

6-1-2012

Investigating the translation of Cobra1: canonical expression is alternatively initiated from a non-AUG codon

Mohamed Abouelsoud

Follow this and additional works at: <https://fount.aucegypt.edu/etds>

Recommended Citation

APA Citation

Abouelsoud, M. (2012). *Investigating the translation of Cobra1: canonical expression is alternatively initiated from a non-AUG codon* [Master's Thesis, the American University in Cairo]. AUC Knowledge Fountain.

<https://fount.aucegypt.edu/etds/1186>

MLA Citation

Abouelsoud, Mohamed. *Investigating the translation of Cobra1: canonical expression is alternatively initiated from a non-AUG codon*. 2012. American University in Cairo, Master's Thesis. *AUC Knowledge Fountain*.

<https://fount.aucegypt.edu/etds/1186>

This Master's Thesis is brought to you for free and open access by the Student Research at AUC Knowledge Fountain. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of AUC Knowledge Fountain. For more information, please contact thesisadmin@aucegypt.edu.



The American University in Cairo
School of Sciences and Engineering

Investigating The Translation Of Cobra1: Canonical Expression Is Alternatively Initiated From A Non- AUG Codon

A Thesis Submitted to
The Biotechnology Program

In partial fulfillment of the requirements for
the degree of Master of Science

By Mohamed Mahmoud Abouelsoud

Under the supervision of
Dr. Rong Li
Dr. Asma Amleh

May 2012

DEDICATION

TO THE BEST PARENTS AND GREATEST PRIVILEGE ANYONE CAN HAVE...

TO MY CARING BROTHER...KHALED

TO MY LOVELY WIFE...YASMINE

ACKNOWLEDGEMENTS

I would like to thank: **Dr.Asma Amleh** for her continuous guidance and support, **Dr.Rong Li** for hosting and guiding me for six months, **Haihui pan** for being an outstanding trainer, **Dr.Rania Siam** for her support, **Yasmine Mustafa** for reviewing the manuscript and **Ahmed Youssef** for helping with the final steps of the project.

This work was funded by department of molecular medicine, **University of Texas, health science center in San Antonio**, Texas and the **American university in Cairo**. Last but not least, I am very grateful for **Al-Alfi foundation** for funding me throughout the program duration and during my stay in UTHSCSA.

ABSTRACT

The American University in Cairo

Investigating the translation of *Cobra1*: Canonical expression is alternatively initiated from a non-AUG codon

By Mohamed M. Abouelsoud

COBRA1, co-factor of BRCA1, is a transcriptional regulator and a subunit of the Negative elongation complex also known as NELF-B. Although this protein was first designated as a cofactor of BRCA1 and hence acts accordingly, it was found later that it elicits a battery of response genes overlapping those regulated by BRCA1 in absence of BRCA1 itself. *Cobra1* deletion is embryonic lethal and results in embryonic stem cells (ESC) differentiation independent of the typical pluripotency machinery. Moreover, it was found that it has a role in suppression of tumors' growth and patients with poor prognosis of breast cancer had decreased levels of COBRA1. Paradoxically, levels of COBRA1 was found elevated in some upper gastro-intestinal tract tumors.

Our understanding of the regulation of gene expression has been evolving as an important venue to explain gene product's diversification. Alternative initiation of translation has been observed in many important genes and showed different subsequent phenotypes. In some cases, the discovered protein isoforms are not generated from the classically recognized Kozak/ATG system (i.e. Canonical initiation). Alternatively, their expression is initiated using a non-canonical mechanism resembling viral internal ribosomal entry site (IRES) pathway. Generation of different protein isoforms has been linked to paradoxes in the associated genes' functions. Among the different functions observed are resistance to degradation, altered cellular localization and regulation of different cell cycle phases.

In this study we have substantiated the hypothesis that *Cobra1* has two protein isoforms, which might be one of the possible reasons for the associated paradoxes. We have used *in-silico* prediction analyses to verify that the 5' un-translated region (5'UTR) of *Cobra1* has the required sequences and complex RNA structures for non-canonical initiation. We also could detect these isoforms in endogenous mouse tissues from different strains and ages. Finally, we were able to induce the expression of the two isoforms *ex-vivo* and still could recognize the isoforms in flag-tag based systems.

Table of Content

DEDICATION	II
ACKNOWLEDGEMENTS	III
ABSTRACT.....	IV
LIST OF TABLES AND FIGURES	VII
LIST OF ABBREVIATIONS	VIII
1. LITERATURE REVIEW	1
1.1. INTRODUCTION.....	1
1.2. THE UNDER-ESTIMATED CAPABILITIES OF MAMMALIAN GENOMES.....	3
1.3. REGULATION OF GENE EXPRESSION	3
1.3.1. <i>Epigenetic control</i>	4
1.3.2. <i>Transcriptional control</i>	4
1.3.3. <i>Post-transcriptional control</i>	5
1.3.4. <i>Translational control</i>	7
1.3.5. <i>Post-translational modifications</i>	13
1.4. COBRA1; FROM AN UNDERMINED COFACTOR TO A PLURIPOTENCY REGULATOR.....	14
2. SIGNIFICANCE, SCOPE AND DESIGN OF THE STUDY.....	18
3. MATERIALS AND METHODS	19
3.1. IN-SILICO ANALYSES.....	19
3.2. BACTERIAL STRAINS AND CULTURE	19
3.3. POLYMERASE CHAIN REACTION (PCR)	19
3.4. RESTRICTION DIGESTION	20
3.5. LIGATION AND TRANSFORMATION.....	20
3.6. SCREENING FOR RECOMBINANTS AND VERIFICATION OF PLASMIDS.....	20
3.7. CELL CULTURE AND CELL LINES	20
3.8. TRANSFECTION	21
3.9. RETROVIRAL INFECTION	22
3.10. EMBRYONIC STEM CELLS DIFFERENTIATION	22
3.11. MOUSE TISSUE ANALYSIS	22
3.12. WESTERN BLOTTING.....	23
3.13. FLAG-IMMUNOPRECIPTATION.....	23
3.14. CELLULAR FRACTIONATION.....	23
3.15. IMMUNOHISTOCHEMISTRY	24
4. RESULTS.....	25

4.1. DISSECTING THE SEQUENCE OF THE 5'UTR OF HUMAN AND MOUSE COBRA1.	25
4.2. DETECTION OF ENDOGENOUS COBRA1 ISOFORMS	26
4.3. CONFIRMATION OF THE EXISTENCE OF COBRA1 ISOFORMS USING <i>EX-VIVO</i> GENE EXPRESSION	27
5. DISCUSSION AND CONCLUSION.....	29
6. FUTURE PROSPECTS	34
7.TABLES.....	35
8. FIGURES	36
9. REFERENCES.....	46

LIST OF TABLES AND FIGURES

Table.1 List of primers used for cloning Full *Cobra1/COBRA1* one transcripts and coding sequences for human and mouse

Figure.1 Nucleotide sequence of the 5'UTR of *Mus Musculus*.

Figure.2 Amino acid sequence of the 5'UTR in *Mus Musculus*

Figure.3 Blastn results for 5'UTRs alignment

Figure.4 Nucleotide sequence of the initiation near-cognate loci

Figure.5 Secondary RNA structures predicted via CARNAC

Figure.6 Secondary RNA structure predicted via Mfold

Figure.7 Curve showing SignalP 4.0 score of COBRA1 versus positive sample

Figure.8 Immunohistochemistry using COBRA1 rabbit anti-mouse antibody

Figure.9.a Western blotting on mouse tissues (Intestine) using COBRA1 anti-mouse antibody

Figure.9.b Western blotting on mouse tissues (Spleen) using COBRA1 anti-mouse antibody

Figure.9.c Western blotting on mouse tissues (Kidney, Liver, Muscle) using COBRA1 anti-mouse antibody

Figure.10 Western blotting on mouse tissues (Kidney, Spleen, Testis) using COBRA1 anti-human antibody

Figure.11 Western blotting showing Comparison of expression pattern of COBRA1 isoforms

Figure.12 Schematic diagram for *ex-vivo* expression of *COBRA1/Cobra1* in mammalian cell lines

Figure.13 Diagram for cloning fragments of *Cobra1*

Figure.14 Amplification of mouse cloning fragments of *Cobra1*

Figure.15.a Transformants verification by colony PCR

Figure.15.b Restriction digestion of the human cloning fragments prior to cloning

Figure.16 Expression of human *COBRA1* constructs in HEK293T cells

Figure.17.a Expression of mouse *Cobra1* constructs in iMEF cells before selection

Figure.17.b Expression of mouse *Cobra1* constructs in iMEF cells after selection

Figure. 18 Flag immune-precipitation for mouse *Cobra1* constructs

Figure.19.a Photos for differentiation phases of ESC differentiation

Figure.19.b Western blotting using COBRA1 anti-mouse antibody throughout differentiation

Figure.20 Western blotting analysis for cytoplasmic and nuclear fractions of the infected cell lines with *Cobra1* construct

List Of Abbreviations

Ab – Antibody

BSA – Bovine serum albumin

ceRNA – Competitive endogenous RNA

ChIP – Chromatin immune-precipitation

Cobra1 – Annotation of mouse Cobra1 gene

COBRA1 – Annotation of human Cobra1 gene

COBRA1 – Annotation of COBRA1 protein

Cre – Cre recombinase

CTD – C-terminal domain

Ctrl – Control

Cyto – Cytoplasmic fraction

D.P.C – days post coitum

EB – Embryoid bodies

eIF – Eukaryotic initiation factor

ESC – Embryonic stem cells

EV – Empty vector

FBS – Fetal bovine serum

FL – Full length

GIT – Gastrointestinal tract

HEK – Human embryonic kidney cells

HGP – Human genome project

IHC – Immunohistochemistry

IP – Immuno-precipitation

IRES – Internal ribosomal entry site

IS – Initiation start site

ITAF – IRES trans-activating factor

KO – knockout

LIF – Leukemia inhibitory factor

LSB – Laemmli sample buffer

LTR – Long tandem repeats

MEF – Mouse embryonic fibroblast

miRNA – Micro RNA

MMTV – mouse mammary tumor virus

NELF-B – Negative elongation factor subunit B

Nuc – Nuclear fraction

ORF – Open reading frame

PCR – Polymerase chain reaction

PIC – Pre-initiation complex

PolII – RNA polymerase II

PTGS – Post-transcriptional gene silencing

Puro – Puromycin

RISC – RNA induced silencing complexes

siRNA – Short interference RNA

WB – Western blotting

3UTR – 3' Un-translated region

5'UTR – 5' Un-translated region

qPCR – Quantitative real time PCR

1. Literature Review

1.1. Introduction

COBRA1 was first discovered as a COfactor of BRCA1, the breast cancer susceptibility gene. After the initial discovery, studies have revealed more of the gene's function especially in tumor formation. The studies conducted on this gene, as is the case with most of genes, are more focused on its functions and role in different fields such as Cancer and stem cells. However, the biology surrounding *Cobra1* and its regulation mechanisms remains an unexplored mystery. Thorough studies have been conducted and showed that unraveling the regulation of gene expression can offer a better understanding to genes' functions. Hence, it is logical to first visit the mechanisms by which genes are regulated and how can regulation contribute to gene products' diversity and functions in general, then focus on *Cobra1*'s importance and regulation.

One might think that the real start of dating for studying regulation of gene expression is the completion of genome projects. However, gene regulation studies have been conducted for decades. Yet, no one can deny that genome projects such as the human genome project (HGP) revolutionized scientific and health related fields since 2003. The revelation of the complete map for the human genome did provide the scientific community with a great platform upon which thousands of studies could be based. However, cell biology related studies have shown that regulation of gene expression needs much more than nucleotide sequence, assembly or even chromosomal mapping to solve regulation mysteries. A gene's sequence might be known but this would only represent the first steps in a thousand miles road.

In other words, the genomes' – especially eukaryotic – capabilities would be remarkably undermined if it were dealt with as a four-digit code. “Organic chemistry is the chemistry of carbon compounds. Biochemistry is the study of carbon compounds that crawl.” Says Mike Adams. Similarly, it is very erroneous to deal with DNA from an organic chemistry point of view. The message encrypted within the four-digit code is

magnificently adding layers of complexity to the regulation of this genome and creates feedback loops of control over the source DNA. This regulation of gene expression and the added complexity is very necessary especially for higher organisms. It explains how one type of a starting DNA sequence – within a developing embryonic stem cell – is capable of producing an organism that has trillions of cells among which we can find astonishing discrepancies. Moreover, regulation of expression can account for the outstanding abilities of cells to change fate or withstand stress. Also, alteration of expression of some genes has led to transformation of the harboring cells into cancerous cells. Consequently, it has become a trend to study the mechanism of gene regulation whenever this gene is linked to some of the aforementioned events affected by alteration of expression. Types, levels and consequences of regulation of gene expression will be discussed further in details in Section 2.

An interesting observation was made during investigations done on COBRA1. COBRA1 was detected as a doublet band or two bands in western blotting. Hence, we were intrigued to launch a battery of experiments to mine for the exact mechanisms that control expression of COBRA1. Perhaps, this would help answer some of the questions and paradoxes surrounding COBRA1 (discussed in section 1.4). It was then the challenge of pinpointing the most probable level of expression regulation at which we should start hypothesizing.

The goal of this study was to investigate if *Cobra1* was regulated in a mechanism different from typical regulation of gene expression. We chose to start at the translational level and the initiation step in specific. This was based on some preliminary data deduced from mutagenesis experiments (discussed later in section 2). These results allowed us to confidently hypothesize that *Cobra1* might be following a non-canonical mechanism of translation initiation from an alternative start codon resulting in a second protein isoform.

The next section will focus on studies that shed more light on the importance of this initiative, others that were used to complete our hypothesis and finally how our battery of experiments was designed and justified.

1.2. The under-estimated capabilities of mammalian genomes

The Human Genome Project (HGP) estimated that the functional sequences of the human genome represents no more than 1.5% of the whole genome [1]. However, a linear relationship, between the estimated number of coded proteins from this percentage and the diverse phenotypes observed within different cells, is undeducible. This owes to the complexity of eukaryotic phenotypes and sophisticated cellular machinery. Thanks to some recent advances, new approach of genomic studies has appeared namely:

Comparative genomic studies. These studies are aimed at identifying evolutionary constraints between closely related species e.g. placental (Eutherian) mammals. The studies have succeeded to show that, for example, 5% of the human genome is conserved with other mammals such as mouse [2] and rat [3]. Postulating that this percentage has appeared due to purifying selection, the consortium of authors suggested that the 3.5% increase are new protein-coding sequences that was undermined by the initial estimation of the HGP[4]. Such studies ultimately concluded that the initial understanding of the mammalian genome capabilities, especially in human, is still immature.

Yet, we have learned from rather earlier studies that the initial limitations for genomic studies might not be the only reason for the aforementioned immature estimation of mammalian genome's capabilities. The mammalian genome-encrypted messages are enormously diversified by the regulation of gene expression through many mechanisms such as: epigenetic control [5], transcriptional control [6], post-translational control [7], [8] among others. In order to have a better understanding of the role of gene regulation in diversity, some of the levels of control will be briefly overviewed in the light of diversification of gene product and not their specific mechanisms.

1.3. Regulation of gene expression

The eukaryotic genome is characterized by multiple sophisticated levels of gene regulation. This complexity is more prominent as we ascend the evolutionary ladder as is clear in mammalian genomes. All the mentioned regulatory processes discussed below will address eukaryotic gene regulation unless otherwise stated. The main steps of gene regulation is classified as follows:

1.3.1. Epigenetic control

1.3.2. Transcriptional control

1.3.3. Post-Transcriptional control

1.3.4. Translational control

1.3.5. Post-translational control

1.3.1. Epigenetic control

Epigenetics is the control of gene expression through modifications done on the DNA molecule or associated proteins after replication [5]. Epigenetics is not concerned with any modifications taking place at the nucleotide level though. The main processes that affect epigenetic control are: Histone modifications, DNA methylations and subsequent nucleosomal positioning on promoter regions. Perhaps the control of gene expression through promoter control is the major pathway through which the above epigenetic modification functions [9]. Some promoters (around 40%) can be turned on or shut off by the state of methylation of their CpG islands [10]. Obviously, these DNA modifications add more to the diversity of gene products. Histone de-acetylases and histone methyl-transferases also play a great role in gene regulation. Where the histone acetylation and methylation states would confer an open or closed access to a gene's promoter, respectively. In this sense the diversity is more of a temporal control and differential expression between tissues [11].

1.3.2. Transcriptional control

Transcriptional control is – to a great extent – dependent on the epigenetic state of the promoter DNA sequence of the regulated gene. However, there are other structures & phenomena that are considered crucial for normal gene transcription: Transcriptional regulatory sequences and RNA Polymerase II (PolII) pausing. While promoters and PolII binding (the actual first steps in gene expression) have been explored for a while now, other phenomena such as stalled PolII are still relatively new to the pages of scientific journals. Promoters depend on the presence of specific motifs to be recognized for the docking of Polymerases specially RNA PolII such as TATA box (reviewed in [12] [13]). Other structures might be needed for the stability of the transcription process such as:

Downstream promoter element. In fact, after PolII identifies the promoter region of a gene, bind and assemble its pre-initiation complex (PIC), it might abort this transcription around 10bp after the transcription start (+1). Only when the CTD domain of PolII is phosphorylated, through TFIIH and other (general and specific) transcription factors, is it possible to clear the promoter region and form a stable transcriptional bubble (Transcriptional machinery + single stranded DNA + Nascent RNA molecule). Only then, the transcriptional machinery stabilizes the stochastic nature of PolII and it does not dissociate until the end of the template of the –being – transcribed DNA [14-16]. Recently, it was found that the transcriptional bubble run into what some scientists have named “Speed bumps”. This results in what appears to be a “Stalling” process of PolII approximately between +30 & +50 [17].

1.3.3. Post-transcriptional control

Post-transcriptional control is relatively the easiest to be linked to regulation of gene expression and is believed to be the highest contributing to gene products diversity. It includes direct RNA maturation after transcription (capping and tailing), RNA processing and modifications (alternative splicing) in addition to PTGS – Post transcriptional gene silencing (RNA interference mediated regulation).

The final transcript made ready for a cell’s machinery to translate is the mature and modified final transcript (mRNA). This mRNA is not the initial product of transcription (pre-mRNA or primary transcript) but it is a result of many processes of modifications and adjustments. In order to sustain the message encrypted within a transcript should be protected by special structures [18], [19]. This is done mainly through two processes: capping and tailing. These modifications are also very necessary for other molecular machineries, the most important of which is cap-dependent translation. A group of enzymes act on the 5’ end of the mRNA and add a specialized 7’methylguanylate through a 5’, 5’ tri-phosphate link. The poly (A) polymerase acts on the 3’end of the nascent pre-mRNA after an endonuclease leaves an uncovered 3’-OH in the last nucleotide. The polymerase catalyzes the addition of 100-250 residues of adenylic acid forming a poly (A) tail [19].

Capping and tailing of pre-mRNA are very important steps. However, alternative splicing of the mature mRNA is considered the most important step in post-transcriptional control. Alternative splicing's impact on gene product can be seen in the astonishing case of DSCAM, an axon guidance receptor in *Drosophila* [20]. This protein is responsible for the extremely sophisticated process of neuron wiring and assembly of the nervous system. It is estimated that the splicing isoforms possible for the pre-mRNA of *DSCAM* exceeds 38,000 isoforms [20], [21]. This number is astonishing because the *Drosophila*'s genome has only 13,600 genes! In other words, the number of isoforms generated by alternative splicing of *DSCAM* is approximately thrice the number of genes within the *Drosophila* genome. This might be the answer long sought for to explain how can *Drosophila* – a fly – have an equivalent functional diversity to the – nematode – *C. elegans* of the 20,316 genes [22].

The journey of the mRNA has not ended yet. It has to get past the tackling of the Post-Transcriptional Gene Splicing (PTGS). PTGS is mediated through short RNA molecules of 20-30 bases categorized – at least in mammals – into three main groups; siRNA, miRNA and piRNA [7], [23]. The process also, known as RNA interference, is carried out through the RISC complex (RNA-induced silencing complexes) containing three types of macromolecules; dicers, short RNA molecules and Argonaute or Piwi [7], [24]. The silencing effector function is mediated through Argonaute (miRNA/siRNA) or Piwi (piRNA) that carries out destruction of double stranded RNA molecules generated from RISC complex [7], [24], [25]. Recently, a new rival to the interfering RNA species has been discovered; ceRNA. The competitive endogenous RNA plays a totally different role in regulation of gene expression [26-28]. The newly – in animals – discovered RNA species act as decoy targets for miRNA, siRNA or piRNA. Thus, target mRNA molecules are now available for translation more readily. This changes everything again! It is quite easy to comprehend now that – as is the case with almost all cellular processes – a balance is needed between the two opposing (miRNA vs. ceRNA) RNA regulatory species to eventually result in the desired mRNA and subsequent protein levels [29].

PTGS is a very important and critical level of regulation. However, its contribution to diversity of gene product wasn't revealed in the aforementioned facts. In

order to appreciate the role of RNA interference from the scope of this review, one has to put back these machineries in context, simulate and hypothesize scenarios to elucidate some of the complexities. A simple – enough - system is a gene regulated through all posttranscriptional controls. If we assume a system in which a gene has two splicing isoforms (full transcript or missing an intron) then is subjected to PTGS. If the miRNA, targeting the mRNA for silencing, were complementary to the spliced out intron, then the miRNA wouldn't exactly be functional. Conversely, It would be conceivable if such a miRNA inhibit one isoform and not the intron-missing isoform. To sophisticate things more, the affected gene's mRNA levels might further be regulated through the presence of a ceRNA that mimics the targeted intron and divert miRNA-mediated PTGS. This scenario was possible in reality because of some recent experiments that succeeded to target specific splicing isoforms for silencing and to spare only needed isoforms [30], [31]. These experiments opened a new venue for therapeutics and added to the specificity of siRNA-based therapies that considered diversity of gene product and the usually – different physiological role [32].

1.3.4. Translational control

Having passed the refining machineries of transcriptional and post-transcriptional controls, mRNA is now engineered qualitatively and quantitatively to be translated into a certain level of the corresponding protein. The story of diversification has a very interesting chapter yet to be told. Translation – in general – is accomplished in three stages; initiation, elongation and termination [33]. From the diversity point of view, Initiation sits at the top of the sources of producing protein isoforms from a single mRNA [33]. In contrast, elongation is a very simple process –or so we believe – that involves only stabilizing the translation complex and the correct codon/anti-codon interaction. Termination comes in second place with only one mechanism for diversification through the recognition of different poly-adenylation sites. These altered poly-adenylation sites result in somewhat truncated or elongated proteins at the 3' un-translated regions (3'UTR). The main focus in this section and literature review is the regulation carried out at the initiation step [33].

Despite the great advances in techniques used to dissect cellular machineries, most of the – so-called – established concepts about initiation of translation remain elusive. It has been observed that eukaryotes initiate translation of their mature mRNA through cap-dependent machinery, internal ribosomal entry site (IRES) pathway [cap-Independent] or alternative initiation from a near-cognate codon [33-37].

Cap-dependent machinery is the most studied pathway for initiation of translation in Eukaryotes [33]. In prokaryotes, (lack the cap and tail of mRNA), ribosomal subunits dock on the sequences just before the initiation start site (IS). Eukaryotes, on the other hand, tend usually to make use of its mRNA special structures especially the cap [18], [38], [39]. Translation, according to this model, starts by the recognition and binding of specific complexes to the cap followed by mRNA scanning. In general, it involves the formation of three main complexes; cap recognition complex, pre-initiation complex (PIC) followed by the Scanning complex that stops when it recognizes a start codon and gives way to the actual 80S Initiation complex [37], [40-42].

When the wandering around factors: eIF4E, eIF4G and eIF4A also collectively known as eIF4F run into a capped mRNA, it interacts with the cap through eIF4E. To this activated mRNA, the small (40S) ribosomal subunit binds and recruits the ternary complex (eIF2–GTP–Metⁱ-tRNA) and others such as eIF2, eIF3 & eIF5 [38]. The scanning powers of the – now – 43S scanning complex is gained after the binding of the eIF1& eIF1A. The scanning 43S complex will stop only when it identifies a cognate AUG codon [37], [42], [43]. A cognate AUG is the first AUG, within a favorable context, that an initiation complex runs into. In Eukaryotes, this favorable context is known as Kozak sequence [43]. This recognition will stimulate the hydrolysis of the GTP in the ternary complex into GDP allowing the PIC to disassemble. Being cleared, the way is now paved for the large 60S ribosomal subunit to bind. It docks on the mRNA to form an 80S ribosomal translating complex where the Metⁱ-tRNA is oriented to the P-site (Peptidyl site) of the 60S subunit [37], [41].

The previously mentioned scenario has been substantiated and supported by many studies. It is estimated that 90-95% of cellular mRNA follow cap dependent mechanisms to initiate translation from an AUG codon [40]. Yet, there are growing numbers of

opinions and evidences showing that what has been elucidated might not be the most accurate way to describe the actual translation process [34-36], [44]. One of the strongest evidences debating against the canonical cap-dependent understanding is the study done by Terenin *et al.* [35]. They showed through a series of experiments that eIF1, eIF2 & eIF4 are dispensable factors. Taking into consideration the –previously stated - roles these factors play, their absence means that the Initiation complex will assemble and start translation in a prokaryotic-like fashion. One might argue against the fact that they used HCV (Hepatitis C Virus) IRES to test their hypothesis and not usual sequences that normally use the cap dependent machinery. However, the inclusion of the factors previously-proved dispensable succeeded in driving translation as efficient. Still, it is ironic to argue using this caveat because the mere existence of IRES is a stab in the back of the canonical cap-dependent pathway [35].

Picornaviruses gave us the first lessons on Internal ribosome entry site (IRES) [42], [45], [46]. The famous group of viruses (including Rhinovirus and Poliovirus), along with other devastating groups such as Adenoviruses (such as HCV and HIV), were found to initiate translation, of their mRNA in host cells, without the need for all the aforementioned sophisticated machinery [36], [46]. Later, few numbers of genes within the human genome were identified as genes harboring “Cellular IRES”. Initiation of translation of those genes does not abide by the canonical cap-dependent initiation. Surprisingly, the knowledge we have on the mechanisms of regulation of IRES is very humble given that the discovery of IRES has aged more than a decade now. GC-rich regions and subsequent complex RNA structures (such as stems, loops, knots or pseudo-knots) were the only observed features that characterize IRES. However, attempts to characterize a conserved secondary RNA structure or specific motifs have not been successful so far [36], [39], [46]. This might be due to technology limitations but can also be due to the fact that there’re no actual conserved motifs for such phenomenon [36], [46]. The evidence of which stems from the fact that even members of the same protein family such as *L-myc* and *c-myc* showed very distant complex RNA structures in their 5’UTR [47], [48]. Moreover, no specific translation initiation factors were linked to this pathway. Some studies suggested the interaction of some factors with the complex RNA structure known as IRES Trans-Acting Factors (ITAFs) [49] such as GRSF1 [50] and LA

autoantigen [51]. Their binding to IRES was only linked to stabilizing the complex secondary RNA structure but not directly to the recruitment of special initiation complexes or ribosomal subunits [37].

There have been many attempts to describe why, when there's a robust and accurate machinery, would translation be initiated from - what might look like - an "irregular" pathway?

The evolutionary model seems like the most logical and credible justification for this phenomenon[34], [36]. It has been suggested that all organisms have the ability to initiate translation without the need for the sophisticated machinery. Upon evolution, eukaryotic cells developed a system that was more robust and reliable in producing consistent and steady rate of gene expression to be able to cope with the sophisticated tasks within higher organisms[34], [44].

This hypothesis was substantiated when correlation studies between IRES and virulence of viruses came out[39], [52]. For instance, after Poliovirus infects a cell, it brings about a global translational "shut off" in two hours[52], [53]. Yet, all the genes carried within the RNA genome (i.e. Polio is an RNA virus) of the virus are efficiently translated using the machinery of the same cell that just lost 90-95% of its proteome! The answer lies within the percentage of the paralyzed gene transcripts. It is more or less the percentage of mRNA that's initiated through cap-dependent machinery[37], [52]. In a nutshell, viruses kill the major translational machinery within the cell because it can drive the translation of its genes through the IRES pathway[52], [54]. If we consider the evolutionary model here, we would think that Poliovirus would only be able to infect organisms down the evolution ladder. The fact is; Poliovirus is a strict human pathogen that can sometimes infect closely related primates. IRES-mediated take over and virulence are not the main reasons for this strict pathogenesis but is considered among the important explanations.

Interestingly, some cellular mRNAs were found to have the ability to drive translation from an upstream near-cognate sequence in a cap-independent fashion such as p53 [55], bFGF2v[56], OSTEOPONTIN [57], OCT4 [58] and p15^{INK4B} [59]. The term

IRES is now less strictly used to describe cap-independent translation from a non-AUG codon upstream of the canonical AUG initially thought to be the IS. If we consider that it is upstream the open reading frame (ORF), IRES now is not so internal. This means that the word Internal in IRES now might be used to refer to the ability to start translation from sequences within 5'UTR dispensing the usual need for the “terminal” 5' cap [46], [60]. In contrast, studies on some genes observed that translation of some genes was initiated alternatively from an AUG downstream of the canonical AUG following a cap-dependent fashion. Only one speculation was reported for the etiology of alternative IS; Leaky scanning [46], [61], [62]. This hypothesis claimed that the reason for such non-canonical pathway for initiation is because of the inability of the scanning ribosome to start translation from a downstream AUG or a near-cognate sequence (GUG or CUG).

The diversity generated from the aforementioned mechanism can be huge. This is easier to fathom in genes within which IRES activity was observed. The implication of the presence of IRES or alternative initiation is profound. Some examples will be discussed to appreciate how, only through alternative initiation of translation, different protein isoforms are generated. Subsequently, this diversity results in an altered function and eventually a different phenotype.

p53, the master guardian of the cell, is a very important protein implicated in many cellular activities such as: cell cycle arrest, tumor suppression, apoptosis and senescence. It is evident that p53 is implicated in distantly related cell functions. The canonical pathway of translation is the prevalent option for p53 expression where the cap-dependent machinery starts from an AUG within a favorable sequence context [55]. During the course of a very extensive study, translation of this crucial tumor suppressor was found to initiate from a downstream AUG using IRES pathway resulting in a 47KDa protein. The implications of such discovery widened our understanding for *p53* regulation and function. p53 acts through homo-tetramerization to form an active complex. The dwarf p53 known as p53/47 can still integrate into these complexes. Yet, it confers new characteristics for the new tetramer [63]. The truncated form is deprived of the MDM2 binding domain. MDM2 binds to p53 and recruits U3 ubiquitin ligase to induce degradation of p53. The resultant complex –being the active form of p53- now functions

in terms of the proportions of each form where proteins with higher levels of p53/47 showed increased resistance to apoptosis for instance [63]. The change in the balance between the two isoforms also yielded very different phenotypes due to the alteration of p53 responsive genes. Finally, it was not surprising to find that the cell has the capacity to shift from the full length to the shorter form in stress conditions. The battery of response genes, generated through the activation of the p53 IRES to produce the dwarf p53/47, is thought to have an implication in the resistance to the stress condition [55],[63]. Moreover, Ray *et al.* proved that there's an intriguing difference of isoforms prevalence in different cell cycle phases. p53 was found elevated relative to p53/47 in G2-M while p53/47 prevails in the G1-S transition [55],[64].

OCT4 sits at the top of the hierarchy of Embryonic Stem Cells' (ESC) pluripotency regulators. OCT4 knockdown results in an inevitable loss of pluripotency in ESC [65]. It has been known for some time now that OCT4 has alternative splicing-generated isoforms; OCT4A & OCT4B [58], [66]. In a recent study, OCT4 was investigated for being controlled also at the translational level. It was concluded that OCT4 generates four isoforms from a single splicing isoform; OCT4B [58]. The fact that the gene produces diversity through translational control is, by now, no more surprising. It is the diverse functions that a single mRNA generates that will keep us amazed. In a very robust group of reporter experiments, OCT4B-190 – the isoform generated by IRES utilization and IS at 190bp in mRNA – showed a preferential initiation in stress response in contrast to a very decreased level of expression of other post-transcriptional and translational isoforms [58]. At the same time, OCT4B-190 showed an increase of the exquisite ability of OCT4 to resist apoptosis [58]. Taken together, this scenario and similar gene responses are very important in ESC. ESC stress-induced damage – if it was, otherwise, that easy to differentiate or die – might be devastating for embryos exposed to any kind of stress even if subtle, defying all notions of survival and continuity [58].

Other examples include bFGF2 [67] (involved in proliferation, cancer formation and stem cells development. It is also considered the prototype of cellular alternative

initiation of translation), OSTEOPONTIN [68](the major immunological signal in Dendritic cells) and p15^{INK4B} [59](an important piece of the senescence puzzle).

bFGF2 does not generate any isoforms through alternative splicing. Yet it was discovered that there are six isoforms generated through alternative initiation of translation. The 34KDa and 18KDa are generated through a cap-dependent mechanism. While, the 24, 22.5, 22 and 16KDa isoforms are generated through IRES mechanisms[56], [67]. The interesting trivia here is the unique properties of some of the isoforms. 34, 24, 22.5 & 22KDa bFGF2 isoforms induce cell transformation through a receptor independent pathway but does not promote cell migration in transfected cells although its expression is associated with prognosis in prostate and pancreatic cancer [56], [69]. On the other hand, the only secreted 18KDa bFGF2 induces cell transformation, proliferation and transformation by binding to bFGF2 receptors on cell surfaces [56]. Osteopontin (*OPN*) is another very important example in which the utilization of another start site generates a distinct phenotype. When a truncated OPN is generated through activation of a downstream AUG from the canonical IS, conventional dendritic cells tend to grow podosomes and migrate. On the other hand, the interaction of the full form from plasmacytoid dendritic cells results in secretion of Interleukin17 from T-helper cells following its maturation [57]. Yet, there are other examples of alternative initiation of translation that we still do not have a clue about the significance of their existence. A very good example is the generation of p15^{INK4B} and p15.5^{INK4B} from the same mRNA molecule. There's no known difference between the two isoforms in localization, biological function or response to stress. However, their existence was substantiated in a very neat study [59]. We should learn from our previous lessons and not jump to conclusions of denial of any significant difference between those isoforms. The magnificent examples we have just revisited teach us that sometimes there's a very fine line of discrimination between some isoforms. It might be our narrow scope, limited technology or inefficient methodology that's keeping us away from this elucidation.

1.3.5. Post-translational modifications

After a gene passes the hurdles of starting transcription, bind to an allowed-to-transcribe PolIII, generate pre-mRNA, produce shuffled mRNA molecules, some of which

avoid interfering short RNA molecules to reach a filtered mature mRNA that's still capable of diversifying through initiating translation from atypical IS, there remains one level of regulation for the generated proteins from all those levels of control; Post-translational control.

Briefly, a polypeptide sequence can be acted upon by modifying enzymes attracted by special motifs encrypted within - what we considered – undermined genomic sequence of the gene. These modifications include phosphorylation, ubiquitination, sumoylation, acetylation, glycation, oligomerization and peptide cleavage. The consequences of which include folding, activation, inactivation to act as switches, shuttling to different cellular compartments and labeling for degradation. A very good example is p53, which necessitates phosphorylation for its activation following DNA damage and acetylation to be able to recruit acetylases such as CBP/p300 to p53-activated promoters. Both acetylation and phosphorylation are thought to protect p53 from binding to MDM2. MDM2 binding recruits E3 ubiquitin ligase that labels p53 for proteasomal degradation [70].

1.4. Cobra1; from an undermined cofactor to a pluripotency regulator

It is evident now that the aforementioned mechanisms of alternative initiation are crucial for a functional cell in many aspects such as tumor suppression, cell cycle regulation, immunological responses and stress responses. For example, the cellular stress response is not just a matter of expressing stress resistant genes such as Heat Shock Proteins (HSP) or unshackling the apoptotic machinery but it involves also the recruitment of diversifying mechanisms within the genome to induce a stress reflex even on the single gene level. This intrigues us to consider more the possibility of generation of isoforms through alternative initiation of translation in light of how crucial or dispensable the investigated gene is. Hence, it is – now – of no doubt that we need to get to know more about COBRA1 before investigating the biology surrounding its expression. The unraveling of which might in turn solve puzzles and paradoxes correlated to its expression.

COBRA1, CO-factor of BRCA1, was first identified through yeast two-hybrid system associated with the BRCT domain of the famous breast cancer susceptibility protein; BRCA1 [71]. It was indicated that COBRA1 is able to induce similar reorganizational changes elicited by BRCA1 to the chromatin structure. Meanwhile, a different group characterized the same protein to interact with the C-terminal domain of RNA PolII acting within a bigger complex to stall the bound PolII. COBRA1 was referred to – by this group – as the B-subunit of the negative elongation complex (NELF-B) [72]. Studies followed to explore the new mystery molecule that appeared in a very hot niche; regulation of transcription through BRCA1 pathway regulation and RNA PolII stalling. In 2004, Aiyar *et al.* showed that COBRA1 and its NELF brothers are able to bind/repress transcription of specifically regulated by Estrogen through its nuclear receptor Estrogen Receptor α (ER α). The findings offered the first real physiological substantiation for RNA PolII pausing role in regulating hormone responsive genes with COBRA1 as a very important player [73]. The study also helped us take out COBRA1 from the undermined role of being BRCA1's cofactor. COBRA1 overexpression in BRCA1-deficient cells showed similar repression of Estrogen responsive gene activation. COBRA1's role through NELF complex was also substantiated when knockdown of the – *in vivo* proven – interacting NELF-E abolished COBRA1's repressive functions. It might have sounded presumable that a gene regulating hormone responsive transcription would definitely affect the development and morphogenesis of breast cells. Yet, Aiyar *et al* ran three-dimensional cell culture experiment – to mimic to the greatest extent the microenvironment of the mammary stroma – to compare between estrogen dependent cell growth in control and *COBRA1* knockdown cells (augmented with NELF-E knockdown). While there was no difference between control and knockdown in absence of Estrogen, T47D cells showed increased proliferation and subsequent cell sizes in Cobra1 knockdown cells. This finding opened a new venue for COBRA1 to be characterized as a putative tumor suppressor [73]. Estrogen response was not the only hormonal effect affected by COBRA1. In 2007, COBRA1 was found to bind the hormonal receptors: Androgen, Glucocorticoid and to a lesser extent Progesterone receptors. The molecule we are discussing for the diversity of its translational products was also recognized in that same study as the first co-regulator of alternative splicing of the hormone responsive

genes [74]. Moreover, COBRA1 was found to regulate not only single promoters of genes, but also clusters of genes specially those of the chromosomal loci 21q22 and Xp11. The gene clusters were regulated in estrogen dependent and independent fashion [75]. What's intriguing was the fact that the members of the regulated/repressed gene cluster specially at the locus 21q22 were associated with breast cancer and in some instances with metastasis to bone[75].

By then a growing number of evidences was converging to direct the interested investigators to walk the path of characterizing COBRA1 role in tumor suppression. The ground for such hypothesis was laid by the aforementioned studies; COBRA1 is a cofactor of a cancer susceptibility gene, stalls RNA PolII of selected genes, interacts with steroid hormone receptors and its absence led to increased growth and proliferation rates of cancerous cell lines. The first evidence encountered in the pursuit of COBRA1's role in tumor suppression were concluded from Immunohistochemistry (IHC) experiments on mammary epithelia. While mammary epithelia showed intense COBRA1 staining relative to myoepithelia/stroma [73], [76], cancerous mammary epithelia showed reduced staining relative to normal epithelia [76]. Upon analyzing tissues from breast cancer patients, COBRA1 depletion or decreased levels were observed in patients with poor prognosis or with metastatic tumors.

In contrast, it was shown in previous studies that COBRA1 levels in epithelia of upper gastrointestinal tract tumors were elevated relative to their normal counterparts [77], [78]. Although this finding is paradoxical: it is not the first time to find a tumor suppressor elevated in an oncogene-like pattern in tumors. A similar paradoxical finding was observed with p53 analysis in some tumors. Later, it was proved that the oncogenic p53 had a point mutation and caused the associated cellular overgrowth. Normally, p53 is a tumor suppressor and is considered the master guardian of the cells. Showing that there was more than one p53 solved the puzzle. Hence, the pursuit we plan to embark upon, to characterize whether there's more than one COBRA1, might solve the paradoxes surrounding COBRA1 and its correlation to tumor formation.

Other very important functions have been attributed to COBRA1 expression. Response genes due to deletion of COBRA1 overlapped with those of BRCA1 deletion.

This indicated that COBRA1 collaborates with BRCA1 to regulate transcription of their common target genes [79] coinciding with previous findings [71], [73]. Its role –as a major transcriptional regulator in stalling PolIII for activation or repression- was also substantiated in a recent study [12]. It also showed a very important consequence for COBRA1 ablation; shortened survival and cell death. This coincided with a very important study that came out in 2009 and gave COBRA1 a very important and different prospect. Amleh *et al.* has proved in a neat unconventional knockout system through a Cre-LoxP system that Cobra1 deletion led to the failure to development of the embryo post-implantation; Cobra1 deletion is embryonic lethal [80]. Mouse models heterozygous for COBRA1 showed normal growth and survival though. In ESC, knockdown of COBRA1 resulted in the loss of pluripotency and differentiation albeit the reduction was 50% only. This differentiation took place although the levels of the master regulation of pluripotency were still intact. The results have opened a new venue for the investigation of the biology surrounding COBRA1 [80].

The presence of COBRA1 at a junction between cancer and stem cells makes it a very interesting molecule to follow as a turning point in the fate of normal stem cells to malignant stem cells and normal embryonic stem cells development.

Currently, studies – in action – are pursuing the exact molecular pathways that *COBRA1* might be involved in. The physiological processes under investigation in light of *COBRA1* expression are cancer development in general and through cancer stem cell formation, mammary gland development in addition to longevity and premature aging [unpublished data].

2. Significance, Scope and Design of the study

During the aforementioned investigations, western blotting (WB) was an indispensable molecular technique for COBRA1 analyses. Western blotting can be tricky sometimes though. One of the most common caveats created by antibodies used in this technique is non-specific binding. It results in detection of proteins of non-specific sizes. This was the case with WB using COBRA1 antibodies. Only this time, the non-specific bands were very reproducible and consistent in different samples and lysates. This fact intrigued Dr.Jianlong Sun and Haihui Pan in Dr.Rong Li lab to initiate a group of point mutagenesis experiments to detect if there's an altered codon used for initiation of COBRA1 translation. The result of which was the elucidation of a mutation site at 134bp of the full mouse *Cobra1* transcript that abolished completely the expression (Figure.1). It was concluded that this site might be a site for translation initiation for *Cobra1*. This site was different from the ATG codon thought to be the canonical site within a favorable context (Kozak sequence) at position 281bp.

Following this discovery and with collaboration with Dr.Li's lab, a study was designed to fully characterize and identify the exact isoforms based on the fact that the newly characterized codon might give rise to an isoform different from the one initiated from the typical ATG. The elucidation of such phenomenon – if proven to exist – might be the key to solve some of the questions about COBRA1. Dealing with one isoform separately will have totally different implications on the investigated processes as indicated in examples of other proteins discussed earlier within translational control (Figure.2).

The study is divided into three different phases. The first approach is to mine the 5'UTR of mouse and human COBRA1 to investigate if it has the properties of IRES to alternatively initiate translation from the atypical site. The second phase is to investigate whether those isoforms occur endogenously in mouse tissues and if a link to any physiological function can be made with the detected isoforms. The final phase is to ectopically express both isoforms by cloning different fragments of *Cobra1* into mammalian expression vectors to investigate if the generation of isoforms is possible *ex-vivo*.

3. Materials and Methods

3.1. In-Silico analyses

The sequences of the full transcripts, of both mouse and human, were obtained from NCBI using the gene IDs 58202 and 25920, respectively. Both full transcripts (human and mouse mRNA sequences retrieved from NCBI database) were submitted to blastn in a multiple alignment interface to determine the degree of similarity between both sequences. The 5'UTR was obtained from upstream of the first canonical ATG and applied to the GeeCee web interface to the database to determine its GC content and the degree of similarity between both of them.

The amino acid sequences of the 5'UTR of both species were submitted to the signal peptide databases: WOLF Psort, SIG-Pred, SignalP 4.0 and SPdb.

For prediction of the secondary RNA structure, two approaches were used. The first used the software CARNAC. This software assumes evolutionary constraints in the submitted sequences and constructs secondary RNA structures compromising –only to acceptable limits – the enthalpy of folding. The second approach was done using the Mfold web server. This approach outputs possible secondary RNA structures with the least enthalpy in an absolute manner devoid of any comparisons.

3.2. Bacterial strains and culture

The bacterial strain used for transforming all constructs was *E. coli* BL21. The bacterial cells were rendered chemically competent before transformation.

Untransformed cells were cultured in LB broth or Agar while transformed cells were cultured in LB supplemented with 100ug/ml Ampicillin.

3.3. Polymerase Chain Reaction (PCR)

Colony and conventional PCR was performed using the following reagents: *Taq* polymerase (NEB, M0273L), dNTPs (Roche, 1581295) and primers from iDT DNA technologies. Cloning for expression included using the high fidelity polymerase *Pfu* (Biovision, 9003-2500). Primers used during the study are included in table.1. The PCR program used in all runs except overlapping PCR is; Step1 is 94° for 5', Step2 [94° for 30sec, Annealing temperature for 30sec, 72 for 1min] for 35 cycles and Step3 72° for 5'. Assembly PCR was carried out by including the overlapping fragments (1:1) with the

whole PCR recipe but not the primers. Step2 was adjusted to three cycles and step3 cancelled. After the end of the third cycle, primers were included and amplified where step2 was repeated for 25 cycles.

3.4. Restriction digestion

For verification of cloning or insert, 200ng of the plasmid were incubated with 1ul of the restriction enzyme used equivalent to 20U EcoRI (NEB, R0101S) and 10U Sall (NEB, 0138S). BSA was added to the buffer prior to incubation at 37°C for 2-4 hours. For ligation and cloning, 10ug of plasmid DNA were incubated with 2ul of the same enzymes mentioned previously at 37°C for 20 hours.

3.5. Ligation and transformation

Restriction digestion products were checked using Agarose gel electrophoresis. Expected bands were excised and extracted using gel extraction kit (Qiagen, 28706). The eluted DNA was quantified using NanoDrop 2000 (Spectrophotometer). The vector used for expression in mammalian cell lines was pBABE-Puro (Cell biolabs, RTV-001-puro). The digested and purified vector and inserts were added to a ligation reaction in a 1:3 molar ration, respectively.

The ligation reaction was supplied with T4 DNA ligase (NEB, M0202S) to seal the nicks at the sticky ends and incubated at room temperature for five hours. The ligation reaction was then used to transform chemically competent bacterial cells.

3.6. Screening for Recombinants and verification of plasmids

For verification of the colonies, colony PCR was performed on 30 picked colonies using the LTR primers supplied with the vector. Two of the positive clones were cultured and subjected to plasmid extraction (Qiagen, 27104). The plasmid extracted was verified by restriction digestion (using EcorRI and Sall) and then sent out for sequencing to the sequencing core facility, UTHSCSA.

3.7. Cell culture and cell lines

For packaging of viral pseudo-particles or transfection of human constructs, HEK293T cell was used and cultured in a DMEM High Glucose, with sodium pyruvate media (life technologies, 11995-065) supplemented with 10% FBS (life technologies,

16141-079), 1/100 Non-essential Amino Acids (Life technologies, 11140-076), 1/100 L-Glutamine (Life technologies, 25030-081). For culturing purposes other than transfection and packaging, 1% of Penicillin/Streptomycin antibiotic mixture and 1mg/mL were added to the media. For freezing cell lines, 100uL DMSO was added to 900uL of single cell suspension with the desired count in its usual culture media. Immortalized Mouse Embryonic Fibroblasts (MEF) was cultured in DMEM High Glucose, with sodium pyruvate supplemented with 10% FBS and 1% Penicillin/Streptomycin. Embryonic Stem Cells were cultured in DMEM High Glucose, with sodium pyruvate media (life technologies, 11995-065) supplemented with 10% FBS (life technologies, 16141-079), 1/100 Non-essential Amino Acids (Life technologies, 11140-076), 1/100 L-Glutamine (Life technologies, 25030-081), 1/100 of Penicillin/Streptomycin antibiotic mixture (Life technologies, 15140-163), 1000U/mL Leukemia Inhibitory Factor (LIF) (Chemicon/ESGRO, ESG1107) and 0.1mM 2-Mercaptoethanol. The vessels used for ESC culture were coated with 0.1% Gelatin for at least two hours before use. All cell lines were incubated in a humid incubator at 37°C and 5%CO₂.

3.8. Transfection

Transfection was carried by plating 3×10^6 HEK293T cells in its culture media (4mL), with no antibiotics or Geneticin, on a 60mm cell culture dish one day before transfection. Next day, the media was changed with only 3mL of the aforementioned recipe. In a 0.5mL Opti-MEM (Life technologies, 31985070), a total of 12ug plasmid DNA was added (6ug if two plasmids are co-transfected in case of packaging) and incubated for five minutes. In a 0.5mL Opti-MEM (Life technologies, 31985070), 20uL of Lipofectamine 2000 (Life technologies, 11668-019) was added and incubated for five minutes. The two 0.5mL Opti-MEM preparations were mixed and incubated for 20 minutes. This was followed by addition of the 1mL mixture to the HEK293T cells drop wise. After 36 hours, part of the cells was lysed for verification of the transfection with western blotting.

3.9. Retroviral Infection

For packaging of the retroviral particles the desired plasmid and a helper retroviral plasmid were co-transfected to HEK293T cells according to the previously mentioned protocol. The second day the media was changed and the cell line was incubated for two days to enrich the media with the viral particles. A day before harvesting the viral particles and starting the infection immortalized MEF cells were incubated in a 6-well cell culture plate at a cell count 7×10^4 cells/ well. On the infection day, the viral particles were purified using a 0.45um filter and incubated with previously prepared immortalized MEF. To the supernatant containing the viral particles, polybrene (Millipore, TR-1003-G) was added with a final concentration 8ug/mL. The plate was spun at 1500rpm (Eppendorf, 5810) at 4°C for four hours. The plate then incubated overnight in a humid incubator at 37°C and 5%CO₂. Next day, the media was changed for the infected cells and incubated for 36hours before adding Puromycin selection at a concentration 2ug/ml.

3.10. Embryonic stem cells differentiation

Mouse embryonic stem cells were used (AB2.2 and 129) for differentiation experiments. The cells were cultured in a 6-well ultra-low attachment cell culture plate without LIF at a count 4×10^6 cells/ well. After 4-5 days, embryoid bodies were collected carefully, moved to a 15mL conical tube and left to settle for two hours. The supernatant was then re-suspended for immortalized MEF media and plated on a gelatin-coated plate. The cells were left to differentiate for 12-15 days before analysis.

3.11. Mouse tissue analysis

All utensils, tubes and tissues used were incubated on dry ice for 15 minutes before use. Mortars and pestles were pre-chilled by liquid Nitrogen. The cells were crushed in liquid nitrogen and the powder was moved to the pre-chilled labeled tubes using cell lifters or spatulas. The powder was weighed, re-suspended in three parts 1XPBS with protease inhibitor cocktail (PIC) (Santa Cruz technologies, sc29130) and finally lysed using 2X Laemmli Sample Buffer (LSB), (50mg tissue powder +150uL 1XPBS/PIC +150uL 2XLSB). The mixture was boiled for 10 minutes, centrifuged at 10,000x g at 4°C (very important for fatty tissues) and finally the supernatant moved to a

clean tube. Quantification of tissue and cell lysates was carried out using BCA assay kit (Pierce, 23225).

3.12. Western Blotting

20-30ug of Lysates were run on SDS-PAGE gels. The dye front was migrated off the gel followed by transferring the gel to a nitrocellulose membrane (Amersham, RPN203D) using semi-dry blotter (Bio-Rad, 170-3940). The antibodies used were mainly against Cobra1. Either polyclonal rabbit anti-mouse or monoclonal mouse anti-human was used for detection of COBRA1 signals. Both antibodies are developed in Dr.Rong Li's lab. Incubations with 10% non-fat dry milk in TBST blocking solution was done for two hours at room temperature, with primary overnight at 4°C and secondary two hours at room temperature. Both Primary and secondary were incubated in 5% non-fat dry milk in TBST. For loading control, pan-Actin (Cell signaling, 4968) was used as a loading control between similar tissue lysates. Chemiluminescent detection of the membranes was done using ECL (Pierce, 32106) or ECL plus (Pierce, 32132) substrates.

3.13. Flag-Immunoprecipitation

The buffer used for lysis (NETN buffer) had 150mM NaCl, 1mM EDTA, 20mM Tris and 0.5% NP-40. Protease and phosphatase inhibitor cocktails were added just before use. The phosphatase inhibitors used were 0.5M NaF, 200mM Na₂P₂O₇·10H₂O and 200mM Na₃VO₄. The cells were resuspended in NETN, rotated for 20 minutes and passed through a 21G needle. Flag beads (Sigma, A2220) were washed NETN buffer before being added to the lysates and incubated overnight. The next day the samples were washed four times (rotated for 15 minutes each). Two of the washes were done with normal NETN buffer followed by two washes with high stringency NETN (500mM NaCl). The beads were lysed in Laemmli sample buffer and analyzed through western blotting.

3.14. Cellular Fractionation

Two buffers were used for fractionating infected iMEF cells for both lines expressing Cobra1 isoforms. The buffer used to obtain cytoplasmic fraction had 10mM HEPES, 10mM KCl, 1.5mM MgCl₂, protease inhibitor cocktail and 1mM DTT (just before use). The insoluble fraction was re-suspended in a buffer containing 25% glycerol,

0.42M NaCl, 1.5mM MgCl₂, 2mM EDTA, 0.4mM HEPES, protease inhibitor cocktail and 1mM DTT (just before use).

3.15. Immunohistochemistry

The experimented mammary glands sections (Thickness= 0.4microns) were Cobra1 f/f luminal mammary epithelial cells' sections. A knockout was induced in the animals where these cells were obtained from via Tamoxifen injection. The experiments were carried with two types of controls: Cobra1 f/f luminal cells' sections with no primary antibody (One for the whole experiment) and Cobra1 f/f luminal cells' sections with primary antibody but was lacking the induction of the knockout with Tamoxifen (One for each dilution). Antigen unmasking solution (Vector labs, H-3300) was used before processing the samples. Using 1%H₂O₂, endogenous peroxidase was blocked. A polyclonal Rabbit anti-mouse raised against Cobra1 was used for this analysis. The binding of the primary antibody and the secondary biotinylated antibody was detected using the M.O.M IHC peroxidase kit (Vector labs, PK-2200).

4. Results

The pursuit of the alternative initiation of Cobra1 through a non-canonical start site to generate two translational isoforms was carried out using three approaches; (1) Using computationally-based predictions and analysis of the 5'UTR of the gene to see if it has what it takes to initiate translation, (2) Investigating the occurrence of this phenomenon endogenously in mouse tissues after validation of the antibodies used for the analysis and finally (3) optimizing the conditions in *ex-vivo* expression systems to substantiate what was concluded from prediction and endogenous observations.

4.1. Dissecting the sequence of the 5'UTR of human and mouse Cobra1.

To elucidate the significance and importance of the sequences within the 5'UTR of Cobra1 in human and mouse, we aligned both sequences in search for constraints or any conserved sequences. The sequences of both mouse and human Cobra1 5'UTR showed 70% of similarity indicating the existence of a relatively high evolutionary constraint between the sequences as shown in Figure.3. It's worth mentioning that aligning between 500bp upstream of exon-1 of *Cobra1/COBRA1* from mouse and human yielded no significant similarity. Furthermore, both sequences showed a very high percentage of GC content specifically 84% and 71% in human and mouse 5'UTRs, respectively. Zooming in to the specific loci of the detected near cognate start site and the corresponding site in human 5'UTR showed very similar context as shown in figure.4. Moreover, these sequences showed a similar and highly conserved predicted secondary RNA structures. The output of CARNAC folding software, which assumes evolutionary constraints between the submitted queries, showed very similar secondary RNA structures as illustrated in figure.5. To confirm the results generated by CARNAC, Mfold folding web interface, which does not assume evolutionary constraints of the submitted queries, was used. Mfold generated similar secondary RNA structures for the separately submitted 5' UTR upstream of the suspected CTG (Figure.6). The possibility of folding into the generated structures was strengthened through the low energy of folding. The change of free energy calculated by the software were -55.03J and -17.94J for mouse and human sequences, respectively. To investigate whether the peptide coded by the extra 144bp in mouse sequence generated a signal peptide responsible for sorting the protein to

its known nuclear position, the 5'UTR sequence of mouse was aligned against four different databases yet yielded no predicted activity for sorting (figure.7).

4.2. Detection of endogenous Cobra1 isoforms

The second phase started by searching for a method to validate if the antibody used can detect only COBRA1 signals or relatively specific when compared to non-specific signals. Thanks to the inducible knockout mouse models developed in Li lab, we were able to test these antibodies by comparing tissues from control against knockout animals. We used COBRA1 antibodies in Immunohistochemistry and showed its very high specificity (figure.8). The rabbit anti-mouse antibody did recognize the signals of the antibody as indicated by the brownish color of the hydrolyzed TMB substrate in control samples. The signals stained specifically the nucleus substantiating more the fact that the detected signal is COBRA1 as reported in previous studies [73]. The *Cobra1*-knocked out mammary gland showed only the blue color of the counter stain Hematoxylin. Further analysis was needed to confirm that this antibody has relatively high affinity for COBRA1 protein in western blotting. A battery of western blotting experiments was run (aided by qPCR results from Haihui Pan) on pairs of tissues from control and knockout mice (littermates) to decisively indicate that the isoforms hypothesized are not a matter of antibody non-specificity. During these experiments we were able to identify the investigated isoforms. Because we were using an inducible knockout system, the knockout model had –sub-optimally– decreased levels of COBRA1 as indicated by qPCR results and not completely abolished ones. This was beneficial for us because it decreased overlapping between the signals of the two isoforms on western blots. Intestine lysates showed –especially in lower exposures – very clear positive and negative results, in the 53-79KDa region of the molecular weight marker, indicative of Cobra1 size. On higher exposures, the decreased levels helped us identify both COBRA1 isoforms (Figure.9.a). The bands in higher exposures in knockout lane were of the same size of a larger band in the control lane indicating that COBRA1 doublet might have been overlooked due to the overlapping between the two isoforms. The unaltered protein levels of COBRA1 in the spleen deduced from only 10% decrease of *Cobra1* mRNA (qPCR results) was reflected in western blotting results that showed no difference between control and knockout (Figure.9.b). Pertaining our quest, we couldn't detect COBRA1

doublet in mouse spleens. On the other hand, analysis of pairs of different tissue lysates showed the sought for isoforms (COBRA1 doublet) especially in kidney and liver (figure.9.c) with an increased abundance of the 70 KDa isoform relative to the 65 KDa (i.e. The isoforms are referred to here as 65 & 70 KDa for the sake of facilitating the discussion. The sizes are only based on computational predictions and await substantiation through mass spectrometry). Paradoxically, when a different strain of wild type mice tissues were used (not subjected to knockout manipulations or drugs), kidney lysates showed an increased levels of the shorter isoform in contrast to the predicted pattern in spleen and testis lysates (Figure.10). WB on kidney lysates was one of the most reproducible results and showed both isoforms clearly (Also, kidney showed 80% decrease in *Cobra1* mRNA levels in Cobra1- knocked out tissues as reported in qPCR results). We tried other approaches to clarify the existence of the different isoforms, so we analyzed kidney lysates from old and young mice (the mice ages were not available as the tissues were supplied by the Barshop longevity center, UTHSCA after characterization). Based on previous reports, some genes shift to different isoform as an effect of aging. Although both isoforms were detected in old and young kidney lysates, there was no alteration of expression among Cobra1 isoforms (Figure.11).

4.3. Confirmation of the existence of Cobra1 isoforms using *ex-vivo* gene expression

To validate the results obtained from the prediction analysis and endogenous observations, *ex-vivo* expression systems were planned as illustrated in figure.12. Amplification of mouse and human *Cobra1/COBRA1* sequences was carried out in two steps due to complications offered by the GC rich content (Figure.13). The subsequent complex RNA secondary structure urged us to amplify the 5'UTR and coding sequence separately but with an overlapping fragment between them. After purification, both fragments were overlapped through assembly PCR to finally yield the full transcript as shown in figures.13 &14. Cloning and bacterial transformation were confirmed by colony PCR (figure.15a) followed by plasmid extraction and verification by restriction digestion (figure.15b) before sending the plasmid out for sequencing.

After verifications of the plasmids and quantification, transfection using the human constructs of *COBRA1* to the human cell line HEK293T. Three constructs were

used: An empty vector (pBABE-EV), a coding sequence (pBABE-ATG) and (pBABE-FL). The transfectants showed different sizes of expression of COBRA1 and no signal in case of empty vector (figure.16). The isoforms were detected using anti-flag antibody to exclude the fact that the signal is because of the endogenous COBRA1 expression. Still, the membrane was stripped and detected using anti-COBRA1 to verify COBRA1 identity. To simulate this experiment using mouse sequences, retroviral infection was required due to the resistance of iMEF to liposome-mediated transfection (passive). Pseudo-retroviral particles – containing the three constructs but using mouse sequences this time – were packaged in HEK293T and their lysates used in WB to ensure successful packaging. Retroviral infection (active) using the purified pseudo-retroviral particles of immortalized MEF showed different sizes of Cobra1 proteins before selection (Figure.17.a). We selected for the successfully infected cells using puromycin and re-run the WB using the packaging lines' lysates as controls in both cases (Figur.17.b). Eventually, we established three mouse cell lines stably expressing the desired constructs. Because all of the expressed proteins had flag tags, we were able to pull down the proteins for further analysis. Using Flag beads, the different isoforms were successfully pulled down and purified (Figure.18). Unfortunately, the results of Mass Spectrometry are still in progress and including them in this manuscript wasn't possible because of the limited timeline of the study.

Since COBRA1 was previously linked to maintenance of ESC pluripotency. Our last approach to characterize Cobra1 isoforms was to detect Cobra1 expression during the differentiation pathway of embryonic stem cells (Figure.19.a). We were fishing for any alteration of the pattern of expression of the COBRA1 doublet. However, COBRA1 doublet was detected in ESC, embryoid bodies and terminally differentiated MEF cells in the usual pattern; high 70:65 KDa ratio (Figure.19.b).

Although we ran the extra peptide (5 KDa) through many signaling peptide databases and found no matches, we were still interested in verifying the computational prediction. The infected iMEF cells were fractionated into cytoplasmic and nuclear fractions. Surprisingly, we could identify the full length tagged COBRA1 in the

cytoplasmic fraction. In contrast, we identified the shorter isoform signal in the nuclear fraction. The preliminary data for the fractionation experiment is shown in figure.20.

5. Discussion and Conclusion

There is a growing interest in the role of different protein isoforms and their different unique properties. This was mainly because of the existence of some unknown isoforms for certain biological or chemical compounds that had devastating effects in some instances. Even the subtlest differences, as in stereoisomers, can implicate profound differences in function. In 1961, the drug thalidomide was retracted from the market due to its correlation to birth defects in a tragedy known now as the “Thalidomide crisis”. Thalidomide was a drug used as a sedative and in morning sickness. The only problem is that thalidomide wasn’t one molecule; it was a twin. The racemic mixture of thalidomide contained R-Thalidomide – the morning sickness drug – but also had in equal amounts the teratogen S-Thalidomide that mainly caused “Phocomelia”. Many studies tried to figure out why did this isoform had these obnoxious effects. It was hypothesized that only the S-enantiomer – but not the R-enantiomer – was able to integrate into the major DNA groove and inhibit promoters of developmentally crucial genes such as *IGF1* and *FGF2* [81]. When the existence of unknown protein isoforms is suspected in a gene that is developmentally crucial, a putative tumor suppressor or a transcriptional regulator, it becomes a must to clarify any ambiguity surrounding this existence and any related functions. *Cobra1* is a crucially important gene for development, a putative tumor suppressor and a transcriptional regulator. However, it was found elevated in upper GIT carcinomas. These functions and paradox made it indispensable for Li and Amleh labs to investigate if COBRA1 has a twin brother. The presence of a different isoform or altered ratio between both isoforms might be adding another level of regulation to COBRA1’s functions. Point mutagenesis studies revealed an atypical near-cognate start site in the 5’UTR upstream the once-thought typical start at a cognate ATG. We fished for the twin Cobra1 in this study by three approaches; *In-silico* prediction analyses, endogenous observation and *ex-vivo* gene expression. Our computational prediction revealed that the 5’UTR of *Cobra1* has what it takes to initiate translation from an atypical start site

independent of mRNA cap. One of the golden standards now used to find out and verify exon existence or coding abilities of genome loci is the existence of evolutionary constraints among different species. Evolutionary constraints within 5'UTR of mammals have been estimated not to exceed 4% of similarity[4]. This implicated that these regions have relatively reduced importance, which coincided with their nature being un-translated regions. Alignment of 5'UTR of both mouse and human *Cobra1/COBRA1* showed 70% similarity. Even on the codon scale, the context of the detected near cognate start codon was conserved. This allowed us to conclude the importance of the 5'UTR region of *Cobra1/COBRA1*. This importance was hypothesized to be related to IRES activity through two main characters: high GC content and complex secondary RNA structure. The GC content of the mouse and human 5'UTR of *Cobra1/COBRA1* was very high: 84% and 71%, respectively. This was translated into a complex RNA structure enough to dock ribosomes to initiate translation from the suspected near-cognate codon. It is worth mentioning that we made use of the predicted evolutionary constraint between the two species in the generation of secondary RNA structures. The mere fact that the CARNAC algorithm was able to generate similar secondary structure –when considering an evolutionary constraint – of the two species indicates that this pattern of folding is plausible. Mfold –in turn – confirmed these results but with more zooming on the region upstream of the hypothesized near cognate start. In conclusion, our *in-silico* prediction analyses certifies *Cobra1*'s 5'UTR for the ability to initiate translation independent of the 5'cap and upstream of the cognate AUG. It was then intriguing to explore mouse tissues – being ethically and technically more accessible than human tissues – for the existence of more than one isoform of *Cobra1*. For proteomics professionals, it is well known that “shadows” or non-specific bands can be detected in western blotting (in polyclonal antibodies even more than monoclonals). We tried to rule out the possibility that the hypothesized isoforms of COBRA1 are not the artifact of COBRA1 antibodies' non-specificity. COBRA1 polyclonal and monoclonal antibodies showed high specificity in detection of *Cobra1* deletion in the epithelia of mammary glands of an MMTV-Cre knockout (mammary gland specific) mouse model for *Cobra1* deletion. The fact that antibody stained the nucleus and minimally cytoplasmic margins also reflected very high specificity (i.e. based on earlier reports that COBRA1 localizes mainly in the nucleus

[73]). Although the background generated by the antibody in IHC was minimal, the application of this finding to WB needed further substantiation. This was the part when we turned to the ER-Cre inducible whole body Cobra1 knockout. Simply, when we have a pair of two similar tissues (same type, from the same sex even from the same littermates), with the knockout induced in only one of them, we can see the specific COBRA1 signals disappear from the knockout lysates in our WB analysis. While trying these experiments, high exposures of WB on intestine lysates (control and knockout) showed the long sought-for “COBRA1 doublet”. A doublet that completely disappeared in lower exposures of knockout lysates but not control lysates. The doublet detected in the tissue with the decreased levels of Cobra1 gave us a hint why this phenomenon might have been overlooked until now. Especially in higher percentage gels, the signals from both isoforms can overlap and appear as a thick band and not a doublet.

The specificity of the antibodies used by then was no more in question. A battery of western blotting experiments was run on many tissues to elucidate the isoforms every tissue expresses. It is worth mentioning that the control and knockout mice, from which we got the tissues for analysis, always showed an increased level of the 70 KDa isoform relative to the 65 KDa. The tissues analyzed included kidney, spleen, heart, liver and small intestine (duodenum). A very interesting twist of events came about when we analyzed tissues from another strain of mice that was not injected with Tamoxifen (Control mice used previously were injected with Tamoxifen to mimic the knockout induction in the experimental mice). Kidney tissues showed an increased abundance of the shorter 65 KDa isoform. This might be due to the different strain used or absence of the Tamoxifen stress. It was reported for some proteins such as p53 that a special isoform (p53/44) appears more abundantly in older organisms [82]. So, we tried comparing the expression of six pairs of old and young kidney tissues but we couldn't find any difference in the relative expression of COBRA1 isoforms. Because tissue lysates are swarming with different proteins, we tried substantiating our results with an *ex-vivo* expression system. In this system we use the full Cobra1 transcript and a truncated one (missing the 5'UTR) to investigate if we can express the different isoforms in different systems. It is worth mentioning that the antibody used in these experiments was against

the tag and, not only Cobra1, to make sure that the detected form is from the induced expression and not the endogenous forms. Both human and mouse expression systems showed two isoforms of tagged-Cobra1. Interestingly, the Flag antibody used to detect the tag could pick up a shadow-like signal synonymous of the one detected by Cobra1 antibodies with the shorter isoform (Figure.20). The signal of the short isoform detected, in the cell line infected with the full transcript, was considered another proof that the antibody is in fact specific and a very important point in the favor of the presence of “COBRA1 dwarf isoform”. In recent reports, some proteins were reported to alter their specific isoforms expressed according to cell fate. That is, the terminally differentiated cell would, still express a protein that was expressed in its predecessor ESC, but have a different isoform of that protein[83]. We cultured and differentiated ESC through a protocol that did not include the addition of any growth factors to avoid the inclusion of more variables to the experiment. Although three of the differentiation phases showed the COBRA1 doublet, we were never able to detect a change of the high 70 KDa: 65 KDa ratio. Even in our mouse tissue analysis, we could only observe a high 70:65 KDa ratio. This might be due to the fact that bound and elongating ribosomes –from an upstream IS – might inhibit the initiation from IRES or even a typical cognate ATG. This might be a reason of the now “activated for translation” mRNA. In turn, this makes it very hard for another machinery to bind to a site within the elongation path of the activated machinery. This was very evident when we cloned only the coding sequence and the results were the successful expression of the short form. This indicates that the deletion of the 5’UTR and its intrinsic initiation abilities gave way for the initiation machinery to start very robustly from the cognate ATG. In some preliminary fractionation experiments, the 70 KDa form showed preferential localization to the cytoplasm while the 65 KDa was mainly nuclear. A possible explanation is the fact that the protein needs to be cleaved to be able to shuttle to the nucleus to do its known functions. Another less probable reason might be an undiscovered role of COBRA1 cytoplasmically. However, this does not fit well with the high 70:65 KDa ratio that would, otherwise, imply an increased activity in the cytoplasm versus nucleus.

In conclusion, we have substantiated the hypothesis and preliminary data suggesting the existence of two isoforms for *Cobra1*. Prediction software helped us conclude that *Cobra1* sequences, and the subsequent secondary RNA structure, has what it takes to generate two COBRA1 isoforms. We used mouse tissue lysates from different strains of mouse to mine for the hypothesized forms and could eventually observe their presence. The experiments on tissue lysates needed substantiation with artificial expression systems. Through these systems, we have shown that it is possible to generate both forms in separate lines.

Although we could prove the presence of the two isoforms (computationally, endogenously and through ectopic expression), we did not observe any difference in the function (not the scope of this study) of both isoforms. For instance, in the mouse infected cell lines, we couldn't observe any change of morphology or rate of growth. Nonetheless, we can postulate that the reason *Cobra1* is using such a non-canonical system for expression/translation is the fact that it is an indispensable gene. A cell cannot afford losing such a crucial protein for stress that would, otherwise, destroy the machinery needed for the canonical, sophisticated, cap-dependent, ATG initiated translation. This was reported for other important proteins such as p53 and OCT4 [55],[58],[63].

6. Future prospects

The elucidation of the difference of the two isoforms' functions, if any, should be at the top of the priorities of any studies to follow. This can be the doorway to the actual characterization of the two isoforms. Exact sequences retrieved after mass spectrometry would allow the exact identification of the added peptide upstream of the AUG codon. Localization studies should be the first to be done given that some preliminary fractionation attempts done during the course of this study showed an intriguing finding. Cobra1 –alleged – 70 KDa isoform showed higher abundance in the cytoplasmic fraction versus increased levels of the 65 KDa isoform in the nuclear fraction (Figure.25). Also, transfected or infected cells can be put under oxidative and thermal stress to identify any stress related alteration of expression ration of these isoforms. Moreover, reporter and deletion assays can be used to verify the ability of the 5'UTR to drive translation and to zoom in on the exact sequences required for translation initiation from the non-canonical site. Last but not least, the pulled down proteins through Flag-IP can be used to identify the interactome of both isoforms.

7.Tables

Table.1 List of primers used for cloning Full *Cobra1/COBRA1* one transcripts and coding sequences for human and mouse

Primer	Sequence
HCbr1A	CCG GAA TTC CAT CCA GGT GCG GGG CGG AAG TGG GC
HCbr1B	GCA GCC CCG CGA ACA TGG CCG AG
HCbr1C	CTC GGC CAT GTT CGC GGG GCT GCA GGA C
HCbr1D	ACG CGT CGA CAT TCA CTT ATC ATC ATC GTC TTT GTA ATC GAG CGG GGC AGG GGC GGG
HCbr1E	CCG GAA TTC CAT CCT CGG CCA TGT TCG CGG GGC T
MCbr1A	CCG GAA TTC CAT TAA TGG GTA GCA AGT CGC TG
MCbr1B	AGT TGG TCA GGG TCT CTT TCA GGT CCT CGC CAT T
MCbr1C	AAT GGC GAG GAC CTG AAA GAG ACC CTG ACC AAC T
MCbr1D	ACG CG CGA CAT TCA CTT ATC ATC ATC GTC TTT GTA ATC CAG TGT GGC TGG AGC AGG
MCbr1E	CCG GAA TTC CAT TAA AAT GGC GAG GAC CTG AAA GAG

8. Figures

• **NUCLEOTIDE SEQUENCE:**

TGGGTAGCAAGTCGCTGCTCGGAACCTGCGGGCGGCGGAGGCGTTTTCCCGGGGCCTATTGGCTGCTGGA
 CCTAGGCAGTCAAGGGCGGAAGTGGGAGGTGTGGGTCCGTGCGAGCTGCTCTGCTGACGCGATCTGGCCA
 CACTGGAGGCGGCGGGCGAACGGGGCTTGGGGGGTCCTCGTGGCACCGTGGAGAGGGCTTCTGGGGCTCC
 GTCAGGGTCCGCTACGGCTCCAGCGGAGCGCGGTGGGGATGGGGTACACAGCCGAGCGTCCGCCGGAGCC
 TCCGCCATGTTTCGCGGGGCTCCAGGACCTGGGCGTGGCCAATGGCGAGGACCTGAAAGAGACCCTGACCA
 ACTGCACCGAGCCTCTCAAGGCCATTGAACAATTCAGACAGAGAATGGCGTGTGCTGCCTTCCCTGCAA

CTG is at position 134

ATG is at position 281

Figure.1 **Nucleotide sequence of the 5'UTR of *Mus Musculus***. The typical (within a favorable context) ATG (highlighted in red) is the ATG once thought to be the canonical start codon. The near-cognate CTG (highlighted in blue) is the canonical start codon according to point mutagenesis and the outcome of this study.

• **PROTEIN SEQUENCE:**

(70KDa)
L A T L E A A G E R G L G G P
 R G T V E R A S G A P S G S
 A T A P A E R G G D G V H S
 R A S A G A S A **M** . . . 65KDa

Figure.2 **Amino acid sequence of the 5'UTR in *Mus Musculus***. The recognized CTG adds 51 amino acids to the protein sequence equivalent to the 5KDa detected in all western blotting done in this study. The start of translation from the typical ATG would result in a 65KDa protein also detected in our western blotting as a faint band.

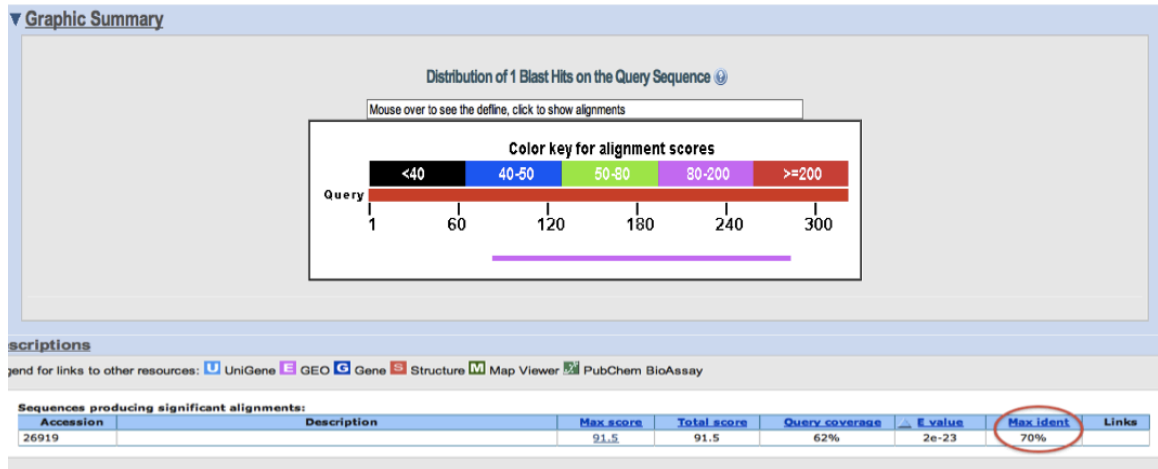


Figure.3 Blastn results for 5'UTRs alignment. Aligning 5'UTR of Cobra1 of *Mus Musculus* and *Homo-Sapiens* showed 70% similarity.

- Cobra1: TGC TGA CGC GAT CTG GCC ACA CTG GAG GCG GCG
- COBRA1: GGC GGG CGG GAC CTG GCC GAG CTG GAG GGC GCC

Figure.4 Nucleotide sequence of the initiation near-cognate loci. The context of the near-cognate CTG detected in point mutagenesis studies in mouse cell lines is similar to a CTG found in the 5'UTR of the human *COBRA1* sequence.

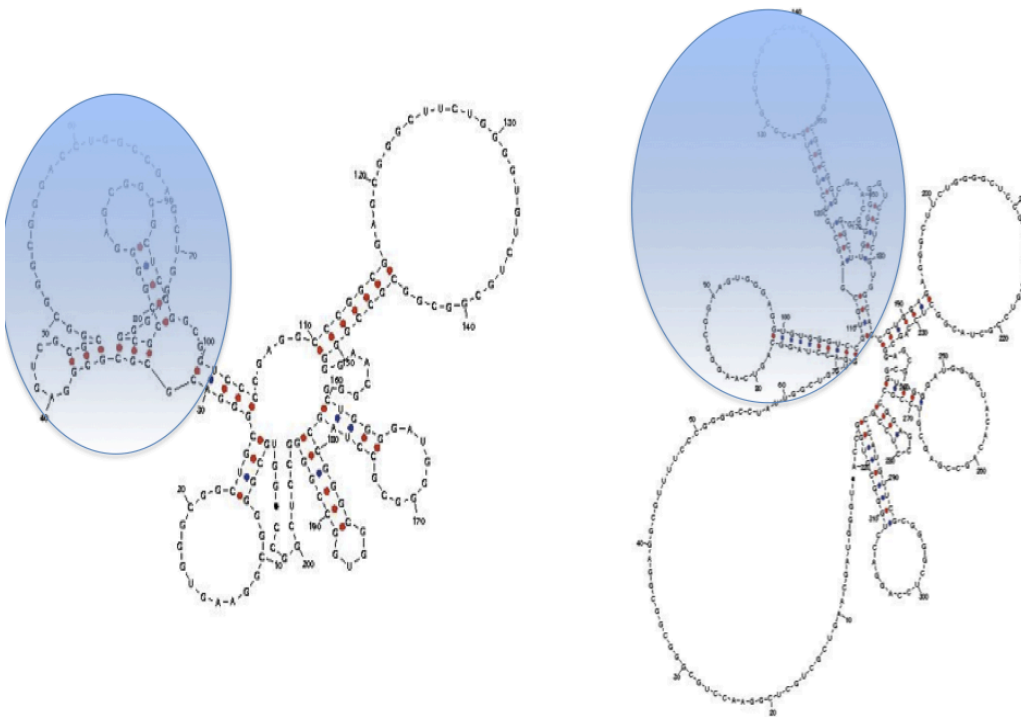


Figure.5 Secondary RNA structures predicted via CARNAC. CARNAC output implies that –in case both human (left) and mouse (right) sequence are evolutionary conserved – there might be a foldable (with low change in free energy) complex secondary RNA structure similar in both species. The highlighted sequences are for the 5'UTR regions of both species.



Figure.6 **Secondary RNA structure predicted via Mfold.** Mfold output shows comparable secondary RNA structures (Stems – loops – knots and pseudo-knots) for the regions upstream of the investigated CTG codon.

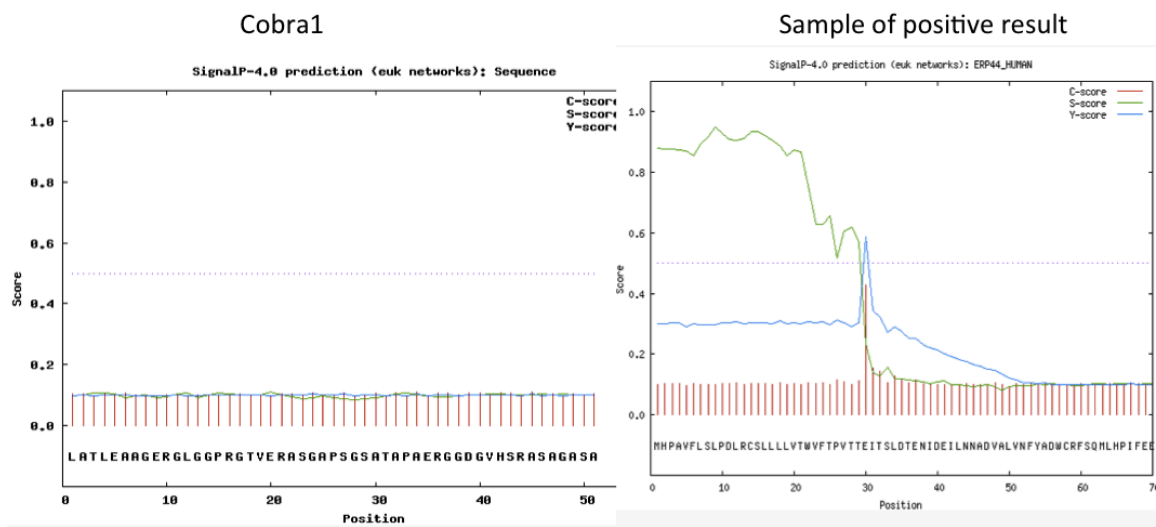
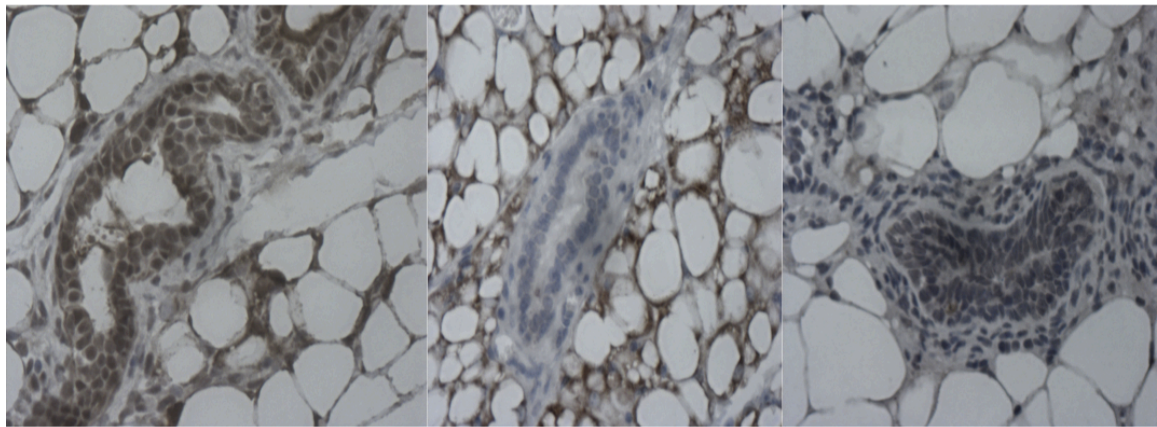


Figure.7 **Curve showing SignalP 4.0 score of COBRA1 vs. positive sample.** SignalP-4.0 output showed no predicted evidence of signaling activity for the 5'UTR of the *Mus Musculus Cobra1* sequence.



Positive Control
Cobra1^{fl/fl} / No Tam

Negative Control
No primary Antibody

Knockout
Cobra1^{fl/fl} /Tam

Figure.8 **Immunohistochemistry using COBRA1 rabbit anti-mouse antibody.** Immunohistochemistry showing highly stained mammary luminal epithelia (left) from control animals indicating positive reactivity with COBRA1 – Ab. Mammary gland from knockout animal showed no reactivity as indicated with its relatively lower staining (blue color of Hematoxylin counter stain) and when compared to a section from control mouse sample that was not reacted with primary COBRA1 antibody.

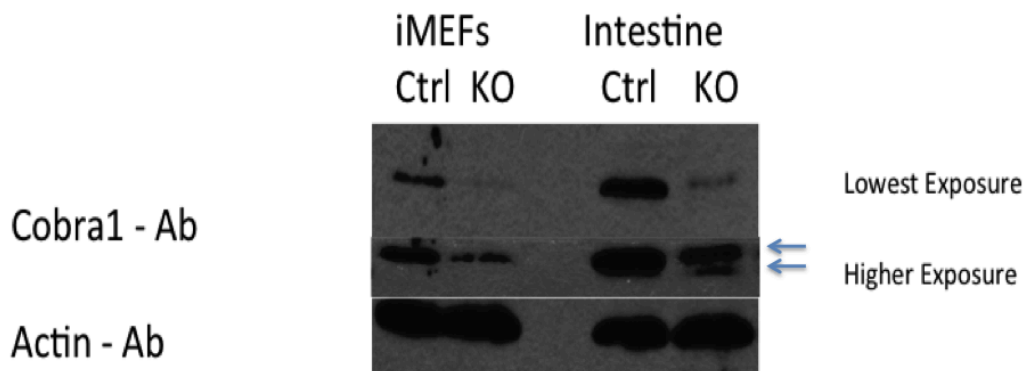


Figure.9.a **Western blotting on mouse tissues using COBRA1 anti-mouse antibody.** Intestine tissues from control and knockout mice analyzed by western blotting. All exposures indicate COBRA1 – Ab specificity. Higher exposures show the COBRA1 doublet in Intestine tissues from knockout mice. Control and knockout iMEFs were used as internal controls. Ctrl=control & KO=Knockout. Actin – Ab was used to ensure equal loading. Protein lysates were run on 17% SDS-PAGE gel.

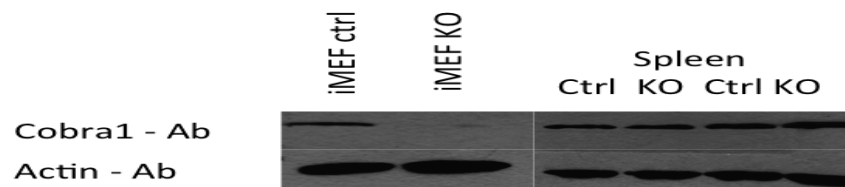


Figure.9.b **Western blotting on mouse tissues using Cobra1 anti-mouse antibody.** Spleen tissues from control and knockout animals did not show any change in Cobra1 expression levels. Actin levels were detected as a loading control. Ctrl=control & KO=Knockout. Protein lysates were run on 17% SDS-PAGE gel.

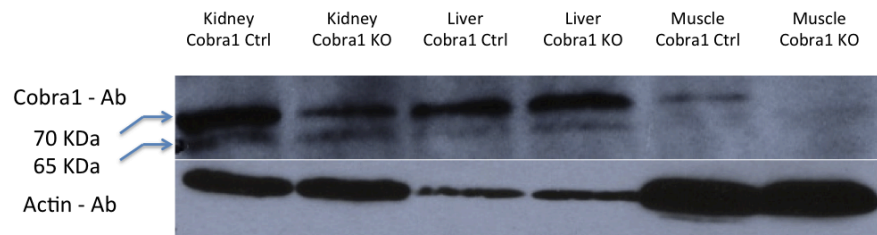


Figure.9.c **Western blotting on mouse tissues using COBRA1 anti-mouse antibody.** Comparison between patterns of expression of COBRA1 in different tissues. Actin levels between similar tissues show equal or similar loading levels. Ctrl=control & KO=Knockout. Protein lysates were run on 17% SDS-PAGE gel.



Figure.10 **Western blotting on mouse tissues using COBRA1 anti-human antibody.** Comparison between COBRA1 isoforms' expression in mouse tissues. Kidney shows an expression of a different isoform other than that expressed in spleen and testis. Very faint expression of the smaller isoform in spleen. Protein lysates were run on 10% SDS-PAGE gel.



Figure.11 **Western blotting showing Comparison of expression pattern of COBRA1 isoforms between mice of different age.** Comparison of patterns of expression of Cobra1 between old and young kidney tissues shows Cobra1 doublet. Protein lysates were run on 17% SDS-PAGE gel.

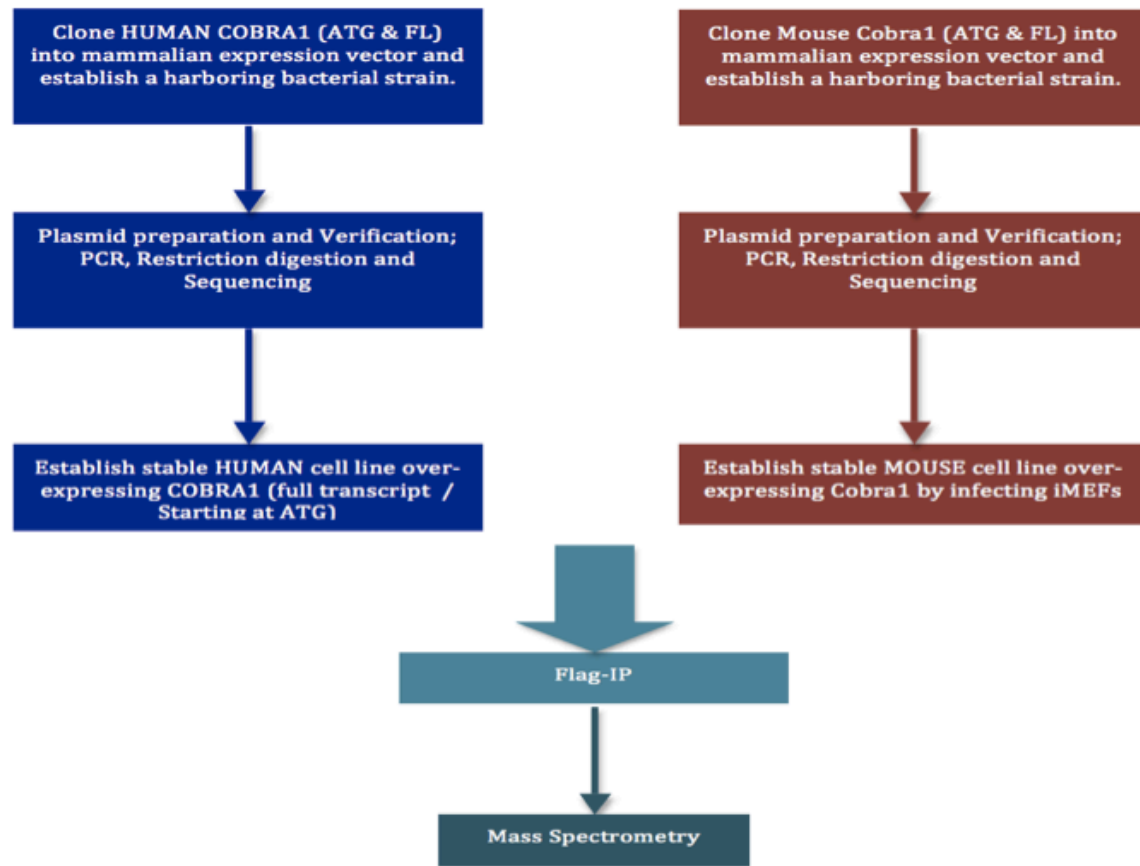


Figure.12 **Schematic diagram for *ex-vivo* expression of *COBRA1/Cobra1* in mammalian cell lines.** ATG constructs contains only the coding sequence of the gene proceeded by its endogenous Kozak sequence. FL constructs are the full transcripts generated by assembly PCR. To exclude the contribution of the expression vector pBABE to the *ex-vivo* expression system results, we used the empty vector and named its constructs and subsequent cell lines EV.

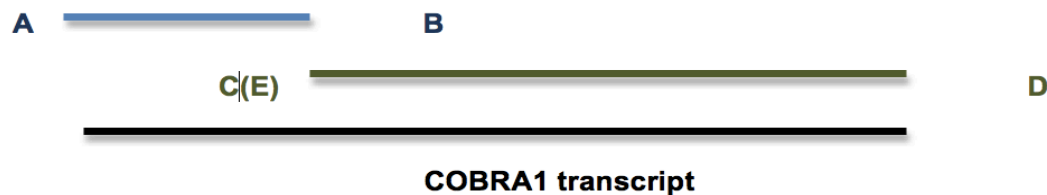


Figure.13 **Diagram for cloning fragments of *Cobra1*.** AB= 5'UTR, CD=Coding sequence for assembly PCR to generate full transcript, ED=Coding sequence preceded by recognition sites for EcoRI for short form construct and AD= full transcript.

Cloning fragments

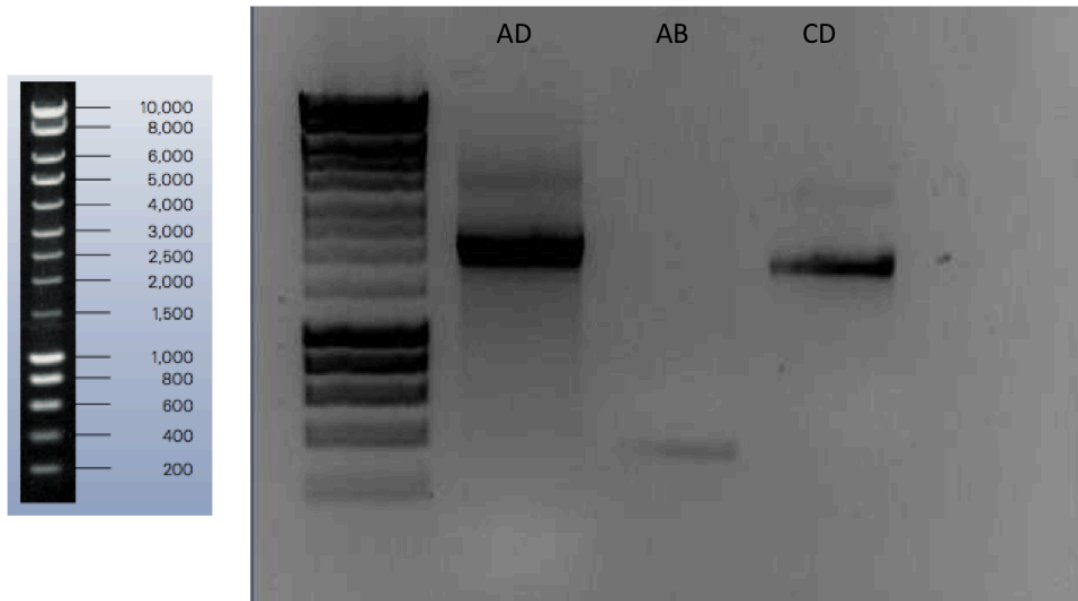


Figure.14 **Amplification of mouse cloning fragments of *Cobra1*.** Cobra1-FL (AD fragment~2200bp) was generated by overlapping fragments AB ~300bp and CD~1900bp in an assembly PCR reaction.

Plasmid verification

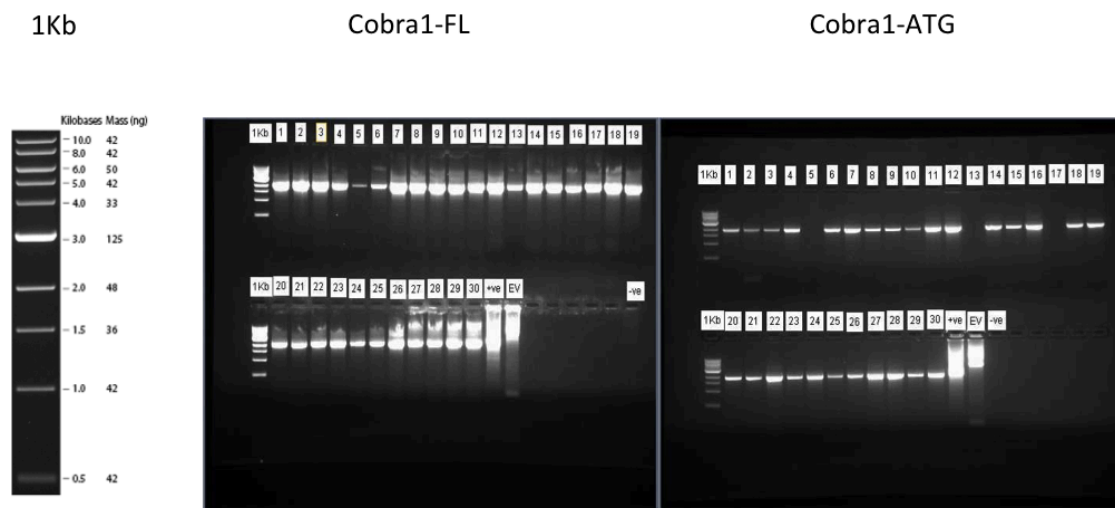


Figure.15.a **Transformants verification by colony PCR.** A Colony PCR using pBABE's LTR primers on both plasmids with COBRA1 constructs (full form~2200bp and coding sequence~2000bp) confirming correct sizes of transcripts.

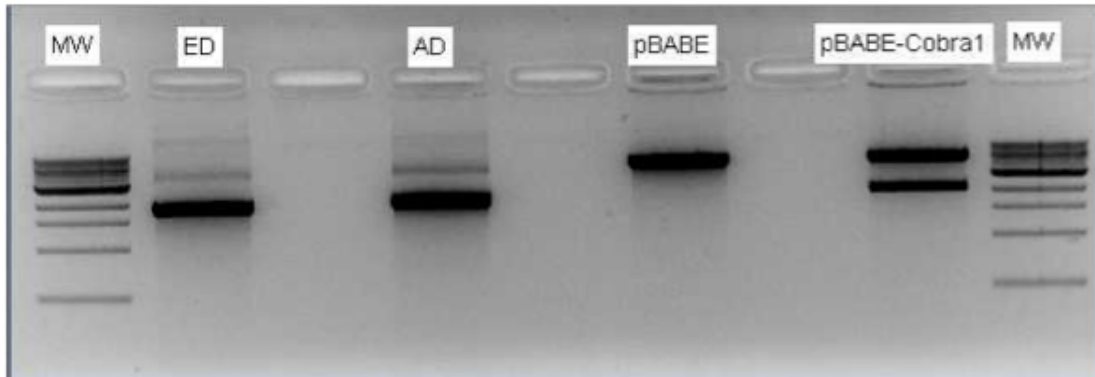


Figure.15.b **Restriction digestion of the human cloning fragments prior to cloning.** Fragment ED(~1900bp) was amplified using a primer similar to primer C but had EcoRI recognition site in it to be used in the short isoform construct(primer E). pBABE-*Cobra1* (Mouse transcript) was digested simultaneously to ensure successful and complete digestion. Restriction enzymes used were EcoRI and Sall.

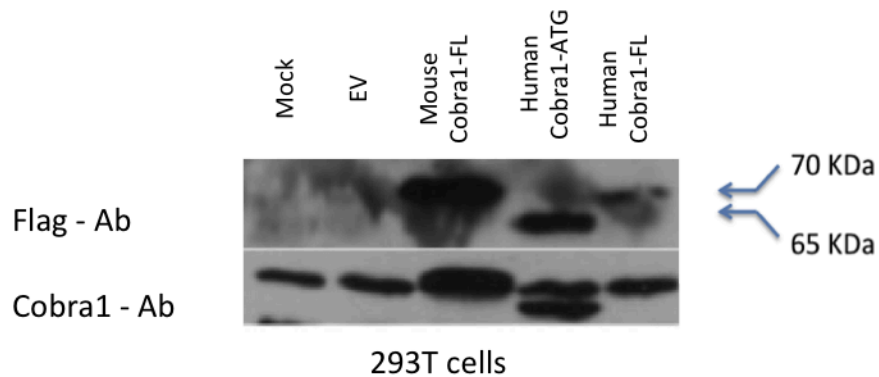


Figure.16 **Expression of human *COBRA1* constructs in HEK293T cells.** Western blotting showing altered size of COBRA1 expressed in the human cell line HEK293T. The mouse full form was used as a control for WB with Flag – Ab. Protein lysates were run on 10% SDS-PAGE gel.

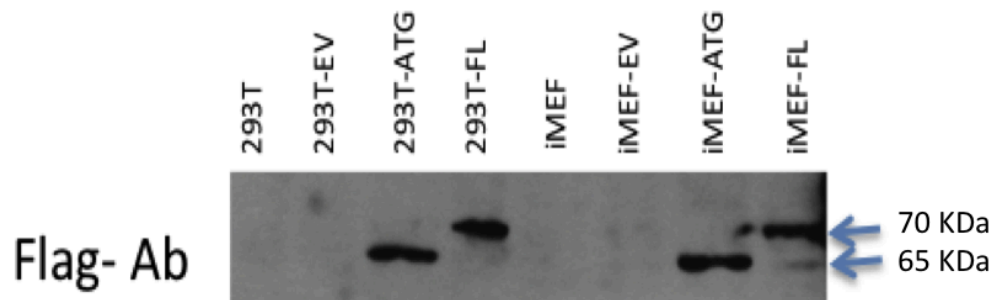


Figure.17.a **Expression of mouse *Cobra1* constructs in iMEF cells before selection.** Packaging of retroviral particles in 293T cells followed by infection of iMEF mouse cell line. The levels of expression after infection were analyzed before addition of Puromycin for selection. Protein lysates were run on 10% SDS-PAGE gel.

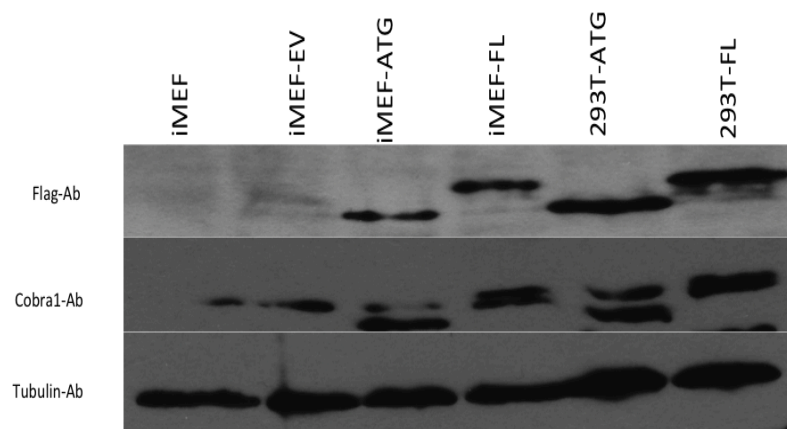


Figure.17.b **Expression of mouse *Cobra1* constructs in iMEF cells after selection.** Levels of expression of flagged COBRA1 after infection were analyzed after scaling up and addition of Puromycin for selection to ensure sustainability of expression and limited retroviral promoter silencing. Tubulin levels were detected as a loading control. Protein lysates were run on 10% SDS-PAGE gel.

- **Flag - Immunoprecipitation**

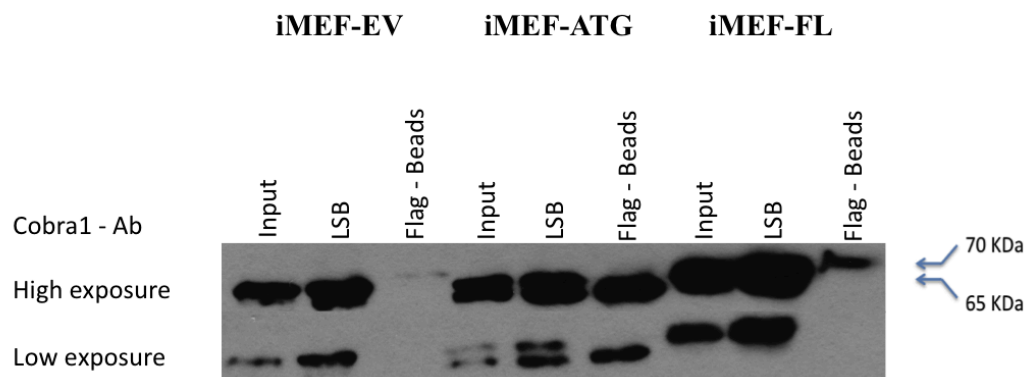


Figure. 18 **Flag immune-precipitation for mouse *Cobra1* constructs.** COBRA1 expressed flagged proteins were pulled down using Flag beads. iMEF cells were also lysed traditionally using Laemmli sample buffer (LSB) and loaded as internal control. iMEFs with *Cobra1* constructs showed positive reactivity with the flag beads in contrast to lysates of cells infected with empty vector. Protein lysates were run on 10% SDS-PAGE gel.

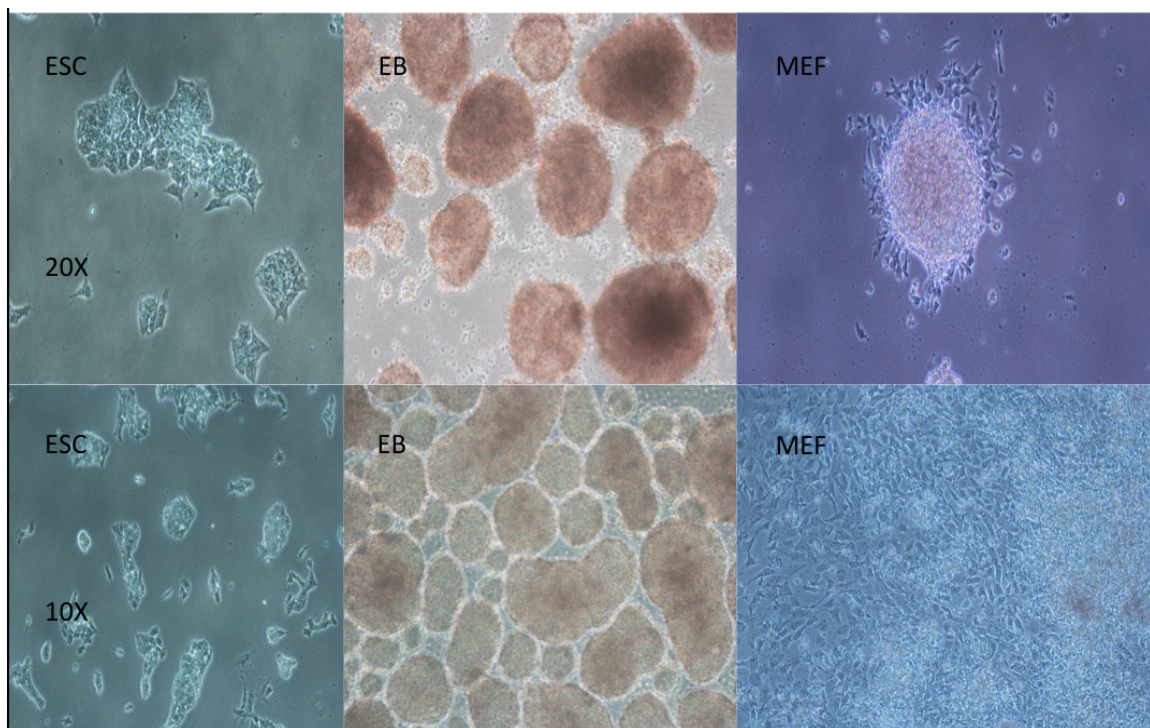


Figure.19.a **Photos for differentiation phases of ESC differentiation.** Panel 1 show pluripotent, colonized and undifferentiated AB2.2 embryonic stem cells . Cells in panel 2 are the outcome of suspending ESC for 14 days to eventually form embryoid bodies. Panel 3 show the differentiated mouse embryonic fibroblasts after culturing the aggregated EB on gelatin coated plates.

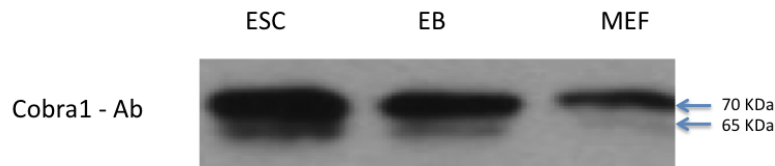


Figure.19.b **Western blotting using COBRA1 anti-mouse antibody throughout differentiation.** Western blotting analysis showing doublet of COBRA1 throughout the differentiation fate of the cells to MEF. Protein lysates were run on 17% SDS-PAGE gel.

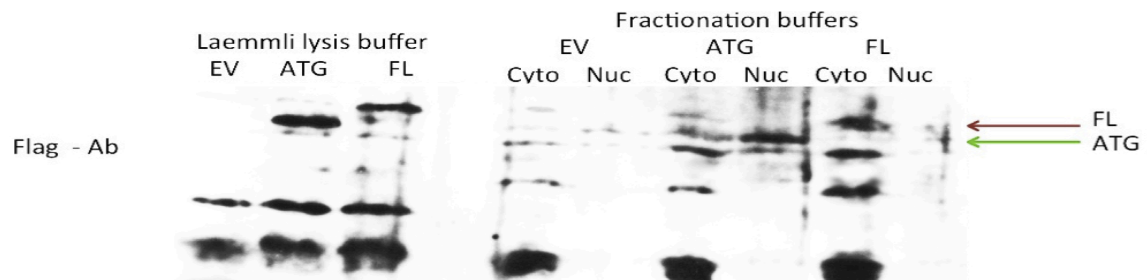


Figure.20 **Western blotting analysis for cytoplasmic and nuclear fractions of the infected cell lines with COBRA1 constructs.** Cytoplasmic fractions show higher abundance of COBRA-FL while nuclear fraction shows predominance of COBRA1-ATG. The shown result is a preliminary fractionation attempt. Although nuclear (lamin A/C) and cytoplasmic (tubulin) were needed to confirm the identity of the fractions (not done due to time and budget constraints), cytoplasmic fractions' identities from all lines can be preliminary inferred from the non-specific signals as is the case with nuclear fractionates. Protein lysates were run on 10% SDS-PAGE gel.

9. References

- [1] E. S. Lander, L. M. Linton, B. Birren *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001.
- [2] J. Thompson, E. Trevaskis, and R. Waterston, "Initial sequencing and comparative analysis of the mouse genome" *Nature*, Vol. 420, no. 6915, pp.520-562, Dec. 2002.
- [3] R. Gibbs, G. Weinstock, M. Metzker, and D. Muzny *et al.*, "Genome sequence of the Brown Norway rat yields insights into mammalian evolution" *Nature*, Vol. 428, no.6982, pp.493-521, Apr. 2004.
- [4] Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Genome Institute at Washington University, Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Genome Institute at Washington University, "A high-resolution map of human evolutionary constraint using 29 mammals," *Nature*, vol. 478, no. 7370, pp. 476–481, Oct. 2011.
- [5] R. Jaenisch and A. Bird, "Epigenetic regulation of gene expression: how the

- genome integrates intrinsic and environmental signals,” *Nat Genet*, vol. 33, no. 3, pp. 245–254, Mar. 2003.
- [6] D. A. Kleinjan and V. van Heyningen, “Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease,” *The American Journal of Human Genetics*, vol. 76, no. 1, pp. 8–32, Jan. 2005.
 - [7] R. W. Carthew and E. J. Sontheimer, “Origins and Mechanisms of miRNAs and siRNAs,” *Cell*, vol. 136, no. 4, pp. 642–655, Feb. 2009.
 - [8] B. Blencowe, “Alternative Splicing: New Insights from Global Analyses,” *Cell*, Vol.126, no.1, pp.37-47, Jul. 2006.
 - [9] S. Saxonov, P. Berg, and D. L. Brutlag, “A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 5, pp. 1412–1417, Jan. 2006.
 - [10] A. Razin and H. Cedar, “DNA methylation and gene expression,” *Microbiol. Rev.*, vol. 55, no. 3, pp. 451–458, Sep. 1991.
 - [11] A. P. Wolffe, “Transcriptional control. Sinful repression,” *Nature*, vol. 387, no. 6628, pp. 16–17, May 1997.
 - [12] J. Sun, H. Pan, C. Lei *et al.*, “Genetic and genomic analyses of RNA polymerase II-pausing factor in regulation of mammalian transcription and cell growth,” *Journal of Biological Chemistry*, vol. 286, no. 42, pp. 36248–36257, Oct. 2011.
 - [13] M. Levine, “Paused RNA polymerase II as a developmental checkpoint,” *Cell*, vol. 145, no. 4, pp. 502–511, May 2011.
 - [14] Y. Hirose and Y. Ohkuma, “Phosphorylation of the C-terminal Domain of RNA Polymerase II Plays Central Roles in the Integrated Events of Eucaryotic Gene Expression,” *Journal of Biochemistry*, vol. 141, no. 5, pp. 601–608, Feb. 2007.
 - [15] A. N. Boettiger and M. Levine, “Synchronous and stochastic patterns of gene activation in the Drosophila embryo,” *Science*, vol. 325, no. 5939, pp. 471–473, Jul. 2009.
 - [16] S. Boeing, C. Rigault, M. Heidemann, D. Eick, and M. Meisterernst, “RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion,” *Journal of Biological Chemistry*, vol. 285, no. 1, pp. 188–196, Jan. 2010.
 - [17] S. Nechaev, D. C. Fargo, G. dos Santos, L. Liu, Y. Gao, and K. Adelman, “Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila,” *Science*, vol. 327, no. 5963, pp. 335–338, Jan. 2010.
 - [18] D. R. Gallie, “The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency,” *Genes & Development*, vol. 5, no. 11, pp. 2108–2116, Nov. 1991.
 - [19] C. J. Wilusz, M. Wormington, and S. W. Peltz, “The cap-to-tail guide to mRNA turnover,” *Nat. Rev. Mol. Cell Biol.*, vol. 2, no. 4, pp. 237–246, Apr. 2001.
 - [20] D. Schmucker, J. C. Clemens, H. Shu *et al.*, “Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity,” *Cell*, vol. 101, no. 6, pp. 671–684, Jun. 2000.
 - [21] C. Lee, N. Kim, M. Roy, and B. R. Graveley, “Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod

- evolution.” *RNA*, vol. 16, no. 1, pp. 91–105, Jan. 2010.
- [22] M. Adams, S. Celniker, R. Holt *et al.*, “The Genome Sequence of *Drosophila melanogaster*,” *Science*, Vol.287, no.5461, pp. 2185–2195, Mar. 2000.
 - [23] V. Ambros, “A uniform system for microRNA annotation,” *RNA*, vol. 9, no. 3, pp. 277–279, Mar. 2003.
 - [24] Y. Tomari and P.D. Zamore, “Perspective: machines for RNAi,” *Genes & Development*, Vol.19, pp.517–529, 2005.
 - [25] G. Meister and T. Tuschl, “Mechanisms of gene silencing by double-stranded RNA,” *Nature*, vol. 431, no. 7006, pp. 343–349, Sep. 2004.
 - [26] J. M. Franco-Zorrilla, A. Valli, M. Todesco *et al.*, “Target mimicry provides a new mechanism for regulation of microRNA activity.” *Nat Genet*, vol. 39, no. 8, pp. 1033–1037, Aug. 2007.
 - [27] F. A. Karreth, Y. Tay, D. Perna, *et al.* “In vivo identification of tumor-suppressive PTEN ceRNAs in an oncogenic BRAF-induced mouse model of melanoma.” *Cell*, vol. 147, no. 2, pp. 382–395, Oct. 2011.
 - [28] M. Cesana, D. Cacchiarelli, I. Legnini, *et al.*, “A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA.” *Cell*, vol. 147, no. 2, pp. 358–369, Oct. 2011.
 - [29] I. Rubio-Somoza, D. Weigel, J. M. Franco-Zorrilla *et al.*, “ceRNAs: miRNA Target Mimic Mimics,” *Cell*, vol. 147, no. 7, pp. 1431–1432, Dec. 2011.
 - [30] M. Kisielow, S. Kleiner, M. Nagasawa, A. Faisal, and Y. Nagamine, “Isoform-specific knockdown and expression of adaptor protein ShcA using small interfering RNA,” *Biochem. J.*, vol. 363, no. 1, pp. 1–5, Apr. 2002.
 - [31] J. L. Clancy, G. H. Wei, N. Echner *et al.*, “mRNA isoform diversity can obscure detection of miRNA-mediated control of translation.” *RNA*, vol. 17, no. 6, pp. 1025–1031, Jun. 2011.
 - [32] R. K. Gaur, “RNA interference: a potential therapeutic tool for silencing splice isoforms linked to human diseases,” *BioTechniques*, vol. Suppl, pp. 15–22, Apr. 2006.
 - [33] R. Groppo and J. D. Richter, “Translational control from head to tail,” *Current opinion in cell biology*, Vol. 21, no. 3, pp. 1–8, Jun. 2009.
 - [34] T. Nakamoto, “Evolution and the universality of the mechanism of initiation of protein synthesis,” *Gene*, vol. 432, no. 1, pp. 1–6, Mar. 2009.
 - [35] I. M. Terenin, S. E. Dmitriev, D. E. Andreev, and I. N. Shatsky, “Eukaryotic translation initiation machinery can operate in a bacterial-like mode without eIF2,” *Nat Struct Mol Biol*, vol. 15, no. 8, pp. 836–841, Jul. 2008.
 - [36] E. Marti nez-Salas, D. Pineiro, and N. Fernandez, “Alternative Mechanisms to Initiate Translation in Eukaryotic mRNAs,” *Comparative and functional genomics*, Vol. 2012, pp. 1–12, Mar. 2012.
 - [37] N. Sonenberg and A. G. Hinnebusch, “Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets,” *Cell*, vol. 136, no. 4, pp. 731–745, Feb. 2009.
 - [38] J. D. Gross, N. J. Moerke, T. von der Haar *et al.*, “Ribosome loading onto the mRNA cap is driven by conformational coupling between eIF4G and eIF4E.” *Cell*, vol. 115, no. 6, pp. 739–750, Dec. 2003.
 - [39] N. Sonenberg and A. G. Hinnebusch, “Regulation of Translation Initiation in

- Eukaryotes: Mechanisms and Biological Targets,” *Cell*, vol. 136, no. 4, pp. 731–745, Feb. 2009.
- [40] M. Kozak, “The scanning model for translation: an update.” *J. Cell Biol.*, Vol. 108, pp.229-241, 1989.
 - [41] T. Pestova, J. Lorsch and C. Hellen in *Translational Control in Biology And Medicine* (Michael Mathews, Nahum Sonenberg, John W. B. Hershey, eds), pp.87-128, Cold Spring Harbor Laboratory Press, Woodbury, NY, 2007.
 - [42] N. Sonenberg and A. G. Hinnebusch, “New Modes of Translational Control in Development, Behavior, and Disease,” *Molecular Cell*, vol. 28, no. 5, pp. 721–729, Dec. 2007.
 - [43] M. Kozak, “Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs.,” *Nucleic Acids Research*, vol. 12, no. 2, pp. 857–872, Jan. 1984.
 - [44] T. Nakamoto, “Mechanisms of the initiation of protein synthesis: in reading frame binding of ribosomes to mRNA,” *Mol Biol Rep*, vol. 38, no. 2, pp. 847–855, May 2010.
 - [45] J. Pelletier and N. Sonenberg, “Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA.,” *Nature*, vol. 334, no. 6180, pp. 320–325, Jul. 1988.
 - [46] M. Stoneley and A. E. Willis, “Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression,” *Oncogene*, vol. 23, no. 18, pp. 3200–3207, Apr. 2004.
 - [47] J. Le Quesne, M. Stoneley, G. Fraser and A. Willis, “Derivation of a structural model for the c-myc IRES,” *Journal of Molecular Biology*, Vol. 310, pp.111-126, 2001.
 - [48] C. L. Jopling, K. A. Spriggs, S. A. Mitchell, M. Stoneley, and A. E. Willis, “L-Myc protein synthesis is initiated by internal ribosome entry.,” *RNA*, vol. 10, no. 2, pp. 287–298, Feb. 2004.
 - [49] E. Martínez-Salas and R. Ramos, “Functional interactions in internal translation initiation directed by viral and cellular IRES elements,” *Journal of General Virology*, Vol. 82, no. 5, pp. 973-984, May 2001.
 - [50] L. Cobbold, K. Spriggs, S. Haines *et al.* “Identification of Internal Ribosome Entry Segment (IRES)-trans-Acting Factors for the Myc Family of IRESs,” *Molecular and Cellular Biology*, Vol. 28, no. 1, 40-49, Jan. 2008.
 - [51] P. S. Ray, “La autoantigen is required for the internal ribosome entry site-mediated translation of Coxsackievirus B3 RNA,” *Nucleic Acids Research*, vol. 30, no. 20, pp. 4500–4508, Oct. 2002.
 - [52] E. Diaz-Guerra, R. Vernal, W. Cantero, E. W. Müllner, and J. A. Garcia-Sanz, “Translation controlled mRNAs: new drug targets in infectious diseases?,” *Infect Disord Drug Targets*, vol. 8, no. 4, pp. 252–261, Dec. 2008.
 - [53] D. F. Summers and J. V. Maizel, “Disaggregation of HeLa cell polysomes after infection with poliovirus.,” *Virology*, vol. 31, no. 3, pp. 550–552, Mar. 1967.
 - [54] G. A. Otto and J. D. Puglisi, “The Pathway of HCV IRES-Mediated Translation Initiation,” *Cell*, vol. 119, no. 3, pp. 369–380, Oct. 2004.
 - [55] R. Grover, M. M. Candeias, R. F. A. hraeus, and S. Das, “p53 and little brother p53/47: linking IRES activities with protein functions,” *Oncogene*, vol. 28, no.

- 30, pp. 2766–2772, Jun. 2009.
- [56] P.-J. Yu, G. Ferrari, A. C. Galloway, P. Mignatti, and G. Pintucci, “Basic fibroblast growth factor (FGF-2): The high molecular weight forms come of age,” *J. Cell. Biochem.*, vol. 100, no. 5, pp. 1100–1108, 2007.
 - [57] M. L. Shinohara, H. J. Kim, J. H. Kim, V. A. Garcia, and H. Cantor, “Alternative translation of osteopontin generates intracellular and secreted isoforms that mediate distinct biological activities in dendritic cells,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 20, p. 7235, 2008.
 - [58] X. Wang, Y. Zhao, Z. Xiao *et al.*, “Alternative Translation of OCT4 by an Internal Ribosome Entry Site and its Novel Function in Stress Response,” *Stem Cells*, vol. 27, no. 6, pp. 1265–1275, Jun. 2009.
 - [59] L. Chaussade, B. Eymin, E. Brambilla, and S. Gazzeri, “Expression of p15 and p15. 5 products in neuroendocrine lung tumours: relationship with p15 (INK4b) methylation status,” *Oncogene*, vol. 20, no. 45, p. 6587, 2001.
 - [60] R. J. Jackson, C. U. T. Hellen, and T. V. Pestova, “The mechanism of eukaryotic translation initiation and principles of its regulation,” *Nature Reviews Molecular Cell Biology*, Vol. 11, pp. 113–127, Feb. 2010.
 - [61] M. Kozak, “Adherence to the first-AUG rule when a second AUG codon follows closely upon the first,” *Proceedings of the National Academy of Sciences*, vol. 92, no. 15, pp. 7134–2666, Jul. 1995.
 - [62] C. Touriol, S. Bornes, S. Bonnal *et al.*, “Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons,” *Biology of the Cell*, vol. 95, no. 3, pp. 169–178, Jan. 2012.
 - [63] M. M. Candeias, D. J. Powell, E. Roubalova *et al.*, “Expression of p53 and p53/47 are controlled by alternative mechanisms of messenger RNA translation initiation,” *Oncogene*, vol. 25, no. 52, pp. 6936–6947, Sep. 2006.
 - [64] P. S. Ray, R. Grover, and S. Das, “Two internal ribosome entry sites mediate the translation of p53 isoforms,” *EMBO Rep.*, vol. 7, no. 4, pp. 404–410, Apr. 2006.
 - [65] J. Nichols, B. Zevnik, K. Anastassiadis *et al.*, “Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4,” *Cell*, vol. 95, no. 3, pp. 379–391, Oct. 1998.
 - [66] J. Takeda, S. Seino, and G. I. Bell, “Human Oct3 gene family: cDNA sequences, alternative splicing, gene organization, chromosomal location, and expression at low levels in adult tissues,” *Nucleic Acids Research*, vol. 20, no. 17, pp. 4613–4620, Sep. 1992.
 - [67] S. Vagner, M. C. Gensac, A. Maret *et al.*, “Alternative translation of human fibroblast growth factor 2 mRNA occurs by internal entry of ribosomes,” *Mol. Cell. Biol.*, vol. 15, no. 1, pp. 35–44, Jan. 1995.
 - [68] M. L. Shinohara, H. J. Kim, J. H. Kim, V. A. Garcia, and H. Cantor, “Alternative translation of osteopontin generates intracellular and secreted isoforms that mediate distinct biological activities in dendritic cells,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 20, pp. 7235–7239, May 2008.
 - [69] Y. Yamanaka, H. Friess, M. Buchler, H. G. Beger, E. Uchida, M. Onda, M. S. Kobrin, and M. Korc, “Overexpression of acidic and basic fibroblast growth factors in human pancreatic cancer correlates with advanced tumor stage,” *Cancer Research*, vol. 53, no. 21, pp. 5289–5296, Nov. 1993.

- [70] Y. Xu, "Regulation of p53 responses by post-translational modifications" *Cell Death and Differentiation*, Vol. 10, pp. 400-403, 2003.
- [71] Q. Ye, Y. F. Hu, H. Zhong *et al.*, "BRCA1-induced large-scale chromatin unfolding and allele-specific effects of cancer-predisposing mutations." *J. Cell Biol.*, vol. 155, no. 6, pp. 911–921, Dec. 2001.
- [72] T. Narita, Y. Yamaguchi, K. Yano *et al.*, "Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex." *Mol. Cell. Biol.*, vol. 23, no. 6, pp. 1863–1873, Mar. 2003.
- [73] S. E. Aiyar, J. Sun, A.L. Blair *et al.*, "Attenuation of estrogen receptor -mediated transcription through estrogen-stimulated recruitment of a negative elongation factor," *Genes & Development*, vol. 18, no. 17, pp. 2134–2146, Sep. 2004.
- [74] J. Sun, A. L. Blair, S. E. Aiyar, and R. Li, "Cofactor of BRCA1 modulates androgen-dependent transcription and alternative splicing," *J. Steroid Biochem. Mol. Biol.*, vol. 107, no. 3, pp. 131–139, Nov. 2007.
- [75] S. E. Aiyar, A. L. Blair, D. A. Hopkinson *et al.*, "Regulation of clustered gene expression by cofactor of BRCA1 (COBRA1) in breast cancer cells," *Oncogene*, vol. 26, no. 18, pp. 2543–2553, Oct. 2006.
- [76] J. Sun, G. Watkins, A. L. Blair, C. Moskaluk, S. Ghosh, W. G. Jiang, and R. Li, "Deregulation of cofactor of BRCA1 expression in breast cancer cells," *J. Cell. Biochem.*, vol. 103, no. 6, pp. 1798–1807, 2008.
- [77] P. A. McChesney, S. E. Aiyar, O. J. Lee *et al.*, "Cofactor of BRCA1: A Novel Transcription Factor Regulator in Upper Gastrointestinal Adenocarcinomas," *Cancer Research*, vol. 66, no. 3, pp. 1346–1353, Feb. 2006.
- [78] Y. Midorikawa, S. Tsutsumi, H. Taniguchi *et al.*, "Identification of genes associated with dedifferentiation of hepatocellular carcinoma with expression profiling analysis." *Jpn. J. Cancer Res.*, vol. 93, no. 6, pp. 636–643, Jun. 2002.
- [79] S. E. Aiyar, H. Cho, J. Lee, and R. Li, "Concerted transcriptional regulation by BRCA1 and COBRA1 in breast cancer cells.," *Int. J. Biol. Sci.*, vol. 3, no. 7, pp. 486–492, 2007.
- [80] A. Amleh, S. J. Nair, J. Sun, A. Sutherland, P. Hasty, and R. Li, "Mouse Cofactor of BRCA1 (Cobra1) Is Required for Early Embryogenesis," *PLoS ONE*, vol. 4, no. 4, pp.1-8 (e5034), Apr. 2009.
- [81] T. Stephens, C. Bunde and B. Fillmore, "Mechanism of action in thalidomide teratogenesis." *Biochemical pharmacology*, Vol. 59, no.12, Jun. 2000.
- [82] B. Maier, W. Gluba, B. Bernier *et al.*, "Modulation of mammalian life span by the short isoform of p53," *Genes & Development*, vol. 18, no. 3, pp. 306–319, Feb. 2004.
- [83] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, "Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes," *Cell*, vol. 147, no. 4, pp. 789–802, Nov. 2011.