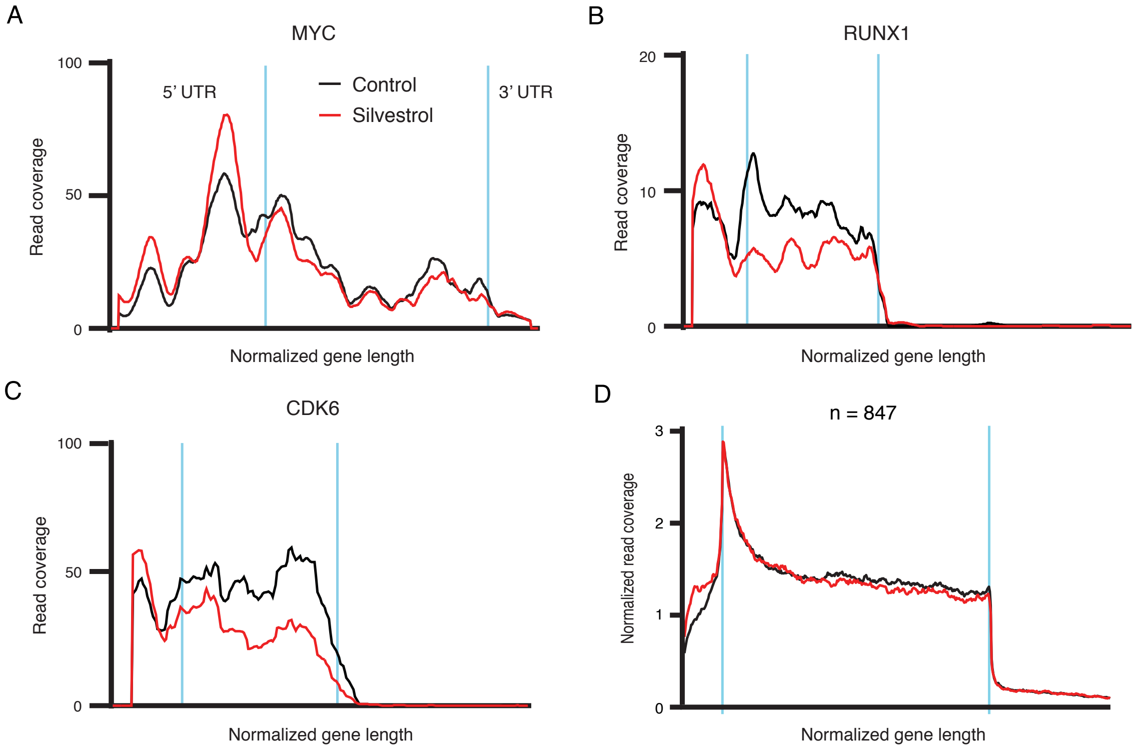

Figure 4.11: Ribosome footprint densities for indicated genes (A, B and C), and the overall density of all rDiff positive genes (D).

5' UTR and coding region in the TE up genes were presented (figure 4.12B). Based on these observations, we speculated a common feature exists in the 5' UTR of TE down and rDiff positive genes.

5' UTR length has been implicated in translational control [133]. Comparing the 5' UTR length across TE up, TE down and background transcripts, we observed that mRNAs with longer 5' UTRs were significantly enriched among the most Silvestrol sensitive mRNAs (average length: 368 nt in TE down; 250 nt in background), whereas the TE up group showed no significant difference in 5' UTR length (average length: 265 nt) (Figure 4.13). The background was randomly selected from the genes that neither belong to TE down, TE up nor rDiff positive genes.

We also searched for known translation regulatory elements, for example, TOP [134] / TOP-like sequences [63], internal ribosome entry sites (IRES) [135], and pyrimidine rich translational elements (PRTEs) [134]. We found no predilection for TOP, TOP-like, PRTE, or IRES elements in TE down genes. Whereas the TE up group showed an enrichment for IRES elements.

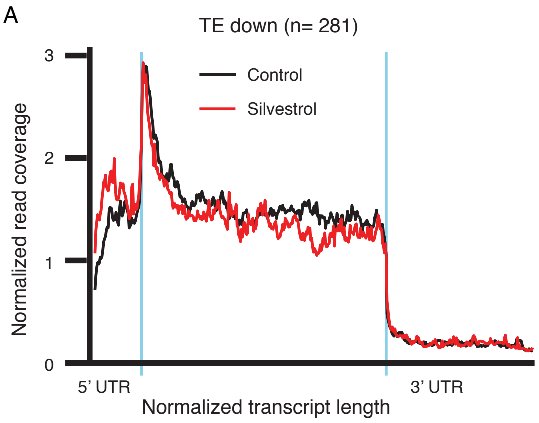

Figure 4.12: Ribosome footprint densities for TE down (A) and TE up (B) gene. Before sum up the RF densities of all the transcripts, the density was normalized by the mean coverage of each transcript. Therefore, the density curve indicates the distribution change along itself instead of the absolute coverage change between Silvestrol treated and control samples.

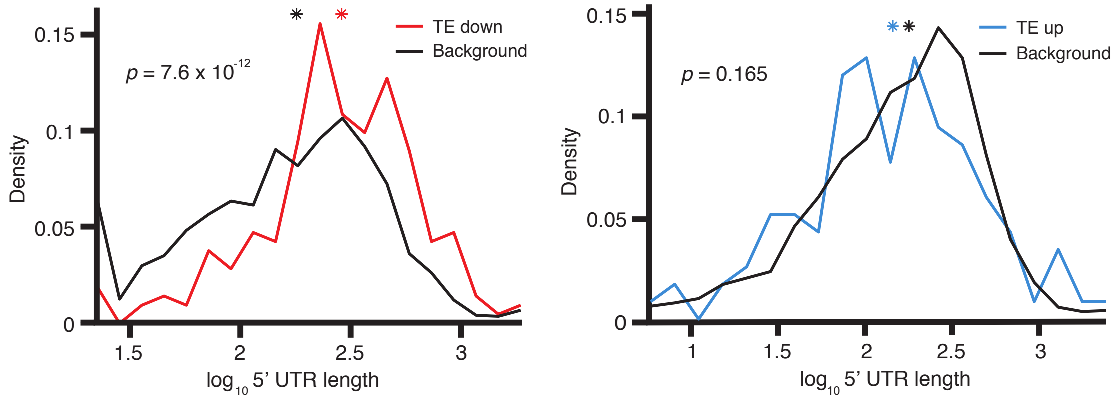

Figure 4.13: Comparison of 5' UTR lengths for TE down or up versus background transcripts. Asterisks indicate mean of transcript lengths. *P* values were calculated from two-sample Kolmogorov-Smirnov test.

Next we asked the question whether a shared motif could be identified in the 5' UTR of TE down and up genes. In order to illustrate this, we firstly quantified the transcripts for every gene based on the RNA-Seq data using MISO [136]. The 5' UTR of the dominated transcript was collected for predicting motifs. We used DREME [137] to search for significantly enriched sequence motifs in the TE down and TE up groups compared to the background list. Over represented motifs were determined with two different settings: searching for *k*-mer length greater than or equal to nine and twelve base pairs. We considered the predicted consensus sequences with $p < 1 \times 10^{-4}$ as significant motifs.

This analysis retrieved a 12-mer $(\mathrm{GCC})_4$. The GC-rich sequence pattern that was significantly over represented among the TE down transcripts (94 out of 281 , $p = 6.19 \times 10^{-5}$) (Figure 4.14). In addition, we found 14 similar 9-mer variations of this motif that were similarly enriched in the TE down group where 177 out of 281 transcripts harbored at least one and often multiple occurrences of these 9-mer motifs ($p = 9.28 \times 10^{-5}$) (Figure 4.14). On the other hand, the TE up mRNAs did not share a recognizable motif.

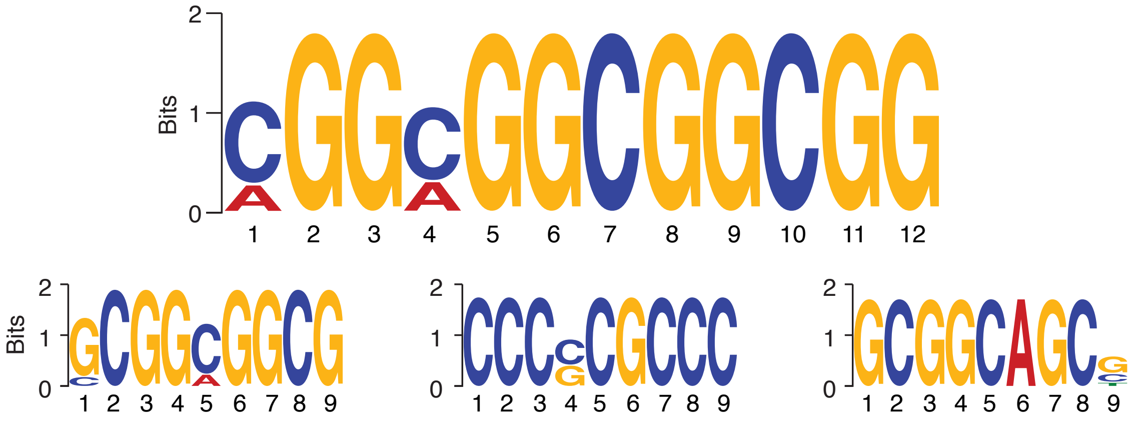

Figure 4.14: The twelve- and nine-nucleotide motifs enriched in TE down transcripts.

Next, we investigated whether rDiff positive genes share the common features with TE down group. Similar to the TE down genes, the rDiff positive mRNAs are significantly enriched for longer 5' UTRs ($p = 0.004$). Further, the rDiff positive genes showed no significant enrichment of TOP, PRTE, or IRES elements. Surprisingly, DREME analysis for a recurrent motif identified a single 12-mer motif in 233 out of 826 transcripts with only one variable nucleotide ($p = 5.45 \times 10^{-5}$) (Figure 4.15). We also identified three additional 9-mer motifs that with similar enrichment and similar GC-rich composition (Figure 4.15).

4.3.7 From Motif to Structure

We also wondered whether genes with TE change have specific structural feature that set them apart from the rest transcripts. Using RNAfold [138] we observed a striking enrichment for G-quadruplex structure [139] among the TE down genes

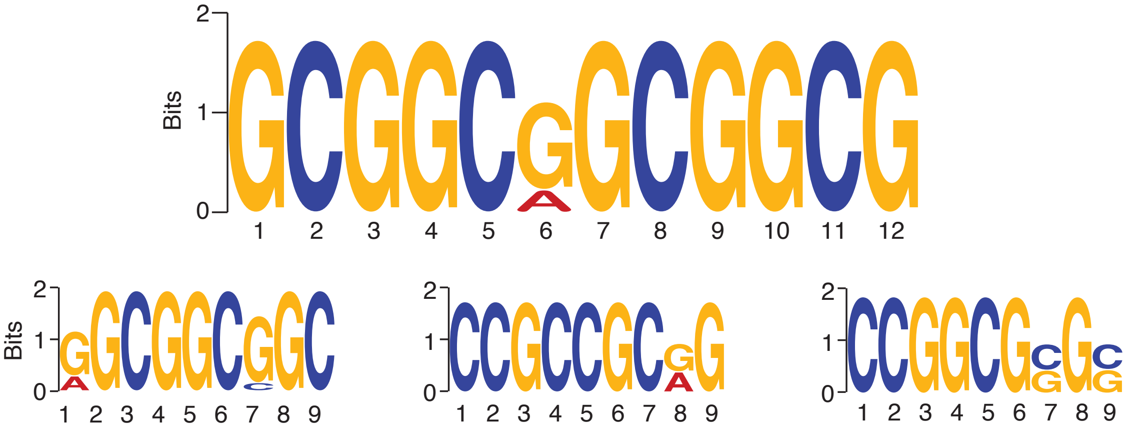

Figure 4.15: The twelve- and nine-nucleotide motifs enriched in rDiff positive transcripts.

($p = 0.89x10^{-10}$, Figure 4.16A). Specifically, 69 of 220 TE down mRNA harbored at least one G-quadruplex in their 5' UTR. Moreover, 48 out of 79 G-quadruplex structures perfectly localized to the $(GCC)_4$ 12-mer motif, and 50 out of 79 localized to the 9-mer motifs and the neighboring nucleotides (Figure 4.16B). In contrast, the TE up mRNAs did not show significant difference compared to background mRNAs (Figure 4.16A).

As described before, the enriched 12-mer motif in rDiff positive genes was nearly identical to the TE down motif, and showed a similar pattern of alternating guanines with one linking cytosine (Figure 4.14 and Figure 4.15). The G-quadruplex prediction also indicated an enrichment of this structure in the 5' UTR of rDiff positive set ($p = 0.0038$, Figure 4.16A). Similarly, the majority of predicted G-quadruplexes exactly localized to the 12-mer and 9-mer motifs in the 5' UTR of the positive mRNAs (Figure 4.16B).

G-quadruplex structures are based on non-Watson-Crick interactions between at least four paired guanine nucleotides that align in different planes and are connected by a linker nucleotide [139] (Figure 4.17A). In our study, we most often observed two guanines separated by an intervening cytosine and sometimes an adenine. The ADAM10 5' UTR provides an example showing the RNA secondary structure containing four predicted G-quadruplex (Figure 4.17B). Two of them match the typical 12-mer motif sequence and the other two are formed by a longer sequence including elements that are similar although not identical to the canonical $(GCC)_4$ motif.

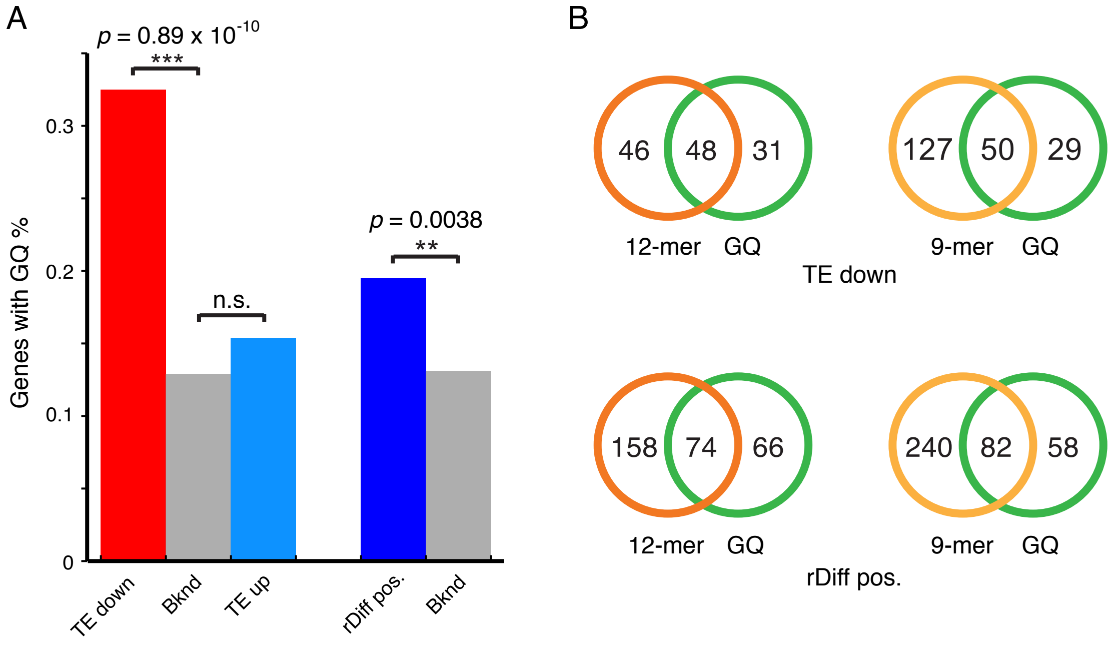

Figure 4.16: Computational analysis of G-quadruplex. A, Enrichment of G-quadruplex in the 5' UTR of TE down and rDiff positive mRNAs. B, Majority of G-quadruplex is located at the same position with GC rich motifs.

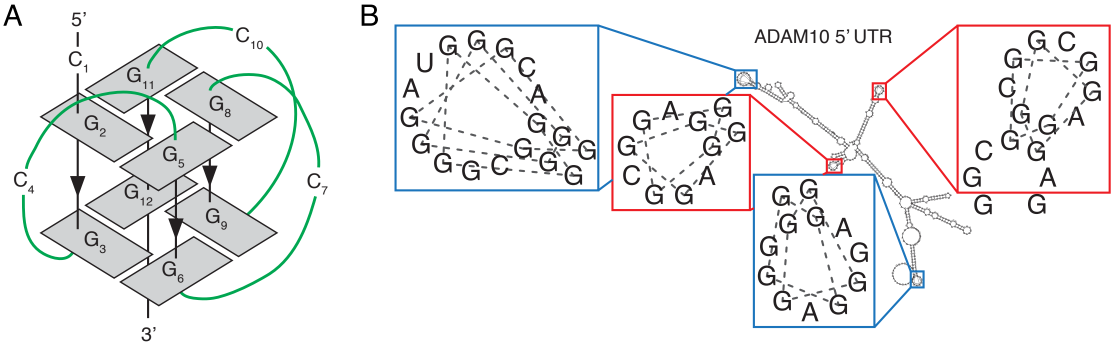

Figure 4.17: The example of G-quadruplex structure. A, Diagram of parallel G-quadruplex conformation. Subscript numbers denote the nucleotide position. B, The structure of ADAM10 5' UTR with canonical motif formed G-quadruplex (red box) and non-canonical motif formed G-quadruplex (blue box).

4.4 Hallmark of Silvestrol Affected mRNAs

4.4.1 $(\mathrm{GCC})_4$ Motif Forms G-quadruplex

To validate whether the 12-mer and many extended 9-mer motifs can form RNA G-quadruplex structure, we performed Circular Dichroism experiment. We compared the molar ellipticity of three different RNA oligomers including the $(\mathrm{GCC})_4$ motif, a known G-quadruplex found in the human telomeric RNA [140] and a control oligomer with equal GC content and length but reshuffled guanine and cytosine order. We observed the positive and negative molar ellipticity peaks at 264 nm and 240 nm for both oligomers of $(\mathrm{GCC})_4$ and human telomeric sequence, whereas the control oligomer showed a shift in peak wavelengths to 270 and 233 nm (Figure 4.18A). We also examined RNA oligomers encoding the 9-mer motifs and included two flanking nucleotides exactly as they occurred in the 5' UTRs of genes with predicted 9-mer motifs (MTA2, TGFB1, MAPKAP1, ADAM10). These oligomers showed the same pattern and the typical molar ellipticity peaks at 264 nm and 240 nm indicative of a parallel GQ structure (Figure 4.18B).

Circular Dichroism combined with thermal unfolding study revealed that the melting temperature for the $(\mathrm{GCC})_4$ motif was higher (56 °C) than the control oligomer (49 °C). Also, the free energy of unfolding was higher for the $(\mathrm{GCC})_4$ motif compared to the control oligomer, with a difference of -32 kcal/mol (Figure 4.18C). Similarly, computational prediction of the complete 5' UTR sequences showed a decrease of the predicted minimum free energy with increasing number of predicted G-quadruplex structures (Figure 4.18D). Together, these results demonstrated that both 12-mer and 9-mer motifs can contribute to form GQ structure that represent a more stable state of the 5' UTR.

4.4.2 G-quadruplex and eIF4A Affect Translation

We directly tested the effects of the $(\mathrm{GCC})_4$ motif in a translation reporter assay. We constructed a luciferase reporter system to compare the translational effect of G-quadruplex to the previously used control oligomer of the same length and GC content. Briefly, four $(\mathrm{GCC})_4$ or control oligomer in tandem were cloned into the 5' UTR of Renilla luciferase plasmid pGL4.73 and HCV-IRES firefly were used as internal control. As shown in Figure 4.19A and B, treatment with Silvestrol (25 nM)

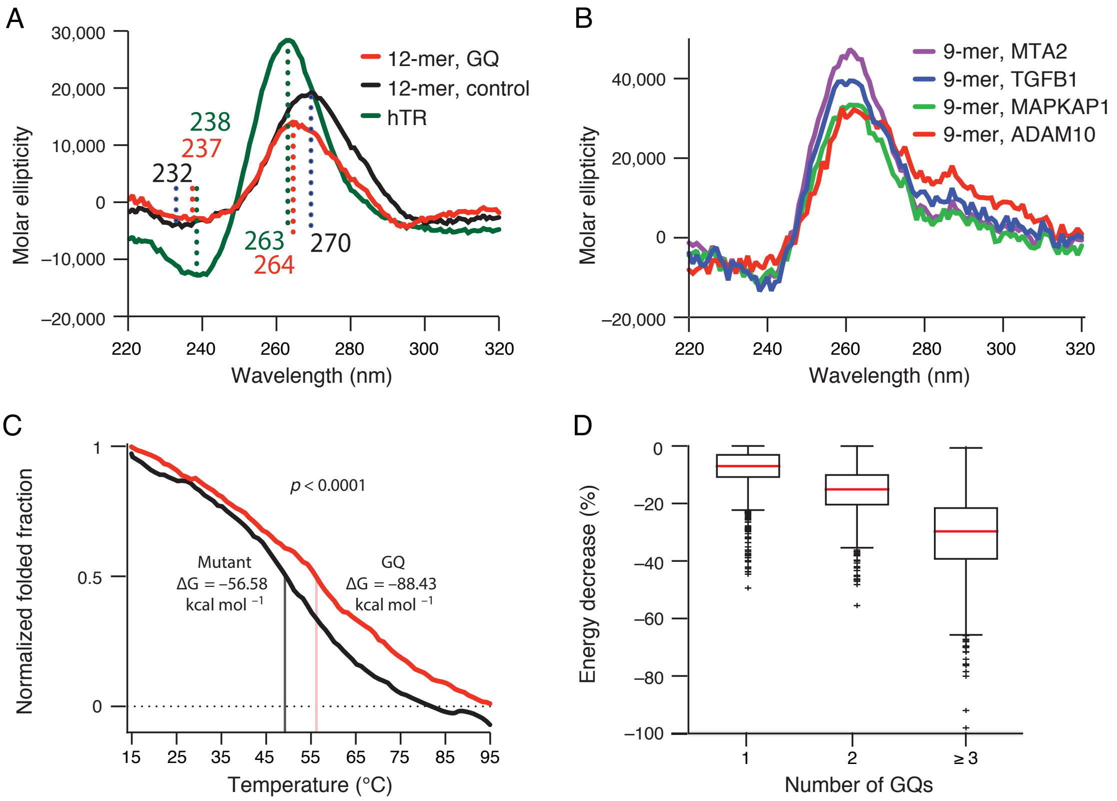

Figure 4.18: Circular dichroism spectra of the oligomers. A, $(GCC)_4$ motif, 12-mer control oligomer and human telomeric RNA (hTR) folded in KCl. B, Nine-mer motifs with flanking nucleatides from the 5' UTR of indicated genes folded in KCl. C, Melting curve for circular dichroismspectra scan at a wavelength of 264nm for the $(GCC)_4$ and the control oligomer. ΔG, free energy of unfolding. D, Box plot of the free energy decrease for UTRs with 1, 2 or ≥ 3 motifs.

reduced the translation of the G-quadruplex construct and did not affect the control, whereas cycloheximide (20 nM) equally suppressed both reporters. Although other RNA helicases (DHX9, DHX36) have been reported in resolving G-quadruplex structure [140, 141], RNAi-mediated eIF4A knockdown in the same reporter assay confirmed an eIF4A dependent effect on the G-quadruplex reporter with little effect on the control sequence (Figure 4.19C).

Given all evidences together, we conclude that long 5' UTR and the $(GCC)_4$ motif or highly similar sequence patterns that can form G-quadruplex structure are the hallmarks of eIF4A-dependent and Silvestrol sensitive translation. Our results demonstrate that, in T-ALL cells, the RNA helicase activity of oncogenic eIF4A is required to unwind the G-quadruplex structure in the 5' UTR of many can-

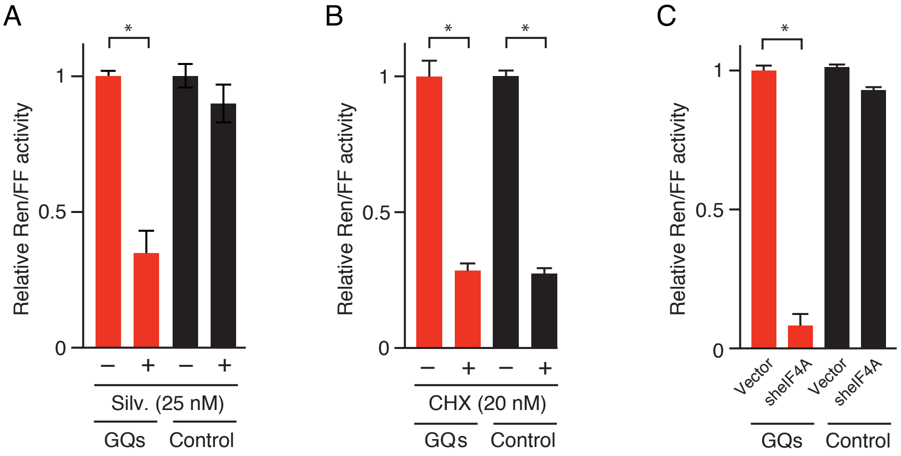

Figure 4.19: Relative *Renilla* luciferase expressed from the G-quadruplex (red bars) or control oligomer (black bars), treated with Silvestrol (Silv., panel A), cyclohexamide (CHX. panel B) or knockdown eIF4A (panel C). *, $p < 0.05$.

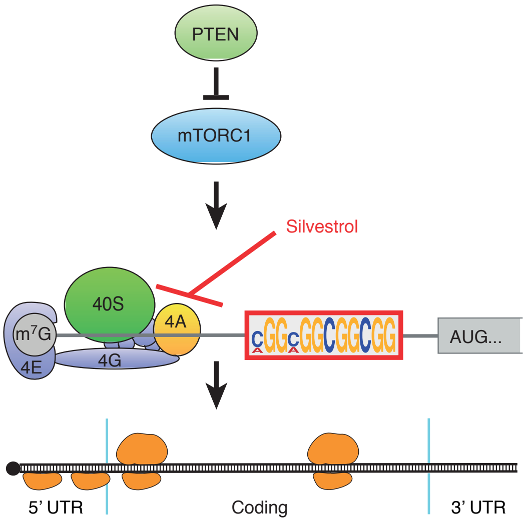

Figure 4.20: Diagram showing the mechanism of eIF4A-dependent translational control.

cer related genes and initiates their translation. The anticancer drug, Silvestrol, selectively blocks the translation by inhibiting the oncogenic eIF4A, causing the ribosomes accumulate in the 5' UTR and a down-regulated translation efficiency, which consequently suppresses the T-ALL leukaemia development Figure 4.20. The cancer related genes are highlighted in Figure 4.21.

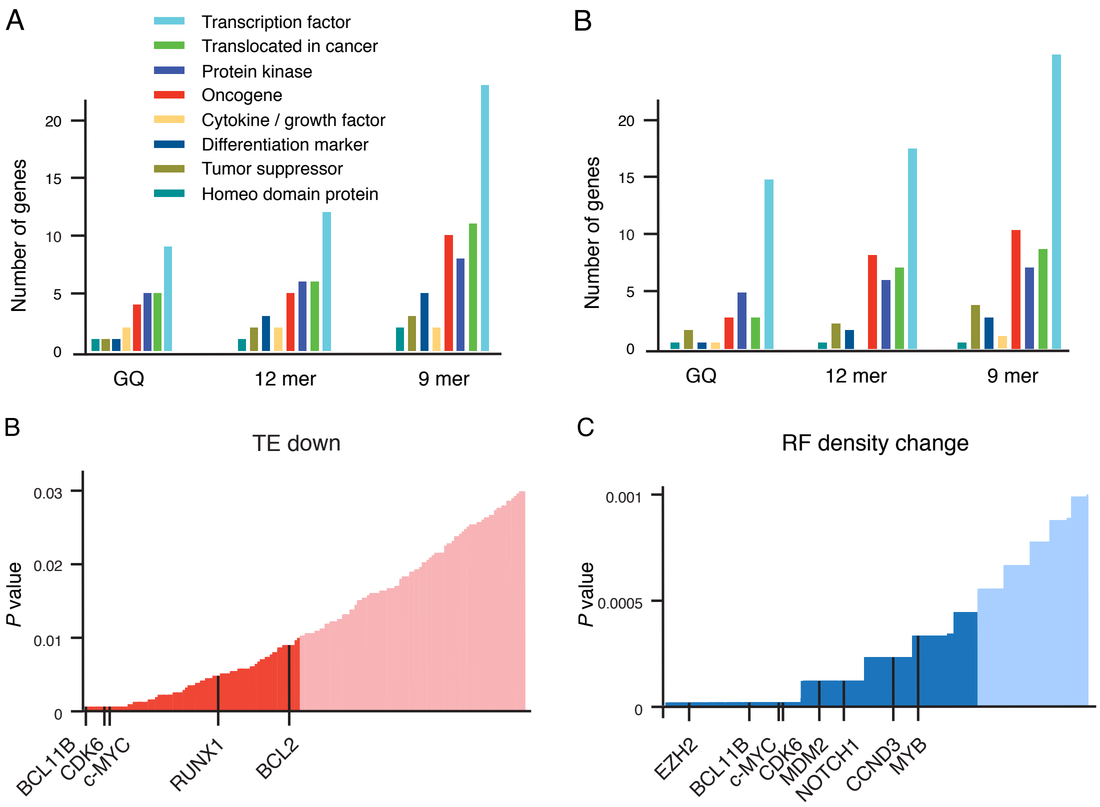

Figure 4.21: Silvestrol sensitive mRNAs. A and B, Gene ontology classification for TE down and rDiff positive groups, respectively. C, TE down genes ranked by P value. D, rDiff positive genes ranked by P value.

4.5 Discussion

Recent studies have described individual examples of G-quadruplex structures in the N-RAS gene, where it also limits translation, or in the VEGF IRES-element where it is thought to enhance IRES dependent translation [142, 143, 144]. Additional computational analyses have suggested that these structures may have a broader role in translational control although the molecular mechanism has remained unclear [145].

In this study, we used ribosome footprinting strategy and RNA-Seq to identify the key features in the 5’ UTR that confer a requirement for the eIF4A RNA helicase for translation. Sepcifically, longer 5’ UTRs and a 12-mer $(GCC)_4$ motif confer eIF4A dependence. In some instances, the computational analysis also identified enrichment of several 9-mer motifs with variation in their neighboring nucleotides. Importantly, the 12-mer and 9-mer motifs precisely localize to more than half of all predicted RNA G-quadruplex structures. This result is very striking, even taking into account the limitations of available computational methods to identify sequence motifs and predict RNA structures.

Our findings indicate that the GC-rich motifs or its structural form G-quadruplex represents a translational control element that is encoded in the 5' UTR of several hundred mRNAs and imparts a requirement for eIF4A RNA helicase activity.

Note that we identified the genes showing translation efficiency change induced by the drug treatment. This is not equivalent to identifying distinct ribosome footprint profiles. Because any transcriptional changes leads to the alteration in translation. On the other hand, even no TE change is observed, it is still not sufficient to conclude the absence of translational control, as the ribosome density distribution along the mRNA can be completely different.

We started this study when the *RiboDiff* has not been developed. Although *DEXSeq* [70] framework is capable of addressing the confounding issue, it aims to detect the differential read counts from RNA-Seq experiments. Therefore, some technical limitations exist. A further investigation using *RiboDiff* on the same data indicates twice as many as candidate genes from *DEXSeq* are detected, and more than 90% genes from *DEXSeq* are included in the new gene sets, providing an opportunity to examine the comprehensive drug-sensitive gene profile.

5 Discussion

Summary and Outlook

My doctoral research mainly focused on the study of protein translational control by using high-throughput ribosome profiling and RNA-Seq methods. In collaboration with Wendel lab, we discovered the role of RNA G-quadruplex in eIF4A-dependent oncogene translation in leukaemia. The eukaryotic translation initiation factor 4A (eIF4A) is an RNA helicase that is required to promote the protein translation for hundreds of genes including oncogenes and transcriptional factors. Many of those harbor the guanine quartet $(GCC)_4$ motif in the 5' UTR of their mRNAs. The guanine quartet forms into an RNA G-quadruplex structure that needs RNA helicase eIF4A to unwind in order to facilitate the ribosome 40S small subunit to go through and initiate the translation at the start codon. The anti-cancer compound Silvestrol suppresses the leukaemia development by interacting with eIF4A and blocking its RNA helicase activity, therefore inhibits the oncogenes' translation. Note that after treating the lymphoblastic leukaemia cell line with Silvestrol, we still observed ribosome footprints in the coding regions although the footprint density decreases (see some examples in Figure 4.11A-C). This indicates the ribosome small subunit can slide along the 5' UTR even without the assistant of eIF4A. In other words, RNA G-quadruplex serves as a rate limiter to control the translation of the oncogenes. In a normal state, these genes are mildly translated for the cells to satisfy their basic physiological requirements and maintain a balance between proliferation and apoptosis. However, if the eIF4A is activated under certain circumstances, the consequence of the up-regulation of the translation for these oncogenes may lead to the development of cancer.

Study on the biology of protein translation motivated us to develop a computational approach that can identify genes under specific translational regulation in different conditions. As the translational landscape provided by the ribosome profiling method is fundamentally confounded by the transcription abundance, we need a statistical framework that can distinguish the translational differences between conditions while takes both transcription level variability and the large variance of read count data across biological replicates into account. In the second half of my doctoral study live, I worked on developing *RiboDiff*, an approach and software that use the negative binomial distribution in a generalized linear model to estimate the over dispersion across samples for RNA-Seq and ribosome profiling data separately, and detects the genes with differential translational regulation while controls the transcriptional differences. Similar to other tools for differential gene expression analysis, such as *DESeq* and *edgeR*, *RiboDiff* detects the *relative* up- and down-regulations from the quantification measurements. This is because the count data of all samples are first sequencing library size corrected in order to eliminate the differences on sequencing depth for each sample. However, if a certain experimental treatment has a global effect on almost all the genes' translation, the normalization of library size neutralizes the global effect and we obtain the relative translational efficiency changes for a gene compared to other genes after the treatment. Hence, the assumption of *RiboDiff* is only a subset of genes are under differential translational regulation.

In the next two sections, I will briefly introduce two other projects that I also contributed to during my dotoral study.

MYC Shapes Cellular Metabolism through Selective Translation Targets

This project is a collaboration with Wendel lab at MSKCC studying the gene MYC in cancer (manuscript submitted). It has been widely recognized that MYC is a transcriptional factor that regulates cell cycle, apoptosis and cellular transformation [146]. Constitutively expressed MYC in human leads to many types of cancer [146]. It has also been reported that MYC can affect protein translation by altering the ribosome biogenesis through transcription regulation of rRNA and ribosomal proteins [147]. In this study, we ask the question of whether MYC can modulate translation through certain specific ways except globally regulating ribosome biogenesis. We switch off MYC in a B-cell lymphoma cell line whose MYC is conditional regulated by tetracycline (Figure 5.1A and B), and performed the RNA-Seq and ribosome profiling experiments in parallel (Figure 5.1C). The sequencing data were processed through our computational pipeline (Chapter 3). After preprocessed the

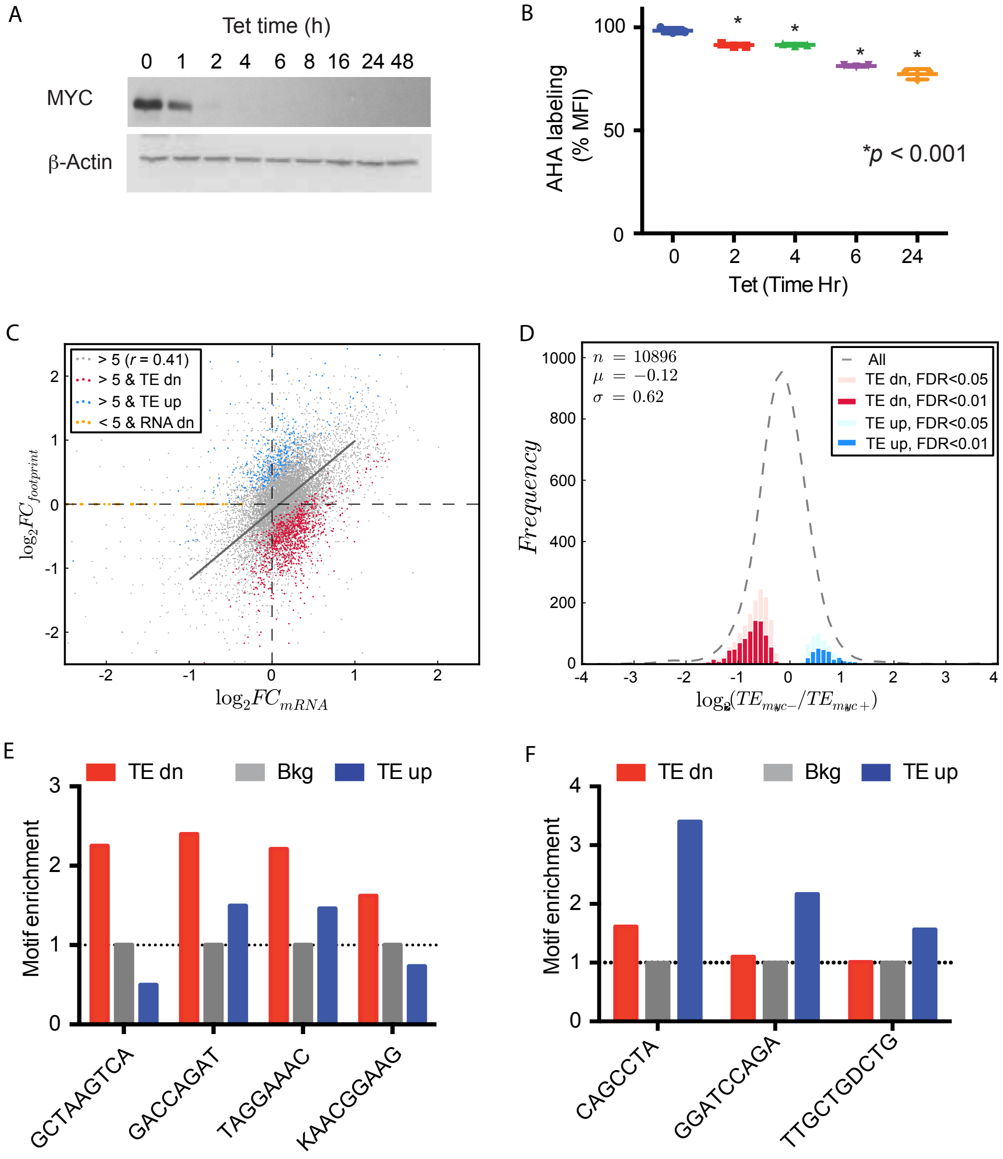

Figure 5.1: The global translational effect of knocking down MYC. A, Western blot of MYC after being switched off. B, Global translation measured by L-azidohomoalanine (AHA) decreased in the time course after switched off MYC. C, Scatter plot of fold change of RNA-Seq against fold change of ribosome profiling. Each dot represents a gene. A linear correlation between the two measurements is observed. The red and blue dots are genes identified as TE down and up. Lowly translated genes (footprint read count smaller than 5) but show transcriptional difference are labelled in yellow. D, Histogram of TE fold-change for all genes. E and F, Enriched motifs identified in TE down and up gene sets.

data by our computatioanl pipeline and statistical test by *RiboDiff*, we identified $2,115$ genes shows significant difference in the translation efficiency at $FDR < 0.05$. Among these genes, $1,530$ genes are TE down-regulated, whereas 585 genes are up-regulated (Figure 5.1D). Further computational analysis revealed that four RNA motifs are enriched in the 5' UTR of the TE down genes, while three motifs are enriched in TE up genes (Figure 5.1E and F). Currently, our work is mainly focusing on identifying the specific proteins (RBM42 and SRSF1 *etc.*) that bind to the predicted motifs and investigating how MYC governs those RNA binding proteins.

Interferon-γ Regulates mRNA Translation to Activate Macrophage

In addition to the described projects, I also participated in a collaboration with Ivashkiv lab at Hospital for Special Surgery. In this study, we elucidated the interferon-γ (IFN-γ), a cytokine produced by natural killer (NK) cells, potentiates macrophage activation by regulating mRNA translation and cellular metabolism. Briefly, IFN-γ suppresses both MAPK-MNK-eIF4E and mTORC1 signaling pathways, which prime macrophages for enhanced bacterial and virus killing and inflammatory activation. Ribosome profiling study of primary human monocytes and macrophages indicated the co-existence of translationally down-regulated (TE down) genes and up-regulated (TE up) genes. In addition to the TE down observation, a consequence of the suppression of the two signaling pathway, we performed miRNA-Seq for human primary macrophages to investigate the global miRNA expression profile. We found that IFN-γ suppressed the expression of 54 miRNAs, which supports the hypothesis that IFN-γ increases the translation of genes by suppressing miRNA expression. Further miRNA target analysis indicated many TE up genes contain the reverse complimentary seed sequence of the suppressed miRNA in their 3' UTR regions. Figure 5.2 shows an example of TE up mRNAs that are targeted by transcriptionally suppressed miR-146b-3p. This observation indicates miRNAs potentially be involved in translational control. Further experimental validatons are needed for detailed studies.

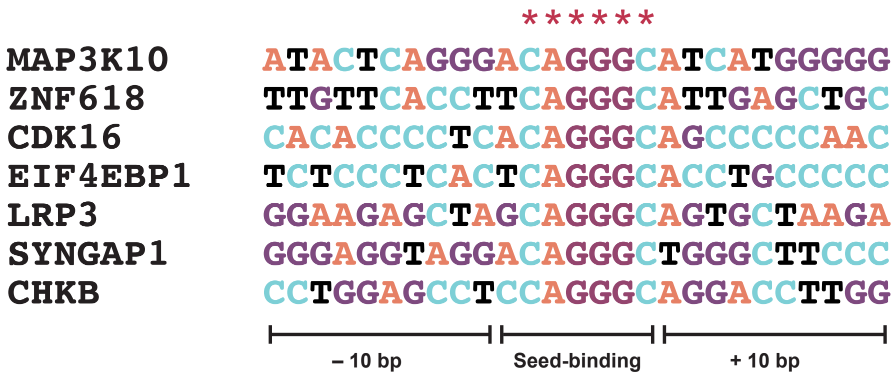
* * * * * *
MAP3K10 ATACTCAGGGACAGGGCATCATGGGGG
ZNF618 TTGTTCACCTTCAGGGCATTGAGCTGC
CDK16 CACACCCCTCACAGGGCAGCCCCCAAC
EIF4EBP1 TCTCCCTCACTCAGGGCACCTGCCCCC
LRP3 GGAAGAGCTAGCAGGGCAGTGCTAAGA
SYNGAP1 GGGAGGTAGGACAGGGCTGGGCTTCCC
CHKB CCTGGAGCCTCCAGGGCAGGACCTTGG
– 10 bp
Seed-binding
+ 10 bp

Figure 5.2: Potential mRNA target of miR-146b-3p that were translationally up-regulated by IFN-γ. Asterisks indicate the 3' UTR sequence complementary to the seed sequence of miR-146b-3p. Source: *Nature Immunology.* 2015 Aug;16(8):838-49.

Future work

RiboDiff is a software facilitates researchers studying translational regulation. The idea of developing this statistical framework was gained from the tight working with biologists. As I am approaching to the end of the doctoral study, it is the time to plan for the future career and decide the direction to go in the next step. Until now, I have already worked on genomics, proteomics, mRNA translation. It would be great if I can get one step back to the transcriptomics and fill the gap between genomics and mRNA translation. In addition, I would like to continue working at the border of method development and biological research. Immunology is a subject that studies the molecular and cellular components that comprise the immune system in both health and disease conditions. Furthermore, immunotherapy has been used in cancer treatment together with chemotherapy and radiotherapy, which has been proved to be a promising strategy for certain cases. After finish my doctoral research, I will work as a postdoc with Prof. Rudensky at Memorial Sloan Kettering Cancer Center. My goal is to use the computational methods and statistical modeling on epigenomic data to understand the molecular mechanism governing the differentiation and function of CD4 T lymphocytes and their role in immunity and tolerance. More specifically, I will study the roles of regulatory T cells in control of tumor immunity and immunity to infections, and in maintenance of immune homeostasis at environmental interfaces.

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 6 edition, Nov 2014.

[2] Anne M. Coghill and Lorrin R. Garson. *The ACS Style Guide: Effective Communication of Scientific Information*. American Chemical Society, 3 edition, July 2006.

[3] D A Steinhauer and J J Holland. Rapid evolution of rna viruses. *Annu Rev Microbiol*, 41:409–33, 1987.

[4] J D Watson and F H Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, Apr 1953.

[5] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–45, Oct 2004.

[6] Sybil P. Parker and McGraw-Hill Companies. *McGraw-Hill Encyclopedia of Science and Technology*. McGraw-Hill, 8 edition, Feb 1997.

[7] K Luger, A W Mäder, R K Richmond, D F Sargent, and T J Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–60, Sep 1997.

[8] R D Kornberg. Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139):868–71, May 1974.

[9] C E Ford and J L Hamerton. The chromosomes of man. *Nature*, 178(4541):1020–3, Nov 1956.

[10] Geoffrey M. Cooper and Robert E. Hausmann. *The Cell: A Molecular Approach*. Sinauer Associates, Inc, 6 edition, Feb 2013.

[11] Tamar Juven-Gershon and James T Kadonaga. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol*, 339(2):225–9, Mar 2010.

[12] Anjanabha Saha, Jacqueline Wittmeyer, and Bradley R Cairns. Chromatin remodelling: the industrial revolution of dna around histones. *Nat Rev Mol Cell Biol*, 7(6):437–47, Jun 2006.

[13] Cedric R Clapier and Bradley R Cairns. The biology of chromatin remodeling complexes. *Annu Rev Biochem*, 78:273–304, 2009.

[14] I Grummt. Regulation of mammalian ribosomal gene transcription by rna polymerase i. *Prog Nucleic Acid Res Mol Biol*, 62:109–54, 1999.

[15] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. Microrna genes are transcribed by rna polymerase ii. *EMBO J*, 23(20):4051–60, Oct 2004.

[16] I M Willis. Rna polymerase iii. genes, factors and transcriptional specificity. *Eur J Biochem*, 212(1):1–11, Feb 1993.

[17] A M Lesk. Why does dna contain thymine and rna uracil? *J Theor Biol*, 22(3):537–40, Mar 1969.

[18] W Schul, B Groenhout, K Koberna, Y Takagaki, A Jenny, E M Manders, I Raska, R van Driel, and L de Jong. The rna 3' cleavage factors cstf 64 kda and cpsf 100 kda are concentrated in nuclear domains closely associated with coiled bodies and newly synthesized rna. *EMBO J*, 15(11):2883–92, Jun 1996.

[19] S Bienroth, W Keller, and E Wahle. Assembly of a processive messenger rna polyadenylation complex. *EMBO J*, 12(2):585–94, Feb 1993.

[20] Nick J Proudfoot, Andre Furger, and Michael J Dye. Integrating mrna processing with transcription. *Cell*, 108(4):501–12, Feb 2002.

[21] A J Shatkin. Capping of eucaryotic mrnas. *Cell*, 9(4 PT 2):645–53, Dec 1976.

[22] Hagen Tilgner, David G Knowles, Rory Johnson, Carrie A Davis, Sudipto Chakrabortty, Sarah Djebali, João Curado, Michael Snyder, Thomas R Gingeras, and Roderic Guigó. Deep sequencing of subcellular rna fractions shows

splicing to be predominantly co-transcriptional in the human genome but inefficient for lncrnas. *Genome Res*, 22(9):1616–25, Sep 2012.

[23] Stefanie A Mortimer, Mary Anne Kidwell, and Jennifer A Doudna. Insights into rna structure and function from genome-wide studies. *Nat Rev Genet*, 15(7):469–79, Jul 2014.

[24] M W Nirenberg and J H Matthaei. The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A*, 47:1588–602, Oct 1961.

[25] M Nirenberg, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman, and C O'Neal. Rna codewords and protein synthesis, vii. on the general nature of the rna code. *Proc Natl Acad Sci U S A*, 53(5):1161–8, May 1965.

[26] Julius Rabl, Marc Leibundgut, Sandro F Ataide, Andrea Haag, and Nenad Ban. Crystal structure of the eukaryotic 40s ribosomal subunit in complex with initiation factor 1. *Science*, 331(6018):730–6, Feb 2011.

[27] Sebastian Klinge, Felix Voigts-Hoffmann, Marc Leibundgut, Sofia Arpagaus, and Nenad Ban. Crystal structure of the eukaryotic 60s ribosomal subunit in complex with initiation factor 6. *Science*, 334(6058):941–8, Nov 2011.

[28] Adam Ben-Shem, Nicolas Garreau de Loubresse, Sergey Melnikov, Lasse Jenner, Gulnara Yusupova, and Marat Yusupov. The structure of the eukaryotic ribosome at 3.0 å resolution. *Science*, 334(6062):1524–9, Dec 2011.

[29] Marc Thiry and Denis L J Lafontaine. Birth of a nucleolus: the evolution of nucleolar compartments. *Trends Cell Biol*, 15(4):194–9, Apr 2005.

[30] R M Seiser and C V Nicchitta. The fate of membrane-bound ribosomes following the termination of protein synthesis. *J Biol Chem*, 275(43):33820–7, Oct 2000.

[31] Andrija Finka, Vishal Sood, Manfredo Quadroni, Paolo De Los Rios, and Pierre Goloubinoff. Quantitative proteomics of heat-treated human cells show an across-the-board mild depletion of housekeeping proteins to massively accumulate few hsps. *Cell Stress Chaperones*, 20(4):605–20, Jul 2015.

[32] A Rich and U L RajBhandary. Transfer rna: molecular structure, sequence, and properties. *Annu Rev Biochem*, 45:805–60, 1976.

[33] M Ibba and D Soll. Aminoacyl-trna synthesis. *Annu Rev Biochem*, 69:617–50, 2000.

[34] Nahum Sonenberg and Alan G Hinnebusch. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–45, Feb 2009.

[35] George W Rogers, Jr, Anton A Komar, and William C Merrick. eif4a: the godfather of the dead box helicases. *Prog Nucleic Acid Res Mol Biol*, 72:307–31, 2002.

[36] Mamatha Bhat, Nathaniel Robichaud, Laura Hulea, Nahum Sonenberg, Jerry Pelletier, and Ivan Topisirovic. Targeting the translation machinery in cancer. *Nat Rev Drug Discov*, 14(4):261–78, Apr 2015.

[37] Sarah E Kolitz and Jon R Lorsch. Eukaryotic initiator trna: finely tuned and ready for action. *FEBS Lett*, 584(2):396–404, Jan 2010.

[38] Norbert Polacek and Alexander S Mankin. The ribosomal peptidyl transferase center: structure, function, evolution, inhibition. *Crit Rev Biochem Mol Biol*, 40(5):285–311, 2005.

[39] F J LaRiviere, A D Wolfson, and O C Uhlenbeck. Uniform binding of aminoacyl-trnas to elongation factor tu by thermodynamic compensation. *Science*, 294(5540):165–8, Oct 2001.

[40] E Scolnick, R Tompkins, T Caskey, and M Nirenberg. Release factors differing in specificity for terminator codons. *Proc Natl Acad Sci U S A*, 61(2):768–74, Oct 1968.

[41] Tommy Alain, Masahiro Morita, Bruno D Fonseca, Akiko Yanagiya, Nadeem Siddiqui, Mamatha Bhat, Domenick Zammit, Victoria Marcus, Peter Metrakos, Lucie-Anne Voyer, Valentina Gandin, Yi Liu, Ivan Topisirovic, and Nahum Sonenberg. eif4e/4e-bp ratio predicts the efficacy of mtor targeted therapies. *Cancer Res*, 72(24):6468–76, Dec 2012.

[42] A G Rowlands, R Panniers, and E C Henshaw. The catalytic mechanism of guanine nucleotide exchange factor action and competitive inhibition by phosphorylated eukaryotic initiation factor 2. *J Biol Chem*, 263(12):5526–33, Apr 1988.

[43] C U Hellen and P Sarnow. Internal ribosome entry sites in eukaryotic mrna molecules. *Genes Dev*, 15(13):1593–612, Jul 2001.

[44] Fátima Gebauer and Matthias W Hentze. Molecular mechanisms of translational control. *Nat Rev Mol Cell Biol*, 5(10):827–35, Oct 2004.

[45] Michael L Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, Jan 2010.

[46] Fatih Ozsolak and Patrice M Milos. Rna sequencing: advances, challenges and opportunities. *Nat Rev Genet*, 12(2):87–98, Feb 2011.

[47] Yongjun Chu and David R Corey. Rna sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther*, 22(4):271–4, Aug 2012.

[48] Samuel Marguerat and Jürg Bähler. Rna-seq: from technology to biology. *Cell Mol Life Sci*, 67(4):569–79, Feb 2010.

[49] Jonathan M Rothberg and John H Leamon. The development and impact of 454 sequencing. *Nat Biotechnol*, 26(10):1117–24, Oct 2008.

[50] Jingyue Ju, Dae Hyun Kim, Lanrong Bi, Qinglin Meng, Xiaopeng Bai, Zengmin Li, Xiaoxu Li, Mong Sano Marma, Shundi Shi, Jian Wu, John R Edwards, Aireen Romu, and Nicholas J Turro. Four-color dna sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A*, 103(52):19635–40, Dec 2006.

[51] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden,

Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–8, Jan 2009.

[52] Anton Valouev, Jeffrey Ichikawa, Thaisan Tonthat, Jeremy Stuart, Swati Ranade, Heather Peckham, Kathy Zeng, Joel A Malek, Gina Costa, Kevin McKernan, Arend Sidow, Andrew Fire, and Steven M Johnson. A high-resolution, nucleosome position map of c. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res*, 18(7):1051–63, Jul 2008.

[53] Zhoutao Chen and Xiaoping Duan. Ribosomal rna depletion for massively parallel bacterial rna-sequencing applications. *Methods Mol Biol*, 733:93–103, 2011.

[54] Shaomei He, Omri Wurtzel, Kanwar Singh, Jeff L Froula, Suzan Yilmaz, Susannah G Tringe, Zhong Wang, Feng Chen, Erika A Lindquist, Rotem Sorek, and Philip Hugenholtz. Validation of two ribosomal rna removal methods for microbial metatranscriptomics. *Nat Methods*, 7(10):807–12, Oct 2010.

[55] Jeffrey A Martin and Zhong Wang. Next-generation transcriptome assembly. *Nat Rev Genet*, 12(10):671–82, Oct 2011.

[56] Nicholas T Ingolia, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–23, Apr 2009.

[57] Nicholas T Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet*, 15(3):205–13, Mar 2014.

[58] Nicholas T Ingolia, Gloria A Brar, Silvia Rouskin, Anna M McGeachy, and Jonathan S Weissman. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mrna fragments. *Nat Protoc*, 7(8):1534–50, Aug 2012.

[59] Tilman Schneider-Poetsch, Jianhua Ju, Daniel E Eyler, Yongjun Dang, Shridhar Bhat, William C Merrick, Rachel Green, Ben Shen, and Jun O Liu. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat Chem Biol*, 6(3):209–217, Mar 2010.

[60] J Gregory Caporaso, Christian L Lauber, William A Walters, Donna Berg-Lyons, James Huntley, Noah Fierer, Sarah M Owens, Jason Betley, Louise Fraser, Markus Bauer, Niall Gormley, Jack A Gilbert, Geoff Smith, and Rob Knight. Ultra-high-throughput microbial community analysis on the illumina hiseq and miseq platforms. *ISME J*, 6(8):1621–4, Aug 2012.

[61] Gloria A Brar and Jonathan S Weissman. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol*, 16(11):651–64, Nov 2015.

[62] Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, Nov 2011.

[63] Carson C Thoreen, Lynne Chantranupong, Heather R Keys, Tim Wang, Nathanael S Gray, and David M Sabatini. A unifying model for mtorc1-mediated regulation of mrna translation. *Nature*, 485(7396):109–13, May 2012.

[64] Andrew L Wolfe, Kamini Singh, Yi Zhong, Philipp Drewe, Vinagolu K Rajasekhar, Viraj R Sanghvi, Konstantinos J Mavrakis, Man Jiang, Justine E Roderick, Joni Van der Meulen, Jonathan H Schatz, Christina M Rodrigo, Chunying Zhao, Pieter Rondou, Elisa de Stanchina, Julie Teruya-Feldstein, Michelle A Kelliher, Frank Speleman, John A Porco, Jr, Jerry Pelletier, Gunnar Rätsch, and Hans-Guido Wendel. Rna g-quadruplexes cause eif4a-dependent oncogene translation in cancer. *Nature*, 513(7516):65–70, Sep 2014.

[65] Xiangwei Gao, Ji Wan, Botao Liu, Ming Ma, Ben Shen, and Shu-Bing Qian. Quantitative profiling of initiating ribosomes in vivo. *Nat Methods*, 12(2):147–53, Feb 2015.

[66] X Gu, Y X Fu, and W H Li. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol*, 12(4):546–57, Jul 1995.

[67] Z Yang. Statistical properties of a dna sample under the finite-sites model. *Genetics*, 144(4):1941–50, Dec 1996.

[68] Ben Langmead, Kasper D Hansen, and Jeffrey T Leek. Cloud-scale rna-sequencing differential expression analysis with myrna. *Genome Biol*, 11(8):R83, 2010.

[69] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic Acids Res*, 40(10):4288–97, May 2012.

[70] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from rna-seq data. *Genome Res*, 22(10):2008–17, Oct 2012.

[71] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–8, Jul 2008.

[72] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[73] Tobias Maier, Marc Güell, and Luis Serrano. Correlation of mrna and protein in complex biological samples. *FEBS Lett*, 583(24):3966–73, Dec 2009.

[74] Keith A Spriggs, Martin Bushell, and Anne E Willis. Translational regulation of gene expression during conditions of cell stress. *Mol Cell*, 40(2):228–37, Oct 2010.

[75] Nahum Sonenberg and Alan G Hinnebusch. New modes of translational control in development, behavior, and disease. *Mol Cell*, 28(5):721–9, Dec 2007.

[76] Ciriaco A Piccirillo, Eva Bjur, Ivan Topisirovic, Nahum Sonenberg, and Ola Larsson. Translational control of immune responses: from transcripts to translatomes. *Nat Immunol*, 15(6):503–11, Jun 2014.

[77] O Donzé, R Jagus, A E Koromilas, J W Hershey, and N Sonenberg. Abrogation of translation initiation factor eif-2 phosphorylation causes malignant transformation of nih 3t3 cells. *EMBO J*, 14(15):3828–34, Aug 1995.

[78] Maria Barna, Aya Pusic, Ornella Zollo, Maria Costa, Nadya Kondrashov, Eduardo Rego, Pulivarthi H Rao, and Davide Ruggero. Suppression of myc oncogenic activity by ribosomal protein haploinsufficiency. *Nature*, 456(7224):971–5, Dec 2008.

[79] Denis M Schewe and Julio A Aguirre-Ghiso. Inhibition of eif2alpha dephosphorylation maximizes bortezomib efficiency and eliminates quiescent multiple myeloma cells surviving proteasome inhibitor therapy. *Cancer Res*, 69(4):1545–52, Feb 2009.

[80] Deborah Silvera, Silvia C Formenti, and Robert J Schneider. Translational control in cancer. *Nat Rev Cancer*, 10(4):254–66, Apr 2010.

[81] Mitchell Guttman, Pamela Russell, Nicholas T Ingolia, Jonathan S Weissman, and Eric S Lander. Ribosome profiling provides evidence that large noncoding rnas do not encode proteins. *Cell*, 154(1):240–51, Jul 2013.

[82] Adam B Olshen, Andrew C Hsieh, Craig R Stumpf, Richard A Olshen, Davide Ruggero, and Barry S Taylor. Assessing gene-level translational control from ribosome profiling. *Bioinformatics*, 29(23):2995–3002, Dec 2013.

[83] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, Florence Jaffrézic, and French StatOmique Consortium. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Brief Bioinform*, 14(6):671–83, Nov 2013.

[84] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11(3):R25, 2010.

[85] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[86] Philipp Drewe, Oliver Stegle, Lisa Hartmann, André Kahles, Regina Bohnert, Andreas Wachter, Karsten Borgwardt, and Gunnar Rätsch. Accurate detec-

tion of differential rna processing. *Nucleic Acids Res*, 41(10):5189–98, May 2013.

[87] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, Jan 2010.

[88] P. McCullagh and John A. Nelder. *Generalized Linear Models, Second Edition*. Chapman and Hall, 1989.

[89] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–32, Apr 2008.

[90] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 2014.

[91] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14(2):232–43, Apr 2013.

[92] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, Mar 1961.

[93] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.

[94] Christian Gonzalez, Jennifer S Sims, Nicholas Hornstein, Angeliki Mela, Franklin Garcia, Liang Lei, David A Gass, Benjamin Amendolara, Jeffrey N Bruce, Peter Canoll, and Peter A Sims. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J Neurosci*, 34(33):10924–36, Aug 2014.

[95] Simon Andrews. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[96] Nathan L Clement, Quinn Snell, Mark J Clement, Peter C Hollenhorst, Jahnvi Purwar, Barbara J Graves, Bradley R Cairns, and W Evan Johnson. The

gnumap algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*, 26(1):38–45, Jan 2010.

[97] Suying Bao, Rui Jiang, WingKeung Kwan, BinBin Wang, Xu Ma, and You-Qiang Song. Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet*, 56(6):406–14, Jun 2011.

[98] Nuno A Fonseca, Johan Rung, Alvis Brazma, and John C Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–77, Dec 2012.

[99] Malachi Griffith, Jason R Walker, Nicholas C Spies, Benjamin J Ainscough, and Obi L Griffith. Informatics for rna sequencing: A web resource for analysis on the cloud. *PLoS Comput Biol*, 11(8):e1004393, Aug 2015.

[100] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, 41(Database issue):D590–6, Jan 2013.

[101] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 43(Database issue):D6–17, Jan 2015.

[102] Fiona Cunningham, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E Hunt, Sophie H Janacek, Nathan Johnson, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Fergal J Martin, Thomas Maurel, William McLaren, Daniel N Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P Wilder, Amonida Zadissa, Bronwen L Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M J Searle, Giulietta Spudich, Stephen J Trevanion, Andy Yates, Daniel R Zerbino, and Paul Flicek. Ensembl 2015. *Nucleic Acids Res*, 43(Database issue):D662–9, Jan 2015.

[103] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[104] The SAM/BAM Format Specification Working Group. http://samtools.github.io/hts-specs/samv1.pdf.

[105] Generic Model Organism Database project. http://gmod.org/wiki/gff2/.

[106] Lindsay I Smith. http://www.cs.otago.ac.nz/cosc453/.

[107] Sebastian Raschka. http://sebastianraschka.com/articles/.

[108] Davide Ruggero. Translational control in cancer etiology. *Cold Spring Harb Perspect Biol*, 5(2), Feb 2013.

[109] Y Mamane, E Petroulakis, O LeBacquer, and N Sonenberg. mtor, translation initiation and cancer. *Oncogene*, 25(48):6416–22, Oct 2006.

[110] A Lazaris-Karatzas, K S Montine, and N Sonenberg. Malignant transformation by a eukaryotic initiation factor subunit that binds to mrna 5' cap. *Nature*, 345(6275):544–7, Jun 1990.

[111] D Rousseau, A C Gingras, A Pause, and N Sonenberg. The eif4e-binding proteins 1 and 2 are negative regulators of cell growth. *Oncogene*, 13(11):2415–20, Dec 1996.

[112] Svetlana Avdulov, Shunan Li, Van Michalek, David Burrichter, Mark Peterson, David M Perlman, J Carlos Manivel, Nahum Sonenberg, Douglas Yee, Peter B Bitterman, and Vitaly A Polunovsky. Activation of translation complex eif4f is essential for the genesis and maintenance of the malignant phenotype in human mammary epithelial cells. *Cancer Cell*, 5(6):553–63, Jun 2004.

[113] Hong Y Zhou and Alice S T Wong. Activation of p70s6k induces expression of matrix metalloproteinase 9 associated with hepatocyte growth factor-mediated invasion in human ovarian cancer cells. *Endocrinology*, 147(5):2557–66, May 2006.

[114] C Bauer, I Diesinger, N Brass, H Steinhart, H Iro, and E U Meese. Translation initiation factor eif-4g is immunogenic, overexpressed, and amplified in patients with squamous cell lung carcinoma. *Cancer*, 92(4):822–9, Aug 2001.

[115] N Brass, D Heckel, U Sahin, M Pfreundschuh, G W Sybrecht, and E Meese. Translation initiation factor eif-4gamma is encoded by an amplified gene and induces an immune response in squamous cell lung carcinoma. *Hum Mol Genet*, 6(1):33–9, Jan 1997.

[116] Hiroyuki Okamoto, Kohichiroh Yasui, Chen Zhao, Shigeki Arii, and Johji Inazawa. Ptk2 and eif3s3 genes may be amplification targets at 8q23-q24 and are associated with large hepatocellular carcinomas. *Hepatology*, 38(5):1242–9, Nov 2003.

[117] C S Pramesh and R C Mistry. Surgical treatment for cancer of the oesophagus and gastric cardia in hebei, china (br j surg 2004; 91: 90-98). *Br J Surg*, 91(4):511, Apr 2004.

[118] Teresa Palomero, Maria Luisa Sulis, Maria Cortina, Pedro J Real, Kelly Barnes, Maria Ciofani, Esther Caparros, Jean Buteau, Kristy Brown, Sherrie L Perkins, Govind Bhagat, Archana M Agarwal, Giuseppe Basso, Mireia Castillo, Satoru Nagase, Carlos Cordon-Cardo, Ramon Parsons, Juan Carlos Zúñiga-Pflücker, Maria Dominguez, and Adolfo A Ferrando. Mutational loss of pten induces resistance to notch1 inhibition in t-cell leukemia. *Nat Med*, 13(10):1203–10, Oct 2007.

[119] Andrew P Weng, John M Millholland, Yumi Yashiro-Ohtani, Marie Laure Arcangeli, Arthur Lau, Carol Wai, Cristina Del Bianco, Carlos G Rodriguez, Hong Sai, John Tobias, Yueming Li, Michael S Wolfe, Cathy Shachaf, Dean Felsher, Stephen C Blacklow, Warren S Pear, and Jon C Aster. c-myc is an important direct target of notch1 in t-cell acute lymphoblastic leukemia/lymphoma. *Genes Dev*, 20(15):2096–109, Aug 2006.

[120] Priscila P Zenatti, Daniel Ribeiro, Wenqing Li, Linda Zuurbier, Milene C Silva, Maddalena Paganin, Julia Tritapoe, Julie A Hixon, André B Silveira, Bruno A Cardoso, Leonor M Sarmento, Nádia Correia, Maria L Toribio, Jörg

Kobarg, Martin Horstmann, Rob Pieters, Silvia R Brandalise, Adolfo A Ferrando, Jules P Meijerink, Scott K Durum, J Andrés Yunes, and João T Barata. Oncogenic il7r gain-of-function mutations in childhood t-cell acute lymphoblastic leukemia. *Nat Genet*, 43(10):932–9, Oct 2011.

[121] Na Liu, Jingru Zhang, and Chunyan Ji. The emerging roles of notch signaling in leukemia and stem cells. *Biomark Res*, 1(1):23, 2013.

[122] Konstantinos J Mavrakis, Joni Van Der Meulen, Andrew L Wolfe, Xiaoping Liu, Evelien Mets, Tom Taghon, Aly A Khan, Manu Setty, Manu Setti, Pieter Rondou, Peter Vandenberghe, Eric Delabesse, Yves Benoit, Nicholas B Socci, Christina S Leslie, Pieter Van Vlierberghe, Frank Speleman, and Hans-Guido Wendel. A cooperative microrna-tumor suppressor gene network in acute t-cell lymphoblastic leukemia (t-all). *Nat Genet*, 43(7):673–8, Jul 2011.

[123] N Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50(3):163–70, Mar 1966.

[124] Liwei Rong, Mark Livingstone, Rami Sukarieh, Emmanuel Petroulakis, Anne-Claude Gingras, Katherine Crosby, Bradley Smith, Roberto D Polakiewicz, Jerry Pelletier, Maria A Ferraiuolo, and Nahum Sonenberg. Control of eif4e cellular localization by eif4e-binding proteins, 4e-bps. *RNA*, 14(7):1318–27, Jul 2008.

[125] Marie-Eve Bordeleau, Francis Robert, Baudouin Gerard, Lisa Lindqvist, Samuel M H Chen, Hans-Guido Wendel, Brigitte Brem, Harald Greger, Scott W Lowe, John A Porco, Jr, and Jerry Pelletier. Therapeutic suppression of translation initiation modulates chemosensitivity in a mouse lymphoma model. *J Clin Invest*, 118(7):2651–60, Jul 2008.

[126] Christina M Rodrigo, Regina Cencic, Stéphane P Roche, Jerry Pelletier, and John A Porco. Synthesis of rocaglamide hydroxamates and related compounds as eukaryotic translation inhibitors: synthetic and biological studies. *J Med Chem*, 55(1):558–62, Jan 2012.

[127] Regina Cencic, Marilyn Carrier, Gabriela Galicia-Vázquez, Marie-Eve Bordeleau, Rami Sukarieh, Annie Bourdeau, Brigitte Brem, Jose G Teodoro, Harald

Greger, Michel L Tremblay, John A Porco, Jr, and Jerry Pelletier. Antitumor activity and mechanism of action of the cyclopenta[b]benzofuran, silvestrol. *PLoS One*, 4(4):e5223, 2009.

[128] Jonathan H Schatz, Elisa Oricchio, Andrew L Wolfe, Man Jiang, Irina Linkov, Jocelyn Maragulia, Weiji Shi, Zhigang Zhang, Vinagolu K Rajasekhar, Nen C Pagano, John A Porco, Jr, Julie Teruya-Feldstein, Neal Rosen, Andrew D Zelenetz, Jerry Pelletier, and Hans-Guido Wendel. Targeting cap-dependent translation blocks converging survival signals by akt and pim kinases in lymphoma. *J Exp Med*, 208(9):1799–807, Aug 2011.

[129] Géraldine Jean, André Kahles, Vipin T Sreedharan, Fabio De Bona, and Gunnar Rätsch. Rna-seq read alignments with palmapper. *Curr Protoc Bioinformatics*, Chapter 11:Unit 11.6, Dec 2010.

[130] Paul Flicek, Ikhlak Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Carlos García-Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Monika Komorowska, Eugene Kulesha, Ian Longden, Thomas Maurel, William M McLaren, Matthieu Muffato, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet Singh Riat, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sheppard, Daniel Sobral, Kieron Taylor, Anja Thormann, Stephen Trevanion, Simon White, Steven P Wilder, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Nathan Johnson, Rhoda Kinsella, Anne Parker, Giulietta Spudich, Andy Yates, Amonida Zadissa, and Stephen M J Searle. Ensembl 2013. *Nucleic Acids Res*, 41(Database issue):D48–55, Jan 2013.

[131] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–57, Jul 2006.

[132] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf,

and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, Mar 2012.

[133] Nissim Hay and Nahum Sonenberg. Upstream and downstream of mtor. *Genes Dev*, 18(16):1926–45, Aug 2004.

[134] O Meyuhas. Synthesis of the translational apparatus is regulated at the translational level. *Eur J Biochem*, 267(21):6321–30, Nov 2000.

[135] J Pelletier and N Sonenberg. Internal initiation of translation of eukaryotic mrna directed by a sequence derived from poliovirus rna. *Nature*, 334(6180):320–5, Jul 1988.

[136] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12):1009–15, Dec 2010.

[137] Timothy L Bailey. Dreme: motif discovery in transcription factor chip-seq data. *Bioinformatics*, 27(12):1653–9, Jun 2011.

[138] Ivo L Hofacker. Vienna rna secondary structure server. *Nucleic Acids Res*, 31(13):3429–31, Jul 2003.

[139] Anthony Bugaut and Shankar Balasubramanian. 5'-utr rna g-quadruplexes: translation regulation and targeting. *Nucleic Acids Res*, 40(11):4727–41, Jun 2012.

[140] E P Booy, M Meier, N Okun, S K Novakowski, S Xiong, J Stetefeld, and S A McKenna. The rna helicase rhau (dhx36) unwinds a g4-quadruplex in human telomerase rna and promotes the formation of the p1 helix template boundary. *Nucleic Acids Res*, 40(9):4110–24, May 2012.

[141] Prasun Chakraborty and Frank Grosse. Human dhx9 helicase preferentially unwinds rna-containing displacement loops (r-loops) and g-quadruplexes. *DNA Repair (Amst)*, 10(6):654–65, Jun 2011.

[142] Sunita Kumari, Anthony Bugaut, Julian L Huppert, and Shankar Balasubramanian. An rna g-quadruplex in the 5' utr of the nras proto-oncogene modulates translation. *Nat Chem Biol*, 3(4):218–21, Apr 2007.

[143] Mark J Morris, Yoichi Negishi, Cathy Pazsint, Joseph D Schonhoft, and Soumitra Basu. An rna g-quadruplex is essential for cap-independent translation initiation in human vegf ires. *J Am Chem Soc*, 132(50):17831–9, Dec 2010.

[144] Ramla Shahid, Anthony Bugaut, and Shankar Balasubramanian. The bcl-2 5' untranslated region contains an rna g-quadruplex-forming motif that modulates protein expression. *Biochemistry*, 49(38):8300–6, Sep 2010.

[145] Julian Leon Huppert, Anthony Bugaut, Sunita Kumari, and Shankar Balasubramanian. G-quadruplexes: the beginning and end of utrs. *Nucleic Acids Res*, 36(19):6260–8, Nov 2008.

[146] Arianna Sabò, Theresia R Kress, Mattia Pelizzola, Stefano de Pretis, Marcin M Gorski, Alessandra Tesi, Marco J Morelli, Pranami Bora, Mirko Doni, Alessandro Verrecchia, Claudia Tonelli, Giovanni Fagà, Valerio Bianchi, Alberto Ronchi, Diana Low, Heiko Müller, Ernesto Guccione, Stefano Campaner, and Bruno Amati. Selective transcriptional regulation by myc in cellular growth control and lymphomagenesis. *Nature*, 511(7510):488–92, Jul 2014.

[147] Jan van Riggelen, Alper Yetil, and Dean W Felsher. Myc as a regulator of ribosome biogenesis and protein synthesis. *Nat Rev Cancer*, 10(4):301–9, Apr 2010.

Appendix

RiboDiff Manual

This document shows how to use *RiboDiff* to detect the protein translational efficiency change from ribosome footprint profile (Ribo-Seq) and RNA-Seq data in different experimental conditions.

To run *RiboDiff*, please provide two input files:

1) A comma delimited text file (CSV) that describes the experimental design. Here is an example:

```
Samples,DataType,Conditions
RbCtlR1,Ribo-Seq,Control
RbCtlR2,Ribo-Seq,Control
RbTrtR1,Ribo-Seq,DrugTreated
RbTrtR2,Ribo-Seq,DrugTreated
RnaCtlR1,RNA-Seq,Control
RnaCtlR2,RNA-Seq,Control
RnaCtlR3,RNA-Seq,Control
RnaTrtR1,RNA-Seq,DrugTreated
RnaTrtR2,RNA-Seq,DrugTreated
RnaTrtR3,RNA-Seq,DrugTreated
```

In this experimental design file, there are three columns organized underneath a header line. The first column indicates the sample name. The second column indicates the data type. It only accepts “Ribo-Seq” and “RNA-Seq” to represent the ribosome footprint and RNA-Seq data, respectively. The last column indicates to which condition the sample belongs. More than two condition keywords are not accepted here. Note the columns are comma separated.

2) A text file containing the raw read count for each gene as follows:

Entry	RbCtlR1	RbCtlR2	RbTrtR1	RbTrtR2	RnaCtlR1	RnaCtlR2	RnaTrtR1	RnaTrtR2
G0001	69	77	81	98	2914	2931	2520	2566
G0002	159	141	145	139	1285	1285	1242	1246
G0003	246	239	236	259	806	847	862	819
G0004	70	59	71	84	1413	1490	1464	1499
...								

In this file the sample name in header line must agree with that in experimental design file. The columns are tab-separated. Only including protein coding genes may give a better global estimation.

Assuming you are at the *RiboDiff* package directory, the command to run *RiboDiff* is:

```
python ./scripts/TE.py -e <exp_outline.txt> -c <cnt_table.txt> -o <result.txt>
```

RiboDiff loads the input files and normalizes the raw count of each replicate sample according to its sequencing library size. Then, it estimates the dispersion parameter for each gene using a generalized linear model, assuming that read counts follow a negative binomial distribution. Note, users can enable or disable the different dispersion estimation for Ribo-Seq and RNA-Seq separately. Next, *RiboDiff* performs a statistical test based on the H_0 and H_1 model fitting. Finally, multiple test correction is performed to generate an adjusted P value for every gene.

RiboDiff also calculates the $\log_2$ fold change of translational efficiency for all genes. In addition, *RiboDiff* can plot figures to visualize the data and the statistical test results. The figures includes

- A scatter plot of dispersion against the mean count across replicates of Ribo-Seq and RNA-Seq.
- A scatter plot of $\log_2$ fold change of translational efficiency against the mean count across replicates of Ribo-Seq.
- A histogram of log fold change of translational efficiency with significant proportion marked in red (TE down) and blue (TE up).

The dispersion used in the fisrt scatter plot is the empirical dispersion. It is calculated as the relationship of $Var = mean + (mean \times disp^2)$

The help information for running *RiboDiff* is listed below. Users can display it by typing “python TE.py -h” in command line.

Options:		
-h	- -help	show this help message and exit.
Required:		
-e	ExptOutline	Text file describing experiment Outline. Must follow required format, please see the manual.
-c	CntFile	Text file containing the count data. Header line must be consistent with information in experiment Outline.
-o	OutFile	Tab delimited text file containing the results.
Optional:		
-d	DispDiff	Allow different dispersions for Ribo-seq and RNA-Seq count data. On: 1; Off: 0. [default: 1]
-s	SumCntCutoff	Set the sum of normalized read count as the threshold to do the test. This option applies for both Ribo-seq and RNA-Seq data. [default: 10]
-i	DispInitial	Set the initial dispersion to start the estimation. [default: 0.01]
-m	MultiTest	Method for multiple test correction. Options: BH (Benjamini-Hochberg); Bonferroni. [default: BH]
-r	RankResult	Rank the result table in ascending order by a specific column. Adjusted p value: 1; TE change: 2; Gene id: 3; Keep the order as in count file: 0. [default: 0]
-p	Plots	Make plots to show the data and results. Plots are in pdf format. On: 1; Off: 0. [default: 0]
-q	CutoffFDR	Set the FDR cutoff for significant case to plot. [default: 0.1]

Similar to many other RNA-Seq based tools, *RiboDiff* uses negative binomial distribution to model the read count, which handles larger variation across samples than Poisson. However, if the randomness of count data from certain types of samples are extremely large, limited number of replicates cannot provide a good estimation on dispersion, which ends up with less significant results.

Source Journals of the Thesis

- Yi Zhong, Theofanis Karaletsos, Philipp Drewe, *et al.* RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints. *Bioinformatics.* 2016 Sep 14. doi: 10.1093/bioinformatics/btw585

- Andrew Wolfe, Kamini Singh, Yi Zhong, *et al.* RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature.* 2014 Sep 4;513(7516):65-70.

- Xiaodi Su, Yingpu Yu, Yi Zhong, *et al.* Interferon-gamma regulates cellular metabolism and mRNA translation to potentiate macrophage activation. *Nature Immunology.* 2015 Aug;16(8):838-49.

Personal Contribution

- I developed the computational tool *RiboDiff*, including design the project and writing the codes. Theofanis Karaletsos and Philipp Drewe participated in this project by providing supports on math theory. Vipin Sreedharan helped to wrap the code. Gunnar Rätsch provided supervision. The manuscript of this work is preprinted in *bioRxiv* and published in *Bioinformatics*.

- I built the computational pipeline for analysis of the ribosome profiling and RNA-Seq data. This part of the work is incorporated into *RiboDiff* manuscript and published as software on github.com.

- I contributed to the collaborating project with Wendel lab where we found the protein translational control through G-quadruplex. I did most of the computational analysis with significant help from Philipp Drewe. This part of the work was supervised by Gunnar Rätsch. The work has been published in *Nature*.

Publications

- **Yi Zhong**, Theofanis Karaletsos, Philipp Drewe, Vipin Sreedharan, Kamini Singh, Hans-Guido Wendel, Gunnar Rätsch. RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints. *Bioinformatics.* 2016 Sep 14.
- Andrew Wolfe, Kamini Singh, **Yi Zhong**, Philipp Drewe, Vinagolu Rajasekhar, Viraj Sanghvi, Konstantinos Mavrakis, Man Jiang, Justine Roderick, Joni Van der Meulen, Jonathan Schatz, Christina Rodrigo, Chunying Zhao, Pieter Rondou, Elisa de Stanchina, Julie Teruya-Feldstein, Michelle Kelliher, Frank Speleman, John Porco, Jerry Pelletier, Gunnar Rätsch, Hans-Guido Wendel. RNA G-quadruplexes cause eIF4A-dependent oncogene translation in cancer. *Nature.* 2014 Sep 4;513(7516):65-70.
- **Yi Zhong**, Phillip Drewe, Andrew Wolfe, Kamini Singh, Hans-Guido Wendel, Gunnar Rätsch. Protein translational control and its contribution to oncogenesis revealed by computational methods. *BMC Bioinformatics* 2015, 16(Suppl 2):A6.
- Xiaodi Su, Yingpu Yu, **Yi Zhong**, Eugenia Giannopoulou, Xiaoyu Hu, Hui Liu, Justin Cross, Gunnar Rätsch, Charles Rice, Lionel Ivashkiv. Interferon-gamma regulates cellular metabolism and mRNA translation to potentiate macrophage activation. *Nature Immunology.* 2015 Aug;16(8):838-49.
- The Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell.* 2015, Nov 5;163(4):1011-25
- Jonas Behr, Andre Kahles, **Yi Zhong**, Vipin Sreedharan, Philipp Drewe, Gunnar Rätsch. MITIE: Simultaneous RNA-Seq-based Transcript Identification and Quantification in Multiple Samples. *Bioinformatics*, 29(20):2529-38, 2013.

- **Yi Zhong***, Xiao Chang*, Xing-Jun Cao*, Yan Zhang, Huajun Zheng, Yongzhang Zhu, Chengsong Cai, Yuan-Yuan Li, Guo-Ping, Zhao, Shengyue Wang, Yixue Li, Rong Zeng, Xuan Li, Xiao-Kui Guo. Comparative Proteogenomic Analysis of the Leptospira interrogans Virulence Attenuated Strain IPAV against the Pathogenic Strain 56601. *Cell Res*, 2011 Aug;21(8): 1210-29.
- Wei Zhao*, **Yi Zhong***, Hua Yuan*, Jin Wang*, Huajun Zheng, Ying Wang, Xufeng Cen, Feng Xu, Jie Bai, Xiaobiao Han, Gang Lu, Yongqiang Zhu, Zhihui Shao, Han Yan, Chen Li, Nanqiu Peng, Zilong Zhang, Yunyi Zhang, Wei Lin, Yun Fan, Zhongjun Qin, Yongfei Hu, Baoli Zhu, Shengyue Wang, Xiaoming Ding, Guo-Ping Zhao. Complete genome sequence of the rifamycin SV-producing Amycolatopsis mediterranei U32 revealed its genetic characteristics in phylogeny and metabolism. *Cell Res*, 2010 Oct;20(10):1096-108.
- Liang-dong Lu, Qing Sun, Xiao-yong Fan, **Yi Zhong**, Yu-feng Yao and Guo-Ping Zhao. Mycobacterial MazG is a novel NTP pyrophosphohydrolase involved in oxidative stress response. *J Biol Chem.* 2010 Sep 3;285(36):28076-85.
- **Yi Zhong**, Ke Dong, Yang Yang, Chunyan Feng, Picardeau M and Xiaokui Guo. Cloning, expression and preliminary characterization of a hemolysin gene tlyA from Leptospira biflexa serovar Patoc. *Chinese Journal of Zoonoses.* 2007 Issue 8, Page 739-744.

* Authors contributed equally to the work.