Freie Universität Berlin

Global survey of *cis*-regulation in mammalian translation

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

Submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Jingyi Hou

Submitted 15.03.2016

Time period of doctoral studies: October 2012 to February 2016

Supervisor: Prof. Dr. Wei Chen

Institution: Max-Delbrück-Center for Molecular Medicine (MDC)

1st Reviewer: Prof. Dr. Wei Chen
Berlin Institute for Medical Systems Biology (BIMSB)
Max-Delbrück-Center for Molecular Medicine (MDC)
Robert-Rössle-Strasse 10
13125 Berlin
Tel.: +49 (0) 30 9406 2995
Email: wei.chen@mdc-berlin.de

2nd Reviewer: Prof. Dr. Markus Wahl
Institute of Chemistry and Biochemistry – Biochemistry
Freie Universität Berlin
Takustrasse 6
14195 Berlin
Tel.: +49 (0) 30 8385 3456
Email: mwahl@chemie.fu-berlin.de

Date of defense: 13. 06. 2016

ACKNOWLEDGEMENT

A Doctorate of Philosophy, is combined with the Latin word “Docēre”, which is “to teach”, and “Philosophy”, which in Greek is “love of wisdom”. This section is dedicated to the people who have helped me obtain this special, valuable skill.

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. Wei Chen for his phenomenal supervision on my Ph.D study. Wei has created a very inspiring and creative research environment for me and never fails to be the most helpful supervisor one can wish for. Without Wei’s continuous support and immense idea and knowledge, I could not have imagined having finished my research projects and this thesis.

Besides my advisor, I would like to thank Dr. Xi Wang for his excellent and dedicated collaborations in the past three years. Xi developed the data analysis pipelines in my both projects and was also actively involved in all the experimental discussion. Without his immense contributions, both projects would not have been fulfilled. Xi is a very pleasant person to work with and the most fantastic computational expert a biologist could imagine to collaborate with.

My sincere thanks also goes to Claudia Quedenau, Wei Sun, Erik McShane, Dr. Henrik Zauber, Dr. Matthias Selbach for their valuable help in my projects. Without their support it would not be possible to conduct my research.

I thank my colleagues for their stimulating daily discussions and emotional support. I will also treasure all the happy moments we have shared together in the past four years. Special thanks also goes to Sabrina Deter for her immense administrative support and Claudia Quedenau for all her professional translation service.

I am also grateful to Prof. Dr. Markus Wahl for being my external advisor and taking his time in reviewing this thesis.

Last but not the least, I would like to thank my family in China as well as in Germany for their unconditional love, endless support and accompany throughout all stages of my Ph.D. Without you all is nothing.

STATEMENT OF CONTRIBUTIONS

This thesis consists of two individual studies. While the first part has already been published on the journal of Molecular Systems Biology, the second part is still under revision when this thesis is completed. The doctoral student is the first author on each of these researches and the presented results emerged directly from her doctoral studies.

Nevertheless, extensive collaboration with bioinformatics experts was key to the success of both studies. In the first study *"Cis*-regulatory Control of Translation in Hybrid Mice", the doctoral student did all the experiments, interpreted the data under supervision of Dr. Wei Chen and with the computational support from Dr. Xi Wang. In the second study "*Cis*-regulatory Impact of Transcript Leaders on Translation", the doctoral student developed the technology CAPTRE, designed and performed all related experiments, interpreted the data under supervision of Dr. Wei Chen and with the computational support from Dr. Xi Wang.

Computational analyses in the discussion section were performed by Dr. Xi Wang. All unpublished parts of this thesis as well as introduction and discussion sections are the sole work of the author and no other than indicated aid and sources have been used.

TABLE OF CONTENT

ABBREVIATIONS

4SU	4-thiouridine
5'TOP	5' Terminal Oligopyrimidine Tract
A	Adenine
AD	Alzheimer's Disease
ADTE	Allelic Divergence in Translational efficiency
APP	Amyloid-β Precursor Protein
ASE	Allele-Specific Expression
A-site	Acceptor Site
ATP	Adenosine Triphosphate
C	Cytosine
CAGE	Cap Analysis of Gene Expression
CAI	Codon Adaptation Index
Capture	CAp Profiling of TRanslational Efficiency
cDNA	Complementary DNA
CDS	Coding Sequence
CHX	Cycloheximide
CLIP	Cross-Linking and Immunoprecipitation
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide triphosphate
dsDNA	Double Stranded DNA
DTT	Dithiothreitol
eIF2-TC	eIF2-guanosine Triphosphate (GTP)/Met-transfer RNA$(tRNA)_i^{Met}$ Ternary Complex
eIF4A	Eukaryotic Initiation Factor 4A
eIF4E	Eukaryotic Initiation Factor 4E
eIF4G	Eukaryotic Initiation Factor 4G
eQTL	Expression Quantitative Trait Locus
eRF1	Eukaryotic Release Factor 1
eRF3	Eukaryotic Release Factor 3
E-site	Exit Site
FBS	Fetal Bovine Serum
FDR	False Discovery Rate
G	Guanine
GC rich	Guanine-Cytosine rich

HA	Hemagglutinin
hr	Hour
IRE	Iron Response Element
IRE-BP	IRE Binding Protein
IRES	Internal Ribosome Entry Site
ITAF	IRES *trans*-activating Factor
kb	Kilobase
Kcal	Kilocalorie
LTM	Lactimidomycin
M	Molar
m	Millie
m^6A	*N6*-Methyladenosine
m^7G	7-methylguanylate
MFE	Minimum Free Energy
mRNA	Messenger RNA
miRNA	MicroRNA
mRNP	Messenger Ribonucleoproteins
MS	Mass Spectrometry
mTOR	Mammalian Target of Rapamycin
n	Nano
NMD	Nonsense-Mediated Decay
NP-40	Nonyl Phenoxypolyethoxylethanol - 40
nt	Nucleotide
N-terminal	Amino-terminal
ORF	Open Reading Frame
PAR-CLIP	Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation
PCR	Polymerase Chain Reaction
PIC	Preinitiation Complex
Poly(A)	Polyadenylation
Pol II	RNA Polymerase II
pQTL	Protein Quantitative Trait Locus
pQTL	Protein-specific QTL
PTC	Premature Termination Codon
QTL	Quantitative Trait Locus
RACE	Rapid Amplification of cDNA Ends

RBP	RNA Binding Protein
RIP	RNA Immunoprecipitation
RIP-Chip	RNA immunoprecipitation and Microarray Analysis
RIP-Seq	RNA Immunoprecipitation and Sequencing
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic Acid
RNase	Ribonuclease
RNP	Ribonucleoprotein
RPF	Ribosome Protected Fragment
rRNA	Ribosomal RNA
rQTL	Ribosome Occupancy QTL
rsQTL	Ribosome-specific QTL
RT	Reverse Transcription
RT-PCR	Reverse Transcription Polymerase Chain Reaction
S	Svedberg, unit for sedimentation rate
SD Sequence	Shine-Dalgarno Sequence
SNP	Single Nucleotide Polymorphism
ssDNA	Single-Stranded DNA
T	Thymine
T4 PNK	T4 Polynucleotide Kinase
TE	Translational efficiency
TL	Transcript Leader
TIS	Translation Initiation Site
tRNA	Transfer RNA
TRAP	Translating Ribosome Affinity Purification
TSS	Transcription Start Site
UTR	Untranslated Region
uAUG	Upstream AUG
uORF	Upstream ORF
UV	Ultraviolet
µ	Micro

SUMMARY

This doctoral thesis consist of two parts: The first part describes a global survey of *cis*-regulatory divergence in mammalian translation, where I applied mRNA sequencing and deep sequencing-based polysome profiling to quantify translational efficiency in F1 hybrid mice. The F1 progeny between *Mus musculus* C57BL/6J and *Mus spretus* SPRET/EiJ was chosen as a model system because the two have the largest number of genetic variants among all mouse strains with high-quality genome assemblies available. This large genomic divergence 1) provides a large number of potential regulatory variants between the two strains and 2) enables a sequencing-based approach to distinguish allelic RNA transcripts. The high quality of the data was demonstrated by employing two independent validation approaches, PacBio full-length sequencing and ribosome profiling. In total, 1008 genes (14.1%) were identified exhibiting significant allelic difference in translational efficiency. Several sequence features were associated with the observed allelic divergence in translation, including local RNA secondary structure near the start codon and proximal out-of-frame upstream AUGs. Finally, *cis*-effects are quantitatively comparable between transcriptional and translational regulation and these effects are more frequently compensatory between the two processes, suggesting a role of the translational regulation in buffering transcriptional noise and thereby maintaining the robustness of protein expression.

In the second part, I developed novel technology CAPTRE to measure the translational status of distinct mRNA TL isoforms. In mouse fibroblasts, a total of 22,357 TSSs derived from 10,875 protein-coding genes were identified. Among 4153 genes expressing multiple TSSs, 745 exhibited significant TE difference between their alternative TL isoforms. Longer isoforms were more frequently associated with lower TE and the global impact of several regulatory elements was also revisited, such as uORFs, cap-adjacent stable RNA secondary structures as well as 5'-terminal oligopyrimidine tract. In addition, several novel sequence motifs that can affect translation activity were identified and their effect was validated using two reporter systems. Finally, quantitative models combining different features identified in this study explained approximately 60% of the variance of the TE difference observed between TL isoforms.

This study provides novel mechanistic insights into translational regulation and characterizes the potential coupling between translational and transcriptional regulation in mammalian cells.

ZUSAMMENFASSUNG

Diese Dissertation setzt sich aus zwei Teilen zusammen: Der erste Teil beschreibt eine globale Studie von *cis*-regulatorischen Divergenzen in der mRNA Translationseffizienz von Säugetierzellen. Hierzu habe ich Polyribosomen Profile erstellt und anschließend mRNA Sequenzierungstechnologien verwendet, um die Translationseffizienz in einem Maus F1-Hybridmodelystem zu bestimmen. Die F1-Nachkommen von *Mus musculus* C57BL/6J und *Mus spretus* SPRET/EiJ wurden hierzu als Modelsystem gewählt, da diese Spezies die größte Zahl an genetischen Variationen in allen Maus Modellen mit qualitativ hochwertigen Genomsequenzierdaten aufweist. Die hohe genomische Divergenz stellt 1) eine große Zahl an potentiell regulatorischen Varianten zwischen beiden Maus Arten dar und ermöglicht 2) eine allelspezifische Zuordnung von mRNA Transkripten durch Sequenzbestimmung. Die hohe Qualität der so gewonnenen Daten wurde mit zwei unabhängigen Methoden validiert: Sequenzbestimmung der mRNA in voller Läng mit Hilfe eines „PacBio“ Instruments, sowie Bestimmung von Translationsraten durch Erstellung von Ribosomen Fußabdrücken (sogenanntes „ribosome profiling“). Insgesamt konnten so 1008 Gene ermittelt werden (14.1%), die einen signifikanten Unterschied in der allelspezifischen Translationsrate aufweisen. Mehrere Sequenzeigenschaften konnten mit allelspezifischen divergenten Translationsraten assoziiert werden: Lokale RNA Sekundärstrukturen in der Nähe des Startcodons, sowie vorgelagerte AUG Startcodons außerhalb des offenen Leserahmens. Schließlich konnte gezeigt werden, dass *cis*-Effekte auf transkriptionaler sowie translationaler Ebene quantitativ vergleichbar sind und häufig eine kompensatorische Wirkung zwischen beiden Prozessen aufweisen. Dies suggeriert eine Puffer-ähnliche Rolle der Translation, wodurch Schwankungen in Transkriptionsraten kompensiert werden können, was wiederum robuste Genexpressionsmuster gewährleistet.

In dem zweiten Teil dieser Arbeit habe ich eine neuartige Technologie namens CAPTRE entwickelt um den Translationsstatus von mRNA Isoformen mit unterschiedlichen vorgelagerten nicht-translatierten Sequenzbereichen zu messen. In Maus Fibroblasten wurde zunächst eine Gesamtzahl von 22.357 Transkriptionsstartpositionen von 10.875 Genen ermittelt. Von 4153 Genen, die alternative Transkriptionsstartpositionen nutzen, zeigten 745 signifikante Unterschiede zwischen Isoformen mit alternativen vorgelagerten nicht-translatierten Sequenzen. Hiervon zeigten längere Isoformen häufig eine Assoziation mit niedrigeren Translationsraten. Weiterhin wurde der globale Einfluss mehrerer regulatorischer Elemente, wie beispielsweise vorgelagerter offener Leserahmen (sogenannte „uORFs“),

stabiler RNA Sekundärstrukturen in der Nähe des Transkriptanfangs, sowie terminalen Oligopyrimidin Sequenzen untersucht. Darüberhinaus wurden mehrere neue Sequenzmotive, welche die Translation beeinflussen können identifiziert und deren Einfluss auf Translationsraten mit Hilfe von zwei unterschiedlichen Reportersystemen validiert. Schließlich wurden quantitative Computermodelle entwickelt um die in dieser Studie gefundenen regulatorischen Elemente zu beschreiben. Unter Verwendung dieser Modelle konnten 60% der beobachteten Varianz in Translationsraten zwischen verschiedenen Isoformen, welche alternative vorgelagerte nicht-translatierte Sequenzen aufweisen, erklärt werden.

Zusammenfassend konnten in dieser Studie wichtige neue mechanistische Erkenntnisse hinsichtlich der translationalen Genregulation gewonnen werden. Insbesondere konnte auf eine mögliche regulatorische Kopplung zwischen transkriptionaler und translationaler Regulation hingewiesen werden.

1. INTRODUCTION

1.1 Translational regulation in mammalian cells

Gene expression in eukaryotes is a complex process orchestrated by multiple regulatory steps, including chromatin remodeling, mRNA transcription, pre-mRNA splicing, mRNA export, localization, mRNA decay, translation, post-translational modification and protein decay. For decades, mRNA abundance levels were widely used as a proxy of protein expression, yet, in various eukaryotes from yeast to human, only approximately 50% of the variation in protein level can be explained by variation in mRNA abundance (de Sousa Abreu *et al*, 2009). Translational regulation of existing mRNAs plays a crucial role in determining cellular protein concentration dynamics, resulting in not only long-term adjustment in cell physiology, but also rapid control of protein changes during conditions of stress (reviewed in Sonenberg and Hinnebusch, 2009; reviewed in Spriggs *et al*, 2010). Recent genome-wide studies further highlight the predominant role of translation in controlling cellular protein concentrations, in both yeast and mammalian cells (Schwanhäusser et al, 2011; Marguerat et al, 2012). In addition, translational dysregulation frequently leads to pathogenic physiology in many human diseases (Cazzola & Skoda, 2000; Reynolds, 2002).

1.2 Eukaryotic translation initiation

Mammalian translational regulation is composed of three stages: initiation, elongation and termination. It is in general agreed that translation initiation is the rate-limiting step and the majority of known regulatory processes occur at this stage. In the canonical model of cap-dependent translation, a 43S preinitiation complex (PIC), which contains the small (40S) ribosomal subunit, methionine initiator tRNA, an eIF2-guanosine triphosphate (GTP)/Met-transfer RNA $(tRNA)_i^{Met}$ ternary complex (eIF2-TC), initiation factors eIF1,1A, 3, is first recruited to the mRNA 5'- end cap structure (the m^7G cap) by a complex of cap-binding protein eIF4E, a large scaffold protein eIF4G, and an ATP-dependent RNA helicase eIF4A. The 43S complex then scans the entire transcript leader sequence (historically named as 5' untranslated region) in a 5' to 3' direction in search for an AUG codon (sometimes near-cognate AUG) by the anticodon of the initiator tRNA. The large (60S) ribosomal subunit is then joined to assemble the elongation-competent 80S ribosome, and translation is initiated. Enormous progress in the last several decades has demonstrated that general or gene-specific

control of translation initiation can take place at numerous regulatory points, including modulating the recruitment of 43S PIC to the 5'cap, scanning of the 43S small ribosome, joining of ribosomal subunits and selection of start codons (Sonenberg & Hinnebusch, 2009). Under certain conditions, translation can also start in a cap-independent manner through a set of specialized elements referred to as internal ribosome entry sites (discussed in section 1.4.7).

1.3 Eukaryotic translation elongation and termination

At the end of translation initiation step, an 80S ribosome is positioned on an mRNA with the anticodon of Met-tRNAi in its P (peptidyl)-site base-paired with the start codon AUG. The second codon occupies the A (acceptor)-site of the ribosome to receive the cognate aminoacyl-tRNA. During the chain elongation, each additional amino acid is added to the nascent polypeptide chain in a three-step cycle: positioning the cognate aminoacyl-tRNA in the A-site of the ribosome, forming the peptide bond and then shifting the mRNA by one codon relative to the ribosome.

Translation can be influenced during the elongation step by ribosomal pausing, which triggers endonucleolytic attack of the mRNA in the vicinity of the stalled ribosome, a process termed mRNA no-go decay (Harigaya & Parker, 2010). Ribosomal pausing can also help co-translational folding of the nascent polypeptide on the ribosome, and delays protein translation while it is decoding mRNA (Buchan & Stansfield, 2007).

Translation termination occurs when a stop codon (UAG, UAA or UGA) is reached by the ribosome entering the A-site. Unlike initiation, translation termination in eukaryotes is not dependent on tRNAs, instead, is catalyzed by two protein factors, eRF1 and eRF3. eRF1 is responsible for high-fidelity stop codon recognition and promotes the hydrolysis of the ester bond linking the peptide chain with the peptidyl-tRNA. eRF3 is a ribosome-dependent guanosine triphosphatase that helps eRF1 to release the completed polypeptide during translation termination (reviewed in Dever & Green, 2012).

1.4 *Cis*-regulatory elements in translational regulation

Gene expression regulation, including translational regulation, is frequently mediated through the interaction of *cis*- and *trans*-acting elements, which are classified depending on how they exert their effect on their target genes. Factors with *trans*-regulatory effects are located

anywhere in the genome, like miRNAs and RNA binding proteins. *Cis*-regulatory elements tend to be located close to the gene they affect, in the case of translational regulation, located within mRNA sequences.

A typical mature mRNA in eukaryotes is structured with 5' untranslated region (5' UTR), followed by a coding region and a 3' untranslated region (3' UTR). Although in lack of protein-coding capacity, UTRs play crucial roles in regulating gene expression by providing structural basis for post-transcriptional control, including modulating mRNA localization (Jansen, 2001), half-life (Bashirullah *et al*, 2001) and translation (Sonenberg, 1994) . The average length of 5'UTRs of human mRNA is around 200 nucleotides, and 3'UTRs is approximately 800 nucleotides (Chatterjee & Pal, 2009). In the mammalian genome, the length of both 3' and 5' UTRs vary considerably among genes as well as among transcript isoforms, thus providing a great diversity of regulation through *cis*-regulatory elements, including RNA secondary structures, upstream open reading frames (uORFs), upstream AUGs (uAUGs), binding sites for RNA binding proteins and microRNAs (miRNAs) or internal ribosome entry sites (IRESs) (Fig.1.1). All these features control the protein synthesis by employing a single or combinatorial mechanism. In the following sections, I will review how *cis*-regulatory elements in UTR regions influence translation in mammals.

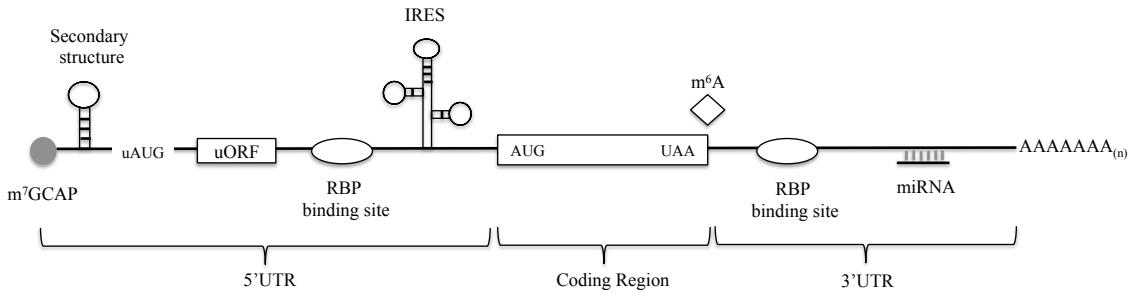

Fig 1.1 Overview on *cis*-regulatory mRNA elements that control translation. uORF : upstream ORF, RBP:RNA binding proteins, IRES: internal ribosome entry site, miRNA : microRNA, m^6A:*N6*-mehyladenosine

1.4.1 Kozak sequence

The optimal sequence context at translation initiation sites contains a characteristic stretch of bases commonly referred to as Kozak sequence (GCC**A/G**CCaugG) (Kozak, 1991b). This conserved sequence element is exquisitely sensitive to nucleotide mutations that lead to changes in translational efficiency. A purine to pyrimidine change at the −3 position (usually

an adenine (A)), as well as mutation of the guanine (G) at +4 position can lead to decreased efficiency of translation initiation (Kozak, 1986a; Dvir *et al*, 2013). Start codon flanked by the most favorable Kozak sequence with A -3 and G +4 (GG**A**TTaug**G**) can be translates over 20-fold more efficiently compared to the weakest sequence element (GG**T**TTaug**T**) (Kozak, 1986b). This striking sequence preference for an A at -3 position and a G at +4 position is highly conserved among vertebrates, drosophila, plants and yeast (Kozak, 1991b). UV-crosslinking experiments have shown that the A at the -3 position can bind to eIF2α subunit and the G at the +4 position to helix 44 in the 18S rRNA. Both interactions were demonstrated to promote AUG recognition at the start sites (Pisarev *et al*, 2006).

In addition to modulating translational efficiency, Kozak context is also critical for the choice of translation initiation sites, especially for non-AUG start codons, CUG, GUG, UUG (Kozak, 1989a; Peabody, 1989). In the scanning model, 40S ribosomal subunit binds to the 5'cap structure, and migrates along the 5' UTR linearly in the 5' to 3' direction in search for the first start codon. However, translation does not exclusively occur at the first AUG codon. In certain cases, the 'first-rule' is replaced by alternative mechanisms. One of such mechanisms is known as 'leaky scanning'. When the 5' nearest AUG is not within the optimal Kozak context, the ribosome can bypass the first AUG and thus continues to search for a start codon further downstream (Kozak, 2002). In some cases, the non-AUG codons can served as alternative initiation sites when positioned upstream of canonical AUG codons, therefore ribosome initiates translation at both the non-canonical codon and canonical codons (Carroll & Derse, 1993; Fuxe *et al*, 2000; Chang & Wang, 2004).

1.4.2 Upstream open reading frames and upstream AUGs

Among *cis*-regulatory elements that control mRNA translation, upstream open reading frames (uORFs) are considered to be of particular importance. uORFs can be partitioned into two classes according to their relative position to the main ORFs: one class is completely upstream of the main ORFs and entirely located within transcript leader sequences, while the other class is overlapping with the coding sequence (CDS) and is lacking in-frame stop codons preceding the main ORFs. Here, I refer to the former class as uORFs, and the latter as uAUGs.

Upstream ORFs and AUGs are very prevalent in mammalian transcriptomes. Several bioinformatic analyses have estimated that 40-50% of the mammalian mRNAs harbor at least one uORF or uAUG (Matsui *et al*, 2007; Calvo, 2009). The presence of uORFs typically

reduces the protein abundance through facilitating mRNA degradation mediated by NMD pathway or more frequently, directly interfering with translation initiation by serving as "decoy" initiation sites that reduce the number of ribosomes initiating at the downstream authentic ORFs (Morris & Geballe, 2000; Somers *et al*, 2013). The effect size of upstream ORFs can be profound. A recent genome-wide analysis of protein and mRNA measurement showed uORFs and uAUGs typically reduce the protein level by 30-80% in several mammalian cells lines under steady state conditions, thus constituting a major regulatory element controlling mRNA translation (Calvo, 2009).

The mechanism by which uORFs and uAUGs regulate translation is not entirely understood. It appears that uORFs tend to repress translation under physiological conditions, while the repression can be alleviated under certain pathophysiological conditions (Somers *et al*, 2013). Several uORF properties have been reported to be associated with greater translational inhibition, including strong uAUG context, evolutionary conservation, increased distance from the cap, and multiple uORFs in the transcript leaders (Calvo, 2009). How all these features contribute to uORF function appears highly complex and the exact mechanism of action remains elusive. However, accumulating experiment data suggests that the complexity of uORF-mediated regulation may result from numerous mechanisms acting together. The best understood mechanism is translation re-initiation, which was demonstrated for the historically first discovered mammalian uORFs, present in the ATF4 mRNA. *ATF4* encodes two uORFs in its transcript leader sequence. In non-stressed cells where eIF2-TC is sufficiently available, ribosomes are able to scan the leader sequence until they reach uORF2, thus blocking the main ORF from being translated. Under conditions of cellular stress, when eIF2-GTP abundance is insufficient, ribosomes requires longer time to re-aquire eIF2-TC to reinitiate translation after termination at the uORF1 and therefore bypass the second uORF, leading to the release of repression of the main ORF (Vattem & Wek, 2004).

Besides re-initiation, several other mechanisms underlying uORF function have been described. uORFs may affect downstream translation via their nascent peptide products that can have *cis*-regulatory functions, for instance, leading to ribosome stalling. S-adenosyl-methionine decarboxylase (*AdoMetDC*) encodes a 6-codon uORF. Detailed studies support the conclusion that the *AdoMetDT* uORF protein sequence is responsible for its regulation, with the codon identity at the fourth and fifth positions as well as the length of the peptide being essential for its repressive function (Ruan *et al*, 1994; Mize, 1998). Examples of uORF peptide as *trans*-regulators have also been reported. Transfection of an exogenous construct that contains the argininosuccinate synthase (*AS*) uORF into bovine aortic endothelial cells

was shown to be capable of repressing endogenous *AS* protein. Mutational analysis showed that the sequence of the *AS* uORF is essential for mediating the repressive effect (Pendleton *et al*, 2002, 2005).

1.4.3 Binding sites for RNA binding proteins

In humans, computational approaches suggested that more than 1500 proteins may interact with all classes of RNA (Gerstberger *et al*, 2014). Interestingly, two recent studies jointly identified 1330 proteins that interact with polyadenylated transcripts in crosslinking and affinity purification experiments (Baltz *et al*, 2012; Castello *et al*, 2012). Many of these proteins are functionally implicated in translational regulation. One of the best-characterized examples of translational regulation mediated through the interaction of an RNA binding protein (iron response element protein (IRP)) and its target sites. This iron-dependent regulatory mechanism is important for maintaining cellular iron homeostasis, since many mRNAs that are responsible for iron storage and metabolism contain this element in their transcript leader regions, for example, ferritin, iron-exportin molecular ferroportin (*FPN1*), succinate dehydrogenase-iron protein, erythroid 5-amniolevulinate synthetase (*eALAS*) (Hentze & Kühn, 1996; Aisen *et al*, 2001). The iron response element (IRE) is a highly conserved stem loop structure of approximately 30 nucleotides, which can be recognized by cytoplasmic RNA binding proteins, IRE-binding protein (IRE-BP). IRE-BP can work as iron-sensing factor with their iron-sulfur center binding iron as 4Fe-4S clusters. The binding sites for iron and IRE are largely overlapping, therefore gain or loss of iron triggers significant protein conformational changes for RNA binding. When cellular iron is deprived, IRE-BP bind IRE and block translation of the downstream ORF of ferritin mRNA. Under condition where the cellular iron level is high, the IRE-BP bind iron and its RNA binding capacity is reduced. The ferritin mRNA is released from the IRE-BP and translated into ferritin protein, which in turn sequesters the excess iron (Hentze & Kühn, 1996).

1.4.4 MicroRNA binding sites

microRNAs (miRNAs) are a class of endogenous, noncoding RNAs approximately 20-22 nucleotides long that are found in plants, animals and some viruses. miRNAs are crucial post-transcriptional regulators of gene expression in animals and plants, by inducing mRNA

destabilization and translation repression. To date, over 2000 miRNAs have been identified in the human genome (Kozomara & Griffiths-Jones, 2014), and are predicted to regulate 60% of all protein coding genes (Friedman *et al*, 2009). Thus miRNAs are involved in almost all cellular processes, including development, differentiation, proliferation and stress responses (Shenoy & Blelloch, 2014; Bushati & Cohen, 2007; Ambros, 2011). In spite of the extensive studies on miRNA-mediated gene silencing, its relative contribution to mRNA decay or translation repression towards their overall regulatory effect is still in debate (Guo *et al*, 2010; Bazzini *et al*, 2012; Eichhorn *et al*, 2014). Although the exact mechanism is still unclear, many studies using different biochemical methods and genome-wide analyses have suggested that miRNA inhibit translation at the initiation step (Braun *et al*, 2012; Fabian, 2010; Fabian & Sonenberg, 2012; Huntzinger & Izaurralde, 2011; Guo *et al*, 2010; Bazzini *et al*, 2012; Eichhorn *et al*, 2014) .

miRNAs function by forming a ribonucleoprotein complex, which is also termed as RNA-induced silencing complex (RISC) or miRISC (Kawamata & Tomari, 2010; Ameres & Zamore, 2013). The most crucial and best-characterized components of mammalian RISC is a miRNA and Argonaute proteins (Wilson & Doudna, 2013). Via RISC, miRNA target recognition is achieved. In contrast to plants, where miRNAs form nearly perfect Watson-Crick base-pair interactions with mRNAs, followed by endonucleotyic mRNA cleavage (Jones-Rhoades *et al*, 2006), the general rule for animal miRNA targeting is considered to be basing pair to mRNA with imperfect complementarity. The most essential requirement for metazoan miRNA target recognition is the nucleotides 2-7 or 2-8 of a miRNA's 5' end, which provides the highest specificity for base-pairing and are therefore known as the 'seed' sequence (Filipowicz *et al*, 2008; Bartel, 2009). Such imperfect base pairing avoids the animal RISC cleavage activity. Instead it allows for a cleavage-independent translation repression and/or mRNA decay by recruiting additional effector proteins.

The target sites in mRNAs that are perfectly complementary to the miRNA seed sequences with are referred to as 'canonical' sites, which constitute the main focus of computational microRNA target site prediction. Recent experimental approaches based on high-throughput sequencing technology revealed many sites in mRNAs showing miRNA binding without following the canonical paradigm, that is, binding occurs at sites without perfect seed matches (Chi *et al*, 2009; Hafner *et al*, 2010; Chi *et al*, 2012; Loeb *et al*, 2012; Helwak *et al*, 2013; Betel *et al*, 2010). It appears that the protein products of mRNAs with non-canonical miRNA target sites undergo on average smaller changes when their cognate

miRNA expression is perturbed, compared to those with canonical sites (Khorshid *et al*, 2013; Helwak *et al*, 2013).

In addition to the sequence complementarity, location on an mRNA is also a factor to determine target site functionality. With few exceptions, miRNA binding sites are located in the 3'UTRs in metazoan mRNAs, and are usually present in multiple copies (Doench & Sharp, 2004; Brennecke *et al*, 2005; Lewis *et al*, 2003; Grimson *et al*, 2007; Nielsen *et al*, 2007). Although miRNA target sites can be predicted within transcript leaders and coding regions of endogenous mRNAs, they are less frequent and it seems that they do not have comparable functional effects, relative to those located in 3'UTRs (Farh *et al*, 2005; Lewis *et al*, 2003; Lim *et al*, 2005). Interestingly, in certain cases, miRNA binding sites when positioned in transcript leaders, may activate translation (Vasudevan & Steitz, 2007; Vasudevan *et al*, 2007; Ørom *et al*, 2008). Ørom and coworkers found miR-10a could interact with a transcript leader region immediately downstream of the 5' TOP (5' terminal oligopyrimidine tract) (see section 1.4.6) sequence of many mRNAs encoding ribosomal proteins, and activate their translation in response to stress or amino acid deprivation (Ørom *et al*, 2008). Intriguingly, the interaction of miR-10a with transcript leaders also seems to follow a non-canonical pattern, that does not involve perfect base-pairing at seed regions (Ørom *et al*, 2008).

1.4.5 RNA Secondary Structures

RNA secondary structures in transcript leaders are important translation regulators. Early in the 1990s evidence emerged that many mammalian cellular mRNAs encoding proto-oncogenes, transcription factors, growth factors and their receptors, harbor stable RNA secondary structures in their transcript leaders, suggesting functional relevance in gene expression (Kozak, 1991a). *In vitro* experiments demonstrated that such secondary structures are indeed strong regulators of translation, likely by interfering with translation initiation through impeding recruitment or scanning of the small ribosomal subunit at transcript leaders (Gray & Hentze, 1994).

Subsequent systematic studies of transcript leader structures further revealed that more than 90% of proto-oncogenes, transcription factors as well as growth factors and their receptors are embedded with stable RNA secondary structure with free energies less than -50 kilocalories (kcal)/mole, and more than two thirds of such stable structures are located close to cap structures (Davuluri, 2000). The extent to which RNA structures can interfere with

translation initiation is dependent on the position and stability of these structures, which in turn largely determines the regulatory mechanism. The previously mentioned study demonstrated that hairpin structures with free energies around -30kcal/mole positioned adjacent to the m^7G cap, are sufficient to block translation initiation. Thus suggesting that even a moderately stable secondary structures in the vicinity of cap structure is sufficient to block formation of the PIC (Kozak, 1989b). Intriguingly, the same structure failed to exert its repressive function when positioned 50 nucleotides further downstream (Kozak, 1989b). In contrast, when this stem loop was replaced by a more stable structure (-61 kcal/mole), translation repression was regained, suggesting that stable RNA secondary structures can be partially overcome by the unwinding activity of eIF4A, a component of the scanning 40S ribosome (Kozak, 1989b).

1.4.6 RNA 5' terminal oligopyrimidine tract

RNA 5' terminal oligopyrimidine tract (5' TOP) is a highly-conserved sequence stretch consisting of a C residue at the cap site, followed by a stretch of 4-14 pyrimidines (Meyuhas *et al*, 1996). 5' TOP is a sequence hallmark of most vertebrate mRNAs that encode ribosomal proteins and translation elongation factors (Meyuhas, 2000). Importantly, the protein synthesis rate of these TOP mRNAs is highly sensitive to the growth rate of cells. Growth arrest results in the shift of TOP mRNAs from polyribosome (polysome) (actively translating fraction) to the sub-polyribosome (non-translating fraction) and leads to an inhibition of TOP mRNA translation (Meyuhas, 2000). This common sequence feature of TOP mRNAs enables a coordinated control of protein synthesis of genes encoding the translation apparatus, which consumes a substantial fraction of cellular energy during growth and proliferation. Therefore the coordinated reduction of global translation is essential for cell viability upon cell-cycle arrest or nutrient deprivation.

Intriguingly, the growth-dependent mode of regulation for TOP mRNAs is strictly dependent on the integrity of their TOP sequence as well as on the position of TOP sequence proximal to the cap. The effect of TOP sequence is completely abolished when its first C residue is replaced with A, or when the partial pyrimidines are replaced with purines (Levy *et al*, 1991). Furthermore, the regulatory effect of TOP sequences are diminished when positioned internally, instead of the 5'ends (Avni *et al*, 1994).

Although an extensive body of studies has accumulated, little has been revealed regarding the exact mechanism how the translation of TOP mRNAs is specifically regulated

under different cellular states. Several upstream signaling pathways are reported to be responsible for the activation of TOP mRNA translation. A number of studies have implicated the role of mTOR (Mammalian Target of Rapamycin) signaling pathway, however contradictory data points towards a role of other pathways, such as the PI3K/Akt signaling pathway in the specific regulation of TOP-containing mRNAs under cellular stress conditions (Patursky-Polischuk *et al*, 2009; Stolovich *et al*, 2002; Tang *et al*, 2001).

In addition, several lines of evidence suggest that TOP mRNAs make use of their specific interaction with a defined set of *trans*-acting factors to facilitate their recruitment to the ribosome by mechanisms that differ from the recruitment of other capped mRNAs. Several *trans*-factors have been experimentally shown to bind TOP mRNAs and can potentially regulate their translation, for example, polypyrimidine-binding protein (PTB), La antigen (La), cellular nucleic acid binding protein (CNBP), hnRPD/AUF1 and TIAR-TIA1(Sawicka *et al*, 2008; Pellizzoni *et al*, 1997, 1998; Kakegawa *et al*, 2007; Damgaard & Lykke-Andersen, 2011). TIAR and TIA1 are stress granule-associated proteins, which upon amino acids starvation, are specifically recruited to the oligo-pyrimidine part of TOP mRNAs, and re-localize the mRNAs from polysome to stress granules, therefore cause inhibition of translation. Intriguingly, this process is dependent on mTOR pathways inactivation (Damgaard & Lykke-Andersen, 2011).

1.4.7 Internal ribosome entry sites (IRESs)

Although start codon recognition via 40S ribosome scanning is the dominant mechanism of translational initiation in eukaryotes, a subset of mRNAs can under certain circumstances bypass this scanning mechanism and start translation in a cap-independent manner. Similar to the initiation procedure in bacterial translation, where the bacterial ribosome is recruited through the Shine-Dalgarno/anti-Shine-Dalgarno (SD/anti-SD) interaction, the eukaryotic cap-independent initiation mechanism is also mediated by several *cis*-regulatory elements embedded in transcript leaders, which can directly recruit the PIC. Several *cis*-regulatory elements were shown to facilitate cap-independent translation initiation, most prominently internal ribosome entry sites (IRES), and very recently RNA methylation sites (Meyer *et al*, 2015; Zhou *et al*, 2015).

IRES elements were first discovered in picornaviral genes in the late 1980s (Pelletier *et al*, 1988), and subsequently identified in the transcript leader regions of cellular mRNAs encoding the protein chaperone BiP in the early 1990s (Macejak & Sarnow, 1991). It was

suggested that up to 10% of the human transcriptome are likely to harbor IRESs (Spriggs & Stoneley, 2008). Interestingly, the majority of the identified IRESs were found in mRNAs of proto-oncogenes, growth factors, and proteins associated with the control of cell growth and cell death. IRES-mediated translation initiation appears to be of particular importance for the selective translation of certain mRNAs under stressed or pathophysiological conditions where the cap-dependent translation is globally compromised. Such conditions include but are not limited to: endoplasmic reticulum (ER) stress, hypoxia, nutrient limitation, mitosis and cell differentiation. Interestingly, groups of IRES containing mRNAs are distinct for each stress condition, indicating a specialized translation re-programming in response to different stress stimuli.

Surprisingly, no common sequences or structural motifs are shared among the currently identified cellular IRES elements. Therefore, the presence or absence of IRES elements in a particular mRNA must be experimentally determined in each individual case. The vast majority of cellular IRESs are located immediately upstream of start codons. There are also cases where IRESs are located further downstream and promote the translation of N-terminal truncated proteins (Weingarten-Gabbay *et al*, 2014). Although many of the cellular IRESs identified so far are GC-rich and hence are likely to be rich in RNA secondary structures, no evolutionary conserved secondary structure motifs have been observed thus far (Stoneley & Willis, 2004).

In addition to having diverse sequences and structures, the mechanism by which IRESs exert their cellular function is highly complex as well. In some extreme cases, translation initiation relies entirely on the interaction between the IRES and the small ribosomal subunit, without any participation of canonical initiation factors (Pisarev *et al*, 2005; Kieft, 2008). Furthermore, as one might expect, distinct classes of RNA binding proteins, termed as IRES trans-acting factors (ITAFs) have been identified as regulators of non-canonical internal initiation. In certain cases, ITAFs can function as RNA chaperons, remodeling the mRNA-ITAF structure into structures that are more accessible to other ITAFs or the 40S ribosomal subunit (Mitchell *et al*, 2003; Pickering *et al*, 2004). Finally, several other mechanisms of IRES-mediated internal initiation have been reported. For example, a short 9-nt IRES from mRNA of human homeodomain protein Gtx and a 90-nt IRES from human proto-oncogene IGF1R mRNA may function through an SD-like interaction between the IRES and the18S ribosomal RNA (Chappell *et al*, 2004; Meng *et al*, 2010).

1.5 *Cis*-regulatory elements in human diseases and associated targeted therapies

As summarized above, *cis*-regulatory elements are playing an integral part in translational regulation. Genetic variants that either disrupt or create these elements may alter protein synthesis, and thus have the potential to result in pathological phenotypes (summarized in Table 1).

Consequently, a handful of therapeutics have been developed to specifically target the *cis*-regulatory elements involved in disease development. Among them, an important example is related to the previously described iron-responsive mechanism (see section 1.4.3). The iron response element has been identified in the transcript leader regions of mRNAs that are associated with human diseases, for example, the mRNA of amyloid-β precursor protein (APP) (Rogers *et al*, 2002). Over-expression of APP is implicated in Alzheimer's disease (AD) and Down's syndrome. Consistently, AD patients were shown to have higher cellular metal ion level in their cerebral cortex. Clinical studies have further shown that copper and iron chelation can decrease APP protein levels (Rogers *et al*, 2002b). A screening using 1200 FDA-proved drug has identified several metal ion chelators that suppress the translation of APP in a luciferase assay, by targeting the iron response element in the transcript leader of APP mRNA, showing potential therapeutic value to decrease the APP level among AD patients (Rogers *et al*, 2002b; Payton *et al*, 2003).

Another example worthy of discussing is Ataluren (Translarn^{TM}), which is also known as PTC124, a licensed small-molecule compound for the treatment of patients with genetic diseases that are caused by a nonsense mutation such as cystic fibrosis, DMD (Duchenne muscular dystrophy), haemophilia and several forms of cancer (Peltz *et al*, 2013). PTC124 is capable of helping the ribosome to skip over the premature termination codons (PTCs) and restoring functional protein production in genes otherwise disrupted by these nonsense mutations. More importantly, PTC124 allows the ribosome to selectively read through premature stop, without disturbing the physiological translation termination (Welch *et al*, 2007). Due to the similar structure of uORFs and main ORFs, this drug may also be used to target upstream ORFs, whose termination and re-initiation may be linked to disease and pathology (Chatterjee & Pal, 2009).

Table 1 Overview of human diseases caused by alterations in *cis*-elements in translational regulation

cis-regulatory element type	Diseases	Affected genes	Affected gene functions	mechanisms of *cis*-elements change	References
uORFs	several forms of tumors	MDM2	promote tumor formation by targeting tumor suppressor proteins, such as p53, for proteasomal degradation.	alternative transcript leader	Brown et al.,1999
uORFs	breast and ovarian cancer	BRCA1	tumor suppressor,with functions in cell cyle, apoptosis and DNA damage repair	alternative transcript leader	Sobczak et al.,2002
uORFs	hereditary thrombocythaemia	THPO	a potent humoral regulator of platelet formation	alternative transcript leader/mutation	Wiestner et al.,1998; Ghilardi et al.,1999
uORFs	Alzheimer's disease	BACE1	a protease responsible for the production of amyloid-beta peptides that accumulate in the brain of Alzheimer's diseases patients	unknown	Zhou and Song, 2006; Mihailovich et al.,2007
uORFs	Gonadal dysgenesis	SRY	a testis-dependent factor which initiates male sex determination	mutation	Calvo et al.,2009
uORFs	Van der Woude syndrome	IRF6	encodes a member of the interferon regulatory transcription factor	mutation	Calvo et al.,2009
uORFs	familial hypercholes terolemia	LDLR	cell surface proteins involved in receptor-mediated endocytosis of specific ligands	mutation	Sözen et al.,2005
uORFs	cystic fibrosis	CFTR	involved in the transport of chloride ions	mutation	Lukowski et al.,2011
uORFs	congenital huperinsulinism	KCNJ11	encodes an integral membrane protein and inward-rectifier type potassium channel.	mutation	Huopio et al.,2002
uORFs	rhizomelic chondrodysplasia punctata	PEX7	encodes the cytosolic receptor for the set of peroxisomal matrix enzymes targeted to the organelle by the peroxisome targeting signal 2 (PTS2)	mutation	Braverman et al.,2002
uORFs	proopiomelanocortin deficiency	POMC	encodes a polypeptide hormone precursor	mutation	Krude et al.,1998
uORFs	levodopa responsive dystonia	GCH1	encodes a member of the GTP cyclohydrolase family.	mutation	Tassin et al.,2000
uORFs	juvenile hemochromatosis	HAMP	involved in the maintenance of iron homeostasis, and it is necessary for the regulation of iron storage in macrophages, and for intestinal iron absorption.	mutation	Rideau et al., 2007
uORFs	hereditary pancreatitis	SPINK1	function in the prevention of trypsin-catalyzed premature activation of zymogens within the pancreas and the pancreatic duct.	mutation	Calvo et al.,2009; Witt et al.,2000
uORFs	carney complex type 1	PRKAR1A	a signaling molecule important for a variety of cellular functions	mutation	Calvo et al.,2009
uORFs	β-thalassemia	HBB	determines the structure of the 2 types of polypeptide chains in adult hemoglobin	mutation	Calvo et al.,2009; Oner et al.,1991
uORFs	Schizophrenia predisposition	DRD3	encodes the D3 subtype of the five (D1-D5) dopamine receptors.	mutation	Sivagnanasundaram et al.,2000
uORFs	Aspirin-exacerbated respiratory disease	WDR46	Scaffold component of the nucleolar structure.	mutation	Pasaje et al., 2012
uORFs	arrhythmogenic right ventricular cardiomyopathy/dysplasia (ARVC)	TGF-β3	involved in cellular proliferation, migration, wound repair ,development , tumorigenesis and immunosupression	mutation	Beffagna et al.,2005
uORFs	melanoma	CDKN2A	encodes a cdk4/cdk6 kinases inhibitor that constrains cells from progressing through G1 restraiction point	mutation	Liu et al., 1999
uORFs	bipolar disorder	HTR3A	encodes a ligand-gated ion channel implicated in behavioural disorder	mutation	Niesler et al., 2001
uORFs	Marie Unna hereditary hypotrichosis	HR	regulate the riming of Wnt signaling required for hair follicle cycling and activating the regerateion of hair follicles	mutation	Wen et al., 2009
RNA secondary structures	diabetic nephrophathy	TGF-β1	a multifunctional cytokine involved in cellular proliferation, differentiation, migration and survival	alternative transcript leader	Jenkins et al.,2010
RNA secondary structures	breast and ovarian cancer	BRCA1	tumor suppressor,with functions in cell cyle, apoptosis and DNA damage repair	alternative transcript leader	Sobczak et al.,2002
IRES	X-linked Charcot-Marie-Tooth disease (CMTX)	GJB1	can form gap junction channels that facilitate the transfer of ions and small molecules between cells	mutation	Hudder and Werner, 2000
IRES	myeloma	c-MYC	encodes a transcription factor that plays a role in cell cycle progression, apoptosis and cellular transformation	mutation	Chappel et al.,2000
IRES	fragile X syndrome (FXS)	FMR1	involved in mRNA trafficking from the nucleus to the cytoplasm	expanded transcript leader	Chiang et al.,2001
Kozak sequence	breast and ovarian cancer	BRCA1	tumor suppressor,with functions in cell cyle, apoptosis and DNA damage repair	mutation	Signori et al.,2001
RNA binding protein binding sites (IRE)	hereditary hyperferritinemia-cataract syndrome(HHCS)	FTL	encodes the light subunit of the ferritin protein. A major intracellular iron storage protein	mutation	Girelli et al.,1997

1.6 Identifying *cis*-regulatory elements in translation

Unlike transcriptional regulation, where numerous genome-wide studies based on microarray and next-generation sequencing have been applied to dissect *cis*-regulatory elements in multiple organisms, global analyses of translational *cis*-regulation are still limited. In this chapter I will discuss the currently available approaches in characterizing *cis*-elements that function in translation.

1.6.1 Gene-gene and isoform-isoform based comparison

Assessing the regulatory significance of a potential *cis*-acting sequence feature in translation requires simultaneous measures of protein and mRNA level for the transcripts harboring this feature.

A recent study based on large-scale measurement of absolute protein and mRNA abundance in medulloblastoma cells, has assessed the relative importance of approximately 200 sequence features. With the sequence features identified as dominant regulators in translation, the authors built a combined model, which can explain up to nearly 70% of the protein variance in their system (Vogel *et al*, 2010).

However since these methods largely focused on the relationship of mRNA and protein abundance at the gene-level, where the translation status of individual transcript isoform was averaged out, the contribution of each sequence feature to translational efficiency may not be accurately evaluated. Moreover, sequence features in one genomic region (i.e., transcript leaders) could be confounded by features in other regions (i.e., 3' UTRs).

To address this problem, several studies with dedicated focus on transcript isoforms could to some extent avoid these limitations (Sterne-Weiler *et al*, 2013; Arribere & Gilbert, 2013; Spies *et al*, 2013). For example, Spies et al. have investigated the role of 3' UTRs in translation by comparing the translational efficiency between 3' tandem UTR isoforms. Since 3' UTR isoforms typically share the same ORF and transcript leaders, the effect of *cis*-element in transcript leaders (i.e., uORFs) was minimized.

1.6.2 Quantitative trait locus mapping

Quantitative trait locus (QTL) mapping is a technique to correlate DNA variants (such as SNPs) in a certain genomic region to phenotype traits (such as height, skin color), with the

underlying hypothesis that a QTL is typically linked to the genes that control the phenotype. Substituting phenotype traits with gene expression, expression QTL (eQTL) is regarded as a variant of QTL. As eQTLs would contribute to the variation in expression levels of mRNAs, these loci would either contain *cis*-regulatory elements if eQTLs mapped to the approximate locations of their associated genes, or encode *trans*-regulatory factors if eQTLs mapped to different chromosomes or far from the locations of their associated genes. The former are termed as *cis*-eQTLs or local eQTLs, and the latter are known as *trans*-eQTLs or distant eQTLs.

At the translational level, taking advantage of mass spectrometry and ribosome profiling based approaches (explained in greater detail in section 1.7.3), protein QTL (pQTL) and ribosome QTL (rQTL) are recently developed with analogous concepts (Skelly *et al*, 2013; Wu *et al*, 2013; Battle *et al*, 2015; Ghazalpour *et al*, 2011). With the aim to study genomic regions that regulate translation, pQTL/rQTL mapping is usually performed together with eQTL mapping, to exclude those loci associated with differences in protein abundance or translational-status, that are the result of mRNA expression changes. This defines the protein-/ribosome-specific QTLs (psQTLs/rsQTLs)(Battle *et al*, 2015).

The local subtype of such QTLs usually requires the loci located in the same transcripts as their associated genes (or exonic regions of the same gene-loci). It is reasonable to assume that such loci may contain sequence variants affecting *cis*-elements involved in translational regulation. Therefore, mapping local psQTLs/rsQTLs can in principle be used to study *cis*-elements in translational regulation. However, due to limited number of instances detected in current studies and the complexity of regulatory mechanism, it is still challenging to identify translational *cis*-elements by gathering general rules from sequence variants in pQTLs/rQTLs.

1.6.3 F1 hybrid models

By measuring the allele-specific expression (ASE), F1 progeny of inbred genotypes can serve as a versatile system for studying *cis*-regulation. By definition, *cis*-elements exclusively affect expression of only the allele of a gene that is located on the same chromosome, whereas *trans*-factors are able to affect expression of both alleles of a gene within a cell. In F1 hybrids, two parental alleles are subjected to the same *trans*-environment. Thus, divergent expression patterns between two alleles reflect the difference arising from the *cis*-regulatory

elements in such a system. Therefore, measuring the differences in expression between the two alleles can tell differences in relative *cis*-element activities (Cowles *et al*, 2002).

Measuring ASE at the mRNA level in F1progeny has been used to investigate regulatory variants intra- species as well as inter- species in plants (Guo *et al*, 2008; Zhang & Borevitz, 2009), yeast (Tirosh *et al*, 2009), and animals (Lawniczak *et al*, 2008; Tirosh *et al*, 2009; Wittkopp *et al*, 2004). More recently, a similar framework has been used in yeast to study the impact of *cis*-variants in translation, with protein level measured by either mass spectrometry or ribosome occupancy (Khan *et al*, 2012; McManus *et al*, 2014; Artieri & Fraser, 2014).

1.6.4 High-throughput mutagenesis reporter

Recently, a massively parallel, high-throughput *in vivo* method for testing systematically mutagenized regulatory variants has been reported in characterizing enhancers and promoters in yeast, mouse and human (Patwardhan *et al*, 2012; Melnikov *et al*, 2012; Sharon *et al*, 2012). By using synthetic libraries, these reverse genetics approaches can readily compensate for single experiment that previously required lab intensive work, and enabled rapid accumulation of functional information for thousands of regulatory elements in a single screening experiment.

Similar strategies have been implemented to study translation. In a recent study, Dvir and colleagues set out to identify *cis*-elements in transcript leaders for translational regulation in yeast by constructing a large-scale library of mutants that differ only in the ten nucleotides preceding the translation initiation sites of a fluorescent reporter (Dvir *et al*, 2013). In a more recent paper from the same group, a synthetic oligonucleotide library of thousands of designed sequences taken from hundreds of viruses and human genome was cloned into a bi-cistronic reporter construct. After FACS sorting followed by deep-sequencing, thousands of sequence with IRES-like activity that could potential confer cap-independent translation were uncovered (Weingarten-Gabbay *et al*, 2016).

While these state-of-art studies facilitate the high-throughput dissection and functional annotation of individual regulatory elements, they also suffer from several limitations: inserting the element in a plasmid preceding a non-native gene sequence may result in an artificial sequence feature, i.e., secondary structure, that differs considerably from the native mRNA context, leading to high false positive rates. In addition, the high inter-

cellular variation of mRNA and protein levels makes it more difficult to identify *cis*-elements that have a relatively small effect size.

1.6.5 Biochemical methods to study *cis*-regulatory elements

One of the important facets of *cis*-element characterization is mapping binding sites for RNA binding proteins (RBPs). Conventional RNA immunoprecipitation (RIP) followed by low-throughput identification of binding sites has been long used to capture RNA-protein interaction. The throughput of these traditional methods, was significantly increased by introducing microarray (RIP-Chip) (Tenenbaum *et al*, 2000) and RNA-sequencing (RIP-Seq) technologies (Zhao *et al*, 2010). However, several limitations are common to these methods: They fail to pinpoint the RBP recognition elements (RRE) in the context of the whole transcript; High background binding cannot readily be differentiated from true binding events; Low-affinity RNA-protein interaction cannot be captured effectively.

An important technical advancement was the introduction of UV-mediated protein-RNA crosslinking in living cells. UV-crosslinking can stabilize RNA-protein interactions by formation of covalent bonds only at direct contact sites between protein and RNA, without promoting protein–protein crosslinking (Greenberg 1979), thus enabling stringent identification of true binding events. Combining the principle of *in vivo* UV crosslinking with immunoprecipitation of protein-RNA complexes leads to the development of CLIP (cross-linking and immunoprecipitation). A recent adaptation of CLIP called PAR-CLIP (photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation) uses photoactivatable ribonucleosides to enhance crosslinking efficiency (Hafner *et al*, 2010). In PAR-CLIP, photoactive ribonucleoside 4-thiouridine (4SU) or 6-thioguanosine (6SU) are supplemented into cell culture medium and incorporated into nascent RNAs. The crosslinking between RNA and protein is induced by low energy (365-nm) UV light irradiation, a condition under which inter- or intra- RNA, protein-DNA and protein-protein crosslinks are avoided. RNA-protein complexes are then captured using specific antibodies against the RBP under investigation. The isolated complex is then subjected to RNase trimming, and RNA fragments bound by the bait RBP are then reverse transcribed into cDNA. During this procedure, the conformational change of 4SU as a result of protein crosslinking frequently leads to a thymidine to cytidine (T to C) conversion, and hence helps with precise RRE identification.

As the approaches discussed above can only characterize binding sites for single RBP at one time, a genome-wide variation has been developed to globally characterize the sites of protein-mRNA interactions (Baltz *et al*, 2012), providing more systematic insights into the *cis*-acting sequence elements of the so called "post-transcriptional regulatome".

1.7 Methods for translation assay

1.7.1 Mass spectrometry-based proteomics

Unbiased proteomic measurements have advanced remarkably in recent years, among which mass spectrometry (MS)-based proteomics methods are the most direct way to measure protein abundance from complex samples. MS by definition is an analytical chemistry technique that helps to identify the amount and type of chemicals present in a sample by measuring the mass-to-charge ratio (*m/z*) and abundance of gas-phase ions (Sparkman, 2000). In a typical MS-based proteomics experiment, the proteins of interest are first isolated from cell lysate or tissue. To achieve higher sensitivity, proteins are then fragmented into peptides by enzymatic digestion. Peptides are separated by one or several steps of high-pressure liquid chromatography and are frequently eluted into an electrospray ion source from which they can be further nebulized into tiny charged droplets. After evaporation, the charged peptides enter the mass spectrometer, the mass spectrum of each peptide is then detected. The output of the experiment is the identity of the peptides and therefor the constituents of the protein population of interest (Aebersold & Mann, 2003).

Steady-state measurements of protein abundance by mass spectrometry are sensitive to both protein degradation and synthesis, which cannot directly reflect translational efficiency. Metabolic pulse labeling allows measurements of protein synthesis and decay rates separately, but these experiments are technically more challenging than normal protein abundance profiling. Moreover, a key limitation common to all mass spectrometry-based proteomics methods is that they currently cannot provide as deep measurements of the cellular proteome as RNA-sequencing based methods can for the transcriptome.

1.7.2 Polysome profiling

Among the classic approaches used to monitor *in vivo* translation is the analysis of polysome profiles. Typically, living cells are first treated with translation inhibitor cycloheximide

(CHX) that can halt elongating ribosomes by binding to the Exit (E) site of 60S ribosomal subunit when a deacylated tRNA is present and blocking the eEF2-mediated tRNA translocation (Schneider-Poetsch *et al*, 2010). Cells are then lysed and cellular extracts are loaded onto linear sucrose density gradients and subjected to ultracentrifugation. Free mRNAs and those bound with different numbers of ribosomes can be separated on gradients based on their velocity in a density gradient (as measured in their sedimentation rate Svedberg (S)). The ribosome bound fractions represent a steady-state balance between the steps of translation initiation, elongation and termination: faster initiation leads to more bound ribosomes, and faster elongation and termination lead to fewer bound ribosomes (Arava *et al*, 2003).

A genomic adaption of this method was achieved through combining density gradient centrifugation and microarray or high-throughput sequencing (Arava *et al*, 2003; Sterne-Weiler *et al*, 2013; Arribere & Gilbert, 2013). Through application of this method, translation status can be estimated by analyzing the mRNA constituents associated with each fraction that contains a different number of ribosomes. One limitation of this method results from the difficulty in resolving the exact number of ribosomes bound to highly ribosome-loaded mRNAs, the most-actively translated mRNAs cannot be entirely separated according to their ribosome number, which may result in a compromised resolution for translational efficiency estimate. Moreover, broad application of polysome profiling is to some extent hindered by the technical challenge of polysome fractionation experiments, and in many cases, by the need to collect and analyze many fractions for each sample.

1.7.3 Ribosome profiling

Ribosome profiling is a novel technology that provides a genome-wide *in vivo* occupancy of bound ribosomes by deep-sequencing ribosome-protected mRNA fragments (Ingolia *et al*, 2009, 2012). In brief, cells are first lysed and cellular lysate is subjected to nuclease digestion to trim away the RNA regions that are not enclosed by ribosomes. After digestion, the intact monoribosome (monosome)-footprint complex can be either recovered through sucrose density gradient fractionation, or through sucrose cushion, or more recently, using spin-column chromatography (Ingolia *et al*, 2012). Ribosome protected fragments (RPFs) are then collected and converted into sequencing library. Compared to polysome profiling that measures the degree of mRNA association with the polysome, ribosome profiling provides the exact number of ribosomes translating on each transcript. Given that each ribosome

footprint represents each elongating ribosome and therefore each peptide being synthesized, the ribosome profiling technology enables more quantitative measurement of translational efficiency. However, since the footprints generated by ribosome are short (28-32nucleotides) and are derived only from coding regions, the sequence reads generated in ribosome profiling experiments is not informative for dissecting the translational status of transcript isoforms, especially of those differed in their 5'ends and 3'ends.

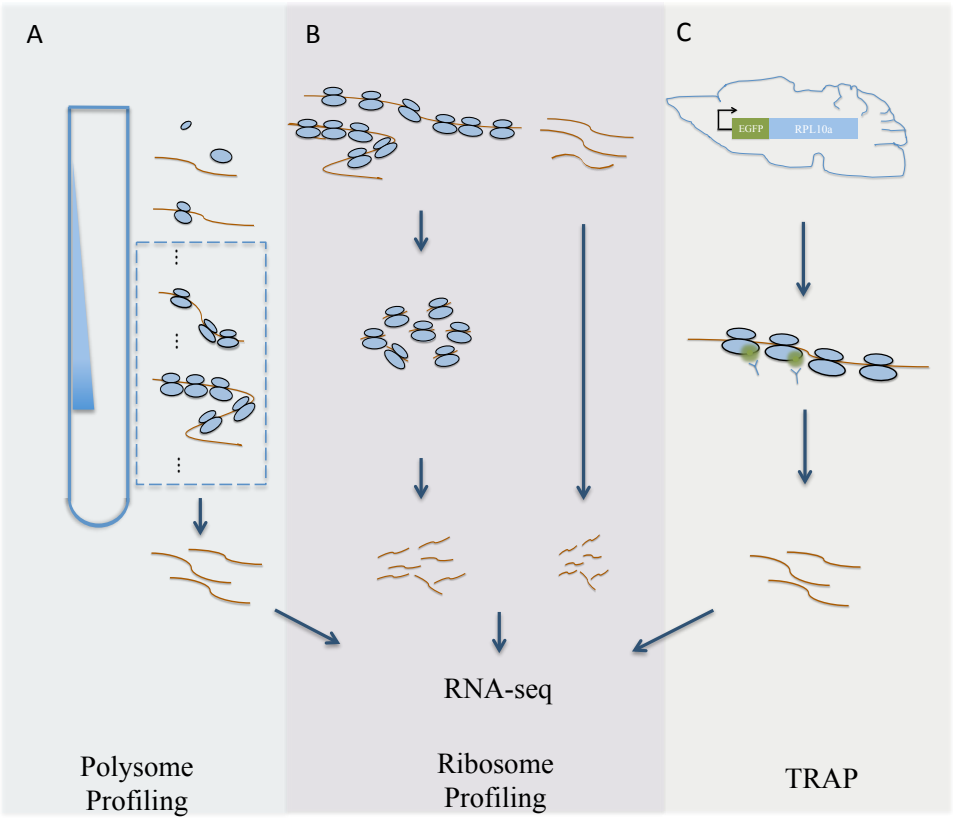

Fig 1.2 Schematic of polysome profiling, ribosome profiling and TRAP. A. Cellular lysate is loaded onto a linear sucrose gradient. After ultracentrifugation, mRNAs bound with polysome are collected and sequenced. B. Cellular extract is first digested with RNase I to remove the unprotected mRNA fragments by ribosome. Monosomes are then recovered and ribosome protected fragments (RPFs) are collected and sequenced. In parallel, total RNA prepared in the same cellular lysate is fragmented and deep-sequenced. C. Engineered bacTRAP mice drive expression of EGFP (green)-tagged-L10a (blue), a ribosomal protein found in polysome, from promoters that are activated in specific cells of the central nervous system. EGFP-L10a-mRNAcomplexes (ribosomes with green dot) are immune-purified from brain tissue from bacTRAP mice, and the ribosome-bound mRNAs are deep-sequenced. Adapted from Kapeli & Yeo, 2012.

Ribosome profiling also enables identification of protein isoforms beyond simply measuring the rate of protein synthesis. Since the footprints of ribosome reflects the genomic regions that is being translated, ribosome profiling can be used to predict open reading frames (ORFs), including novel ORFs that could encode small peptides (Ingolia *et al*, 2011; Ingolia,

2010; Brar *et al*, 2012). In addition, due to the sub-codon resolution of ribosome profiling, it can be applied to annotate non-canonical or alternative translation initiation sites (TISs) by enriching the RPFs at the start codons when cells are pretreated with harringtonine or lactimidomycin (LTM) (Ingolia *et al*, 2011; Lee *et al*, 2012) . LTM blocks translation in a similar but different mechanism as CHX by only binding to the empty E-site of large ribosomal subunit during the translation initiation step when the deacylated tRNA is absent (Schneider-Poetsch *et al*, 2010). Harringtonine binds to free 60S ribosomal subunit and forms an 80S ribosome with the initiator tRNA but blocks aminoacyl-tRNA binding in the A-site and peptide bond formation (Fresno *et al*, 1977).

Another important application of ribosome profiling is to study mechanistic of translational regulation. As ribosome occupancy on a certain codon reflects the time that ribosome spends on that codon, the stacking footprint reads can be used to study ribosome stalling. Ingolia and colleagues have used this excess of ribosome footprints to detect peptide-mediated translational stalling in mammalian cells and RNA-mediated stalling in bacteria (Ingolia *et al*, 2011; Li *et al*, 2012).

1.7.4 Translating ribosome affinity purification

As the above methods are performed on bulk cells without taking into consideration that many tissues are composed of multiple cell types and each of them is of unique gene expression pattern. For example, brain is among the tissues showing the highest heterogeneity. Therefore to study the translational status of a specific group of cell type in brain is a challenging task. To tackle this problem, another immunoprecipitation-based method, termed as Translating Ribosome Affinity Purification (TRAP), has been developed in mice by genetically introducing an epitope tag EGFP (enhanced green fluorescent protein) to a ribosome protein L10a (Heiman *et al*, 2008) (Fig 1.2C). The expression of the engineered ribosome protein is under the control of defined promoters therefore are only active in specific cell types of the central nervous system. Affinity purification of the epitope tag-labeled ribosomes and their associated translating mRNAs allows a specified study of translatome in corresponding cell types.

A similar version of TRAP is named as RiboTag, by tagging HA (hemagglutinin) to Rlp22 followed by affinity purification of HA-tagged polysome (Sanz *et al*, 2009). The RiboTag mice carry an Rpl22 allele with a floxed wild-type C-terminal Exon followed by an identical exon that has three copy of HA-tag. When the RiboTag mouse is crossed to a mouse

expressing Cre-recombinase in a cell-type specific manner, Cre-recombinase activates the expression of HA-tagged Rpl22, which further incorporates into ribosomes.

The methodologies and technologies discussed above have greatly advanced our understanding of translational regulation in recent years and have provided the basis for a systematic characterization of *cis*-regulation in translation. However, a complete understanding of the *cis*-acting regulatome in translation requires the combinatorial application of a diverse set of methods in different biological systems as well as development of new technologies.

In this thesis, I will present the application of a unique combination of methods and technology introduced in sections 1.6 and 1.7 to systematically dissect *cis*-regulation in mammalian translation.

The specific questions addressed in this thesis include:

1. What is the global impact of *cis*-regulatory effects on translation in mammals and what are the *cis*-acting features involved?
2. What is the impact of alternative usage of transcript leaders in translation? What sequence features in transcript leaders are functionally implicated in translational regulation and what is their relative contribution to translational regulation?

2. *Cis*-regulatory Control of Translation in Hybrid Mice

Note: Results in this chapter have been published in Molecular Systems Biology (Hou *et al*, 2015). DOI: 10.15252/msb.156240

Online link: http://dx.doi.org/10.15252/msb.156240

Changes in translational regulation represent one of the major dynamic processes during evolution and such changes largely arise from the divergence in *cis*-regulatory elements (Artieri & Fraser, 2014). Compared to transcriptional regulation, where numerous genome-wide studies have been conducted to dissect *cis*-regulatory divergence in different organisms, global analysis of *cis*-regulation in translation still lags behind. As discussed above, one important approach that can directly address the *cis*-effects in translation is to compare allelic differences in translational efficiencies in an F1 hybrid. Recently, inspired by ribosome profiling technology, several studies sought to investigate allele-specific translational efficiency in F1 hybrid yeast (Albert *et al*, 2014; McManus *et al*, 2014; Artieri & Fraser, 2014). While all these studies revealed a pervasive *cis*-regulation at the translational level, which is comparable to the *cis*-effect that acts on transcription, it is still controversial whether allelic divergence in translational regulation more frequently compensates or reinforces the divergence resulting from allelic mRNA abundance (McManus *et al*, 2014; Muzzey *et al*, 2014). Compared to unicellular organisms, more complex regulation is required in mammalian cells for achieving multi-cellular functions. However, genome-wide profiling of allele-specific translational pattern is still lacking in mammalian systems.

2.1 Study design

To investigate *cis*-effects in mRNA translation, we used an F1 hybrid between two inbred mouse strains, *Mus musculus* C57BL/6J (B6) and *Mus spretus* SPRET/EiJ (SPRET) (Fig 2.1A). The two parental strains diverged ~1.5 million years ago, that result in ~35.4 million single nucleotide polymorphisms (SNPs) and ~4.5 million insertion and deletions (indels) between their genomes (Keane *et al*, 2011). Such a high sequence divergence enabled us to unambiguously determine the allelic origin for a large fraction of sequencing reads and at the same time, provides large number of *cis*-variants that could potentially influence translation.

To monitor translation, we performed quantitative assays for allele-specific translational efficiency by applying mRNA sequencing and deep sequencing-based polysome

profiling (Fig 2.1B). We measured mRNA abundance (total-mRNA) by sequencing the polyadenylated RNAs, and quantified the abundance of mRNA transcripts associated with polyribosome (polysome-mRNA) as well as ribosome profiling (ribosome-mRNA).

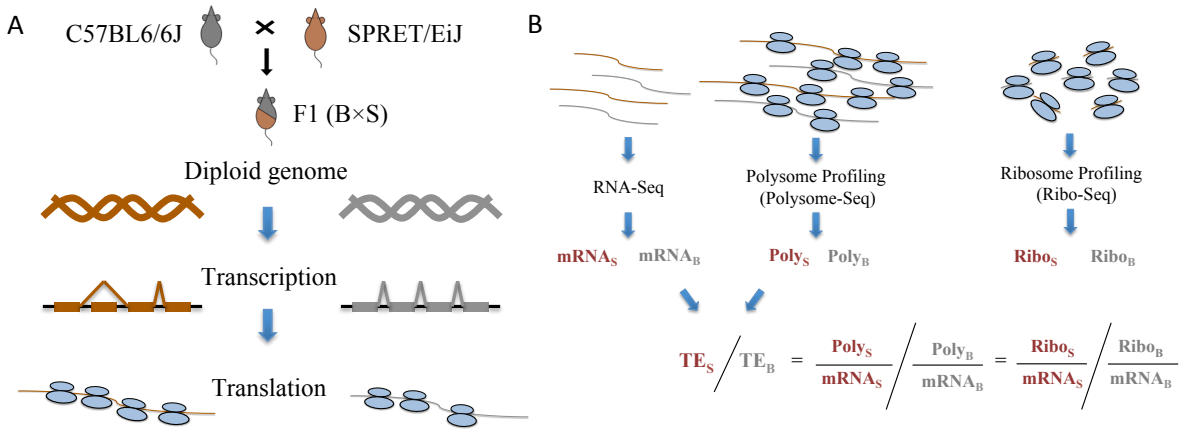

Fig.2.1 Study design. A. Fibroblast cell line was derived from an F1 hybrid mouse between C57BL/6J and SPRET/EiJ inbred strains. B. Using F1 fibroblasts, we deep-sequenced the polyadenylated RNAs to measure mRNA abundance (total-mRNA. In parallel, we performed deep sequencing-based polysome profiling and ribosome profiling to estimate the translation status by quantifying the abundance of mRNA associated with polysome (polysome-mRNA) and with ribosome (ribo-mRNA) respectively.

2.2 Pervasive Allelic Divergence in Translational Efficiency (ADTE)

From two biological replicates, we obtained on average 158.5 million and 94.6 million 100-nt read pairs for total- and polysome-mRNA respectively, with an average of 61% total-mRNA and 65% polysome-mRNA uniquely mapping to B6 and SPRET transcriptome (Fig 2.2A). Translational efficiency (TE) was defined as the abundance ratio between polysome-mRNA and total-mRNA, and only the reads assigned with unambiguous allelic origin were used. Figure 2.2B shows two representative examples with significant ADTE, biased towards the C57BL/6J or the SPRET/EiJ allele, respectively. While the Cnppd1 mRNA was transcribed from both alleles with similar abundance, mRNAs associated with the polysome contained a higher amount of C57BL/6J-derived transcripts, indicating the higher translational efficiency of the C57BL/6J allele. In contrast, transcripts derived from the C57BL/6J allele of the gene *Lbp* was translated at lower efficiency than SPRET/EiJ-derived transcripts.

By using a bootstrapping strategy similar to the method reported by Muzzy *et al*, (Muzzey *et al*, 2014), out of 7156 genes with reliable quantification of both alleles, we found

1008 (14.1%) exhibiting significant allelic divergence in their translational efficiency (Fig 2.2C).

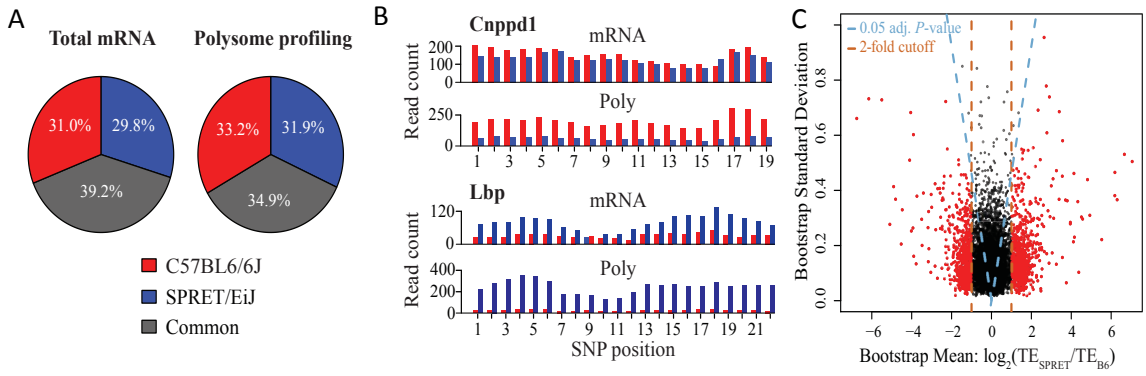

Fig 2.2 Pervasive Allelic Divergence in Translational efficiency (ADTE). A. Percentage of uniquely mapped reads from total mRNA sequencing (left) and polysome profiling (right) that were unambiguously assigned to C57BL/6J (red) and SPRET/EiJ (blue) alleles, or assigned to the two allele with equal probability (common, grey). B. Barplots showing the number of sequencing reads from total-mRNA (Total) or polysome-mRNA (Poly) assigned to C57BL/6J (red), SPRET/EiJ (blue) alleles (y-axis) at different SNP loci (x-axis) across the coding region of genes *Cnppd1* (up) and *Lbp* (low). C. Scatterplot showing the bootstrap means (x-axis) and standard deviations (y-axis) in estimating ADTE for the 7156 genes containing at least five coding SNPs supported with sufficient allelic reads. Dashed blue lines indicated the Benjamini-Hochberg adjusted P-value of 0.05, and dashed brown lines indicated the 2-fold divergence. Genes with significant ADTE (Benjamini-Hochberg adjusted P value < 0.05, allelic TE bias > 2 fold) were depicted as red dots. Adapted from Hou *et al*, 2015.

2.3 Validating ADTE by ribosome profiling and PacBio sequencing

To estimate the accuracy of our ADTE measurements, we performed ribosome profiling to assess mRNA translational status at a higher resolution by directly measuring the number of ribosomes bound by different mRNAs (Ingolia *et al*, 2009). Due to the short length of ribosome protected mRNA fragments (28-32 nucleotide), only 19% uniquely mapped RPF reads could be unambiguously assigned to either allele. Among the 1008 genes with significant ADTE identified based on polysome data, 688 had sufficient allelic ribosome profiling data. Among them, 460 genes (66.9%) showed significant ADTE bias towards the same allele as estimated by polysome profiling (Fig 2.3A). Importantly, no single gene showed significant ADTE towards the different allele between polysome profiling and ribosome profiling.

To assess the accuracy of ADTE quantification based on short reads generated by Illumina sequence platform, we randomly selected 33 genes for independent validation using PacBio RS system. We sequenced RT-PCR products (500-600 bp, spanning at least 3 SNPs)

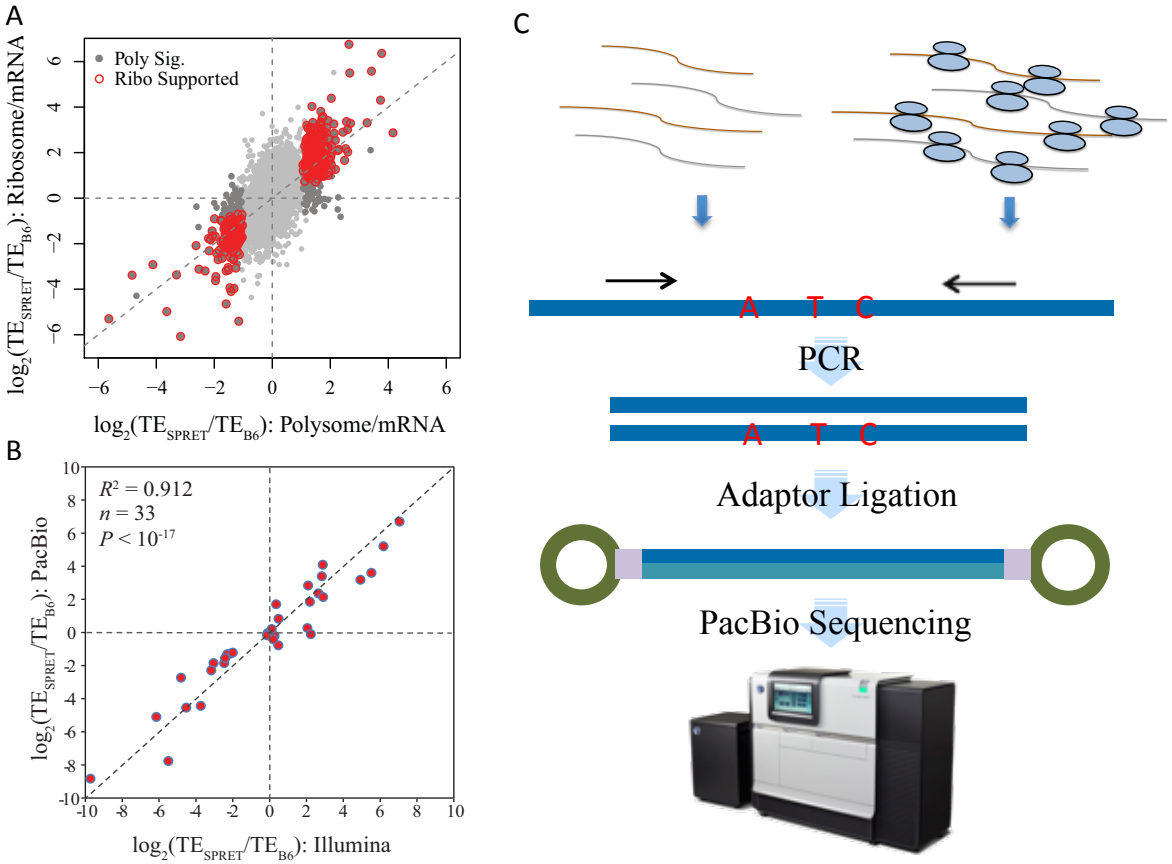

Fig 2.3 Validating ADTE by ribosome profiling and PacBio sequencing. A. Scatterplot comparing the ADTE estimated based on polysome profiling (x-axis) to that based on ribosome profiling (y-axis). All dots represent the 4511 genes with both sufficient polysome profiling and ribosome profiling data. Among them, the 688 genes with significant ADTE based on polysome profiling are depicted in dark grey, of which the 460 genes that were also estimated with significant ADTE based on ribosome profiling are depicted in red circles. B. Schematic of ADTE validation with PacBio sequencing. A, T, C represent SNPs between alleles. C. Scatterplot comparing ADTE estimated based on Illumina sequencing data (x-axis) to that based on PacBio sequencing (y-axis) for the 33 genes. The ADTE estimated based on PacBio sequencing are significantly correlated with that determined by Illumina approach (R^2=0.912, P<10^{-17}). Adapted from Hou *et al*, 2015.

amplified from both total- and polysome-mRNA using primers targeting regions without sequence variance between two alleles (Fig 2.3C). The longer read length is expected to facilitate the unambiguous assignment of the PacBio reads to the parental alleles. Allelic ratios of both total- and polysome-mRNA abundance can thus be calculated with higher

precision. As shown in Fig 2.3B, the ADTE estimated in this way were correlated well with that determined by our Illumina sequencing-based approach (R^2=0.912, $P<10^{-17}$).

2.4 Genes with ADTE contain higher sequence variants in 5'UTRs

As discussed above, the ADTE observed in the F1 hybrid reflects the impact of the allelic differences in *cis*-elements present on mRNA sequence. To confirm this, we first calculated the density of sequence variants between the two parental genomes for 634 genes with significant ADTE and 1291 control genes without ADTE (restricted to single-isoform genes with unambiguous 5'and 3' UTR annotations). As shown in Figure 2.4A, the genes with significant ADTE harbored higher density of sequence variants than the control genes ($P=1.7\times10^{-5}$, Kolmogorov–Smirnov test).

Next, we sought to explore how sequence variants in different positions along the transcripts contribute to allelic TE divergence. For this purpose, each gene was separated into 5'UTR, CDS and 3'UTR regions. SNP density was calculated in each region and then normalized against the overall SNP density of the same gene. Compared to the 1291 control genes, the 634 genes with significant ADTE showed higher enrichment of SNPs in 5'UTRs (Fig 2.4B).

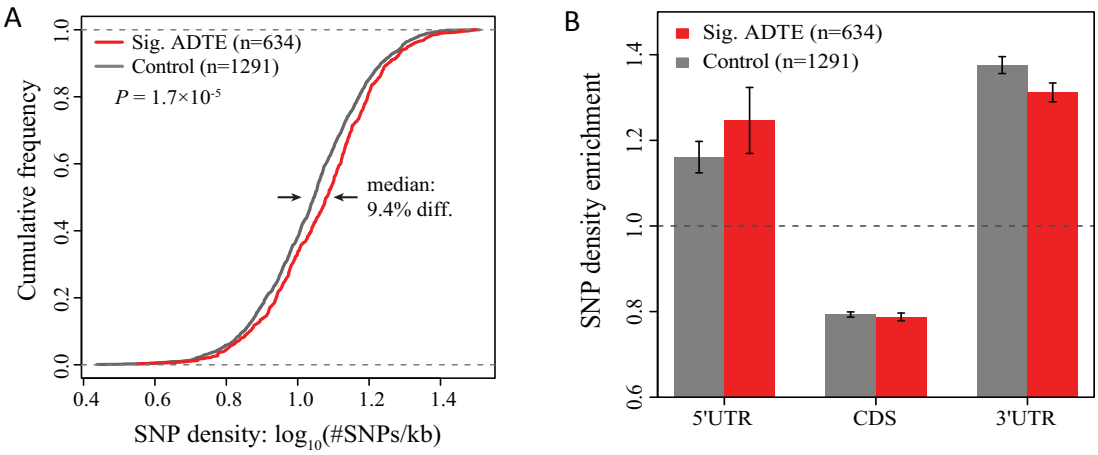

Fig 2.4 Higher SNP density in 5'UTRs for genes with ADTE. A. The cumulative distribution function of SNP density (number of SNPs per kb) for genes with significant ADTE (red) and without (controls genes, grey). Compared to the control genes, the genes with significant ADTE showed significantly higher SNP density ($P=1.7\times10^{-5}$, Kolmogorov–Smirnov test), with the median SNP density for the genes with significant ADTE being approximately 9.4% higher than that for the control genes. B. Barplots showing the regional SNP density enrichment for the genes with significant ADTE (red) and the control genes (grey). Compared to the 1291 control genes, the 634 genes with significant ADTE showed relatively higher enrichment of SNPs in 5'UTRs. Barplot with error bars, mean ± SE. Adapted from Hou *et al*, 2015.

2.5 mRNA secondary structures proximal to start codons contribute to ADTE

Inspired by the observation above, we further examined the SNP enrichment within 5'UTRs close to the start codon. As shown in Figure 2.5A, compared to the control group, the genes with significant ADTE contained higher SNP in the region proximal to the start codon.

Further, mRNA secondary structures in the vicinity of the start codon have been reported to affect translation in *Escherichia.coli* and yeast (Kudla *et al*, 2009; Dvir *et al*, 2013). We therefore asked if they could account for the observed allelic translation divergence in mammalian cells. The minimum free energy (MFE), which represents the most stable structure on an RNA sequence, was calculated for RNA sequences with the length from 20 to 50 nucleotides surrounding the start codon. We compared the MFE between two alleles and then correlated the MFE difference to the observed ADTE. Alleles with less stable local secondary structure surrounding the start codon tend to show higher TE (Fig 2.5B).

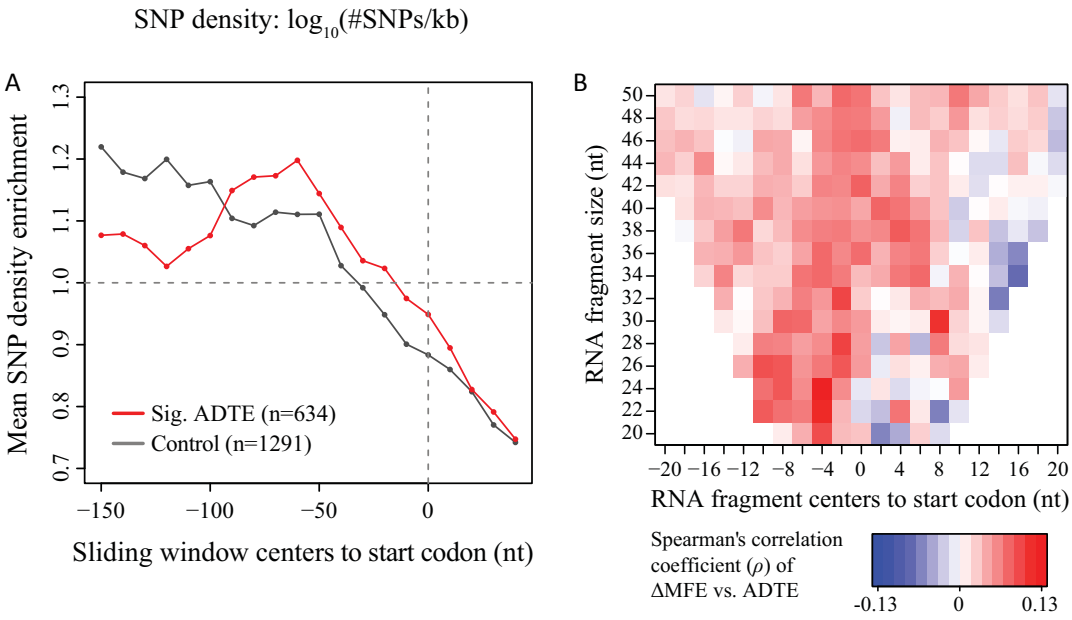

Fig 2.5 Genes with ADTE show larger difference in RNA folding near start codons. A. SNP density enrichment in 5'UTR proximal to the start codon for the genes with significant ADTE (red) and the control genes (grey). The distance of window center to start codon is indicated on the x-axis and the mean SNP density enrichment from the two gene groups is indicated on the y-axis. Although the SNP enrichment difference in five windows had a nominal $P < 0.05$, after Benjamini-Hochberg correction for multiple testing, no windows remained significant (adjusted $P < 0.05$). B. Heatmap showing the Spearman's correlation coefficient (ρ) between ADTE and the allelic difference in the minimum free energy (MFE) of mRNA segments surrounding the start codon. For each mRNA segment, its length is indicated on the y-axis and the distance of its center to start codon is indicated on the x-axis. Color keys for ρ were shown below the heatmap. Note that ρ in none of the segments achieved statistical significance (FDR<0.05). Adapted from Hou *et al*, 2015.

2.6 Proximal out-of-frame upstream AUGs has impact on ADTE

Another class of potent translational regulators in 5'UTRs is uORFs/uAUGs. To investigate whether the presence of uORFs/uAUGs contributed to the observed allelic TE bias, we separated 1640 (695) genes with uORFs (uAUGs) into two groups, one group consists of 1597 (618) genes with uORFs (uAUGs) on both alleles, while the other group contains 43 (77) genes with uORFs (uAUGs) only on one allele. Comparing the distribution of ADTE between the two groups, we did not observe significant differences between the two groups for either uORFs (Fig 2.6B; P=0.32, Mann–Whitney U test) or uAUGs (Fig 2.6C; P=0.72, Mann–Whitney U test). A previous study suggested that uAUGs located within the same frame as the main ORF (in-frame) or not (out-of-frame) may have different influence on translation of the main ORFs (Dvir *et al*, 2013). Therefore, we separated the genes with uAUGs into two sets, each of which containing only in-frame or out-of-frame uAUGs, as illustrated in Figure 2.6A. Interestingly, while we observed no significant correlation between ADTE and presence/absence of the in-frame uAUGs (Fig 2.6D; P=0.30, Mann–Whitney U test), we found that, for genes with proximal (<=100-nt upstream of the main ORF) out-of-frame uAUGs in only one allele, ADTE significantly differed from that of genes with proximal out-of-frame uAUGs in both alleles (Fig 2.6D; P=0.038, Mann–Whitney U test). The observation indicates that the presence of a proximal out-of-frame uAUG may hamper the translation of the main ORF.

A number of other sequence features are known to potentially affect translation, including GC content, codon bias (measured by codon adaptation index, CAI) and miRNA target sites (Vogel *et al*, 2010; Santhanam *et al*, 2009; Plotkin & Kudla, 2010; Mayr & Bartel, 2009; Sandberg *et al*, 2008). However, we did not observe any correlation between these features and allelic translational efficiency difference in our study.

2.7 Comparable allelic regulation of translation versus transcription, and their coordination

Allelic divergence in protein abundance in F1 hybrids is governed by the allele-specific transcriptional as well as translational regulation. To explore the relative contribution of the two processes, we first calculated allelic bias in mRNA abundance, which can be expected to largely result from allelic transcriptional regulation (data not shown). Based on total-mRNA-sequencing data, and when applying the same bootstrapping strategy at identical thresholds as

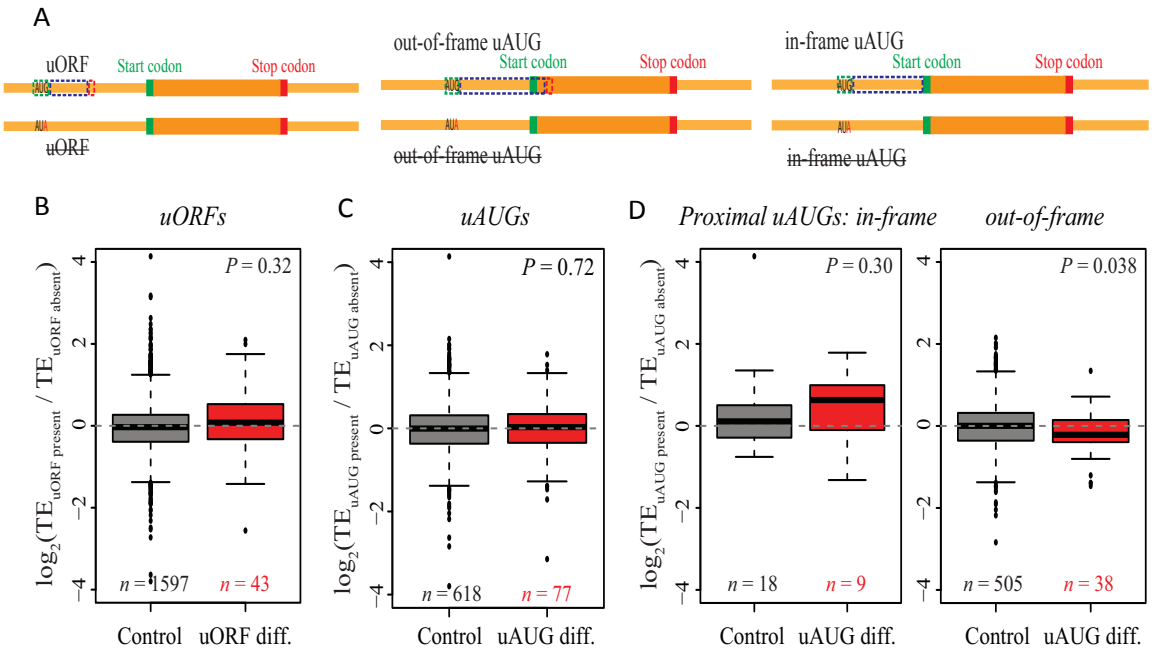

Fig 2.6 Impact of uORFs/uAUGs on ADTE. A. Schematic representation of genes with uORF, out-of-frame uAUG and in-frame uAUG in one allele, but without in the other allele. B. Boxplots comparing the distribution of ADTE between 1597 genes with uORF present in both alleles (grey) and 43 genes with uORF present in only one allele (red). No significant differences between the two groups were observed (P=0.32, Mann–Whitney U test). C. Boxplots comparing the distribution of ADTE between 618 genes with uAUG presence in both alleles (grey) and 77 genes with uAUG presence in only one allele (red). No significant differences between the two groups were observed (P=0.72, Mann–Whitney U test). D. Boxplots comparing the distribution of ADTE between 18 (505) genes with proximal in-frame (out-of-frame) uAUG presence in both alleles (grey) and 9 (38) genes with proximal in-frame (out-of-frame) uAUG presence in only one allele (red). While no significant correlation was observed between ADTE and presence or absence of the proximal in-frame uAUGs (P=0.30, Mann–Whitney U test) for genes with proximal out-of-frame uAUGs in only one allele, ADTE significantly differed from that of genes with proximal out-of-frame uAUGs in both alleles (P=0.038, Mann–Whitney U test). Adapted from Hou *et al*, 2015.

we used for polysome-mRNA sequencing (adjusted P-value < 0.05 and allelic divergence > 2-fold, FDR = 4.74%, Fig 2.7A), we identified 1041 out of 7892 genes with significant allelic difference in mRNA abundance. As shown in Figure 2.7A, the proportion of genes exhibiting allelic bias at mRNA abundance and at translational efficiency was similar (Fig 2.7A; 13.2% vs. 14.1%, P=0.11, Fisher's exact test). Additionally, the allelic difference in mRNA abundance could only explain up to 43% of the allelic divergence in polysome-bound mRNA abundance (Fig 2.7B). These observations suggested that allelic regulation at the two levels were of comparable importance in determining the cumulative allelic bias in protein abundance.

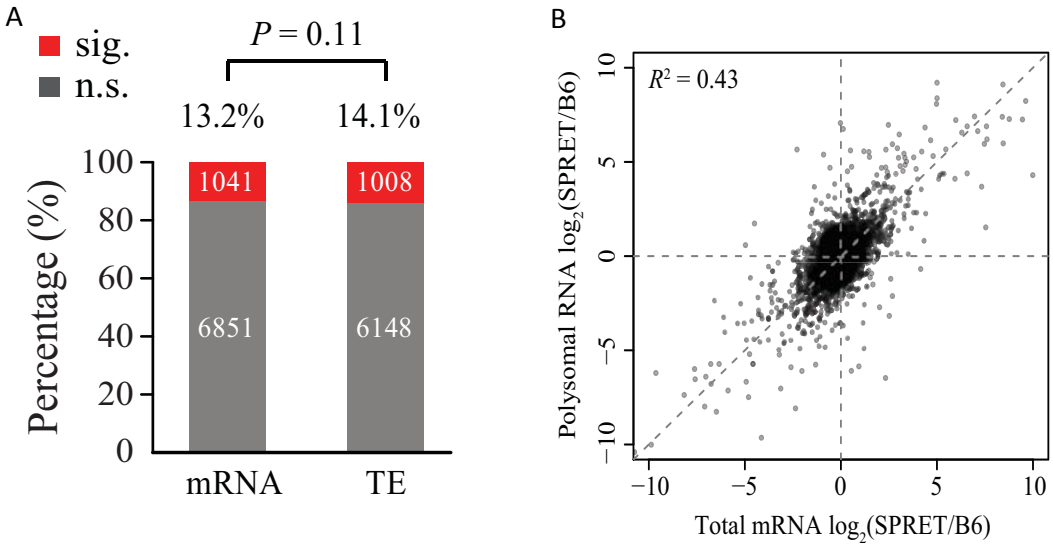

Fig 2.7 Comparable allelic regulation of translation and transcription. A. Comparable *cis*-effects at transcriptional and translational levels. Barplots show 13.2% and 14.1% of genes with significant allelic bias at transcriptional or translational levels, respectively. The difference between both of these fractions is statistically non-significant (P=0.11, Fisher's exact test). B. Scatterplot comparing the log2 transformed allelic ratio of cellular mRNA abundance (x- axis) versus that of polysome associated mRNA abundance (y-axis). Each dot represents one gene. The R^2 of 0.43 indicates that less than half of the allelic divergence in polysome associated RNA abundance can be explained by the allelic divergence in mRNA cellular abundance. Adapted from Hou *et al*, 2015.

Several recent studies in yeast have shown that allelic translation and transcription were regulated in a coordinated fashion; however, it is still in debate whether the regulatory effects at these two levels reinforce or compensate each other (Artieri & Fraser, 2014; McManus *et al*, 2014; Muzzey *et al*, 2014). That is, the allelic bias at the transcriptional level that favors one allele might be further enhanced by translational efficiency that favors the same allele (reinforcing). Alternatively, translational efficiency that favors the other allele would lead to compensatory effects. Here we sought to use our mammalian hybrid system to investigate how allele-specific translation and allele-specific mRNA abundance are coordinated. As shown in Figure 2.7A, out of 7892 genes, 1041 and 1008 showed significant allelic bias in either mRNA abundance or translational efficiency, respectively. Among them, 185 genes displayed allelic biases at both levels. 137 out of the 185 overlapped genes showed compensatory effects between the two processes (mRNA abundance and TE divergence in opposite direction), nearly two times more frequently than those with reinforcing effects (mRNA abundance and TE divergence in the same direction) (n=48) (Fig 2.8A).

We then classified the 7892 genes into three groups according to their allelic bias at transcriptional or translational levels, and asked whether genes with or without allelic bias in transcriptional and/or translational regulation had distinct biological functions. As shown in Figure 2.8B, the genes without allelic biases in either process were highly enriched in constitutive cellular processes, for example chromatin modification and transcription. While compensatory genes showed enrichment of certain essential functions, such as regulation of proteolysis, reinforcing genes were enriched in two specific functional categories, i.e., cartilage development and sensory perception of sound.

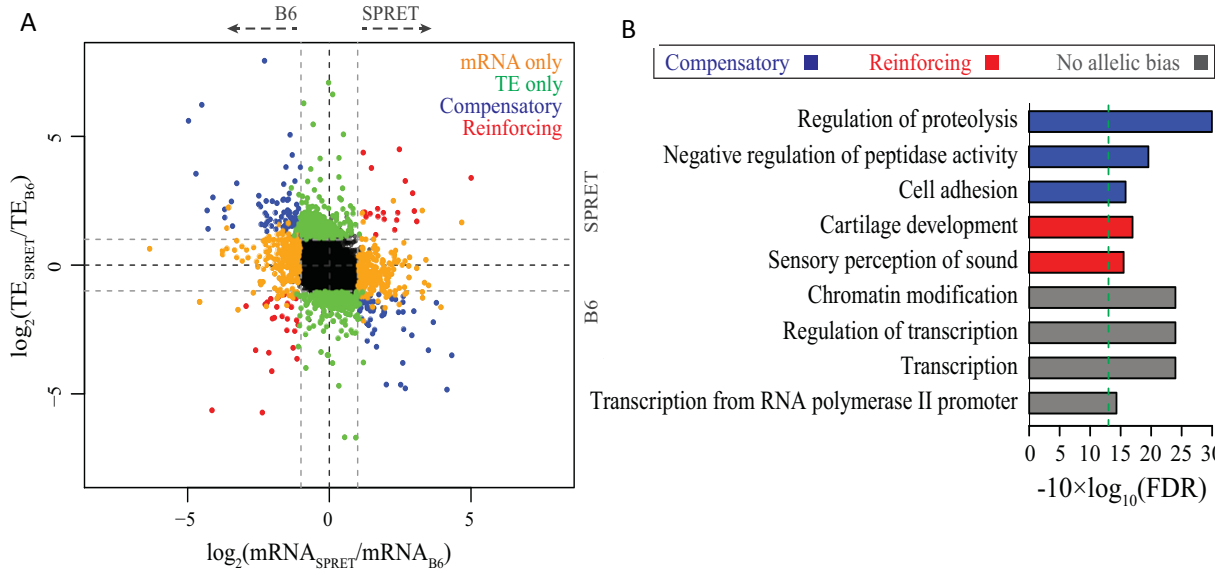

Fig 2.8 *cis*-effects at transcription and translational level are more frequently compensatory. A. Scatterplot comparing allelic divergence of each gene (log2 transformed-fold change) at transcriptional (x-axis) and translational (y-axis) levels. Grey dash lines indicate 2-fold divergence at either level. Compensatory and reinforcing genes are depicted as blue and red dots, respectively. Genes with significant allelic bias at only mRNA level and only TE level are depicted in orange and green, respectively. B. Gene Ontology (GO) enrichment of compensatory genes (blue), reinforcing genes (red), and genes without allelic bias at either level (grey). All shown GO term analyses were performed with an FDR < 0.05. Adapted from Hou *et al*, 2015.

2.8 Summary

To globally investigate *cis*-divergence in translational regulation in mammals, we applied mRNA sequencing and deep sequencing-based polysome profiling to quantify translational efficiency. We chose the F1 progeny between *Mus musculus* C57BL/6J and *Mus spretus* SPRET/EiJ as our model system because the two have the largest number of genetic variants among all the mouse strains with high quality genome assembly available. With over 60% of

mapped mRNA-sequencing as well as polysome profiling reads unambiguously assigned to the parental alleles, 7156 genes could be analyzed with reliable quantification of both alleles. Importantly, we validated our results with two independent approaches, 1) PacBio full-length sequencing to assess the accuracy of allelic read mapping; 2) ribosome profiling to support the allelic translational status estimated based on polysome profiling. This multilayered validation demonstrated high quality of our data. In total, we identified 1008 genes (14.1%) exhibiting significant allelic difference in translational efficiency. Further analysis of sequence features of these genes with biased allelic translation revealed a statistically significant impact on translational efficiency by local RNA secondary structure near the start codon as well as proximal out-of-frame upstream AUGs. Finally, we observed that the *cis*-effect was quantitatively comparable between transcriptional and translational regulation. Moreover, *cis*-effects in the two processes were more frequently compensatory, suggesting a role of the translational regulation in buffering transcriptional noise and thereby maintaining the robustness of protein expression.

3. *Cis*-regulatory Impact of Transcript Leaders on Translation

3.1 Transcription start site (TSS) heterogeneity and transcript leader isoforms in mammals

In the work on allele-specific translation in F1 hybrid mice described above, we observed that SNPs associated with translational efficiency divergence were more enriched in transcript leader (TL) regions compared to other genomic regions, indicating a crucial role of TLs in regulating translation. Previous studies in yeast have shown that around two hundred yeast genes express multiple isoforms with different TLs, with many of which displaying diverse translational status (Arribere & Gilbert, 2013). Both *in vitro* and *in vivo* analyses have demonstrated that different TL sequences derived from the same yeast genes cause large differences in translational efficiency (Rojas-Duran & Gilbert, 2012). Compared to unicellular organisms like yeast, core promoter architecture in mammals displays much higher complexity and transcription can initiate over much broader genomic regions (Lenhard *et al*, 2012). Moreover, amounting evidences suggest that approximately 50% of human and mouse genes have multiple TSSs (Kimura *et al*, 2006; Cooper *et al*, 2006; Baek *et al*, 2007), with many of which displaying a highly dynamic and cell type-specific expression manner (Forrest *et al*, 2014). The prevalence of alternative TSS usage enables a highly dynamic and specialized transcriptional control of mRNA production and more importantly, substantially diversifies the repertoire of transcript variants, conferring great potential for differential translational regulation. Previous studies of individual mammalian genes have demonstrated that transcription initiation at alternative sites can drastically alter the TL length, resulting in enhanced or diminished protein synthesis rates (Pozner *et al*, 2000; Courtois *et al*, 2003; Blaschke *et al*, 2003). Such TSS switches are usually of great functional significance, and are frequently associated with pathologic phenotypes (Arrick *et al*, 1991; Sobczak & Krzyzosiak, 2002) (refer to Table 1 for more examples).

3.2 Genome-wide assessment of translational status of TL isoforms with CAp Profiling of TRanslational Efficiency (CAPTRE)

In light of the previously described insights, it becomes a question of great concern to what extent highly divergent TL isoforms can affect translational regulation at a genome-wide scale. However, global profiling of differentially translated alternative TLs in mammals is largely limited due to the technical challenge of accurately assembling TL isoforms and

quantitatively measuring their translational status at the same time. Very recently, several studies tried to explore the underlying mechanisms for isoform-specific translational regulation by combining polysome profiling and RNA-seq, with the analyses focused on annotated transcript isoforms (Dieudonné et al, 2015; Floor & Doudna, 2016). However, the incomplete transcript annotation used in these studies, largely obstructed the identification of TL isoforms and therefore their impact on translation.

Here, we advance current technologies by developing CAp Profiling of TRanslational Efficiency (CAPTRE), which combines polysome fractionation and Cap-profiling, a 5'-end sequencing strategy (see section 3.3). Polysome profiling is a well-established and widely-used method to assess the *in vivo* translational status of mRNAs (Arava et al, 2003; Spies et al, 2013; Arribere & Gilbert, 2013). In this study, mRNAs bound by different number of ribosomes were separated into seven fractions on a sucrose density gradient through velocity sedimentation (Fig 3.5B). Following polysome fractionation, each fraction is subjected to Cap-profiling. Translational status of different TL isoforms is estimated by measuring the relative abundance of captured 5' end sequences across sucrose gradient fractions (Fig 3.1). To enable normalization of sequencing reads across different density fractions and to control for loss of RNA at each manipulation step, an identical amount of *D. melanogaster* total RNA is added as a spike-in control immediately after collecting the samples (Fig 3.1). To measure translational efficiency, the average number of ribosomes per mRNA transcript associated with each corresponding TSS is calculated (Spies et al, 2013) (Fig 3.5A).

3.3 Global identification of mRNA 5'ends by Cap-profiling

For genome-wide annotation of TLs of all capped RNAs, we developed a method called Cap Profiling or CAP (Fig 3.2), by adapting the cap-trapper full-length cDNA cloning method for 5'end-sequencing that is compatible with high-throughput sequencing platforms (Carninci *et al*, 1996). Total RNA is reverse-transcribed using a random primer fused to an Illumina sequencing primer under optimized reverse transcription condition where even highly structured RNA can be efficiently reverse transcribed. In subsequent biochemical modification steps, the two adjacent hydroxyl groups at 5'cap structures of Pol II transcripts and at 3'ends of all RNA species are first oxidized into ketones and then are biotinylated

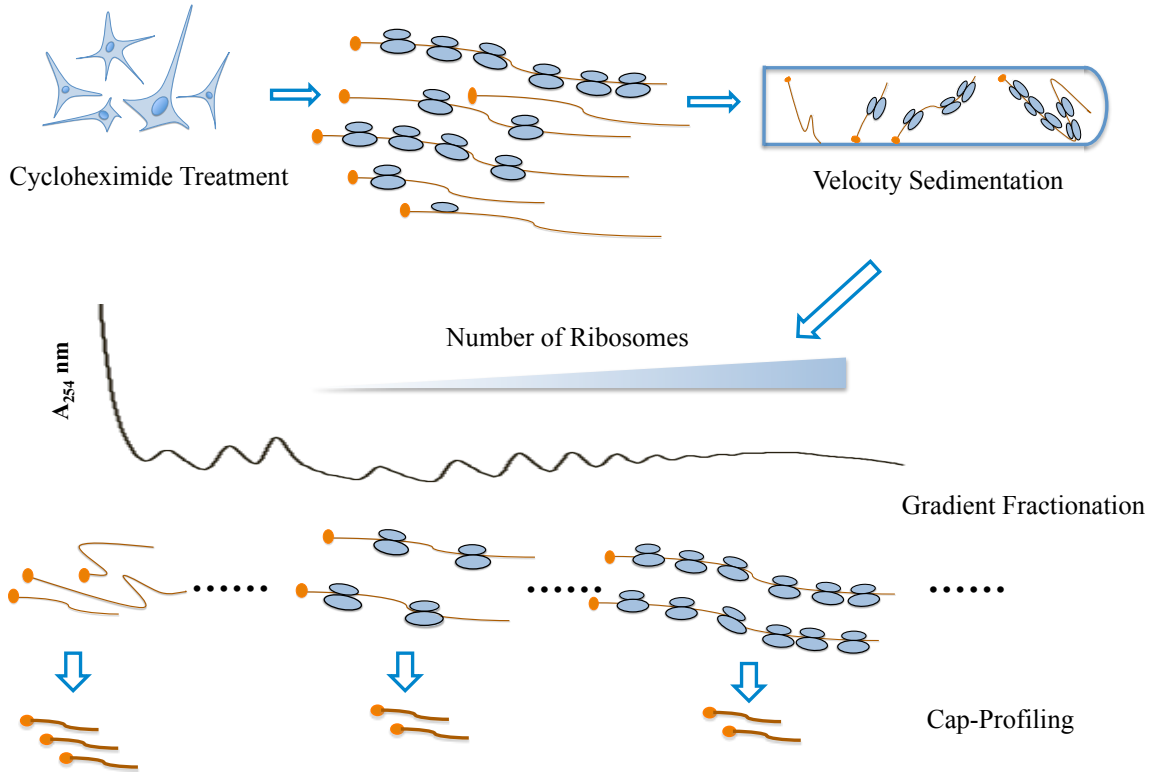

Fig 3.1 Schematic of study design. CAp Profiling of TRanslational Efficiency (CAPTRE), combining polysome fractionation and Cap-profiling. The translation of murine fibroblast cells was arrested *in vivo* by cycloheximide. Cells were lysed and cellular lysate was fractionated in a sucrose gradient through velocity sedimentation. RNAs associated with different number of ribosomes were fractionated and collected. The 5'ends of RNAs were captured by Cap-profiling and quantified by high-throughput sequencing.

by biotin hydrazide. To specifically select only cDNA fragments that extend all the way to the 5'end of each RNA template, 5' truncated cDNAs and RNAs that are not reverse transcribed have to be removed. To achieve this goal, the single-stranded region of RNAs that are not protected by synthesized cDNA are subjected to RNase I digestion. Completely reverse transcribed cDNA/RNA hybrids protected from RNase I treatment are then selected by streptavidin coated magnetic beads. sscDNAs are then released from the beads by alkaline hydrolysis and ligated to a double-stranded adaptor with random nucleotide overhangs. Ligation products are then amplified and sequenced on Illumina platforms. After mapping to genomes, uniquely mapped CAP tags are then clustered, and each cluster corresponds to one TSS. Compared to cap analysis of gene expression (CAGE), which is also a cap-trapping based method designed to survey the 5'ends of RNAs by tagging the first 27 nucleotides (nt) (Takahashi *et al*, 2012), Cap-profiling generates longer and paired-end sequencing reads (2×100 nt in this study) that substantially enhance mapping efficiency and therefore facilitate

the annotation of 5' boundaries of transcript. Based on its application in this study, Cap-profiling yielded 2.9% more uniquely mapped reads in total compared to 27-nt tags (the theoretical length of CAGE tags), and 13.1% more reads in annotated TLs. Moreover, Cap-profiling simplifies the library preparation procedure by skipping conventional and labor intense restriction enzyme digestion.

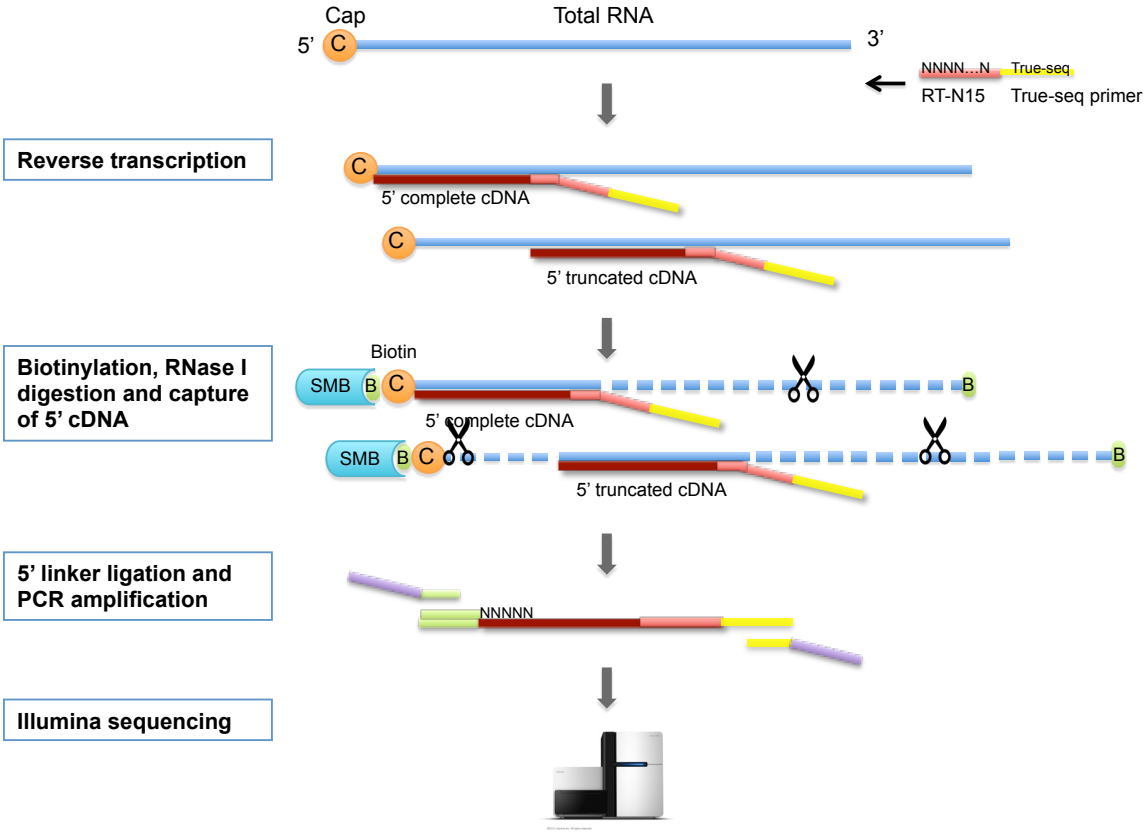

Fig 3.2 Schematic of Cap-profiling. Total RNA was reverse-transcribed using random primers (N15) (light red) fused to the 3' part of Illumina TruSeq Universal Adaptor sequence (yellow). Cap structure (orange) and 3' ends of all RNAs were biotinylated (green). Single-stranded RNA regions that are not protected by synthesized cDNAs (dark red) including the 3' ends were cleaved by RNase I (scissors). The 5' complete cDNA containing the biotinylated cap structure was then captured by Streptavadin coated magnetic beads (blue). Single-stranded cDNA was then ligated with double-stranded 5' linkers with random overhangs (green). cDNAs were amplified for 18 cycles with PCR primers containing Illumia sequencing primers (purple). The amplified libraries were sequenced using 2 x 100 nt cycles (paired-end protocol) on an Illumina HiSeq2000 platform. C, 5'cap; B, biotin; SMB, strepavadin-coated magnetic beads; RT-N15, reverse transcription random primer. Adapted from Takahashi *et al*, 2012.

Using Cap-profiling, we first characterized TSS usage in murine fibroblast cells. On average, we obtained 46.2 million paired-end reads per fraction, after filtering non-coding RNA, 78% of the reads could be uniquely aligned to the mouse genome, with 92.3% of

which mapped to the annotated TL regions and 1Kb upstream of the annotated TSSs (Fig 3.3A). To increase the sensitivity in detecting expressed TSSs, we combined all the sequencing data together and determined TSSs by clustering mapped reads. In total, we identified 22,357 TSSs that were assigned to 10,875 protein-coding genes annotated in the RefSeq database. The read counts for each TSS correlated very well between two biological replicates for each of the seven fractions (R=0.95~0.98). Out of this set, 17,033 (76.2%) TSSs were mapped within gross TL regions of 9,951 protein-coding genes. More specifically, 13,705 sites mapped to the annotated TL regions and 3,328 sites mapped to a 1 kb window upstream of annotated TSSs (UP-1kb; Fig 3.3B).

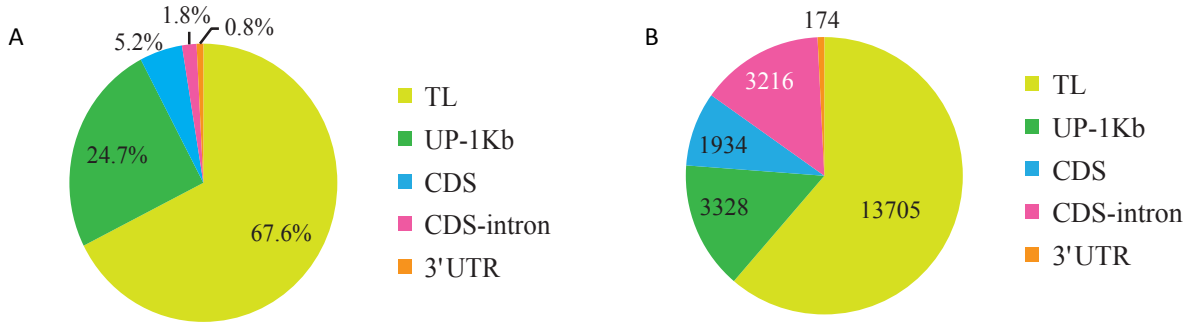

Fig 3.3 Distribution of uniquely mapped reads and TSS clusters relative to known TSSs and other genomic regions. A. Piechart showing the distribution of reads identified in this study in different regions of protein-coding genes. 92.3% of the reads were mapped to the annotated TL regions and 1kb upstream of the annotated TSSs. B. Piechart showing the regional enrichment for TSS clusters identified in this study. 17,033 (76.2%) TSSs were mapped within gross TL regions of 9,951 protein-coding genes, including both annotated TL region (n=13,705) and 1kb upstream of the annotated TSSs (UP-1kb; n=3,328).

Out of 9,951 protein-coding genes with at least one TSS detected in the gross TL regions, 4,153 (41.7%) have multiple TSSs (Fig 3.4A). While genes with a single TSS showed higher expression and were enriched in genes encoding proteins with essential functions (e.g. nucleosome assembly and translation), genes with multiple TSSs were enriched in regulation of transcription (Fig 3.4B). Figure 3.4C shows two representative examples with alternative TSSs. Ribosome number increases from gradient fraction 1 (free ribosomal fraction) to fraction 7 (ribosome number ≥9) (Fig 3.5B). For *Itpr3*, the mRNA isoform transcribed from the distal TSS showed a weaker ribosome association compared to the mRNA transcribed from the proximal TSS, indicating that the proximal TSS-associated TL isoform translated more efficiently. In contrast, transcripts derived from the proximal TSS of gene *Peol* showed lower translational efficiency compared to those transcribed from the distal TSS (Fig 3.4C).

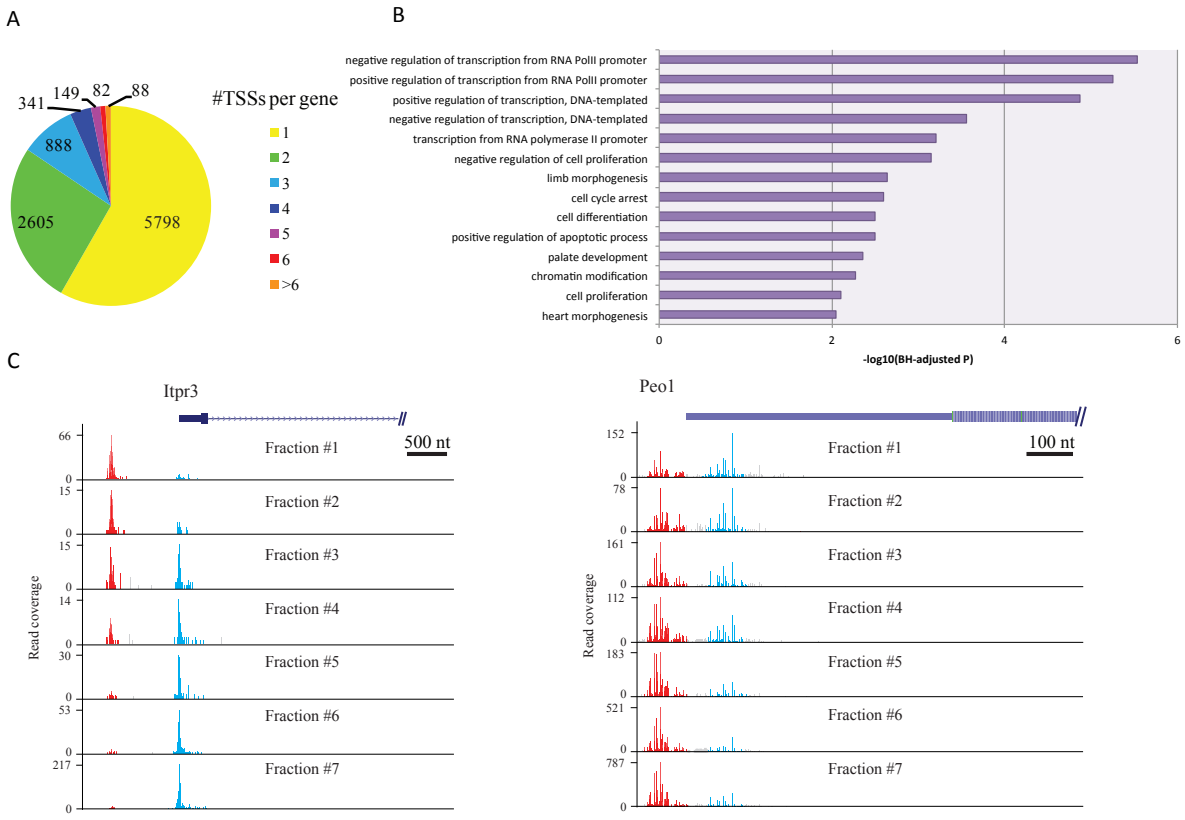

Fig 3.4 4,153 (41.7%) genes expressed multiple TSSs. A. Piechart showing the number of TSSs in the gross TL regions per gene. Out of the 9,951 genes with at least one TSS detected, 4,153 (41.7%) expressed multiple TSSs. B. GO enrichment for multi-TSS genes over all expressed genes. C. Genome browser view showing two representative genes with alternative TSSs. Positions of the distal TSS isoforms and the proximal TSS isoforms are indicated with red and blue arrows respectively.

3.4 Alternative TSSs usage leads to differential TE in 745 genes

For each of the 17,033 TSSs that mapped to the gross TL regions, we estimate relative translational efficiency (TE) by calculating the average number of associated ribosomes based on their normalized sequencing read counts from different gradient fractions (Fig 3.5A, B). To test if our polysome fractionation strategy correctly estimates the translation status of TSS-associated TL isoforms, we compared the TE values obtained from our data to those of published ribosome profiling datasets (Eichhorn *et al*, 2014) and protein synthesis rates based on proteomics measurements (Schwanhäusser *et al*, 2011) in the same cell line. To compute TE for each gene, we combined counts for alternative TL isoforms and then normalized the average number of ribosomes per mRNA against ORF length. These values correlated well between our data and proteomics measurements (r_s=0.46) and even better with that derived from ribosome profiling (r_s=0.57).

To investigate the global impact of alternative TSS usage on translational regulation, for each of the 4,153 genes with multiple TSSs, we compared TE fold changes between any pair of alternative start sites. Due to the potential uncertainty associated with small number of reads derived from low abundant mRNAs, we applied a bootstrapping strategy to estimate the confidence of TE fold changes. In brief, for each of the seven gradient fractions, we sampled mapped reads with replacement to generate a pseudo dataset and repeated the sampling 1000 times. For each of the 1000 bootstrap replicates, TE fold changes between alternative TSS-associated TL isoforms were calculated in the same manner as for the experimental data and the resulting bootstrap distribution was summarized with a mean and a standard deviation. The greater the bootstrap mean deviates from zero, the larger the TE diverges between the two isoforms. By contrast, lower bootstrap standard deviation gives more confidence in the estimation of TE difference (Fig 3.5C). After applying a threshold of Benjamini-Hochberg adjusted P-value < 0.01 and TE divergence > 1.5 in both replicates (FDR = 5.2%), we identified 745 genes exhibiting significant TE difference in 1618 pairs of TL isoforms.

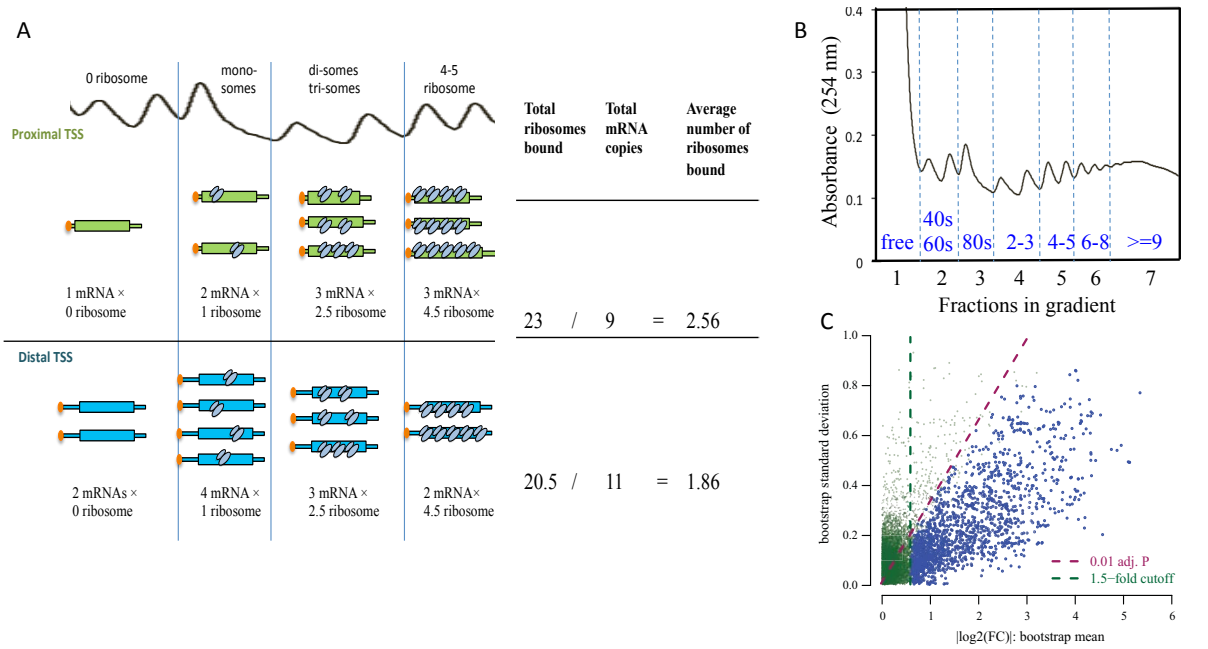

Fig 3.5 Alternative TSSs lead to differential TE in 745 genes. A. Schematic of polysome fractionation, illustrating calculation of translational efficiency for individual TL isoform. In this example, the TL isoform with proximal TSS has larger average number of ribosomes bound (2.5 ribosomes/mRNA) compared to the TL isoform with distal TSS (1.86 ribosomes/mRNA). B. Polysome profiling used to separate mRNAs into seven fractions based on the number of bound ribosomes. C. Scatterplot showing the bootstrap means (x-axis) and standard deviations (y-axis) in estimating TE divergence for 1618 pairs of TL isoforms. Dashed purple line indicated the Benjamini-Hochberg adjusted P-value of 0.01, and dashed green line indicated the 1.5-fold divergence. Genes with significant TE divergence (Benjamini-Hochberg adjusted P value < 0.01, TE bias > 1.5 fold) are depicted in blue. Adapted from Spies *et al,* 2013.

To validate the identified TL isoforms and their associated translational status, we randomly picked four genes with different TE between TL isoforms. To verify whether Cap-profiling identified the exact 5'-ends of mRNAs, we chose an alternative cap-capturing strategy on specific cap-dependent linker ligation (see section Materials and Methods). To simultaneously verify the translational status for each TL isoform of those genes, RNA was extracted from non-ribosomal fractions and polysomal fractions and converted into cDNA separately. After the ligation with unique linkers, cDNA was amplified with gene-specific primers. All PCR products were of the size corresponding to the respective TSS (Fig 3.6). Furthermore, the relative abundance of transcript isoforms in both non-ribosomal and

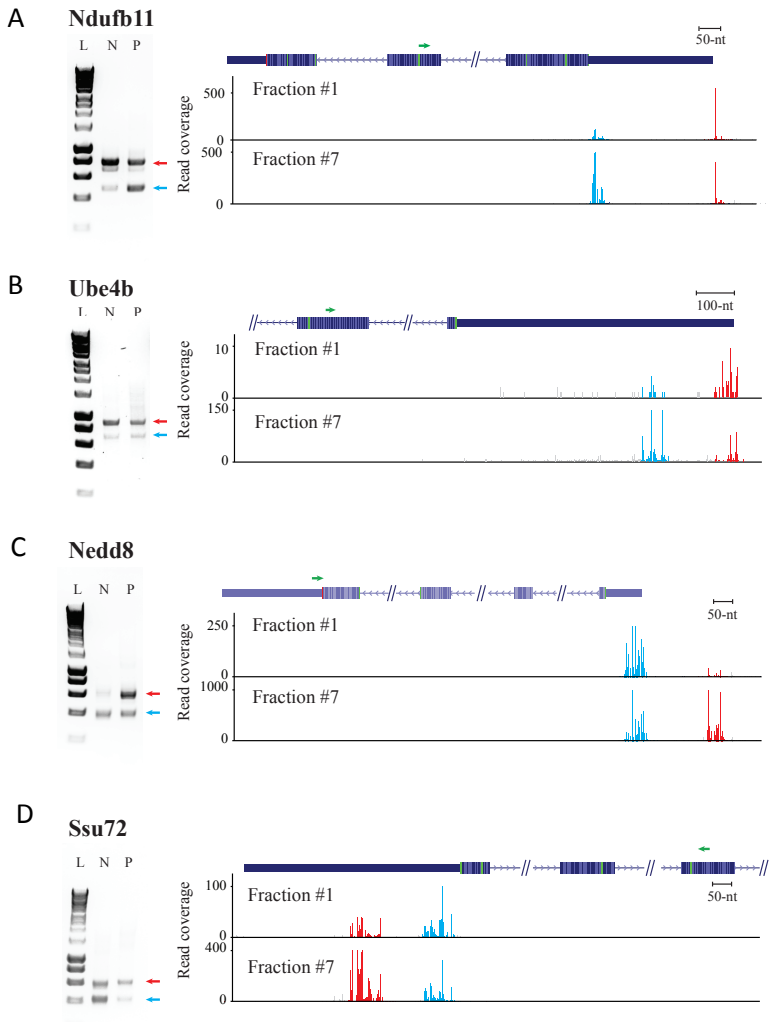

Fig 3.6 Validation of TL isoforms and their translation status in genes *Ndufb11, Ube4b, Nedd8, Ssu72*. Left: 1% agrose gel electrophoresis of amplified products of mRNA 5'ends obtained from non-ribosomal and polysomal fractions. Positions of the distal TSS isoforms and the proximal TSS isoforms are indicated with red and blue arrows respectively. L, HyperLadder I; N, non-ribosomal fraction; P, polysomal fraction. Right: read coverage for the two TSS clusters in gradient fractions 1 and 7, with gene structure showing on the top. The position of reverse PCR-primer used is indicated with green arrow.

polysomal fractions was consistent with that determined by CAPTRE in fraction 1 and fraction 7 from polysome gradient. For example, *Nedd8* expressed two TL isoforms, in the non-ribosomal fraction 1, the shorter isoform was predominant, whereas in the actively-translating fraction 7, the longer isoform was more abundant. This abundance difference in the non-translating and translating pool was supported by our validation experiment (Fig 3.6C). In contrast, in the case of *Ndufb11*, the longer TL isoform was more predominant in the fraction 1, whereas the shorter TL isoform was more abundant in the fraction 7, indicating that the shorter isoform exhibited higher translational efficiency (Fig 3.6A).

3.5 Longer TL isoforms tend to have lower TE

To decipher the rules by which TL isoforms affect TE, we first sought to check for the effect of TL length by only focusing on genes without alternative splicing in their TL regions.

In 6,536 pair-wised TE comparisons between TL isoforms of the same genes, we found a global tendency that longer TLs were associated with lower TE by plotting the relative TE for long and short TLs for each comparison (Fig 3.7A). Intriguingly, as shown in Figure 3.7B, with the increase in length difference between isoforms, the fraction of genes showing TE divergence also increases. This could be explained by the likelihood that with

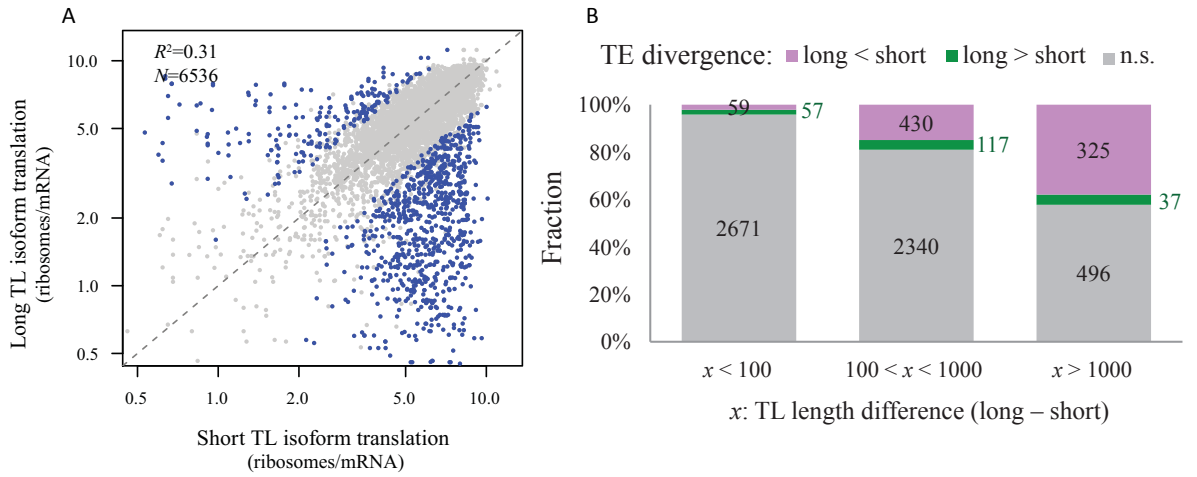

Fig 3.7 Longer TL isoforms tend to have lower TE. A. Longer TL isoforms by and large translated less efficiently than shorter TL isoforms of the same gene. The number of ribosomes per mRNA between shorter TL isoforms (x-axis) and longer TL isoforms (y-axis) was correlated far from perfect (R^2=0.31), indicating a large TE diversity between TL isoforms. Blue dots were isoform pairs with significant differential TE. B. Larger the length difference between the two isoforms, the higher the fraction associated with significant TE divergence. In addition, with the increase of length difference, the trend became more prominent that longer TL isoforms tend to have lower TE.

the increase in TL length difference, more *cis*-elements in the divergent part can be exclusively used by the long TL isoform. Among the 1025 pairs of significant differential TE for long/short TLs, nearly 80% (814) showed a bias towards lower TE for longer TLs, suggesting that TL sequences in general comprised of more translational repressive elements than enhancing ones.

3.6 Alternative TL sequences are sufficient to confer the TE divergence between TL isoforms

To further exclude the effect from other mRNA features that may also influence translation and to directly examine whether the sequence of TL isoforms are sufficient to confer the observed TE divergence, we used an *in vivo* reporter system to compare the TE of a *Renilla* luminescent reporter gene led by the long and short TL isoforms from eight genes, respectively (Fig 3.8A). Here, TE is calculated as the reporter gene's luciferase activity normalized against the corresponding mRNA abundance measured by RT-qPCR. For each of the eight genes, sequences of different TL isoforms were inserted immediately upstream of

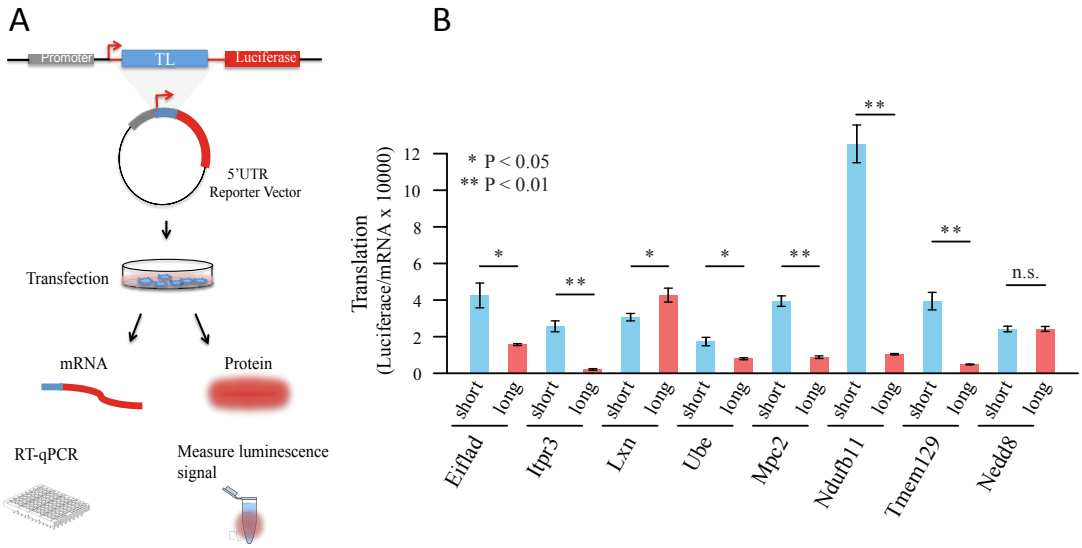

Fig 3.8 TL sequences are sufficient to cause the TE divergence between TL isoforms. A. Schematic of the experiment. TL sequences (blue) were inserted downstream of the promoter (grey) and upstream of the start codon of the luciferase gene (red). After transfection of the reporters, mRNA levels of the luciferase gene were measured by RT-qPCR and in parallel, protein levels were measured by luminescent signal. B. TE is calculated by luciferase activity normalized by mRNA abundance. Seven out of eight genes showed significant TE divergence between TL isoforms. Barplots with error bars, mean ± SE; n=3. * P < 0.05, ** P < 0.01 (student's *t*-test).

the start codon of the luciferase gene, resulting in reporter gene constructs shared the same ORFs and 3'UTRs but only differed in their TLs. Seven out of the eight genes showed TE

biased towards the same TL isoform as observed using CAPTRE (Fig 3.8B). Notably, the shorter TL isoform from *Ndufb11* resulted in TE eleven times higher than the longer isoform (Fig 3.8B), indicating that alternative TLs can lead to impressive differences in TE. Taking together, the reporter assay demonstrated that TL sequence alone was sufficient to confer translational difference between TL isoforms *in vivo*.

3.7 Sequence features associated with TE difference among TL isoforms

3.7.1 uORFs and out-of-frame uAUGs reduce TE

uORFs and uAUGs have been reported to negatively affect the translation of the main ORFs (see section 1.4.2). To test whether the presence of uORFs between the divergent part of TL isoforms contributed to the observed TE difference, we first separated the TL isoform pairs into two groups, one group containing 940 with uORFs in the divergent TL sequences, and the other 1,874 pairs without. Comparing the distribution of TE difference between the two groups, we observed that the presence of uORFs led to larger TE decrease in longer TL isoform (Fig 3.9A). A previous study reported that the number of uORFs is associated with the degree of translation inhibition (Calvo, 2009). Consistently, we observed that the more uORFs are present in sequence regions specific to long TLs, the larger the TE diverged between TL isoforms (Fig 3.9B).

In our previous work (see section 2.6), we observed that out-of-frame and in-frame uAUGs conferred different effects on translation. While out-of-frame uAUGs tend to decrease TE, in-frame uAUGs have no significant impact. Applying the same analyses as described above for uORFs, we tested the effect of the two uAUG subtypes on TE separately. Consistent with our previous finding, the presence of out-of-frame uAUGs, but not in-frame uAUGs within the divergent TL sequence led to decreased TE of the long TL isoforms (Fig 3.9C).

Encouraged by the above findings, we further restricted our analyses to uORFs/uAUGs that were used under the same cellular context. We collected publically available ribosome profiling data (Shalgi *et al*, 2013) and used ribosome footprints generated from initiating ribosome profiling of harringtonine pretreated cells (see section 1.7.3) to map uORFs/uAUGs used in our cells (see Materials and Methods). Using the ORF-RATER approach (Fields *et al*, 2015), we generated a list of 163 canonical uORFs (started with AUG) and 9 out-of-frame uAUGs. Following the analyses described above, we revealed the same

regulatory tendency for uORFs and out-of-frame uAUGs (Fig 3.9D), confirming that uORFs and out-of-frame uAUGs indeed suppressed translation.

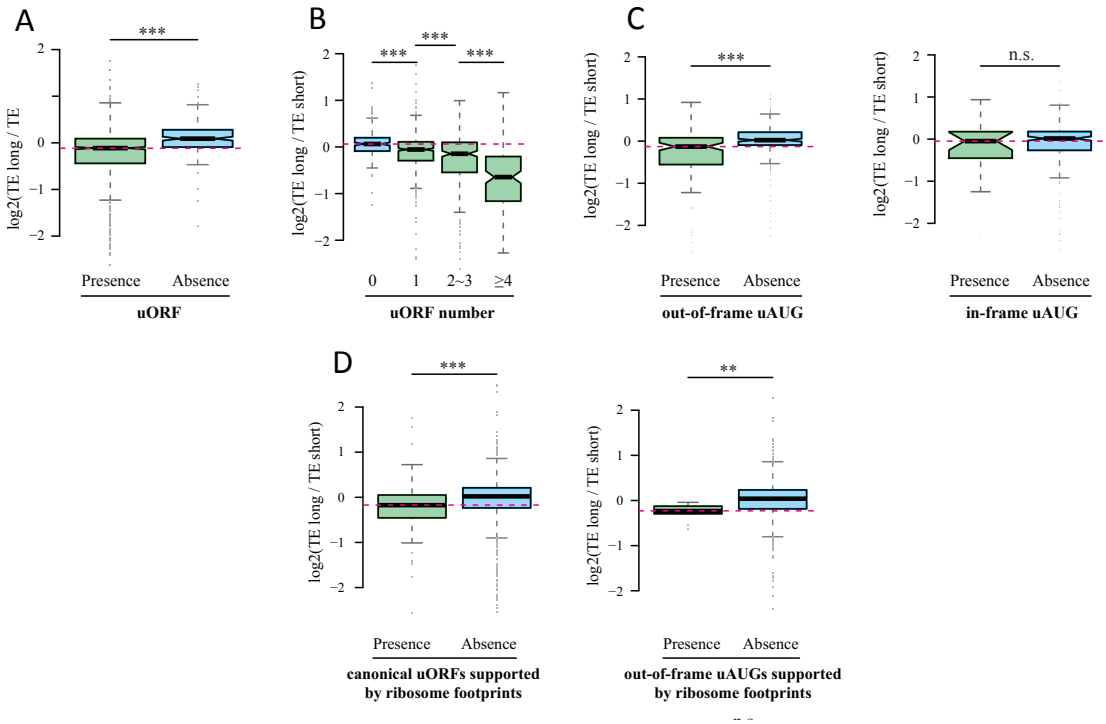

Fig 3.9 uORFs and out-of-frame uAUGs reduce TE. A. Boxplots comparing the log2 TE fold changes between two groups of long and short TL isoform pairs, one group with at least one uORF present in each of the isoform-divergent parts and the other without. B. The presence groups in A were further split into three subgroups according to the number of uORFs present. C. Left: Same to A, but the sequence feature of interest is out-of-frame uAUGs. Right: Same to A, the sequence feature of interest is in-frame uAUGs. D. Left: Same to A, the sequence feature of interest is canonical uORFs supported by ribosome footprints. Same to A, the sequence feature of interest is out-of-frame uAUGs supported by ribosome footprints. In all panels, ** $P < 0.01$, *** $P < 0.001$; Mann–Whitney U test.

3.7.2 Cap-adjacent RNA secondary structures decrease TE

Stable RNA secondary structures in TLs were shown to be capable of diminishing translational initiation *in vitro* (Kozak, 1989). It has also been reported that stable RNA secondary structures are embedded in 5'ends of TLs for the majority of proto-oncogenes, transcription factors and growth factors, whose expression is tightly controlled (Davuluri, 2000). To examine whether stable RNA secondary structures also contribute to the observed TE difference between TL isoforms, we calculated and compared the minimum free energy (MFE) between their TL sequences. Out of 2,185 isoform pairs, 28 and 24 were found with only their long or short TLs having stable structures (MFE < -30kcal/mole for 50nt RNA fragments) immediately after 5'-cap, whereas 2,133 pairs with stable structures in both or

neither of the TL isoforms. Compared to mRNAs with stable structures in both the long or short TLs or neither of them, transcripts with strong RNA folding near the 5'-cap showed reduced TE in the structured TL isoforms specifically for either the long or the short TLs (Fig 3.10A). Thus, our results suggest that stable RNA secondary structure at 5'-cap tends to decrease TE *in vivo*, mostly likely by inhibiting the entry of ribosomal 43S pre-initiation complex (Gray & Hentze, 1994).

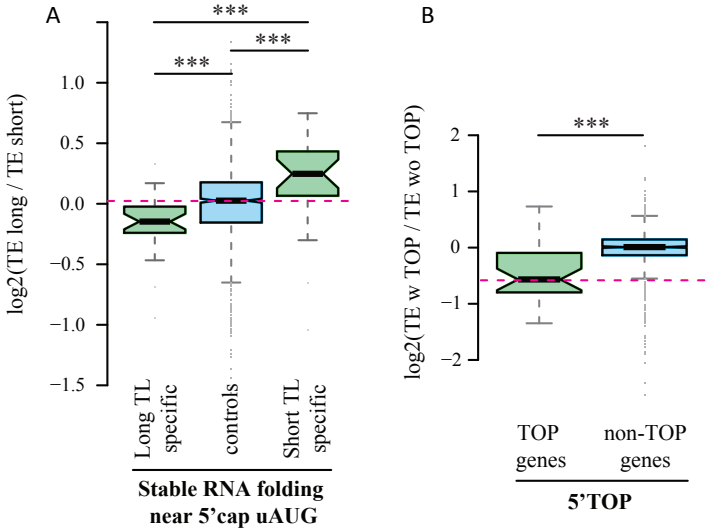

Fig 3.10 Roles of cap-adjacent stable RNA structures and 5'TOP sequences in translational regulation. A. Boxplots comparing the log2 TE fold changes between three groups of TL long and short isoform pairs, the first group with 5'cap-adjacent stable RNA secondary structures present only in long TL isoforms, the second group with 5'cap-adjacent stable RNA structure present/absent in both isoforms, and the last group with 5'cap-adjacent stable RNA structure present only in short TL isoforms. B. Boxplots comparing the log2 TE fold changes between TOP genes and non-TOP genes. For TOP genes, the TE fold changes were the ratios of isoforms with TOP sequences present over isoforms without TOP sequences, and for non-TOP genes, isoforms were randomly assigned as numerators and denominators. *** $P < 0.001$; Mann–Whitney U test.

3.7.3 RNA 5' terminal oligopyrimidine tract (5' TOP) reduce translation

Another sequence feature at the 5' ends of an mRNA is 5' TOP sequences, a highly-conserved *cis*-elements in translational regulation (see section 1.4.6). We examined 166 known TOP genes from the literature (Thoreen *et al*, 2012; Hsieh *et al*, 2012), of which 33 genes expressed multiple TSSs and with one TSS harboring 5'TOP sequences (C followed by at least 4 pyrimidines). Comparing to isoforms of the same genes without TOP sequences, we found that TL isoforms with TOP sequences translated significantly less efficiently, when using non-TOP genes as controls (Fig 3.10B). Given that our data was generated from cells

under normal growth condition, the phenomenon observed here suggests that TOP sequences may to some extent repress translation even when the cellular growth is unperturbed.

3.7.4 Sequence motifs associated with translational repression

For the systematic discovery of potential sequence elements that affect translational regulation, we extended the investigation by correlating the appearance of all possible hexamers within TL divergent regions to observed TE differences. As AUG-containing hexamers may reflect the presence uORFs or uAUGs, we excluded them from further analyses. We found 137 hexamers negatively correlated with TE difference with Benjamini-Hochberg -corrected P-value < 0.01. Intriguingly, AAAAAU matched to the binding sites of PABPC1, a cytoplasmic poly (A) binding protein, which typically binds to 3' poly (A) tail of eukaryotic mRNAs (Paz *et al*, 2014). Interestingly, binding of PABPC1 to an adenine-rich elements in its own TL has been shown to inhibit its translation (de Melo Neto *et al*, 1995; Melo *et al*, 2003). To validate the regulatory role of other hexamer motifs, two repressing motifs (AAUCCC and CAAGAU) were inserted into the TL region of a *Renilla* luminescent reporter construct with five copies of the respective hexamer sequence (Fig 3.11A). Using a similar approach as described above (see section 3.6) for comparing long and short TLs, we determined the TE of these hexamer constructs. As illustrated in Figure 3.11 B, C, the two motifs indeed reduced TE compared to control constructs with reverse complementary sequences or randomly shuffled sequences.

3.8 Quantitative models predict approximately 60 % of the TE divergence in TL isoforms

To further understand the relative contribution of these *cis*-elements to the observed TE divergence, alone or in combination, we trained non-linear regression models, first with individual sequence features separately, and then in combination. As shown in Figure 3.12A, the number of uORFs in the divergent part between long and short TLs and their length

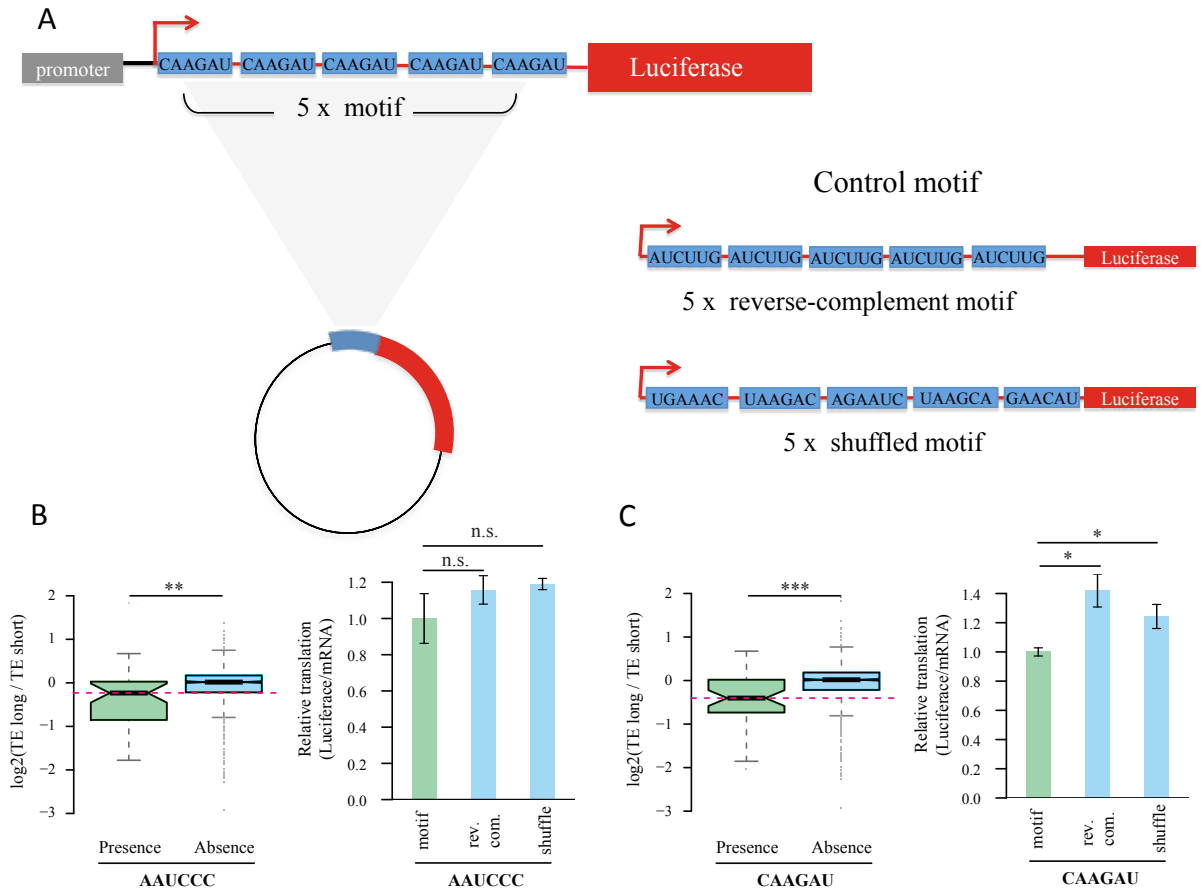

Fig 3.11 Sequence motifs associated with translational repression. A. Schematic of the experiment. B. Left: Boxplots comparing the log2 TE fold changes between two groups of TL long and short isoform pairs, one group with the motif AAUCCC present in long isoform-specific regions, and the other without. Right: luciferase assay comparing the relative TE between reporter genes with five copies of motif AAUCCC, reverse-complementary of motif AAUCCC, and randomly shuffled sequences in their TLs. (n=3; mean ± SEM; n.s., not significant). C. Similar to (B), motif is CAAGAU. (n=3; mean ± SEM; * $P < 0.05$; student's t-test).In boxplots, ** $P < 0.01$, *** $P < 0.001$; Mann–Whitney U test.

difference were the two best single predictors for TE difference, which explained 35.5% and 35.1% of its variance respectively. Out-of-frame AUGs and stable RNA secondary structures near 5'caps had less predictive power, yet explaining a fraction of observed difference of 3.7% and 3.5 %, respectively (Fig 3.12A). This mild influence might be explained by their limited occurrence in TL-divergent sequences. To understand the combinatory contribution of all sequence features investigated, we trained multi-variable regression models by using multiple features as the predictors. Using 29 features including uORFs, uAUGs, TL length, RNA secondary structures, TOP sequences and hexamers, the model explained 57% the variance of observed TE difference (Fig 3.12B).

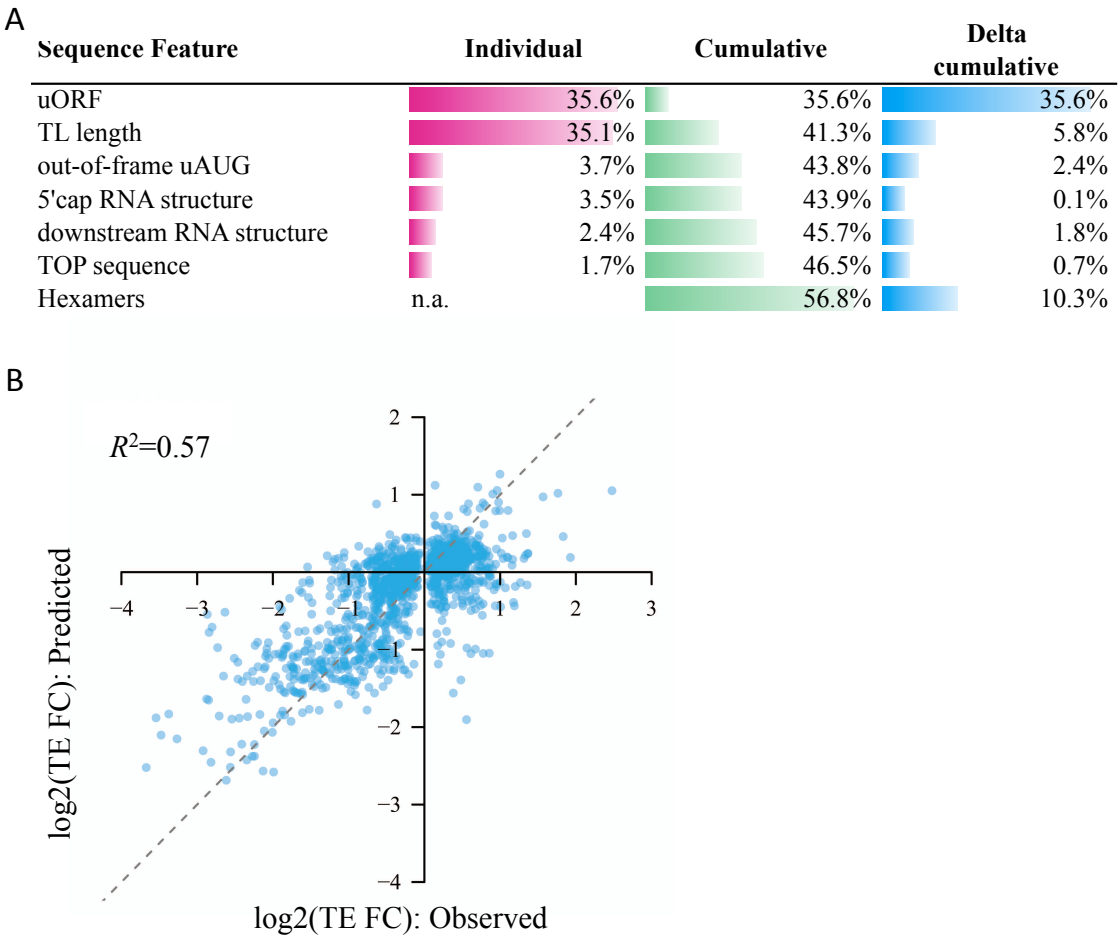

Fig 3.12 Quantitative models predict approximately 60% of the TE divergence in TL isoforms. A. Barplots showing the individual and cumulative contribution for sequence features in explaining the TE difference between TL isoforms. B. The combinatory non-linear regression model based on all sequences features investigated in this study explained 57% variance of TE difference between TL isoforms.

3.9 Summary

We developed a quantitative method to investigate the translational status of different TL isoforms. CAPTRE (CAp Profiling of TRanslational Efficiency) combines polysome fractionation and a 5'end sequencing strategy based on cap-trapper, which together constitutes a novel method with unprecedented power to accurately measure the translation status of TL isoforms. Applying CAPTRE to NIH 3T3 murine fibroblast cells, we captured more than 4000 genes expressing multiple TSSs, and in 745 (~18%) of the multi-TSS genes alternative TL isoforms led to significantly differential translational efficiency. Using this data we demonstrated that *cis*-elements such as uORFs/uAUGs, cap-adjacent stable RNA secondary structures and 5'TOP sequence had significantly negative impact on translation.

Furthermore, we identified several novel sequence motifs that can significantly reduce translational activity. Finally, with statistical modeling, close to 60% of the variance in translational activity changes between TL isoforms can be explained using only the *cis*-elements we identified.

4. DISCUSSION

4.1 Emerging importance of TL choice on translational regulation

Although the majority of known *cis*-regulatory elements that act as translational regulators are located in TL sequences, in the recent years, most studies on translational regulation have focused on 3'UTR sequences exclusively. In a recent genome-wide study, the translational status of mRNAs with distinct 3'UTR isoforms was compared. Importantly, the 3'end choice was shown to have only a limited impact on translation, with proximal and distal tandem 3' UTR isoforms being often translated with similar efficiencies (Spies *et al*, 2013). This observation suggests that mRNA-specific translational regulation might occur primarily through *cis*-features located elsewhere in the transcripts, e.g., TLs or CDS. Indeed, we observed a weaker correlation between translational efficiency of long and short TL isoforms of all identified genes (R^2=0.31, n=6536) (Fig 3.7A), when compared to that of distal and proximal tandem 3'UTR isoforms in the same cell line (R^2=0.58, n=4298) (Spies *et al*, 2013), suggesting that TL regions have a greater impact on translation than 3'UTRs.

In the first part of this thesis we observed that the SNP density in TLs rather than that in 3'UTRs or CDS, showed the strongest association with allelic translational efficiency divergence in an F1 hybrid mouse model system. This suggests a potential regulatory importance of this region. The role of TLs in regulating translation was underestimated in previous studies, which may be attributed to the imprecise and incomplete annotation of 5'ends of many mRNA transcripts (Vogel *et al*, 2010; Floor & Doudna, 2016). To follow up on these findings, we developed a novel technology to globally annotate the 5'ends of all expressed genes and assembled the TL isoforms of all protein-coding genes in the second part of this thesis. Based on our own 5' end annotation, we compared the translational efficiency between TL isoforms and found that around 40% of protein-coding genes with multiple TSSs have different translational efficiency between isoforms. The exact numbers we report in this study may underestimate the actual effect size. This is because the resolution of sucrose density gradients is limited for the most-actively translated mRNAs that have the highest number of ribosomes bound to them. Thus, the exact number of ribosomes bound to highly ribosome-loaded mRNAs cannot be accurately estimated, which results in a compromised resolution for estimating translational efficiency divergence (Ingolia et al, 2012). Therefore the actual number of affected genes and the corresponding effect size in translational efficiency of TL isoforms may be even more profound than reported here.

Moreover, translational regulation is primarily attributed to the step of initiation. However, regulation at the level of initiation may not always explain the full extent of regulation (Shalgi *et al*, 2013). The current estimate of translational efficiency based on ribosome association fails to take into account the translation elongation rate. In theory, slow elongation rate and fast initiation rate are both reflected in a high degree of ribosome occupancy. However, both effects will result in opposing protein synthesis efficiencies. To this end, a more direct estimate of translation can be achieved by combining polysome /ribosome profiling with other technologies, such as measuring the newly synthesized proteins using mass spectrometry based proteomics.

Another feature of eukaryotic transcripts that was not investigated in this study, but may contribute considerably to translational regulation, is alternative splicing within TLs. Approximately 30% of human transcripts contain introns within their TLs, which is much more frequent than the occurrence of introns in 3'UTRs (roughly 10%) (Pesole *et al*, 2001). Consequently, alternative splicing within TLs is estimated to affect 20% of genes in the mammalian transcriptomes, compared to only 4% that are estimated be affected by alternative splicing in 3'UTRs (Modrek, 2001). Splicing in TLs is often coupled with alternative usage of promoters, which in turn results in alternative 5' boundaries of mRNAs, thus further diversifying the sequence space in TLs available for translational regulation. Future studies that fully dissect the contribution of splice isoforms may provide better insights into the impact of TLs on translation.

In addition to a change in protein production rates, transcripts derived downstream of canonical translation initiation sites can also yield variant protein isoforms. Based on our CAPTRE data, we found that 502 TSSs were associated with heavy polysome fractions (≥4 ribosomes), and 71 of them contained downstream translation initiation sites that were further supported by ribosome occupancy (Fig 4.1). Collectively, these observations suggest that transcripts led by these downstream TSSs were active in translation, presumably yielding N-terminally truncated proteins. Given that the N-terminus of proteins is often essential for proper protein function and/or their cellular localization (Chen *et al*, 2002; Arce *et al*, 2006; Zhang *et al*, 2015), protein isoforms generated by alternative TSS can serve as important regulatory mechanism for protein localization and functions.

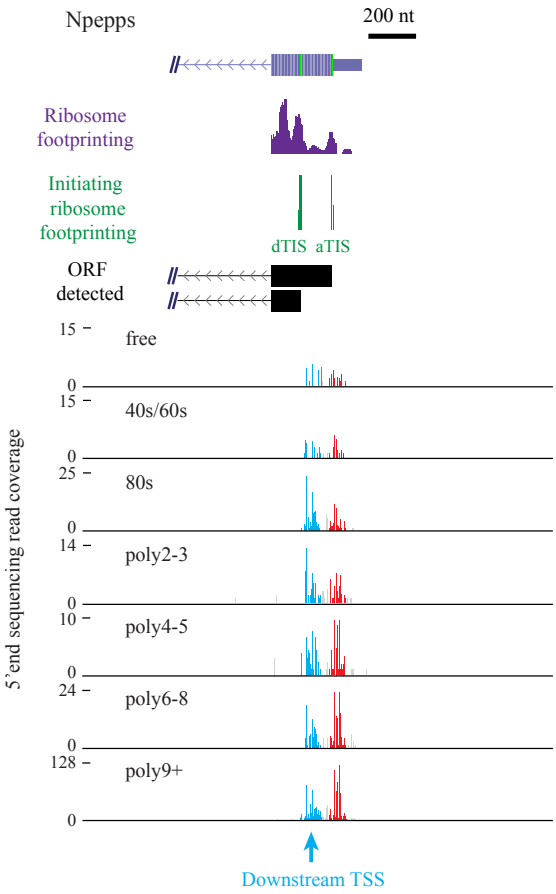

Fig 4.1 Example of alternative TSSs for protein N-terminal changes. Downstream TSSs could lead to N-terminal truncated proteins in gene *Npepps*. Cumulative reads from the seven gradient fractions were plotted under the gene structure. dTIS, downstream translation initiation site; aTIS, annotated translation initiation site.

4.2 Revisiting the regulatory roles of known *cis*-elements in translation

Several sequence features we identified in this work associated with allele- or isoform-specific translational efficiency divergence have previously been implicated in translational regulation. Despite this prior knowledge, our analyses still provided novel insights into the mechanisms of mammalian translational regulation.

In the first part of this thesis, we found that in murine fibroblasts sequence variants affecting local RNA secondary structures that surround translation initiation sites, influence allelic divergence in translation. This observation is largely in agreement with previous findings in yeast (Shah et al, 2013; Dvir et al, 2013; Muzzey et al, 2014). Several recent genome-wide surveys of RNA secondary structures have found that in *Saccharomyces cerevisiae, Arabidopsis thaliana* and human cell lines, RNA fragments in the vicinity of start

codons do not tend to form stable secondary structures (Kertesz et al, 2010; Wan et al, 2012, 2014; Ding et al, 2014). It is possible that the structure-less context near start codons is required for efficient ribosome assembly. Given that most of the predicted structural alterations between alleles are caused by single nucleotide polymorphisms (SNPs) (data not shown), it is tempting to speculate that individual SNPs are sufficient to alter local RNA structure to an extent that ribosome assembly can be influenced. Indeed, a recent study investigating the variations of RNA secondary structures in a human family trio (mother, father and child), reported that over 1900 transcribed single nucleotide variants (approximately 15% of all transcribed single nucleotide variants) alter local RNA structures (Wan *et al*, 2014).

RNA secondary structures can also influence translation when positioned entirely within TLs. We observed that stable RNA structures (MFE < -30kcal/mole) when present in immediate proximity to the cap (within 50 nt), were associated with translational repression. When positioned further downstream from the cap, RNA structures with slightly higher stability appeared to have diminished repressive effects (Fig 4.2). Similar observations were reported in a previous *in vitro* study (see section 1.4.5), thus substantiating our findings. Interestingly, the observed effects might be explained by two different underlying mechanisms. While moderately stable cap-proximal RNA structures are sufficient to block 43S ribosome entry, the downstream secondary structures that function by stalling scanning ribosomes might require higher stability to not be disrupted by the 43S ribosome, although the exact mechanisms are still unknown (Gray & Hentze, 1994).

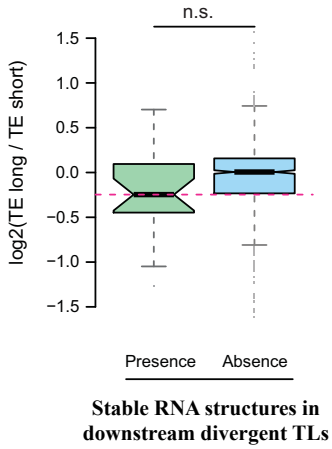

Fig 4.2 Boxplots comparing the log2 TE fold changes between two groups of TL isoform pairs, one group with stable RNA secondary structures (MFE < -35kcal/mole in any 50-nt RNA fragments) present in the downstream divergent TL region, and the other without.

Among all the features we identified in TLs, the presence of uORFs can best predict the translation divergence between isoforms. Although the best studied examples of uORFs in the mouse *ATF4* (see section 1.4.2) and the yeast *GCN4* genes (Hinnebusch, 1997) are known to positively regulate translation from the downstream main ORFs under stress conditions, our study revealed that under normal conditions, uORFs more frequently repress translation. We also observed that mRNAs with uORFs are still associated with actively translating polysome, indicating that in mammalian cells uORFs are generally permissive for translation reinitiation at downstream start codons. This is in contrast to uORFs in yeast, which have been shown to more frequently block reinitiation at the downstream ORFs (Somers *et al*, 2013). Interestingly, while out-of-frame uAUGs negatively affect TE, we did not observe such an effect for in-frame uAUGs, possibly because the latter more frequently generate N-terminal extended protein isoforms (Medenbach *et al*, 2011; Dvir *et al*, 2013; Kozak, 2005).

As a result of the recent advent of ribosome profiling as a generally applicable methodology, an unprecedented prevalence of non-canonical translation initiation sites has been revealed (Ingolia *et al*, 2011; Lee *et al*, 2012; Fritsch *et al*, 2012). Surprisingly, some of these studies found that the number of uORFs with non-canonical start codons even exceeded the number of uORFs with canonical ones (Ingolia *et al*, 2011; Fritsch *et al*, 2012). By combining initiating ribosome profiling in cells pre-treated with harringtonine (see section 1.7.3) with our translational efficiency measurement, we found that although canonical uORFs and uAUGs reduce translation from downstream ORFs, such effects were neither observed for uORFs led by non-canonical start codons (CUG, GUG or UUG) (Fig 4.3A), nor for out-of-frame non-canonical upstream start codons (Fig 4.3B). This result is consistent

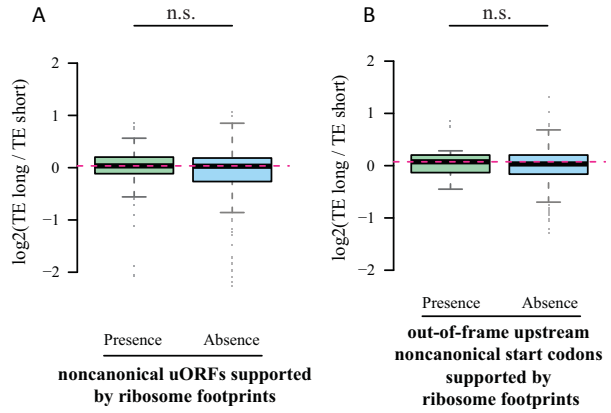

Fig 4.3 Upstream translation started at non-canonical start codons. A. Boxplots comparing the log2 TE fold changes between two groups of alternative isoform pairs, one group with at least one uORF with non-canonical start codons supported by ribosome footprints and the other without. B. Same as A, but the sequence feature of interest is out-of-frame upstream non-canonical start codons

supported by ribosome footprints.
with previous observations that non-optimal Kozak context at the first start codon can promote ribosome bypass and thus reach the start codon further downstream, a mechanism referred to as “leaky scanning” (Kozak, 2005).

TL length is another feature we identified as critical for translational regulation. Specifically we showed that longer TL isoforms tend to have lower translational efficiency. One possible explanation for this observation is that the length of TLs is to some extent correlated with the number of *cis*-regulation elements, which by and large play a repressive role in translational regulation. Importantly, in our multi-variable models where the effects of known *cis*-elements were excluded, the length of TL still correlated with translation divergence between isoforms. Despite our multilayered analyses, it is likely that there are other TL *cis*-elements that remain uncharacterized (see section 4.3). Specifically, whether the length of a TL sequence *per se* could also influence translation, i.e., 43S subunits scanning, is still unknown.

4.3 Other *cis*-elements in translational regulation

To estimate the relative contribution as well as the prediction power of each specific sequence element to explain differences in translational efficiency between TL isoforms, we built a multi-variable non-linear regression model. Using this model, we were able to explain nearly 60% of the observed variation between TL isoforms with the regulatory impact of the specific *cis*-elements we identified in TLs. The remaining unexplained variations could come from other sequence features in this region, particularly IRESs, which were not investigated in our analyses due to the lack of common sequences or structural motifs shared among the currently identified cellular IRESs (see section 1.4.7).

In the first part of this study (using F1 hybrid mice as a model system), we did not find significant impact of several known *cis*-acting features, such as the number of miRNA binding sites and codon usage bias, on ADTE. We reasoned that observed ADTE might largely result from a combined effect of several distinct regulatory mechanisms acting together to achieve regulatory outcomes. Therefore, the contribution of individual features with small individual effect size, may not be sufficient to reach statistical significance. It is important to consider that miRNA-mediated gene regulation can occur through pathways that lead to mRNA degradation as well as translational repression (Bartel, 2009) (see section 1.4.4). However many recent studies only show a modest influence on translational efficiency

(Guo *et al*, 2010; Mukherji *et al*, 2011; Eichhorn *et al*, 2014). Furthermore, our study uses computational algorithms to predict miRNA target sites, which may suffer from low biological accuracy that is inherent to the prediction methodology or used parameters. It has been shown that at most 60-70% computationally predicted miRNA target sites are functionally relevant in a biological context (Lewis *et al*, 2003; Selbach *et al*, 2008).

Furthermore, optimal codon context can have regulatory effects by modulating translational elongation. Thus codon optimality may not influence mRNA-polysome association and hence is not susceptible to our translational efficiency measurement using polysome profiling (Tuller *et al*, 2010; Novoa & Ribas de Pouplana, 2012; Presnyak *et al*, 2015). On the contrary, increasing codon adaptation will lead to faster elongation that in theory reduces ribosome occupancy. In this regard, ribosome association is inferior to measuring the abundance of newly-synthesized protein in estimating translational efficiency (Shah *et al*, 2013). These considerations may further help to explain why several previous ribosome profiling studies also failed to detect differences in translational efficiency as a result of optimal and non-optimal codon usage (Ingolia *et al*, 2009; Qian *et al*, 2012; Charneski & Hurst, 2013). Moreover, since the correlation between codon usage and translation was mostly described in *E.coli* and yeast, this effect may not readily apply in multicellular organisms. Indeed, it was suggested that in humans, codon usage bias could be more influenced by GC content and RNA secondary structure rather than by the number of available tRNA genes (Chamary & Hurst, 2005; Vogel *et al*, 2010). Another possibility is that in studies where large gains in protein production upon optimizing transgene codon adaptation are reported, the transgenes are usually overexpressed and consume a large fractions of cellular free ribosomes (Shah *et al*, 2013). The increased codon adaptation may help to release ribosomes engaged on these transgenes and in turn increases their translation initiation by increasing the pool of free ribosomes. However, endogenous genes are normally expressed to a level that accumulates below 1% of the transcriptome and therefore is unlikely to have an overall effect on the pool of free ribosomes upon optimizing codon adaptation (Shah *et al*, 2013).

Albeit the importance of the Kozak sequence in translation is well recognized, we could not assess its regulatory potential in translation either by using F1 hybrid mice or by comparing different TL isoforms. The inherent challenge is that with increased essentiality of a sequence feature, the conservation of this sequence among species also increases, resulting in too few allelic variations in this sequence feature to achieve statistical significance in our study. In fact, we found that the third nucleotide upstream of the start codon (position -3),

which typically is expected to be a purine (A or G for optimal translation), was indeed a purine on both alleles for approximately 90% of the genes. Therefore, a complete modeling of how *cis*-elements regulate translation still requires complementary methods, such as systematic mutagenic reporter systems (Dvir *et al*, 2013).

4.4 Translation in a cap-independent manner

The current understanding of cap-independent translation is largely limited to its role during stressed cellular conditions, where cap-dependent translation is compromised, for example, during viral infections or diseases. However emerging evidences demonstrate that cap-independent translation may as well be employed under conditions when the cap-dependent translation machinery is still intact (Du *et al*, 2013; Xue *et al*, 2014).

Several recent studies that utilize large-scale systematic screening approaches have identified tens of thousands of sequences that could serve as translation initiation signals. Using an *in vitro* mRNA display method, Wellensiek et al. identified more than 12,000 translation enhancing elements that initiate translation in a cap-independent manner (Wellensiek *et al*, 2013). More recently, applying a bicistronic assay combined with fluorescence-activated cell sorting and deep-sequencing (FACS-seq), Segal and colleagues identified and characterized thousands of human and viral sequences with *in vivo* cap-independent translational activity and expanded the set of IRES sequences known to date by approximately 50-fold (Weingarten-Gabbay *et al*, 2016). These two studies together revealed that a large fraction of genes may utilize cap-independent translation and suggested a previously underestimated functional significance of cap-independent translation.

Intriguingly, an internal RNA modification, namely *N6*-methyladenosine (m^6A) has recently been reported in an *in vitro* study to initiate translation in a cap-independent manner under stress cellular condition (Meyer *et al*, 2015). The enhanced m^6A level within transcript leader sequences of a group of heat shock response genes was shown to be responsible for their increased translation *in vivo*, thus ensuring the appropriate cellular response for stimulation (Zhou *et al*, 2015). However, this proposed m^6A-based initiation mechanism differs fundamentally from IRES-driven translation, as a free 5' terminus and ribosome scanning are still required (Meyer *et al*, 2015). Moreover, the generality and magnitude that m^6A may exert on regulating translation, particularly translation initiation is largely

unknown. In this context, it will be of great interest to show whether m^6A status also has an impact on translational efficiency under normal cellular conditions.

4.5 *Cis*-regulation under different cellular conditions

A recent publication compared isoform-specific translational regulation in multiple cell types and reported that 5'UTR sequences conferred robust translational regulation, while 3'UTRs seemed to exert cell-type specific effects (Floor & Doudna, 2016). This is largely consistent with our observations that most *cis*-regulatory elements identified in both studies of this thesis have regulatory capacities that enable them to directly interfere with the translational machinery, the activity of which is relatively more stable across different cell types under normal growth conditions. In contrast, RNA binding proteins and miRNAs, which are mostly 3'UTR regulators, may have large fluctuation in their activities and abundance among different cell types and therefore result in cell-type specific patterns of 3'UTR regulation. We thus reason that our models for explaining TL isoform translational efficiency divergence could be largely generalized to other cell types under normal conditions.

However, due to the altered activity of the translational machinery during conditions of cellular stress (Spriggs *et al*, 2010), the activity of certain *cis*-elements (i.e. uORFs) may be altered. In some extreme cases, the mode of translation initiation of mRNAs can indeed switch from a cap-dependent to a cap-independent mechanism (see section 1.4.2 and 1.4.7). Finally, the usage of alternative transcription start sites also displays a highly-dynamic and cell type-specific regulatory scheme (Forrest *et al*, 2014) that in turn complicates the impact of TL isoforms on translation. Therefore, simultaneous measurement of TSS activity and TSS-associated translation across different cellular conditions will be imperative for a better understanding of the underlying mechanisms of translational regulation mediated by TL choice.

4.6 Interplay between eukaryotic gene regulatory steps

Transcription and translation in prokaryotic cells are closely coupled, with translation occurring alongside transcription. In eukaryotes, translation and transcription occur in separate cellular compartments and thus appear to operate independently. Transcription takes place in the nucleus, while translation of the processed transcript occurs only in the cytoplasm. The spatial and temporal separation of transcription and translation renders a

much more complicated and sophisticated mode of gene regulation in eukaryotes by introducing multiple layers of regulatory processes.

However, recent genetic and biochemical analyses are starting to recognize that instead of operating entirely independent, eukaryotic translation and transcription, as well as other levels of gene regulation are extensively coupled, by either tethering together machineries that are responsible for different regulatory processes (Maniatis & Reed, 2002) or by regulating multiple-processes with the same machinery (Komili & Silver, 2008).

Another interesting point of view supporting a potential coordination is that nucleus transcription not only determines the abundance of cellular mRNAs, but also assembles in their untranslated regions various post-transcriptional regulatory elements that can influence their protein production in the cytoplasm. In yeast, mRNAs from many genes involved in the responses to pheromone, nitrogen starvation, and osmotic stress are poorly translated under non-stressed condition. Upon stimulation, however, structural changes in TLs that arose from promoter switch profoundly increases the translational efficiency of these genes (Law *et al*, 2005). This coordinated change in transcription and translation renders a rapid response to environmental stimulation and is especially essential for yeast cellular maintenance and survival under challenging conditions.

In the abovementioned example, *cis*-regulation exerts its function through structural changes triggered by alterations in mRNA sequences. However, since cellular mRNAs typically function as components of mRNPs (messenger ribonucleoproteins), rather than "naked" ribonucleic acids, *cis*-regulatory elements can be further attributed to the protein coats that are associated with mRNAs. In fact, nascent mRNAs are co-transcriptionally assembled into mRNPs, whose composition and structure are highly dynamic and precisely regulated. Transporting from nucleus to cytoplasm, mRNPs mediate a myriad of regulatory steps throughout the entire life cycle of mRNAs. One of the best studied mRNP complexes is the EJC (exon junction complex), which is loaded onto pre-mRNAs in the nucleus during splicing and regulates export and translation in the cytoplasm (Komili & Silver, 2008). A recently unveiled phenomenon that perfectly fits this mRNP-centric model suggests that promoter-dependent downstream RNA metabolism may act by co-transcriptionally assembling the regulatory *trans*-factors onto mRNAs within the nucleus. After exiting the nucleus, mRNAs that carry the pre-loaded *trans*-regulators (e.g. RBPs) further undergo post-transcriptional regulation in the cytoplasm. Importantly, yeast promoter sequences have been shown to direct both mRNA localization and translation during glucose starvation (Zid & O'Shea, 2014). Such crosstalk between different cellular compartments is far from being an

exception. In mammalian cells, the translation elongation factor eEF1A facilitates transcription, nuclear export, and stabilization of Hsp70 mRNA during the heat shock response in addition to its well-defined role in protein synthesis (Vera *et al*, 2014). Yeast upstream activating sequences (UAS), analog to enhancers in higher eukaryotes, have been shown to be capable of regulating mRNA half-lives in the cytoplasm through an unknown mechanism (Bregman *et al*, 2011). It is highly possible that such coordination between different regulatory processes is merely the tip of an iceberg. With the development of genome-wide technologies and integration of various datasets and analyses, our understanding about the interplay between individual components of gene regulatory networks and the global picture of their functional connections will be transformed.

5. MATERIALS AND METHODS

5.1 F1 hybrid mouse fibroblast cell cultures

Female F1 hybrid mice were derived from crossing C57BL/6 J and SPRET/EiJ. Adult mouse fibroblast cells were isolated and cultured according to the protocol from ENCODE project (https://genome.ucsc.edu/ENCODE/protocols/cell/mouse/Fibroblast Stam protocol.pdf) with modification of cell culture medium (RPMI 1640 Medium, GlutaMAX™ Supplement with 10% FBS and 1% Penicillin/Streptomycin).

5.2 mRNA sequencing

Total RNAs from mouse fibroblast cells were extracted using TriZOL reagent (Life Technologies) following the manufacturer's protocol. Truseq Stranded mRNA sequencing libraries were prepared with 500 ng total RNA according to the manufacturer's protocol (Illumina). The libraries were sequenced in 2 x 100 nt manner on HiSeq 2000 platform (Illumina).

5.3 Polysome profiling of fibroblast cells from F1 mice

Mouse fibroblast cells were grown to 80% confluency. Prior to lysis, cells were treated with cycloheximide (100 μg/ml) for 10 min at 37°C. Then cells were washed with ice-cold PBS (supplemented with 100 μg/ml cycloheximide) and further lysed in 300 μl of lysis buffer (10 mM HEPES pH 7.4, 150 mM KCl, 10 mM $MgCl_2$, 1 % NP-40, 0.5 mM DTT, 100 μg/ml cycloheximide). After lysing the cells by passing 8 times through 26-gauge needle, the nuclei and the membrane debris were removed by centrifugation (13,000 rpm, 10 mins, 4°C). The supernatant was then layered onto a 10 mL linear sucrose gradient (10%-50% [w/v], supplemented with 10 mM HEPES pH 7.4, 150 mM KCl, 10 mM $MgCl_2$, 0.5 mM DTT, 100 μg/ml cycloheximide), and centrifuged (36,000 rpm, 120 min, 4°C) in an SW41Ti rotor (Beckman). Fractions were collected and digested with 200 μg proteinase K in 1% SDS and for 30 min at 42°C. RNA from polysome fractions were recovered by extraction with an equal volume of acid phenol-chloroform (pH 4.5), followed by ethanol precipitation. TruSeq

Stranded Total RNA libraries were prepared with 500 ng RNA according to the manufacturer's protocol (Illumina). The libraries were sequenced in 2 x 100 nt manner on HiSeq 2000 platform (Illumina).

5.4 Ribosome profiling of fibroblast cells from F1 mice

Mouse fibroblast cells were cultured and lysed in the same way as for polysome profiling (see above). After lysis, ribosome-protected fragments were collected as described in Ingolia *et al* (Ingolia *et al*, 2012), with minor modifications. In brief, cell lysate was treated with RNase I at room temperature for 45 min. The nuclease digestion was stopped by adding SUPERase • In™ RNase inhibitor (Invitrogen) and then loaded onto a linear sucrose gradient (10%-50%). After ultra-centrifugation, mono ribosome was recovered and RNA was isolated as described for polysome profiling (see above). rRNA was removed using Ribo-Zero™ Magnetic Kit (Human/Mouse/Rat) (Epicentre). The 28-32 nt ribosome-protected fragments were purified through 15% (wt/vol) polyacrylamide TBE-urea gel. The size-selected RNA was end-repaired by T4 PNK for 1 hr at 37°C. The sequencing libraries were then generated using TruSeq Small RNA Sample Preparation Kit (Illumina) and sequenced in 1 x 50 nt manner on Illumina HiSeq2000 platform.

5.5 PacBio sequencing

Starting from 500 ng total RNA or polysomal RNA, reverse transcription (RT) was performed using random hexamer and SuperScript II reverse transcriptase. PCR was followed using 1ul of RT product as template in 50µl of Phusion High-Fidelity DNA Polymerase system (NEB). PCR primers were designed for amplifying the genic region covering ≥3 sequence variants between C57BL/6J and SPRET/EiJ transcripts. PCR program was as following, 30 s at 98 °C, followed by 30 cycles of 10 s at 98 °C, 30 s at 60 °C, and 45 s at 72 °C, and a final elongation of 5 min at 72 °C. The amplified RT-PCR products from total RNA or polysomal RNA were mixed separately. The mixed products were then purified using Agencourt AMPure XP system (Beckman Coulter) and quantified by Qubit HS dsDNA measurement system (Life Technology). These mixed PCR products were sequenced on

PacBio RS SMRT platform according to the manufacturer's instruction. All the primer sequences were listed in Appendix Table S1.

5.6 Polysome profiling of NIH 3T3 cells

Murine NIH 3T3 cells were grown to 80% confluency. Lysate was fractionated as described above. Fractions were manually collected according to the A_{254} peaks that indicate the number of ribosomes. 50 ng fly total RNAs were added into each fraction as spike-in immediately. The collected fractions were first digested with 200 μg proteinase K in 1% SDS for 30 min at 42°C. RNA from each fraction was recovered by extraction with an equal volume of acid phenol-chloroform (pH 4.5), followed by ethanol precipitation.

5.7 Cap-Profiling

3μg total RNA collected from each fraction (see above) were reverse transcribed using random primer (N15-oligo) tailed with 3' part of Illumina TruSeq Universal Adaptor sequence (P5). 5' complete single-stranded cDNAs were captured based on a protocol from Takahashi et al, 2012 with several modifications. In brief, cap structure and 3' ends of all RNAs were oxidized by $NaIO_4$ on ice for 45 min, followed by an overnight biotinylation with a long-arm biotin hydrazide at room temperature. Single-stranded RNA regions that were not covered by synthesized cDNAs including the 3' ends were cleaved by RNase I. The 5' complete cDNAs containing the biotinylated cap site were then captured with Dynabeads® M-280 Streptavadin (Life Technologies). RNAs were hydrolyzed with 50 mM NaOH and single-stranded cDNAs were therefore released from the beads. After ligation with double-stranded 5' linkers with random overhangs (containing 3' part of Illumina TruSeq Universal Adaptor P7), cDNAs were amplified for 18 cycles using Cap Forward Primer (containing P5) and Cap Reverse Primer with barcode embedded. The amplified libraries were sequenced in 2 x 100 nt manner on Illumina HiSeq2000 platform. All the primer and adaptor sequences were listed Appendix in Table S2.

5.8 Validation of TL isoforms and their associated translational status

To validate our findings based on CAP, we used the TeloPrime Full-Length cDNA Amplification Kit (Lexogen) to independently determine the 5' end of capped mRNA. In brief, a gene-specific primer was used to synthesize the complementary DNA (see Appendix Table S3). In the subsequent ligation, a double-stranded adapter with a 5'C overhang allows for an atypical base-pairing with the inverted G of the cap structure. The ligation can only take place if the RT has really reached the 5' end of the mRNA (Lexogen´s unique Cap-Dependent Linker Ligation (CDLL)). After second-strand synthesis the dsDNA was amplified in a 30-cycled PCR using 5' Lexogen primer (FP: 5' – TGGATTGATATGTAATACGACTCACTATAG) and 3' gene specific primers (Appendix Table S3). Amplified products of four genes of non-ribosomal (pool of free ribosomal, 40-60s sub-ribosomal fractions) and polysomal fractions (pool of fractions with at least 2 ribosomes) were loaded onto an agarose-gel (1%).
All the primer sequences were listed in Appendix Table S3.

5.9 Luciferase reporter assay

To validate the effect of TL length on translation, longer and shorter versions of transcript leaders derived from eight genes were PCR amplified from genomic DNAs or cDNAs, if there is an intron within the transcript leader regions. During PCR, an NcoI and a BglII restriction site were introduced to the upstream and downstream of the TL sequences, respectively. Each TL fragment was then inserted into the Multiple Cloning Site of the pLightSwitch_5'UTR vector (Active Motif) downstream of an ACTB promoter and upstream of *RenSP* luciferase reporter gene. All constructs were validated by Sanger sequencing. Plasmids were transfected into 3T3 cells by using Lipofectamine® 2000 Transfection Reagent (Life Technologies) following the manufacturers' instructions. Luciferase assay was conducted using the LighSwitch Luciferase Assay Reagent™ (Active Motif) and the luciferase activity was measured by Infinite® M200 (Tecan) plate reader and normalized by the absorbance of lysate at 260 nm. Total RNA was extracted from the same lysate using TRIzol® LS Reagent (Life Technologies) and Direct-zol™ RNA Kits (Zymo Research) following the manufacturers' instructions. DNA was removed by *in-column* DNase I digestion. RT-qPCR was performed to measure the *RenSP* mRNA level, which was then normalized by the mRNA level of housekeeping gene *ActB*. Translational efficiency of

different constructs was estimated as the normalized luciferase activity divided by normalized *RenSP* mRNA level.

To validate the effect of putative motifs on translational regulation, ~100 nt oligos containing five copies of specific hexmer motif in were synthesized. An AflII site and a BglII site were also included in the 5' and 3' ends. As negative control, the oligos containing the reverse complement sequence and shuffled sequence of the hexamer motifs were used. The test and control oligos were then amplified by PCR. After restriction enzyme digestion, each TL was cloned into the Multiple Cloning Site of the pLightSwitch_5'UTR vector. Translational efficiency of different constructs was measured as described above.
All the primer sequences were listed in Appendix Table S4.

5.10 Initiating ribosome profiling

Mouse 3T3 cells were cultured in the same way as for polysome profiling (see above). Harringtonine was added to cell culture at a final concentration of 2 μg/mL. Cells were incubated at 37°C for 120 s. Cycloheximide was then added at cell culture to a final concentration of 100 ug/mL. Cells were immediately lysed in the same way as described for polysome profiling (see above). After lysis, ribosome-protected fragments were collected as described in Ingolia et al (Ingolia *et al*, 2012), with minor modifications. In brief, cell lysate was treated with RNase I at room temperature for 45 min. The nuclease digestion was stopped by adding SUPERase • In™ RNase inhibitor (Invitrogen). Monosomes were purified using illustra™ MicroSpin S-400 HR columns (GE Healthcae) following the instruction of ARTseq™ Ribosome Profiling Kit (Epicentre). RNA was isolated as described for polysome profiling (see above). rRNA was removed using Ribo-Zero™ Magnetic Kit (Human/Mouse/Rat) (Epicentre). The 28-32 nt ribosome-protected fragments were purified through 15% (wt/vol) polyacrylamide TBE-urea gel. The size-selected RNA was end-repaired by T4 PNK for 1 hr at 37°C followed by heat inactivation at 70°C for 10min. The dephosphorylated RNA was precipitated by ethanol and then ligated with a preadenylated FTP-3'adaptor for 2.5h at room temperature. The ligation product was purified through 15% (wt/vol) polyacrylamide TBE-urea gel and then reverse transcribed by FTP-RT primer using SuperScript III (Invitrogen) according to the manufacturers' instructions. Reverse transcription product was ethanol precipitated and further purified through 15% (wt/vol) polyacrylamide TBE-urea gel. Circularization of the reverse transcription product was

performed in the reaction containing 1x CircLigase Buffer, 50 mM ATP, 2.5 mM $MnCl_2$ and 100 U CircLigase (Epicentre) at 60°C for 1h, and the reaction was heat inactivated at 80°C for10 min. Circularized cDNA template was amplified by PCR for 12 cycles using the Phusion High-Fidelity DNA Polymerase using barcoded PCR primers. The final libraries were sequenced in 1 x 50 nt manner on Illumina HiSeq2000 platform. All the primer and adaptor sequences were listed in Appendix Table S5.

6. REFERENCES

Aebersold R & Mann M (2003) Mass spectrometry-based proteomics. *Nature* **422:** 198–207

Aisen P, Enns C & Wessling-Resnick M (2001) Chemistry and biology of eukaryotic iron metabolism. *Int. J. Biochem. Cell Biol.* **33:** 940–59

Albert FW, Muzzey D, Weissman JS & Kruglyak L (2014) Genetic influences on translation in yeast. *PLoS Genet.* **10:** e1004692

Ambros V (2011) MicroRNAs and developmental timing. *Curr. Opin. Genet. Dev.* **21:** 511–7

Ameres SL & Zamore PD (2013) Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.* **14:** 475–88

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO & Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.* **100:** 3889–94

Arce L, Yokoyama NN & Waterman ML (2006) Diversity of LEF/TCF action in development and disease. *Oncogene* **25:** 7492–504

Arribere JA & Gilbert W V. (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.* **23:** 977–987

Arrick BA, Lee AL, Grendell RL & Derynck R (1991) Inhibition of translation of transforming growth factor-beta 3 mRNA by its 5' untranslated region. *Mol. Cell. Biol.* **11:** 4306–4313

Artieri CG & Fraser HB (2014) Evolution at two levels of gene expression in yeast. *Genome Res.* **24:** 411–421

Avni D, Shama S, Loreni F & Meyuhas O (1994) Vertebrate mRNAs with a 5'-terminal pyrimidine tract are candidates for translational repression in quiescent cells: characterization of the translational cis-regulatory element. *Mol. Cell. Biol.* **14:** 3822–33

Baek D, Davis C, Ewing B, Gordon D & Green P (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.* **17:** 145–55

Baltz AG, Munschauer M & Schwanhäusser B (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46:** 674–690

Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* **136:** 215–233

Bashirullah A, Cooperstock RL & Lipshitz HD (2001) Spatial and temporal control of RNA stability. *Proc. Natl. Acad. Sci. U. S. A.* **98:** 7025–8

Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK & Gilad Y (2015) Impact of regulatory variation from RNA to protein. *Science (80-.).* **347:** 664–667

Bazzini AA, Lee MT & Giraldez AJ (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336:** 233–7

Beffagna G, OCCHI G, NAVA A, VITIELLO L, DITADI A, BASSO C, BAUCE B, CARRARO G, THIENE G & TOWBIN J (2005) Regulatory mutations in transforming growth factor-?3 gene cause arrhythmogenic right ventricular cardiomyopathy type 1. *Cardiovasc. Res.* **65:** 366–373

Betel D, Koppal A, Agius P, Sander C & Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11:** R90

Blaschke RJ, Töpfer C, Marchini A, Steinbeisser H, Janssen JWG & Rappold GA (2003) Transcriptional and translational regulation of the Leri-Weill and Turner syndrome homeobox gene SHOX. *J. Biol. Chem.* **278:** 47820–6

Brar GA, Yassour M, Friedman N & Regev A (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science (80-.).* **335:** 552–557

Braun JE, Huntzinger E & Izaurralde E (2012) A molecular link between miRISCs and deadenylases provides new insight into the mechanism of gene silencing by microRNAs. *Cold Spring Harb. Perspect. Biol.* **4:**

Braverman N, Chen L, Lin P, Obie C, Steel G, Douglas P, Chakraborty PK, Clarke JTR, Boneh A, Moser A, Moser H & Valle D (2002) Mutation analysis of PEX7 in 60 probands with rhizomelic chondrodysplasia punctata and functional correlations of genotype with phenotype. *Hum. Mutat.* **20:** 284–97

Bregman A, Avraham-Kelbert M & Barkai O (2011) Promoter elements regulate cytoplasmic mRNA decay. *Cell* **147:** 1473–1483

Brennecke J, Stark A, Russell RB & Cohen SM (2005) Principles of microRNA-target recognition. *PLoS Biol.* **3:** e85

Brown CY, Mize GJ, Pineda M, George DL & Morris DR (1999) Role of two upstream open reading frames in the translational control of oncogene mdm2. *Oncogene* **18:** 5631–7

Buchan JR & Stansfield I (2007) Halting a cellular production line: responses to ribosomal pausing during translation. *Biol. Cell* **99:** 475–87

Bushati N & Cohen SM (2007) microRNA functions. *Annu. Rev. Cell Dev. Biol.* **23:** 175–205

Calvo SE (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* **106:** 7507–7512

Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, Kamiya M, Shibata K, Sasaki N, Izawa M, Muramatsu M, Hayashizaki Y & Schneider C (1996) High-Efficiency Full-Length cDNA Cloning by Biotinylated CAP Trapper. *Genomics* **37:** 327–336

Carroll R & Derse D (1993) Translation of equine infectious anemia virus bicistronic tat-rev mRNA requires leaky ribosome scanning of the tat CTG initiation codon. *J. Virol.* **67:** 1433–40

Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J & Hentze MW (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149:** 1393–406

Cazzola M & Skoda RC (2000) Translational pathophysiology: a novel molecular mechanism of human disease. *Blood* **95:** 3280–8

Chamary J V & Hurst LD (2005) Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6:** R75

Chang K-J & Wang C-C (2004) Translation initiation from a naturally occurring non-AUG codon in Saccharomyces cerevisiae. *J. Biol. Chem.* **279:** 13778–85

Chappell SA, Edelman GM & Mauro VP (2004) Biochemical and functional analysis of a 9-nt RNA sequence that affects translational efficiency in eukaryotic cells. *Proc. Natl. Acad. Sci. U. S. A.* **101:** 9590–4

Chappell SA, LeQuesne JP, Paulin FE, deSchoolmeester ML, Stoneley M, Soutar RL, Ralston SH, Helfrich MH & Willis AE (2000) A mutation in the c-myc-IRES leads to enhanced internal ribosome entry in multiple myeloma: a novel mechanism of oncogene de-regulation. *Oncogene* **19:** 4437–40

Charneski CA & Hurst LD (2013) Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol.* **11:** e1001508

Chatterjee S & Pal JK (2009) Role of 5′-and 3′-untranslated regions of mRNAs in human diseases. *Biol. Cell* **101:** 251–262

Chen T, Ueda Y, Xie S & Li E (2002) A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation. *J. Biol. Chem.* **277:** 38746–54

Chi SW, Hannon GJ & Darnell RB (2012) An alternative mode of microRNA target recognition. *Nat. Struct. Mol. Biol.* **19:** 321–7

Chi SW, Zang JB, Mele A & Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460:** 479–86

Chiang PW, Carpenter LE & Hagerman PJ (2001) The 5'-untranslated region of the FMR1 message facilitates translation by internal ribosome entry. *J. Biol. Chem.* **276:** 37916–21

Cooper SJ, Trinklein ND, Anton ED, Nguyen L & Myers RM (2006) Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res.* **16:** 1–10

Courtois V, Chatelain G, Han Z-Y, Le Novère N, Brun G & Lamonerie T (2003) New Otx2 mRNA isoforms expressed in the mouse brain. *J. Neurochem.* **84:** 840–853

Cowles CR, Hirschhorn JN, Altshuler D & Lander ES (2002) Detection of regulatory variation in mouse genes. *Nat. Genet.* **32:** 432–7

Damgaard CK & Lykke-Andersen J (2011) Translational coregulation of 5'TOP mRNAs by TIA-1 and TIAR. *Genes Dev.* **25:** 2057–68

Davuluri R V. (2000) CART Classification of Human 5' UTR Sequences. *Genome Res.* **10:** 1807–1816

Dever TE & Green R (2012) The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **4:** a013706

Doench JG & Sharp PA (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.* **18:** 504–11

Du X, Wang J, Zhu H, Rinaldo L, Lamar K-M, Palmenberg AC, Hansel C & Gomez CM (2013) Second cistron in CACNA1A gene encodes a transcription factor mediating cerebellar development and SCA6. *Cell* **154:** 118–33

Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A & Segal E (2013) Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **110:** E2792–801

Eichhorn SW, Guo H, McGeary SE, Rodriguez-Mias RA, Shin C, Baek D, Hsu S-H, Ghoshal K, Villén J & Bartel DP (2014) mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Mol. Cell* **56:** 104–115

Fabian MR (2010) Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* **79:** 351–379

Fabian MR & Sonenberg N (2012) The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat. Struct. Mol. Biol.* **19:** 586–93

Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB & Bartel DP (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science (80-.).* **310:** 1817–1821

Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, Raychowdhury R, Hacohen N, Carr SA, Ingolia NT, Regev A & Weissman JS (2015) A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol. Cell* **60:** 816–827

Filipowicz W, Bhattacharyya SN & Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9:** 102–114

Floor SN & Doudna JA (2016) Tunable protein synthesis by transcript isoforms in human cells. *Elife* **5:** e10921

Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, Lassmann T, Itoh M, Summers KM, Suzuki H, Daub CO, Kawai J, Heutink P, Hide W, Freeman TC, Lenhard B, Bajic VB, Taylor MS, Makeev VJ, Sandelin A, Hume DA, et al (2014) A promoter-level mammalian expression atlas. *Nature* **507:** 462–70

Fresno M, Jiménez A & Vázquez D (1977) Inhibition of translation in eukaryotic systems by harringtonine. *Eur. J. Biochem.* **72:** 323–30

Friedman RC, Farh KK-H, Burge CB & Bartel DP (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19:** 92–105

Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, Schreiber S, Platzer M, Krawczak M, Hampe J & Brosch M (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res.* **22:** 2208–18

Fuxe J, Raschperger E & Pettersson RF (2000) Translation of p15.5INK4B, an N-terminally extended and fully active form of p15INK4B, is initiated from an upstream GUG codon. *Oncogene* **19:** 1724–8

Gerstberger S, Hafner M & Tuschl T (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15:** 829–45

Ghazalpour A, Bennett B & Petyuk VA (2011) Comparative analysis of proteome and transcriptome variation in mouse. *PLoS Genet.* **7:** e1001393

Ghilardi N, Wiestner A, Kikuchi M, Ohsaka A & Skoda RC (1999) Hereditary thrombocythaemia in a Japanese family is caused by a novel point mutation in the thrombopoietin gene. *Br. J. Haematol.* **107:** 310–6

Gray NK & Hentze MW (1994) Regulation of protein synthesis by mRNA structure. *Mol. Biol. Rep.* **19:** 195–200

Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP & Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27:** 91–105

Guo H, Ingolia NT, Weissman JS & Bartel DP (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466:** 835–840

Guo M, Yang S, Rupe M, Hu B, Bickel DR, Arthur L & Smith O (2008) Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSS) reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. *Plant Mol. Biol.* **66:** 551–63

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

Harigaya Y & Parker R (2010) No-go decay: a quality control mechanism for RNA in translation. *Wiley Interdiscip. Rev. RNA* **1:** 132–141

Heiman M, Schaefer A, Gong S, Peterson JD, Day M, Ramsey KE, Suárez-Fariñas M, Schwarz C, Stephan DA, Surmeier DJ, Greengard P & Heintz N (2008) A translational profiling approach for the molecular characterization of CNS cell types. *Cell* **135:** 738–48

Helwak A, Kudla G, Dudnakova T & Tollervey D (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* **153:** 654–65

Hentze MW & Kühn LC (1996) Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* **93:** 8175–82

Hinnebusch AG (1997) Translational Regulation of Yeast GCN4: A WINDOW ON FACTORS THAT CONTROL INITIATOR-tRNA BINDING TO THE RIBOSOME. *J. Biol. Chem.* **272:** 21661–21664

Hou J, Wang X, Mcshane E, Zauber H, Sun W, Selbach M & Chen W (2015) Extensive allele-specific translational regulation in hybrid mice. **:** 1–16

Hsieh AC, Liu Y, Edlind MP & Ingolia NT (2012) The translational landscape of mTOR

signalling steers cancer initiation and metastasis. *Nature* **485:** 55–61
Hudder A & Werner R (2000) Analysis of a Charcot-Marie-Tooth disease mutation reveals an essential internal ribosome entry site element in the connexin-32 gene. *J. Biol. Chem.* **275:** 34586–91
Huntzinger E & Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat. Rev. Genet.* **12:** 99–110
Huopio H, Jääskeläinen J, Komulainen J, Miettinen R, Kärkkäinen P, Laakso M, Tapanainen P, Voutilainen R & Otonkoski T (2002) Acute insulin response tests for the differential diagnosis of congenital hyperinsulinism. *J. Clin. Endocrinol. Metab.* **87:** 4502–7
Ingolia NT (2010) Genome-wide translational profiling by ribosome footprinting. *Methods Enzymol.* **470:** 119–42
Ingolia NT, Brar GA & Rouskin S (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7:** 1534–1550
Ingolia NT, Ghaemmaghami S, Newman JR & Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (80-.).* **324:** 218–223
Ingolia NT, Lareau LF & Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147:** 789–802
Jansen RP (2001) mRNA localization: message on the move. *Nat. Rev. Mol. Cell Biol.* **2:** 247–56
Jenkins RH, Bennagi R, Martin J, Phillips AO, Redman JE & Fraser DJ (2010) A conserved stem loop motif in the 5'untranslated region regulates transforming growth factor-β(1) translation. *PLoS One* **5:** e12283
Jones-Rhoades MW, Bartel DP & Bartel B (2006) MicroRNAS and their regulatory roles in plants. *Annu. Rev. Plant Biol.* **57:** 19–53
Kakegawa T, Ohuchi N & Hayakawa A (2007) Identification of AUF1 as a rapamycin-responsive binding protein to the 5′-terminal oligopyrimidine element of mRNAs. *Arch. Biochem. ...* **465:** 274–281
Kapeli K & Yeo GW (2012) Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. *Front. Neurosci.* **6:** 144
Kawamata T & Tomari Y (2010) Making RISC. *Trends Biochem. Sci.* **35:** 368–76
Keane TM, Goodstadt L & Danecek P (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477:** 289–294
Khan Z, Bloom JS, Amini S, Singh M, Perlman DH, Caudy AA & Kruglyak L (2012) Quantitative measurement of allele-specific protein expression in a diploid yeast hybrid by LC-MS. *Mol. Syst. Biol.* **8:**
Khorshid M, Hausser J, Zavolan M & van Nimwegen E (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods* **10:** 253–5
Kieft JS (2008) Viral IRES RNA structures and ribosome interactions. *Trends Biochem. Sci.* **33:** 274–83
Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, Ishii S, Sugiyama T, Saito K, Isono Y, Irie R, Kushida N, Yoneyama T, Otsuka R, Kanda K, Yokoi T, et al (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16:** 55–65
Komili S & Silver PA (2008) Coupling and coordination in gene expression processes: a systems biology view. *Nat. Rev. Genet.* **9:** 38–48

Kozak M (1986a) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–292
Kozak M (1986b) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44:** 283–92
Kozak M (1989a) Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. *Mol. Cell. Biol.* **9:** 5073–80
Kozak M (1989b) Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Mol. Cell. Biol.* **9:** 5134–42
Kozak M (1991a) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* **266:** 19867–70
Kozak M (1991b) An analysis of vertebrate mRNA sequences: intimations of translational control. *J. Cell Biol.* **115:** 887–903
Kozak M (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene* **299:** 1–34
Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361:** 13–37
Kozomara A & Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42:** D68–73
Krude H, Biebermann H, Luck W, Horn R, Brabant G & Grüters A (1998) Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nat. Genet.* **19:** 155–7
Kudla G, Murray AW, Tollervey D & Plotkin JB (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324:** 255–258
Law GL, Bickel KS, MacKay VL & Morris DR (2005) The undertranslated transcriptome reveals widespread translational silencing by alternative 5′ transcript leaders. *Genome Biol* **6:** R111
Lawniczak MKN, Holloway AK, Begun DJ & Jones CD (2008) Genomic analysis of the relationship between gene expression variation and DNA polymorphism in Drosophila simulans. *Genome Biol.* **9:** R125
Lee S, Liu B, Lee S, Huang S-X, Shen B & Qian S-B (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **109:** E2424–32
Lenhard B, Sandelin A & Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13:** 233–45
Levy S, Avni D, Hariharan N, Perry RP & Meyuhas O (1991) Oligopyrimidine tract at the 5' end of mammalian ribosomal protein mRNAs is required for their translational control. *Proc. Natl. Acad. Sci. U. S. A.* **88:** 3319–23
Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP & Burge CB (2003) Prediction of mammalian microRNA targets. *Cell* **115:** 787–798
Li G-W, Oh E & Weissman JS (2012) The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484:** 538–541
Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS & Johnson JM (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433:** 769–73
Liu L, Dilworth D, Gao L, Monzon J, Summers A, Lassam N & Hogg D (1999) Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat. Genet.* **21:** 128–32
Loeb GB, Khan AA, Canner D, Hiatt JB, Shendure J, Darnell RB, Leslie CS & Rudensky AY (2012) Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol. Cell* **48:** 760–70

Lukowski SW, Bombieri C & Trezise AEO (2011) Disrupted post-transcriptional regulation of the cystic fibrosis transmembrane conductance regulator (CFTR) by a 5'UTR mutation is associated with a CFTR-related disease. *Hum. Mutat.* **32:** E2266–82

Macejak DG & Sarnow P (1991) Internal initiation of translation mediated by the 5' leader of a cellular mRNA. *Nature* **353:** 90–4

Maniatis T & Reed R (2002) An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506

Matsui M, Yachie N, Okada Y, Saito R & Tomita M (2007) Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Lett.* **581:** 4184–8

Mayr C & Bartel DP (2009) Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138:** 673–684

McManus CJ, May GE, Spealman P & Shteyman and A (2014) Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.* **24:** 422–430

Medenbach J, Seiler M & Hentze MW (2011) Translational control via protein-regulated upstream open reading frames. *Cell* **145:** 902–13

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, Kellis M, Lander ES & Mikkelsen TS (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30:** 271–7

Melo EO, de Melo Neto OP & Martins de Sá C (2003) Adenosine-rich elements present in the 5′-untranslated region of PABP mRNA can selectively reduce the abundance and translation of CAT mRNAs in vivo. *FEBS Lett.* **546:** 329–334

de Melo Neto OP, Standart N & Martins de Sa C (1995) Autoregulation of poly(A)-binding protein synthesis in vitro. *Nucleic Acids Res.* **23:** 2198–205

Meng Z, Jackson NL, Shcherbakov OD, Choi H & Blume SW (2010) The human IGF1R IRES likely operates through a Shine-Dalgarno-like interaction with the G961 loop (E-site) of the 18S rRNA and is kinetically modulated by a naturally polymorphic polyU loop. *J. Cell. Biochem.* **110:** 531–44

Meyer KD, Patil DP, Zhou J, Zinoviev A, Skabkin MA, Elemento O, Pestova TV, Qian S-B & Jaffrey SR (2015) 5′ UTR m6A Promotes Cap-Independent Translation. *Cell*

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

Meyuhas O, Avni D & Shama S (1996) 13 Translational Control of Ribosomal Protein mRNAs in Eukaryotes. *Cold Spring Harb. Monogr. Arch.* **30:** 363–364

Mihailovich M, Thermann R, Grohovaz F, Hentze MW & Zacchetti D (2007) Complex translational regulation of BACE1 involves upstream AUGs and stimulatory elements within the 5' untranslated region. *Nucleic Acids Res.* **35:** 2975–85

Mitchell SA, Spriggs KA, Coldwell MJ, Jackson RJ & Willis AE (2003) The Apaf-1 Internal Ribosome Entry Segment Attains the Correct Structural Conformation for Function via Interactions with PTB and unr. *Mol. Cell* **11:** 757–771

Mize GJ (1998) The Inhibitory Upstream Open Reading Frame from Mammalian S-Adenosylmethionine Decarboxylase mRNA Has a Strict Sequence Specificity in Critical Positions. *J. Biol. Chem.* **273:** 32500–32505

Modrek B (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29:** 2850–2859

Morris DR & Geballe AP (2000) Upstream Open Reading Frames as Regulators of mRNA Translation. *Mol. Cell. Biol.* **20:** 8635–8642

Mukherji S, Ebert MS, Zheng GXY, Tsang JS, Sharp PA & van Oudenaarden A (2011) MicroRNAs can generate thresholds in target gene expression. *Nat. Genet.* **43:** 854–9

Muzzey D, Sherlock G & Weissman JS (2014) Extensive and coordinated control of allele-specific expression by both transcription and translation in Candida albicans. *Genome Res.* **24:** 963–973

Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J & Burge CB (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13:** 1894–910

Niesler B, Flohr T, Nöthen MM, Fischer C, Rietschel M, Franzek E, Albus M, Propping P & Rappold GA (2001) Association between the 5' UTR variant C178T of the serotonin receptor gene HTR3A and bipolar affective disorder. *Pharmacogenetics* **11:** 471–5

Novoa EM & Ribas de Pouplana L (2012) Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet.* **28:** 574–81

Oner R, Oner C, Wilson JB, Tamagnini GP, Ribeiro LM & Huisman TH (1991) Dominant beta-thalassaemia trait in a Portuguese family is caused by a deletion of (G)TGGCTGGTGT(G) and an insertion of (G)GCAG(G) in codons 134, 135, 136 and 137 of the beta-globin gene. *Br. J. Haematol.* **79:** 306–10

Ørom UA, Nielsen FC & Lund AH (2008) MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol. Cell* **30:** 460–71

Pasaje CFA, Bae JS, Park B-L, Cheong HS, Kim J-H, Uh S-T, Park C-S & Shin HD (2012) WDR46 is a Genetic Risk Factor for Aspirin-Exacerbated Respiratory Disease in a Korean Population. *Allergy. Asthma Immunol. Res.* **4:** 199–205

Patursky-Polischuk I, Stolovich-Rain M, Hausner-Hanochi M, Kasir J, Cybulski N, Avruch J, Rüegg MA, Hall MN & Meyuhas O (2009) The TSC-mTOR pathway mediates translational activation of TOP mRNAs by insulin largely in a raptor- or rictor-independent manner. *Mol. Cell. Biol.* **29:** 640–9

Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, May D, Lee C, Andrie JM, Lee S-I, Cooper GM, Ahituv N, Pennacchio LA & Shendure J (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30:** 265–70

Payton S, Cahill CM, Randall JD, Gullans SR & Rogers JT (2003) Drug discovery targeted to the Alzheimer's APP mRNA 5'-untranslated region: the action of paroxetine and dimercaptopropanol. *J. Mol. Neurosci.* **20:** 267–75

Paz I, Kosti I, Ares M, Cline M & Mandel-Gutfreund Y (2014) RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.* **42:** W361–7

Peabody DS (1989) Translation initiation at non-AUG triplets in mammalian cells. *J. Biol. Chem.* **264:** 5031–5

Pelletier J, Kaplan G, Racaniello VR & Sonenberg N (1988) Cap-independent translation of poliovirus mRNA is conferred by sequence elements within the 5' noncoding region. *Mol. Cell. Biol.* **8:** 1103–12

Pellizzoni L, Lotti F, Maras B & Pierandrei-Amaldi P (1997) Cellular nucleic acid binding protein binds a conserved region of the 5' UTR of Xenopus laevis ribosomal protein mRNAs. *J. Mol. Biol.* **267:** 264–75

Pellizzoni L, Lotti F, Rutjes SA & Pierandrei-Amaldi P (1998) Involvement of the Xenopus laevis Ro60 autoantigen in the alternative interaction of La and CNBP proteins with the 5'UTR of L4 ribosomal protein mRNA. *J. Mol. Biol.* **281:** 593–608

Peltz SW, Morsy M, Welch EM & Jacobson A (2013) Ataluren as an agent for therapeutic nonsense suppression. *Annu. Rev. Med.* **64:** 407–25

Pendleton LC, Goodwin BL, Flam BR, Solomonson LP & Eichler DC (2002) Endothelial argininosuccinate synthase mRNA 5'-untranslated region diversity. Infrastructure for tissue-specific expression. *J. Biol. Chem.* **277:** 25363–9

Pendleton LC, Goodwin BL, Solomonson LP & Eichler DC (2005) Regulation of endothelial argininosuccinate synthase expression and NO production by an upstream open reading

frame. *J. Biol. Chem.* **280:** 24252–60
Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F & Liuni S (2001) Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* **276:** 73–81
Pickering BM, Mitchell SA, Spriggs KA, Stoneley M & Willis AE (2004) Bag-1 internal ribosome entry segment activity is promoted by structural changes mediated by poly(rC) binding protein 1 and recruitment of polypyrimidine tract binding protein 1. *Mol. Cell. Biol.* **24:** 5595–605
Pisarev A V, Kolupaeva VG, Pisareva VP, Merrick WC, Hellen CUT & Pestova T V (2006) Specific functional interactions of nucleotides at key -3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. *Genes Dev.* **20:** 624–36
Pisarev A V, Shirokikh NE & Hellen CUT (2005) Translation initiation by factor-independent binding of eukaryotic ribosomes to internal ribosomal entry sites. *C. R. Biol.* **328:** 589–605
Plotkin JB & Kudla G (2010) Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12:** 32–42
Pozner A, Goldenberg D, Negreanu V, Le S-Y, Elroy-Stein O, Levanon D & Groner Y (2000) Transcription-Coupled Translation Control of AML1/RUNX1 Is Mediated by Cap- and Internal Ribosome Entry Site-Dependent Mechanisms. *Mol. Cell. Biol.* **20:** 2297–2307
Presnyak V, Alhusaini N, Chen Y-H, Martin S, Morris N, Kline N, Olson S, Weinberg D, Baker KE, Graveley BR & Coller J (2015) Codon Optimality Is a Major Determinant of mRNA Stability. *Cell* **160:** 1111–1124
Qian W, Yang J-R, Pearson NM, Maclean C & Zhang J (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet.* **8:** e1002603
Reynolds PR (2002) In sickness and in health: the importance of translational regulation. *Arch. Dis. Child.* **86:** 322–324
Rideau A, Mangeat B, Matthes T, Trono D & Beris P (2007) Molecular mechanism of hepcidin deficiency in a patient with juvenile hemochromatosis. *Haematologica* **92:** 127–8
Rogers JT, Randall JD, Cahill CM, Eder PS, Huang X, Gunshin H, Leiter L, McPhee J, Sarang SS, Utsuki T, Greig NH, Lahiri DK, Tanzi RE, Bush AI, Giordano T & Gullans SR (2002) An iron-responsive element type II in the 5'-untranslated region of the Alzheimer's amyloid precursor protein transcript. *J. Biol. Chem.* **277:** 45518–28
Rogers JT, Randall JD, Eder PS, Huang X, Bush AI, Tanzi RE, Venti A, Payton SM, Giordano T, Nagano S, Cahill CM, Moir R, Lahiri DK, Greig N, Sarang SS & Gullans SR Alzheimer's disease drug discovery targeted to the APP mRNA 5'untranslated region. *J. Mol. Neurosci.* **19:** 77–82
Ruan H, Hill JR, Fatemie-Nainie S & Morris DR (1994) Cell-specific translational regulation of S-adenosylmethionine decarboxylase mRNA. Influence of the structure of the 5' transcript leader on regulation by the upstream open reading frame. *J. Biol. Chem.* **269:** 17905–10
Sandberg R, Neilson JR & Sarma A (2008) Proliferating cells express mRNAs with shortened 3'untranslated regions and fewer microRNA target sites. *Science (80-.).* **320:** 1643–1647
Santhanam AN, Bindewald E, Rajasekhar VK, Larsson O, Sonenberg N, Colburn NH & Shapiro BA (2009) Role of 3'UTRs in the translation of mRNAs regulated by oncogenic eIF4E--a computational inference. *PLoS One* **4:** e4868
Sanz E, Yang L, Su T, Morris DR, McKnight GS & Amieux PS (2009) Cell-type-specific isolation of ribosome-associated mRNA from complex tissues. *Proc. Natl. Acad. Sci. U.*

S. A. **106:** 13939–44
Sawicka K, Bushell M, Spriggs KA & Willis AE (2008) Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem. Soc. Trans.* **36:** 641–7
Schneider-Poetsch T, Ju J, Eyler DE, Dang Y, Bhat S, Merrick WC, Green R, Shen B & Liu JO (2010) Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat. Chem. Biol.* **6:** 209–217
Schwanhäusser B, Busse D & Li N (2011) Global quantification of mammalian gene expression control. *Nature* **473:** 337–342
Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R & Rajewsky N (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature* **455:** 58–63
Shah P, Ding Y, Niemczyk M, Kudla G & Plotkin JB (2013) Rate-limiting steps in yeast protein translation. *Cell* **153:** 1589–601
Shalgi R, Hurt JA, Krykbaeva I, Taipale M, Lindquist S & Burge CB (2013) Widespread Regulation of Translation by Elongation Pausing in Heat Shock. *Mol. Cell* **49:** 439–452
Sharon E, Kalma Y & Sharp A (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30:** 521–530
Shenoy A & Blelloch RH (2014) Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat. Rev. Mol. Cell Biol.* **15:** 565–76
Sivagnanasundaram S, Morris AG, Gaitonde EJ, McKenna PJ, Mollon JD & Hunt DM (2000) A cluster of single nucleotide polymorphisms in the 5'-leader of the human dopamine D3 receptor gene (DRD3) and its relationship to schizophrenia. *Neurosci. Lett.* **279:** 13–6
Skelly DA, Merrihew GE, Riffle M, Connelly CF, Kerr EO, Johansson M, Jaschob D, Graczyk B, Shulman NJ, Wakefield J, Cooper SJ, Fields S, Noble WS, Muller EGD, Davis TN, Dunham MJ, Maccoss MJ & Akey JM (2013) Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res.* **23:** 1496–504
Sobczak K & Krzyzosiak WJ (2002) Structural Determinants of BRCA1 Translational Regulation. *J. Biol. Chem.* **277:** 17349–17358
Somers J, Pöyry T & Willis AE (2013) A perspective on mammalian upstream open reading frame function. *Int. J. Biochem. Cell Biol.* **45:** 1690–700
Sonenberg N (1994) mRNA translation: influence of the 5' and 3' untranslated regions. *Curr. Opin. Genet. Dev.* **4:** 310–5
Sonenberg N & Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136:** 731–745
de Sousa Abreu R, Penalva LO, Marcotte EM & Vogel C (2009) Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* **5:** 1512–26
Sözen MM, Whittall R, Oner C, Tokatli A, Kalkanoğlu HS, Dursun A, Coşkun T, Oner R & Humphries SE (2005) The molecular basis of familial hypercholesterolaemia in Turkish patients. *Atherosclerosis* **180:** 63–71
Sparkman O (2000) Mass spectrometry desk reference Pittsburgh Pa.: Global View Pub.
Spies N, Burge CB & Bartel DP (2013) 3′ UTR-isoform choice has limited influence on the stability and translational efficiency of most mRNAs in mouse fibroblasts. *Genome Res.* **23:** 2078–2090
Spriggs KA, Bushell M & Willis AE (2010) Translational regulation of gene expression during conditions of cell stress. *Mol. Cell* **40:** 228–237
Spriggs KA & Stoneley M (2008) Re-programming of translation following cell stress allows IRES-mediated translation to predominate. *Biol. Cell* **100:** 27–38
Sterne-Weiler T, Martinez-Nunez RT, Howard JM, Cvitovik I, Katzman S, Tariq MA,

Pourmand N & Sanford JR (2013) Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.* **23:** 1615–1623

Stolovich M, Tang H, Hornstein E, Levy G, Cohen R, Bae SS, Birnbaum MJ & Meyuhas O (2002) Transduction of growth or mitogenic signals into translational activation of TOP mRNAs is fully reliant on the phosphatidylinositol 3-kinase-mediated pathway but requires neither S6K1 nor rpS6 phosphorylation. *Mol. Cell. Biol.* **22:** 8101–13

Stoneley M & Willis AE (2004) Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression. *Oncogene* **23:** 3200–7

Takahashi H, Lassmann T, Murata M & Carninci P (2012) 5 end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat. Protoc.* **7:** 542–561

Tang H, Hornstein E, Stolovich M, Levy G, Livingstone M, Templeton D, Avruch J & Meyuhas O (2001) Amino acid-induced translation of TOP mRNAs is fully dependent on phosphatidylinositol 3-kinase-mediated signaling, is partially inhibited by rapamycin, and is independent of S6K1 and rpS6 phosphorylation. *Mol. Cell. Biol.* **21:** 8671–83

Tassin J, Dürr A, Bonnet AM, Gil R, Vidailhet M, Lücking CB, Goas JY, Durif F, Abada M, Echenne B, Motte J, Lagueny A, Lacomblez L, Jedynak P, Bartholomé B, Agid Y & Brice A (2000) Levodopa-responsive dystonia. GTP cyclohydrolase I or parkin mutations? *Brain* **123 (Pt 6:** 1112–21

Tenenbaum SA, Carson CC, Lager PJ & Keene JD (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 14085–90

Thoreen CC, Chantranupong L & Keys HR (2012) A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature* **485:** 109–113

Tirosh I, Reikhav S, Levy A & Barkai N (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *Science (80-.).* **324:** 659–662

Tuller T, Carmi A, Vestsigian K & Navon S (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141:** 344–354

Vasudevan S & Steitz JA (2007) AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. *Cell* **128:** 1105–18

Vasudevan S, Tong Y & Steitz JA (2007) Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318:** 1931–4

Vattem KM & Wek RC (2004) Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc. Natl. Acad. Sci. U. S. A.* **101:** 11269–74

Vera M, Pani B, Griffiths LA, Muchardt C, Abbott CM, Singer RH & Nudler E (2014) The translation elongation factor eEF1A1 couples transcription to translation during heat shock response. *Elife* **3:** e03164

Vogel C, Abreu RDS & Ko D (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.* **6:** 400

Wan Y, Qu K, Zhang QC, Flynn RA & Manor O (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505:** 706–709

Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A & Segal E (2016) Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science (80-.).* **351:** 240–240

Weingarten-Gabbay S, Khan D, Liberman N, Yoffe Y, Bialik S, Das S, Oren M & Kimchi A (2014) The translation initiation factor DAP5 promotes IRES-driven translation of p53 mRNA. *Oncogene* **33:** 611–8

Welch EM, Barton ER, Zhuo J, Tomizawa Y, Friesen WJ, Trifillis P, Paushkin S, Patel M, Trotta CR, Hwang S, Wilde RG, Karp G, Takasugi J, Chen G, Jones S, Ren H, Moon Y-

C, Corson D, Turpoff AA, Campbell JA, et al (2007) PTC124 targets genetic disorders caused by nonsense mutations. *Nature* **447:** 87–91

Wellensiek BP, Larsen AC, Stephens B, Kukurba K, Waern K, Briones N, Liu L, Snyder M, Jacobs BL, Kumar S & Chaput JC (2013) Genome-wide profiling of human cap-independent translation-enhancing elements. *Nat. Methods* **10:** 747–50

Wen Y, Liu Y, Xu Y, Zhao Y, Hua R & Wang K (2009) Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat. Genet.* **41:** 228–233

Wiestner A, Schlemper RJ, van der Maas AP & Skoda RC (1998) An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nat. Genet.* **18:** 49–52

Wilson RC & Doudna JA (2013) Molecular mechanisms of RNA interference. *Annu. Rev. Biophys.* **42:** 217–39

Witt H, Luck W, Hennies HC, Classen M, Kage A, Lass U, Landt O & Becker M (2000) Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. *Nat. Genet.* **25:** 213–6

Wittkopp PJ, Haerum BK & Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature* **1678:** 2010–2012

Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, Tang H & Snyder M (2013) Variation and genetic control of protein abundance in humans. *Nature* **499:** 79–82

Xue S, Tian S, Fujii K, Kladwang W, Das R & Barna M (2014) RNA regulons in Hox 5′ UTRs confer ribosome specificity to gene regulation. *Nature* **517:** 33–38

Zhang X & Borevitz JO (2009) Global analysis of allele-specific expression in Arabidopsis thaliana. *Genetics* **182:** 943–54

Zhang X, Gao X, Coots RA, Conn CS, Liu B & Qian S-B (2015) Translational control of the cytosolic stress response by mitochondrial ribosomal protein L18. *Nat. Struct. Mol. Biol.* **22:** 404–10

Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M & Lee JT (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* **40:** 939–53

Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR & Qian S-B (2015) Dynamic m(6)A mRNA methylation directs translational control of heat shock response. *Nature* **526:** 591–594

Zhou W & Song W (2006) Leaky scanning and reinitiation regulate BACE1 gene expression. *Mol. Cell. Biol.* **26:** 3353–64

Zid BM & O'Shea EK (2014) Promoter sequences direct cytoplasmic localization and translation of mRNAs during starvation in yeast. *Nature*

7. APPENDIX

Table S1. PCR primers for PacBio validation

Gene Symbols	Forweard Primers	Reverse Primers
Ankrd1	CCGAGCATGCTTAGAAGGAC	GCTCTTCTGTTGGGAAATGC
Itih2	TCATTTACCTGCCCAAAAGC	ATGTCCTTTCACCTCCATGC
Sgk1	GCCTGAGTATCTGGCTCCTG	ATCCACAGGAGGTGCATAGG
Tnfaip2	ATCATGGCCAACATCAACAA	GTATGTGGCCACCTCGATCT
Cyp7b1	CCTGCAGTCAACAGGTCAAA	GCCACACTTTCAGCTTCTCC
Irf5	GCTGTGCCCTTAACAAAAGC	TTGCTCCTGGGTAGCCTCTA
Macrod1	GAAGGAGGCCAAATCCTTTC	AGGTCCAGGCTGCTCAAGTA
Shcbp1	TGGACTTTCCATCCCTGAAG	ATGACCTTCTGGCCATTTTG
Icam1	AGTTGTTTTGCTCCCTGGAA	GTCTGCTGAGACCCCTCTTG
Impact	CATTTATGGCGAGGAGTGGT	TGAGCCTGAAAAGTGCTCCT
Kcnj15	ACCCCGAGTCATGTCAAAGA	ACCTGGATGACCAGGCATAG
Serpinb2	GGGCTTTATCCTTTCCGTGT	CATGGCCAGTTCTTCCTGTC
Ociad2	AGTGTCCACTCATGGGAACC	AAAACGGTTGGAAACCACAG
Rarres2	TGAGGTGAAGCCATGAAGTG	CTGGAGAAGGCAAACTGTCC
Tmtc4	TGTGATCCCCTTTCTTCCTG	CAACGACAGCAGCTCTTCAG
Acta1	TTGTGTGTGACAACGGCTCT	GAAGGAATAGCCACGCTCAG
Cd55	TCGAAAACAACCTCCACTCC	TGAGGGGGTTCCTGTACTTG
Ehd3	AGAGGATCAGCCGAGGGTAT	TTTTGGTGTCCTTCCCAAAC
Gstt2	TTCTCCCAGGTGAACTGCTT	TGCTCAGGATGGTGCTATGA
Ifi204	GCTGCTCCTGACCAAATGAT	AACCCATTGCACCCAAAATA
Serpinb6b	ATCCACTGCTGGAAGCAAAT	TCACAAGGACCAGTGGAGTG
Edil3	GGAACTTCTTGGCTGTGAGC	AGCTCTGACCGCAGAGTGAT
Nmnat2	TTCGAGAGAGCCAGGGATTA	TCCCCAACAATCACTTCCAT
Raph1	TGGCCAACTTTTCTTACCGC	CCCTGGTGTGTGGTCAAATC
Col2a1	GCCAAGACCTGAAACTCTGC	GGAGGTCTTCTGTGATCGGT
Epb4.1l3	GTACCCGAGGAGACCAAACA	ACACTCGTGCTTTCTACCCT
Mapk13	GCAACCTGGCTGTGAATGAA	GGCATCATCAAAAGGCTGCT
Mmp16	CTGAGACCCGGAGAGCAATT	CCTGTCATGTCTCCTTGGGT
Arhgap22	GCCAACTACAACCTGCTCAG	TATGAGCCAGTTCCCACCAG
Thbs2	AGCACAGATCGACACAGACA	TGTTCTCAGGGCACACATCA
Calr	CGCCAAATTCGAACCCTTCA	GGAATCTGTGGGGTCATCGA
Ltbp3	GGAGAGGACGGCATGTGTAT	GGTCAGGAGCAAAGGATGTAC
Psmd6	AAATCCCTCGACTGGCAGAT	CTGCTGGAAGACTGTGCAAC

Table S2. Oligos for Cap-Profiling

N15-oligo	5’-TACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNN-3’
CAP GN5 up	5’ CAGACGTGTGCTCTTCCGATCTGNNNNN-P 3’
CAP N6 up	5’ CAGACGTGTGCTCTTCCGATCTNNNNNN-P 3’
CAP 5'adptor down	5’ P-AGATCGGAAGAGCACACGTCTG-NH2 3’
CAP forward PCR primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC
Cap Reverse barcode 1 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG
Cap Reverse barcode 2 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG
Cap Reverse barcode 3 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG
Cap Reverse barcode 4 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG
Cap Reverse barcode 5 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
Cap Reverse barcode 6 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG
Cap Reverse barcode 7 primer	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG

Table S3. Oligos for validation of TL isoforms and their associated translational status

Nedd8	Nedd8-RT:ACATTCTCCCACCAGTAGA	Nedd8-GSP-R:GCAAGGAGGTAAACGGAAC
Ssu72	Ssu72-RT:TTGTCCTGGATGTCCACATT	Ssu72-GSP-R:CAGTTCTGGAACCTTTCTGGA
Ube4b	Ube4b-RT:TCATGTTGTGGACATTGAGA	Ube4b-GSP-R:CTGATGCAGCTATTGGAGGT
Ndufb11	Ndufb11-RT:TCGGTAAGCCATCAGTCA	Ndufb11-GSP-R:GATCGAAGTAGTTGGATTCCAT

Table S4. Oligos for Luciferase Assay

Oligo table for TL isoform translation validation

Gene Name	Long TL PCR primers	Short TL PCR primers
Lxn	F:GTCACTAGATCTAAGGAGGAAGAGGGAAGGAAGGCGCTGA	F:GTCACTAGATCTCCCACTTGGACACCCACTCGGCTG
	R:ACTGTCCCATGGCTTGGGGGAGACAGCGCGGGCT	R:ACTGTCCCATGGCTTGGGGGAGACAGCGCGGGCT
Ube4b	F:GTCACTAGATCTGACCCCTTTCAAAGATGGCCGCCCT	F:GTCACTAGATCTTTTAGAGGGGAGGGGCTTCCCGGT
	R:ACTGTCCCATGGCGCTTTCCTCTTAATGGTGAAAGGCGTTAGA	R:ACTGTCCCATGGCGCTTTCCTCTTAATGGTGAAAGGCGTTAGA
Eif1ad	F:GTCACTAGATCTCCCGGACACACCGCGCAT	F:GTCACTAGATCTGAATCGCAATTCCCGGCGCGGT
	R:ACTGTCCCATGGGCTGGTTTCTGTCCAGGGTTGTTAGG	R:ACTGTCCCATGGGCTGGTTTCTGTCCAGGGTTGTTAG
Itpr3	F:GTCACTAGATCTTATCTCAGGAGTTCAAACCAAAGCTCTAGGAGGAAGCAAAC	F:GTCACTAGATCTCAGACTTCCTGCTCCTTCCAGGCTGCA
	R:ACTGTCCCATGGGGCTTCGGCCCTCCGGGG	R:ACTGTCCCATGGGGCTTCGGCCCTCCGGGGCT
Nedd8	F:GTCACTAGATCTGTTTGTTCCGTTCCAGCTCG	F:GTCACTAGATCTAGTGTTCCTTGCCGTGGAGT
	R:ACTGTCCCATGGCTTCTTCCCAGGTTGGGGTT	R:ACTGTCCCATGGCTTCTTCCCAGGTTGGGGTT
Mpc2	F:GTCACTAGATCTTGCTGAGCTCCGCCCCCCTGA	F:GTCACTAGATCTGAAGCCGCTGTGCGTCACGATT
	R:ACTGTCCCATGGCGCGGCGGCCTAGGGAT	R:ACTGTCCCATGGCGCGGCGGCCTAGGGAT
Tmem129	F: GTCACTAGATCTCGATCTGACGGCGGTGGCT	F:GTCACTAGATCTGCACAGTGGGAGCGTTGG
	R: ACTGTCCCATGGCCCGCCACCGCTCACTG	R:ACTGTCCCATGGCCCGCCACCGCTCACTG
Ndufb11	Template: TAATACGACTCACTATAGGGACGAAGAAAATGAACAGACTCTAGATCTCCCAGGACTCCGCAGTACAAGCTGTCCCATGGACTCTCTCCCAGACAACAGAAACTATAGTGTCACCTAAAT	F:GTCACTAGATCTACAACTAGAAGCTCCACCTCTTTC
	F:TAATACGACTCACTATAGGG	R:ACTGTCCCATGGGACAGCTTGTACTGCGGAGTC
	R:ATTTAGGTGACACTATAG	
RenSp	F:GGTCAGAAGACCAACCCTCA R:CACGATAGCGTTGCTGAAGA	
Actb	F:CTGAACCCTAAGGCCAACCG R:TGGCTACGTACATGGCTGGG	

Oligo Table for Motif Validation

vMotif-F	GTACTCGATCATGACGTCACTAGATCT
vMotif-R	CTCAGAACTTGACGTACTGCTACTTAA
AAAAAT TL	GTACTCGATCATGACGTCACTAGATCTAAAAATTCACAAAAATTCAAAAA AATCAGTAAAAATAGACAAAAATCTTAAGTAGCAGTACGTCAAGTTCTG AG
AAAAAT rev-com TL	GTACTCGATCATGACGTCACTAGATCTATTTTTTCACATTTTTTCAAATTTT TCAGTATTTTTAGACATTTTTCTTAAGTAGCAGTACGTCAAGTTCTGAG
AATCCC TL	GTACTCGATCATGACGTCACTAGATCTAATCCCTCACAATCCCTTCAAATC CCCAGTAATCCCAGACAATCCCCTTAAGTAGCAGTACGTCAAGTTCTGAG
AATCCC rev-com TL	GTACTCGATCATGACGTCACTAGATCTAGGGATTTCACGGGATTTTCAGG GATTCAGTGGGATTAGACGGGATTCTTAAGTAGCAGTACGTCAAGTTCTG AG
AATCCC shuffled TL	GTACTCGATCATGACGTCACTAGATCTCCATCATCACCTCAACTTCAACCT ACCAGTCTCCAAAGACCACATCCTTAAGTAGCAGTACGTCAAGTTCTGA G
CAAGAT TL	GTACTCGATCATGACGTCACTAGATCTCAAGATTCACCAAGATTCAACAA GATCAGTCAAGATAGACCAAGATACCTTAAGTAGCAGTACGTCAAGTTCT GAG
CAAGAT rev-com TL	GTACTCGATCATGACGTCACTAGATCTATCTTGTCACATCTTGTCAAATCT TGCAGTATCTTGAGACATCTTGACCTTAAGTAGCAGTACGTCAAGTTCTG AG
CAAGAT shuffled TL	GTACTCGATCATGACGTCACTAGATCTTGAAACTCACTAAGACTCAAAGA ATCCAGTTAAGCAAGACGAACATACCTTAAGTAGCAGTACGTCAAGTTCT GAG
RenSp	F:GGTCAGAAGACCAACCCTCA R:CACGATAGCGTTGCTGAAGA
Actb	F:CTGAACCCTAAGGCCAACCG R:TGGCTACGTACATGGCTGGG

Table S5. Oligos for Initiating Ribosome Profiling

FTP 3'adaptor	5’-/5rApp/GATCGGAAGAGCACACGTCT/3ddC/-3’
FTP RT Primer	5’-(phos)-AGATCGTCGGACTGTAGAACTCTGAACGTGTAGATCTC GGTGGTCGAGACGTGTGCTCTTCCGATC-3’
FTP forward PCR primer	5’-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGT TCTACAGTCCGA-3’
FTP reverse PCR primer	see table S2 "Cap reverse barcode primer "